AgEcon SEARCH
RESEARCH IN AGRICULTURAL & APPLIED ECONOMICS

The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search
http://ageconsearch.umn.edu
aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

The Stata Journal

Editors

H. Joseph Newton
Department of Statistics
Texas A&M University
College Station, Texas
editors@stata-journal.com

Nicholas J. Cox
Department of Geography
Durham University
Durham, UK
editors@stata-journal.com

Associate Editors

Christopher F. Baum, Boston College
Nathaniel Beck, New York University
Rino Bellocco, Karolinska Institutet, Sweden, and University of Milano-Bicocca, Italy
Maarten L. Buis, WZB, Germany
A. Colin Cameron, University of California–Davis
Mario A. Cleves, University of Arkansas for Medical Sciences
William D. Dupont, Vanderbilt University
Philip Ender, University of California–Los Angeles
David Epstein, Columbia University
Allan Gregory, Queen's University
James Hardin, University of South Carolina
Ben Jann, University of Bern, Switzerland
Stephen Jenkins, London School of Economics and Political Science
Ulrich Kohler, University of Potsdam, Germany
Frauke Kreuter, Univ. of Maryland–College Park
Peter A. Lachenbruch, Oregon State University
Jens Lauritsen, Odense University Hospital
Stanley Lemeshow, Ohio State University
J. Scott Long, Indiana University
Roger Newson, Imperial College, London
Austin Nichols, Urban Institute, Washington DC
Marcello Pagano, Harvard School of Public Health
Sophia Rabe-Hesketh, Univ. of California–Berkeley
J. Patrick Royston, MRC Clinical Trials Unit, London
Philip Ryan, University of Adelaide
Mark E. Schaffer, Heriot-Watt Univ., Edinburgh
Jeroen Weesie, Utrecht University
Ian White, MRC Biostatistics Unit, Cambridge
Nicholas J. G. Winter, University of Virginia
Jeffrey Wooldridge, Michigan State University

Stata Press Editorial Manager

Lisa Gilmore

Stata Press Copy Editors

David Culwell, Deirdre Skaggs, and Shelbi Seiner

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

Subscriptions are available from StataCorp, 4905 Lakeway Drive, College Station, Texas 77845, telephone 979-696-4600 or 800-STATA-PC, fax 979-696-4601, or online at

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

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.

Copyright © 2014 by StataCorp LP

Copyright Statement: The *Stata Journal* and the contents of the supporting files (programs, datasets, and help files) are copyright © by StataCorp LP. The contents of the supporting files (programs, datasets, and help files) may be copied or reproduced by any means whatsoever, in whole or in part, as long as any copy or reproduction includes attribution to both (1) the author and (2) the *Stata Journal*.

The articles appearing in the *Stata Journal* may be copied or reproduced as printed copies, in whole or in part, as long as any copy or reproduction includes attribution to both (1) the author and (2) the *Stata Journal*.

Written permission must be obtained from StataCorp if you wish to make electronic copies of the insertions. This precludes placing electronic copies of the *Stata Journal*, in whole or in part, on publicly accessible websites, fileservers, or other locations where the copy may be accessed by anyone other than the subscriber.

Users of any of the software, ideas, data, or other materials published in the *Stata Journal* or the supporting files understand that such use is made without warranty of any kind, by either the *Stata Journal*, the author, or StataCorp. In particular, there is no warranty of fitness of purpose or merchantability, nor for special, incidental, or consequential damages such as loss of profits. The purpose of the *Stata Journal* is to promote free communication among Stata users.

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.

The Stata Journal (2014)
14, Number 2, pp. 280–291

Regression models for count data based on the negative binomial(p) distribution

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

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

Abstract. We present new Stata commands for estimating several regression models suitable for analyzing overdispersed count outcomes. The `nbregp` command nests the `dispersion(constant)` and `dispersion(mean)` versions of Stata's `nbreg` command in a model for negative binomial(p) regression. The `zignbreg` command extends Stata's `gnbreg` command for zero inflation, and the `zinbregp` command fits a negative binomial(p) regression model with zero inflation. The new commands for zero-inflated models allow specification of links within the `glm` command's collection for the Bernoulli model of zero inflation. These commands will optionally calculate a Vuong test, which compares the zero-inflated model with the nonzero-inflated model.

