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	
1-year subscription	$ 79	1-year subscription	$115
2-year subscription	$155	2-year subscription	$225
3-year subscription	$225	3-year subscription	$329
3-year subscription (electronic only)	$210	3-year subscription (electronic only)	$210
1-year student subscription	$ 48	1-year student subscription	$ 79
1-year university library subscription	$ 99	1-year university library subscription	$135
2-year university library subscription	$195	2-year university library subscription	$265
3-year university library subscription	$289	3-year university library subscription	$395
1-year institutional subscription	$225	1-year institutional subscription	$259
2-year institutional subscription	$445	2-year institutional subscription	$510
3-year institutional subscription	$650	3-year institutional subscription	$750

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 (2012)
12, Number 3, pp. 447–453

A generalized Hosmer–Lemeshow goodness-of-fit test for multinomial logistic regression models

Morten W. Fagerland
Unit of Biostatistics and Epidemiology
Oslo University Hospital
Oslo, Norway
morten.fagerland@medisin.uio.no

David W. Hosmer
Department of Public Health
University of Massachusetts–Amherst
Amherst, MA

Abstract. Testing goodness of fit is an important step in evaluating a statistical model. For binary logistic regression models, the Hosmer–Lemeshow goodness-of-fit test is often used. For multinomial logistic regression models, however, few tests are available. We present the `mlogitgof` command, which implements a goodness-of-fit test for multinomial logistic regression models. This test can also be used for binary logistic regression models, where it gives results identical to the Hosmer–Lemeshow test.

Keywords: st0269, mlogitgof, goodness of fit, logistic regression, multinomial logistic regression, polytomous logistic regression

1 Introduction

Regression models for categorical outcomes should be evaluated for fit and adherence to model assumptions. There are two main elements of such an assessment: discrimination and calibration. Discrimination measures the ability of the model to correctly classify observations into outcome categories. Calibration measures how well the model-estimated probabilities agree with the observed outcomes, and it is typically evaluated via a goodness-of-fit test.

The (binary) logistic regression model describes the relationship between a binary outcome variable and one or more predictor variables. Several goodness-of-fit tests have been proposed (Hosmer and Lemeshow 2000, chap. 5), including the Hosmer–Lemeshow test (Hosmer and Lemeshow 1980), which is available in Stata through the postestimation command `estat gof`.

 st0269

The multinomial (or polytomous) logistic regression model is a generalization of the binary model when the outcome variable is categorical with more than two nominal (unordered) values. In Stata, a multinomial logistic regression model can be fit using the estimation command `mlogit`, but there is currently no goodness-of-fit test available. In this article, we will describe a Stata implementation of the multinomial goodness-of-fit test proposed by Fagerland, Hosmer, and Bofin (2008). Available through the command `mlogitgof`, this test can be used after both logistic regression (`logistic`) and multinomial logistic regression (`mlogit`). If used after `logistic`, it produces results identical to the Hosmer–Lemeshow test obtained from `estat gof`.

2 The goodness-of-fit test

Let Y denote an outcome variable with c unordered categories, coded $(0,\ldots,c-1)$. Assume that the outcome $Y=0$ is the reference (or baseline) outcome. Let $\mathbf{x}$ be a vector of p independent predictor variables, $\mathbf{x}=(x_1,x_2,\ldots,x_p)$. For details of the multinomial logistic regression model, we refer the reader to Hosmer and Lemeshow (2000, chap. 8) and to the Stata manual entry [R] **mlogit**.

Suppose that we have a sample of n independent observations, $(\mathbf{x}_i,y_i)$, $i=1,\ldots,n$. Recode y_i into binary indicator variables $\widetilde{y}_{ij}$, such that $\widetilde{y}_{ij}=1$ when $y_i=j$ and $\widetilde{y}_{ij}=0$ otherwise ($i=1,\ldots,n$ and $j=0,\ldots,c-1$). After fitting the model, let $\widehat{\pi}_{ij}$ denote the estimated probabilities for each observation ($i=1,\ldots,n$) for each possible outcome ($j=0,\ldots,c-1$).

The test is based on a strategy of sorting the observations according to $1-\widehat{\pi}_{i0}$, the complement of the estimated probability of the reference outcome. We then form g groups, each containing approximately n/g observations. For each group, we calculate the sums of the observed and estimated frequencies for each outcome category,

$$\begin{aligned} O_{kj} &= \sum_{l\in\Omega_k}\widetilde{y}_{lj} \\ E_{kj} &= \sum_{l\in\Omega_k}\widehat{\pi}_{lj} \end{aligned}$$

where $k=1,\ldots,g$; $j=0,\ldots,c-1$; and Ω_k denotes indices of the n/g observations in group k. A useful summary of the model's goodness of fit can be obtained by tabulating the values of O_{kj} and E_{kj} as shown in table 1.

