Title

stata.com

discrim logistic — Logistic discriminant analysis

Description

`discrim logistic` performs logistic discriminant analysis. See [MV] **discrim** for other discrimination commands.

Quick start

Logistic discriminant analysis of `v1`, `v2`, `v3`, and `v4` for groups defined by `catvar`

```
discrim logistic v1 v2 v3 v4, group(catvar)
```

As above, but use prior probabilities that are proportional to group size

```
discrim logistic v1 v2 v3 v4, group(catvar)  ///
     priors(proportional)
```

As above, but suppress iteration log

```
discrim logistic v1 v2 v3 v4, group(catvar)  ///
     priors(proportional) nolog
```

Menu

Statistics > Multivariate analysis > Discriminant analysis > Logistic

Syntax

```
discrim logistic varlist [if] [in] [weight], group(groupvar) [options]
```

options	Description
Model	
*group(*groupvar*)	variable specifying the groups
priors(*priors*)	group prior probabilities
ties(*ties*)	how ties in classification are to be handled
Reporting	
notable	suppress resubstitution classification table
nolog	suppress the mlogit log-likelihood iteration log

priors	Description
equal	equal prior probabilities; the default
proportional	group-size-proportional prior probabilities
matname	row or column vector containing the group prior probabilities
matrix_exp	matrix expression providing a row or column vector of the group prior probabilities

ties	Description
missing	ties in group classification produce missing values; the default
random	ties in group classification are broken randomly
first	ties in group classification are set to the first tied group

*group() is required.

statsby and xi are allowed; see [U] **11.1.10 Prefix commands**.

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

See [U] **20 Estimation and postestimation commands** for more capabilities of estimation commands.

Options

Model

group(*groupvar*) is required and specifies the name of the grouping variable. *groupvar* must be a numeric variable.

priors(*priors*) specifies the prior probabilities for group membership. The following *priors* are allowed:

priors(equal) specifies equal prior probabilities. This is the default.

priors(proportional) specifies group-size-proportional prior probabilities.

priors(*matname*) specifies a row or column vector containing the group prior probabilities.

priors(*matrix_exp*) specifies a matrix expression providing a row or column vector of the group prior probabilities.

`ties(`*ties*`)` specifies how ties in group classification will be handled. The following *ties* are allowed:

`ties(missing)` specifies that ties in group classification produce missing values. This is the default.

`ties(random)` specifies that ties in group classification are broken randomly.

`ties(first)` specifies that ties in group classification are set to the first tied group.

Reporting

`notable` suppresses the computation and display of the resubstitution classification table.

`nolog` suppress the `mlogit` log-likelihood iteration log.

Remarks and examples

stata.com

Albert and Lesaffre (1986) explain that logistic discriminant analysis is a partially parametric method falling between parametric discrimination methods such as LDA and QDA (see [MV] **discrim lda** and [MV] **discrim qda**) and nonparametric discrimination methods such as kth-nearest-neighbor (KNN) discrimination (see [MV] **discrim knn**). Albert and Harris (1987) provide a good explanation of logistic discriminant analysis. Instead of making assumptions about the distribution of the data within each group, logistic discriminant analysis is based on the assumption that the likelihood ratios of the groups have an exponential form; see *Methods and formulas*. Multinomial logistic regression provides the basis for logistic discriminant analysis; see [R] **mlogit**. Multinomial logistic regression can handle binary and continuous regressors, and hence logistic discriminant analysis is also appropriate for binary and continuous discriminating variables.

