Title

discrim qda — Quadratic discriminant analysis

Syntax

`discrim qda` *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
`lootable`	display leave-one-out classification table

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**.

`fweight`s are allowed; see **[U] 11.1.6 weight**.

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

Menu

Statistics > Multivariate analysis > Discriminant analysis > Quadratic (QDA)

Description

discrim qda performs quadratic discriminant analysis. See [MV] **discrim** for other discrimination 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.

lootable displays the leave-one-out classification table.

Remarks and examples

stata.com

Quadratic discriminant analysis (QDA) was introduced by Smith (1947). It is a generalization of linear discriminant analysis (LDA). Both LDA and QDA assume that the observations come from a multivariate normal distribution. LDA assumes that the groups have equal covariance matrices. QDA removes this assumption, allowing the groups to have different covariance matrices.

One of the penalties associated with QDA's added flexibility is that if any groups have fewer observations, n_i, than discriminating variables, p, the covariance matrix for that group is singular and QDA cannot be performed. Even if there are enough observations to invert the covariance matrix, if the sample size is relatively small for a group, the estimation of the covariance matrix for that group may not do a good job of representing the group's population covariance, leading to inaccuracies in classification.

▷ Example 1: QDA classification tables and error rates

We illustrate QDA with a small dataset introduced in example 1 of [MV] **manova**. Andrews and Herzberg (1985, 357–360) present data on six apple tree rootstock groups with four measurements on eight trees from each group.

We request the display of the leave-one-out (LOO) classification table in addition to the standard resubstitution classification table produced by `discrim qda`.

```
. use http://www.stata-press.com/data/r13/rootstock
(Table 6.2 Rootstock Data, Rencher and Christensen (2012))
. discrim qda y1 y2 y3 y4, group(rootstock) lootable
Quadratic discriminant analysis
Resubstitution classification summary
```

Key
Number Percent

True rootstock	Classified 1	2	3	4	5	6	Total
1	8 100.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	8 100.00
2	0 0.00	7 87.50	0 0.00	1 12.50	0 0.00	0 0.00	8 100.00
3	1 12.50	0 0.00	6 75.00	0 0.00	1 12.50	0 0.00	8 100.00
4	0 0.00	0 0.00	1 12.50	7 87.50	0 0.00	0 0.00	8 100.00
5	0 0.00	3 37.50	0 0.00	0 0.00	4 50.00	1 12.50	8 100.00
6	2 25.00	0 0.00	0 0.00	0 0.00	1 12.50	5 62.50	8 100.00
Total	11 22.92	10 20.83	7 14.58	8 16.67	6 12.50	6 12.50	48 100.00
Priors	0.1667	0.1667	0.1667	0.1667	0.1667	0.1667	

```
Leave-one-out classification summary
```

Key
Number Percent

True rootstock	Classified 1	2	3	4	5	6	Total
1	2 25.00	0 0.00	0 0.00	3 37.50	1 12.50	2 25.00	8 100.00
2	0 0.00	3 37.50	0 0.00	2 25.00	2 25.00	1 12.50	8 100.00
3	1 12.50	2 25.00	4 50.00	0 0.00	1 12.50	0 0.00	8 100.00
4	1 12.50	1 12.50	3 37.50	2 25.00	0 0.00	1 12.50	8 100.00
5	0 0.00	4 50.00	1 12.50	0 0.00	2 25.00	1 12.50	8 100.00
6	3 37.50	1 12.50	0 0.00	0 0.00	2 25.00	2 25.00	8 100.00
Total	7 14.58	11 22.92	8 16.67	7 14.58	8 16.67	7 14.58	48 100.00
Priors	0.1667	0.1667	0.1667	0.1667	0.1667	0.1667	

Compare the counts on the diagonal of the resubstitution classification table with those on the LOO table. The LOO table has fewer of the observations with correct classifications. The resubstitution classification table is overly optimistic.

