The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

The *Stata Journal* is indexed and abstracted by *CompuMath Citation Index*, *Current Contents/Social and Behavioral Sciences*, *RePEc: Research Papers in Economics*, *Science Citation Index Expanded* (also known as *SciSearch*), *Scopus*, and *Social Sciences Citation Index*.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

**Subscription rates** listed below include both a printed and an electronic copy unless otherwise mentioned.

| U.S. and Canada | | Elsewhere | |
|---|---|---|---|
| **Printed & electronic** | | **Printed & electronic** | |
| 1-year subscription | $ 98 | 1-year subscription | $138 |
| 2-year subscription | $165 | 2-year subscription | $245 |
| 3-year subscription | $225 | 3-year subscription | $345 |
| 1-year student subscription | $ 75 | 1-year student subscription | $ 99 |
| 1-year institutional subscription | $245 | 1-year institutional subscription | $285 |
| 2-year institutional subscription | $445 | 2-year institutional subscription | $525 |
| 3-year institutional subscription | $645 | 3-year institutional subscription | $765 |
| **Electronic only** | | **Electronic only** | |
| 1-year subscription | $ 75 | 1-year subscription | $ 75 |
| 2-year subscription | $125 | 2-year subscription | $125 |
| 3-year subscription | $165 | 3-year subscription | $165 |
| 1-year student subscription | $ 45 | 1-year student subscription | $ 45 |

Back issues of the *Stata Journal* may be ordered online at

http://www.stata.com/bookstore/sjj.html

Individual articles three or more years old may be accessed online without charge. More recent articles may be ordered online.

http://www.stata-journal.com/archives.html

The *Stata Journal* is published quarterly by the Stata Press, College Station, Texas, USA.

Address changes should be sent to the *Stata Journal*, StataCorp, 4905 Lakeway Drive, College Station, TX 77845, USA, or emailed to sj@stata.com.



The *Stata Journal* (ISSN 1536-867X) is a publication of Stata Press. Stata, Stata, Stata Press, Mata, Mata, and NetCourse are registered trademarks of StataCorp LP.

# Modeling count data with generalized distributions

Tammy Harris
Institute for Families in Society
University of South Carolina
Columbia, SC
harris68@mailbox.sc.edu

Joseph M. Hilbe
School of Social and Family Dynamics
Arizona State University
Tempe, AZ
hilbe@asu.edu

James W. Hardin
Institute for Families in Society
Department of Epidemiology and Biostatistics
University of South Carolina
Columbia, SC
jhardin@sc.edu

**Abstract.** We present motivation and new commands for modeling count data. While our focus is to present new commands for estimating count data, we also discuss generalized binomial regression and present the zero-inflated versions of each model.

**Keywords:** st0351, gbin, zigbin, nbregf, nbregw, zinbregf, zinbregw, binomial, Waring, count data, overdispersion, underdispersion

## 1 Introduction

We introduce programs for regression models of count data. Poisson regression analysis is widely used to model such response variables because the Poisson model assumes equidispersion (equality of the mean and variance). In practice, equidispersion is rarely reflected in data. In most situations, the variance exceeds the mean. This occurrence of extra-Poisson variation is known as overdispersion (see, for example, Dean [1992]). In situations where the variance is smaller than the mean, data are characterized as being underdispersed. Modeling underdispersed count data with inappropriate models can lead to overestimated standard errors and misleading inference. While there are various approaches for modeling overdispersed count data, such as the negative binomial distributions and other mixtures of Poisson (Yang et al. 2007; Hilbe 2014), there are few models for underdispersed count data. Harris, Yang, and Hardin (2012) introduced a generalized Poisson regression command to handle underdispersed count data.

As stated earlier, count data can be analyzed using regression models based on the Poisson distribution. However, in this article, we will discuss other discrete regression models that can be used, such as the generalized negative binomial distribution, which was described by Jain and Consul (1971) and later by Consul and Gupta (1980). The distribution was also investigated by Famoye (1995), who illustrated a use for analyzing grouped binomial data.

 

The generalized binomial regression model is a simplification based on the generalized negative binomial distribution for which we treat one of the parameters as the known denominator of proportional (grouped binomial) outcomes. The properties and utility of the distribution for regression models for count and grouped binomial data are discussed in Jain and Consul (1971), Consul and Gupta (1980), and Famoye (1995).

Another extension of the negative binomial distribution is the univariate generalized Waring distribution, or the beta negative binomial distribution. The present generalized Waring distribution was proposed and used by Irwin (1968) to model accident count data. An advantage of this model over the negative binomial model is that investigators can separate the unobserved heterogeneity from the internal factors of each individual's characteristics and external factors (covariates) that may affect the variability of data (confounding). For more technical and historical information on the distribution and associated regression models, see Rodríguez-Avi et al. (2009), Irwin (1968), and Hilbe (2011).

To distinguish the origins of specific regression models, we use NBREGF for count models based on the generalized negative binomial distribution, GBIN for grouped binomial models based on a simplification of the generalized negative binomial distribution, and NBREGW for count models based on the generalized Waring distribution.

Many applications of the NBREGF regression model have been illustrated in studies involving medicine, ecology, physics, etc. Wang et al. (2012) used the NBREGF model to analyze a rehabilitation program study that evaluated brain function in stroke patients by using functional magnetic resonance imaging. Hardin and Hilbe (2012) presented an example that used microplot data of carrot fly damage. For this example, the authors analyzed these data by using Stata's suite of `ml()` functions and developed syntax for the GBIN regression. Lastly, Rodríguez-Avi et al. (2009) used the NBREGW regression model to model the number of goals scored by football players, and they compared the results with the results of a regression model based on the negative binomial distribution.

