Title

stata.com

discrim estat — Postestimation tools for discrim

Postestimation commands

The following postestimation commands are of special interest after `candisc`, `discrim knn`, `discrim lda`, `discrim logistic`, and `discrim qda`:

Command	Description
`estat classtable`	classification table
`estat errorrate`	classification error-rate estimation
`estat grsummarize`	group summaries
`estat list`	classification listing
`estat summarize`	estimation sample summary

There are more postestimation commands of special interest after `discrim lda` and `discrim qda`; see [MV] **discrim lda postestimation** and [MV] **discrim qda postestimation**.

Description for estat

`estat classtable` displays a cross-tabulation of the original groups with the classification groups. Classification percentages, average posterior probabilities, group prior probabilities, totals, and leave-one-out results are available.

`estat errorrate` displays error-rate estimates for the classification. Count-based estimates and both stratified and unstratified posterior-probability-based estimates of the error rate are available. These estimates can be resubstitution or leave-one-out estimates.

`estat grsummarize` presents estimation sample summary statistics for the discriminating variables for each group defined by the grouping variable. Means, medians, minimums, maximums, standard deviations, coefficients of variation, standard errors of the means, and group sizes may be displayed. Overall sample statistics are also available.

`estat list` lists group membership, classification, and probabilities for observations.

`estat summarize` summarizes the variables in the discriminant analysis over the estimation sample.

Quick start for estat

Classification table

Classification table computed using proportional prior probabilities

```
estat classtable, priors(proportional)
```

Leave-one-out classification table showing average posterior probability of classification into each group

```
estat classtable, looclass probabilities
```

Classification error-rate estimation

Error-rate table estimated from leave-one-out error count

```
estat errorrate, looclass
```

Error rates estimated from posterior probabilities using proportional prior probabilities

```
estat errorrate, pp priors(proportional)
```

Group summaries

Summary statistics by group

```
estat grsummarize
```

Mean, median, standard deviation, minimum, and maximum by group

```
estat grsummarize, mean median sd min max
```

Classification listing

Listing of group membership, classification, and probabilities including the leave-one-out results

```
estat list, classification(looclass) probabilities(loopr)
```

As above, but suppress the resubstitution classification

```
estat list, classification(looclass noclass)   ///
     probabilities(loopr nopr)
```

Estimation sample summary

Summary of variables from the most recent discriminant analysis and displaying variable labels

```
estat summarize, labels
```

Menu for estat

Statistics > Postestimation

Syntax for estat

Classification table

```
estat classtable [if] [in] [weight] [, classtable_options]
```

Classification error-rate estimation

```
estat errorrate [if] [in] [weight] [, errorrate_options]
```

Group summaries

```
estat grsummarize [, grsummarize_options]
```

Classification listing

```
estat list [if] [in] [, list_options]
```

Estimation sample summary

```
estat summarize [, labels noheader noweights]
```

classtable_options	Description
Main	
`class`	display the classification table; the default
`looclass`	display the leave-one-out classification table
Options	
`priors(`*priors*`)`	group prior probabilities; defaults to `e(grouppriors)`
`nopriors`	suppress display of prior probabilities
`ties(`*ties*`)`	how ties in classification are to be handled; defaults to `e(ties)`
`title(`*text*`)`	title for classification table
`probabilities`	display the average posterior probability of being classified into each group
`nopercents`	suppress display of percentages
`nototals`	suppress display of row and column totals
`norowtotals`	suppress display of row totals
`nocoltotals`	suppress display of column totals

priors	Description
`equal`	equal prior probabilities
`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
random	ties in group classification are broken randomly
first	ties in group classification are set to the first tied group
nearest	ties in group classification are assigned based on the closest observation, or missing if this still results in a tie; after `discrim knn` only

errorrate_options	Description
Main	
class	display the classification-based error-rate estimates table; the default
looclass	display the leave-one-out classification-based error-rate estimates table
count	use a count-based error-rate estimate
pp[(*ppopts*)]	use a posterior-probability-based error-rate estimate
Options	
priors(*priors*)	group prior probabilities; defaults to `e(grouppriors)`
nopriors	suppress display of prior probabilities
ties(*ties*)	how ties in classification are to be handled; defaults to `e(ties)`
title(*text*)	title for error-rate estimate table
nototal	suppress display of total column