Keywords: st0336, nbregp, zignbreg, zinbregp, Vuong test, zero inflation

1 Introduction

Regression modeling of count outcomes is supported in several Stata commands. Missing from the official collection of commands is support for a regression model based on a generalization of the negative binomial (NB) distribution discussed in Greene (2008). A simple version of this model (without support for `predict` or zero inflation) was illustrated in Hardin and Hilbe (2012). This illustration used a simple `lf` style program callable from Stata's `ml` command. However, the command lacked the full support enjoyed by Stata's other built-in commands.

We present Stata estimation commands to evaluate negative binomial(p) (NB-P) regression, zero-inflated generalized NB regression, and zero-inflated NB-P regression. This article is organized as follows: in section 2, we review the regression models; in section 3, we present Stata syntax for the new commands; and in section 4, we present examples.

© 2014 StataCorp LP

2 Two extensions of NB regression

The NB probability mass function is given by

$$f(y;\alpha,\delta) = \frac{\Gamma(y+1/\alpha)}{\Gamma(1/\alpha)\Gamma(y+1)}\left(\frac{1}{1+\delta\alpha}\right)^{1/\alpha}\left(1-\frac{1}{1+\delta\alpha}\right)^{y}$$

with mean $E(y) = \delta$, and variance $V(y) = \delta(1+\delta\alpha)$. Stata users have access to two parameterizations of the NB distribution. The two results of the parameterizations are called the NB-1 (constant dispersion) and the NB-2 (mean dispersion) models. The numerals used in naming these two models correspond to the nature of the variance (as a function of the power of the mean). The NB-1 model results from introducing coefficients via $\alpha = \theta\exp(X\beta) = \theta\mu$ and $\delta = \exp(X\beta) = \mu$ so that the mean is μ, the variance is $\mu(1+\theta)$, and the dispersion is $(1+\theta)$. The NB-2 model results from introducing regressors X via $\alpha = \theta$ and $\delta = \exp(X\beta) = \mu$ so that the mean is μ, the variance is $\mu(1+\mu\theta)$, and the dispersion is $1+\mu\theta$.

Stata software has included the `gnbreg` command since at least the release of version 4.0. The `gnbreg` command includes an observation-specific dispersion parameter via a linear combination of predictors (separate from the linear combination of predictors for the mean). Instead of being a scalar value constant over all observations, as assumed in the `nbreg` command, this generalization allows the dispersion to change even within a specific covariate pattern for the mean. Thus the `gnbreg` command generalizes the treatment of the dispersion parameter in the regression model. Specifically, regressors X are introduced via $\alpha = \theta$ and $\delta = \exp(X\beta)$ as in the NB-2 specification, and a second set of regressors Z is used to replace the dispersion parameter $\theta = \exp(Z\gamma)$. This is not the only generalization of the NB regression model.

Greene (2008) discusses the implementation of a second generalization to the underlying NB probability distribution for which the variance is a function of a parameter power of the mean; also see Cameron and Trivedi (2013). In this NB-P model, regressors X are introduced via $\alpha = \theta\exp(X\beta)^{P-2} = \theta\mu^{P-2}$ and $\delta = \exp(X\beta) = \mu$ so that the mean is μ, the variance is $\mu(1+\mu^{P-1}\theta)$, and the dispersion is $(1+\mu^{P-1}\theta)$. In this presentation, we see that the distribution is equal to NB-1 when $P = 1$ and to NB-2 when $P = 2$.

2.1 Zero-inflated count models

Similar to the manner in which the zero-inflated Poisson and the zero-inflated NB models are derived, we can imagine two separate processes generating outcomes such that the outcome of the two processes are partially visible.

