



The statistical analysis of count data / El análisis estadístico de los datos de recuento

Joseph M. Hilbe

To cite this article: Joseph M. Hilbe (2017) The statistical analysis of count data / El análisis estadístico de los datos de recuento, *Cultura y Educación*, 29:3, 409-460, DOI: 10.1080/11356405.2017.1368162

To link to this article: <https://doi.org/10.1080/11356405.2017.1368162>



Published online: 23 Oct 2017.



Submit your article to this journal 



Article views: 6252



[View related articles](#)



[View Crossmark data](#)



Citing articles: 18 View citing articles



The statistical analysis of count data / *El análisis estadístico de los datos de recuento*

Joseph M. Hilbe

Arizona State University

(Received 23 November 2016; accepted 26 February 2017)

Abstract: This monograph provides an overview of the various regression models used to analyse count response models. We begin by defining counts and the methods used to model count data. We then discuss the basic count model — Poisson regression — focusing on the nature of equi-dispersion, which occurs when the mean and variance are identical in value. Equi-dispersion is a distributional assumption of the Poisson model. We examine how to determine when this assumption is violated, which results in extra-dispersion; i.e., either under- or overdispersion. Extra-dispersion biases the Poisson model standard errors, leading us to accept or reject a model when we should not. The negative binomial model is generally used to model generic overdispersion, but if we know the cause of the overdispersion we can select an alternative count model that appropriately adjusts for it. The same is the case with under-dispersion. Aside from looking at the Poisson and negative binomial models, we also evaluate models such as generalized Poisson, Poisson inverse Gaussian, two-part hurdle models, zero-inflated mixture models and other varieties of count model. Finally, we provide a brief look at Bayesian count models, showing how to estimate a Bayesian negative binomial model.

Keywords: count; Poisson; negative binomial; generalized Poisson; zero-inflated; hurdle

Resumen: Este artículo ofrece una descripción general de los varios modelos de regresión que se emplean en el análisis de modelos de respuesta de recuento. En primer lugar definiremos qué se entiende por recuentos y los métodos que se emplean para modelar los datos de recuento. A continuación hablaremos del modelo de recuento básico, la regresión de Poisson, centrándonos concretamente en la naturaleza de la equidispersión, la cual se produce cuando la media y la varianza tienen un valor idéntico. La equidispersión es una asunción distribucional del modelo de Poisson.

English version: pp. 409–433 / Versión en español: pp. 434–459

References / Referencias: pp. 459–460

Translated from English / Traducción del inglés: Joaquim Siles i Borràs

Corresponding address on behalf of the late Professor Hilbe / Dirección de correspondencia a nombre del difunto Profesor Hilbe: Gabriel Liberman, Data Graph – Research and Statistical Consulting, 39 Prof. Shor St., Holon, Israel.

E-mail: gabriel@data-graph.com

Examinaremos cómo determinar cuándo se infringe la asunción, lo cual resulta en una dispersión extra ya sea por infradispersión o por sobredispersión. La sobredispersión sesga los errores típicos del modelo de Poisson, llevándonos a aceptar o rechazar un modelo cuando no deberíamos hacerlo. El modelo binomial negativo suele utilizarse para modelar la sobredispersión genérica, aunque si conocemos la causa de la sobredispersión podremos seleccionar un modelo de recuento alternativo que se le ajuste adecuadamente. Lo mismo sucede con la infradispersión. Además de fijarnos en los modelos binomial negativo y de Poisson, también evaluaremos modelos como el Poisson generalizado, el Poisson-inverso gaussiano, los modelos valla de dos partes, los modelos de cero inflado mixtos y otras variaciones del modelo de recuento. Finalmente, ofreceremos un breve análisis de los modelos de recuento bayesianos a través de los cuales mostraremos cómo realizar una estimación de un modelo binomial negativo bayesiano.

Palabras clave: recuento; Poisson; binomial negativo; Poisson generalizado; cero inflado; valla

Preliminary background

Interest in the statistical modelling of count data began in earnest during the 1970s. Analysts working in the fields of insurance and transportation perhaps initiated this interest. Modelling the number of insurance claims and automobile accidents and deaths was of considerable interest to analysts at the time. Those in other fields generally preferred to model counts as if they were instances of a continuous normal or lognormal distribution. A typical strategy was to log the count variable to be modelled, perhaps adding 0.5, 0.1 or 0.01 to zero counts so that they are defined when logged (the log of 0 is undefined), and modelling the data with an ordinary least squares regression. There are still some analysts who believe that this is an acceptable manner of how to model count data — but it is not. Using least squares regression, which is based on the Gaussian or normal probability distribution, on count data violates the distributional assumptions upon which the normal model is based and produces biased statistical results. The central problem is that the normal model assumes that negative values are possible and that the variance of the variable being modelled is constant across its range of values. These assumptions are not possible for count data. Moreover, adding a small value to zero and then logging it does not necessarily result in a value close to zero. The log of 0.5 is -0.693 ; the log of 0.1 is -2.303 ; and the log of 0.01 is -4.605 . These values are not close to zero.

Count data consists of discrete non-negative counts ranging from zero to infinity. Count data is also typically right skewed with the variance defined in terms of the mean of the distribution.

Examples of count data in the context of education include statistics based on counting the number of absences in school districts or in schools, per day, week, month or academic year, counting the instances of bullying, acts of violence or other events of interest that occur in schools throughout a district, the number of

spelling errors made by each student taking a writing examination and the number of teachers or administrators at schools of various sizes. In fact, counting events that occur over different periods of time or over varying geographical areas is common when analysing count data. Count models should be used when the events, subjects or observations to be modelled are both non-negative and discrete.

The standard or paradigm count model is known as Poisson regression. Poisson regression is based on the Poisson probability distribution, which itself assumes that the observations being counted in the model are independent and not correlated in any way. A central criterion of the Poisson distribution is that mean and variance of the counts being modelled are identical. When this occurs the model is said to be equi-dispersed.

Until the early to mid-1990s the majority of researchers used Poisson regression to model count data, unless they incorrectly used some form of normal model. However, many analysts realized that the data they were modelling were not equi-dispersed; i.e., the variance of the count data being modelled exceeded its mean. The result of using a Poisson model on such data is that the standard errors of the model are biased. The explanatory predictors of the model may appear to significantly contribute to the understanding of the counts, when in fact they do not. This realization has generated a number of different types of count models, which are aimed at handling extra-dispersed count data.

In this brief overview we shall look at many of the alternative models used by researchers to model non-Poisson count data. First, however, I shall describe the Poisson model and how it is to be interpreted. The logic used in Poisson regression generally runs through to alternative count models. Keep in mind, though, that each alternative to Poisson regression is an attempt to deal with data that in some way violate the Poisson criterion of equi-dispersion.

Methods of estimation

The majority of count models we discuss in this overview are estimated using maximum likelihood. This is the standard regression method used in frequentist-based statistical modelling. When counts are clustered into groups or panels the data are many times modelled using some variety of quadrature, or at times using an Expectation-Maximization (EM) algorithm. I shall briefly touch on this class of models later in the overview. Bayesian modelling is another entirely separate way of estimating the parameters of a count model. I shall describe this class of count model as well since it will be in common use in the near future.

Almost all count models are parametric models. That is, they are based on an underlying probability distribution, which theoretically generates the data being evaluated. The distribution is characterized by parameters that specify the form taken by the data. The model attempts to estimate these parameters in as unbiased a manner as possible.

The Poisson model is based on the Poisson probability distribution, which has a single parameter, the mean. The Poisson model is also a member of the single parameter exponential family of distributions, which underlie a family of models known as Generalized Linear Models, or GLM. In fact, the Poisson model is usually found in the GLM procedure or function of a statistical package. In SAS, the Poisson model is part of the GENMOD procedure; in SPSS it is part of the GENLIN procedure; in R it is part of the *glm* function; and in Stata it is a member of the *glm* command. Stata also has a *poisson* command, which is modelled using a full maximum likelihood algorithm. GLM models are also based on maximum likelihood, but in a more simplified version (Hilbe, 2011).

Poisson regression

The Poisson probability distribution function (PDF) may be characterized as:

$$f(y; \mu) = \frac{e^{-\mu}(\mu)^y}{y!} \quad (1)$$

Where y is the variable containing the observed model counts and μ is the predicted or fitted mean of the distribution of counts. A well-fitted model is one where the observed and predicted counts are close throughout the entire range of counts.

When modelling data we are interested in determining the parameter(s) of the distribution on the basis of the adjusted response term, which for this discussion are counts. We may re-parameterize the Poisson PDF to likelihood form, given in exponential family form as:

$$\mathcal{L}(\mu; y) = \prod_{i=1}^n \exp\{y_i \log(\mu) - \mu_i - \log(y_i!)\} \quad (2)$$

By taking the natural log of both sides of the likelihood we obtain the log-likelihood function, which is used as the foundation of maximum likelihood modelling.

$$\mathcal{L}(\mu; y) = \sum_{i=1}^n \{y_i \log(\mu) - \mu_i - \log(y_i!)\} \quad (3)$$

You may recall from basic statistics that the linear predictor of a normal or linear regression model is symbolized as xb . It is the same as the fitted or predicted value of a linear model. The formula representing this relationship is given as:

$$xb = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n \quad (4)$$

Where the betas are coefficients and X terms are predictor values. The first term, β_0 , is the intercept. However, for nearly all of the count models we discuss here the predicted value differs from the linear predictor. In order to calculate the predicted mean, the linear predictor must be transformed to log (μ) form.

$$\ln(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n \quad (5)$$

The fitted value, μ , may then be determined by taking the exponentiation of both sides of the equation.

$$\mu = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n} \quad (6)$$

or

$$\mu = \exp(xb) \quad (7)$$

The linear relationship of the predictors to the fit is through the natural log of μ . $\ln(\mu)$ is referred to as the Poisson link function, and $\exp(xb)$ the inverse link, which defines μ .

Most maximum likelihood software estimates parameters in terms of β . These are the parameters being estimated. The re-parameterized Poisson log-likelihood function therefore takes the form:

$$\mathcal{L}(\beta; y) = \sum_{i=1}^n \{y_i (x'_i \beta) - \exp(x'_i \beta) - \ln\Gamma(y_i + 1)\} \quad (8)$$

The maximum likelihood estimates of β can be determined by taking the first derivative of (8) with respect to β , setting the result to 0 and solving.

$$\frac{\partial(\mathcal{L}(\beta; y))}{\partial \beta} = \sum_{i=1}^n (y_i - \exp(x'_i \beta)) x'_i \quad (9)$$

We shall use educational data from a sociological study authored by Quine (1973) and later re-evaluated in Aitkin (1978). Australian aboriginee (1) and white (0) children in four age groups were classified as normal (0) or slow (1) learners. The count response consists of the days a student was absent from school during the academic year. Age groups and how they correspond to year in school are displayed in the table below. The predictors *slow*, *aborig* and *girl* are binary with 1 = variable name; 0 = other category. Stata is used for the analysis.

```
. use absenteeism
. su
```

Variable	Obs	Mean	Std. Dev.	Min	Max
days	146	16.4589	16.25322	0	81
age	146	2.541096	1.03808	1	4
slow	146	.4315068	.4969914	0	1
aborig	146	.4726027	.5009674	0	1
girl	146	.5479452	.4994092	0	1
id	146	73.5	42.29066	1	146

age	Freq.	Percent	Cum.	
F0	27	18.49	18.49	last yr primary school (8th gr)
F1	46	31.51	50.00	1st yr secondary school (9th gr)
F2	40	27.40	77.40	2nd yr secondary school (10th gr)
F3	33	22.60	100.00	3rd yr secondary school (11th gr)
Total	146	100.00		

The data can be modelled using Stata's *poisson* or *glm* commands. I prefer *glm* since it allows ready access to a host of ancillary residuals and diagnostics. *Poisson* can be used as well for other tests as desired. Only the *glm* command displays the Pearson dispersion statistic, which is necessary to determine if there is possible extradispersion in the data. The 'i.' prefix to the variable *age* factors the levels of *age*, providing the first level as the default reference level.

. glm days i.age slow aborig girl, fam(poi) nolog	
Generalized linear models	No. of obs = 146
Optimization : ML	Residual df = 139
	Scale parameter = 1
Deviance = 1696.706552	(1/df) Deviance = 12.20652
Pearson = 1830.191125	(1/df) Pearson = 13.16684
Variance function: V(u) = u	[Poisson]
Link function : g(u) = ln(u)	[Log]
Log likelihood = -1142.591815	AIC = 15.74783
	BIC = 1003.985
days	OIM
	Coef. Std. Err. z P> z [95% Conf. Interval]
age	
F1 -.3339014 .0700935 -4.76 0.000 -.4712821 -.1965206	
F2 .2578284 .0624194 4.13 0.000 .1354886 .3801681	
F3 .4276938 .0676864 6.32 0.000 .295031 .5603567	
slow .348943 .0520431 6.70 0.000 .2469403 .4509456	
aborig .5336043 .0418831 12.74 0.000 .4515149 .6156937	
girl -.1615966 .0425346 -3.80 0.000 -.2449628 -.0782304	
_cons 2.343372 .0603757 38.81 0.000 2.225038 2.461707	
. abic	
AIC Statistic = 15.76153	AIC*n = 2301.1836
BIC Statistic = 15.87983	BIC(Stata) = 2325.0525

The Pearson dispersion is displayed as 13.16684. The statistic is defined as the Pearson Chi² statistic divided by the residual degrees of freedom. Simulation studies (Hilbe, 2011) demonstrate that this dispersion is preferred to using the deviance dispersion, which is biased. If a Poisson model is equi-dispersed the Pearson dispersion statistic has a value of 1.0. Values greater than 1.0 are termed overdispersed; values under 1.0 are under-dispersed. Extra-dispersed data refers to data that are not equi-dispersed; i.e., that are either under- or overdispersed.

Note that all of the predictors appear to significantly contribute to the model; i.e., they contribute to the understanding of days absent. However, the fact that the dispersion is so high gives evidence that the data are highly correlated — overdispersed — and that the variance exceeds the mean. Excess correlation many times is a result of the data being clustered; e.g., modelling students across schools where teaching methods, demographic profile and so forth may differ between schools. Student results (whatever is being tested) may then be more similar within each school than between schools. This produces overdispersion. If test results are clumped into various ranges of results, if the data are highly skewed, if there is a needed interaction term, if there are excessive zero counts in the data, or at times if no zero counts are possible — all of these situations, and others, can give rise to overdispersion in the data. When the data are over- or underdispersed, the Poisson model standard errors need adjustment. If we know the cause of extra-dispersion, or if the extra-dispersion is high, a count model other than Poisson will likely need to be employed.

A simple summary of days supports the assessment that the variance of the days absent is greater than the mean. Again, this is an essential assumption for using the Poisson model on count data.

```
. su days
Variable |       Obs        Mean      Std. Dev.       Min       Max
-----+-----+-----+-----+-----+-----+-----+
  days |     146    16.4589    16.25322         0        81
. di r(sd)^2
264.16726
```

The observed variance is some twenty times greater than the mean of *days*. Stata users can learn how statistics are stored following a simple descriptive test by typing *return list* and following a model by typing *ereturn list*. Here I use *r(sd)* for the saved standard deviation of days. In any case, when the variance of count variable being modelled is substantially greater than its mean, the variable is overdispersed. Explanatory predictors may ameliorate or even eliminate the overdispersion, which can in general be evaluated by checking the value of the dispersion statistic.

The other components of the count model output above are the *z*-statistic, *p*-value, and confidence intervals. It is important to understand how these are defined.