ppopts	Description
stratified	present stratified results
unstratified	present unstratified results

grsummarize_options	Description
Main	
n[(%*fmt*)]	group sizes
mean[(%*fmt*)]	means
median[(%*fmt*)]	medians
sd[(%*fmt*)]	standard deviations
cv[(%*fmt*)]	coefficients of variation
semean[(%*fmt*)]	standard errors of the means
min[(%*fmt*)]	minimums
max[(%*fmt*)]	maximums
Options	
nototal	suppress overall statistics
transpose	display groups by row instead of column

list_options	Description
Main	
misclassified	list only misclassified and unclassified observations
classification(*clopts*)	control display of classification
probabilities(*propts*)	control display of probabilities
varlist[(*varopts*)]	display discriminating variables
[no]obs	display or suppress the observation number
id(*varname* [format(%*fmt*)])	display identification variable
Options	
weight[(*weightopts*)]	display frequency weights
priors(*priors*)	group prior probabilities; defaults to e(grouppriors)
ties(*ties*)	how ties in classification are to be handled; defaults to e(ties)
separator(#)	display a horizontal separator every # lines

clopts	Description
noclass	do not display the standard classification
looclass	display the leave-one-out classification
notrue	do not show the group variable
nostar	do not display stars indicating misclassified observations
nolabel	suppress display of value labels for the group and classification variables
format(%*fmt*)	format for group and classification variables; default is %5.0f for unlabeled numeric variables

propts	Description
nopr	suppress display of standard posterior probabilities
loopr	display leave-one-out posterior probabilities
format(%*fmt*)	format for probabilities; default is format(%7.4f)

varopts	Description
first	display input variables before classifications and probabilities
last	display input variables after classifications and probabilities
format(%*fmt*)	format for input variables; default is the input variable format

weightopts	Description
none	do not display the weights
format(%*fmt*)	format for the weight; default is %3.0f for weights $<$ 1,000, %5.0f for 1,000 $<$ weights $<$ 100,000, and %8.0g otherwise

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

Options for estat

Options are presented under the following headings:

Options for estat classtable

Main

`class`, the default, displays the classification table. With in-sample observations, this is called the resubstitution classification table.

`looclass` displays a leave-one-out classification table, instead of the default classification table. Leave-one-out classification applies only to the estimation sample, and so, in addition to restricting the observations to those chosen with `if` and `in` qualifiers, the observations are further restricted to those included in `e(sample)`.

Options

`priors(`*priors*`)` specifies the prior probabilities for group membership. If `priors()` is not specified, `e(grouppriors)` is used. If `nopriors` is specified with `priors()`, prior probabilities are used for calculation of the classification variable but not displayed. The following *priors* are allowed:

`priors(equal)` specifies equal prior probabilities.

`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.

`nopriors` suppresses display of the prior probabilities. This option does not change the computations that rely on the prior probabilities specified in `priors()` or as found by default in `e(grouppriors)`.

`ties(`*ties*`)` specifies how ties in group classification will be handled. If `ties()` is not specified, `e(ties)` determines how ties are handled. The following *ties* are allowed:

`ties(missing)` specifies that ties in group classification produce missing values.

`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.

`ties(nearest)` specifies that ties in group classification are assigned based on the closest observation, or missing if this still results in a tie. `ties(nearest)` is available after `discrim knn` only.

`title(`*text*`)` customizes the title for the classification table.

`probabilities` specifies that the classification table show the average posterior probability of being classified into each group. `probabilities` implies `norowtotals` and `nopercents`.

`nopercents` specifies that percentages are to be omitted from the classification table.

`nototals` specifies that row and column totals are to be omitted from the classification table.

`norowtotals` specifies that row totals are to be omitted from the classification table.

`nocoltotals` specifies that column totals are to be omitted from the classification table.

