Title

stata.com

discrim — Discriminant analysis

Description

`discrim` performs discriminant analysis, which is also known as classification. kth-nearest-neighbor (KNN) discriminant analysis, linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and logistic discriminant analysis are available.

Syntax

`discrim` *subcommand* ... [, ...]

subcommand	Description
`knn`	kth-nearest-neighbor discriminant analysis
`lda`	linear discriminant analysis
`logistic`	logistic discriminant analysis
`qda`	quadratic discriminant analysis

Remarks and examples

stata.com

Remarks are presented under the following headings:

Introduction

Discriminant analysis is used to describe the differences between groups and to exploit those differences in allocating (classifying) observations of unknown group membership to the groups. Discriminant analysis is also called classification in many references. However, several sources use the word classification to mean cluster analysis.

Some applications of discriminant analysis include medical diagnosis, market research, classification of specimens in anthropology, predicting company failure or success, placement of students (workers) based on comparing pretest results to those of past students (workers), discrimination of natural versus man-made seismic activity, fingerprint analysis, image pattern recognition, and signal pattern classification.

Most multivariate statistics texts have chapters on discriminant analysis, including Rencher (1998), Rencher and Christensen (2012), Johnson and Wichern (2007), Mardia, Kent, and Bibby (1979), Anderson (2003), Everitt and Dunn (2001), Tabachnick and Fidell (2013), and Albert and Harris (1987). Books dedicated to discriminant analysis include Lachenbruch (1975), Klecka (1980), Hand (1981), Huberty (1994), McLachlan (2004), and Afifi, May, and Clark (2012). Of these, McLachlan (2004) gives the most extensive coverage, including 60 pages of references.

If you lack observations with known group membership, use cluster analysis to discover the natural groupings in the data; see [MV] **cluster**. If you have data with known group membership, possibly with other data of unknown membership to be classified, use discriminant analysis to examine the differences between the groups, based on data where membership is known, and to assign group membership for cases where membership is unknown.

Some researchers are not interested in classifying unknown observations and are interested only in the descriptive aspects of discriminant analysis. For others, the classification of unknown observations is the primary consideration. Huberty (1994), Rencher (1998), Rencher and Christensen (2012), and others split their discussion of discrimination into two parts. Huberty labels the two parts descriptive discriminant analysis and predictive discriminant analysis. Rencher and Christensen reserve discriminant analysis for descriptive discriminant analysis and uses the label classification for predictive discriminant analysis.

There are many discrimination methods. `discrim` has both descriptive and predictive LDA; see [MV] **discrim lda**. If your interest is in descriptive LDA, `candisc` computes the same thing as `discrim lda`, but with output tailored for the descriptive aspects of the discrimination; see [MV] **candisc**.

The remaining `discrim` subcommands provide alternatives to LDA for predictive discrimination. [MV] **discrim qda** provides quadratic discriminant analysis (QDA). [MV] **discrim logistic** provides logistic discriminant analysis. [MV] **discrim knn** provides kth-nearest-neighbor (KNN) discrimination.

The discriminant analysis literature uses conflicting terminology for several features of discriminant analysis. For example, in descriptive LDA, what one source calls a classification function another source calls a discriminant function while calling something else a classification function. Check the *Methods and formulas* sections for the `discrim` subcommands for clarification.

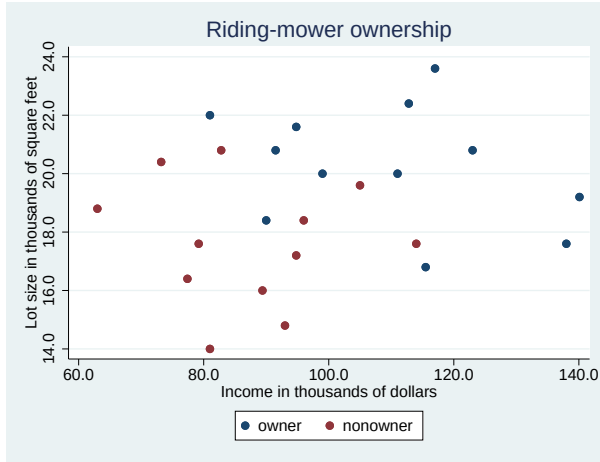
A simple example

We demonstrate the predictive and descriptive aspects of discriminant analysis with a simple example.