The *z*-statistic is the ratio of the predictor coefficient and standard error. For the predictor *slow* we have:

```
di .348943 /.0520431
6.704885
```

or using saved estimation symbols for the coefficient (beta) and standard error:

```
. di _b[slow] / _se[slow]
6.7048791
```

The *z*-statistic is assumed to follow a cumulative standard normal distribution. This is unlike the normal or least squares model that uses a *t*-statistic. The probability for *slow* is calculated as $2 * N(-|z|)$, where N is a cumulative standard normal function and z is the calculated *z*-statistic.

```
. di %9.8f normal(-abs(_b[slow]/_se[slow]))*2
0.00000000
```

Confidence intervals are based on a significance level. If we use a 95% confidence interval, the significance is based on the two-sided inverse cumulative standard normal distribution:

```
. di invnorm(0.975)
1.959964
```

The confidence intervals for *slow* are therefore determined using the formulae:

```
. di _b[slow] - invnorm(0.975) * _se[slow]
.24694028
. di _b[slow] + invnorm(0.975) * _se[slow]
.45094564
```

Note that these values correspond to the model output displayed above. It should be mentioned that use of a *p*-value to define the statistical significance of a predictor is now being widely discouraged by statisticians and social scientists. The preferred method is to check the predictor confidence intervals. If the confidence interval includes zero within its range, the predictor is considered not significant. A significant predictor does not contain zero within its confidence interval.

Modifying Poisson standard errors

The problem with this model is that the standard errors are biased, and this leads an analyst to believe that the predictors are significant when they may not be. R users will remodel the data with the *glm* function using what is termed the quasipoisson distribution in place of Poisson. This is identical to scaling the standard errors by multiplying the given model standard errors by the square root of the Pearson dispersion. What this technique does is to adjust the standard errors to what they would be if the dispersion statistic were 1.0. From about 1973 until the mid 1990s this method was the most popular method of adjusting an extra-dispersed Poisson model.

Calculation of scaled standard errors occurs after estimation of the model. The software, whether it is Stata, SAS, SPSS or R, estimates the standard errors, calculates the adjustment and runs one more iteration with the adjustment displayed. For the predictor *slow*, scaled standard errors are calculated as:

SE OF SLOW

```
. di _se[slow]
.05204314
```

SQUARE ROOT OF THE PEARSON DISPERSION STATISTIC

```
. di sqrt(e(dispers_p))
3.6286144
```

SCALED STANDARD ERROR

```
. di _se[slow] * sqrt(e(dispers_p))
.18884449
```

In Stata, the *scale(x2)* option in *glm* produces scaled standard errors. The *nohead* option directs the software to block displaying header statistics. *nolog* inhibits the display of the iteration log.

```
. glm days i.age slow aborig girl, fam(poi) nolog nohead scale(x2)
```

days	OIM					
	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
age						
F1	-.3339014	.2543423	-1.31	0.189	-.8324031	.1646003
F2	.2578284	.2264959	1.14	0.255	-.1860955	.7017522
F3	.4276938	.2456077	1.74	0.082	-.0536885	.9090762
slow	.348943	.1888445	1.85	0.065	-.0211854	.7190714
aborig	.5336043	.1519776	3.51	0.000	.2357336	.831475
girl	-.1615966	.1543415	-1.05	0.295	-.4641004	.1409072
_cons	2.343372	.2190801	10.70	0.000	1.913983	2.772761

(Standard errors scaled using square root of Pearson X2-based dispersion.)

Note that the scaled standard error for *slow* is the same as we calculated by hand. R users may type the following for the identical results:

```
> mypoi <- glm(days ~ factor(age)+slow+aborig+girl,family=quasipoisson, data=absent)
> summary(mypoi)
```

The *z*, *p*-value and confidence intervals must also be adjusted when scaling standard errors. The following code will obtain the correct results.

P-VALUE

```
. di %9.8f normal(-abs(_b[slow])/_se[slow])*2
0.06371483
```

SCALED CI

```
. di _b[slow] - invnorm(0.975) * _se[slow]
-.02118543
. di _b[slow] + invnorm(0.975) * _se[slow]
.71907136
```

Perhaps the most popular and advised method to adjust standard errors when there is evidence of extra-dispersion is to use some type of sandwich variance estimator. There are various methods of calculating what are called robust, sandwich or Huber-White standard errors. See Hardin and Hilbe (2012) for examples. Applying robust standard errors to the above model results in:

POISSON – ROBUST SE

```
. glm days i.age slow aborig girl, fam(poi) nolog nohead vce(ro)
```

days	Robust					
	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
age						
F1	-.3339014	.2675844	-1.25	0.212	-.8583571	.1905543
F2	.2578284	.2505294	1.03	0.303	-.2332002	.7488569
F3	.4276938	.2485347	1.72	0.085	-.0594252	.9148129
slow	.348943	.1881936	1.85	0.064	-.0199097	.7177957
aborig	.5336043	.1538529	3.47	0.001	.2320582	.8351504
girl	-.1615966	.1555387	-1.04	0.299	-.4664468	.1432536
_cons	2.343372	.2433279	9.63	0.000	1.866459	2.820286

Poisson rate ratio

In a similar manner to how analysts use odds ratios with logistic models, the exponentiation of count model coefficients results in *incidence rate ratios*. The statistic relates to the rate at which counts occur for one level of a predictor compared to another level. The *eform* option exponentiates coefficients and appropriately adjusts other associated statistics. Significance of a predictor should be evaluated on whether the confidence interval contains the value of one. If the confidence interval does not include one, the predictor may be considered significant, but only if the other distributional assumptions of the model have been satisfied, e.g., equidispersion and independence of observations.

```
. glm days i.age slow aborig girl, fam(poi) nolog eform nohead
```

days	OIM					
	IRR	Std. Err.	z	P> z	[95% Conf. Interval]	
age						
F1	.7161244	.0501957	-4.76	0.000	.6242015	.8215844
F2	1.294117	.080778	4.13	0.000	1.145096	1.46253
F3	1.533716	.1038117	6.32	0.000	1.343168	1.751297
slow	1.417568	.0737747	6.70	0.000	1.280103	1.569796
aborig	1.705067	.0714135	12.74	0.000	1.57069	1.85094
girl	.8507844	.0361877	-3.80	0.000	.7827337	.9247513
_cons	10.41631	.6288915	38.81	0.000	9.253838	11.7248

The interpretation of the rate ratios can be given as:

- Students in the first year of secondary school are absent some 28% less often than are last-year primary students. $(1 - .716) * 100$
- Students in the second year of secondary school are absent some 30% more often than are last year primary students. $(.29 * 100)$
- Students in the third year of secondary school are absent some 53% more often than are last-year primary students. $(.53 * 100)$
- Slow learners are absent some 42% more often than average learners. $(.418 * 100)$

- Students with an aborigine ethnic background are absent more than 70% more often than students who are ethnically identified as white. (.705 * 100)
- Girls are absent some 15% less than are boys. (1 - .85)* 100

Note that when interpreting a predictor, we assume that the other predictors are held at a constant value.

Researchers rarely interpret count model coefficients. When doing so the reference is to log-counts, which are difficult to interpret. Researchers generally use standard count models, such as the Poisson, to predict counts given predictor profiles. The rate ratio parameterization is used when predictors are being interpreted in terms of the differential percentage of counts for levels in a predictor.

Many researchers error when they calculate the standard errors and confidence intervals of risk ratios. The so-called delta method should be used for calculating standard errors. Essentially, for risk ratios, standard errors are calculated by multiplying the model standard error by the risk ratio. Confidence intervals are not based on this standard error, but rather are the exponentiation of the model confidence intervals. Compare the calculations below with the statistical output above. Not all software provides these values.

RATE RATIO STANDARD ERROR

```
. di exp(_b[slow])*_se[slow]
.07377471
```

RATE RATIO CONFIDENCE INTERVAL

```
. di exp(_b[slow] - invnorm(0.975) * _se[slow])
1.2801027
. di exp(_b[slow] + invnorm(0.975) * _se[slow])
1.569796
```

Parameterization by area or time

Poisson models are often used to analyse count data that are collected over different periods of time or in different areas. When this is desired the Poisson model is adjusted to have an underlying PDF appearing as:

$$f(y; \mu) = \frac{e^{-t\mu} (t\mu)^y}{y!} \quad (10)$$

Where t is a variable indicating the differences in time over which each observation is taken, or the differences in area. To model the data one enters the time or area variable as an offset. It is important, however, to log the offset variable before entering it into the model. The reason is that the offset is added to the linear predictor, which is internally logged by the estimation algorithm. The offset variable must be put into the same metric as the other terms in the linear predictor. Note, though, that the coefficient of the logged offset is set as one. It enters the linear predictor as a constant.

For an example, consider data on days sick by teachers at various schools in the city of Phoenix, Arizona, in 2015. Sickdays is the count response, with predictors gender (female = 1; male = 0), white (teacher identified as racially white = 1; nonwhite = 0), tenured (yes = 1; no = 0) and involved in extra-curricular activities, e.g., coaching, (yes = 1, no = 0). The schools in the data were of varying sizes, ranging from a small charter school of 45 to a large school of 3,353 students. The number of teachers was not recorded for these data. The counts of days sick are adjusted by the size of school, which is entered into the model as a logged offset. Using Stata the data can be modelled as:

```
. glm sickdays gender white tenured extra, exposure(schoolsize) fam(poi) nolog vce(ro)
eform

Generalized linear models
Optimization : ML
No. of obs      =      27
Residual df    =      22
Scale parameter =      1
Deviance       =  42.89127063
(1/df) Deviance =  1.949603
Pearson        =  41.54589787
(1/df) Pearson =  1.88845

Variance function: V(u) = u          [Poisson]
Link function   : g(u) = ln(u)      [Log]

AIC            =  4.208466
BIC            = -29.61714
Log pseudolikelihood = -51.81429589

-----+
           | Robust
sickdays |   IRR   Std. Err.      z   P>|z|   [95% Conf. Interval]
-----+
  gender |  1.137691  .3173939   0.46  0.644   .6584996  1.965591
  white |  4.685963  2.118087   3.42  0.001   1.932191  11.36443
tenured |  5.492599  1.777975   5.26  0.000   2.912331  10.35893
  extra |  3.53922   1.556543   2.87  0.004   1.494688  8.380399
  _cons |  .0008261  .0002398  -24.45  0.000   .0004676  .0014593
ln(school~e) |          1 (exposure)
-----+
```

The gender of the teacher was not a significant predictor. Teachers identifying themselves as white were some four and a half times more likely to be sick than nonwhite teachers. Tenured teachers were some five and a half times more likely to be sick than non-tenured teachers, and teachers who were involved in extra-curricular activities were some three and a half times more likely to be sick than those teachers who were not involved in extra-curricular activities. Remember that for each predictor, the other predictors are held at a constant value, otherwise there would be an interaction effect. The model does appear to be somewhat over-dispersed (1.888). Since the standard errors have been adjusted by a robust sandwich variance estimator, it may be wise to model the data using another count model. Thirty percent of the teachers had no sick days at all, which violates the Poisson distributional assumption that only 2.2% of the teachers have no sick days (given a mean of 3.815 sick days per school). Refer to the section on zero inflated models for more information regarding this assumption and how to deal with it.

Model evaluation

The Poisson model can best be evaluated as to its goodness of fit by assessing the dispersion statistic as discussed earlier, and by testing the difference between the

observed and predicted counts. The smaller the difference between the two across the range of counts, the better the fit. A Chi² test with a degree of freedom equal to the number of counts being compared can be used to determine the significance of the fit at the .05 level. Values of the probability greater than .05 traditionally indicate a well-fitted model, however one must be wary of interpreting *p*-values too strictly. The test results should simply be reported as displayed without interpretation of significance unless the Chi² statistic is substantially high. The test below indicates a relatively poor fit, which corresponds to the high dispersion in the data. A graphic of the difference between observed and predicted counts can be instructional in visually assessing fit.

```
. chi2gof, cells(15) table
```

Chi-square Goodness-of-Fit Test for Poisson Model:

```
Chi-square chi2(14) = 70.52
Prob>chi2 = 0.00
```

Cells	Abs.	Freq.	Fitted			Abs.	Dif.
			Rel.	Freq.	Rel. Freq.		
0	9		.0616		9.7e-05	.0615	
1	4		.0274		6.7e-04	.0267	
2	6		.0411		.0024	.0387	
3	6		.0411		.0059	.0352	
4	1		.0068		.0115	.0047	
5	19		.1301		.0187	.1114	
6	8		.0548		.0267	.0281	
7	6		.0411		.0347	.0064	
8	4		.0274		.0419	.0145	
9	2		.0137		.0477	.034	
10	4		.0274		.0521	.0247	
11	7		.0479		.0549	.007	
12	2		.0137		.0563	.0426	
13	3		.0205		.0563	.0358	
14 or more	57		.4452		.59	.1448	

Negative binomial regression

The data we modelled above using a Poisson model are highly overdispersed. The standard errors are biased. Statisticians generally turn to negative binomial regression when faced with modelling overdispersed Poisson data. The negative binomial is a mixture of the Poisson and gamma distributions, with an ancillary parameter, called the dispersion parameter, adjusting for overdispersion. The negative binomial can only be used on overdispersed count data. However, almost all real count data are in fact overdispersed. Therefore the negative binomial model has become a central model in the evaluation of count data.

The negative binomial probability function is expressed in a variety of ways. The definition of the distribution used in Stata, SAS and SPSS software is given as:

$$f(y; \mu, \alpha) = \binom{y_i + \frac{1}{\alpha} - 1}{\frac{1}{\alpha} - 1} \left(\frac{1}{1 + \alpha\mu_i} \right)^{\frac{1}{\alpha}} \left(\frac{\alpha\mu_i}{1 + \alpha\mu_i} \right)^{y_i} \quad (11)$$

The log-likelihood function for the above PDF may be expressed in exponential mean format as:

$$\begin{aligned} \mathcal{L}(\mu; y, \alpha) = & \sum_{i=1}^n y_i \log \left(\frac{\alpha\mu_i}{1 + \alpha\mu_i} \right) - \frac{1}{\alpha} \log(1 + \alpha\mu_i) + \log\Gamma \left(y_i + \frac{1}{\alpha} \right) \\ & - \log\Gamma(y_i + 1) - \log\Gamma \left(\frac{1}{\alpha} \right) \end{aligned}$$

With the $\log\Gamma(*)$ function symbolizing the log-gamma function. Recall that the first term in the negative binomial PDF is a choose function which can be converted to factorial form. Factorials can themselves be transformed to gamma functions. When the PDF is parameterized into log-likelihood form, the gamma functions become log-gamma functions (Hilbe, 2011).

The mean of the negative binomial model is μ , as is the Poisson mean. However, the variance is:

$$V(\mu) = \mu + \alpha\mu^2 \quad (12)$$

In comparison to the Poisson variance of μ . The variance can be interpreted as a mixture of the Poisson variance, μ , and the gamma variance, μ^2/γ . However, the model is parameterized so that there is a direct relationship of the mean and amount of dispersion in the data. The dispersion parameter is therefore established as $\alpha = 1/\gamma$. This parameterization is currently used in Stata, SAS, SPSS, Genstat, Limdep and in all commercial statistical software. R users must be cautioned, though, since the default *glm.nb* and *glm* functions indirectly parameterize the dispersion parameter, calling the dispersion parameter theta, θ . For these R functions, the negative binomial variance is $\mu + \mu^2/\theta$. Note that when $\alpha = 0$, the negative binomial is Poisson. Greater values of α in a negative binomial model indicate that more Poisson overdispersion was adjusted. For R, though, a model is Poisson when θ is infinitely high. Low values of θ indicate high dispersion or correlation in the data. However, the *gamlss* package in R, which provides R users with a host of modelling capabilities, uses the direct parameterization, α . The parameterization of the negative binomial dispersion parameter does not have a bearing on the model coefficients or standard errors. Care must be taken, though, to specify the software or type of dispersion used when reporting negative binomial study results in publications.