Options for estat errorrate

Main

`class`, the default, specifies that the classification-based error-rate estimates table be presented. The alternative to `class` is `looclass`.

`looclass` specifies that the leave-one-out classification error-rate estimates table be presented.

`count`, the default, specifies that the error-rate estimates be based on misclassification counts. The alternative to `count` is `pp()`.

`pp`[`(`*ppopts*`)`] specifies that the error-rate estimates be based on posterior probabilities. `pp` is equivalent to `pp(stratified unstratified)`. `stratified` indicates that stratified estimates be presented. `unstratified` indicates that unstratified estimates be presented. One or both may be specified.

Options

`priors(`*priors*`)` specifies the prior probabilities for group membership. If `priors()` is not specified, `e(grouppriors)` is used. If `nopriors` is specified with `priors()`, prior probabilities are used for calculation of the error-rate estimates but not displayed. The following *priors* are allowed:

`priors(equal)` specifies equal prior probabilities.

`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.

`nopriors` suppresses display of the prior probabilities. This option does not change the computations that rely on the prior probabilities specified in `priors()` or as found by default in `e(grouppriors)`.

`ties(`*ties*`)` specifies how ties in group classification will be handled. If `ties()` is not specified, `e(ties)` determines how ties are handled. The following *ties* are allowed:

`ties(missing)` specifies that ties in group classification produce missing values.

`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.

`ties(nearest)` specifies that ties in group classification are assigned based on the closest observation, or missing if this still results in a tie. `ties(nearest)` is available after `discrim knn` only.

`title(`*text*`)` customizes the title for the error-rate estimates table.

`nototal` suppresses the total column containing overall sample error-rate estimates.

Options for estat grsummarize

Main

`n`[`(`%*fmt*`)`] specifies that group sizes be presented. The optional argument provides a display format. The default options are `n` and `mean`.

`mean`[`(`%*fmt*`)`] specifies that means be presented. The optional argument provides a display format. The default options are `n` and `mean`.

`median`[`(`%*fmt*`)`] specifies that medians be presented. The optional argument provides a display format.

`sd`[(%*fmt*)] specifies that standard deviations be presented. The optional argument provides a display format.

`cv`[(%*fmt*)] specifies that coefficients of variation be presented. The optional argument provides a display format.

`semean`[(%*fmt*)] specifies that standard errors of the means be presented. The optional argument provides a display format.

`min`[(%*fmt*)] specifies that minimums be presented. The optional argument provides a display format.

`max`[(%*fmt*)] specifies that maximums be presented. The optional argument provides a display format.

Options

`nototal` suppresses display of the total column containing overall sample statistics.

`transpose` specifies that the groups are to be displayed by row. By default, groups are displayed by column. If you have more variables than groups, you might prefer the output produced by `transpose`.

Options for estat list

Main

`misclassified` lists only misclassified and unclassified observations.

`classification(`*clopts*`)` controls display of the group variable and classification. By default, the standard classification is calculated and displayed along with the group variable in `e(groupvar)`, using labels from the group variable if they exist. *clopts* may be one or more of the following:

`noclass` suppresses display of the standard classification. If the observations are those used in the estimation, classification is called resubstitution classification.

`looclass` specifies that the leave-one-out classification be calculated and displayed. The default is that the leave-one-out classification is not calculated. `looclass` is not allowed after `discrim logistic`.

`notrue` suppresses the display of the group variable. By default, `e(groupvar)` is displayed. `notrue` implies `nostar`.

`nostar` suppresses the display of stars indicating misclassified observations. A star is displayed by default when the classification is not in agreement with the group variable. `nostar` is the default when `notrue` is specified.

`nolabel` specifies that value labels for the group variable, if they exist, not be displayed for the group or classification or used as labels for the probability column names.

`format(`%*fmt*`)` specifies the format for the group and classification variables. If value labels are used, string formats are permitted.

`probabilities(`*propts*`)` controls the display of group posterior probabilities. *propts* may be one or more of the following:

`nopr` suppresses display of the standard posterior probabilities. By default, the posterior probabilities are shown.