Herein, we illustrate modeling count data using the NBREGF, GBIN, and NBREGW regression models. This article is organized as follows. In section 2, we review the three count-data regression models and their zero-inflated versions. In section 3, we present the syntax for the new commands. In section 4, we present a real-world data example. Finally, in section 5, we give a summary. We also present software that we enhanced from Hardin and Hilbe (2012) to fit NBREGF and GBIN models.

# 2 The models

## 2.1 Generalized negative binomial: Famoye

As implemented in the accompanying software, the NBREGF model assumes that $\theta$ is a scalar unknown parameter. Thus the probability mass function (PMF), mean, and variance are given by

$$P(Y=y)=\frac{\theta}{\theta+\phi y}\binom{\theta+\phi y}{y}\mu^y\left(1-\mu\right)^{\theta-y+\phi y} \tag{1}$$

where $0<\mu<1$, $1\leq\phi<\mu^{-1}$ for $\theta>0$ and nonnegative outcomes $y_i\in(0,1,2,\ldots)$.

$$\begin{aligned} E(Y) &= \theta\mu(1-\phi\mu)^{-1} \\ V(Y) &= \theta\mu\left(1-\mu\right)(1-\phi\mu)^{-3} \end{aligned}$$

The main differences from the GBIN model are that the parameter $\theta$ is an unknown parameter in (1) but a known parameter in (2) and that $\sigma=\phi>1$. In the limit $\phi\to 1$, the variance approaches that of the negative binomial distribution. Thus the $\phi$ parameter generalizes the negative binomial distribution in the NBREGF model to have greater variance than is allowed in a negative binomial regression model. To construct a regression model, we implemented the log link $\log(\mu)=x\beta$ to make results comparable to Poisson and negative binomial models.

## 2.2 Generalized binomial

The generalized binomial regression model is based on a simplification of the generalized negative binomial distribution. We assume that the $\theta$ parameter in (1) is a vector of observation-specific known constants $n$ (they are the denominators of grouped binomial data), $\sigma=\phi$, and $\mu$ is replaced with $\pi/(1+\phi\pi)$. When $\theta$ is known, the $\sigma$ parameter is nonnegative, while in the generalized negative binomial distribution, $\phi>1$. Under these changes, the PMF, mean, and variance are given by

$$P(Y=y)=\frac{n}{n+\sigma y}\binom{n+\sigma y}{y}\left(\frac{\pi}{1+\sigma\pi}\right)^y\left(1-\frac{\pi}{1+\sigma\pi}\right)^{n-y+\sigma y} \tag{2}$$

$$\begin{aligned} E(Y) &= n\frac{\pi}{1+\sigma\pi}\left(1-\frac{\pi}{1+\sigma\pi}\sigma\right)^{-1} \\ &= n\pi \\ V(Y) &= n\frac{\pi}{1+\sigma\pi}\left(1-\frac{\pi}{1+\sigma\pi}\sigma\right)(1+\pi\sigma)^{-3} \\ &= n\pi(1+\sigma\pi)(1+\sigma\pi-\pi) \end{aligned}$$

Parameterizing $g(\pi)=x\beta$, where $g(\cdot)$ is a suitable link function assuming that $\pi$ plays the role of the probability of success, we obtain results that coincide with a grouped data binomial model. The variance is equal to binomial variance if $\sigma=0$, and it is equal to negative binomial variance if $\sigma=1$. Thus the $\sigma>0$ parameter generalizes the binomial distribution in the GBIN regression model.

## 2.3 Generalized Waring

As illustrated in Irwin (1968), the generalized Waring distribution can be constructed under the following specifications:

i. $Y|x, \lambda_x, v \sim \text{Poisson}(\lambda_x)$

ii. $\lambda_x|v \sim \text{Gamma}(a_x, v)$

iii. $v \sim \text{Beta}(\rho, k)$

In the author's presentation for accident data, he specifies $\lambda|v$ as "accident liability" and $v$ as "accident proneness". The PMF is ultimately given by

$$P(Y = y) = \frac{\Gamma(a_x + \rho)\Gamma(k + \rho)}{\Gamma(\rho)\Gamma(a_x + k + \rho)} \frac{(a_x)_y (k)_y}{(a_x + k + \rho)_y} \frac{1}{y!}$$

where $k, \rho, a_x > 0$, $a_x = \mu(\rho - 1)/k$, and $(a)_w$ is the Pochhammer notation for $\Gamma(a + w)/\Gamma(w)$ if $a > 0$. The expected value and variance of the distribution are

$$\begin{aligned} E(Y) &= \frac{a_x k}{\rho - 1} = \mu \\ V(Y) &= \mu + \mu\left(\frac{k + 1}{\rho - 2}\right) + \mu^2\left\{\frac{k + \rho - 1}{k(\rho - 2)}\right\} \end{aligned} \tag{3}$$

where $a_x, k > 0$ and $\rho > 2$ (to ensure nonnegative variance). To construct a regression model, we implemented the log link $\log(\mu) = x\beta$ to make results comparable to Poisson and negative binomial models. A unique characteristic of this model occurs when the data are from a different underlying distribution. For instance, when the data are from a Poisson distribution with $V(Y) = \mu$, it indicates that $(k + 1)/(\rho - 2) \to 0$ and $\{k + \rho - 1\}/\{k(\rho - 2)\} \to 0$ then $k, \rho \to \infty$. Also, if the data have an underlying NB-2 (negative binomial-2) distribution with $V(Y) = \mu + \alpha\mu^2$ (where $\alpha$ is the dispersion parameter), it indicates that $(k + 1)/(\rho - 2) \to 0$ and $\{k + \rho - 1\}/\{k(\rho - 2)\} \to \alpha$, where $k \to 1/\alpha$ and $\rho \to \infty$.