We next model the example data using a negative binomial model.

```
. glm days i.age slow aborig girl, fam(nb ml) nolog

Generalized linear models
Optimization : ML
No. of obs = 146
Residual df = 139
Scale parameter = 1
Deviance = 167.9518008
Pearson = 137.7760369
(1/df) Deviance = 1.208286
(1/df) Pearson = .9911945

Variance function: V(u) = u + (.7844)u^2 [Neg. Binomial]
Link function : g(u) = ln(u) [Log]

AIC = 7.583226
BIC = -524.7695
Log likelihood = -546.5755091
-----
```

days	OIM					
	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
age						
F1	-.4484282	.2376006	-1.89	0.059	-.9141168	.0172605
F2	.0880802	.2415453	0.36	0.715	-.3853399	.5615002
F3	.356901	.2466194	1.45	0.148	-.1264641	.8402661
slow	.2921092	.1829335	1.60	0.110	-.066434	.6506523
aborig	.5693717	.1576082	3.61	0.000	.2604654	.878278
girl	-.0823203	.164681	-0.50	0.617	-.4050892	.2404486
_cons	2.407529	.2088926	11.53	0.000	1.998107	2.816951

```
. abic
AIC Statistic = 7.596925 AIC*n = 1109.151
BIC Statistic = 7.71522 BIC(Stata) = 1133.0199
```

The Akaike Information Criterion (*AIC*) and Bayesian Information Criterion (*BIC*) statistics are substantially lower than for the Poisson model, which were 2,301 and 2,325 respectively. Models on the same data having lower *AIC* and *BIC* values are considered as comparatively better fitted models. Usually, if the values differ by more than 10, we can be tentatively confident that the model with the lower values for *AIC* and *BIC* is to be preferred. There are a number of different *AIC* and *BIC* statistics that have been developed by statisticians, but the ones used in Stata, SAS, SPSS and R are the same. Note though that SAS refers to the *BIC* statistic as the Schwarz Criterion (SC) and is provided in most model output.

The Pearson dispersion statistic for the above negative binomial model is .99, very close to 1.0. This indicates that the negative binomial has fully accommodated the Poisson overdispersion. The standard errors displayed in the model are fairly consistent with the values given in the scaled and robust Poisson models. The model coefficients and rate ratios are interpreted in the same manner as for a Poisson model.

When modelling the categorical predictor, *age*, we have found that the levels of *age* are not significant with reference to the lower level, called F1. However, by re-structuring the age levels so that the highest level is modelled against all of the other levels, the resulting binary predictor is significant. I call the new binary predictor *a4*, with the reference level being the combined levels 1–3 of *age*. The code to create *a4* is given as:

```
. gen byte a4=age==4
```

A full maximum likelihood negative binomial model of the adjusted data with robust standard errors and coefficients parameterized as rate ratios is given below.

```
. nbreg days a4 aborig girl slow, nolog vce(ro) irr

Negative binomial regression                               Number of obs      =      146
                                                       Wald chi2(4)      =     23.77
Dispersion      = mean                                Prob > chi2       =    0.0001
Log pseudolikelihood = -550.00597                    Pseudo R2        =    0.0163
-----
|          Robust
days |      IRR   Std. Err.      z   P>|z| [95% Conf. Interval]
-----+
a4 |  1.581041 .3126757   2.32  0.021  1.073012  2.329601
aborig |  1.817805 .2756562   3.94  0.000  1.350421  2.446953
girl |   .8055  .1268699  -1.37  0.170  .5915582  1.096816
slow |  1.312772 .2367492   1.51  0.131  .9218941  1.86938
_cons | 10.55794  1.82515  13.63  0.000  7.523712 14.81584
-----+
/lnalpha | -.1929157  .125174                   -.4382523  .0524209
-----+
alpha |   .8245515  .1032124                   .645163   1.053819
-----
. abic
AIC Statistic = 7.61652          AIC*n      = 1112.012
BIC Statistic = 7.681596         BIC(Stata) = 1129.9136
```

The *AIC* and *BIC* statistics for this model are a bit greater than for the model having all levels of *age*. However, the difference in the statistics is not significant. Note that the predictors *girl* and *slow* have confidence intervals including the value 1, indicating that they do not significantly contribute to the model or the understanding of days absent from school. If these are dropped from the model the *AIC* and *BIC* values become smaller. The *AIC* statistic is 0.3 smaller, but the *BIC* statistic is 6.3 smaller, indicating a possible better fit. The fact that there are only 146 observations in the model lead us to be cautious about giving too much weight to small drops in *AIC* and *BIC* statistic values.

Heterogeneous negative binomial

We saw evidence of rather substantial overdispersion in the data when we modelled them using Poisson regression. The negative binomial model, however, appeared to properly adjust for it. However, we can still check to determine which, if any, predictors are contributing to any excess dispersion in the negative binomial model. The model with *a4* above has a dispersion statistic of 1.009295 (not shown) so we should not expect contributing predictors. However, most negative binomial models do not completely adjust for all of the extra-dispersion in the data.

We can determine which predictors add to extra-dispersion using a heterogeneous negative binomial model.

```
. gnbreg days a4 aborig slow girl , nolog lnalpha(a4 aborig slow girl)

Generalized negative binomial regression
Number of obs = 146
LR chi2(4) = 16.65
Prob > chi2 = 0.0023
Pseudo R2 = 0.0150
Log likelihood = -547.99619
-----
```

	days	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
days	a4	.3906054	.2133284	1.83	0.067	-.0275107 .8087215
	aborig	.5861604	.1616776	3.63	0.000	.2692781 .9030427
	slow	.2931937	.1768329	1.66	0.097	-.0533924 .6397797
	girl	-.174187	.1599756	-1.09	0.276	-.4877334 .1393593
	_cons	2.345858	.1732254	13.54	0.000	2.006342 2.685373
lnalpha	a4	.280441	.3567383	0.79	0.432	-.4187531 .9796352
	aborig	-.4404478	.258165	-1.71	0.088	-.946442 .0655464
	slow	.3046848	.2938908	1.04	0.300	-.2713306 .8807003
	girl	-.0665591	.261742	-0.25	0.799	-.579564 .4464458
	_cons	-.1484573	.2755071	-0.54	0.590	-.6884414 .3915267

Significant predictors in the *lnalpha* component of the model contribute to extra-dispersion. As suspected, none of the model predictors are significant. If any predictors were, though, we would know that they should be checked to determine how they might contribute to extra-dispersion.

Zero inflated models

Many count models have an excess of zero counts. The Poisson and negative binomial distributions underlying well-fitted models assume that the observed number of counts in the model are close to what is predicted based on the distributional mean. The mean of *days* is 16.4589. The number of 0 counts predicted for the Poisson model is therefore 0.

POISSON: PREDICTED NUMBER OF DAYS WITH 0 COUNTS

```
. di exp(-16.4589)*16.4589^0 * 146
.00001038
```

For the above negative binomial model of the data we must also know the value of the model dispersion parameter. Here it is 0.8245515. The product of the mean and *alpha* is 13.571211. With that information we may calculate the predicted zero using the NB PDF:

```
. di exp(0* log( 13.571211/(1+ 13.571211)) - (1/.8245515) * log(1+ 13.571211) /*
*/ + lngamma(0 +1/.8245515) /*
*/ - lngamma(0+1) - lngamma(1/.8245515))
.03880924
```

NEGATIVE BINOMIAL: PREDICTED NUMBER OF DAYS WITH 0 COUNTS

```
. di 146 * .03880924
5.666149
```

There are 5.7 predicted zero counts for the negative binomial model. There are in fact nine zero counts in the data, as observed in the section on model evaluation. There is not a wide discrepancy in the two values, so likely a zero-inflated model is not necessary.

A zero-inflated model is a mixture model that models the zero counts as a binary model — usually a logistic regression — and all of the counts using a count model. The mixture occurs in that both the binary and count components of the model estimate zero counts. The main point when using a zero-inflated model on data is to adjust for overdispersion in the data caused by excessive zero counts.

```
zinb days a4 aborig slow girl, nolog inflate(days a4 aborig slow girl) irr

Zero-inflated negative binomial regression          Number of obs      =      146
                                                Nonzero obs       =      137
                                                Zero obs        =         9

Inflation model = logit                          LR chi2(4)        =     17.91
Log likelihood   = -545.8493                     Prob > chi2       =    0.0013

-----+
days |      IRR   Std. Err.      z   P>|z|   [95% Conf. Interval]
-----+
days |
  a4 |  1.527242  .3082863    2.10  0.036   1.028219  2.268453
  aborig |  1.69585  .2661596    3.37  0.001   1.24679  2.306649
  girl |  .7499481  .1152076   -1.87  0.061   .5549693  1.013429
  slow |  1.305235  .2339429    1.49  0.137   .9185955  1.854613
  _cons |  11.90538  2.003941   14.72  0.000   8.559869  16.55844
-----+
inflate |
  a4 | -.3847747  1.526356   -0.25  0.801   -3.376378  2.606829
  aborig | -21.21131  17819.8   -0.00  0.999   -34947.38  34904.95
  girl | -3.407821  11.99498   -0.28  0.776   -26.91755  20.10191
  slow | -.1653036  1.8618   -0.09  0.929   -3.814365  3.483757
  _cons | -1.660992  .9083664   -1.83  0.067   -3.441357  .1193733
-----+
/lnalpha |  -.3533463  .1556277   -2.27  0.023   -.6583709  -.0483216
-----+
alpha |   .702334  .1093026                               .517694  .9528273
-----+
Likelihood-ratio test of alpha=0: chibar2(01) =  1104.35 Pr>=chibar2 =  0.0000
Vuong test of zinb vs. standard negative binomial: z =  1.30  Pr>z =  0.0976

. abic
AIC Statistic =  7.628073          AIC*n      = 1113.6987
BIC Statistic =  7.838715          BIC(Stata) = 1146.5183
```

The *AIC* and *BIC* statistics are nearly identical to the standard negative binomial model — in fact, the zero-inflated model has a bit greater *AIC* and *BIC* values. If there were excessive zero COUNTS this would not be the case.

Notice also the two tests under the model output. The likelihood ratio test determines if the zero-inflated negative binomial (ZINB) model is a significantly better fit than the zero-inflated Poisson (ZIP) model. *p*-values under .05 indicate that the ZINB model is preferred.

The Vuong test determines if the inflated model is preferred to the non-inflated model. The test statistic is normally distributed, $N(0,1)$, with large positive values favouring the inflated model; large negative values favour the non-inflated model. A *p*-value of .05 specifies that the preferred model is significantly better fit than

the other. Here there is no difference in the ZINB and NB models in terms of fit. We already suspected this to be the case.

Two-part hurdle count models

Like zero-inflated models, hurdle count models are generally used to model data with excessive zero counts. The model consists of two separable parts. Positive counts are modelled using a zero truncated model. The binary component models the data with 1 equal to counts greater than zero and 0 as zero. Usually the binary component uses a logistic or probit model, but this is not necessary. See Hilbe (2014) for a full discussion. Recent studies on the topic indicate that some hurdle models may be superior to zero-inflated models when used to adjust for excessive zero counts in the data.

A primary difference in how zero-inflated and hurdle models are used rests in the fact that the count component of a hurdle model is a zero-truncated model. Zeros are completely evaluated using the binary component. With a zero-inflated model zeros are included in both the count and binary components. Remember also that when interpreting a zero inflated model the binary component models zero whereas for the hurdle model the binary component models one, which consists of all counts in the data greater than zero. The binary component of the hurdle model therefore models positive counts compared to zero counts.

The model below is a negative binomial-logit hurdle model with robust standard errors. It is recommended that robust adjustments be used for all count models as a default. If the data are not extra-dispersed, the standard errors and associated statistics reduce to the regular model standard errors. I have shown variations for pedagogical purposes.

Negative Binomial – Logit Hurdle Model

```

. hnblogit days a4 aborig slow girl, nolog vce(ro)

Negative Binomial-Logit Hurdle Regression           Number of obs      =      146
                                                Wald chi2(4)       =      6.31
Log pseudolikelihood = -546.21056                Prob > chi2      =    0.1770

-----+-----| Robust
-----+-----| Coef.   Std. Err.      z   P>|z|   [95% Conf. Interval]
logit      | 
    a4 | -.1764771   .9443589   -0.19   0.852   -2.027387   1.674432
  aborig |  2.076402   1.089484   1.91   0.057   -.0589462   4.211751
    slow | -.2278658   .7658108   -0.30   0.766   -1.728827   1.273096
    girl |  1.004074   .7218531   1.39   0.164   -.4107321   2.41888
   _cons |  1.842387   .7199687   2.56   0.010   .4312742   3.2535

-----+-----| negbinomial
-----+-----| 
    a4 | .4660117   .192477   2.42   0.015   .0887637   .8432598
  aborig | .525538   .1518729   3.46   0.001   .2278726   .8232033
    slow | .298715   .1819296   1.64   0.101   -.0578605   .6552905
    girl | -.2796566   .1564645   -1.79   0.074   -.5863214   .0270081
   _cons |  2.444588   .1703575   14.35   0.000   2.110694   2.778483

-----+-----| /lnalpha | 
-----+-----| 
   alpha | .7185961   .0992408   .5481896   .9419741

-----+-----| . abic
AIC Statistic      =  7.633021          AIC*n        = 1114.4211
BIC Statistic      =  7.843663          BIC(Stata)  = 1147.2408

```

Given that the negative binomial model has apparently adjusted for the overdispersion in the data, we have found that the zero-inflated model is not preferred to the non-inflated model. Likewise, we would not expect that the above hurdle model would be better fit than the NEGATIVE binomial model we used on the data. The *AIC* and *BIC* statistics confirm this expectation.

Alternative count models

There are a number of other count models that have been developed by statisticians to model data that may not be appropriate for Poisson or negative binomial models. The generalized Poisson (GP) and Poisson inverse Gaussian (PIG) models are prime examples. Each of the models discussed below were authored and published by James Hardin (Univ. of South Carolina) and myself (Hardin & Hilbe, 2012).

Generalized Poisson

The generalized Poisson (GP) model is a mixture of Poisson distributions. Unlike the negative binomial, the generalized Poisson can model underdispersed Poisson data. Negative values of *delta*, the GP dispersion parameter, indicate adjustment for Poisson underdispersed data.

```
. gpoisson days a4 aborig slow girl, nolog irr vce(ro)

Generalized Poisson regression                               Number of obs = 146
                                                               Wald chi2(4) = 19.45
Prob > chi2 = 0.0006                                         = 0.0006
Dispersion = .7659908                                         Prob > chi2 = 0.0006
Log pseudolikelihood= -552.06227                           Pseudo R2 = 0.0169
-----
|          Robust
days |      IRR   Std. Err.      z    P>|z| [95% Conf. Interval]
-----+
a4 |    1.21606   .253484    0.94   0.348   .8082105   1.829723
aborig |   1.712981   .2406284   3.83   0.000   1.300713   2.255921
slow |     1.1055   .1671772   0.66   0.507   .821935   1.486895
girl |    .8738746   .1269624  -0.93   0.353   .657326   1.161763
_cons |   12.09324   1.905104  15.82   0.000   8.880741   16.46783
-----+
/tanhdelta |   1.010553   .0474949   .9174651   1.103642
-----+
delta |    .7659908   .0196276   .724696   .8018033
-----+
Likelihood-ratio test of delta=0:  chi2(1) = 1293.28      Prob>=chi2 = 0.0000

. abic
AIC Statistic = 7.644689                                AIC*n = 1116.1245
BIC Statistic = 7.709765                                BIC(Stata) = 1134.0262
```

The likelihood ratio test tells us that the GP model is preferred to the Poisson model. However, the *AIC* and *BIC* statistics inform us that the model is no better than the negative binomial used before.