`loopr` specifies that leave-one-out posterior probabilities be displayed. `loopr` is not allowed after `discrim logistic`.

`format(`%*fmt*`)` specifies the format for displaying probabilities. The default is `format(%7.4f)`.

`varlist`[`(`*varopts*`)`] specifies that the discriminating variables found in `e(varlist)` be displayed and specifies the display options for the variables.

`none` specifies that discriminating variables are not to be displayed. This is the default.

`first` specifies variables be displayed before classifications and probabilities.

`last` specifies variables be displayed after classifications and probabilities.

`format(`*%fmt*`)` specifies the format for the input variables. By default, the variable's format is used.

[`no`]`obs` indicates that observation numbers be or not be displayed. Observation numbers are displayed by default unless `id()` is specified.

`id(`*varname* [`format(`*%fmt*`)`]`)` specifies the identification variable to display and, optionally, the format for that variable. By default, the format of *varname* is used.

Options

`weight`[`(`*weightopts*`)`] specifies options for displaying weights. By default, if `e(wexp)` exists, weights are displayed.

`none` specifies weights not be displayed. This is the default if weights were not used with `discrim`.

`format(`*%fmt*`)` specifies a display format for the weights. If the weights are $<$ 1,000, `%3.0f` is the default, `%5.0f` is the default if 1,000 $<$ weights $<$ 100,000, else `%8.0g` is used.

`priors(`*priors*`)` specifies the prior probabilities for group membership. If `priors()` is not specified, `e(grouppriors)` is used. The following *priors* are allowed:

`priors(equal)` specifies equal prior probabilities.

`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. If `ties()` is not specified, `e(ties)` determines how ties are handled. The following *ties* are allowed:

`ties(missing)` specifies that ties in group classification produce missing values.

`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.

`ties(nearest)` specifies that ties in group classification are assigned based on the closest observation, or missing if this still results in a tie. `ties(nearest)` is available after `discrim knn` only.

`separator(#)` specifies a horizontal separator line be drawn every # observations. The default is `separator(5)`.

Options for estat summarize

`labels`, `noheader`, and `noweights` are the same as for the generic `estat summarize`; see [R] **estat summarize**.

Remarks and examples

stata.com

Remarks are presented under the following headings:

There are several `estat` commands that apply after all the `discrim` subcommands. `estat summarize` and `estat grsummarize` summarize the discriminating variables over the estimation sample and by-group. `estat list` displays classifications, posterior probabilities, and more for selected observations. `estat classtable` and `estat errorrate` display the classification table, also known as a confusion matrix, and error-rate estimates based on the classification table.

Discriminating-variable summaries

`estat summarize` and `estat grsummarize` provide summaries of the variables involved in the preceding discriminant analysis model.

▷ Example 1

Example 3 of [MV] **discrim lda** introduces the famous iris data originally from Anderson (1935) and used by Fisher (1936) in the development of linear discriminant analysis. We continue our exploration of the linear discriminant analysis of the iris data and demonstrate the summary `estat` tools available after all `discrim` subcommands.

```
. use http://www.stata-press.com/data/r14/iris
(Iris data)
. discrim lda seplen sepwid petlen petwid, group(iris) notable
```

The `notable` option of `discrim` suppressed display of the classification table. We explore the use of `estat classtable` later.

What can we learn about the underlying discriminating variables? `estat summarize` gives a summary of the variables involved in the discriminant analysis, restricted to the estimation sample.

```
. estat summarize
Estimation sample discrim                 Number of obs =       150
```

Variable	Mean	Std. Dev.	Min	Max
groupvar				
iris	2	.8192319	1	3
variables				
seplen	5.843333	.8280661	4.3	7.9
sepwid	3.057333	.4358663	2	4.4
petlen	3.758	1.765298	1	6.9
petwid	1.199333	.7622377	.1	2.5

`estat summarize` displays the mean, standard deviation, minimum, and maximum for the group variable, `iris`, and the four discriminating variables, `seplen`, `sepwid`, `petlen`, and `petwid`. Also shown is the number of observations. If we had fit our discriminant model on a subset of the data, `estat summarize` would have restricted its summary to those observations.

