The Stata Journal

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	$115	1-year subscription	$145
2-year subscription	$210	2-year subscription	$270
3-year subscription	$285	3-year subscription	$375
1-year student subscription	$ 85	1-year student subscription	$115
1-year institutional subscription	$345	1-year institutional subscription	$375
2-year institutional subscription	$625	2-year institutional subscription	$685
3-year institutional subscription	$875	3-year institutional subscription	$965
Electronic only		**Electronic only**	
1-year subscription	$ 85	1-year subscription	$ 85
2-year subscription	$155	2-year subscription	$155
3-year subscription	$215	3-year subscription	$215
1-year student subscription	$ 55	1-year student subscription	$ 55

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 (2014)
14, Number 4, pp. 798–816

The chi-squared goodness-of-fit test for count-data models

Miguel Manjón
QURE-CREIP Department of Economics
Rovira i Virgili University
Reus, Spain
miguel.manjon@urv.cat

Oscar Martínez
QURE-CREIP Department of Economics
Rovira i Virgili University
Reus, Spain
oscar.martinez@urv.cat

Abstract. In this article, we discuss the implementation of Andrews's (1988a, *Journal of Econometrics* 37: 135–156; 1988b, *Econometrica* 56: 1419–1453) chi-squared goodness-of-fit test as a postestimation command. The new command `chi2gof` reports the test statistic, its degrees of freedom, and its p-value. `chi2gof` can be used after the `poisson`, `nbreg`, `zip`, and `zinb` commands.

Keywords: st0360, chi2gof, Andrews's chi-squared goodness-of-fit test, m-tests, count-data models

1 Introduction

In empirical work, one often fits a model using alternative specifications and then concentrates on the coefficient estimates supported by a goodness-of-fit test (thus ignoring estimates not supported by the test). This practice reflects the use of the goodness-of-fit test to detect specification errors in a model. Goodness-of-fit tests can also be used in model comparison and selection. As Cameron and Trivedi (2013, 225) explain, "competing models [...] are compared and evaluated using model diagnostics and goodness-of-fit measures".

One of these goodness-of-fit tests is the Pearson chi-squared test (see, for example, Cameron and Trivedi [2005, 266] for details). This is implemented in Stata as the postestimation command `estat gof` following use of the `logit`, `logistic`, `probit`, and `poisson` commands.[1] By typing `estat gof` after `logit`, `logistic`, `probit`, or `poisson`, one obtains the χ^2-statistic of the test and its p-value (as well as the number of observations and the number of covariate patterns). Alternatively, the `group(#)` option results in analogous output for the related Hosmer–Lemeshow test (see Hosmer and Lemeshow [1980]; and Hosmer, Lemeshow, and Sturdivant [2013]).

However, the Pearson and Hosmer–Lemeshow tests assume that the estimated coefficients are known. To control for the potential estimation error, Cameron and Trivedi (2010) suggest using the chi-squared diagnostic test developed by Andrews (1988a,b). This chi-squared goodness-of-fit test generalizes Pearson's chi-squared test by comparing the sample relative frequencies of the dependent variable with the predicted fre-

1. Structural modeling (`sem`) and survey data (`svy:`) are other areas of application of this test that are supported by Stata.

 st0360

quencies from the model using a quadratic form and an estimate of the asymptotic variance of the corresponding population moment condition. Unlike Pearson's test (or the Hosmer–Lemeshow test), the chi-squared goodness-of-fit test can be constructed from any regular asymptotically normal estimator of the conditional expectation of the dependent variable. However, this m-test is not yet available in Stata.[2]

In this article, we discuss the implementation of the chi-squared goodness-of-fit test in count-data models as a postestimation command. `chi2gof` reports the test statistic, its degrees of freedom, and its p-value when used after the `poisson`, `nbreg`, `zip`, and `zinb` commands. As an option, the command produces a table with the cells, absolute frequencies, relative frequencies, predicted frequencies, and absolute differences between actual and predicted frequencies.

2 Statistical basis for the chi-squared goodness-of-fit test

2.1 The chi-squared goodness-of-fit test

Let's consider a model given by $f(y|\mathbf{w}, \boldsymbol{\theta})$, with the conditional density of the variable of interest (y) given a set of covariates ($\mathbf{w}$) and a vector of parameters ($\boldsymbol{\theta}$).[3] We are particularly interested in the conditional density of the Poisson, the negative binomial (NB), the zero-inflated Poisson (ZIP), and the zero-inflated negative binomial (ZINB) models. Thus $\mathbf{w} = \mathbf{x}$ in the Poisson and NB models and $\mathbf{w} = (\mathbf{x}, \mathbf{z})$ in the inflated versions (that is, $\mathbf{z}$ is the set of covariates used in the inflated part of the model). Also let J be the number of (mutually exclusive) cells in which the range of the dependent variable y_i is partitioned ($i = 1, \ldots, N$). Finally, let $d_{ij}(y_i) = \mathbf{1}(y_i \in j)$ be an indicator variable that takes value 1 if observation i belongs to cell j and 0 otherwise.

If the model is correctly specified, then

$$E\{d_{ij}(y_i) - p_{ij}(\mathbf{w}_i, \boldsymbol{\theta})\} = 0 \tag{1}$$

where $p_{ij}(\mathbf{w}_i, \boldsymbol{\theta})$ is the probability that observation i falls in cell j according to $f(y|\mathbf{w}, \boldsymbol{\theta})$. In particular, stacking all J moments in vector notation, (1) becomes

$$E\{\mathbf{d}_i(y_i) - \mathbf{p}_i(\mathbf{w}_i, \boldsymbol{\theta})\} = 0$$

Given a sample analog

$$\widehat{\mathbf{m}}_N\left(\widehat{\boldsymbol{\theta}}\right) = \frac{1}{N}\sum_{i=1}^{N}\left\{\mathbf{d}_i(y_i) - \mathbf{p}_i\left(\mathbf{w}_i, \widehat{\boldsymbol{\theta}}\right)\right\}$$

2. According to Cameron and Trivedi (2010, 266), "m-tests such as conditional moment tests are tests of whether moment conditions imposed by a model are satisfied" and "are a general specification testing procedure that encompasses many common specification tests".
3. This subsection is largely based on Greene (1994) and Cameron and Trivedi (2005, 2013).

the chi-squared goodness-of-fit test statistic of Andrews (1988a,b) is

$$N\widehat{\mathbf{m}}_N'\left(\widehat{\boldsymbol{\theta}}\right)\widehat{\mathbf{V}}^{-1}\widehat{\mathbf{m}}_N\left(\widehat{\boldsymbol{\theta}}\right) \tag{2}$$

where $\mathbf{V}$ is a variance–covariance matrix given by $\sqrt{N}\widehat{\mathbf{m}}_N(\widehat{\boldsymbol{\theta}}) \rightarrow N(0,\mathbf{V})$.