## 2.4 Zero inflation

When there is an excess of zeros in count-response data, Poisson (and other) distribution models may not be appropriate to use. Hardin and Hilbe (2012) describe the two origins of zero outcomes: 1) individuals who do not enter into the counting process and 2) individuals who enter into the counting process and have a zero outcome. Therefore, the model must be separated into different parts, one consisting of a zero count $y = 0$ and the other consisting of a nonzero count $y > 0$. The zero-inflated model is given by

$$P(Y = y) = \begin{cases} p + (1 - p)f(y) & y = 0 \\ (1 - p)f(y) & y = 1, 2, \ldots \end{cases}$$

where $p$ is the probability that the binary process results in a zero outcome, $0 \leq p < 1$, and $f(y)$ is the probability function. Zero-inflation models are proposed for the NBREGF, GBIN, and NBREGW distributions.

# 3 Syntax

The accompanying software includes the command files as well as supporting files for prediction and help. In the following syntax diagrams, unspecified *options* include the usual collection of maximization and display options available to all estimation commands. In addition, all zero-inflated commands include the `ilink(`*linkname*`)` option to specify the link function for the inflation model. The generalized binomial model for grouped binomial data also includes the `link(`*linkname*`)` option for linking the probability of success to the linear predictor. Supported *linkname*s include `logit`, `probit`, `loglog`, and `cloglog`.

The syntax for specifying a generalized binomial regression model for grouped data is given by

`gbin` *depvar* [ *indepvars* ] [ *if* ] [ *in* ] [ *weight* ] [ `,` *options* ]

and the syntax for the zero-inflated version is given by

`zigbin` *depvar* [ *indepvars* ] [ *if* ] [ *in* ] [ *weight* ] `,`
`inflate(`*varlist*[`, offset(`*varname*`)`] | `_cons)` [ `vuong` *options* ]

The syntax for fitting a generalized negative binomial regression model where the distribution is assumed to follow Famoye's description is given by

`nbregf` *depvar* [ *indepvars* ] [ *if* ] [ *in* ] [ *weight* ] [ `,` *options* ]

The syntax for fitting a generalized negative binomial regression model where the distribution is derived from the Waring distribution is given by

`nbregw` *depvar* [ *indepvars* ] [ *if* ] [ *in* ] [ *weight* ] [ `,` *options* ]

The syntax for specifying a zero-inflated count model where the count distribution follows that described by Famoye is given by

`zinbregf` *depvar* [ *indepvars* ] [ *if* ] [ *in* ] [ *weight* ] `,`
`inflate(`*varlist*[`, offset(`*varname*`)`] | `_cons)` [ `vuong` *options* ]

The syntax for specifying a zero-inflated count model where the count distribution follows the Waring distribution is given by

`zinbregw` *depvar* [ *indepvars* ] [ *if* ] [ *in* ] [ *weight* ] `,`
`inflate(`*varlist*[`, offset(`*varname*`)`] | `_cons)` [ `vuong` *options* ]

A Vuong test (see Vuong [1989]) evaluates whether the regression model with zero inflation or the regression model without zero inflation is closer to the true model. A

random variable $\omega$ is defined as the vector $\log L_Z - \log L_S$, where $L_Z$ is the likelihood of the zero-inflated model evaluated at its maximum likelihood estimation, and $L_S$ is the likelihood of the standard (nonzero-inflated) model evaluated at its maximum likelihood estimation. The vector of differences over the $N$ observations is then used to define the statistic

$$\boldsymbol{V} = \frac{\sqrt{N}\overline{\omega}}{\sqrt{\sum(\omega - \overline{\omega})^2/(N-1)}}$$

which, asymptotically, is characterized by a standard normal distribution. A significant positive statistic indicates preference for the zero-inflated model, and a significant negative statistic indicates preference for the model without zero inflation. Nonsignificant Vuong statistics indicate no preference for either model. Results of this test are included in a footnote to the estimation of the model when the user includes the `vuong` option in any of the zero-inflated commands. Vuong statistics with corrections based on the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) are also displayed in the output (see Desmarais and Harden [2013] for details). They are displayed for each of the zero-inflated models discussed in this article.

