lcstats — Latent class model-comparison statistics

Description

`lcstats` calculates model-comparison statistics for latent class models fit using `fmm` or `gsem`. You can specify which statistics to show in the reported table, including the number of classes, estimation sample size, log likelihood, rank, entropy, Akaike information criterion (AIC), Schwarz Bayesian information criterion (BIC), corrected AIC (AICc), consistent AIC (CAIC), Vuong–Lo–Mendell–Rubin (VLMR) likelihood-ratio test, and Lo–Mendell–Rubin (LMR)-adjusted likelihood-ratio test.

The VLMR and LMR tests are commonly used to determine the number of latent classes your data supports for similarly specified models. To conduct the VLMR and LMR tests, you must store the estimation results using `estimates store`. `lcstats` also works with the current estimation results.

Quick start

Report the default statistics—number of classes, sample size, log likelihood, rank, and entropy—for a logit outcomes model with two latent classes

```
gsem y* <-, logit lclass(C 2)
lcstats .
```

Compare a logit outcomes model with 1 latent class to a logit outcomes model with 2 latent classes; report default statistics, including the LMR-adjusted likelihood-ratio test for 2 classes versus 1 class

```
gsem y* <-, logit lclass(C 1)
estimate store m1
gsem y* <-, logit lclass(C 2)
lcstats m1 .
```

Same as above, but also show AIC and BIC

```
lcstats m1 ., aic bic
```

Same as above, but split the output into two tables

```
lcstats m1 ., aic bic split
```

Same as above, but specify how to split the output

```
lcstats m1 ., results(N rank aic bic entropy) results(k_classes ll df lmr p_lmr)
```

Specify a single table, and select statistics of interest and column order

```
lcstats m1 ., results(k_classes bic lmr p_lmr entropy)
```

Menu

Statistics > Postestimation

Syntax

```
lcstats [namelist] [ , options ]
```

namelist is a name, a list of names, _all, or *. A name may be ., meaning the current (active) estimates. _all and * mean the same thing. If *namelist* is not specified, the current (active) estimates is used; this is equivalent to specifying *namelist* as ".".

name is the name under which estimation results were stored using estimates store (see [R] **estimates store**), and "." refers to the last estimation results, whether or not these were already stored.

options	Description
Main	
all	report all available statistics
noentropy	do not report entropy
allic	report AIC, BIC, AICc, and CAIC
aic	report AIC
bic	report BIC
aicc	report AICc
caic	report CAIC
noicnotes	suppress notes for information criteria
nolrtests	do not report likelihood-ratio tests
lmr	report the LMR-adjusted likelihood-ratio test
vlmr	report the VLMR likelihood-ratio test
nolrnotes	suppress notes for likelihood-ratio tests
Formats	
* pformat([*%fmt*] [, ...])	specify numeric format for *p*-values
nformat(*%fmt* [*results*][, basestyle])	specify numeric format
Split tables	
split	split output into two tables
results(*results*)	specify results and their order for separate tables
Options	
[no]shownames	show or hide estimates' names
extraspace(#)	specify the number of extra spaces between columns
name(*cname*)	work with collection *cname*; default is name(LCStats)
replace	replace the collection
label(*filename*)	specify the collection labels
style(*filename*[, override])	specify the collection style

*The full specification is pformat([*%fmt*] [, minimum([#][, label(*string*)])]).

results is a list of result names and may include any of the following:

results	Definition
`k_classes`	number of classes
`N`	sample size
`ll`	log likelihood
`rank`	rank of `e(V)`
`aic`	AIC
`bic`	BIC
`aicc`	AICc
`caic`	CAIC
`entropy`	measure of separation between latent classes
`df`	degrees of freedom for the likelihood-ratio tests
`vlmr`	VLMR likelihood-ratio test statistic
`p_vlmr`	p-value for VLMR
`lmr`	LMR-adjusted likelihood-ratio test statistic
`p_lmr`	p-value for LMR

Options

Main

`all` specifies that all available statistics be reported in the output. This option is a shortcut for specifying `aic`, `bic`, `aicc`, `caic`, `entropy`, `lmr`, and `vlmr`.

`noentropy` specifies that entropy not be reported.

`allic`, `aic`, `bic`, `aicc`, `caic`, and `noicnotes` control the reporting of information criteria and their notes. The default is to not report information criteria.