Under the null hypothesis that the moment condition (1) holds, the chi-squared goodness-of-fit test statistic is asymptotically χ^2 distributed with rank($\mathbf{V}$) degrees of freedom. However, $\mathbf{V}$ may not be of full rank. The rank is usually $J-1$ because the sum of the probabilities over all J cells is 1. Moreover, the computation of this variance–covariance matrix is often complicated.

This is why, when the maximum likelihood (ML) estimation is used, it is the outer product of the gradient form of the test that is usually computed. This is N times the (uncentered) R^2 of the following auxiliary regression,

$$1 = \widehat{\mathbf{m}}_i\boldsymbol{\delta} + \widehat{\mathbf{s}}_i\boldsymbol{\gamma} + u_i$$

where 1 is a column vector of N, $\widehat{\mathbf{m}}_i$ includes $d_{ij}(y_i) - p_{ij}(\mathbf{w}_i, \widehat{\boldsymbol{\theta}}^{\mathrm{ML}})$ for $j = 1,\ldots,J-1$ (the last column of $d_i - p_i$ has been dropped), and $\widehat{\mathbf{s}}_i = \{\partial \log f(y_i|\mathbf{w}_i,\boldsymbol{\theta})\}/(\partial\boldsymbol{\theta})|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{\mathrm{ML}}}$ is the matrix of contributions to the score evaluated at the ML estimate of $\boldsymbol{\theta}$. It is easy to see that the test statistic

$$N \times R^2 = 1'\mathbf{H}(\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'1$$

where $\mathbf{H}_i = [\widehat{\mathbf{m}}_i, \widehat{\mathbf{s}}_i]$ is the ith row of matrix $\mathbf{H}$. This asymptotically equivalent version of (2) is used in the `chi2gof` command. Under the null hypothesis of correct specification of the model, this statistic asymptotically follows a χ^2 distribution with $J-1$ degrees of freedom.

To conclude this section, we provide details of the computation of this test regarding both the predicted probabilities (p_{ij}) and the scores ($\widehat{\mathbf{s}}_i$).

2.2 Predicted probabilities

Let $\mu_i = e^{\mathbf{x}_i\boldsymbol{\beta}}$ be the conditional expectation of the Poisson model. This model predicts that the probability that the variable of interest takes the value t is

$$\Pr(y_i = t) = \frac{e^{\mu_i}\mu_i^t}{t!} = P_P(t)$$

Let $\Gamma(\cdot)$ be the gamma function (see, for example, Cameron and Trivedi [2013, 505–506] for details). The predicted probabilities of the NB model with conditional variance $\mu + \alpha\mu^2$ are

$$\Pr(y_i = t) = \frac{\Gamma\left(t+\alpha^{-1}\right)}{\Gamma\left(t+1\right)\Gamma\left(\alpha^{-1}\right)}\left(\frac{1}{1+\mu_i\alpha}\right)^{\alpha^{-1}}\left(\frac{\mu_i\alpha}{1+\mu_i\alpha}\right)^t = P_{\mathrm{NB}}(t)$$

Also let's denote the distribution function used in the inflated versions of the Poisson and NB models by φ. Stata currently supports two functions—the logit and the probit. Thus

$$\varphi_i = \varphi(\mathbf{z}_i, \boldsymbol{\gamma}) = \begin{cases} \Lambda(\mathbf{z}_i'\boldsymbol{\gamma}) = \frac{e^{\mathbf{z}_i'\boldsymbol{\gamma}}}{1+e^{\mathbf{z}_i'\boldsymbol{\gamma}}} & \text{in the logit case} \\ \\ \Phi(\mathbf{z}_i'\boldsymbol{\gamma}) = \int_{-\infty}^{\mathbf{z}_i'\boldsymbol{\gamma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du & \text{in the probit case} \end{cases}$$

Finally, if we denote an indicator function that takes value 1 if the condition in brackets is true and 0 otherwise by $\mathbf{1}(\cdot)$, then the predicted probabilities of the ZIP and NB regression models can be, respectively, expressed as follows:

$$\Pr(y_i = t) = \mathbf{1}(t = 0)\varphi_i + (1 - \varphi_i) P_P(t)$$

and

$$\Pr(y_i = t) = \mathbf{1}(t = 0)\varphi_i + (1 - \varphi_i) P_{\mathrm{NB}}(t)$$

2.3 Scores

The individual contribution to the likelihood function in the models considered is

$$f(y_i|\mathbf{x}_i, \boldsymbol{\theta}) = g(y_i) = \begin{cases} f(y_i|\mathbf{x}_i, \boldsymbol{\beta}) = P_P(y_i) & \text{in the Poisson model} \\ \\ f(y_i|\mathbf{x}_i, \boldsymbol{\beta}, \alpha) = P_{\mathrm{NB}}(y_i) & \text{in the NB model} \end{cases}$$

and

$$f(y_i|\mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\theta}) = \mathbf{1}(y_i = 0)\varphi_i + (1 - \varphi_i) g(y_i) \text{ in the inflated versions}$$

Thus the first derivative of the likelihood function in the Poisson model with respect to the parameters of interest, $\boldsymbol{\theta} = \boldsymbol{\beta}$, is

$$\mathbf{s}_i = \frac{\partial \log f(y_i|\mathbf{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{x}_i (y_i - \mu_i)$$

whereas the first derivative of the likelihood function in the NB model with respect to the parameters of interest, $\boldsymbol{\theta} = (\boldsymbol{\beta}, \alpha)$, is

$$\mathbf{s}_i = \left\{ \begin{array}{c} \frac{\partial \log f(y_i|\mathbf{x}_i, \boldsymbol{\beta}, \alpha)}{\partial \boldsymbol{\beta}} \\ \\ \frac{\partial \log f(y_i|\mathbf{x}_i, \boldsymbol{\beta}, \alpha)}{\partial \alpha} \end{array} \right\} = \left[\begin{array}{c} \mathbf{x}_i \left(\frac{y_i - \mu_i}{1 + \alpha\mu_i} \right) \\ \\ \frac{1}{\alpha^2} \left\{ \log(1 + \mu_i\alpha) - \sum_{t=0}^{y-1} \frac{1}{t + \alpha^{-1}} \right\} + \frac{y_i - \mu_i}{\alpha(1 + \mu_i\alpha)} \end{array} \right]$$