The `estat errorrate` postestimation command provides estimates of the error rates for the groups. We request the count-based estimates, first for the resubstitution classification and then for the LOO classification. We also suppress display of the prior probabilities, which will default to equal across the groups because that is how we estimated our QDA model. See [MV] **discrim estat** for details of the `estat errorrate` command.

```
. estat errorrate, nopriors
Error rate estimated by error count
```

	rootstock 1	2	3	4	5
Error rate	0	.125	.25	.125	.5

	rootstock 6	Total
Error rate	.375	.2291667

```
. estat errorrate, nopriors looclass
Error rate estimated by leave-one-out error count
```

	rootstock 1	2	3	4	5
Error rate	.75	.625	.5	.75	.75

	rootstock 6	Total
Error rate	.75	.6875

The estimated group error rates are much higher in the LOO table.

See example 2 of [MV] **discrim qda postestimation** for an examination of the squared Mahalanobis distances between the rootstock groups. We could also list the misclassified observations, produce group summaries, examine covariances and correlations, and generate classification and probability variables and more; see [MV] **discrim qda postestimation**.

◁

See example 3 of [MV] **discrim estat** and example 1 of [MV] **discrim qda postestimation** for other examples of the use of `discrim qda`.

Stored results

discrim qda stores the following in e():

Scalars	
e(N)	number of observations
e(N_groups)	number of groups
e(k)	number of discriminating variables
Macros	
e(cmd)	discrim
e(subcmd)	qda
e(cmdline)	command as typed
e(groupvar)	name of group variable
e(grouplabels)	labels for the groups
e(varlist)	discriminating variables
e(wtype)	weight type
e(wexp)	weight expression
e(title)	title in estimation output
e(ties)	how ties are to be handled
e(properties)	nob noV
e(estat_cmd)	program used to implement estat
e(predict)	program used to implement predict
e(marginsnotok)	predictions disallowed by margins
Matrices	
e(groupcounts)	number of observations for each group
e(grouppriors)	prior probabilities for each group
e(groupvalues)	numeric value for each group
e(means)	group means on discriminating variables
e(SSCP_W#)	within group SSCP matrix for group #
e(W#_eigvals)	eigenvalues of e(SSCP_W#)
e(W#_eigvecs)	eigenvectors of e(SSCP_W#)
e(sqrtS#inv)	Cholesky (square root) of the inverse covariance matrix for 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' 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

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

QDA assumes that the groups are multivariate normal. Let $\mathbf{S}_i$ denote the within-group sample covariance matrix for group i and $\bar{\mathbf{x}}_i$ denote the sample mean of group i. The squared Mahalanobis distance between observation $\mathbf{x}$ and $\bar{\mathbf{x}}_i$ is

$$D_i^2 = (\mathbf{x} - \bar{\mathbf{x}}_i)'\mathbf{S}_i^{-1}(\mathbf{x} - \bar{\mathbf{x}}_i)$$

Plugging these sample estimates into the multivariate normal density gives

$$P(\mathbf{x}|G_i) = (2\pi)^{-p/2}|\mathbf{S}_i|^{-1/2}e^{-D_i^2/2}$$

Substituting this into the formula for $P(G_i|\mathbf{x})$ and simplifying gives

$$P(G_i|\mathbf{x}) = \frac{q_i|\mathbf{S}_i|^{-1/2}e^{-D_i^2/2}}{\sum_{j=1}^{g} q_j|\mathbf{S}_j|^{-1/2}e^{-D_j^2/2}}$$

as the QDA posterior probability of observation $\mathbf{x}$ belonging to group i.

The squared Mahalanobis distance between group means is produced by `estat grdistances`; see [MV] **discrim qda postestimation**.

Classification functions can be derived from the Mahalanobis QDA; see Huberty (1994, 58). Let $Q_i(\mathbf{x})$ denote the quadratic classification function for the ith group applied to observation $\mathbf{x}$.

$$Q_i(\mathbf{x}) = -D_i^2/2 - \ln|\mathbf{S}_i|/2 + \ln(q_i)$$

An observation can be classified based on largest posterior probability or based on largest quadratic classification function score.

References

Andrews, D. F., and A. M. Herzberg, ed. 1985. *Data: A Collection of Problems from Many Fields for the Student and Research Worker.* New York: Springer.

Huberty, C. J. 1994. *Applied Discriminant Analysis.* New York: Wiley.

Rencher, A. C. 1998. *Multivariate Statistical Inference and Applications.* New York: Wiley.

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

Smith, C. A. B. 1947. Some examples of discrimination. *Annals of Eugenics* 13: 272–282.

Also see

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

[MV] **discrim** — Discriminant analysis

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