`allic` specifies that all information criteria be reported in the output. This option is a shortcut for specifying `aic`, `bic`, `aicc`, and `caic`.

`aic` specifies that AIC be reported.

`bic` specifies that BIC be reported.

`aicc` specifies that AICc be reported. This information criterion is a second-order approximation and is recommended for small sample sizes.

`caic` specifies that CAIC be reported. This information criterion is a consistent version of AIC; that is, the probability of selecting the "true model" approaches 1 as sample size increases.

`noicnotes` suppresses the notes for the information criteria.

`nolrtests`, `lmr`, `vlmr`, and `nolrnotes` control reporting of likelihood-ratio tests comparing models with C versus $C-1$ latent classes. The default is `lmr`.

`nolrtests` specifies that no likelihood-ratio test be reported.

`lmr` specifies that the LMR-adjusted likelihood-ratio test be reported.

`vlmr` specifies that the VLMR likelihood-ratio test be reported.

`nolrnotes` suppresses the likelihood-ratio test notes.

Formats

`pformat([`*%fmt*`] [, minimum([#][, label(`*string*`)])])` changes the numeric format, such as the number of decimal places, for p-value results `p_lmr` and `p_vlmr`.

`minimum([#][, label(`*string*`)])` specifies that p-values less than # be displayed as "<#", where # is formatted according to *%fmt*.

If suboption `label(`*string*`)` is specified, then "*string*" is used instead of "<#". If *string* contains `%s`, then `%s` is replaced by # formatted according to *%fmt*.

The default style is equivalent to `pformat(%6.3f, minimum(.001))`.

`nformat(`*%fmt* [*results*] [`, basestyle`]) changes the numeric format, such as the number of decimal places, for specified results. If *results* are not specified, the numeric format is changed for all results.

This option is repeatable, and when multiple formats apply to one result, the rightmost specification is applied. Note that specifying a `pformat()` option will override any `nformat()` option applied to the p-value results `p_lmr` and `p_vlmr`, regardless of the order that the options are specified.

`basestyle` indicates that the format be applied to results that do not already have their own format instead of overriding the format for all results.

The default style is equivalent to

```
nformat(%9.0g, basestyle)
nformat(%6.4f entropy)
nformat(%21.0fc N k_classes rank df)
nformat(%21.2fc aic bic aicc caic)
nformat(%21.2fc ll lmr vlmr)
```

Split tables

`split` and `results(`*results*`)` control how to split the reported statistics into multiple tables.

`split` is a shortcut for splitting the results into two tables: entropy and the information criteria are reported in the first table; likelihood-ratio test results are reported in the second table.

By default, `split` is a shortcut for

```
results(N rank entropy)
results(k_classes ll df lmr p_lmr)
```

With option `all`, `split` is a shortcut for

```
results(N rank aic bic aicc caic entropy)
results(k_classes ll df vlmr p_vlmr lmr p_lmr)
```

`results(`*results*`)` specifies the results to report in the table columns. This option is repeatable, and each specification defines a separate table. Results not selected in any of the specified `results()` options are suppressed from the output. Repeating results is not allowed.

Options

`shownames` and `noshownames` control reporting of estimates' names in the table row headers. The default is to show the estimates' names in the table row headers.

`extraspace(#)` specifies extra spaces to pad columns in each reported table. The first and middle columns get # extra spaces added on both sides. The last column gets # extra spaces added on the left. The default is `extraspace(1)`.

This column property is also respected by collect export when publishing your collection to SMCL and plain text.

name(*cname*) specifies the collection for lcstats to work with. The default is name(LCStats).

replace permits lcstats to overwrite the existing collection. This option is implied for name(LCStats).

label(*filename*) specifies the *filename* containing the collection labels to use for your table. Labels in *filename* will be loaded for the table, and default labels will be used for any labels not specified in *filename*.

style(*filename*[, override]) specifies the *filename* containing the collection styles to use for your table. This might be a style you saved with collect style save or a predefined style shipped with Stata. The lcstats collection styles will be discarded, and only the collection styles in *filename* will be applied. Note that the layout specification saved in *filename* will not be applied; lcstats will always specify the layout.

If you prefer the lcstats collection styles but also want to apply any styles in *filename*, specify override. If there are conflicts between the default collection styles and those in *filename*, the ones in *filename* will take precedence.

The default is to use only the collection styles defined in style-lcstats.stjson; see [TABLES] **Predefined styles**.