In the inflated versions of these models, $f(y_i|\mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\theta}) = f(y_i|\mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\beta}, \boldsymbol{\gamma})$ for the Poisson and $f(y_i|\mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\theta}) = f(y_i|\mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}, \alpha)$ for the NB. Therefore, the first derivative of the likelihood function with respect to the parameters of interest can be written as

$$\mathbf{s}_i = \left\{ \begin{array}{c} \frac{\partial \log f(y_i|\mathbf{x}_i,\mathbf{z}_i,\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} \\ \\ \frac{\partial \log f(y_i|\mathbf{x}_i,\mathbf{z}_i,\boldsymbol{\theta})}{\partial \boldsymbol{\gamma}} \\ \\ \frac{\partial \log f(y_i|\mathbf{x}_i,\mathbf{z}_i,\boldsymbol{\theta})}{\partial \alpha} \end{array} \right\} = \left\{ \begin{array}{c} \frac{(1-\varphi_i)}{f(y_i|\mathbf{x}_i,\mathbf{z}_i,\boldsymbol{\theta})} \frac{\partial g(y_i)}{\partial \boldsymbol{\beta}} \\ \\ \frac{\mathbf{1}(y_i=0)\varphi_i' - \varphi_i' g(y_i)}{f(y_i|\mathbf{x}_i,\mathbf{z}_i,\boldsymbol{\theta})} \\ \\ \frac{(1-\varphi_i)}{f(y_i|\mathbf{x}_i,\mathbf{z}_i,\boldsymbol{\theta})} \frac{\partial g(y_i)}{\partial \alpha} \end{array} \right\}$$

where

$$\frac{\partial g\left(y_i\right)}{\partial \boldsymbol{\beta}} = \begin{cases} P_P\left(y_i\right)\mathbf{x}_i\left(y_i - \mu_i\right) & \text{in the ZIP model} \\ \\ P_{\mathrm{NB}}\left(y_i\right)\mathbf{x}_i\left(\frac{y_i-\mu_i}{1+\alpha\mu_i}\right) & \text{in the ZINB model} \end{cases}$$

$$\varphi_i' = \frac{\partial \varphi_i}{\partial \boldsymbol{\gamma}} = \begin{cases} \mathbf{z}_i \frac{e^{\mathbf{z}_i'\boldsymbol{\gamma}}}{\left(1+e^{\mathbf{z}_i'\boldsymbol{\gamma}}\right)^2} & \text{in the logit case} \\ \\ \mathbf{z}_i \frac{1}{\sqrt{2\pi}} e^{-\frac{(\mathbf{z}_i'\boldsymbol{\gamma})^2}{2}} & \text{in the probit case} \end{cases}$$

and, in the case of the ZINB model,

$$\frac{\partial g\left(y_i\right)}{\partial \alpha} = P_{\mathrm{NB}}\left(y_i\right)\left[\frac{1}{\alpha^2}\left\{\log(1+\mu_i\alpha) - \sum_{t=0}^{y-1}\frac{1}{t+\alpha^{-1}}\right\} + \frac{y_i-\mu_i}{\alpha(1+\mu_i\alpha)}\right]$$

(Notice that this derivative is not needed in the ZIP model because $\alpha = 0$.)

3 The chi2gof command

3.1 Syntax

`chi2gof, cells(`*numlist*`)` [`prcount table`]

3.2 Options

`cells(`*numlist*`)` specifies a set of ascending integers greater than or equal to zero that determines the (mutually exclusive) cells in which the range of the dependent variable is partitioned to compute the test. `cells()` is required.

In principle, any partition of the dependent variable can be used (Andrews 1988b). For example, if 3 cells are chosen, the following partitions can be used: $\{0,1,2,3\}$, $\{4,5\}$, and $\{6,7,\ldots,\infty\}$; $\{0,1\}$, $\{2,3,4,5\}$, and $\{6,7,\ldots,\infty\}$; $\{0,1,2,3,4,5\}$, $\{6\}$, and $\{7,8,\ldots,\infty\}$; etc. Thus `chi2gof` allows partitions with both single-value elements (except for the last cell) and multiple-value elements. In the first case, *numlist* is the number of cells chosen by the user; in the second case, *numlist* is a set of integers that corresponds to the upper limits of the intervals considered.

- Choosing the number of cells involves using partitions like $\{0\}$ and $\{1, 2, 3, \ldots, \infty\}$ when `cells(2)`; $\{0\}$, $\{1\}$, and $\{2, 3, \ldots, \infty\}$ when `cells(3)`; $\{0\}$, $\{1\}$, $\{2\}$, and $\{3, 4, \ldots, \infty\}$ when `cells(4)`; and so on. In general, for `cells(J)`, the partition that `chi2gof` uses is $\{0\}$, $\{1\}$, $\{2\}, \ldots, \{J-2\}$, and $\{J-1, \ldots, \infty\}$.

- Choosing the upper limits of the intervals involves using partitions like $[0, 1]$, $[2, 5]$, and $[6, \infty)$ when `cells(1 5)`; $[0, 3]$, $[4, 4]$, $[5, 9]$, and $[10, \infty)$ when `cells(3 4 9)`; and $[0, 0]$, $[1, 1]$, $[2, 2]$, $[3, 3]$, and $[4, \infty)$ when `cells(0 1 2 3)`. In general, for `cells(`a_0 $a_1 \ldots a_{J-2}$`)`, the partition that `chi2gof` uses is $[0, a_0]$, $[a_0 + 1, a_1], \ldots, [a_{J-3} + 1, a_{J-2}]$, and $[a_{J-2} + 1, \infty)$.

