Title

stata.com

estat gof — Pearson or Hosmer–Lemeshow goodness-of-fit test

Description

`estat gof` reports the Pearson goodness-of-fit test or the Hosmer–Lemeshow goodness-of-fit test.

`estat gof` requires that the current estimation results be from `logistic`, `logit`, or `probit`; see [R] **logistic**, [R] **logit**, or [R] **probit**. For `estat gof` after `poisson`, see [R] **poisson postestimation**. For `estat gof` after `sem`, see [SEM] **estat gof**.

Quick start

Pearson goodness-of-fit test for current estimation results

```
estat gof
```

As above, but apply to all observations in dataset instead of just those in `e(sample)`

```
estat gof, all
```

Hosmer–Lemeshow goodness-of-fit test

```
estat gof, group(10)
```

As above, and display table of groups used for the test

```
estat gof, group(10) table
```

Menu for estat

Statistics > Postestimation

Syntax

```
estat gof [if] [in] [weight] [, options]
```

options	Description
Main	
group(#)	perform Hosmer–Lemeshow goodness-of-fit test using # quantiles
all	execute test for all observations in the data
outsample	adjust degrees of freedom for samples outside estimation sample
table	display table of groups used for test

estat gof is not appropriate after the svy prefix.

collect is allowed; see [U] **11.1.10 Prefix commands**.

fweights are allowed; see [U] **11.1.6 weight**.

Options

Main

group(#) specifies the number of quantiles to be used to group the data for the Hosmer–Lemeshow goodness-of-fit test. group(10) is typically specified. If this option is not given, the Pearson goodness-of-fit test is computed using the covariate patterns in the data as groups.

all requests that the statistic be computed for all observations in the data, ignoring any if or in restrictions specified by the estimation command.

outsample adjusts the degrees of freedom for the Pearson and Hosmer–Lemeshow goodness-of-fit tests for samples outside the estimation sample. See *Samples other than the estimation sample* later in this entry.

table displays a table of the groups used for the Hosmer–Lemeshow or Pearson goodness-of-fit test with predicted probabilities, observed and expected counts for both outcomes, and totals for each group.

Remarks and examples

stata.com

Remarks are presented under the following headings:

Introduction

estat gof computes goodness-of-fit tests: either the Pearson χ^2 test or the Hosmer–Lemeshow test.

By default, estat gof computes statistics for the estimation sample by using the last model fit by logistic, logit, or probit. However, samples other than the estimation sample can be specified; see *Samples other than the estimation sample* later in this entry.

▷ Example 1

`estat gof`, typed without options, presents the Pearson χ^2 goodness-of-fit test for the fitted model. The Pearson χ^2 goodness-of-fit test is a test of the observed against expected number of responses using cells defined by the covariate patterns; see *predict with the number option* in [R] **logistic postestimation** for the definition of covariate patterns.

```
. use https://www.stata-press.com/data/r17/lbw
(Hosmer & Lemeshow data)
. logistic low age lwt i.race smoke ptl ht ui
  (output omitted)
. estat gof
Goodness-of-fit test after logistic model
Variable: low
      Number of observations =    189
Number of covariate patterns =    182
           Pearson chi2(173) = 179.24
                 Prob > chi2 = 0.3567
```

Our model fits reasonably well. However, the number of covariate patterns is close to the number of observations, making the applicability of the Pearson χ^2 test questionable but not necessarily inappropriate. Hosmer, Lemeshow, and Sturdivant (2013, 157–160) suggest regrouping the data by ordering on the predicted probabilities and then forming, say, 10 nearly equal-sized groups. `estat gof` with the `group()` option does this:

```
. estat gof, group(10)
note: obs collapsed on 10 quantiles of estimated probabilities.
Goodness-of-fit test after logistic model
Variable: low
 Number of observations =    189
       Number of groups =     10
Hosmer-Lemeshow chi2(8) =   9.65
            Prob > chi2 = 0.2904
```

Again we cannot reject our model. If we specify the `table` option, `estat gof` displays the groups along with the expected and observed number of positive responses (low-birthweight babies):

```
. estat gof, group(10) table
note: obs collapsed on 10 quantiles of estimated probabilities.

Goodness-of-fit test after logistic model
Variable: low

  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
```

In this table, the column `Prob` shows the upper boundaries of predicted probabilities for these 10 groups, which are the 10th, 20th, ..., and 100th percentiles in this case.