# 4 Example

We shall use the popular German health data for the year 1984 as example data. The goal of our model is to understand the number of visits made to a physician during 1984. Our predictor of interest is whether the patient is highly educated based on achieving a graduate degree, for example, an MA or MS, an MBA, a PhD, or a professional degree. Confounding predictors are age (from 25–64) and income in German Marks, divided by 10. We first model the data using Poisson regression. The `glm` command is used to determine the Pearson dispersion, or dispersion statistic, which is not available using the `poisson` command.

```
. use rwm1984, clear
(German health data for 1984; Hardin & Hilbe, GLM and Extensions, 3rd ed)
. gen hh = hhninc/10
. glm docvis edlevel4 age hh, nolog eform fam(poisson)
Generalized linear models                         No. of obs      =       3874
Optimization     : ML                             Residual df     =       3870
                                                  Scale parameter =          1
Deviance         =  24369.36065                   (1/df) Deviance =   6.296992
Pearson          =  44032.57716                   (1/df) Pearson  =   11.37793
Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = ln(u)                   [Log]
                                                  AIC             =   8.120749
Log likelihood   = -15725.89176                   BIC             =  -7604.745
```

| docvis | IRR | OIM Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| edlevel4 | .7887207 | .0380651 | -4.92 | 0.000 | .7175343 | .8669693 |
| age | 1.026209 | .0008362 | 31.75 | 0.000 | 1.024571 | 1.027849 |
| hh | .3468308 | .0257417 | -14.27 | 0.000 | .299876 | .4011378 |
| _cons | 1.326749 | .0608884 | 6.16 | 0.000 | 1.212619 | 1.451619 |

```
. estat ic
Akaike´s information criterion and Bayesian information criterion
```

| Model | Obs | ll(null) | ll(model) | df | AIC | BIC |
|---|---|---|---|---|---|---|
| . | 3874 | . | -15725.89 | 4 | 31459.78 | 31484.83 |

Note: N=Obs used in calculating BIC; see [R] BIC note

```
. nbreg docvis edlevel4 age hh, nolog irr
Negative binomial regression                      Number of obs   =       3874
                                                  LR chi2(3)      =     161.23
Dispersion     = mean                             Prob > chi2     =     0.0000
Log likelihood = -8344.5927                       Pseudo R2       =     0.0096
```

| docvis | IRR | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| edlevel4 | .7265669 | .0837908 | -2.77 | 0.006 | .5795774 | .9108351 |
| age | 1.026037 | .0023731 | 11.11 | 0.000 | 1.021397 | 1.030699 |
| hh | .4487569 | .0718929 | -5.00 | 0.000 | .327827 | .6142958 |
| _cons | 1.246529 | .1453412 | 1.89 | 0.059 | .991871 | 1.56657 |
| /lnalpha | .8413514 | .0308101 | | | .7809646 | .9017381 |
| alpha | 2.319499 | .0714641 | | | 2.183578 | 2.463882 |

```
Likelihood-ratio test of alpha=0:  chibar2(01) = 1.5e+04 Prob>=chibar2 = 0.000
```

```
. estat ic
Akaike´s information criterion and Bayesian information criterion
```

| Model | Obs | ll(null) | ll(model) | df | AIC | BIC |
|---|---|---|---|---|---|---|
| . | 3874 | -8425.206 | -8344.593 | 5 | 16699.19 | 16730.5 |

```
Note:  N=Obs used in calculating BIC; see [R] BIC note
```

The AIC and BIC statistics are substantially lower here than they are for the Poisson model, indicating a much better fit than the Poisson model.

```
. display 1/exp(_b[edlevel4])
1.3763358
```

Patients without a graduate education are 38% more likely to see a physician than are patients with a graduate education. We can likewise affirm that patients without a graduate education saw a physician 38% more often in 1984 than patients with a graduate education.

The negative binomial model did not adjust for all the correlation, or dispersion, in the data.

```
. quietly glm docvis edlevel4 age hh, fam(nbin ml)
. display e(dispers_p)
1.4017258
```

This is perhaps due to the excessive number of times a patient in the data never saw a physician in 1984. A tabulation of `docvis` shows that nearly 42% of the 3,874 patients in the data did not visit a physician. This value is far greater than the one accounted for by the Poisson and negative binomial distributional assumptions.

```
. count if docvis==0
 1611
. display "Zeros account for " %4.2f (r(N)*100/3874) "% of the outcomes"
Zeros account for 41.58% of the outcomes
```

Given the excess zero counts in `docvis`, it may be wise to employ a zero-inflated regression model on the data. At the least, we can determine which predictors tend to prevent patients from going to the doctor.

```
. zinb docvis edlevel4 age hh, nolog inflate(edlevel4 age hh) irr
Zero-inflated negative binomial regression        Number of obs   =       3874
                                                  Nonzero obs     =       2263
                                                  Zero obs        =       1611

Inflation model = logit                           LR chi2(3)      =      98.50
Log likelihood  = -8330.799                       Prob > chi2     =     0.0000
```

| docvis | IRR | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| docvis | | | | | | |
| edlevel4 | .9176719 | .1289238 | -0.61 | 0.541 | .6967903 | 1.208573 |
| age | 1.020511 | .0025432 | 8.15 | 0.000 | 1.015538 | 1.025508 |
| hh | .4506524 | .0720932 | -4.98 | 0.000 | .3293598 | .6166132 |
| _cons | 1.768336 | .2419851 | 4.17 | 0.000 | 1.352333 | 2.31231 |
| inflate | | | | | | |
| edlevel4 | 1.174194 | .3519899 | 3.34 | 0.001 | .4843067 | 1.864082 |
| age | -.0521002 | .0115586 | -4.51 | 0.000 | -.0747547 | -.0294458 |
| hh | .2071444 | .570265 | 0.36 | 0.716 | -.9105545 | 1.324843 |
| _cons | -.037041 | .4438804 | -0.08 | 0.933 | -.9070305 | .8329486 |
| /lnalpha | .6203884 | .0662583 | 9.36 | 0.000 | .4905245 | .7502522 |
| alpha | 1.85965 | .1232172 | | | 1.633173 | 2.117534 |

```
. estat ic
Akaike´s information criterion and Bayesian information criterion
```

| Model | Obs | ll(null) | ll(model) | df | AIC | BIC |
|---|---|---|---|---|---|---|
| . | 3874 | -8380.051 | -8330.799 | 9 | 16679.6 | 16735.96 |

```
Note:  N=Obs used in calculating BIC; see [R] BIC note
```