Notice that `cells(0 1 2 ...` $J-2$`)` is equivalent to `cells(`J`)`. Notice also that to construct the partition, one must select an integer 2 or more for the number of cells J. However, the chosen number should prevent cell frequencies from getting too small (Cameron and Trivedi 2005, 2013). Thus users should look at the distribution of the dependent variable to ensure that cells do not have zero or very few observations. Users should also try using alternative values around the number of cells initially chosen.

`prcount` calculates the probability that according to the model, a particular value of the dependent variable belongs to one of the defined cells. By default, the command calculates these predicted probabilities (or predicted frequencies) using the definition of the conditional density of the dependent variable (`direct`). These probabilities can also be computed using the command `prcounts` of Long and Freese (2001). Results are generally the same when using either command. However, differences occur when the number of counts is high, particularly if the ZINB model is used. In this case, an error message results stating "`Missing values encountered when prcount option is used (try direct option)`".[4]

`table` produces a table with the absolute and relative frequencies of each defined cells, the mean fitted value of the relative frequencies (that is, the mean value of the predicted probabilities for each individual of each of the defined cell), and the absolute differences between actual and predicted frequencies. This can be useful in assessing the adequacy of the partition of the dependent variable being used. This may help to detect cells with too few observations. Also the table may help identify the source of misspecification. In the Poisson model, for example, big absolute differences in the zero value may indicate overdispersion.

4. Notice also that the statistic may not be computed for the ZINB model if the α parameter is too small. If it is, an error message states that a `Problem with alpha prevents estimation of predicted probabilities (alpha too small)`. In practice, this does not happen often because of the use of the `lngamma()` function. Ultimately, both error messages occur because of the large numbers that the `lngamma()` function generates (see section 2).

Note that, as expression (2) shows, we can interpret the absolute differences between actual and predicted frequencies as the approximate contribution of each cell to the chi-squared goodness-of-fit test (the exact contribution being a quadratic form in $\mathbf{V}$). Also each cell contributes the absolute differences between actual and predicted frequencies divided by the root of the predicted frequencies (the so-called Pearson residuals) to the chi-squared goodness-of-fit test when $\mathbf{V}$ is a diagonal matrix of the predicted frequencies. In this case, the chi-squared goodness-of-fit test becomes Pearson's chi-squared test. However, this is not the case in the count-data models considered here, nor is it in most regression applications (the multinomial logit model being an exception). This is why Pearson's residuals are generally not useful when analyzing the chi-squared goodness-of-fit test.

3.3 Stored results

`chi2gof` stores the following in `r()`:

Scalars

`r(chi2gof)`	chi-squared test statistic
`r(dof_chi2gof)`	degrees of freedom
`r(p_chi2gof)`	p-value

4 Examples

In applications, the model should be suspected of being misspecified (that is, the model moment conditions are not satisfied) if the resulting test is statistically significant. Otherwise, there is no evidence of misspecification in the model. We illustrate this using the four examples below, in which we show the use of the new command and the interpretation of its output in different settings.

Given the illustrative purpose of this section, we closely follow the sources of the examples (Cameron and Trivedi 2010, 2013) when describing the data and discussing the possible misspecification of the proposed models. We contribute by merely analyzing the results of the chi-squared goodness-of-fit test. We do not address the reasons behind the possible misspecification of the models.

In the first example, we replicate results from chapter 5 of Cameron and Trivedi (2013). In the second example, we replicate and extend results reported in chapter 6 of Cameron and Trivedi (2013). In the third and fourth examples, we replicate and extend results from chapter 17 of Cameron and Trivedi (2010). For all examples, we report the output resulting from both the estimation command (`poisson`, `nbreg`, `zip`, or `zinb`) and the new command (`chi2gof`). In the first and second examples, we also report a table with the cells, absolute frequencies, relative frequencies, predicted frequencies, and absolute differences between actual and predicted frequencies (option `table`).

Our results seem to confirm the original authors' conclusions in respect to the poor fit of the ZIP and the ZINB models in the second example. In the third example, the NB2 model provides a similar fit (in terms of information criteria) to more complex models such as the NB2 hurdle and the NB2 with a finite mixture. However, the chi-squared goodness-of-fit test suggests that the NB2 model is misspecified. In the fourth example, our results seem to confirm the authors' doubts about the NB2 model being outperformed by its inflated version (ZINB).

4.1 Example 1

The first application we consider here is the analysis of the determinants of takeover bids done by Cameron and Trivedi (2013), which uses a sample of 126 U.S. firms taken over between 1978 and 1985. The dependent variable is the number of bids received by the firm after the initial tender offer (`numbids`), while covariates include defensive actions taken by the management of the firm (`leglrest`, `realrest`, `finrest`, and `whtknght`), firm-specific characteristics (`bidprem`, `insthold`, `size`, and `sizesq`), and intervention by federal regulators (`regulatn`). The relation between the dependent and explanatory variables is fit using the Poisson regression model.

```
. infile docno weeks numbids takeover bidprem insthold size leglrest realrest
> finrest regulatn whtknght sizesq constant using
> http://cameron.econ.ucdavis.edu/racd/racd5.asc
(126 observations read)

. poisson numbids leglrest realrest finrest whtknght bidprem insthold size
> sizesq regulatn, nolog

Poisson regression                              Number of obs   =        126
                                                LR chi2(9)      =      33.25
                                                Prob > chi2     =     0.0001
Log likelihood = -184.94833                     Pseudo R2       =     0.0825

------------------------------------------------------------------------------
     numbids |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    leglrest |   .2601464   .1509594     1.72   0.085    -.0357286    .5560213
    realrest |  -.1956597   .1926309    -1.02   0.310    -.5732093    .1818899
     finrest |   .0740301   .2165219     0.34   0.732    -.3503452    .4984053
    whtknght |   .4813822   .1588698     3.03   0.002      .170003    .7927613
     bidprem |  -.6776958   .3767372    -1.80   0.072    -1.416087    .0606956
    insthold |  -.3619912   .4243292    -0.85   0.394    -1.193661    .4696788
        size |   .1785026   .0600221     2.97   0.003     .0608614    .2961438
      sizesq |  -.0075693   .0031217    -2.42   0.015    -.0136878   -.0014509
    regulatn |  -.0294392   .1605682    -0.18   0.855     -.344147    .2852686
       _cons |   .9860598   .5339201     1.85   0.065    -.0604044    2.032524
------------------------------------------------------------------------------
```

```
. chi2gof, cells(6) table
Chi-square Goodness-of-Fit Test for Poisson Model:
        Chi-square chi2(5)  =   48.66
        Prob>chi2           =    0.00
```

