Title

stata.com

candisc — Canonical linear discriminant analysis

Description

`candisc` performs canonical linear discriminant analysis (LDA). What is computed is the same as with [MV] **discrim lda**. The difference is in what is presented. See [MV] **discrim** for other discrimination commands.

Quick start

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

```
candisc v1 v2 v3 v4, group(catvar)
```

As above, but use prior probabilities proportional to group size

```
candisc v1 v2 v3 v4, group(catvar) priors(proportional)
```

Present the leave-one-out classification table in addition to standard output

```
candisc v1 v2 v3 v4, group(catvar) lootable
```

As above, but suppress the resubstitution classification table

```
candisc v1 v2 v3 v4, group(catvar) lootable notable
```

Menu

Statistics > Multivariate analysis > Discriminant analysis > Canonical linear discriminant analysis

Syntax

```
candisc 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
nostats	suppress display of canonical statistics
nocoef	suppress display of standardized canonical discriminant function coefficients
nostruct	suppress display of canonical structure matrix
nomeans	suppress display of group means on canonical variables

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.

lootable displays the leave-one-out classification table.

nostats suppresses the display of the table of canonical statistics.

nocoef suppresses the display of the standardized canonical discriminant function coefficients.

nostruct suppresses the display of the canonical structure matrix.

nomeans suppresses the display of group means on canonical variables.

Remarks and examples

stata.com

See [MV] **discrim** for background on discriminant analysis (classification) and see [MV] **discrim lda** for more information on linear discriminant analysis. What candisc displays by default with

```
. candisc x y z, group(group)
```

you can also obtain with the following sequence of discrim commands and estat postestimation commands.

```
. discrim x y z, group(group) notable
. estat canontest
. estat loadings
. estat structure
. estat grmeans, canonical
. estat classtable
```

The candisc command will appeal to those performing descriptive LDA.

▷ Example 1

Example 2 of [MV] **discrim knn** introduces a head-measurement dataset from Rencher and Christensen (2012, 291) that has six discriminating variables and three groups. The three groups are high school football players, college football players, and nonplayers. The data were collected as a preliminary step in determining the relationship between helmet design and neck injuries.

Descriptive discriminant analysis allows us to explore the relationship in this dataset between head measurements and the separability of the three groups.