In the generalized NB regression model, each observation in the dataset contains information on the number of outcomes (successes); this count can also be thought of as a rate if we consider the amount of time for which each observation was exposed. When we consider zero inflation for binomial or count outcomes, we introduce a Bernoulli process that models the probability of zero successes; this probability of failure is parameterized via a user-specified link function of a linear predictor, $z\gamma$:

$$\begin{aligned} P(Y=0) &= P_{\text{Bernoulli}}(Y=0|z\gamma) \\ &+ \{1-P_{\text{Bernoulli}}(Y=0|z\gamma)\}\,P_{\text{count}}(Y=0|x\beta,n) \\ P(Y=y>0) &= \{1-P_{\text{Bernoulli}}(Y=0|z\gamma)\}\,P_{\text{count}}(Y=y|x\beta,n) \end{aligned}$$

An extension of the likelihood-ratio test called the Vuong test (Vuong 1989) evaluates whether the count model with zero inflation or the count model without zero inflation is closer to the true model.

A random variable ω is defined as the vector $\log L_{\text{Z}} - \log L_{\text{S}}$, where L_{Z} is a vector of the observation-level contributions to the likelihood of the zero-inflated model evaluated at its maximum likelihood estimate, and L_{S} is a vector of the observation-level contributions to the likelihood of the standard (nonzero-inflated) model evaluated at its maximum likelihood estimate. The vector of differences over the N observations is then used to define the statistic

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

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.

Thus zero-inflated versions of the NB-P and the generalized NB model can be developed. Each zero-inflated model can be compared with the associated model without zero inflation via the Vuong test.

Greene (2008) points out that a Vuong statistic could be developed to compare the NB-1 and NB-2 models. When the count model includes only a constant, this statistic is zero, and Greene (2008) reports rarely encountering a significant result for this comparison in practice. Obviously, if one were to generate synthetic data according to one or the other distribution, the statistic would achieve the nominal size for large enough samples. However, Greene (2008) also points out that under the NB-P model, one can perform likelihood-ratio tests against either (or both) of the NB-1 and NB-2 models. These likelihood-ratio tests are included by default in the accompanying software.

3 Stata syntax

The software accompanying this article includes the command files and supporting files for prediction and help. In all the following syntax diagrams, unspecified *options* include

the usual collection of maximization and display options available to all estimation commands. In addition, the zero-inflated commands `zignbreg` and `zinbregp` include the option `ilink(`*linkname*`)` to specify the link function for the inflation model. Supported *linkname*s include `logit`, `probit`, `loglog`, and `cloglog`.

Equivalent in syntax to the `zip` command, the basic syntax for the zero-inflated generalized NB model is

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

Equivalent in syntax to the `nbreg` command, the basic syntax for the NB-P regression command is

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

Equivalent in syntax to the `zip` command, the basic syntax for the zero-inflated NB-P regression command is

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

Help files are included for the estimation and postestimation specifications of these models. The help files include example specifications.

4 Example

We use the included dataset on German health reform to build models similar to those used in the discussion of Riphahn, Wambach, and Million (2003). These data include several variables: the number of days each year the patient visits a physician, `docvis`; the age in years of the patient, `age`; the monthly income in German marks per 1,000, `hhninc`; and the number of years (including partial years) of education, `educ`.