Cells	Abs. Freq.	Rel. Freq.	Fitted Rel. Freq.	Abs. Dif.
0	9	.0714	.2132	.1418
1	63	.5	.2977	.2023
2	31	.246	.2327	.0134
3	12	.0952	.1367	.0414
4	6	.0476	.068	.0204
5 or more	4	.0397	.0517	.012

From these results, reported on pages 185 and 195–196 of their book, Cameron and Trivedi (2013, 196) "[c]onclude that the Poisson is an inadequate fully parametric model, due to its inability to model the relatively few zeros in the sample". The table with the absolute and relative frequencies of each defined cell, the mean fitted value of the relative frequencies, and the absolute differences between actual and predicted frequencies that we report using the option `table` clarifies this. It is also interesting to note that "none of the earlier diagnostics [they performed], such as residual analysis, detected this weakness of the Poisson estimates" (Cameron and Trivedi 2013, 196).

4.2 Example 2

The second application we consider is Cameron and Trivedi's (2013) analysis of the determinants of the number of recreational boating trips to Lake Somerville, Texas, in 1980 (`trips`). Covariates include a subjective quality index of the facility (`so`), a dummy variable to indicate the practice of water-skiing at the lake (`ski`), the household income of the head of the group (`i`), a dummy variable to indicate whether the user paid a fee (`fc3`), dollar expenditure when visiting Lake Conroe (`c1`), dollar expenditure when visiting Lake Somerville (`c3`), and dollar expenditure when visiting Lake Houston (`c4`). In this analysis, they discuss different models (including finite mixtures and hurdle types of the Poisson and the NB models) and goodness-of-fit measures (the G^2 statistic, the pseudo-R^2, etc.). However, here we limit the reported results to the Poisson, NB2, ZIP, and ZINB estimates as well as to the chi-squared goodness-of-fit test.

```
. infile trips so ski i fc3 c1 c3 c4 using
> http://cameron.econ.ucdavis.edu/racd/racd6d2.asc, clear
(659 observations read)

. poisson trips so ski i fc3 c1 c3 c4, nolog

Poisson regression                                Number of obs   =        659
                                                  LR chi2(7)      =    2543.90
                                                  Prob > chi2     =     0.0000
Log likelihood = -1529.4313                       Pseudo R2       =     0.4540

------------------------------------------------------------------------------
       trips |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          so |   .4717259   .0170905    27.60   0.000     .4382291    .5052227
         ski |   .4182137   .0571905     7.31   0.000     .3061224    .5303051
           i |  -.1113232   .0195885    -5.68   0.000    -.1497159   -.0729304
         fc3 |   .8981652   .0789854    11.37   0.000     .7433567    1.052974
          c1 |  -.0034297   .0031178    -1.10   0.271    -.0095405    .0026811
          c3 |  -.0425364   .0016703   -25.47   0.000    -.0458102   -.0392626
          c4 |   .0361336   .0027096    13.34   0.000     .0308229    .0414444
       _cons |   .2649934   .0937224     2.83   0.005     .0813009    .4486859
------------------------------------------------------------------------------

. chi2gof, cells(6) table

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

         Chi-square chi2(5)  =  252.57
         Prob>chi2           =    0.00

------------------------------------------------------------
                                           Fitted
      Cells  Abs. Freq.  Rel. Freq.  Rel. Freq.   Abs. Dif.
------------------------------------------------------------
          0      417         .6328       .4196       .2131
          1       68         .1032       .2208       .1177
          2       38         .0577       .1031       .0454
          3       34         .0516       .0617       .0101
          4       17         .0258       .0449       .0191
  5 or more       72          .129       .1499       .0209
------------------------------------------------------------

. nbreg trips so ski i fc3 c1 c3 c4, nolog

Negative binomial regression                      Number of obs   =        659
                                                  LR chi2(7)      =     478.33
Dispersion     = mean                             Prob > chi2     =     0.0000
Log likelihood = -825.55758                       Pseudo R2       =     0.2246

------------------------------------------------------------------------------
       trips |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          so |    .721999   .0453323    15.93   0.000     .6331493    .8108487
         ski |   .6121388   .1504163     4.07   0.000     .3173282    .9069493
           i |  -.0260589   .0452342    -0.58   0.565    -.1147163    .0625986
         fc3 |   .6691677   .3614399     1.85   0.064    -.0392415    1.377577
          c1 |   .0480086   .0159516     3.01   0.003      .016744    .0792732
          c3 |   -.092691   .0082685   -11.21   0.000    -.1088969   -.0764851
          c4 |   .0388357   .0117139     3.32   0.001     .0158769    .0617945
       _cons |  -1.121936   .2208284    -5.08   0.000    -1.554752   -.6891205
-------------+----------------------------------------------------------------
    /lnalpha |   .3157293   .1060209                      .1079321    .5235264
-------------+----------------------------------------------------------------
       alpha |   1.371259   .1453821                      1.113972     1.68797
------------------------------------------------------------------------------
Likelihood-ratio test of alpha=0:  chibar2(01) = 1407.75 Prob>=chibar2 = 0.000
```

```
. chi2gof, cells(6) table
Chi-square Goodness-of-Fit Test for NegBin Model:
        Chi-square chi2(5)  =   23.54
        Prob>chi2           =    0.00
------------------------------------------------------------
                                           Fitted
      Cells  Abs. Freq.  Rel. Freq.  Rel. Freq.   Abs. Dif.
------------------------------------------------------------
          0      417         .6328       .6419       .0091
          1       68         .1032       .1224       .0192
          2       38         .0577       .0503       .0074
          3       34         .0516       .0303       .0213
          4       17         .0258       .0215       .0043
  5 or more       72          .129       .1336       .0046
------------------------------------------------------------
. zip trips so ski i fc3 c1 c3 c4, inflate(so ski i fc3 c1 c3 c4) robust nolog
Zero-inflated Poisson regression                  Number of obs   =        659
                                                  Nonzero obs     =        242
                                                  Zero obs        =        417
Inflation model      = logit                      Wald chi2(7)    =      75.75
Log pseudolikelihood = -1163.419                  Prob > chi2     =     0.0000
------------------------------------------------------------------------------
             |               Robust
       trips |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
trips        |
          so |   .0396788   .0834161     0.48   0.634    -.1238138    .2031715
         ski |   .4691185   .1763189     2.66   0.008     .1235398    .8146972
           i |  -.0943536   .0477011    -1.98   0.048    -.1878461   -.0008612
         fc3 |   .6050712   .2358399     2.57   0.010     .1428335    1.067309
          c1 |   .0023539   .0144217     0.16   0.870    -.0259121    .0306199
          c3 |  -.0364429   .0108506    -3.36   0.001    -.0577096   -.0151762
          c4 |   .0235891   .0081585     2.89   0.004     .0075987    .0395795
       _cons |   2.113707   .5032877     4.20   0.000     1.127281    3.100133
-------------+----------------------------------------------------------------
inflate      |
          so |  -1.651993   .2076671    -7.96   0.000    -2.059013   -1.244973
         ski |   .0588168   .4614636     0.13   0.899    -.8456352    .9632688
           i |  -.0719113   .1110972    -0.65   0.517    -.2896579    .1458352
         fc3 |  -20.59898   .6327039   -32.56   0.000    -21.83905    -19.3589
          c1 |  -.0058103   .0244693    -0.24   0.812    -.0537693    .0421486
          c3 |   .0723226   .0208863     3.46   0.001     .0313861     .113259
          c4 |  -.0753998   .0251974    -2.99   0.003    -.1247858   -.0260137
       _cons |   3.558284    .532032     6.69   0.000      2.51552    4.601047
------------------------------------------------------------------------------
```

```
. chi2gof, cells(6) table
Chi-square Goodness-of-Fit Test for ZIP Model:
         Chi-square chi2(5)  =  112.39
         Prob>chi2           =    0.00
------------------------------------------------------------
                                           Fitted
      Cells  Abs. Freq.  Rel. Freq.  Rel. Freq.   Abs. Dif.
------------------------------------------------------------
         0      417         .6328       .6354       .0026
         1       68         .1032        .033       .0701
         2       38         .0577        .042       .0156
         3       34         .0516       .0448       .0068
         4       17         .0258       .0431       .0173
  5 or more      72          .129       .2016       .0726
------------------------------------------------------------

. zinb trips so ski i fc3 c1 c3 c4, inflate(so ski i fc3 c1 c3 c4) robust nolog
Zero-inflated negative binomial regression       Number of obs   =        659
                                                 Nonzero obs     =        242
                                                 Zero obs        =        417
Inflation model      = logit                     Wald chi2(7)    =     140.80
Log pseudolikelihood = -719.3693                 Prob > chi2     =     0.0000
------------------------------------------------------------------------------
             |               Robust
       trips |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
trips        |
          so |    .170791   .0566653     3.01   0.003      .059729     .281853
         ski |    .492453   .1459854     3.37   0.001     .2063268    .7785792
           i |  -.0688226   .0414572    -1.66   0.097    -.1500773    .0124321
         fc3 |    .547295   .2205721     2.48   0.013     .1149817    .9796083
          c1 |   .0399582   .0184614     2.16   0.030     .0037745    .0761418
          c3 |  -.0658707   .0104456    -6.31   0.000    -.0863437   -.0453977
          c4 |   .0207245   .0113653     1.82   0.068    -.0015512    .0430001
       _cons |   1.091936   .2919291     3.74   0.000     .5197651    1.664106
-------------+----------------------------------------------------------------
inflate      |
          so |  -38.63617   2.002604   -19.29   0.000     -42.5612   -34.71114
         ski |  -16.06663   1.084321   -14.82   0.000    -18.19186    -13.9414
           i |  -.2029069   .3395567    -0.60   0.550    -.8684258     .462612
         fc3 |  -11.42997   2.420961    -4.72   0.000    -16.17496   -6.684971
          c1 |   -.023586   .0152638    -1.55   0.122    -.0535026    .0063305
          c3 |   .0775286   .0214598     3.61   0.000     .0354681    .1195891
          c4 |  -.0628993   .0214707    -2.93   0.003    -.1049811   -.0208175
       _cons |   20.97537    2.85067     7.36   0.000     15.38816    26.56258
-------------+----------------------------------------------------------------
    /lnalpha |  -.1832683   .1150975    -1.59   0.111    -.4088553    .0423186
-------------+----------------------------------------------------------------
       alpha |   .8325447   .0958238                      .6644104    1.043227
------------------------------------------------------------------------------
```

```
. chi2gof, cells(6) table
Chi-square Goodness-of-Fit Test for ZINegBin Model:
        Chi-square chi2(5)  =   18.34
        Prob>chi2           =    0.00
```