Poisson inverse Gaussian

The Poisson inverse Gaussian (PIG) model is a mixture of Poisson and inverse Gaussian distributions. The modelling algorithm is recursive, and is considerably complex. Only a few software implementations of the model exist.

```
. pigreg days a4 aborig slow girl, nolog irr vce(ro)

Poisson-Inverse Gaussian regression
Number of obs      =      146
Wald chi2(4)      =     2041.09
Log pseudolikelihood = -556.07488
Prob > chi2        =     0.0000

-----
          |      Robust
days |   IRR  Std. Err.      z  P>|z|  [95% Conf. Interval]
-----+
a4 |  1.39857  .0297321  15.78  0.000  1.341494  1.458075
aborig |  2.022934  .052767  27.01  0.000  1.922111  2.129044
slow |  1.186465  .0196503  10.32  0.000  1.14857  1.225611
girl |  .8164273  .0112542 -14.71  0.000  .7946648  .8387858
_cons | 10.73001  .4199487  60.63  0.000  9.937705  11.58549
-----+
/lnalpha |  .2366043  .1886122           -.1330689  .6062774
-----+
alpha |  1.26694  .2389603           .8754048  1.833593
-----+
. abic
AIC Statistic    =  7.699656          AIC*n       = 1124.1498
BIC Statistic    =  7.764732          BIC(Stata) = 1142.0514
```

The *p*-values of the rate ratios, given robust standard errors, are all significant. However, the *AIC* and *BIC* statistics indicate a possible poor fit compared to the negative binomial. Dropping *girl* from the model results in a four-point drop in the *BIC* statistic. On the other hand, to determine the comparative fit of the PIG model to the NB one should compare the differences in observed vs predicted counts across the range of data.

It should be noted that both the GP and PIG models can also be estimated as zero-inflated models. In this case, though, no improvement is made from the non-inflated models.

Three parameter models

A number of three-parameter negative binomial models exist, giving the models a greater range for estimating the parameters of the underlying distribution of the model. Models are available in Stata for the NB-P model, which parameterizes the exponent of the second term in the variance function, $\mu + \alpha\mu^\rho$ (Hardin & Hilbe, 2014), the Waring negative binomial and Famoye negative binomial models (Harris, Hilbe, & Hardin, 2014). The P in the NB-P model represents that the power of the second term in the variance function is a parameter to be estimated. However, the traditional symbol for the parameter is *p*, or rho, and not P.

The key use of three-parameter count models is in modelling negative binomial data that are otherwise extra-dispersed. Usually three-parameter models are used to adjust for data that are highly overdispersed; i.e., data having excess

variability in the data that cannot be handled by the standard negative binomial model, or by alternative two parameter models. But three-parameter count models have been designed to be used for a variety of purposes.

Hierarchical count models

The full complement of random intercept, random slopes and multi-level count models exist in Stata, SAS and R. They are used to adjust for overdispersion resulting in the data being structured in panels or in some longitudinal manner. Poisson and negative binomial GEE models are also available in these packages, as well as in SPSS.

Other count models

Statisticians have also created generalized additive and quantile count models for non-parametric count data, exact Poisson regression for small unbalanced count data, finite mixture models for counts that are generated from two or more separate sources, and truncated and censored count models (Hardin & Hilbe, 2015). Each of these models is discussed in Hilbe (2011).

Bayesian count models

Bayesian methodology is fast becoming popular among statisticians and researchers in a wide variety of disciplines. Before the early years of this century, the great majority of researchers engaged in Bayesian modelling were forced to solve Bayesian posteriors using analytic methods. This could be very difficult, and typically relegated modelling to rather simple tasks. However, beginning several years later various software packages began to offer Bayesian modelling based on Markov Chain Monte Carlo (*MCMC*), which allows estimation of the Bayesian posterior distribution by means of a remarkable sampling algorithm. JAGS, WinBUGS, OpenBUGS, Stan, Python, MLwiN, SAS, SPSS, Stata and other packages have become popular, and have expanded Bayesian modelling throughout the research community.

WinBUGS was the first comprehensive Markov Chain Monte Carlo (*MCMC*) sampling package for Bayesian modelling, released in 2000. Other similar packages followed. Stata offered its first Bayesian package in 2015.

Bayesian modelling can pretty much duplicate maximum likelihood models when vague, diffuse or so-called non-informative priors are used to mix with the model likelihood. Unlike frequentist-based maximum likelihood modelling, Bayesian model parameters are considered to be randomly distributed. A sampling algorithm is used to determine the posterior distribution of each parameter in the model.

I shall demonstrate the use of Stata's *bayesmh* command to estimate the posterior parameters of the negative binomial using non-informative or flat priors.

The top model is a standard negative binomial model; the Bayesian model of the data follows.

MLE NEGATIVE BINOMIAL

```
. nbreg days a4 aborig girl slow, nolog vce(ro)
```

days	Robust					
	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
a4	.4580832	.1977657	2.32	0.021	.0704694	.8456969
aborig	.59763	.1516423	3.94	0.000	.3004165	.8948435
slow	.2721409	.180343	1.51	0.131	-.0813249	.6256066
girl	-.2162921	.1575045	-1.37	0.170	-.5249952	.092411
_cons	2.356878	.1728699	13.63	0.000	2.01806	2.695697
/lnalpha	-.1929157	.125174			-.4382523	.0524209
alpha	.8245515	.1032124			.645163	1.053819

BAYESIAN NEGATIVE BINOMIAL

bayesmh days, nocons likelihood(llf(///
ln(nbinomialp(exp(-{lnalpha}), days,	///
1/(1+exp({days: a4 aborig slow girl}+{days:_cons}))))))	///
prior({days:a4 aborig slow girl _cons}{lnalpha}, flat) rseed(1944)	
Bayesian regression	MCMC iterations = 12,500
Random-walk Metropolis-Hastings sampling	Burn-in = 2,500
	MCMC sample size = 10,000
	Number of obs = 146
	Acceptance rate = .1822
	Efficiency: min = .03343
	Equal-tailed
	Mean Std. Dev. MCSE Median [95% Cred. Interval]
lnalpha -.1520669 .1209699 .006491 -.1578817 -.3710215 .1024264	
days	
a4 .4510724 .2278986 .013421 .4435438 .0004725 .8943773	
aborig .5889158 .1644661 .01003 .582773 .2859969 .9122398	
slow .2566763 .1899522 .011947 .2546729 -.1253877 .626903	
girl -.2139218 .1642201 .006388 -.2112521 -.5433723 .1210683	
_cons 2.232879 .2258881 .011743 2.225916 1.795334 2.715911	

```
. di exp(-.1520669 )
.85893082
```

The means of the respective posterior distributions are close to the maximum likelihood parameter estimates. The Deviance Information Criterion (*DIC*) statistic is displayed in the table below. *DIC* is similar to the *AIC* and *BIC* statistics, but it is used specifically for Bayesian models. A listing of predictor effective sample sizes (*ESS*), correlation times and efficiencies for each of the model parameters are displayed in the second table below. *ESS* statistics represent the effective sample size for each parameter, and are compared to the *MCMC* sampling size of the model, which here is 10,000. The posteriors are determined on the basis of the *ESS*. Ideally the closer the *ESS* value is to the *MCMC* sample size the better.

Correlation times are the inverse of the efficiencies, which are listed in the right-most column. In general, lower correlation times and corresponding higher efficiency values are preferred. Efficiencies above 10% are considered particularly good. For our model the highest is *aborig* with an efficiency of 5.9%. Note that running the model again with a different seed will result in sometimes rather substantial differences in efficiencies. I recommend resampling with more iterations, and perhaps having a *burnin* sample size that is greater than the default. *Burnin* values are sampling iterations that are not used in calculating the posterior distribution. This is important since early iterations likely differ considerably from the underlying distribution and can bias results.

```
. bayesstats ic
```

Bayesian information criteria

	DIC	log (ML)	log (BF)
active	1111.725	-556.4729	.

Note: Marginal likelihood (ML) is computed
using Laplace-Metropolis approximation.

```
. bayesstats ess
```

Efficiency summaries MCMC sample size = 10,000

	ESS	Corr. time	Efficiency
lnalpha	334.28	29.91	0.0334
days			
a4	351.39	28.46	0.0351
aborig	590.44	16.94	0.0590
slow	390.83	25.59	0.0391
girl	539.99	18.52	0.0540
_cons	523.93	19.09	0.0524

Finally, graphical displays of the shape of each posterior parameter, as well as trace plots, are important diagnostics. A Stata command to display diagnostic graphs for *aborig* and the dispersion parameter, *lnalpha*, is given below. However, the output is not shown here.

```
. bayesgraph diagnostics {days: aborig} {lnalpha} , histopts(normal)
```

R's *MCMCpack* family of functions provides R users with a number of Bayesian models. Most R users, however, employ JAGS or Stan from within the R environment. This gives users much more modelling power than by using JAGS and Stan in standalone form. SAS's Genmod and *MCMC* procedures provide excellent SAS Bayesian modelling tools. INLA, or Integrated Nested Laplace Approximation, is a new Bayesian method which is optimal for the Bayesian modelling of hierarchical and spatial data. The field is new and growing,

and can only benefit those who wish to engage in the sophisticated modelling of count data.

Summary

This overview of modelling count data provides only a basic look at how count models are developed and interpreted. When faced with count data to model, it is important to first assess the mean and range of the count variable being modelled and to determine its dispersion status. Selecting model predictors is done in the standard manner. Dispersion can be most easily determined by observing the Pearson dispersion statistic of a Poisson model of the data. Values of the dispersion that are substantially greater than 1.0 indicate that the data are overdispersed. In this case it is likely better to model the data using a negative binomial model. If the data are under-dispersed, a generalized Poisson or hurdle model can be used for parameter estimation. If the model displays no extra-dispersion, then a Poisson model is likely preferred.

When the data have more zero counts than are acceptable based on the distributional assumptions of the model, then a zero-inflated or hurdle model should be tried. Diagnostic tests tell the analyst if the zero-inflated model is better fitted than the non-inflated standard model. Moreover, tests comparing the observed vs predicted counts across the range of counts in the model should be used to assess fit. *AIC* and *BIC* statistics can be used to select the best fitted model. However, it is important to check various fit tests rather than only relying on one.

At times neither a Poisson nor negative binomial model fits the data. Other models can be used; e.g., generalized Poisson, Poisson inverse Gaussian or one of the three-parameter count models. Other models exist as well, including some we did not mention in this overview. I recommend that analysts become familiar with the various count models and select the one that best fits their study data. More in-depth discussion of the full range of count models with worked-out examples can be found in Hilbe (2011) and Hilbe (2014). Examples are provided in Stata, R and SAS.

El análisis estadístico de los datos de recuento

Contexto preliminar

Fue en la década de 1970 cuando el modelado estadístico de los datos de recuento empezó a generar un auténtico interés, especialmente entre los analistas que trabajaban en los campos de los seguros y el transporte, los cuales en aquellos momentos mostraban un especial interés en modelar tanto la cantidad de reclamaciones de seguros como de muertes y accidentes de automóvil. Aquellos que trabajaban en otros campos preferían, por norma general, modelar recuentos como si se tratara de casos de distribución continua o log-normal. La estrategia habitual era calcular el logaritmo de la variable de recuento que debía modelarse, añadiendo quizás 0.5, 0.1 o 0.01 a los valores cero, con el fin de definirlos para el logaritmo (el logaritmo de 0 es indefinido) y modelar los datos con una regresión de mínimos cuadrados común. Aunque todavía hay analistas que creen que esta es una manera válida de modelar los datos de recuento, no lo es. El uso de una regresión de mínimos cuadrados en los datos de recuento, basada en la distribución gaussiana o de probabilidad normal, infringe las asunciones distribucionales sobre las que está basado el modelo normal, y produce resultados estadísticos sesgados. El problema principal es que el modelo normal asume que los valores negativos son posibles y que la varianza de la variable que debe modelarse es constante en todo su rango de valores. Estas asunciones no son posibles en los datos de recuento. Es más, añadir un valor a cero y, a continuación, obtener su logaritmo no siempre resulta en un valor cercano a cero. El logaritmo de 0.5 es -0.693 , el logaritmo de 0.1 es -2.303 y el logaritmo de 0.01 es -4.605 . Estos valores no son cercanos a cero.

Los datos de recuento consisten en recuentos discretos no negativos que van de cero a infinito. Los datos de recuento también suelen mostrar una asimetría a la derecha con una varianza definida en términos de la media de la distribución. Entre los ejemplos de datos de recuento en el contexto de la educación se incluye estadísticas basadas en el recuento del número de ausencias en distritos escolares o en colegios por día, semana, mes o año académico, el recuento de casos de acoso escolar, actos violentos u otros acontecimientos de interés que se produzcan en los colegios de todo un distrito, el número de errores ortográficos que comete un estudiante cuando realiza un examen escrito y el número de profesores o administradores en colegios de varios tamaños. De hecho, el recuento de los acontecimientos que se producen a lo largo de diferentes períodos de tiempo, o en áreas geográficas distintas, es una práctica común cuando se analizan datos de recuento. Los modelos de recuento deberían usarse cuando los acontecimientos, temas u observaciones que deban modelarse sean discretos y no negativos.

El modelo de recuento estándar o paradigmático recibe el nombre de regresión de Poisson. La regresión de Poisson está basada en la distribución de probabilidad de Poisson, la cual asume que las observaciones que son objeto del recuento en el modelo son independientes y en ningún caso están correlacionadas. Uno de los criterios centrales de la distribución de Poisson es que la media y la varianza de los recuentos que son objeto de la modelación son idénticas. Cuando este es el caso, se dice que el modelo es equidisperso.

Hasta la mitad de la década de 1990, la mayoría de los investigadores utilizaba la regresión de Poisson para modelar datos de recuento, salvo en aquellos casos en que, de manera incorrecta, se utilizaba algún tipo de modelo normal. No obstante, muchos analistas se dieron cuenta de que los datos que estaban modelando no eran equidispersos; eso es, que la varianza de los datos de recuento que eran modelados excedía su media. El resultado de usar un modelo de Poisson en dichos datos es que los errores típicos del modelo están sesgados. Aunque pueda parecer que los predictores explicativos del modelo contribuyen de manera significativa a la comprensión de los recuentos, de hecho no es así. Esto ha generado varios tipos diferentes de modelos de recuento que pretenden gestionar de manera eficiente los datos de recuento sobredispersos.

Esta breve presentación está dedicada a los modelos alternativos que los investigadores utilizan para modelar los datos de recuento que no tienen una distribución de Poisson. No obstante, en primer lugar describiré el modelo de Poisson y cómo debe interpretarse. Por norma general, la lógica empleada en la regresión de Poisson es el hilo conductor de los modelos de recuento alternativos. Téngase en cuenta que todas las alternativas a la regresión de Poisson son un intento de tratar aquellos datos que de una u otra manera infringen el criterio de equidispersión de Poisson.

Los métodos de estimación

La mayoría de los modelos de recuento a los que se hace referencia en esta descripción general se han estimado utilizando la máxima verosimilitud. Este es el método de regresión estándar que se emplea en el modelado estadístico de base frecuentista. En muchas ocasiones, cuando los recuentos están aglomerados en grupos o paneles, los datos se modelan utilizando alguna variedad de cuadratura o, a veces, un algoritmo esperanza-maximización (EM). Más adelante me referiré brevemente a esta clase de modelos. La modelación bayesiana es otro modo totalmente diferente de realizar estimaciones de parámetros de un modelo de recuento. También describiré esta clase de modelo de recuento, dado que en el futuro será de uso común.

La mayoría de los modelos de recuento son modelos paramétricos. Eso es, están basados en una distribución de probabilidad subyacente que, en principio, genera los datos que se están evaluando. La distribución se caracteriza mediante parámetros que especifican la forma que toman los datos. El modelo intenta estimar estos parámetros de la manera menos sesgada posible.

El modelo de Poisson está basado en la distribución de probabilidad de Poisson, la cual cuenta con un solo parámetro: la media. El modelo de Poisson es también miembro de la familia exponencial de distribuciones de parámetro único que sirve

de base a una familia de modelos conocidos por el nombre de Modelos Lineales Generalizados (GLM, según sus siglas en inglés). De hecho, el modelo de Poisson suele encontrarse en la función o el procedimiento GLM de un paquete estadístico. En SAS, el modelo de Poisson forma parte del procedimiento GENMOD, en SPSS forma parte de GENLIN, en R forma parte de la función *glm* y en Stata es miembro del comando *glm*. Stata también tiene un comando *poisson* que se modela empleando un algoritmo de verosimilitud máxima. Los modelos GLM también están basados en la máxima verosimilitud, aunque en una versión mucho más simplificada (Hilbe, 2011).