```
. use http://www.stata-press.com/data/r15/head
(Table 8.3 Head measurements, Rencher and Christensen (2012))
. candisc wdim circum fbeye eyehd earhd jaw, group(group)
Canonical linear discriminant analysis
```

Fcn	Canon. Corr.	Eigen-value	Variance Prop.	Variance Cumul.	Like-lihood Ratio	F	df1	df2	Prob>F	
1	0.8107	1.91776	0.9430	0.9430	0.3071	10.994	12	164	0.0000	e
2	0.3223	.115931	0.0570	1.0000	0.8961	1.9245	5	83	0.0989	e

Ho: this and smaller canon. corr. are zero; e = exact F

Standardized canonical discriminant function coefficients

	function1	function2
wdim	.6206412	.9205834
circum	-.0064715	-.0009114
fbeye	-.0047581	-.021145
eyehd	-.7188123	.5997882
earhd	-.3965116	-.3018196
jaw	-.5077218	-.9368745

Canonical structure

	function1	function2
wdim	.1482946	.3766581
circum	-.2714134	.1305383
fbeye	-.1405813	-.061071
eyehd	-.824502	.5363578
earhd	-.5177312	.1146999
jaw	-.2119042	-.3895934

Group means on canonical variables

group	function1	function2
high school	-1.910378	-.0592794
college	1.16399	-.3771343
nonplayer	.7463888	.4364137

Resubstitution classification summary

Key
Number Percent

True group	Classified high school	college	nonplayer	Total
high school	26	1	3	30
	86.67	3.33	10.00	100.00
college	1	20	9	30
	3.33	66.67	30.00	100.00
nonplayer	2	8	20	30
	6.67	26.67	66.67	100.00
Total	29	29	32	90
	32.22	32.22	35.56	100.00
Priors	0.3333	0.3333	0.3333	

As seen in the canonical correlation table, the first linear discriminant function accounts for almost 95% of the variance. The standardized discriminant function coefficients (loadings) indicate that two of the variables, `circum` (head circumference) and `fbeye` (front-to-back measurement at eye level), have little discriminating ability for these three groups. The first discriminant function is contrasting `wdim` (head width at widest dimension) to a combination of `eyehd` (eye-to-top-of-head measurement), `earhd` (ear-to-top-of-head measurement), and `jaw` (jaw width).

The canonical structure coefficients, which measure the correlation between the discriminating variables and the discriminant function, are also shown. There is controversy on whether the standardized loadings or the structure coefficients should be used for interpretation; see Rencher and Christensen (2012, 301) and Huberty (1994, 262–264).

The group means on the canonical variables are shown, giving some indication of how the groups are separated. The means on the first function show the high school group separated farthest from the other two groups.

The resubstitution classification table, also known as a confusion matrix, indicates how many observations from each group are classified correctly or misclassified into the other groups. The college and nonplayer groups appear to have more misclassifications between them, indicating that these two groups are harder to separate.

All the postestimation tools of `discrim lda` are available after `candisc`; see [MV] **discrim lda postestimation**. For example, `estat grsummarize` can produce discriminating-variable summaries for each of our three groups.

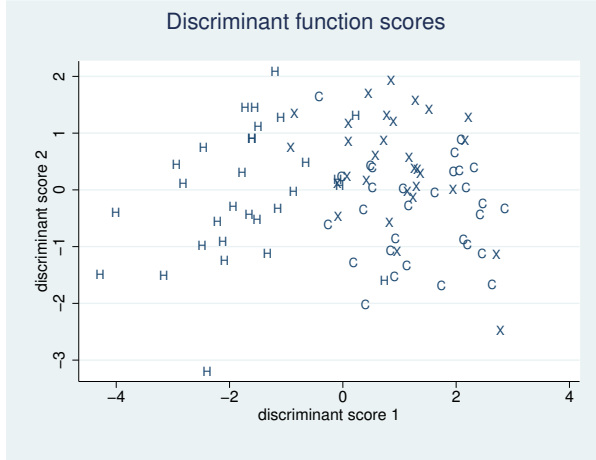
```
. estat grsummarize
Estimation sample candisc
Summarized by group
```

Mean	group high school	college	nonplayer	Total
wdim	15.2	15.42	15.58	15.4
circum	58.937	57.37967	57.77	58.02889
fbeye	20.10833	19.80333	19.81	19.90722
eyehd	13.08333	10.08	10.94667	11.37
earhd	14.73333	13.45333	13.69667	13.96111
jaw	12.26667	11.94333	11.80333	12.00444
N	30	30	30	90

A score plot graphs observation scores from the first two discriminant functions; see [MV] **scoreplot**. After `candisc`, `scoreplot` automatically labels the points with the value labels assigned to the groups. The value labels for our three groups are long—the resulting graph is too crowded.

To overcome this, we create a new label language (see [D] **label language**), define one letter labels for the groups, assign this label to our group variable, and then call `scoreplot`. We then reset the label language back to the default containing the longer, more descriptive value labels.

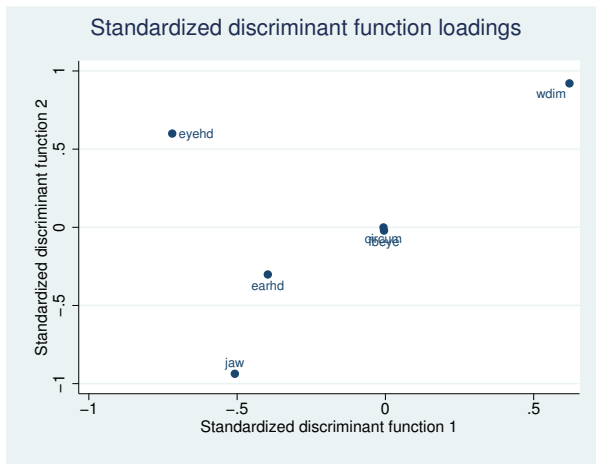
```
. label language short, new
(language short now current language)
. label define fball 1 "H" 2 "C" 3 "X"
. label values group fball
. scoreplot, msymbol(i) aspect(.625)
. label language default
```



The score plot illustrates the separation due to the first and second canonical linear discriminant functions. As expected from our examination of the earlier descriptive output, the high school group (labeled `H`) is reasonably well separated from the college (labeled `C`) and nonplayer (labeled `X`) groups. There is some separation in the second dimension between the college and nonplayer groups, but with substantial overlap.

A loading plot provides a graphical way of looking at the standardized discriminant function coefficients (loadings) that we previously examined in tabular form.

```
. loadingplot
```



`circum` and `fbeye` are near the origin, indicating that they provide almost no discriminating ability in comparison to the other discriminating variables. The relative locations of the remaining variables indicate their contribution to the discriminant functions.

Stored results

candisc stores the following in e():

Scalars

e(N)	number of observations
e(N_groups)	number of groups
e(k)	number of discriminating variables
e(f)	number of nonzero eigenvalues

Macros

e(cmd)	candisc
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 eigen
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)	pooled within-group SSCP matrix
e(SSCP_B)	between-groups SSCP matrix
e(SSCP_T)	total SSCP matrix
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(S)	pooled within-group covariance matrix
e(Sinv)	inverse of e(S)
e(sqrtSinv)	Cholesky (square root) of e(Sinv)
e(Ev)	eigenvalues of $\mathbf{W}^{-1}\mathbf{B}$
e(L_raw)	eigenvectors of $\mathbf{W}^{-1}\mathbf{B}$
e(L_unstd)	unstandardized canonical discriminant function coefficients
e(L_std)	within-group standardized canonical discriminant function coefficients
e(L_totalstd)	total-sample standardized canonical discriminant function coefficients
e(C)	classification coefficients
e(cmeans)	unstandardized canonical discriminant functions evaluated at group means
e(canstruct)	canonical structure matrix
e(candisc_stat)	canonical discriminant analysis statistics

Functions

e(sample)	marks estimation sample

Methods and formulas

See *Methods and formulas* in [MV] **discrim lda** for information.

References

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

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

Also see

[MV] **discrim lda** — Linear discriminant analysis