Cells	Abs. Freq.	Rel. Freq.	Fitted Rel. Freq.	Abs. Dif.
0	417	.6328	.6564	.0237
1	68	.1032	.0719	.0313
2	38	.0577	.0536	.0041
3	34	.0516	.0403	.0113
4	17	.0258	.0308	.005
5 or more	72	.129	.1469	.0179

Cameron and Trivedi (2013) initially analyze results from the Poisson and NB2 models. In the Poisson model, they notice that "the chi-squared goodness-of-fit test based on cells for $0, \ldots, 4$ and 5 or more trips [...] leads to a value of 252.6, much larger than the $\chi^2(5)$ critical value, [...] indicating a poor fit of the Poisson to the data"(Cameron and Trivedi 2013, 248). In the NB2 model, "[t]he statistic [...] is 23.5. Although this is a substantial improvement on the Poisson, the model is still rejected because the 5% critical value for $\chi^2(5)$ is 11.07"(Cameron and Trivedi 2013, 248–249). Thus none of these models fit the data well, and other specifications should be considered.

They also state, "Plausible alternatives to the models considered above are hurdle models, zero-inflated models, and finite-mixture models" (Cameron and Trivedi 2013, 250). However, because the `chi2gof` command does not cover either hurdle or finite-mixture models, here we concentrate on zero-inflated models (ZIP and ZINB). In the ZIP model, the chi-squared goodness-of-fit test shows a value much larger than that found in the NB2 model. In the ZINB model, the test indicates a better fit than that of the NB2, but it still rejects the null hypothesis of correct specification of the model.