Table 1. Contingency table of observed (O_{kj}) and estimated (E_{kj}) frequencies

Group	$Y=0$		$Y=1$		$\cdots$	$Y=c-1$	
1	O_{10}	E_{10}	O_{11}	E_{11}	$\cdots$	$O_{1,c-1}$	$E_{1,c-1}$
2	O_{20}	E_{20}	O_{21}	E_{21}	$\cdots$	$O_{2,c-1}$	$E_{2,c-1}$
$\vdots$	$\vdots$		$\vdots$		$\ddots$	$\vdots$	
g	O_{g0}	E_{g0}	O_{g1}	E_{g1}	$\cdots$	$O_{g,c-1}$	$E_{g,c-1}$

The multinomial goodness-of-fit test statistic is the Pearson's chi-squared statistic from the table of observed and estimated frequencies:

$$C_g = \sum_{k=1}^{g} \sum_{j=0}^{c-1} (O_{kj} - E_{kj})^2 / E_{kj}$$

Under the null hypothesis that the fitted model is the correct model and the sample is sufficiently large, Fagerland, Hosmer, and Bofin (2008) showed that the distribution of C_g is chi-squared and has $(g-2) \times (c-1)$ degrees of freedom.

3 The mlogitgof command

The `mlogitgof` command is a postestimation command that can be used after multinomial logistic regression (`mlogit`) or binary logistic regression (`logistic`). The syntax, options, and output of the command are similar to those of the postestimation command `estat gof`.

3.1 Syntax

`mlogitgof` [*if*] [*in*] [, `group(#)` `all` `outsample` `table`]

3.2 Options

`group(#)` specifies the number of quantiles to be used to group the observations. The default is `group(10)`.

`all` requests that the goodness-of-fit test be computed for all observations in the data, ignoring any `if` or `in` qualifiers specified with `mlogit` or `logistic`.

`outsample` adjusts the degrees of freedom for the goodness-of-fit test for samples outside the estimation sample.

`table` displays a table of the groups used for the goodness-of-fit test that lists the predicted probabilities, observed and expected counts for all outcomes, and totals for each group.

3.3 Saved results

`mlogitgof` saves the following in `r()`:

Scalars

`r(N)`	number of observations
`r(g)`	number of groups
`r(chi2)`	χ^2
`r(df)`	degrees of freedom
`r(P)`	probability greater than χ^2

4 Examples

```
. use http://www.stata-press.com/data/r12/sysdsn1
(Health insurance data)
. mlogit insure age nonwhite
  (output omitted)
. mlogitgof, table
Goodness-of-fit test for a multinomial logistic regression model
Dependent variable: insure
Table: observed and expected frequencies
```

Group	Prob	Obs_3	Exp_3	Obs_2	Exp_2	Obs_1	Exp_1	Total
1	0.4557	2	4.51	26	22.74	34	34.75	62
2	0.4737	6	4.45	27	23.93	28	32.62	61
3	0.4874	6	4.53	30	25.26	26	32.21	62
4	0.4996	7	4.45	21	25.72	33	30.82	61
5	0.5073	1	4.52	24	26.69	37	30.78	62
6	0.5170	5	4.45	24	26.78	32	29.77	61
7	0.5250	3	4.51	22	27.78	37	29.71	62
8	0.5479	6	4.43	32	28.14	23	28.43	61
9	0.6503	7	4.68	28	33.71	27	23.61	62
10	0.6914	2	4.46	43	36.25	16	20.29	61

```
  number of observations =    615
number of outcome values =      3
      base outcome value =      1
        number of groups =     10
   chi-squared statistic =     25.043
      degrees of freedom =     16
      Prob > chi-squared =      0.069
```

```
. mlogitgof if age < 40, group(8) table
Goodness-of-fit test for a multinomial logistic regression model
Dependent variable: insure
Table: observed and expected frequencies
```

Group	Prob	Obs_3	Exp_3	Obs_2	Exp_2	Obs_1	Exp_1	Total
1	0.5061	1	2.70	15	15.96	20	18.34	37
2	0.5115	3	2.63	11	15.71	19	17.67	36
3	0.5175	2	2.70	16	16.34	18	17.97	37
4	0.5217	2	2.62	12	16.08	20	17.30	36
5	0.5281	1	2.62	14	16.26	21	17.12	36
6	0.5372	2	2.69	21	17.00	11	17.32	37
7	0.6651	4	2.63	19	19.18	12	14.19	36
8	0.6961	1	2.61	24	21.74	7	11.64	36

```
  number of observations =    291
number of outcome values =      3
      base outcome value =      1
        number of groups =      8
    chi-squared statistic =     14.387
       degrees of freedom =     12
       Prob > chi-squared =      0.277
```

When used after `logistic`, `mlogitgof` produces results identical to the `estat gof` command:

```
. use http://www.stata-press.com/data/r12/lbw
(Hosmer & Lemeshow data)
. logistic low age lwt i.race smoke ptl ht ui
  (output omitted)
. estat gof, group(10) table
Logistic model for low, goodness-of-fit test
  (Table collapsed on quantiles of estimated probabilities)
```