The AIC statistic is 20 points lower in the zero-inflated model but 5 points higher for the BIC statistic. However, variables `edlevel4` and `age` appear to affect zero counts, with younger graduate patients more likely to not see a physician at all during the year. Given the zero-inflated model, patients without a graduate education see the physician 9% more often than patients with a graduate education.

```
. display 1/exp(_b[edlevel4])
1.0897141
```

Because excess zero counts did not appear to bear on extra correlation in the data, there may be other factors. We employ a generalized Waring negative binomial model to further identify the source of extra dispersion.

## 4.1 Generalized negative binomial: Waring

```
. nbregw docvis edlevel4 age hh, nolog eform
Generalized negative binomial-W regression      Number of obs   =       3874
                                                LR chi2(3)      =     163.80
Log likelihood = -8315.421                      Prob > chi2     =     0.0000
```

| docvis | IRR | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| edlevel4 | .6910153 | .0865378 | -2.95 | 0.003 | .5406164 | .8832549 |
| age | 1.027732 | .0024925 | 11.28 | 0.000 | 1.022859 | 1.032629 |
| hh | .4693135 | .086958 | -4.08 | 0.000 | .3263967 | .674808 |
| _cons | 1.142679 | .1431097 | 1.06 | 0.287 | .8939621 | 1.460593 |
| /lnrhom2 | .9045584 | .1992573 | | | .5140212 | 1.295096 |
| /lnk | -.6113509 | .0521974 | | | -.7136559 | -.5090458 |
| rho | 4.470841 | .4923331 | | | 3.672001 | 5.651345 |
| k | .5426174 | .0283232 | | | .4898501 | .6010688 |

```
. estat ic
Akaike´s information criterion and Bayesian information criterion
```

| Model | Obs | ll(null) | ll(model) | df | AIC | BIC |
|---|---|---|---|---|---|---|
| . | 3874 | -8397.319 | -8315.421 | 6 | 16642.84 | 16680.41 |

```
Note:  N=Obs used in calculating BIC; see [R] BIC note
```

The AIC and BIC statistics are substantially lower here than for either the negative binomial or zero-inflated version. For the calculated $\rho$ and $k$, the $V(Y) = \mu + 0.624\mu + 2.994\mu^2$, where $\mu$ is the mean. Here we see that the term $\{k + \rho - 1\}/\{k(\rho - 2)\} = 2.994$, from (3), is close to the dispersion parameter $\alpha = 2.319$ when using an NB-2 regression model from above. More information on the background of this model can be found in Hilbe (2011).

To address the excess zeros in the outcome, we also fit a zero-inflated Waring model.

```
. zinbregw docvis edlevel4 age hh, nolog inflate(edlevel4 age hh) eform vuong
Zero-inflated gen neg binomial-W regression       Number of obs   =       3874
Regression link:                                  Nonzero obs     =       2263
Inflation link : logit                            Zero obs        =       1611
                                                  Wald chi2(3)    =      66.10
Log likelihood = -8262.174                        Prob > chi2     =     0.0000
```

| docvis | IRR | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| docvis | | | | | | |
| edlevel4 | .9414482 | .1406355 | -0.40 | 0.686 | .7024933 | 1.261684 |
| age | 1.017108 | .0024842 | 6.95 | 0.000 | 1.012251 | 1.021989 |
| hh | .4841428 | .0964645 | -3.64 | 0.000 | .3276222 | .7154409 |
| _cons | 2.457403 | .3313549 | 6.67 | 0.000 | 1.886691 | 3.200751 |
| inflate | | | | | | |
| edlevel4 | .613575 | .2222675 | 2.76 | 0.006 | .1779387 | 1.049211 |
| age | -.026716 | .0048778 | -5.48 | 0.000 | -.0362763 | -.0171558 |
| hh | -.0137845 | .3544822 | -0.04 | 0.969 | -.7085569 | .6809879 |
| _cons | .1834942 | .245023 | 0.75 | 0.454 | -.2967421 | .6637305 |
| /lnrhom2 | .1842115 | .0856861 | | | .0162699 | .3521532 |
| /lnk | 1.071457 | .2498257 | | | .581808 | 1.561107 |
| rho | 3.20227 | .1030178 | | | 3.016403 | 3.422126 |
| k | 2.919632 | .7293992 | | | 1.789271 | 4.764092 |

```
Vuong test of zinbregw vs. gen neg binomial(W):   z =     0.55   Pr>z = 0.2897
    Bias-corrected (AIC) Vuong test:              z =     0.13   Pr>z = 0.4482
    Bias-corrected (BIC) Vuong test:              z =    -1.20   Pr>z = 0.8845
. estat ic
Akaike´s information criterion and Bayesian information criterion
```

| Model | Obs | ll(null) | ll(model) | df | AIC | BIC |
|---|---|---|---|---|---|---|
| . | 3874 | . | -8262.174 | 10 | 16544.35 | 16606.97 |

```
Note:  N=Obs used in calculating BIC; see [R] BIC note
```

Note that introducing the zero-inflation component into the regression model results in losing significance of the education level in the model of the mean outcomes. However, that variable does play a significant role (along with age) in determining whether a person has zero visits to the doctor.