We illustrate a Poisson model of `docvis` on `age`, `hhninc`, and `educ`. Using Stata's `glm` command, we see evidence of overdispersion in the Pearson statistic. Recall that the Poisson distribution assumes that the mean and variance of the response variable are equal for a given set of covariates. When the mean and variance are equal, the data are said to be equidispersed. When the variance is greater than the mean, the data are said to be overdispersed. Evaluating whether there is overdispersion in data is indicated when the `(1/df) Pearson` statistic is greater than one.

```
. use rwm
. keep if age != . & hhninc != . & educ != . & docvis != .
(0 observations deleted)
. glm docvis age hhninc edu, nolog family(poisson)
Generalized linear models                         No. of obs      =     27326
Optimization     : ML                             Residual df     =     27322
                                                  Scale parameter =         1
Deviance         =  156589.5963                   (1/df) Deviance =  5.731264
Pearson          =   256396.682                   (1/df) Pearson  =  9.384257
Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = ln(u)                   [Log]
                                                  AIC             =  7.671674
Log likelihood   = -104814.0886                   BIC             = -122520.9
------------------------------------------------------------------------------
             |                 OIM
      docvis |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0212508   .0003047    69.75   0.000     .0206536     .021848
      hhninc |  -.0532375   .0022036   -24.16   0.000    -.0575564    -.0489186
        educ |  -.0420873   .0017279   -24.36   0.000     -.045474    -.0387006
       _cons |   .8523131   .0254907    33.44   0.000     .8023521     .902274
------------------------------------------------------------------------------
. estat ic
-----------------------------------------------------------------------------
       Model |   Obs    ll(null)   ll(model)     df          AIC         BIC
-------------+---------------------------------------------------------------
           . | 27326           .  -104814.1      4     209636.2      209669
-----------------------------------------------------------------------------
                Note:  N=Obs used in calculating BIC; see [R] BIC note
```

The dispersion statistic is 9.38, which is far greater than would be estimated if the data were equidispersed. Likely reasons for overdispersion in these data are that the data are longitudinal and that there are an excess of zero outcomes. For illustration, we are ignoring these important facts in some of these analyses. A first step in addressing overdispersion is to consider fitting an NB regression model. This model allows overdispersion such that the conditional variance of the outcome is assumed to be a quadratic function of the conditional mean.

```
. glm docvis age hhninc edu, nolog family(nbinomial ml)
Generalized linear models                         No. of obs      =     27326
Optimization     : ML                             Residual df     =     27322
                                                  Scale parameter =         1
Deviance         =  28510.91449                   (1/df) Deviance =  1.043515
Pearson          =  36242.35265                   (1/df) Pearson  =   1.32649
Variance function: V(u) = u+(1.9363)u^2           [Neg. Binomial]
Link function    : g(u) = ln(u)                   [Log]
                                                  AIC             =  4.415284
Log likelihood   = -60322.02105                   BIC             = -250599.5
```

docvis	Coef.	OIM Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.0204292	.0008006	25.52	0.000	.0188601	.0219984
hhninc	-.0476814	.0052278	-9.12	0.000	-.0579278	-.0374351
educ	-.0459575	.0042257	-10.88	0.000	-.0542398	-.0376752
_cons	.9132608	.0633757	14.41	0.000	.7890467	1.037475

```
Note: Negative binomial parameter estimated via ML and treated as fixed once
      estimated.
. estat ic
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	27326	.	-60322.02	4	120652	120684.9

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

The dispersion statistic for the NB regression model is 1.33. This is a substantial improvement but still indicates unaccounted overdispersion. In fact, a substantial number of zero outcomes in the data may reflect a completely separate data-generating mechanism. We will explore that idea with zero-inflated models. Before investigating zero-inflated models, we first investigate two alternatives.

Negative binomial regression is a common first strategy for addressing overdispersed data. The scalar heterogeneity parameter in the NB model can often appropriately adjust for the extra correlation in the data. Thus it is necessary to assess the heterogeneity parameter to determine whether it is different from zero. If not, then the NB model is no different from the Poisson model. We can assess the parameter as part of the standard output of the `nbreg` command.