We also report a table with the cells, absolute frequencies, relative frequencies, predicted frequencies, and absolute differences between actual and predicted frequencies. This partially replicates results reported in table 6.14 in Cameron and Trivedi (2013). We can see that the Poisson model performs poorly, underpredicting zeros and overpredicting positive outcomes. Its inflated version, the ZIP model, does a better job in predicting the zeros, and this substantially improves the fit (the statistic is 112.39). However, it still performs worse than the NB2. Finally, the ZINB yields the lower goodness-of-fit test (the statistic is 18.34) despite not predicting much better than the NB2.

4.3 Example 3

Using data from the U.S. Medical Expenditure Panel Survey for 2003, Cameron and Trivedi (2010) analyze the determinants of the annual number of doctor visits (`docvis`) for a sample of the Medicare population aged 65 and higher. Covariates include having private insurance that supplements Medicare (`private`), having public Medicaid insurance for low-income individuals that supplements Medicare (`medicaid`), age (`age`), squared age (`age2`), the years of education (`educyr`), the presence of an activity limitation (`actlim`), and the number of chronic conditions (`totchr`). They estimate the relationship between `docvis` and the covariates by using alternative estimators and specifications. However, again we restrict the analysis to the results from Poisson and NB2 models.

They first fit a Poisson regression model using the ML estimator, as follows:

```
. use http://www.stata-press.com/data/mus/mus17data, clear
. poisson docvis private medicaid age age2 educyr actlim totchr, nolog
Poisson regression                                Number of obs   =       3677
                                                  LR chi2(7)      =    4477.98
                                                  Prob > chi2     =     0.0000
Log likelihood =  -15019.64                       Pseudo R2       =     0.1297

------------------------------------------------------------------------------
      docvis |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     private |   .1422324   .0143311     9.92   0.000      .114144    .1703208
    medicaid |   .0970005   .0189307     5.12   0.000     .0598969     .134104
         age |   .2936722   .0259563    11.31   0.000     .2427988    .3445457
        age2 |  -.0019311   .0001724   -11.20   0.000    -.0022691   -.0015931
      educyr |   .0295562    .001882    15.70   0.000     .0258676    .0332449
      actlim |   .1864213    .014566    12.80   0.000     .1578726    .2149701
      totchr |   .2483898   .0046447    53.48   0.000     .2392864    .2574933
       _cons |  -10.18221   .9720115   -10.48   0.000    -12.08732   -8.277101
------------------------------------------------------------------------------

. chi2gof, cells(5)
Chi-square Goodness-of-Fit Test for Poisson Model:
         Chi-square chi2(4)  = 1011.40
         Prob>chi2           =    0.00
```

Results show that all the explanatory variables are statistically significant and have the expected sign. In particular, "`docvis` is increasing in age, education, number of chronic conditions, being limited in activity, and having either type of supplementary health insurance"(Cameron and Trivedi 2010, 574). However, the likelihood-ratio test reported after `nbreg` clearly shows that the parameter α is statistically significant. Thus the null hypothesis of equidispersion that the Poisson model implies is rejected by the data.[5]

5. Actually, Cameron and Trivedi (2010) use an auxiliary regression between $\{(y-\widehat{\mu})^2 - y\}/\widehat{\mu}$ and $\widehat{\mu}$ to test for equidispersion.

```
. use http://www.stata-press.com/data/mus/mus17data
. nbreg docvis private medicaid age age2 educyr actlim totchr, nolog
Negative binomial regression                      Number of obs   =       3677
                                                  LR chi2(7)      =     773.44
Dispersion     = mean                             Prob > chi2     =     0.0000
Log likelihood = -10589.339                       Pseudo R2       =     0.0352

------------------------------------------------------------------------------
      docvis |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     private |   .1640928   .0332186     4.94   0.000     .0989856    .2292001
    medicaid |    .100337   .0454209     2.21   0.027     .0113137    .1893603
         age |   .2941294   .0601588     4.89   0.000     .1762203    .4120384
        age2 |  -.0019282   .0004004    -4.82   0.000    -.0027129   -.0011434
      educyr |   .0286947   .0042241     6.79   0.000     .0204157    .0369737
      actlim |   .1895376   .0347601     5.45   0.000      .121409    .2576662
      totchr |   .2776441   .0121463    22.86   0.000     .2538378    .3014505
       _cons |  -10.29749   2.247436    -4.58   0.000    -14.70238   -5.892595
-------------+----------------------------------------------------------------
    /lnalpha |  -.4452773   .0306758                     -.5054007   -.3851539
-------------+----------------------------------------------------------------
       alpha |   .6406466   .0196523                      .6032638    .6803459
------------------------------------------------------------------------------
Likelihood-ratio test of alpha=0:  chibar2(01) = 8860.60 Prob>=chibar2 = 0.000
. chi2gof, cells(5)
Chi-square Goodness-of-Fit Test for NegBin Model:
         Chi-square chi2(4)  =   39.72
         Prob>chi2           =    0.00
```

Cameron and Trivedi (2010) then consider alternative models for handling the observed overdispersion, including the NB model, the Poisson and NB hurdle models, and the Poisson and NB finite-mixture models. They also compare their goodness of fit using the Akaike and Bayes criteria. These analyses lead them to conclude "that the NB2 hurdle model provides the best fitting and the most parsimonious specification" (Cameron and Trivedi 2010, 598). Still, the differences in fit between the NB2 hurdle model and the NB or the NB2 finite-mixture model are very small. On this basis, the NB2 model can be chosen to make inferences. The chi-squared goodness-of-fit test suggests, however, that this model is misspecified.

4.4 Example 4

Using the same dataset as in the previous example, Cameron and Trivedi (2010, 600–605) analyze the determinants of the number of emergency room visits by the survey respondent (`er`). They state, "The full set of explanatory variables in the model was initially the same as that used in the `docvis` example. However, after some preliminary analysis, this list was reduced to just three health-status variables—`age`, `actlim`, and `totchr`—that appeared to have some predictive power for `er`" (Cameron and Trivedi 2010, 600).