## 4.2 Generalized negative binomial: Famoye

We can also attempt to understand the relationship of doctor visits and the high education of patients with the additional factors age and income by using another parameterization of negative binomial. This model was discussed in Famoye (1995), but it has had little notice in the literature, which is probably because of the lack of associated software support.

```
. nbregf docvis edlevel4 age hh, nolog eform
Generalized negative binomial-F regression      Number of obs   =       3874
                                                LR chi2(3)      =     166.51
Log likelihood = -8337.884                      Prob > chi2     =     0.0000
```

| docvis | IRR | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| edlevel4 | .7205452 | .0831698 | -2.84 | 0.005 | .5746591 | .9034669 |
| age | 1.025957 | .0024634 | 10.67 | 0.000 | 1.02114 | 1.030796 |
| hh | .4596616 | .0743405 | -4.81 | 0.000 | .3347915 | .6311055 |
| _cons | 2.366462 | .3349416 | 6.09 | 0.000 | 1.793177 | 3.123028 |
| /lnphim1 | -3.252403 | .4280259 | | | -4.091318 | -2.413488 |
| /lntheta | -.6445887 | .0760764 | | | -.7936957 | -.4954816 |
| phi | 1.038681 | .0165565 | | | 1.016717 | 1.089503 |
| theta | .5248784 | .0399309 | | | .4521706 | .6092774 |

```
. estat ic
Akaike´s information criterion and Bayesian information criterion
```

| Model | Obs | ll(null) | ll(model) | df | AIC | BIC |
|---|---|---|---|---|---|---|
| . | 3874 | -8421.139 | -8337.884 | 6 | 16687.77 | 16725.34 |

```
Note:  N=Obs used in calculating BIC; see [R] BIC note
```

Note that the risk ratios are nearly identical to the NB-2 negative binomial model. The AIC and BIC statistics are lower than NB-2, but only by about 12 and 5 points, respectively. Because of the excessive zero counts, we model a zero-inflated model.

```
. zinbregf docvis edlevel4 age hh, nolog inflate(edlevel4 age hh) eform vuong
Zero-inflated gen neg binomial-F regression       Number of obs   =       3874
Regression link:                                  Nonzero obs     =       2263
Inflation link : logit                            Zero obs        =       1611
                                                  LR chi2(3)      =     176.08
Log likelihood = -8292.015                        Prob > chi2     =     0.0000
```

| docvis | IRR | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| docvis | | | | | | |
| edlevel4 | .9125286 | .1191361 | -0.70 | 0.483 | .7065079 | 1.178626 |
| age | 1.017058 | .0024233 | 7.10 | 0.000 | 1.012319 | 1.021818 |
| hh | .4915087 | .0753322 | -4.63 | 0.000 | .3639736 | .6637315 |
| _cons | .0010836 | .2112138 | -0.04 | 0.972 | 1.3e-169 | 8.9e+162 |
| inflate | | | | | | |
| edlevel4 | .7118035 | .2073926 | 3.43 | 0.001 | .3053213 | 1.118286 |
| age | -.0380198 | .0054111 | -7.03 | 0.000 | -.0486254 | -.0274142 |
| hh | .2529651 | .3447803 | 0.73 | 0.463 | -.422792 | .9287221 |
| _cons | .368429 | .2425669 | 1.52 | 0.129 | -.1069933 | .8438514 |
| /lnphim1 | 6.826485 | 195.023 | | | -375.4115 | 389.0645 |
| /lntheta | 7.679818 | 194.9173 | | | -374.3511 | 389.7107 |
| phi | 922.9442 | 179800.3 | | | 1 | 9.3e+168 |
| theta | 2164.225 | 421844.9 | | | 2.6e-163 | 1.8e+169 |

```
Vuong test of zinbregf vs. gen neg binomial(F):   z =     6.23   Pr>z = 0.0000
    Bias-corrected (AIC) Vuong test:              z =     5.68   Pr>z = 0.0000
    Bias-corrected (BIC) Vuong test:              z =     3.99   Pr>z = 0.0000
. estat ic
Akaike´s information criterion and Bayesian information criterion
```

| Model | Obs | ll(null) | ll(model) | df | AIC | BIC |
|---|---|---|---|---|---|---|
| . | 3874 | -8380.053 | -8292.015 | 10 | 16604.03 | 16666.65 |

```
Note:  N=Obs used in calculating BIC; see [R] BIC note
```

The AIC and BIC statistics are substantially lower than the nonzero-inflated parameterization, and they are also lower than the Waring regression model. Here we find that younger patients without a graduate education see physicians more frequently than patients with a graduate education (as we discovered before) and that the important statistics are $\phi$ and $\theta$.

## 4.3 Generalized binomial regression

If the outcomes are bounded counts (for which the bounds are known), then the data can be addressed by grouped binomial models. Rather than introducing a new dataset for these models as we did before, we illustrate how to generate synthetic data.