```
. nbreg docvis age hhninc edu, nolog
Negative binomial regression                    Number of obs   =      27326
                                                LR chi2(3)      =    1027.40
Dispersion     = mean                           Prob > chi2     =     0.0000
Log likelihood = -60322.021                     Pseudo R2       =     0.0084

------------------------------------------------------------------------------
      docvis |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0204292   .0008006    25.52   0.000     .0188601    .0219984
      hhninc |  -.0476814   .0052279    -9.12   0.000    -.0579279    -.037435
        educ |  -.0459575   .0042257   -10.88   0.000    -.0542398   -.0376752
       _cons |   .9132608   .0633758    14.41   0.000     .7890465    1.037475
-------------+----------------------------------------------------------------
    /lnalpha |   .6608039   .0115374                      .638191    .6834168
-------------+----------------------------------------------------------------
       alpha |   1.936348   .0223404                      1.893053    1.980634
------------------------------------------------------------------------------
Likelihood-ratio test of alpha=0:  chibar2(01) = 8.9e+04 Prob>=chibar2 = 0.000
. estat ic

-----------------------------------------------------------------------------
       Model |   Obs    ll(null)   ll(model)     df          AIC         BIC
-------------+---------------------------------------------------------------
           . | 27326   -60835.72  -60322.02      5       120654    120695.1
-----------------------------------------------------------------------------
                Note:  N=Obs used in calculating BIC; see [R] BIC note
```

Clearly, the dispersion parameter (labeled `alpha`) is significantly different from zero. Thus the NB model seems to fit the data better than the Poisson model. When one uses the default parameterization of the NB regression model, `dispersion(mean)`, the conditional variance of the outcome is a quadratic function of the conditional mean $\mu(1+\theta\mu)$—this is the so-called NB-2 model. One could specify `dispersion(constant)`, in which case the parameterization of the NB model would specify a conditional variance of the outcome that was a linear function of the conditional mean $\mu(1+\theta)$—this is the so-called NB-1 model.

A generalized (three-parameter) NB distribution yielding an alternative regression model is known as the NB-P model. In this generalization, the relationship of the conditional variance in terms of the conditional mean is a parameter.

The generalized NB-P is often used to determine between choosing the NB-1 or the NB-2 model for a given set of count data. The NB-P model incorporates an additional scalar parameter to the standard NB symbolizing the term "power", where the conditional variance of the conditional mean is given by $\mu(1+\theta\mu^{P-1})$. Clearly, these NB models allow for overdispersion but not underdispersion. Also note that when data are adequately modeled as Poisson, the NB-P model can encounter numeric difficulties in estimation because θ and P are near zero.

Parameterized in this fashion, the θ parameter in the model is such that higher values reflect greater correlation in the data. Therefore, the NB-2 model can adjust for a greater degree of correlation than the NB-1 model. See Cameron and Trivedi (2013), Greene (2012), Hilbe and Greene (2008), and Hilbe (2011). Note that Hardin and Hilbe (2012) also treat this model but parameterize θ in the reciprocal.

While some statisticians prefer to model overdispersed data using only powers associated with the well-known NB-1 or NB-2, allowing the NB-P model to be used as a means to select between the two forms of the NB model, one can use the NB-P model in its own right.

```
. nbregp docvis age hhninc edu, nolog
Negative binomial-P regression                  Number of obs   =      27326
                                                Wald chi2(3)    =    1059.09
Log likelihood = -60258.97                      Prob > chi2     =     0.0000
```

docvis	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.0217776	.000775	28.10	0.000	.0202587	.0232965
hhninc	-.0387497	.0053987	-7.18	0.000	-.049331	-.0281684
educ	-.0412764	.0042127	-9.80	0.000	-.0495332	-.0330196
_cons	.7702929	.0622765	12.37	0.000	.6482332	.8923525
/P	1.544368	.0425538	36.29	0.000	1.460964	1.627772
/lntheta	1.187102	.0507474			1.087639	1.286565
theta	3.27757	.166328			2.967261	3.620331

```
Likelihood-ratio test of P=1:     chi2 =   115.72 Prob > chi2     =     0.0000
Likelihood-ratio test of P=2:     chi2 =   126.10 Prob > chi2     =     0.0000
. estat ic
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	27326	.	-60258.97	6	120529.9	120579.2

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

The Akaike information criterion (AIC) and Bayesian information criterion (BIC) statistics are some 130 lower than standard NB-2 regression but about 50 higher than the heterogeneous NB. If we were foremost interested in using NB-P to select between NB-1 or NB-2 for modeling the data, the model does not help much. With an estimated value of the power parameter at 1.54, neither NB-1 nor NB-2 is clearly preferred. Adjusting for zero counts and using a zero-inflated NB-P may well resolve the issue.