La regresión de Poisson

La función de distribución de probabilidad de Poisson (PDF) puede caracterizarse del siguiente modo:

$$f(y; \mu) = \frac{e^{-\mu}(\mu)^y}{y!} \quad (1)$$

siendo y la variable que contiene los recuentos de modelo observados y μ la media esperada o ajustada de la distribución de recuentos. Un modelo bien ajustado es aquel en el que los recuentos observados y esperados están los unos cerca de los otros a lo largo de todo el rango de recuentos.

Cuando modelamos datos, lo que nos interesa es determinar el parámetro o los parámetros de la distribución en base al término de respuesta ajustado, que en nuestro estudio son los recuentos. Podemos reparametrizar la PDF de Poisson en forma de verosimilitud, la cual se representa en forma de familia exponencial como:

$$\mathcal{L}(\mu; y) = \prod_{i=1}^n \exp\{y_i \log(\mu) - \mu_i - \log(y_i!)\} \quad (2)$$

Tomar el log natural de ambos lados de la verosimilitud nos permite obtener la función log-verosimilitud, la cual se emplea como fundamento de la modelación de máxima verosimilitud.

$$\mathcal{L}(\mu; y) = \sum_{i=1}^n \{y_i \log(\mu) - \mu_i - \log(y_i!)\} \quad (3)$$

Como el lector recordará, en estadística básica el predictor lineal de un modelo de regresión normal o lineal se simboliza con las letras xb , que son las mismas que el valor ajustado o esperado de un modelo lineal. La fórmula que representa esta relación es la siguiente:

$$xb = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n \quad (4)$$

siendo las betas los coeficientes y los términos X los valores del predictor. El primer término, β_0 , es el intercepto. No obstante, en casi todos los modelos de recuento que hemos visto aquí, el valor esperado difiere del predictor lineal. A fin de calcular la media esperada, el predictor lineal debe transformarse y adoptar la forma $\log(\mu)$.

$$\ln(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n \quad (5)$$

El valor ajustado, μ , puede a su vez determinarse tomándose la potenciación de ambos lados de la ecuación.

$$\mu = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n} \quad (6)$$

o

$$\mu = \exp(xb) \quad (7)$$

La relación lineal de los predictores con el ajuste es a través del log natural de μ . Se denomina a $\log(\mu)$ como la función de enlace de Poisson y a $\exp(xb)$ como el enlace inverso que define a μ .

La mayoría del software de máxima verosimilitud estima parámetros en términos de β . Estos son los parámetros que son objeto de estimación, por lo cual la función de log-verosimilitud de Poisson reparametrizada se representa como:

$$\mathcal{L}(\beta; y) = \sum_{i=1}^n \{y_i (x'_i \beta) - \exp(x'_i \beta) - \ln\Gamma(y_i + 1)\} \quad (8)$$

Las estimaciones de verosimilitud máxima de β pueden determinarse tomando la primera derivada de (8) con respecto a β , ajustando el resultado a 0 y resolviéndolo.

$$\frac{\partial(\mathcal{L}(\beta; y))}{\partial \beta} = \sum_{i=1}^n (y_i - \exp(x'_i \beta)) x'_i \quad (9)$$

Utilizaremos datos educativos de un estudio sociológico realizado por Quine (1973) y posteriormente reevaluado en Aitkin (1978). Los participantes del estudio, niños aborígenes australianos (1) y de origen caucásico (0), fueron clasificados en cuatro grupos etarios como aprendices normales (0) o lentos (1). La respuesta de recuento eran los días que un alumno no asistió al colegio durante el año académico. En la siguiente tabla se muestran los grupos etarios y cómo se corresponden al curso escolar en el que se encuentran. Los predictores *slow* (lento), *aborig* (aborigen) y *girl* (niña) son predictores binarios en los que 1 = nombre variable y 0 = otra categoría. El análisis se realiza con Stata.

```
. use absenteeism
. su
```

Variable	Obs	Mean	Std. Dev.	Min	Max
days	146	16.4589	16.25322	0	81
age	146	2.541096	1.03808	1	4
slow	146	.4315068	.4969914	0	1
aborig	146	.4726027	.5009674	0	1
girl	146	.5479452	.4994092	0	1
id	146	73.5	42.29066	1	146

age	Freq.	Percent	Cum.	
F0	27	18.49	18.49	last yr primary school (8th gr)
F1	46	31.51	50.00	1st yr secondary school (9th gr)
F2	40	27.40	77.40	2nd yr secondary school (10th gr)
F3	33	22.60	100.00	3rd yr secondary school (11th gr)
Total	146	100.00		

Los datos pueden modelarse mediante el uso de los comandos *poisson* o *glm* de Stata. Personalmente prefiero *glm* porque permite un fácil acceso a una gran cantidad de residuos auxiliares y diagnósticos. *Poisson* también puede usarse en otras pruebas según se deseé. Solo el comando *glm* muestra la estadística de dispersión de Pearson, la cual es necesaria para determinar si existe en los datos una posible extradispersión. El prefijo ‘i.’ en la variable *age* (edad) factorea los niveles de *age*, siendo el primero el nivel de referencia por defecto.

```
. glm days i.age slow aborig girl, fam(poi) nolog
```

Generalized linear models		No. of obs	=	146	
Optimization	: ML	Residual df	=	139	
Deviance	= 1696.706552	Scale parameter	=	1	
Pearson	= 1830.191125	(1/df) Deviance	=	12.20652	
		(1/df) Pearson	=	13.16684	
Variance function:	V(u) = u	[Poisson]			
Link function	: g(u) = ln(u)	[Log]			
Log likelihood	= -1142.591815	AIC	=	15.74783	
		BIC	=	1003.985	
days		OIM			
		Coef.	Std. Err.	z	
age				P> z	[95% Conf. Interval]
F1	-.3339014	.0700935	-4.76	0.000	-.4712821 -.1965206
F2	.2578284	.0624194	4.13	0.000	.1354886 .3801681
F3	.4276938	.0676864	6.32	0.000	.295031 .5603567
slow	.348943	.0520431	6.70	0.000	.2469403 .4509456
aborig	.5336043	.0418831	12.74	0.000	.4515149 .6156937
girl	-.1615966	.0425346	-3.80	0.000	-.2449628 -.0782304
_cons	2.343372	.0603757	38.81	0.000	2.225038 2.461707


```
. abic
```

AIC Statistic	= 15.76153	AIC*n	= 2301.1836
BIC Statistic	= 15.87983	BIC(Stata)	= 2325.0525

La dispersión de Pearson muestra un valor de 13.16684. La estadística se define como la estadística Chi² de Pearson dividida por los grados de libertad de los residuos. Los estudios de simulación (Hilbe, 2011) demuestran que esta dispersión es preferible al uso de la dispersión de la desviación, que está sesgada. Si un modelo de Poisson es equidisperso, la dispersión estadística de Pearson tiene un valor de 1.0. Los valores superiores a 1.0 reciben el nombre de sobredispersos y los valores inferiores a 1.0 el de infradispersos. Los datos extra dispersos hacen referencia a datos que no son equidispersos; es decir, que o bien son infradispersos o bien son sobredispersos.

Téngase en cuenta que aunque pueda parecer que todos los predictores contribuyen al modelo, eso es, a la comprensión de los días de ausencia, el hecho de que la dispersión sea tan elevada demuestra que los datos están altamente correlacionados, son sobredispersos, y que la varianza excede a la media. La correlación excesiva es, muchas veces, el resultado de la aglomeración de los datos; por ejemplo, modelos de estudiantes en todos los colegios en que los métodos de enseñanza, el perfil demográfico, etc., difieren de un colegio a otro. Los resultados de los alumnos (independientemente de lo que se esté examinando) pueden ser más parecidos dentro de cada uno de los colegios que de un colegio a otro. Esto produce una sobredispersión. Si los resultados del examen se aglomeran en varios rangos de resultados, si los datos sufren un sesgo elevado, si hay un término de interacción necesario, si existe un número excesivo de recuentos en los datos, o si a veces no es posible realizar recuentos de cero, todas estas situaciones, y otras, pueden dar pie a la sobredispersión en los datos. Cuando los datos sufren una infradispersión o una sobredispersión, se hace necesario ajustar los errores típicos del modelo de Poisson. Si conocemos la causa de la sobredispersión, o si la sobredispersión es elevada, entonces es probable que sea necesario emplear un modelo de recuento que no sea el de Poisson.

Un simple resumen de los días confirma la evaluación que afirma que la varianza de los días de ausencia es superior a la media. Una vez más, se trata de una asunción esencial que permite usar el modelo de Poisson en los datos de recuento.

```
. su days
Variable |       Obs        Mean      Std. Dev.       Min       Max
-----+-----+-----+-----+-----+-----+-----+
days |      146     16.4589    16.25322         0        81
. di r(sd)^2
264.16726
```

La varianza observada es unas veinte veces mayor que la media de los días. Los usuarios de Stata pueden aprender cómo se almacenan los estadísticos siguiendo una sencilla prueba descriptiva que consiste en escribir *return list* y seguir un modelo a través de escribir *ereturn list*. Aquí he utilizado *r(sd)* para la desviación típica de los días guardada. En cualquier caso, cuando la varianza de la variable del recuento que se está modelando es sustancialmente superior a su media, la variable sufre una sobredispersión. Los predictores explicativos pueden

mejorar o incluso eliminar la sobredispersión, lo cual, en general, puede evaluarse comprobando el valor de los estadísticos de dispersión.

Los demás componentes del resultado del modelo de recuento que aparecen más arriba son el estadístico z , el valor p y los intervalos de confianza. Es importante entender cómo se definen dichos componentes.

El estadístico z es la relación del coeficiente del predictor con el error típico. Para el predictor *slow* tenemos:

```
di .348943 /.0520431
6.704885
```

o bien se usan símbolos de estimación guardados para el coeficiente (beta) y el error típico:

```
. di _b[slow] / _se[slow]
6.7048791
```

Se asume que el estadístico z sigue una distribución normal típica acumulativa, a diferencia del modelo normal o de mínimos cuadrados que se emplea en la estadística t . La probabilidad de la variable *slow* se calcula como $2 * N(-|z|)$, en que N es una función normal típica acumulativa y z es la estadística z calculada.

```
. di %9.8f normal(-abs(_b[slow]/_se[slow]))*2
0.00000000
```

Los intervalos de confianza están basados en un nivel de significación. Si utilizamos un intervalo de confianza del 95%, el nivel de significación se basa en la distribución normal típica acumulativa inversa de dos lados.

```
. di invnorm(0.975)
1.959964
```

Los intervalos de confianza relativos a *slow* se determinan, por lo tanto, utilizando la fórmula:

```
. di _b[slow] - invnorm(0.975) * _se[slow]
.24694028
. di _b[slow] + invnorm(0.975) * _se[slow]
.45094564
```

Nótese que estos valores corresponden al resultado del modelo mostrado anteriormente. Debería mencionarse que, hoy en día, tanto estadísticos como científicos sociales están desaconsejando el uso de un valor p para definir el nivel estadísticamente significativo de un predictor, siendo preferible el método de comprobar los intervalos de confianza del predictor. Si el intervalo de confianza incluye el cero dentro de su rango, se considerará que el predictor no es significativo. Un predictor significativo no contiene el cero en su intervalo de confianza.

La modificación de los errores típicos de Poisson

El problema que encierra este modelo es que los errores estándar están sesgados y conducen al analista a creer que los predictores son significativos, cuando podría

no serlo. Los usuarios de R remodelarán los datos con la función *glm* utilizando lo que se conoce con el término de distribución *quasi-Poisson* en vez de *Poisson*. Esto es idéntico a *escalar* los errores típicos a través de multiplicar los errores típicos del modelo dado por la raíz cuadrada de la dispersión de Pearson. Lo que esta técnica hace es ajustar los errores típicos a lo que serían si la estadística de dispersión fuera 1.0. Desde alrededor de 1973 hasta mediados de la década de 1990, este fue el método más utilizado para el ajuste de un modelo de Poisson sobredisperso.

El cálculo de los errores típicos escalados se produce después de la estimación del modelo. El paquete estadístico, ya sea Stata, SAS, SPSS o R, estima los errores típicos, calcula el ajuste y ejecuta una iteración más con el ajuste mostrado. En el caso del predictor *slow*, los errores típicos escalados se calculan del siguiente modo:

ERROR TÍPICO DE SLOW

```
. di _se[slow]
.05204314
```

RAÍZ CUADRADA DEL ESTADÍSTICO PEARSON DE DISPERSIÓN

```
. di sqrt(e(dispers_p))
3.6286144
```

ERRORES TÍPICOS ESCALADOS

```
. di _se[slow] * sqrt(e(dispers_p))
.18884449
```

En Stata, la opción *scale(x2)* en *glm* produce errores típicos escalados. La opción *nohead* ordena al paquete estadístico bloquear la muestra de los títulos de comandos, mientras que *nolog* impide que se muestre la iteración del log.

```
. glm days i.age slow aborig girl, fam(poi) nolog nohead scale(x2)

-----  

          OIM  

days | Coef. Std. Err.      z   P>|z| [95% Conf. Interval]  

-----+-----  

age |  

F1 | -.3339014 .2543423 -1.31  0.189  -.8324031 .1646003  

F2 | .2578284 .2264959  1.14  0.255  -.1860955 .7017522  

F3 | .4276938 .2456077  1.74  0.082  -.0536885 .9090762  

|  

slow | .348943  .1888445  1.85  0.065  -.0211854 .7190714  

aborig | .5336043 .1519776  3.51  0.000  .2357336 .831475  

girl | -.1615966 .1543415 -1.05  0.295  -.4641004 .1409072  

_cons | 2.343372 .2190801 10.70  0.000  1.913983 2.772761  

-----  

(Standard errors scaled using square root of Pearson X2-based dispersion.)
```

Nótese que el error típico escalado correspondiente a *slow* es el mismo que el calculado a mano. Los usuarios de R pueden escribir lo siguiente para obtener resultados idénticos.

```
> mypoi <- glm(days ~ factor(age)+slow+aborig+girl,family=quasipoisson, data=absent)
> summary(mypoi)
```

El valor p , z , y los intervalos de confianza también deberán ajustarse cuando se escalen los errores típicos. El siguiente código obtendrá los resultados correctos.