They first fit a NB model, as follows:

```
. use http://www.stata-press.com/data/mus/mus17data_z
. nbreg er age actlim totchr, nolog
Negative binomial regression                      Number of obs   =       3677
                                                  LR chi2(3)      =     225.15
Dispersion     = mean                             Prob > chi2     =     0.0000
Log likelihood = -2314.4927                       Pseudo R2       =     0.0464

------------------------------------------------------------------------------
          er |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0088528   .0061341     1.44   0.149    -.0031697    .0208754
      actlim |   .6859572   .0848127     8.09   0.000     .5197274    .8521869
      totchr |   .2514885   .0292559     8.60   0.000     .1941481     .308829
       _cons |  -2.799848   .4593974    -6.09   0.000    -3.700251   -1.899446
-------------+----------------------------------------------------------------
    /lnalpha |   .4464685   .1091535                      .2325315    .6604055
-------------+----------------------------------------------------------------
       alpha |   1.562783   .1705834                       1.26179    1.935577
------------------------------------------------------------------------------
Likelihood-ratio test of alpha=0:  chibar2(01) =  237.98 Prob>=chibar2 = 0.000
. chi2gof, cells(5)
Chi-square Goodness-of-Fit Test for NegBin Model:
         Chi-square chi2(4)  =    1.84
         Prob>chi2           =    0.76
```

Only `age` is not statistically significant in this model. However, because the proportion of zeros is relatively high—"[t]he first four values [...] account for over 99% of the probability mass of `er`" (Cameron and Trivedi 2010, 600)—they also consider the inflated version of the NB2 model.

```
. use http://www.stata-press.com/data/mus/mus17data_z
. zinb er age actlim totchr, inflate(age actlim totchr) vuong nolog
Zero-inflated negative binomial regression        Number of obs   =       3677
                                                  Nonzero obs     =        710
                                                  Zero obs        =       2967
Inflation model = logit                           LR chi2(3)      =      34.29
Log likelihood  = -2304.868                       Prob > chi2     =     0.0000

------------------------------------------------------------------------------
          er |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
er           |
         age |   .0035485   .0076344     0.46   0.642    -.0114146    .0185116
      actlim |   .2743106   .1768941     1.55   0.121    -.0723954    .6210165
      totchr |   .1963408   .0558635     3.51   0.000     .0868504    .3058313
       _cons |  -1.822978   .6515914    -2.80   0.005    -3.100074   -.5458825
-------------+----------------------------------------------------------------
inflate      |
         age |  -.0236763   .0284226    -0.83   0.405    -.0793835    .0320309
      actlim |   -4.22705   18.91192    -0.22   0.823    -41.29372    32.83962
      totchr |  -.3471091   .2052892    -1.69   0.091    -.7494686    .0552505
       _cons |   1.846526   2.071003     0.89   0.373    -2.212565    5.905618
-------------+----------------------------------------------------------------
    /lnalpha |   .1602371    .235185     0.68   0.496    -.3007171    .6211913
-------------+----------------------------------------------------------------
       alpha |   1.173789   .2760576                      .7402871    1.861144
------------------------------------------------------------------------------
Vuong test of zinb vs. standard negative binomial: z =     1.99  Pr>z = 0.0233
. chi2gof, cells(5)
Chi-square Goodness-of-Fit Test for ZINegBin Model:
         Chi-square chi2(4)  =    6.70
         Prob>chi2           =    0.15
```

To compare both models, Cameron and Trivedi (2010) use penalized log-likelihood-base statistics (the Akaike information criterion and Bayesian information criterion). Interestingly, "[t]his example indicates that having many zeros in the dataset does not automatically mean that a zero-inflated model is necessary. For these data, the ZINB model is only a slight improvement on the NB2 model and is actually no improvement at all if Bayesian information criterion is used as the model-selection criterion" (Cameron and Trivedi 2010, 605). Results of the chi-squared goodness-of-fit test confirm these conclusions because, although none of the models show signs of misspecification, the NB model yields a smaller statistic.

5 Concluding remarks

In this article, we discuss the implementation of the chi-squared goodness-of-fit test of Andrews (1988a,b) as a postestimation command. The new command `chi2gof` reports the test statistic, its degrees of freedom, and its p-value. It also stores these scalars as returned results in `r(chi2gof)`, `r(dof_chi2gof)`, and `r(p_chi2gof)`, respectively. As an option, the command produces a table with the actual, predicted, and absolute differences between actual and predicted frequencies. `chi2gof` can be used after the `poisson`, `nbreg`, `zip`, and `zinb` commands.

This specification test compares the sample relative frequencies of the dependent variable with the predicted frequencies of the model using a quadratic form and an estimate of the asymptotic variance of the corresponding population moment condition. Unlike Pearson's test (or the Hosmer–Lemeshow test), the chi-squared goodness-of-fit test can be constructed from any regular asymptotically normal estimator of the conditional expectation of the range of the dependent variable. In particular, `chi2gof` computes the test statistic using the outer product of the gradient form of the test (see Cameron and Trivedi [2005, 2013]).

We illustrate the use of the test in four examples from Cameron and Trivedi (2010, 2013). Under the null hypothesis of correct specification of the model, this statistic asymptotically follows a chi-squared distribution with $J-1$ degrees of freedom, J being the number of cells in which the dependent variable is partitioned. Thus, in applications, the model should be suspected of being misspecified (that is, the model moment conditions are not satisfied) if the resulting test is statistically significant. Otherwise, there is no evidence of misspecification in the model.

6 References

Andrews, D. W. K. 1988a. Chi-square diagnostic tests for econometric models: Introduction and applications. *Journal of Econometrics* 37: 135–156.

———. 1988b. Chi-square diagnostic tests for econometric models: Theory. *Econometrica* 56: 1419–1453.

Cameron, A. C., and P. K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. New York: Cambridge University Press.

———. 2010. *Microeconometrics Using Stata*. Rev. ed. College Station, TX: Stata Press.

———. 2013. *Regression Analysis of Count Data*. 2nd ed. Cambridge: Cambridge University Press.

Greene, W. H. 1994. Accounting for excess zeros and sample selection in Poisson and negative binomial regression models. Working Paper 94-10, Stern School of Business, Department of Economics.

Hosmer, D. W., Jr., and S. Lemeshow. 1980. Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics—Theory and Methods* 9: 1043–1069.

Hosmer, D. W., Jr., S. Lemeshow, and R. X. Sturdivant. 2013. *Applied Logistic Regression*. 3rd ed. Hoboken, NJ: Wiley.

Long, J. S., and J. Freese. 2001. Predicted probabilities for count models. *Stata Journal* 1: 51–57.

About the authors

Miguel Manjón and Oscar Martínez are associate professors at the Department of Economics, Rovira i Virgili University (Spain).