Remarks and examples

lcstats is illustrated in [SEM] **Example 51g** and [SEM] **Example 52g**.

Stored results

lcstats stores the following in r():

Matrices
r(S) latent class statistics

The rows of r(S) correspond with the names of the estimation results in the order they were specified. See the *results* table in *Syntax* for the complete list and order of the columns of r(S).

Methods and formulas

For each estimation result, `lcstats` collects or computes the following:

- `k_classes`: number of classes, `e(lclass_k_levels)`
- `N`: sample size, `e(N)`
- `ll`: log likelihood, `e(ll)`
- `rank`: rank of `e(V)`
- `aic`: AIC
- `bic`: BIC
- `aicc`: AICc
- `caic`: CAIC
- `entropy`: measure of separation between latent classes

Akaike's (1974) information criterion is defined as

$$\texttt{aic} = -2\ln L + 2k$$

where $\ln L$ is the maximized log likelihood of the model and k is the number of parameters estimated (that is, `rank`). Schwarz's (1978) BIC is another measure of fit defined as

$$\texttt{bic} = -2\ln L + k\ln N$$

where N is the sample size. Hurvich and Tsai (1989) derived a second-order variant of AIC called AICc,

$$\texttt{aicc} = \texttt{aic} + \frac{2k(k+1)}{N-k-1}$$

Bozdogan (1987) proposed a consistent version of AIC called CAIC,

$$\texttt{caic} = -2\ln L + k(\ln N + 1)$$

See [R] **estat ic** for a focused discussion of these information criteria.

Entropy is computed from the predicted posterior latent class probabilities, as described by Ramaswamy et al. (1993). Let C be the number of latent classes for a given estimation and $\hat{p}_{ij}$ be the predicted posterior probability for class i in observation j, where $i = 1, \ldots, C$ and $j = 1, \ldots, N$. Then

$$\texttt{entropy} = 1 + \frac{1}{N\ln(C)} \sum_{j=1}^{N} \sum_{i=1}^{C} \hat{p}_{ij} \ln(\hat{p}_{ij})$$

`entropy` ranges from 0 to 1, and values closer to 1 indicate better separation between latent classes.

Let M_1 and M_2 denote estimation results based on the same data and model specifications for the observed outcome variables. Denote their log-likelihood values by $\ln L_1$ and $\ln L_2$ and ranks by k_1 and k_2. Suppose M_1 has $C-1$ latent classes and M_2 has C latent classes. Then, according to Vuong (1989) and Lo, Mendell, and Rubin (2001), the likelihood-ratio test statistic

$$\texttt{vlmr} = 2(\ln L_2 - \ln L_1)$$

is asymptotically distributed as a weighted sum of independent χ^2_1 variables. The LMR-adjusted likelihood-ratio test statistic is

$$\texttt{lmr} = \frac{2(\ln L_2 - \ln L_1)}{1 + 1/\{(k_2 - k_1)\ln N\}}$$

and has the same asymptotic distribution as `vlmr`. The reported degrees of freedom for these likelihood-ratio tests is

$$\texttt{df} = k_2 - k_1$$

The p-values `p_vlmr` and `p_lmr` are computed using a numerical approximation of the distribution of the weighted sum of independent χ^2_1 variables.

References

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19: 716–723. https://doi.org/10.1109/TAC.1974.1100705.

Bozdogan, H. 1987. Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika* 52: 345–370. https://doi.org/10.1007/BF02294361.

Hurvich, C. M., and C.-L. Tsai. 1989. Regression and time series model selection in small samples. *Biometrika* 76: 297–307. https://doi.org/10.1093/biomet/76.2.297.

Lo, Y., N. R. Mendell, and D. B. Rubin. 2001. Testing the number of components in a normal mixture. *Biometrika* 88: 767–778. https://doi.org/10.1093/biomet/88.3.767.

Ramaswamy, V., W. S. Desarbo, D. J. Reibstein, and W. T. Robinson. 1993. An empirical pooling approach for estimating marketing mix elasticities with PIMS data. *Marketing Science* 12: 103–124. https://doi.org/10.1287/mksc.12.1.103.

Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics* 6: 461–464. https://doi.org/10.1214/aos/1176344136.

Vuong, Q. H. 1989. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* 57: 307–333. https://doi.org/10.2307/1912557.

Also see

[SEM] **gsem** — Generalized structural equation model estimation command

[SEM] **gsem postestimation** — Postestimation tools for gsem