Herein, we synthesize the generalized binomial outcome along with a zero-inflated version of the generalized binomial outcome. To highlight the options built in to the

commands, we generate data following a complementary log-log link function for the generalized binomial outcome and a log-log link for the zero-inflation component.

```
. set seed  13092
. drop _all
. set obs 1500
obs was 0, now 1500
. // Linear predictors for zero-inflation
. gen z1 = runiform() < 0.5
. gen z2 = runiform() < 0.5
. gen zg = -0.5+0.25*z1+0.25*z2
. // Note that the zero-inflation link function is in terms of Prob(Y=0)
. gen z  = rbinomial(1,1-exp(-exp(-zg)))  // ilink(loglog)
. // Linear predictors for the outcome
. gen x1 = runiform() < 0.5
. gen xb = -2+0.5*x1
. gen n  = floor(10*runiform()) + 1
. // Note that the outcome link function is in terms of Prob(Y=1)
. gen mu  = 1-exp(-exp(xb))                        // link(cloglog)
```

Once we have defined the components of the outcome and the necessary covariates, we generate the outcome. The zero-inflated version of the outcome is the product of the binomial outcome and the zero-inflation (binary) component.

```
. // Program to generate random outcomes "y"
. gen double yu = runiform() // random quantile
. gen y = 0 // initial outcome
. gen double p = 0 // initial cumulative probability
. capture program drop doit
. program define doit
  1.    args sigma
  2.    local flag 1
  3.    local y = 0
  4.    while `flag´ { // increase cumulative probability if y < n
  5.       quietly replace p = p + exp(lngamma(n+`y´*`sigma´+1)-
>             lngamma(n+`y´*`sigma´-`y´+1)-lngamma(`y´+1)+log(n)+`y´*log(mu) +
>             (n+`y´*`sigma´-`y´)*log(1+mu*`sigma´-mu)-log(n+`y´*`sigma´)-
>             (n+`y´*`sigma´)*log(1+mu*`sigma´)) if `y´ < n
  6.       quietly replace y = y+1 if p <= yu // increase y if cumulative
>             probability <= yu
  7.       quietly replace p = 1 if y >= n
  8.       local y = `y´+1
  9.       quietly count if p <= yu // see if finished
 10.       if `r(N)´==0 {
 11.          local flag = 0  // all done
 12.       }
 13.    }
 14. end
. doit 1.25  // sigma=1.25
. // Zero-inflated outcomes "yo"
. gen yo = y*z
```

Having created an outcome with specified associations to our covariates, we can fit a model to see how closely the sample data match the specifications.

```
. // Nonzero-inflated model of nonzero-inflated outcome
. gbin y x1, link(cloglog) n(n) nolog
```

```
Generalized binomial regression                 Number of obs   =       1500
Link           = cloglog                        LR chi2(1)      =      50.73
Dispersion     = generalized binomial           Prob > chi2     =     0.0000
Log likelihood = -1775.7031                     Pseudo R2       =     0.0141
```

| y | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| x1 | .4411576 | .0681407 | 6.47 | 0.000 | .3076043 | .574711 |
| _cons | -2.000648 | .0503157 | -39.76 | 0.000 | -2.099264 | -1.902031 |
| /lnsigma | .2661259 | .1168846 | | | .0370362 | .4952155 |
| sigma | 1.304899 | .1525227 | | | 1.037731 | 1.640852 |

```
Likelihood-ratio test of sigma=0:  chibar2(01) =  152.55 Prob>=chibar2 = 0.000
```

Before fitting the zero-inflated model for the zero-inflated outcome, we first illustrate how well a zero-inflated model might fit the nonzero-inflated outcome. In this case, we should expect the binomial regression components to estimate the means well, and we should expect the covariate of the zero-inflation component to be nonsignificant.

```
. // Zero-inflated model of nonzero-inflated outcome
. zigbin y  x1, inflate(z1 z2) n(n) link(cloglog) ilink(loglog) vuong nolog
```

```
Zero-inflated generalized binomial regression   Number of obs   =       1500
Regression link: cloglog                        Nonzero obs     =        751
Inflation link : loglog                         Zero obs        =        749
                                                LR chi2(1)      =      42.67
Log likelihood =   -1772.5                      Prob > chi2     =     0.0000
```

| y | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| y | | | | | | |
| x1 | .447438 | .0681432 | 6.57 | 0.000 | .3138797 | .5809963 |
| _cons | -1.958826 | .0540354 | -36.25 | 0.000 | -2.064733 | -1.852918 |
| inflate | | | | | | |
| z1 | .4499741 | .4248806 | 1.06 | 0.290 | -.3827765 | 1.282725 |
| z2 | 2.068714 | 60.05847 | 0.03 | 0.973 | -115.6437 | 119.7812 |
| _cons | -3.264426 | 60.05983 | -0.05 | 0.957 | -120.9795 | 114.4507 |
| /lnsigma | .1366821 | .1379003 | | | -.1335977 | .4069618 |
| sigma | 1.146464 | .1580977 | | | .874942 | 1.502247 |

```
Vuong test of zigbin vs. gen binomial:            z =     1.25   Pr>z = 0.1048
    Bias-corrected (AIC) Vuong test:              z =     0.08   Pr>z = 0.4683
    Bias-corrected (BIC) Vuong test:              z =    -3.04   Pr>z = 0.9988
```