◁

❑ Technical note

`estat gof` with the `group()` option puts all observations with the same predicted probabilities into the same group. If, as in the previous example, we request 10 groups, the groups that `estat gof` makes are $[p_0, p_{10}]$, $(p_{10}, p_{20}]$, $(p_{20}, p_{30}]$, ..., $(p_{90}, p_{100}]$, where p_k is the kth percentile of the predicted probabilities, with p_0 the minimum and p_{100} the maximum.

If there are many ties at the quantile boundaries, as will often happen if all independent variables are categorical and there are only a few of them, the sizes of the groups will be uneven. If the totals in some of the groups are small, the χ^2 statistic for the Hosmer–Lemeshow test may be unreliable. In this case, fewer groups should be specified, or the Pearson goodness-of-fit test may be a better choice.

❑

▷ Example 2

The `table` option can be used without the `group()` option. We would not want to specify this for our current model because there were 182 covariate patterns in the data, caused by including the two continuous variables, `age` and `lwt`, in the model. As an aside, we fit a simpler model and specify `table` with `estat gof`:

```
. logistic low i.race smoke ui
```

```
Logistic regression                                     Number of obs =    189
                                                        LR chi2(4)    =  18.80
                                                        Prob > chi2   = 0.0009
Log likelihood = -107.93404                             Pseudo R2     = 0.0801
```

low	Odds ratio	Std. err.	z	P>|z|	[95% conf.	interval]
race						
Black	3.052746	1.498087	2.27	0.023	1.166747	7.987382
Other	2.922593	1.189229	2.64	0.008	1.316457	6.488285
smoke	2.945742	1.101838	2.89	0.004	1.415167	6.131715
ui	2.419131	1.047359	2.04	0.041	1.035459	5.651788
_cons	.1402209	.0512295	-5.38	0.000	.0685216	.2869447

Note: **_cons** estimates baseline odds.

```
. estat gof, table
```

```
Goodness-of-fit test after logistic model
Variable: low
```

Group	Prob	Obs_1	Exp_1	Obs_0	Exp_0	Total
1	0.1230	3	4.9	37	35.1	40
2	0.2533	1	1.0	3	3.0	4
3	0.2907	16	13.7	31	33.3	47
4	0.2923	15	12.6	28	30.4	43
5	0.2997	3	3.9	10	9.1	13
6	0.4978	4	4.0	4	4.0	8
7	0.4998	4	4.5	5	4.5	9
8	0.5087	2	1.5	1	1.5	3
9	0.5469	2	4.4	6	3.6	8
10	0.5577	6	5.6	4	4.4	10
11	0.7449	3	3.0	1	1.0	4

Group	Prob	race	smoke	ui
1	0.1230	White	Nonsmoker	0
2	0.2533	White	Nonsmoker	1
3	0.2907	Other	Nonsmoker	0
4	0.2923	White	Smoker	0
5	0.2997	Black	Nonsmoker	0
6	0.4978	Other	Nonsmoker	1
7	0.4998	White	Smoker	1
8	0.5087	Black	Nonsmoker	1
9	0.5469	Other	Smoker	0
10	0.5577	Black	Smoker	0
11	0.7449	Other	Smoker	1

```
      Number of observations =    189
Number of covariate patterns =     11
            Pearson chi2(6) =   5.71
                Prob > chi2 = 0.4569
```

❑ Technical note

`logistic`, `logit`, or `probit` and `estat gof` keep track of the estimation sample. If you type, for instance, `logistic ... if x==1`, then when you type `estat gof`, the statistics will be calculated on the `x==1` subsample of the data automatically.

You should specify `if` or `in` with `estat gof` only when you wish to calculate statistics for a set of observations other than the estimation sample. See *Samples other than the estimation sample* later in this entry.

If the `logistic` model was fit with `fweights`, `estat gof` properly accounts for the weights in its calculations. (`estat gof` allows only `fweights`.) You do not have to specify the weights when you run `estat gof`. Weights should be specified with `estat gof` only when you wish to use a different set of weights.

❑

Samples other than the estimation sample

`estat gof` can be used with samples other than the estimation sample. By default, `estat gof` remembers the estimation sample used with the last `logistic`, `logit`, or `probit` command. To override this, simply use an `if` or `in` restriction to select another set of observations, or specify the `all` option to force the command to use all the observations in the dataset.