Group	Prob	Obs_1	Exp_1	Obs_0	Exp_0	Total
1	0.0827	0	1.2	19	17.8	19
2	0.1276	2	2.0	17	17.0	19
3	0.2015	6	3.2	13	15.8	19
4	0.2432	1	4.3	18	14.7	19
5	0.2792	7	4.9	12	14.1	19
6	0.3138	7	5.6	12	13.4	19
7	0.3872	6	6.5	13	12.5	19
8	0.4828	7	8.2	12	10.8	19
9	0.5941	10	10.3	9	8.7	19
10	0.8391	13	12.8	5	5.2	18

```
      number of observations =       189
            number of groups =        10
     Hosmer-Lemeshow chi2(8) =         9.65
                 Prob > chi2 =         0.2904
```

```
. mlogitgof, table
Goodness-of-fit test for a binary logistic regression model
Dependent variable: low
Table: observed and expected frequencies
```

Group	Prob	Obs_1	Exp_1	Obs_0	Exp_0	Total
1	0.0827	0	1.18	19	17.82	19
2	0.1276	2	2.03	17	16.97	19
3	0.2015	6	3.17	13	15.83	19
4	0.2432	1	4.30	18	14.70	19
5	0.2792	7	4.89	12	14.11	19
6	0.3138	7	5.64	12	13.36	19
7	0.3872	6	6.54	13	12.46	19
8	0.4828	7	8.18	12	10.82	19
9	0.5941	10	10.31	9	8.69	19
10	0.8391	13	12.76	5	5.24	18

```
  number of observations =    189
number of outcome values =      2
     base outcome value =      0
       number of groups =     10
   chi-squared statistic =      9.651
     degrees of freedom =      8
     Prob > chi-squared =      0.290
```

5 Discussion

The `mlogitgof` command is designed to work similarly to the `estat gof` command. The main difference is that when `estat gof` is executed without the `group()` option, the ungrouped Pearson's chi-squared test is performed, whereas `mlogitgof` defaults to using $g = 10$ groups when executed without the `group()` option. The ungrouped test was not implemented in `mlogitgof` because it was found to be unsuitable for use in the simulation study by Fagerland, Hosmer, and Bofin (2008). In other aspects, the two commands produce identical results when applied after `logistic`.

As shown in section 2, the goodness-of-fit test is based on a comparison of observed and estimated frequencies in groups of observations defined by the estimated probability of the reference outcome. Different choices for reference outcome could produce different results. The sensitivity of the test to the choice of reference outcome is generally small (Fagerland, Hosmer, and Bofin 2008), but large differences may occur in specific datasets. When in doubt, perform the test for two or more choices for the reference outcome. It might also help to avoid using outcomes with few observations as reference outcome.

Goodness-of-fit tests target model misspecification and may detect a poorly fitting model. Alone, however, they cannot completely assess model fit. Goodness-of-fit tests should be considered as just one of several tools for assessing goodness of fit. Specifically, we cannot conclude that a model fits on the basis of a nonsignificant result from one goodness-of-fit test. The typical goodness-of-fit test analyzes unspecific deviations from model assumptions. To detect a specific departure of interest or the impact of individual observations, other procedures are often more useful, for example, regression diagnostics or certain graphical techniques (Hosmer and Lemeshow 2000, chap. 8).

Furthermore, a goodness-of-fit test is not something we use in the model-building stage to compare different models, such as the Akaike information criterion. We do not use goodness-of-fit tests to grade competing models or as a tool for selecting the best model. Instead, goodness-of-fit tests are used to assess the final model.

One general problem for logistic regression models is the low power of overall goodness-of-fit tests. This means that a large sample size is often necessary to detect small and medium model deviations. We refer the reader to Fagerland, Hosmer, and Bofin (2008) for a discussion on this and other limitations—such as the impact of the choice of groups—of the goodness-of-fit test for multinomial logistic regression.

6 References

Fagerland, M. W., D. W. Hosmer, and A. M. Bofin. 2008. Multinomial goodness-of-fit tests for logistic regression models. *Statistics in Medicine* 27: 4238–4253.

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.

———. 2000. *Applied Logistic Regression*. 2nd ed. New York: Wiley.

About the authors

Morten W. Fagerland is a senior researcher in biostatistics at Oslo University Hospital. His research interests include the application of statistical methods in medical research, analysis of categorical data and contingency tables, and comparisons of statistical methods using Monte Carlo simulations.

David W. Hosmer is a professor (emeritus) of biostatistics at the University of Massachusetts–Amherst and an adjunct professor of statistics at the University of Vermont. He is a coauthor of *Applied Logistic Regression*, of which a third edition is currently being written. His current research includes nonlogit link modeling of binary data, applications of logistic regression to modeling survival among thermally injured patients, and time-to-event modeling of fracture occurrence in an international cohort of elderly women.