Another generalization of the (two-parameter) NB distribution is to allow the overdispersion parameter to vary across observations instead of assuming that it is a fixed-scalar quantity. Stata refers to this model as a generalized NB regression model, though others call it a heterogeneous NB regression model.

The heterogeneous NB model allows us to determine which predictors most influence the value of the dispersion parameter. For these data, the heterogeneous model may tell us which predictors influence the generation of zero counts. Significant coefficients of the scale parameter are those likely influencing zero values.

```
. gnbreg docvis age hhninc edu, nolog lnalpha(age hhninc edu)
Generalized negative binomial regression          Number of obs   =      27326
                                                  LR chi2(3)      =    1039.39
                                                  Prob > chi2     =     0.0000
Log likelihood = -60230.363                       Pseudo R2       =     0.0086

------------------------------------------------------------------------------
      docvis |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
docvis       |
         age |   .0208972   .0008043    25.98   0.000     .0193209    .0224736
      hhninc |  -.0467431   .0050995    -9.17   0.000    -.0567379   -.0367483
        educ |  -.0454817    .004282   -10.62   0.000    -.0538742   -.0370891
       _cons |   .8837919   .0646169    13.68   0.000      .757145    1.010439
-------------+----------------------------------------------------------------
lnalpha      |
         age |  -.0131726   .0010249   -12.85   0.000    -.0151814   -.0111639
      hhninc |  -.0208019   .0070329    -2.96   0.003    -.0345861   -.0070177
        educ |   .0073402   .0056123     1.31   0.191    -.0036597    .0183401
       _cons |   1.239363   .0828105    14.97   0.000     1.077057    1.401668
------------------------------------------------------------------------------

. estat ic

-----------------------------------------------------------------------------
       Model |   Obs    ll(null)   ll(model)     df          AIC         BIC
-------------+---------------------------------------------------------------
           . | 27326   -60750.06  -60230.36      8     120476.7    120542.5
-----------------------------------------------------------------------------
                Note:  N=Obs used in calculating BIC; see [R] BIC note
```

There is evidence that the dispersion varies across age and income categories. However, we may not have suitably addressed a difference in the underlying mechanisms producing zero and count outcomes. We have extended Stata's `gnbreg` command to allow for zero inflation. As with other zero-inflated commands, we have also included the Vuong test to compare the zero-inflated model with the nonzero-inflated model. Also we have extended the `nbregp` command to allow for zero inflation.

```
. zinbregp docvis age hhninc edu, inflate(age hhninc) vuong nolr nolog
Zero-inflated negative binomial-p regression      Number of obs   =      27326
Regression link:                                  Nonzero obs     =      17191
Inflation link : logit                            Zero obs        =      10135
                                                  Wald chi2(3)    =    1002.52
Log likelihood = -60257.39                        Prob > chi2     =     0.0000
```