If you use `estat gof` with a sample that is completely different from the estimation sample (that is, no overlap), you should also specify the `outsample` option so that the χ^2 statistic properly adjusts the degrees of freedom upward. For an overlapping sample, the conservative thing to do is to leave the degrees of freedom the same as they are for the estimation sample.

▷ Example 3

We want to develop a model for predicting low-birthweight babies. One approach would be to divide our data into two groups, a developmental sample and a validation sample. See Lemeshow and Gall (1994) and Tilford, Roberson, and Fiser (1995) for more information on developing prediction models and severity-scoring systems.

We will do this with the low-birthweight data that we considered previously. First, we randomly divide the data into two samples.

```
. use https://www.stata-press.com/data/r17/lbw, clear
(Hosmer & Lemeshow data)
. set seed 101
. generate r = runiform()
. sort r
. generate group = 1 if _n <= _N/2
(95 missing values generated)
. replace group = 2 if group==.
(95 real changes made)
```

Then, we fit a model using the first sample (group = 1), which is our developmental sample.

```
. logistic low age lwt i.race smoke ptl ht ui if group==1
Logistic regression                                     Number of obs =     94
                                                        LR chi2(8)    =  28.03
                                                        Prob > chi2   = 0.0005
Log likelihood = -42.351112                             Pseudo R2     = 0.2487

------------------------------------------------------------------------------
         low | Odds ratio   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |    .922865   .0555349    -1.33   0.182     .8201924     1.03839
         lwt |   .9825782   .0114438    -1.51   0.131     .9604029    1.005265
             |
        race |
       Black |   5.975476   4.936135     2.16   0.030     1.183652    30.16621
       Other |   3.364479   2.760784     1.48   0.139     .6736724      16.803
             |
       smoke |   3.442716    2.53779     1.68   0.094     .8117831    14.60032
         ptl |   3.467274   2.337648     1.84   0.065     .9249222    12.99784
          ht |   5.928512   6.047106     1.74   0.081     .8030021    43.76982
          ui |   4.045883   2.947396     1.92   0.055     .9703295     16.8697
       _cons |   3.120871   5.977489     0.59   0.552     .0731049     133.231
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

To test calibration in the developmental sample, we calculate the Hosmer–Lemeshow goodness-of-fit test by using `estat gof`.

```
. estat gof, group(10)
note: obs collapsed on 10 quantiles of estimated probabilities.

Goodness-of-fit test after logistic model
Variable: low

 Number of observations =     94
       Number of groups =     10
Hosmer-Lemeshow chi2(8) =   5.64
            Prob > chi2 = 0.6871
```

We did not specify an `if` statement with `estat gof` because we wanted to use the estimation sample. Because the test is not significant, we are satisfied with the fit of our model.

Running `lroc` (see [R] **lroc**) gives a measure of the discrimination:

```
. lroc, nograph

Logistic model for low

Number of observations =       94
Area under ROC curve   =   0.8145
```

Now, we test the calibration of our model by performing a goodness-of-fit test on the validation sample. We specify the `outsample` option so that the number of degrees of freedom is 10 rather than 8.

```
. estat gof if group==2, group(10) table outsample
note: obs collapsed on 10 quantiles of estimated probabilities.

Goodness-of-fit test after logistic model
Variable: low

  Table collapsed on quantiles of estimated probabilities
```

Group	Prob	Obs_1	Exp_1	Obs_0	Exp_0	Total
1	0.0276	0	0.2	10	9.8	10
2	0.0496	2	0.4	7	8.6	9
3	0.0875	1	0.7	9	9.3	10
4	0.1536	4	1.1	5	7.9	9
5	0.2283	4	2.0	6	8.0	10
6	0.2842	4	2.2	5	6.8	9
7	0.4190	3	3.6	7	6.4	10
8	0.5248	5	4.3	4	4.7	9
9	0.6413	5	5.8	5	4.2	10
10	0.9787	4	7.3	5	1.7	9

```
  Number of observations =     95
        Number of groups =     10
Hosmer-Lemeshow chi2(10) =  29.30
            Prob > chi2 = 0.0011
```

We must acknowledge that our model does not fit well on the validation sample. The model's discrimination in the validation sample is appreciably lower, as well.