VALORES DE P

```
. di %9.8f normal(-abs(_b[slow]/_se[slow]))*2  
0.06371483
```

INTEVALOS DE CONFIANZA (CI) ESCALADOS

```
. di _b[slow] - invnorm(0.975) * _se[slow]  
-.02118543  
  
. di _b[slow] + invnorm(0.975) * _se[slow]  
.71907136
```

El método más utilizado y recomendado para ajustar errores típicos cuando existente evidencia de sobredispersión es, quizás, el uso de algún tipo de estimador de varianza sándwich. Existen varios métodos para calcular lo que llamamos errores típicos robustos, sándwich o errores típicos Huber-White. Para consultar ejemplos al respecto, véase Hardin y Hilbe (2012). El resultado de aplicar errores típicos robustos al modelo anterior es el siguiente:

POISSON — ERROR TÍPICO ROBUSTO

days		Robust				
		Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
<hr/>						
age						
F1		-.3339014	.2675844	-1.25	0.212	-.8583571 .1905543
F2		.2578284	.2505294	1.03	0.303	-.2332002 .7488569
F3		.4276938	.2485347	1.72	0.085	-.0594252 .9148129
slow		.348943	.1881936	1.85	0.064	-.0199097 .7177957
aborig		.5336043	.1538529	3.47	0.001	.2320582 .8351504
girl		-.1615966	.1555387	-1.04	0.299	-.4664468 .1432536
_cons		2.343372	.2433279	9.63	0.000	1.866459 2.820286

El cociente de tasa de Poisson

De manera parecida a como los analistas utilizan la razón de oportunidad relativa (*odd ratios*) con modelos logísticos, la potenciación de los coeficientes del modelo de recuento resulta en *índices de incidencia*. Este estadístico se relaciona con la tasa en la cual se producen recuentos para un nivel de predictor en comparación con otro nivel. La opción *eform* eleva a la potencia los coeficientes y ajusta de manera apropiada otras estadísticas asociadas. El efecto estadísticamente significativo de un predictor debería evaluarse en función de si el intervalo de confianza contiene el valor de uno. Si el intervalo de confianza no incluye uno, el predictor puede considerarse significativo, aunque solo si se han satisfecho otras asunciones distribucionales del modelo como, por ejemplo, la equidispersión y la independencia de las observaciones.

```
. glm days i.age slow aborig girl, fam(poi) nolog eform nohead
```

days	OIM					
	IRR	Std. Err.	z	P> z	[95% Conf. Interval]	
age						
F1	.7161244	.0501957	-4.76	0.000	.6242015	.8215844
F2	1.294117	.080778	4.13	0.000	1.145096	1.46253
F3	1.533716	.1038117	6.32	0.000	1.343168	1.751297
slow	1.417568	.0737747	6.70	0.000	1.280103	1.569796
aborig	1.705067	.0714135	12.74	0.000	1.57069	1.85094
girl	.8507844	.0361877	-3.80	0.000	.7827337	.9247513
_cons	10.41631	.6288915	38.81	0.000	9.253838	11.7248

La interpretación del cociente de tasas puede ser la siguiente:

- Los alumnos de primer año de secundaria se ausentan un 28% menos que los alumnos de primaria de último año. ($1 - .716$) * 100
- Los alumnos de segundo año de secundaria se ausentan un 30% más que los alumnos de primaria de último año. (.29 * 100)
- Los alumnos de tercer año de secundaria se ausentan un 53% más que los alumnos de primaria de último año. (.53 * 100)
- Los aprendices lentos se ausentan un 42% más que los aprendices de velocidad media. (.418 * 100)
- Los alumnos de origen étnico aborigen se ausentan un 70% más que los alumnos de origen étnico identificado como blanco. (.705 * 100)
- Las niñas se ausentan un 15% menos que los niños. ($1 - .85$)* 100

Nótese que cuando se interpreta un predictor, asumimos que los otros predictores se mantienen a un valor constante.

Los investigadores raramente interpretan los coeficientes del modelo de recuento. Cuando lo hacen, la referencia son los log-recuentos, que son de difícil interpretación. En cambio, los investigadores usan generalmente modelos de recuento estándar, como el de Poisson, con el fin de predecir perfiles de predictor de recuentos. La parametrización del cociente de tasas se utiliza cuando los predictores se interpretan en términos de porcentaje diferencial de recuentos para niveles en un predictor.

Muchos investigadores se equivocan cuando calculan los errores típicos y los intervalos de confianza de los índices de riesgo. El llamado método delta es el método que debería usarse para calcular errores típicos. Esencialmente, por lo que respecta a los índices de riesgo, los errores típicos se calculan multiplicando el error típico del modelo por el índice de riesgo. Los intervalos de confianza no están basados en este error típico, sino que son la potenciación de los intervalos de confianza del modelo. Compárense los cálculos a continuación con el resultado estadístico anterior. No todos los paquetes estadísticos proporcionan estos valores.

COCIENTE DE TASA DEL ERROR TÍPICO

```
. di exp(_b[slow])*_se[slow]
.07377471
```

COCIENTE DE TASA DEL INTERVALO DE CONFIANZA

```
. di exp(_b[slow] - invnorm(0.975) * _se[slow])
1.2801027
```

```
. di exp(_b[slow] + invnorm(0.975) * _se[slow])
1.569796
```

La parametrización por área o tiempo

Los modelos de Poisson a menudo se utilizan para analizar los datos de recuento que se recogen a lo largo de diferentes períodos de tiempo o en diferentes áreas. Cuando este es el objetivo, el modelo de Poisson se ajusta para obtener la siguiente PDF subyacente:

$$f(y; \mu) = \frac{e^{-t\mu}(t\mu)^y}{y!} \quad (10)$$

siendo t una variable que indica las diferencias temporales o de área a lo largo de las cuales se realiza la observación. Para modelar los datos debe introducirse la variable de tiempo o área a modo de compensación. No obstante, es importante calcular el logaritmo de la variable de compensación antes de introducirla en el modelo. La razón para ello es que la compensación se añade al predictor lineal, el cual se calcula internamente mediante el algoritmo de estimación. La variable de compensación debe colocarse en la misma métrica que los demás términos en el predictor lineal. Nótese, no obstante, que el coeficiente de la compensación calculada se ajusta a 1.0. Esto introduce el predictor lineal como una constante.

Como ejemplo de ello, considérense los datos relativos a los días que los profesores estuvieron enfermos en varios colegios de la ciudad de Phoenix, Arizona, en 2015. Los días de enfermedad son la respuesta de recuento, con predictores de género (mujer = 1; hombre = 0), caucásico/a (profesor/a identificado/a como de origen caucásico = 1; origen no caucásico = 0), titular (sí = 1; no = 0), e implicado/a en actividades extracurriculares; e.g., entrenador/a (sí = 1, no = 0). Los colegios que forman parte de los datos son centros de tamaños diferentes, desde pequeños colegios subvencionados con 45 alumnos hasta centros con 3,353 alumnos. Los datos no incluyen el número de profesores. Los recuentos de los días que los profesores estuvieron enfermos se ajustaron en función del tamaño del colegio, el cual se introduce en el modelo a modo de variable de compensación. El uso de Stata permite modelar los datos del siguiente modo:

. glm sickdays gender white tenured extra, exposure(schoolsize) fam(poi) nolog vce(ro) eform	
Generalized linear models	No. of obs = 27
Optimization : ML	Residual df = 22
Deviance = 42.89127063	Scale parameter = 1
Pearson = 41.54589787	(1/df) Deviance = 1.949603
Variance function: V(u) = u	(1/df) Pearson = 1.88845
Link function : g(u) = ln(u)	[Poisson]
	[Log]
Log pseudolikelihood = -51.81429589	AIC = 4.208466
	BIC = -29.61714
-----	-----
	Robust
sickdays	IRR Std. Err. z P> z [95% Conf. Interval]
-----+-----	-----+-----
gender 1.137691 .3173939 0.46 0.644 .6584996 1.965591	
white 4.685963 2.118087 3.42 0.001 1.932191 11.36443	
tenured 5.492599 1.777975 5.26 0.000 2.912331 10.35893	
extra 3.53922 1.556543 2.87 0.004 1.494688 8.380399	
_cons .0008261 .0002398 -24.45 0.000 .0004676 .0014593	
ln(school~e) 1 (exposure)	
-----	-----

El sexo del profesor no es un predictor significativo. Los profesores que se identificaron a sí mismos como personas de origen caucásico resultaron tener cuatro veces y media más de probabilidad de estar enfermos que los profesores de origen no caucásico. Los profesores titulares tienen cinco veces y media más de probabilidad de estar enfermos que los profesores no titulares, y los profesores que realizan actividades extracurriculares tienen tres veces y media más de probabilidad de estar enfermos que aquellos profesores que no participan en actividades extracurriculares. Recuérdese que por cada predictor, los otros predictores se mantienen a un valor constante, ya que de lo contrario se generaría un efecto de interacción. El modelo parece mostrar una ligera sobredispersión (1.888). Dado que los errores típicos se han ajustado mediante un estimador de varianza de sándwich robusto, sería recomendable modelar los datos utilizando otro modelo de recuento. El 30% de los profesores no estuvieron enfermos ningún día, lo cual infringe la asunción distribucional de Poisson que afirma que solo el 2.2% de los profesores no presentan días de ausencia por enfermedad (dada una media de 3.815 días de enfermedad por escuela). Para obtener más información relacionada con esta asunción, y sobre cómo abordarla, véase la sección de los modelos de cero inflado.

La evaluación del modelo

La mejor manera de evaluar la bondad de ajuste del modelo de Poisson es valorando la estadística de dispersión, tal y como se ha mencionado anteriormente, y probando la diferencia entre los recuentos observado y predicho. Cuanto menor sea la diferencia entre los dos a lo largo de todo el rango de recuentos, mejor será el ajuste. Puede utilizarse una prueba Chi² con grados de libertad equivalentes a la cantidad de los recuentos que se están comparando con el fin de determinar la significación del ajuste al nivel .05. Aunque, tradicionalmente, los valores de probabilidad superiores a .05 indican que se trata de un modelo bien

ajustado, será importante asegurarse de que los valores p no se interpreten de manera demasiado estricta. Simplemente deberá informarse sobre los resultados de la prueba describiéndolos tal y como estos se muestren, sin interpretación del nivel de significación, salvo que la estadística Chi² sea substancialmente alta. La siguiente prueba indica un ajuste relativamente deficiente que corresponde a la alta dispersión en los datos. Un gráfico de la diferencia entre el recuento observado y el recuento predicho puede resultar instructivo a la hora de evaluar visualmente el ajuste.

```
. chi2gof, cells(15) table
```

Chi-square Goodness-of-Fit Test for Poisson Model:

Cells	Abs. Freq.	Fitted			Abs. Dif.
		Rel. Freq.	Rel. Freq.	Rel. Freq.	
0	9	.0616	9.7e-05	.0615	
1	4	.0274	6.7e-04	.0267	
2	6	.0411	.0024	.0387	
3	6	.0411	.0059	.0352	
4	1	.0068	.0115	.0047	
5	19	.1301	.0187	.1114	
6	8	.0548	.0267	.0281	
7	6	.0411	.0347	.0064	
8	4	.0274	.0419	.0145	
9	2	.0137	.0477	.034	
10	4	.0274	.0521	.0247	
11	7	.0479	.0549	.007	
12	2	.0137	.0563	.0426	
13	3	.0205	.0563	.0358	
14 or more	57	.4452	.59	.1448	

La regresión binomial negativa

Los datos que hemos modelado más arriba mediante el modelo de Poisson son altamente sobredispersos. Los errores típicos están sesgados. Por norma general, los estadísticos se inclinan por la regresión binomial negativa cuando se enfrentan a una modelación de datos de Poisson sobredispersos. La binomial negativa es una combinación de las distribuciones Gamma y de Poisson, con un parámetro auxiliar que recibe el nombre de parámetro de dispersión, cuya finalidad es ajustar la sobredispersión. La binomial negativa solo puede utilizarse en datos de recuento sobredispersos. No obstante, la mayoría de datos reales de recuento son de hecho sobredispersos. Esto significa que el modelo binomial negativo se ha convertido en un modelo central en la evaluación de datos de recuento.

La función de probabilidad binomial negativa se expresa de varias maneras. La definición de la distribución que se utiliza en los paquetes estadísticos Stata, SAS y SPSS es la siguiente:

$$f(y; \mu, \alpha) = \binom{y_i + \frac{1}{\alpha} - 1}{\frac{1}{\alpha} - 1} \left(\frac{1}{1 + \alpha\mu_i} \right)^{\frac{1}{\alpha}} \left(\frac{\alpha\mu_i}{1 + \alpha\mu_i} \right)^{y_i} \quad (11)$$

La función de log-verosimilitud para la PDF anterior puede expresarse en el siguiente formato de media exponencial:

$$\begin{aligned} \mathcal{L}(\mu; y, \alpha) = & \sum_{i=1}^n y_i \log \left(\frac{\alpha\mu_i}{1 + \alpha\mu_i} \right) - \frac{1}{\alpha} \log(1 + \alpha\mu_i) + \log\Gamma \left(y_i + \frac{1}{\alpha} \right) \\ & - \log\Gamma(y_i + 1) - \log\Gamma \left(\frac{1}{\alpha} \right) \end{aligned}$$

en que la función $\log\Gamma(*)$ simboliza la función log-Gamma. Cabe recordar que el primer término de la PDF binomial negativa es una función de elección que puede convertirse en formato factorial. Las factoriales pueden, a su vez, transformarse en funciones Gamma. Cuando la PDF se parametriza en formato de log-verosimilitud, las funciones Gamma se convierten en funciones log-Gamma (Hilbe, 2011).

La media del modelo binomial negativo es μ , tal como sucede con la media de Poisson. No obstante, la varianza es

$$V(\mu) = \mu + \alpha\mu^2 \quad (12)$$

en comparación con la varianza de Poisson, que es μ . La varianza puede interpretarse como una combinación de la varianza de Poisson, μ , y la varianza Gamma, μ^2/γ . No obstante, el modelo se parametriza a fin de que exista una relación directa entre la media y la cantidad de dispersión en los datos. Por lo tanto, el parámetro de dispersión se establece como $\alpha = 1/\gamma$. Esta parametrización se utiliza actualmente en Stata, SAS, SPSS, Genstat, Limdep y todos los paquetes estadísticos comerciales. Los usuarios de R deberán, no obstante, tener en cuenta que las funciones *glm.nb* y *glm* por defecto parametrizan indirectamente el parámetro de dispersión, el cual recibe el nombre de theta, θ . La varianza binomial negativa para estas funciones R es $\mu + \mu^2/\theta$. Nótese que cuando $\alpha = 0$, la binomial negativa es de Poisson. Esto significa que unos valores altos de α en un modelo binomial negativo son indicativos de que se ha ajustado una mayor dispersión de Poisson. Para R, sin embargo, un modelo es de Poisson cuando θ es infinitamente alta. En este caso, unos valores bajos de θ indican una dispersión o una correlación alta en los datos. No obstante, el paquete *gamlss* en R, que proporciona a los usuarios de R un gran número de capacidades de modelación, utiliza la parametrización directa, α . La parametrización del parámetro de dispersión binomial negativa no tiene repercusión alguna en los coeficientes del modelo ni en los errores típicos. A pesar de ello, será necesario especificar el

paquete estadístico o tipo de dispersión que se emplea a la hora de informar sobre los resultados de los estudios binomiales negativos en publicaciones.

A continuación presentamos los datos que sirven de ejemplo utilizando un modelo binomial negativo.

```
. glm days i.age slow aborig girl, fam(nb ml) nolog

Generalized linear models
Optimization      : ML
No. of obs       =      146
Residual df     =      139
Scale parameter =      1
Deviance        = 167.9518008
(1/df) Deviance = 1.208286
Pearson         = 137.7760369
(1/df) Pearson  = .9911945

Variance function: V(u) = u + (.7844)u^2
Link function   : g(u) = ln(u) [Neg. Binomial]
                  [Log]

Log likelihood   = -546.5755091          AIC            = 7.583226
                                         BIC            = -524.7695
-----

```

	OIM						
days		Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
age							
F1		-.4484282	.2376006	-1.89	0.059	-.9141168	.0172605
F2		.0880802	.2415453	0.36	0.715	-.3853399	.5615002
F3		.356901	.2466194	1.45	0.148	-.1264641	.8402661
slow		.2921092	.1829335	1.60	0.110	-.066434	.6506523
aborig		.5693717	.1576082	3.61	0.000	.2604654	.878278
girl		-.0823203	.164681	-0.50	0.617	-.4050892	.2404486
_cons		2.407529	.2088926	11.53	0.000	1.998107	2.816951

```
. abic
AIC Statistic = 7.596925          AIC*n       = 1109.151
BIC Statistic = 7.71522           BIC(Stata) = 1133.0199
```

Los estadísticos del Criterio de Información Akaike (*AIC*, por sus siglas en inglés) y del Criterio de Información Bayesiano (*BIC*, por sus siglas en inglés) son sustancialmente más bajos que las del modelo de Poisson, que equivalían a 2,301 y 2,325 respectivamente. Se considera que los modelos aplicados sobre los mismos datos con valores *AIC* y *BIC* inferiores son comparativamente modelos mejor ajustados. Por norma general, si los valores difieren en más de 10, podemos en principio tener la seguridad de que el modelo preferible será aquel con valores más bajos para *AIC* y *BIC*. Aunque existen varios estadísticos *AIC* y *BIC* diferentes desarrollados por especialistas, las que se emplean en Stata, SAS, SPSS y R son las mismas. Nótese que SAS hace referencia a la estadística *BIC* con el nombre de Criterio Schwarz (SC, por sus siglas en inglés), siendo este el que se facilita en la mayoría de resultados del modelo.