▷ Example 1: Discriminant analysis for prediction

Johnson and Wichern (2007, 578) introduce the concepts of discriminant analysis with a two-group dataset. A sample of 12 riding-lawnmower owners and 12 nonowners is sampled from a city and the income in thousands of dollars and lot size in thousands of square feet are recorded. A riding-mower manufacturer wants to see if these two variables adequately separate owners from nonowners, and if so to then direct their marketing on the basis of the separation of owners from nonowners.

```
. use https://www.stata-press.com/data/r17/lawnmower2
(Johnson and Wichern (2007) table 11.1)
```



Do these two variables adequately separate riding-mower owners from nonowners so that the riding-mower manufacturer can base predictions of riding-mower ownership on income and lot size? The graph shows some separation of owners from nonowners, but with overlap. With predictive LDA we can quantify our ability to discriminate between riding-mower owners and nonowners.

```
. discrim lda lotsize income, group(owner)

Linear discriminant analysis
Resubstitution classification summary
```

Key
Number Percent

True owner	Classified Nonowner	Owner	Total
Nonowner	10	2	12
	83.33	16.67	100.00
Owner	1	11	12
	8.33	91.67	100.00
Total	11	13	24
	45.83	54.17	100.00
Priors	0.5000	0.5000	

The table presented by `discrim lda` (and the other `discrim` subcommands) is called a classification table or confusion matrix. It is labeled as a resubstitution classification table because the same observations used in estimating the discriminant model were classified using the model. The diagonal elements in the main body of the table show the number and percent correctly classified into each group. The off-diagonal elements show the misclassified number and percent. One owner and two nonowners were misclassified.

The resubstitution classification table provides an overly optimistic assessment of how well the linear discriminant function will predict the ownership status for observations that were not part of the training sample. A leave-one-out classification table provides a more realistic assessment for future prediction. The leave-one-out classification is produced by holding each observation out, one at a

time; building an LDA model from the remaining training observations; and then classifying the held out observation using this model. The leave-one-out classification table is available at estimation time, at playback, or through the `estat classtable` postestimation command.

```
. estat classtable, loo nopriors
Leave-one-out classification table
```

Key
Number Percent

True owner	LOO Classified Nonowner	Owner	Total
Nonowner	9 75.00	3 25.00	12 100.00
Owner	2 16.67	10 83.33	12 100.00
Total	11 45.83	13 54.17	24 100.00

With leave-one-out classification we see that 5, instead of only 3, of the 24 observations are misclassified.

The `predict` and `estat` commands provide other predictive discriminant analysis tools. `predict` generates variables containing the posterior probabilities of group membership or generates a group membership classification variable. `estat` displays classification tables, displays error-rate tables, and lists classifications and probabilities for the observations.

We now use `estat list` to show the resubstitution and leave-one-out classifications and posterior probabilities for those observations that were misclassified by our LDA model.

```
. estat list, class(loo) probabilities(loo) misclassified
```

Obs	Classification True	Class.	LOO Cl.	Probabilities Nonowner	Owner	LOO Probabilities Nonowner	Owner
1	Owner	Nonown *	Nonown *	0.7820	0.2180	0.8460	0.1540
2	Owner	Owner	Nonown *	0.4945	0.5055	0.6177	0.3823
13	Nonown	Owner *	Owner *	0.2372	0.7628	0.1761	0.8239
14	Nonown	Nonown	Owner *	0.5287	0.4713	0.4313	0.5687
17	Nonown	Owner *	Owner *	0.3776	0.6224	0.2791	0.7209

```
* indicates misclassified observations
```

◁

We have used `discrim lda` to illustrate predictive discriminant analysis. The other `discrim` subcommands could also be used for predictive discrimination of these data.