▷ Example 1: A two-group logistic discriminant analysis

Morrison (2005, 443–445) provides data on 12 subjects with a senile-factor diagnosis and 37 subjects with a no-senile-factor diagnosis. The data consist of the Wechsler Adult Intelligence Scale (WAIS) subtest scores for information, similarities, arithmetic, and picture completion. Morrison (2005, 231) performs a logistic discriminant analysis on the two groups, using the similarities and picture completion scores as the discriminating variables.

```
. use http://www.stata-press.com/data/r14/senile
(Senility WAIS subtest scores)
. discrim logistic sim pc, group(sf) priors(proportional)
Iteration 0:   log likelihood = -27.276352
Iteration 1:   log likelihood = -19.531198
Iteration 2:   log likelihood = -19.036702
Iteration 3:   log likelihood = -19.018973
Iteration 4:   log likelihood = -19.018928
Logistic discriminant analysis
Resubstitution classification summary
```

Key
Number Percent

True sf	Classified No-SF	SF	Total
No-SF	37	0	37
	100.00	0.00	100.00
SF	6	6	12
	50.00	50.00	100.00
Total	43	6	49
	87.76	12.24	100.00
Priors	0.7551	0.2449	

We specified the `priors(proportional)` option to obtain proportional prior probabilities for our logistic classification. These results match those of Morrison (2005, 231), though he does not state that his results are based on proportional prior probabilities. If you change to equal prior probabilities you obtain different classification results.

Which observations were misclassified? `estat list` with the `misclassified` option shows the six misclassified observations and the estimated probabilities.

```
. estat list, misclassified varlist
```

Obs.	Data sim	pc	Classification True	Class.	Probabilities No-SF	SF
38	5	8	SF	No-SF *	0.7353	0.2647
41	7	9	SF	No-SF *	0.8677	0.1323
44	9	8	SF	No-SF *	0.8763	0.1237
46	7	6	SF	No-SF *	0.6697	0.3303
48	10	3	SF	No-SF *	0.5584	0.4416
49	12	10	SF	No-SF *	0.9690	0.0310

```
* indicates misclassified observations
```

See example 1 of [MV] **discrim logistic postestimation** for more postestimation analysis with this logistic discriminant analysis.

◁

Example 2: A three-group logistic discriminant analysis

Example 2 of [MV] **discrim knn** introduces a head measurement dataset with six discriminating variables and three groups; see Rencher and Christensen (2012, 290–292). We now apply `discrim logistic` to see how well the logistic model can discriminate between the groups.

```
. use http://www.stata-press.com/data/r14/head
(Table 8.3 Head measurements, Rencher and Christensen (2012))
. discrim logistic wdim circum fbeye eyehd earhd jaw, group(group)
Iteration 0:   log likelihood = -98.875106
Iteration 1:   log likelihood = -60.790737
Iteration 2:   log likelihood = -53.746934
Iteration 3:   log likelihood = -51.114631
Iteration 4:   log likelihood = -50.249426
Iteration 5:   log likelihood = -50.081199
Iteration 6:   log likelihood = -50.072248
Iteration 7:   log likelihood = -50.072216
Logistic discriminant analysis
Resubstitution classification summary
```

Key
Number Percent

True group	Classified high school	college	nonplayer	Total
high school	27 90.00	2 6.67	1 3.33	30 100.00
college	1 3.33	20 66.67	9 30.00	30 100.00
nonplayer	2 6.67	8 26.67	20 66.67	30 100.00
Total	30 33.33	30 33.33	30 33.33	90 100.00
Priors	0.3333	0.3333	0.3333	

The counts on the diagonal of the resubstitution classification table are similar to those obtained by `discrim knn` (see example 2 of [MV] **discrim knn**) and `discrim lda` (see example 1 of [MV] **candisc**), whereas `discrim qda` seems to have classified the nonplayer group more accurately (see example 3 of [MV] **discrim estat**).