```
. lroc if group==2, nograph

Logistic model for low

Number of observations =       95
Area under ROC curve   =   0.6835
```

◁

Stored results

`estat gof` stores the following in `r()`:

Scalars

`r(N)`	number of observations
`r(m)`	number of covariate patterns or groups
`r(df)`	degrees of freedom
`r(chi2)`	χ^2
`r(p)`	p-value for χ^2 test

Methods and formulas

Let M be the total number of covariate patterns among the N observations. View the data as collapsed on covariate patterns $j = 1, 2, \ldots, M$, and define m_j as the total number of observations having covariate pattern j and y_j as the total number of positive responses among observations with covariate pattern j. Define p_j as the predicted probability of a positive outcome in covariate pattern j.

The Pearson χ^2 goodness-of-fit statistic is

$$\chi^2 = \sum_{j=1}^{M} \frac{(y_j - m_j p_j)^2}{m_j p_j (1 - p_j)}$$

This χ^2 statistic has approximately $M - k$ degrees of freedom for the estimation sample, where k is the number of independent variables, including the constant. For a sample outside the estimation sample, the statistic has M degrees of freedom.

The Hosmer–Lemeshow goodness-of-fit χ^2 (Hosmer and Lemeshow 1980; Lemeshow and Hosmer 1982; Hosmer, Lemeshow, and Klar 1988) is calculated similarly, except that rather than using the M covariate patterns as the group definition, the quantiles of the predicted probabilities are used to form groups. Let $G = \#$ be the number of quantiles requested with `group(#)`. The smallest index $1 \leq q(i) \leq M$, such that

$$W_{q(i)} = \sum_{j=1}^{q(i)} m_j \geq \frac{N}{G}$$

gives $p_{q(i)}$ as the upper boundary of the ith quantile for $i = 1, 2, \ldots, G$. Let $q(0) = 1$ denote the first index.

The groups are then

$$[p_{q(0)}, p_{q(1)}], (p_{q(1)}, p_{q(2)}], \ldots, (p_{q(G-1)}, p_{q(G)}]$$

If the `table` option is given, the upper boundaries $p_{q(1)}, \ldots, p_{q(G)}$ of the groups appear next to the group number on the output.

The resulting χ^2 statistic has approximately $G - 2$ degrees of freedom for the estimation sample. For a sample outside the estimation sample, the statistic has G degrees of freedom.

References

Archer, K. J., and S. A. Lemeshow. 2006. Goodness-of-fit test for a logistic regression model fitted using survey sample data. *Stata Journal* 6: 97–105.

Fagerland, M. W., and D. W. Hosmer, Jr. 2012. A generalized Hosmer–Lemeshow goodness-of-fit test for multinomial logistic regression models. *Stata Journal* 12: 447–453.

Hosmer, D. W., Jr., and S. A. Lemeshow. 1980. Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics—Theory and Methods* 9: 1043–1069. https://doi.org/10.1080/03610928008827941.

Hosmer, D. W., Jr., S. A. Lemeshow, and J. Klar. 1988. Goodness-of-fit testing for the logistic regression model when the estimated probabilities are small. *Biometrical Journal* 30: 911–924. https://doi.org/10.1002/bimj.4710300805.

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

Lemeshow, S. A., and J.-R. L. Gall. 1994. Modeling the severity of illness of ICU patients: A systems update. *Journal of the American Medical Association* 272: 1049–1055. https://doi.org/10.1001/jama.1994.03520130087038.

Lemeshow, S. A., and D. W. Hosmer, Jr. 1982. A review of goodness of fit statistics for the use in the development of logistic regression models. *American Journal of Epidemiology* 115: 92–106. https://doi.org/10.1093/oxfordjournals.aje.a113284.

Nattino, G., S. A. Lemeshow, G. Phillips, S. Finazzi, and G. Bertolini. 2017. Assessing the calibration of dichotomous outcome models with the calibration belt. *Stata Journal* 17: 1003–1014.

Tilford, J. M., P. K. Roberson, and D. H. Fiser. 1995. sbe12: Using lfit and lroc to evaluate mortality prediction models. *Stata Technical Bulletin* 28: 14–18. Reprinted in *Stata Technical Bulletin Reprints*, vol. 5, pp. 77–81. College Station, TX: Stata Press.

Also see

[R] **logistic** — Logistic regression, reporting odds ratios

[R] **logit** — Logistic regression, reporting coefficients

[R] **probit** — Probit regression

[R] **estat classification** — Classification statistics and table

[R] **lroc** — Compute area under ROC curve and graph the curve

[R] **lsens** — Graph sensitivity and specificity versus probability cutoff

[U] **20 Estimation and postestimation commands**