El estadístico de dispersión de Pearson relativo al modelo binomial negativo arriba mencionado es 0.99, muy cerca de 1.0. Esto indica que el modelo binomial negativo se ha adaptado plenamente a la sobredispersión de Poisson. Los errores típicos que se muestran en el modelo son bastante más consistentes con los valores dados en los modelos escalado y Poisson robusto. Los coeficientes del modelo y los cocientes de tasas se interpretan del mismo modo que en el caso del modelo de Poisson.

Durante el modelado del predictor categórico, *age*, se observa que los niveles de *age* no eran significativos con referencia al nivel inferior, llamado F1. No obstante, cuando se reestructuran los niveles de edad con el fin de modelar el nivel más alto respecto a los demás niveles, el predictor binario resultante sí que es significativo. Llamo al nuevo predictor binario *a4*, siendo el nivel de referencia los niveles combinados 1–3 de *age*. El código para la creación de *a4* es el siguiente:

```
. gen byte a4=edad==4
```

A continuación se detalla un modelo completo binomial negativo de verosimilitud máxima de los datos ajustados con errores típicos robustos y coeficientes parametrizados como cocientes de tasas.

```
. nbreg days a4 aborig girl slow, nolog vce(ro) irr
```

days	Robust					
	IRR	Std. Err.	z	P> z	[95% Conf. Interval]	
a4	1.581041	.3126757	2.32	0.021	1.073012	2.329601
aborig	1.817805	.2756562	3.94	0.000	1.350421	2.446953
girl	.8055	.1268699	-1.37	0.170	.5915582	1.096816
slow	1.312772	.2367492	1.51	0.131	.9218941	1.86938
_cons	10.55794	1.82515	13.63	0.000	7.523712	14.81584
/lnalpha	-.1929157	.125174			-.4382523	.0524209
alpha	.8245515	.1032124			.645163	1.053819
. abic						
AIC Statistic	=	7.61652			AIC*n	= 1112.012
BIC Statistic	=	7.681596			BIC(Stata)	= 1129.9136

Los estadísticos *AIC* y *BIC* relativos a este modelo son ligeramente superiores a los relativos al modelo que dispone de todos los niveles de *age*. No obstante, la diferencia en los estadísticos no es significativa. Nótese que los predictores *girl* y *slow* tienen intervalos de confianza entre los que se incluye el valor 1, lo que indica que no contribuyen de manera significativa al modelo o a la comprensión de los días de ausencia en el colegio. Si estos son apartados del modelo, los valores *AIC* y *BIC* se verán reducidos. El estadístico *AIC* es 0.3 inferior, aunque el estadístico *BIC* es 6.3 inferior, lo cual indica un ajuste posiblemente mejor. El hecho de que únicamente haya 146 observaciones en el modelo debe llevarnos a actuar con cautela a la hora de dar demasiada importancia a pequeños descensos en los valores de los estadísticos *AIC* y *BIC*.

Modelo binomial negativo heterogéneo

El modelo desarrollado utilizando la regresión de Poisson ha evidenciado una sobredispersión más bien sustancial en los datos. No obstante, el modelo binomial

negativo parecía estar adecuadamente ajustado. A pesar de ello, podemos realizar una comprobación que nos permitirá determinar qué predictores, si alguno, contribuyen a cualquier dispersión excesiva en el modelo binomial negativo. El modelo con *a4* mencionado anteriormente tiene un estadístico de dispersión de 1.009295 (que no se muestra), por lo que no deberíamos esperar la existencia de predictores que contribuyan a dicha dispersión. Sin embargo, la mayoría de los modelos binomiales negativos no se ajustan por completo a todas las sobredispersiones en los datos.

El uso de un modelo binomial negativo heterogéneo permite determinar qué predictores contribuyen a la sobredispersión.

Generalized negative binomial regression						
				Number of obs	=	146
Log likelihood =	-547.99619			LR chi2(4)	=	16.65
				Prob > chi2	=	0.0023
				Pseudo R2	=	0.0150
days		Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
days						
a4		.3906054	.2133284	1.83	0.067	-.0275107 .8087215
aborig		.5861604	.1616776	3.63	0.000	.2692781 .9030427
slow		.2931937	.1768329	1.66	0.097	-.0533924 .6397797
girl		-.174187	.1599756	-1.09	0.276	-.4877334 .1393593
_cons		2.345858	.1732254	13.54	0.000	2.006342 2.685373
lnalpha						
a4		.280441	.3567383	0.79	0.432	-.4187531 .9796352
aborig		-.4404478	.258165	-1.71	0.088	-.946442 .0655464
slow		.3046848	.2938908	1.04	0.300	-.2713306 .8807003
girl		-.0665591	.261742	-0.25	0.799	-.579564 .4464458
_cons		-.1484573	.2755071	-0.54	0.590	-.6884414 .3915267

Los predictores significativos en el componente *lnalpha* del modelo contribuyen a la sobredispersión. Como se sospechaba, ninguno de los predictores del modelo es significativo. No obstante, si cualquiera de los predictores lo fuera, sabríamos que es necesario comprobarlo con el fin de determinar cómo puede contribuir a la sobredispersión.

Los modelos de cero inflado

Muchos modelos de recuento tienen un recuento de ceros excesivo. Las distribuciones binomial negativa y de Poisson subyacentes a los modelos bien ajustados asumen que el número observado de recuentos en el modelo se acerca a lo anticipado en base a la media distribucional. La media de días (*days*) es 16.4589. El número anticipado de días con recuentos 0 para el modelo de Poisson es, por lo tanto, 0.

POISSON: NÚMERO DE DÍAS PREDICHIOS CON RECUENTOS 0

```
. di exp(-16.4589)*16.4589^0 * 146
.00001038
```

Por lo que respecta al modelo binomial negativo arriba mencionado, también deberemos conocer el valor del parámetro de dispersión del modelo. En este caso es 0.8245515. El producto de la media y *alpha* es 13.571211. Con esta información podemos ahora calcular el cero predicho usando la PDF NB (del inglés *negative binomial*, ‘binomial negativo’).

```
. di exp(0* log( 13.571211/(1+ 13.571211)) - (1/.8245515) * log(1+ 13.571211) /*
   */ + lngamma(0 +1/.8245515) /*
   */ - lngamma(0+1) - lngamma(1/.8245515))
.03880924
```

BINOMIAL NEGATIVO: NÚMERO DE DÍAS PREDICIOS CON RECUENTOS 0

```
. di 146 * .03880924
5.666149
```

El número de ceros predicho por el modelo binomial negativo es 5.7. Como puede observarse en la sección de la evaluación del modelo, de hecho hay nueve recuentos de cero en los datos. No existe una gran discrepancia en los dos valores, por lo que probablemente no será necesario emplear un modelo de cero inflado.

Un modelo de cero inflado es un modelo de combinación que modela los recuentos de cero como un modelo binario, generalmente una regresión logística, y todos los recuentos utilizando un modelo de recuento. La combinación se produce en tanto que el componente binario y el recuento del modelo estiman los recuentos de cero. El punto principal al utilizar un modelo de cero inflado en los datos es ajustar la sobredispersión en los datos causada por un exceso de recuentos de cero.

```
zinb days a4 aborig slow girl, nolog inflate(days a4 aborig slow girl) irr

Zero-inflated negative binomial regression          Number of obs      =      146
                                                    Nonzero obs       =      137
                                                    Zero obs        =         9
Inflation model = logit                          LR chi2(4)        =     17.91
Log likelihood   = -545.8493                     Prob > chi2       =  0.0013
-----
          days |      IRR      Std. Err.      z      P>|z|      [95% Conf. Interval]
-----+
days      |
    a4 |  1.527242  .3082863   2.10  0.036  1.028219  2.268453
    aborig |  1.69585  .2661596   3.37  0.001  1.24679  2.306649
    girl |  .7499481  .1152076  -1.87  0.061  .5549693  1.013429
    slow |  1.305235  .2339429   1.49  0.137  .9185955  1.854613
    _cons |  11.90538  2.003941  14.72  0.000  8.559869  16.55844
-----
inflate   |
    a4 |  -.3847747  1.526356  -0.25  0.801  -3.376378  2.606829
    aborig |  -21.21131  17819.8  -0.00  0.999  -34947.38  34904.95
    girl |  -3.407821  11.99498  -0.28  0.776  -26.91755  20.10191
    slow |  -.1653036  1.8618  -0.09  0.929  -3.814365  3.483757
    _cons |  -1.660992  .9083664  -1.83  0.067  -3.441357  .1193733
-----
/lnalpha |  -.3533463  .1556277  -2.27  0.023  -.6583709  -.0483216
-----
alpha |  .702334  .1093026                               .517694  .9528273
-----
Likelihood-ratio test of alpha=0: chibar2(01) =  1104.35 Pr>=chibar2 =  0.0000
Vuong test of zinb vs. standard negative binomial: z =      1.30  Pr>z =  0.0976
.
.abic
AIC Statistic = 7.628073          AIC*n = 1113.6987
BIC Statistic = 7.838715          BIC(Stata) = 1146.5183
```

Los estadísticos *AIC* y *BIC* son casi idénticos a los del modelo binomial negativo típico; de hecho, el modelo de cero inflado goza de unos valores *AIC* y *BIC* ligeramente mayores. Si hubiese excesivos recuentos de cero, este no sería el caso.

Nótense también las dos pruebas realizadas con el resultado del modelo. La prueba de razón de verosimilitud determina si el modelo binomial negativo de cero inflado (ZINB, por sus siglas en inglés) es un ajuste significativamente mejor que el modelo de Poisson de cero inflado (ZIP, por sus siglas en inglés). Los valores *p* por debajo de .05 indican que el modelo ZINB es preferible.

La prueba de Vuong determina si el modelo inflado es preferible al modelo no inflado. La estadística de la prueba suele distribuirse, $N(0,1)$, con los valores positivos altos favoreciendo al modelo inflado y los valores negativos altos favoreciendo al modelo no inflado. Un valor *p* de .05 especifica que el modelo preferible es un ajuste significativamente mejor que el otro. En este caso, no hay diferencia entre los modelos ZINB y NB por lo que respecta al ajuste, lo cual, de hecho, ya era objeto de nuestras sospechas.

Los modelos de recuento *valla* de dos partes

Como los modelos de cero inflado, los modelos de recuento *valla* suelen utilizarse para modelar datos con un exceso de recuentos de cero. El modelo consiste en dos partes separables. Los recuentos positivos se modelan mediante el modelo truncado por cero. El componente binario modela los datos, siendo 1 equivalente a los recuentos superiores a cero y siendo 0 equivalente a cero. Generalmente, el componente binario utiliza un modelo logístico o probit, aunque este no es necesario. Para un examen completo al respecto, véase Hilbe (2014). Recientes estudios sobre esta cuestión han indicado que algunos modelos *valla* pueden ser superiores a los modelos de cero inflado cuando se emplean para ajustar recuentos de cero excesivos en los datos.

La diferencia primordial en cómo se utilizan los modelos de cero inflado y *valla* se basa en que el componente de recuento de un modelo *valla* es un modelo truncado por cero. Los ceros se evalúan completamente utilizando el componente binario. Con los modelos de cero inflado, los ceros se incluyen tanto en el recuento como en el componente binario. Asimismo, es importante recordar que, cuando se interpreta un modelo de cero inflado, el componente binario modela el cero, mientras que en el caso del modelo *valla* el componente binario modela el uno, lo cual incluye a todos los recuentos que se realizan en los datos superiores a cero. El componente binario del modelo *valla* modela los recuentos positivos en comparación a los recuentos de ceros.

El modelo a continuación es un modelo *valla* binomial-logit negativo con errores típicos robustos. Se recomienda el uso por defecto de ajustes robustos en todos los modelos de recuento. Si los datos no son sobredispersos, los errores típicos y las estadísticas asociadas se reducen a los errores típicos del modelo regular. Aquí he mostrado variaciones por motivos pedagógicos.

MODELO VALLA BINOMIAL-LOGIT NEGATIVO

```
. hnblogit days a4 aborig slow girl, nolog vce(ro)

Negative Binomial-Logit Hurdle Regression          Number of obs      =      146
Log pseudolikelihood = -546.21056                Wald chi2(4)       =      6.31
                                                    Prob > chi2        =     0.1770
-----+
|           Robust
|   Coef.  Std. Err.      z    P>|z|  [95% Conf. Interval]
-----+
logit    |
    a4 | -.1764771  .9443589   -0.19  0.852  -2.027387  1.674432
    aborig |  2.076402  1.089484    1.91  0.057  -.0589462  4.211751
    slow | -.2278658  .7658108   -0.30  0.766  -1.728827  1.273096
    girl |  1.004074  .7218531    1.39  0.164  -.4107321  2.41888
    _cons |  1.842387  .7199687    2.56  0.010  .4312742  3.2535
-----+
negbinomial |
    a4 |  .4660117  .192477    2.42  0.015  .0887637  .8432598
    aborig |  .525538  .1518729   3.46  0.001  .2278726  .8232033
    slow |  .298715  .1819296   1.64  0.101  -.0578605  .6552905
    girl |  -.2796566  .1564645  -1.79  0.074  -.5863214  .0270081
    _cons |  2.444588  .1703575   14.35 0.000  2.110694  2.778483
-----+
/lnalpha |  -.3304558  .1381037   -2.39  0.017  -.6011341  -.0597775
-----+
alpha |  .7185961  .0992408                               .5481896  .9419741
-----+
. abic
AIC Statistic =  7.633021          AIC*n      = 1114.4211
BIC Statistic =  7.843663          BIC(Stata) = 1147.2408
```

Dado que el modelo binomial negativo ha sido aparentemente ajustado para la sobredispersión en los datos, hemos descubierto que el modelo de cero inflado no es preferible al modelo no inflado. Asimismo, no se prevé que el modelo valla arriba descrito esté mejor ajustado que el modelo binomial negativo que hemos utilizado en los datos. Las estadísticas *AIC* y *BIC* confirman esta previsión.

Los modelos de recuento alternativos

Existen otros modelos de recuento desarrollados por especialistas con la finalidad de modelar datos que pueden no ser apropiados para los modelos de Poisson o binomial negativo. El modelo Generalizado de Poisson (GP) y el modelo Poisson-inverso Gaussiano (PIG) son ejemplos claros de ello. Todos los modelos examinados a continuación fueron propuestos y publicados por James Hardin (Universidad de Carolina del Sur) y J.M. Hilbe (Hardin & Hilbe, 2012).

El modelo Generalizado de Poisson

El modelo generalizado de Poisson es una combinación de distribuciones de Poisson. A diferencia del binomial negativo, el modelo generalizado de Poisson puede modelar datos de Poisson infradispersos. Los valores negativos de *delta*, el parámetro de dispersión GP, indican el ajuste relativo a los datos infradispersos de Poisson.

```
. gpoisson days a4 aborig slow girl, nolog irr vce(ro)

Generalized Poisson regression                               Number of obs = 146
                                                               Wald chi2(4) = 19.45
Prob > chi2 = 0.0006                                         = 0.0006
Dispersion = .7659908                                         Prob > chi2 = 0.0006
Log pseudolikelihood= -552.06227                           Pseudo R2 = 0.0169
-----
```

days		IRR	Robust Std. Err.	z	P> z	[95% Conf. Interval]
a4	1.21606	.253484	0.94	0.348	.8082105	1.829723
aborig	1.712981	.2406284	3.83	0.000	1.300713	2.255921
slow	1.1055	.1671772	0.66	0.507	.821935	1.486895
girl	.8738746	.1269624	-0.93	0.353	.657326	1.161763
_cons	12.09324	1.905104	15.82	0.000	8.880741	16.46783
/tanhdelta	1.010553	.0474949			.9174651	1.103642
delta	.7659908	.0196276			.724696	.8018033

Likelihood-ratio test of delta=0: chi2(1) = 1293.28 Prob>=chi2 = 0.0000

```
. abic
AIC Statistic = 7.644689          AIC*n = 1116.1245
BIC Statistic = 7.709765          BIC(Stata) = 1134.0262
```

La prueba de razón de verosimilitud nos dice que el modelo GP es preferible al modelo de Poisson. No obstante, la estadística *AIC* y la estadística *BIC* nos informan de que el modelo no es mejor que el binomial negativo utilizado anteriormente.