Postestimation commands after `discrim lda` provide descriptive discriminant analysis; see [MV] **discrim lda postestimation** and [MV] **candisc**.

Example 2: Discriminant analysis for description

The riding-mower manufacturer of the previous example wants to understand how income and lot size affect riding-mower ownership. Descriptive discriminant analysis provides tools for exploring how the groups are separated. Fisher's (1936) linear discriminant functions provide the basis for descriptive LDA; see [MV] **discrim lda** and [MV] **discrim lda postestimation**. The postestimation command `estat loadings` allows us to view the discriminant function coefficients, which are also called loadings.

```
. estat loadings, standardized unstandardized
Canonical discriminant function coefficients
                │  function1
────────────────┼────────────
       lotsize  │   .3795228
        income  │   .0484468
         _cons  │  -11.96094
Standardized canonical discriminant function coefficients
                │  function1
────────────────┼────────────
       lotsize  │   .7845512
        income  │   .8058419
```

We requested both the unstandardized and standardized coefficients. The unstandardized coefficients apply to unstandardized variables. The standardized coefficients apply to variables standardized using the pooled within-group covariance. Standardized coefficients are examined to assess the relative importance of the variables to the discriminant function.

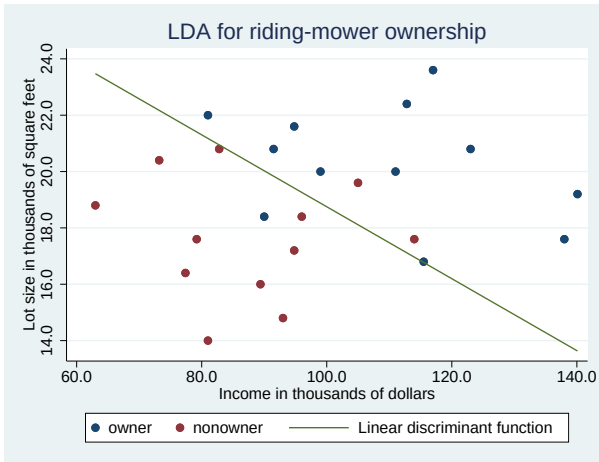
The unstandardized coefficients determine the separating line between riding-mower owners and nonowners.

$$0 = 0.3795228\ \texttt{lotsize} + 0.0484468\ \texttt{income} - 11.96094$$

which can be reexpressed as

$$\texttt{lotsize} = -0.1276519\ \texttt{income} + 31.51574$$

We now display this line superimposed on the scatterplot of the data.



Other descriptive statistics and summaries are available; see [MV] **discrim lda postestimation**.

Prior probabilities, costs, and ties

Classification is influenced by the selection of prior probabilities, assignment of costs to misclassification, and the method of handling ties in classification criteria.

Prior probabilities are the presumptive or a priori probabilities of group membership. Before you flip a balanced coin 10 times, you know the prior probability of getting heads is the same as getting tails—both are 0.5. Group prior probabilities, commonly called priors, must be taken into account in calculations of posterior probabilities; see *Methods and formulas* for details.

If the cost of misclassification is not equal over the groups, an optimal classification into groups must take misclassification cost into account. When there are two groups, members of the first group can be classified into the second, or members of the second group can be classified into the first. The relative undesirability of these two misclassifications may not be the same. Example 3 of [MV] **discrim knn** classifies poisonous and edible mushrooms. Misclassifying poisonous mushrooms as edible is a big deal at dinnertime.