More interesting than an overall summary of the discriminating variables is a summary by our group variable, `iris`.

```
. estat grsummarize
Estimation sample discrim lda
Summarized by iris
```

Mean	iris setosa	versicolor	virginica	Total
seplen	5.006	5.936	6.588	5.843333
sepwid	3.428	2.77	2.974	3.057333
petlen	1.462	4.26	5.552	3.758
petwid	.246	1.326	2.026	1.199333
N	50	50	50	150

By default, `estat grsummarize` displays means of the discriminating variables for each group and overall (the total column), along with group sizes. The summary is restricted to the estimation sample.

The petal length and width of *Iris setosa* appear to be much smaller than those of the other two species. *Iris versicolor* has petal length and width between that of the other two species.

Other statistics may be requested. A look at the minimums and maximums might provide more insight into the separation of the three iris species.

```
. estat grsummarize, min max
Estimation sample discrim lda
Summarized by iris
```

		iris setosa	versicolor	virginica	Total
seplen					
	Min	4.3	4.9	4.9	4.3
	Max	5.8	7	7.9	7.9
sepwid					
	Min	2.3	2	2.2	2
	Max	4.4	3.4	3.8	4.4
petlen					
	Min	1	3	4.5	1
	Max	1.9	5.1	6.9	6.9
petwid					
	Min	.1	1	1.4	.1
	Max	.6	1.8	2.5	2.5

Although this table is helpful, an altered view of it might make comparisons easier. `estat grsummarize` allows a format to be specified with each requested statistic. We can request a shorter format for the minimum and maximum and specify a fixed format so that the decimal point lines up. `estat grsummarize` also has a `transpose` option that places the variables and requested statistics as columns and the groups as rows. If you have fewer discriminating variables than groups, this might be the most natural way to view the statistics. Here we have more variables, but with a narrow display format, the transposed view still works well.

```
. estat grsummarize, min(%4.1f) max(%4.1f) transpose
Estimation sample discrim lda
Summarized by iris
```

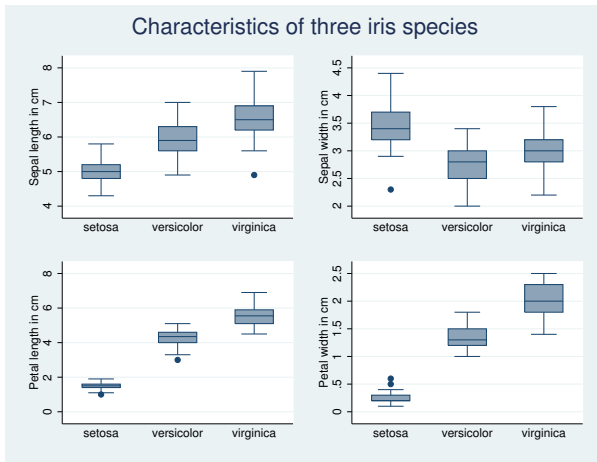
iris	seplen Min	seplen Max	sepwid Min	sepwid Max	petlen Min	petlen Max	petwid Min	petwid Max
setosa	4.3	5.8	2.3	4.4	1.0	1.9	0.1	0.6
versicolor	4.9	7.0	2.0	3.4	3.0	5.1	1.0	1.8
virginica	4.9	7.9	2.2	3.8	4.5	6.9	1.4	2.5
Total	4.3	7.9	2.0	4.4	1.0	6.9	0.1	2.5

The maximum petal length and width for *Iris setosa* are much smaller than the minimum petal length and width for the other two species. The petal length and width clearly separate *Iris setosa* from the other two species.

You are not limited to one or two statistics with `estat grsummarize`, and each statistic may have different requested display formats. The total column, or row if the table is transposed, can also be suppressed.