Note that the Vuong statistic was nonsignificant in this example. Though it fails to provide compelling evidence for one model over the other, we would prefer the nonzero-

inflated model because of the lack of significant covariates in the inflation. When we fit a zero-inflated model for the outcome that was specifically generated to include zero inflation, we see a much better fit.

```
. // Zero-inflated model of zero-inflated outcome
. zigbin yo x1, inflate(z1 z2) n(n) link(cloglog) ilink(loglog) vuong nolog
Zero-inflated generalized binomial regression     Number of obs   =       1500
Regression link: cloglog                          Nonzero obs     =        541
Inflation link : loglog                           Zero obs        =        959
                                                  LR chi2(1)      =      28.34
Log likelihood = -1518.557                        Prob > chi2     =     0.0000
```

| yo | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| yo | | | | | | |
| x1 | .4628085 | .086265 | 5.36 | 0.000 | .2937322 | .6318848 |
| _cons | -1.969505 | .0873894 | -22.54 | 0.000 | -2.140785 | -1.798225 |
| inflate | | | | | | |
| z1 | .2292778 | .1270487 | 1.80 | 0.071 | -.019733 | .4782886 |
| z2 | .3955768 | .1296781 | 3.05 | 0.002 | .1414125 | .6497411 |
| _cons | -.4796692 | .1724896 | -2.78 | 0.005 | -.8177426 | -.1415958 |
| /lnsigma | .0882868 | .2415756 | | | -.3851926 | .5617661 |
| sigma | 1.092301 | .2638733 | | | .6803196 | 1.753767 |

```
Vuong test of zigbin vs. gen binomial:            z =     3.11   Pr>z = 0.0009
    Bias-corrected (AIC) Vuong test:              z =     2.59   Pr>z = 0.0048
    Bias-corrected (BIC) Vuong test:              z =     1.20   Pr>z = 0.1159
```

Here the Vuong test indicates a clear preference for the zero-inflation model, and we note that the estimated coefficients are close to the values we specified in synthesizing these data.

# 5 Discussion and conclusions

In this article, we introduced programs for modeling count data. These count data can be overdispersed (variance is greater than the mean), underdispersed (variance is smaller than the mean), or undispersed (variance equals the mean). We then illustrated the use of the new commands `nbregf`, `zinbregf`, `nbregw`, and `zinbregw` using real-world German health data from 1984. We synthesized data and used it to demonstrate the `gbin` and `zigbin` models. This article is fairly technical, and some readers may desire more background on count-data models such as the Poisson, generalized Poisson, and negative binomial models. For those readers, we recommend Hardin and Hilbe (2012), Cameron and Trivedi (2013), Winkelmann (2008), and Tang, He, and Tu (2012).

# 6 References


Cameron, A. C., and P. K. Trivedi. 2013. *Regression Analysis of Count Data*. 2nd ed. Cambridge: Cambridge University Press.

Consul, P. C., and H. C. Gupta. 1980. The generalized negative binomial distribution and its characterization by zero regression. *SIAM Journal on Applied Mathematics* 39: 231–237.

Dean, C. B. 1992. Testing for overdispersion in Poisson and binomial regression models. *Journal of the American Statistical Association* 87: 451–457.

Desmarais, B. A., and J. J. Harden. 2013. Testing for zero inflation in count models: Bias correction for the Vuong test. *Stata Journal* 13: 810–835.

Famoye, F. 1995. Generalized binomial regression model. *Biometrical Journal* 37: 581–594.

Hardin, J. W., and J. M. Hilbe. 2012. *Generalized Linear Models and Extensions*. 3rd ed. College Station, TX: Stata Press.

Harris, T., Z. Yang, and J. W. Hardin. 2012. Modeling underdispersed count data with generalized Poisson regression. *Stata Journal* 12: 736–747.

Hilbe, J. M. 2011. *Negative Binomial Regression*. 2nd ed. Cambridge: Cambridge University Press.

———. 2014. *Modeling Count Data*. Cambridge: Cambridge University Press.

Irwin, J. O. 1968. The generalized Waring distribution applied to accident theory. *Journal of the Royal Statistical Society* Series A 131: 205–225.

Jain, G. C., and P. C. Consul. 1971. A generalized negative binomial distribution. *SIAM Journal on Applied Mathematics* 21: 501–513.

Rodríguez-Avi, J., A. Conde-Sánchez, A. J. Sáez-Castillo, M. J. Olmo-Jiménez, and A. M. Martínez-Rodríguez. 2009. A generalized Waring regression model for count data. *Computational Statistics and Data Analysis* 53: 3717–3725.

Tang, W., H. He, and X. M. Tu. 2012. *Applied Categorical and Count Data Analysis*. Boca Raton, FL: Chapman & Hall/CRC.

Vuong, Q. H. 1989. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* 57: 307–333.

Wang, X.-F., Z. Jiang, J. J. Daly, and G. H. Yue. 2012. A generalized regression model for region of interest analysis of fMRI data. *Neuroimage* 59: 502–510.

Winkelmann, R. 2008. *Econometric Analysis of Count Data*. 5th ed. Berlin: Springer.

Yang, Z., J. W. Hardin, C. L. Addy, and Q. H. Vuong. 2007. Testing approaches for overdispersion in Poisson regression versus the generalized Poisson model. *Biometrical Journal* 49: 565–584.

### About the authors

Tammy Harris is a senior research associate in the Institute for Families in Society at the University of South Carolina, Columbia, SC. She graduated from the Department of Epidemiology and Biostatistics at the University of South Carolina with a PhD in August 2013.

Joseph M. Hilbe is an emeritus professor (University of Hawaii), an adjunct professor of statistics at Arizona State University, Tempe, AZ, and a Solar System Ambassador at Jet Propulsion Laboratory, Pasadena, CA.

James W. Hardin is an associate professor in the Department of Epidemiology and Biostatistics and an affiliated faculty in the Institute for Families in Society at the University of South Carolina, Columbia, SC.