Stored results

`discrim logistic` stores the following in `e()`:

Scalars

`e(N)`	number of observations
`e(N_groups)`	number of groups
`e(k)`	number of discriminating variables
`e(ibaseout)`	base outcome number

Macros

`e(cmd)`	`discrim`
`e(subcmd)`	`logistic`
`e(cmdline)`	command as typed
`e(groupvar)`	name of group variable
`e(grouplabels)`	labels for the groups
`e(varlist)`	discriminating variables
`e(dropped)`	variables dropped because of collinearity
`e(wtype)`	weight type
`e(wexp)`	weight expression
`e(title)`	title in estimation output
`e(ties)`	how ties are to be handled
`e(properties)`	`b noV`
`e(estat_cmd)`	program used to implement `estat`
`e(predict)`	program used to implement `predict`
`e(marginsnotok)`	predictions disallowed by `margins`

Matrices

`e(b)`	coefficient vector
`e(groupcounts)`	number of observations for each group
`e(grouppriors)`	prior probabilities for each group
`e(groupvalues)`	numeric value for each group

Functions

`e(sample)`	marks estimation sample

Methods and formulas

Let g be the number of groups, n_i the number of observations for group i, and q_i the prior probability for group i. Let $\mathbf{x}$ denote an observation measured on p discriminating variables. For consistency with the discriminant analysis literature, $\mathbf{x}$ will be a column vector, though it corresponds to a row in your dataset. Let $f_i(\mathbf{x})$ represent the density function for group i, and let $P(\mathbf{x}|G_i)$ denote the probability of observing $\mathbf{x}$ conditional on belonging to group i. Denote the posterior probability of group i given observation $\mathbf{x}$ as $P(G_i|\mathbf{x})$. With Bayes's theorem, we have

$$P(G_i|\mathbf{x}) = \frac{q_i f_i(\mathbf{x})}{\sum_{j=1}^{g} q_j f_j(\mathbf{x})}$$

Substituting $P(\mathbf{x}|G_i)$ for $f_i(\mathbf{x})$, we have

$$P(G_i|\mathbf{x}) = \frac{q_i P(\mathbf{x}|G_i)}{\sum_{j=1}^{g} q_j P(\mathbf{x}|G_j)}$$

Dividing both the numerator and denominator by $P(\mathbf{x}|G_g)$, we can express this as

$$P(G_i|\mathbf{x}) = \frac{q_i L_{ig}(\mathbf{x})}{\sum_{j=1}^{g} q_j L_{jg}(\mathbf{x})}$$

where $L_{ig}(\mathbf{x}) = P(\mathbf{x}|G_i)/P(\mathbf{x}|G_g)$ is the likelihood ratio of $\mathbf{x}$ for groups i and g.

This formulation of the posterior probability allows easy insertion of the Multinomial logistic model into the discriminant analysis framework. The multinomial logistic model expresses $L_{ig}(\mathbf{x})$ in a simple exponential form

$$L_{ig}(\mathbf{x}) = \exp(a_{0i} + \mathbf{a}_i'\mathbf{x})$$

see Albert and Harris (1987, 117). Logistic discriminant analysis uses `mlogit` to compute the likelihood ratios, $L_{ig}(\mathbf{x})$, and hence the posterior probabilities $P(G_i|\mathbf{x})$; see [R] **mlogit**. However, `mlogit` and `predict` after `mlogit` assume proportional prior probabilities. `discrim logistic` assumes equal prior probabilities unless you specify the `priors(proportional)` option.

References

Albert, A., and E. K. Harris. 1987. *Multivariate Interpretation of Clinical Laboratory Data*. New York: Dekker.

Albert, A., and E. Lesaffre. 1986. Multiple group logistic discrimination. *Computers and Mathematics with Applications* 12A(2): 209–224.

Morrison, D. F. 2005. *Multivariate Statistical Methods*. 4th ed. Belmont, CA: Duxbury.

Rencher, A. C., and W. F. Christensen. 2012. *Methods of Multivariate Analysis*. 3rd ed. Hoboken, NJ: Wiley.

Also see

[MV] **discrim logistic postestimation** — Postestimation tools for discrim logistic

[MV] **discrim** — Discriminant analysis

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

[R] **mlogit** — Multinomial (polytomous) logistic regression

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