docvis	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
docvis						
age	.0211927	.0008709	24.33	0.000	.0194857	.0228996
hhninc	-.0450366	.0063834	-7.06	0.000	-.0575478	-.0325253
educ	-.0417591	.0042365	-9.86	0.000	-.0500625	-.0334556
_cons	.8357666	.0741013	11.28	0.000	.6905306	.9810025
inflate						
age	-.0368993	.0147789	-2.50	0.013	-.0658654	-.0079332
hhninc	-.5452372	.2344786	-2.33	0.020	-1.004807	-.0856675
_cons	-1.231483	.8693829	-1.42	0.157	-2.935442	.4724763
/P	1.563359	.0451585	34.62	0.000	1.47485	1.651869
/lntheta	1.145579	.0582813			1.031349	1.259808
theta	3.14426	.1832516			2.804848	3.524744

```
Vuong test of zinbregp vs. negative binomial(p):  z =   105.70   Pr>z = 0.0000
. estat ic
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	27326	.	-60257.39	9	120532.8	120606.7

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

The value of alpha, as understood for standard NB-2 regression, is the same as theta reported above. It is rather high, but it is lower than that of the nonzero-inflated model, which indicates that it needed to adjust for the excess zero-response values. The Vuong statistic also informs us that the zero-inflated model is favored over the model that does not adjust for excess zero counts. Predictors influencing the generation of zeros are age and income. Educational level does not appear to contribute to zero counts. The fact that the data are clustered by year undermines the usefulness of both the AIC and the BIC statistics, which assume the independence of observations. Standard errors differ little from the model-based standard errors. In this case, the AIC and BIC statistics are slightly higher than the model that does not adjust for zero counts, so we cannot rely on either of them to tell us much about the models in question.

Note that both the `inflate()` and `lnalpha()` options are required: `inflate()` specifies the predictors we believe may influence the generation of zero counts, and `lnalpha()` defines the predictors we believe bear on the dispersion statistic.

```
. zignbreg docvis age hhninc edu, nolog lnalpha(age hhninc edu)
> inflate(age hhninc) vuong
Zero-inflated generalized binomial regression     Number of obs   =      27326
Regression link:                                  Nonzero obs     =      17191
Inflation link : logit                            Zero obs        =      10135
                                                  LR chi2(3)      =     914.30
Log likelihood = -60230.36                        Prob > chi2     =     0.0000
```

docvis	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
docvis						
age	.0208973	.0008043	25.98	0.000	.019321	.0224737
hhninc	-.0467432	.0050995	-9.17	0.000	-.0567381	-.0367484
educ	-.0454815	.004282	-10.62	0.000	-.0538741	-.037089
_cons	.8837831	.0646171	13.68	0.000	.7571359	1.01043
inflate						
age	-.0036265	17.52539	-0.00	1.000	-34.35276	34.34551
hhninc	.0564263	89.04143	0.00	0.999	-174.4616	174.5744
_cons	-18.27565	1189.44	-0.02	0.988	-2349.536	2312.984
lntheta						
age	-.0131727	.0010249	-12.85	0.000	-.0151815	-.011164
hhninc	-.0208018	.0070329	-2.96	0.003	-.034586	-.0070176
educ	.0073404	.0056123	1.31	0.191	-.0036595	.0183403
_cons	1.239372	.0828105	14.97	0.000	1.077066	1.401677

```
Vuong test of zignbreg vs. gen negative binomial: z =     0.24   Pr>z = 0.4039
. estat ic
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	27326	-60687.51	-60230.36	11	120482.7	120573.1

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

The Vuong test shows no preference for the model with or without zero inflation, so we would prefer the more parsimonious specification. In this example, the generalized (heterogeneous) NB model seems to fit the data best.

5 References

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

Greene, W. 2008. Functional forms for the negative binomial model for count data. *Economics Letters* 99: 585–590.

Greene, W. H. 2012. *Econometric Analysis*. 7th ed. Upper Saddle River, NJ: Prentice Hall.

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

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

Hilbe, J. M., and W. H. Greene. 2008. Count response regression models. In *Handbook of Statistics 27: Epidemiology and Medical Statistics*, ed. C. R. Rao, J. P. Miller, and D. C. Rao, 210–252. Amsterdam: Elsevier.

Riphahn, R. T., A. Wambach, and A. Million. 2003. Incentive effects in the demand for health care: A bivariate panel count data estimation. *Journal of Applied Econometrics* 18: 387–405.

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

About the authors

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

Joseph M. Hilbe is an emeritus professor at the University of Hawaii, an adjunct professor of statistics at Arizona State University in Tempe, AZ, and a Solar System Ambassador with NASA's Jet Propulsion Laboratory in Pasadena, CA.