Using Stata's `graph box` command is another way of seeing the differences among the three iris species for the four discriminating variables.

```
. graph box seplen, over(iris) name(sl)
. graph box sepwid, over(iris) name(sw)
. graph box petlen, over(iris) name(pl)
. graph box petwid, over(iris) name(pw)
. graph combine sl sw pl pw, title(Characteristics of three iris species)
```



The box plots confirm the lack of overlap in the petal lengths and widths for *Iris setosa* compared with the other two iris species. Other differences between the species are also seen.

◁

More summary `estat` commands are available after `discrim lda`; see [MV] **discrim lda postestimation**.

Discrimination listings

Listing the true group, classified group, group posterior probabilities, and discriminating variables for observations of interest after `discrim` is easy with the `estat list` command.

▷ Example 2

Example 1 of [MV] **discrim** introduced the riding-mower data of Johnson and Wichern (2007) and presented a linear discriminant analysis that concluded with the use of `estat list` displaying the misclassified observations.

```
. use http://www.stata-press.com/data/r14/lawnmower2
(Johnson and Wichern (2007) Table 11.1)
. discrim lda income lotsize, group(owner) notable
. estat list, class(loo) pr(loo) misclassified
```

	Classification			Probabilities		LOO Probabilities	
Obs.	True	Class.	LOO Cl.	nonowner	owner	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
```

The `misclassified` option limited the listing to those observations that were misclassified by the linear discriminant model. `class(loo)` and `pr(loo)` added leave-one-out (LOO) classifications and probabilities to the resubstitution classifications and probabilities.

We demonstrate a few other options available with `estat list`. We can limit which observations are displayed with `if` and `in` qualifiers and can add the display of the discriminating variables with the `varlist` option. Here we limit the display to those observations that have `income` greater than $110,000.

```
. estat list if income > 110, varlist
```

	Data		Classification		Probabilities	
Obs.	income	lotsize	True	Class.	nonowner	owner
2	115.5	16.8	owner	owner	0.4945	0.5055
5	117.0	23.6	owner	owner	0.0040	0.9960
6	140.1	19.2	owner	owner	0.0125	0.9875
7	138.0	17.6	owner	owner	0.0519	0.9481
8	112.8	22.4	owner	owner	0.0155	0.9845
10	123.0	20.8	owner	owner	0.0196	0.9804
12	111.0	20.0	owner	owner	0.1107	0.8893
17	114.0	17.6	nonowner	owner *	0.3776	0.6224

```
* indicates misclassified observations
```

Starting with the command above, we specify `sep(0)` to suppress the separator line that, by default, displays after every 5 observations. We eliminate the observation numbers with the `noobs` option. With the `class()` option: the `looclass` suboption adds the LOO classification; the `noclass` suboption suppress the resubstitution classification; and the `nostar` suboption eliminates the marking

of misclassified observations with asterisks. With `pr(loopr nopr)` we specify that LOO probabilities are to be displayed and resubstitution probabilities are to be suppressed.

```
. estat list if income > 110, sep(0) class(looclass noclass nostar)
> pr(loopr nopr) varlist noobs
```

Data		Classification		LOO Probabilities	
income	lotsize	True	LOO Cl	nonowner	owner
115.5	16.8	owner	nonowner	0.6177	0.3823
117.0	23.6	owner	owner	0.0029	0.9971
140.1	19.2	owner	owner	0.0124	0.9876
138.0	17.6	owner	owner	0.0737	0.9263
112.8	22.4	owner	owner	0.0168	0.9832
123.0	20.8	owner	owner	0.0217	0.9783
111.0	20.0	owner	owner	0.1206	0.8794
114.0	17.6	nonowner	owner	0.2791	0.7209

◁

Use the `if e(sample)` qualifier to restrict the listing from `estat list` to the estimation sample. Out-of-sample listings are obtained if your selected observations are not part of the estimation sample.

As an alternative to `estat list`, you can use `predict` after `discrim` to obtain the classifications, posterior probabilities, or whatever else is available for prediction from the `discrim` subcommand that you ran, and then use `list` to display your predictions; see [D] **list** and see example 2 of [MV] **discrim knn postestimation** for an example.

Classification tables and error rates

Classification tables (also known as confusion matrices) and error-rate estimate tables are available with the `estat classtable` and `estat errorrate` commands after `discrim`.

▷ Example 3

Example 2 of [MV] **discrim knn** introduces a head measurement dataset from Rencher and Christensen (2012, 291) with six discriminating variables and three groups. We perform a quadratic discriminant analysis (QDA) on the dataset to illustrate classification tables and error-rate estimation.