The expected misclassification cost is the sum of the products of the cost for each misclassification multiplied by the probability of its occurrence. Let p_{ij} be the probability that an observation from group i is classified into group j, let c_{ij} be the cost of misclassifying an observation from group i into group j, and let q_i be the prior probability that the observation is from group i. The expected cost of misclassification is then

$$\text{cost} = \sum_{i,j\neq i}^{g} c_{ij}p_{ij}q_i$$

It is this expected cost that we wish to minimize. In the two-group case

$$\text{cost} = c_{12}p_{12}q_1 + c_{21}p_{21}q_2$$

and we can use cost-adjusted group prior probabilities, $\widehat{q}_i$, in the place of the prior probabilities to minimize the cost of misclassification.

$$\widehat{q}_1 = \frac{c_{12}q_1}{c_{12}q_1 + c_{21}q_2}$$

$$\widehat{q}_2 = \frac{c_{21}q_2}{c_{12}q_1 + c_{21}q_2}$$

With more than two groups, there is often not a simple rule to take costs into account. More discussion on this topic is provided by McLachlan (2004, 7–9), Huberty (1994, 68–69), Johnson and Wichern (2007, 606–609), and Anderson (2003, chap. 6).

See example 3 of [MV] **discrim knn** for an application of costs.

A tie in classification occurs when two or more group posterior probabilities are equal for an observation. Ties are most common with kth-nearest-neighbor discriminant analysis, though they can occur in other forms of discriminant analysis. There are several options for assigning tied observations. The default is to mark the observation as unclassified, that is, classified to a missing value. Ties can also be broken. For most forms of discriminant analysis ties can be broken in two ways—randomly or assigned to the first group that is tied. For kth-nearest-neighbor discriminant analysis, dissimilarities are calculated, and so ties may also be broken by choosing the group of the nearest of the tied observations. If this still results in a tie, the observation is unclassified.

Methods and formulas

See [MV] **discrim lda** for the methods and formulas for descriptive discriminant analysis.

For predictive discriminant analysis, 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

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

An observation is classified as belonging to the group with the highest posterior probability.

The difference between the `discrim` subcommands is in the choice of $f_i(\mathbf{x})$. LDA, `discrim lda`, assumes that the groups are multivariate normal with equal covariance matrices; see [MV] **discrim lda**. QDA, `discrim qda`, assumes that the groups are multivariate normal, allowing the groups to have unequal covariance matrices; see [MV] **discrim qda**. Logistic discriminant analysis, `discrim logistic`, uses the multinomial logistic model to obtain the posterior probabilities; see [MV] **discrim logistic**. kth-nearest neighbor, `discrim knn`, uses a simple nonparametric estimate of $f_i(\mathbf{x})$, based on examination of the k closest observations; see [MV] **discrim knn**.

References

Afifi, A. A., S. May, and V. A. Clark. 2012. *Practical Multivariate Analysis*. 5th ed. Boca Raton, FL: CRC Press.

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

Anderson, T. W. 2003. *An Introduction to Multivariate Statistical Analysis*. 3rd ed. New York: Wiley.

Everitt, B. S., and G. Dunn. 2001. *Applied Multivariate Data Analysis*. 2nd ed. London: Arnold.

Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7: 179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x.

Hand, D. J. 1981. *Discrimination and Classification*. New York: Wiley.

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

Johnson, R. A., and D. W. Wichern. 2007. *Applied Multivariate Statistical Analysis*. 6th ed. Englewood Cliffs, NJ: Prentice Hall.

Klecka, W. R. 1980. *Discriminant Analysis*. Newbury Park, CA: SAGE.

Lachenbruch, P. A. 1975. *Discriminant Analysis*. New York: Hafner Press.

Mardia, K. V., J. T. Kent, and J. M. Bibby. 1979. *Multivariate Analysis*. London: Academic Press.

McLachlan, G. J. 2004. *Discriminant Analysis and Statistical Pattern Recognition*. 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.

Tabachnick, B. G., and L. S. Fidell. 2013. *Using Multivariate Statistics*. 6th ed. Boston: Pearson.

Also see

[MV] **discrim estat** — Postestimation tools for discrim

[MV] **candisc** — Canonical linear discriminant analysis

[MV] **cluster** — Introduction to cluster-analysis commands

[U] 20 Estimation and postestimation commands