El modelo Poisson-inverso Gaussiano

El modelo Poisson-inverso Gaussiano (PIG) es una combinación de las distribuciones Poisson e inversa-Gaussiana. El algoritmo de modelación es recursivo y considerablemente complejo. De hecho existen muy pocas implementaciones del paquete estadístico del modelo.

```
. pigreg days a4 aborig slow girl, nolog irr vce(ro)

Poisson-Inverse Gaussian regression                         Number of obs = 146
                                                               Wald chi2(4) = 2041.09
                                                               Prob > chi2 = 0.0000
Log pseudolikelihood = -556.07488
```

days		IRR	Robust Std. Err.	z	P> z	[95% Conf. Interval]
a4	1.39857	.0297321	15.78	0.000	1.341494	1.458075
aborig	2.022934	.052767	27.01	0.000	1.922111	2.129044
slow	1.186465	.0196503	10.32	0.000	1.14857	1.225611
girl	.8164273	.0112542	-14.71	0.000	.7946648	.8387858
_cons	10.73001	.4199487	60.63	0.000	9.937705	11.58549
/lnalpha	.2366043	.1886122			-.1330689	.6062774
alpha	1.26694	.2389603			.8754048	1.833593

```
. abic
AIC Statistic = 7.699656          AIC*n = 1124.1498
BIC Statistic = 7.764732          BIC(Stata) = 1142.0514
```

Los valores p de los cocientes de tasas son todos significativos, utilizando errores estándar robustos. No obstante, los estadísticos AIC y BIC indican un posible ajuste deficiente en comparación con el binomial negativo. Omitir *girl* del modelo resulta en una caída de cuatro puntos en la estadística BIC . Por otro lado, para determinar el ajuste comparativo del modelo PIG en NB, será necesario comparar las diferencias entre los recuentos observados y predichos en todo el rango de datos.

Nótese que los modelos GP y PIG también pueden estimarse como modelos de cero inflado. En este caso, sin embargo, no se ha producido ninguna mejora con relación a los modelos no inflados.

Los modelos de tres parámetros

Existen modelos binomiales negativos de tres parámetros que ofrecen a los modelos un mayor rango para la estimación de los parámetros de la distribución subyacente del modelo. Los modelos están disponibles en Stata para el modelo NB-P, el cual parametriza el exponente del segundo término en la función de varianza, $\mu + \alpha\mu^P$ (Hardin & Hilbe, 2014), el modelo binomial negativo Waring y el modelo binomial negativo Famoye (Harris, Hilbe, & Hardin, 2014). La P en el modelo NB-P representa que la potencia del segundo término en la función de varianza es un parámetro a estimar. No obstante, el símbolo tradicional para el parámetro es ρ , o rho, y no P.

El uso principal de los modelos de recuento de tres parámetros reside en la modelación de datos binomiales negativos que, de no ser así, sería sobredispersa. Los modelos de tres parámetros se usan generalmente para realizar ajustes en datos altamente sobredispersos; eso es, datos con una variabilidad excesiva en los datos que el modelo binomial negativo típico o, alternativamente, los modelos de dos parámetros no pueden gestionar. No obstante, los modelos de recuento de tres parámetros se han diseñado para ser utilizados con una variedad de propósitos.

Los modelos de recuento jerárquicos

En Stata, SAS y R existe el complemento completo de modelos de recuento multinivel, de pendiente aleatoria y de intercepción aleatoria que se utilizan para realizar ajustes en la sobredispersión, lo cual causa que los datos se estructuren en paneles o en algún modo longitudinal. Los modelos GEE, de Poisson y binomial negativo también están disponibles en estos paquetes, así como en SPSS.

Otros modelos de recuento

Los estadistas también han creado modelos de recuento adicional y cuantil generalizados para datos de recuentos no paramétricos, una regresión de Poisson exacta para pequeños recuentos de datos desequilibrados, modelos de combinación finita para recuentos generados a partir de dos o tres fuentes autónomas y modelos de recuentos truncados y censurados (Hardin & Hilbe, 2015). Cada uno de estos modelos ha sido examinado en Hilbe (2011).

Los modelos de recuento bayesianos

La metodología bayesiana se está popularizando rápidamente entre estadistas e investigadores en una amplia variedad de disciplinas. A finales del siglo pasado, la gran mayoría de investigadores que se ocuparon de la modelación bayesiana se vieron obligados a solucionar los posteriores bayesianos a través de métodos analíticos. Esto podía resultar realmente difícil y, por lo general, acabar relegando la modelación a tareas más bien sencillas. No obstante, a partir de unos años más tarde, varios paquetes estadísticos empezaron a ofrecer una modelación bayesiana basada en los métodos Montecarlo mediante el uso de cadenas Markov (*MCMC*), los cuales permiten realizar una estimación de la distribución posterior bayesiana a través de un destacable algoritmo de muestreo. JAGS, WinBUGS, OpenBUGS, Stan, Python, MLwiN, SAS, SPSS, Stata y otros paquetes se han convertido en paquetes de gran popularidad que han ampliado la modelación bayesiana a través de la comunidad de investigadores.

Lanzado al mercado en el año 2000, WinBUGS fue el primer paquete general de muestreo que empleó los métodos Montecarlo mediante el uso de cadenas Markov (*MCMC*) para la modelación bayesiana. Otros paquetes parecidos siguieron los mismos pasos. Stata ofreció su primer paquete bayesiano en 2015.

La modelación bayesiana puede prácticamente duplicar los modelos de máxima verosimilitud, cuando los valores a priori vagos, difusos o los así llamados ‘no informativos’ se emplean para crear una combinación con la verosimilitud del modelo. A diferencia del modelado de máxima verosimilitud de base frecuentista, los parámetros del modelo bayesiano están considerados de distribución aleatoria. Es decir, se emplea un algoritmo de muestreo para determinar la distribución posterior de cada uno de los parámetros del modelo.

A continuación se muestra el uso del comando *bayesmh* de Stata para realizar una estimación de los parámetros posteriores del binomial negativo utilizando priores no informativos o planos. El modelo superior (Binomial negativo) es un modelo binomial típico al que le sigue el modelo bayesiano de datos.

EL MODELO BINOMIAL NEGATIVO MLE

```
. nbreg days a4 aborig girl slow, nolog vce(ro)
```

days		Robust					
		Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
a4	.4580832	.1977657	2.32	0.021	.0704694	.8456969	
aborig	.59763	.1516423	3.94	0.000	.3004165	.8948435	
slow	.2721409	.180343	1.51	0.131	-.0813249	.6256066	
girl	-.2162921	.1575045	-1.37	0.170	-.5249952	.092411	
_cons	2.356878	.1728699	13.63	0.000	2.01806	2.695697	
/lnalpha	-.1929157	.125174			-.4382523	.0524209	
alpha	.8245515	.1032124			.645163	1.053819	

EL MODELO BINOMIAL NEGATIVO BAYESIANO

```

bayesmh days, nocons likelihood(llf(
    ln(nbinomialp(exp(-{lnalpha}), days,
    1/(1+exp({days: a4 aborig slow girl}+{days:_cons})))))) ///
prior({days:a4 aborig slow girl _cons}{lnalpha}, flat) rseed(1944)

Bayesian regression
Random-walk Metropolis-Hastings sampling
MCMC iterations = 12,500
Burn-in = 2,500
MCMC sample size = 10,000
Number of obs = 146
Acceptance rate = .1822
Efficiency: min = .03343
-----
| Equal-tailed
| Mean Std. Dev. MCSE Median [95% Cred. Interval]
-----+
lnalpha | -.1520669 .1209699 .006491 -.1578817 -.3710215 .1024264
-----+
days |
    a4 | .4510724 .2278986 .013421 .4435438 .0004725 .8943773
    aborig | .5889158 .1644661 .01003 .582773 .2859969 .9122398
    slow | .2566763 .1899522 .011947 .2546729 -.1253877 .626903
    girl | -.2139218 .1642201 .006388 -.2112521 -.5433723 .1210683
    _cons | 2.232879 .2258881 .011743 2.225916 1.795334 2.715911
-----+
. di exp(-.1520669 )
.85893082

```

Las medias de las respectivas distribuciones posteriores se acercan a las estimaciones de parámetro de verosimilitud máxima. En la siguiente tabla se muestra la estadística del Criterio Informativo de la Desviación (*DIC*). El *DIC* es parecido a las estadísticas *AIC* y *BIC*, aunque se utiliza específicamente para los modelos bayesianos. En la segunda tabla a continuación se muestra un listado de tamaños de muestras efectivas (*ESS*, por sus siglas en inglés), los tiempos de correlación y las eficiencias relativas a cada uno de los parámetros del modelo. La estadística *ESS* representa el tamaño de la muestra efectiva para cada parámetro y se compara con el tamaño de muestreo *MCMC* del modelo, que aquí es de 10,000. Los posteriores se determinan en función del *ESS*. En principio, cuanto más cerca esté el valor *ESS* del tamaño de la muestra de *MCMC*, mejor.

Los tiempos de correlación son la inversa de las eficiencias que aparecen listadas en la última columna de la derecha. Por norma general, serían preferibles unos tiempos de correlación inferiores y unos valores correspondientes de eficiencia más altos. Se considera como especialmente buenas unas eficiencias superiores al 10%. Para nuestro modelo, la mayor es la de *aborig* con una eficiencia del 5.9%. Nótese que, si volvemos a ejecutar el modelo con un valor inicial (en inglés, *seed*) diferente, en ocasiones obtendremos eficiencias con diferencias más bien sustanciales. Se recomienda remuestrear con más iteraciones y, quizás, con un tamaño de muestra *burnin* mayor que el de por defecto. Los valores *burnin* son iteraciones de muestreo que no se utilizan en el cálculo de la distribución posterior. Esto es importante porque es probable que las iteraciones tempranas difieran considerablemente de la distribución subyacente y sesguen los resultados.

```
. bayesstats ic

Bayesian information criteria
-----
|      DIC    log (ML)   log (BF)
-----
active | 1111.725 -556.4729 .
-----

Note: Marginal likelihood (ML) is computed
using Laplace-Metropolis approximation.

. bayesstats ess

Efficiency summaries      MCMC sample size = 10,000
-----
|      ESS   Corr. time   Efficiency
-----
lnalpha | 334.28      29.91      0.0334
-----
days |
    a4 | 351.39      28.46      0.0351
  aborig | 590.44      16.94      0.0590
    slow | 390.83      25.59      0.0391
    girl | 539.99      18.52      0.0540
  _cons | 523.93      19.09      0.0524
-----
```

Por último, tanto los gráficos correspondientes a la forma de cada uno de los parámetros como los diagramas de trazos son diagnósticos importantes. Más abajo se visualiza un comando Stata que muestra gráficos de diagnóstico para *aborig* y el parámetro de dispersión *lnalpha*, aunque sin resultados.

```
. bayesgraph diagnostics {days: aborig} {lnalpha} , histopts(normal)
```

La familia de funciones *MCMCpack* de R ofrece a los usuarios de R varios modelos bayesianos. La mayoría de usuarios de R emplean JAGS o Stan dentro del mismo entorno R. Esto proporciona a los usuarios una capacidad de modelado muy superior a JAGS o Stan de forma independiente. Los procedimientos *MCMC* y Genmod de SAS ofrecen unas excelentes herramientas de modelación bayesiana SAS. La aproximación anidada integrada de Laplace (INLA, por sus siglas en inglés) es un nuevo método bayesiano que resulta óptimo para la modelación bayesiana de datos jerárquicos y espaciales. Este es un nuevo campo en proceso de desarrollo que solo puede ofrecer beneficios a aquellos que estén interesados en la modelación sofisticada de datos de recuento.

Resumen

Esta descripción general del modelado de datos de recuento ha ofrecido un breve resumen sobre cómo se desarrollan e interpretan los modelos de recuento. Cuando nos encontramos con datos de recuento, es importante evaluar primero la media y el rango de la variable de recuento que se está modelando y determinar su estado de dispersión. La selección de los predictores del modelo se lleva a cabo de

manera típica. La manera más sencilla de determinar la dispersión es observando el estadístico de dispersión de Pearson de un modelo de Poisson de datos. Los valores de la dispersión que son sustancialmente superiores a 1 indican que los datos son sobredispersos. En este caso, es probablemente mejor modelar los datos utilizando un modelo binomial negativo. Si los datos son infradispersos, entonces podrá utilizarse un modelo valla o de Poisson generalizado para realizar la estimación de los parámetros. Si el modelo no muestra sobredispersión alguna, entonces será probablemente preferible emplear un modelo de Poisson.

Cuando los datos tienen más recuentos de cero de lo aceptable en base a las asunciones distribucionales del modelo, deberá intentarse utilizar un modelo de valla o de cero inflado. Las pruebas de diagnóstico muestran al analista si el modelo de cero inflado está mejor ajustado que el modelo típico. Es más, y con el fin de evaluar el ajuste, deberán usarse las pruebas que comparan los recuentos observados en relación a los recuentos predichos a lo largo de todo el rango de recuentos en el modelo. Los estadísticos *AIC* y *BIC* pueden emplearse para seleccionar el modelo mejor ajustado. No obstante, es importante comprobar varias pruebas de ajuste en vez de únicamente confiar en una.

A veces, ni el modelo de Poisson ni el modelo binomial negativo se ajustan a los datos. En dichos casos podrán utilizarse otros modelos como, por ejemplo, el modelo de Poisson generalizado, el modelo Poisson-inverso Gaussiano o un modelo de recuento de tres parámetros. También existen otros modelos, algunos de los cuales no se han mencionado en esta descripción general. Es recomendable que los analistas se familiaricen con los varios modelos de recuento existentes y seleccionen aquel que mejor se ajuste a sus datos de estudio. Para consultar un examen en mayor profundidad de todo el rango de modelos de recuento con ejemplos elaborados, véase Hilbe (2011) y Hilbe (2014). Los ejemplos se proporcionan en Stata, R y SAS.

Acknowledgments / Agradecimientos

With sincere gratitude, we would like to thank both Gabriel Liberman and Amir Hefetz from Data Graph - Research and Statistical Consulting, the translation team and the publisher for making the translation of this article to Spanish a reality. Without all of the help, this would not have been possible. The Hilbe Family. / Queremos expresar nuestro más sincero agradecimiento a Gabriel Liberman y Amir Hefetz de Data Graph - Investigación y Consultoría Estadística, así como a los equipos de traducción y edición, sin vuestra ayuda la traducción de este artículo al español no se hubiera hecho realidad. La familia Hilbe.

Disclosure statement

No potential conflict of interest was reported by the author / Los autores no han referido ningún potencial conflicto de interés en relación con este artículo.

References / Referencias

- Aitkin, M. (1978). The analysis of unbalanced cross-classifications. *Journal of the Royal Statistical Society, Series A*, 41, 195–223. doi:[10.2307/2344453](https://doi.org/10.2307/2344453)

- Hardin, J. W., & Hilbe, J. M. (2012). *Generalized linear models and extensions* (3rd ed.). College Station, TX: Stata Press.
- Hardin, J. W., & Hilbe, J. M. (2014). Regression models for count data based on the negative binomial(p) distribution. *Stata Journal*, 14, 280–291.
- Hardin, J. W., & Hilbe, J. M. (2015). Regression models for count data from truncated distributions. *Stata Journal*, 15, 226–246.
- Harris, T., Hilbe, J. M., & Hardin, J. W. (2014). Modeling count data with generalized distributions. *Stata Journal*, 14, 562–579.
- Hilbe, J. M. (2011). *Negative binomial regression* (2nd ed.). Cambridge, UK: Cambridge University Press.
- Hilbe, J. M. (2014). *Modeling count data*. Cambridge, UK: Cambridge University Press.
- Quine, S. (1973). *Achievement orientation of aboriginal and white Australian adolescents* (PhD dissertation). Australian National University, Canberra, Australia.