```
. use http://www.stata-press.com/data/r14/head
(Table 8.3 Head measurements, Rencher and Christensen (2012))
. discrim qda wdim circum fbeye eyehd earhd jaw, group(group)
Quadratic discriminant analysis
Resubstitution classification summary
```

Key
Number Percent

True group	Classified high school	college	nonplayer	Total
high school	27 90.00	1 3.33	2 6.67	30 100.00
college	2 6.67	21 70.00	7 23.33	30 100.00
nonplayer	1 3.33	4 13.33	25 83.33	30 100.00
Total	30 33.33	26 28.89	34 37.78	90 100.00
Priors	0.3333	0.3333	0.3333	

By default, `discrim` displayed the resubstitution classification table. A resubstitution classification table is obtained by classifying the observations used in building the discriminant model. The resubstitution classification table is overly optimistic as an indicator of how well you might classify other observations.

This resubstitution classification table shows that from the high school group 27 observations were correctly classified, 1 observation was classified as belonging to the college group, and 2 observations were classified as belonging to the nonplayer group. The corresponding percentages were also presented: 90%, 3.33%, and 6.67%. The college and nonplayer rows are read in a similar manner. For instance, there were 7 observations from the college group that were misclassified as nonplayers. Row and column totals are presented along with the group prior probabilities. See table 9.4 of Rencher and Christensen (2012, 321) for this same classification table.

There are various ways of estimating the error rate for a classification. `estat errorrate` presents the overall (total) error rate and the error rate for each group. By default, it uses a count-based estimate of the error rate.

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

	group			
	high school	college	nonplayer	Total
Error rate	.1	.3	.166666667	.188888889
Priors	.333333333	.333333333	.333333333	

This is a resubstitution count-based error-rate estimate corresponding to the classification table previously presented. Three of the 30 high school observations were misclassified—a proportion of 0.1; 9 of the 30 college observations were misclassified—a proportion of 0.3; and 5 of the 30 nonplayers were misclassified—a proportion of 0.1667. The total error rate is computed as the sum of the group error rates times their prior probabilities—here 0.1889.

An error-rate estimate based on the posterior probabilities is also available with `estat errorrate`.

```
. estat errorrate, pp
Error rate estimated from posterior probabilities
```

Error Rate	group			
	high school	college	nonplayer	Total
Stratified	.08308968	.337824355	.2030882	.208000745
Unstratified	.08308968	.337824355	.2030882	.208000745
Priors	.333333333	.333333333	.333333333	

Because we did not specify otherwise, we obtained resubstitution error-rate estimates. By default both the stratified and unstratified estimates are shown. The stratified estimates give less weight to probabilities where the group sample size is large compared with the group prior probabilities; see *Methods and formulas* for details. Here the stratified and unstratified estimates are the same. This happens when the prior probabilities are proportional to the sample sizes—here we have equal prior probabilities and equal group sizes.

For this example, the count-based and posterior-probability-based estimates are similar to one another.

Leave-one-out (LOO) estimation provides a more realistic assessment of your potential classification success with observations that were not used in building the discriminant analysis model. The `loo` option of `estat classtable` and `estat errorrate` specify a LOO estimation.

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

Key
Number

True group	LOO Classified		
	high school	college	nonplayer
high school	26	2	2
college	3	16	11
nonplayer	4	9	17

To demonstrate some of the available options, we specified the `nopercents` option to suppress the display of percentages; the `nopriors` option to suppress the display of the prior probabilities; and the `nototals` option to suppress row and column totals.

If you compare this LOO classification table with the resubstitution classification table, you will see that fewer observations appear on the diagonal (were correctly classified) in the LOO table. The LOO estimates are less biased than the resubstitution estimates.

We now examine the LOO error-rate estimates by using the `loo` option with the `estat error` command. We first produce the count-based estimates and then request the posterior-probability-based estimates. In the first case, we use the `nopriors` option to demonstrate that you can suppress the display of the prior probabilities. Suppressing the display does not remove their effect on the computations. In the second `estat errorrate` call, we specify that only the unstratified estimates be presented. (Because the prior probabilities and samples sizes match [are equal], the stratified results will be the same.)

```
. estat err, loo nopriors
Error rate estimated by leave-one-out error count
```

	group			
	high school	college	nonplayer	Total
Error rate	.133333333	.466666667	.433333333	.344444444

```
. estat err, loo pp(unstratified)
Error rate estimated from leave-one-out posterior probabilities
```

	group			
Error Rate	high school	college	nonplayer	Total
Unstratified	.049034154	.354290969	.294376541	.232567222
Priors	.333333333	.333333333	.333333333	

Instead of displaying percentages below the counts in the classification table, we can display average posterior probabilities. The `probabilities` option requests the display of average posterior probabilities. We add the `nopriors` option to demonstrate that the prior probabilities can be suppressed from the table. The classifications are still based on the prior probabilities; they are just not displayed.

```
. estat classtable, probabilities nopriors
Resubstitution average-posterior-probabilities classification table
```

Key
Number Average posterior probability

	Classified		
True group	high school	college	nonplayer
high school	27 0.9517	1 0.6180	2 0.5921
college	2 0.6564	21 0.8108	7 0.5835
nonplayer	1 0.4973	4 0.5549	25 0.7456
Total	30 0.9169	26 0.7640	34 0.7032

Both `estat classtable` and `estat errorrate` allow `if` and `in` qualifiers so that you can select the observations to be included in the computations and displayed. If you want to limit the table to the estimation sample, use `if e(sample)`. You can also do out-of-sample classification tables and error-rate estimation by selecting observations that were not part of the estimation sample.

◁

❑ Technical note

As noted by Huberty (1994, 92), the posterior-probability-based error-rate estimates for the individual groups may be negative. This may happen when there is a discrepancy between group prior probabilities and relative sample size.

Continuing with our last example, if we use prior probabilities of 0.2, 0.1, and 0.7 for the high school, college, and nonplayer groups, the nonplayer stratified error-rate estimate and the high school group unstratified error-rate estimate are negative.

```
. estat error, pp priors(.2, .1, .7)
Error rate estimated from posterior probabilities
```

Error Rate	group high school	college	nonplayer	Total
Stratified	.19121145	.737812235	-.001699715	.110833713
Unstratified	-.36619243	.126040785	.29616143	.146678593
Priors	.2	.1	.7	

❑

More examples of the use of `estat list`, `estat classtable`, and `estat errorrate` can be found in the other `discrim`-related manual entries.

Stored results

`estat classtable` stores the following in `r()`:

Matrices

`r(counts)`	group counts
`r(percents)`	percentages for each group (unless `nopercents` specified)
`r(avgpostprob)`	average posterior probabilities classified into each group (`probabilities` only)

`estat errorrate` stores the following in `r()`:

Matrices

`r(grouppriors)`	row vector of group prior probabilities used in the calculations
`r(erate_count)`	matrix of error rates estimated from error counts (`count` only)
`r(erate_strat)`	matrix of stratified error rates estimated from posterior probabilities (`pp` only)
`r(erate_unstrat)`	matrix of unstratified error rates estimated from posterior probabilities (`pp` only)

`estat grsummarize` stores the following in `r()`:

Matrices

`r(count)`	group counts
`r(mean)`	means (`mean` only)
`r(median)`	medians (`median` only)
`r(sd)`	standard deviations (`sd` only)
`r(cv)`	coefficients of variation (`cv` only)
`r(semean)`	standard errors of the means (`semean` only)
`r(min)`	minimums (`min` only)
`r(max)`	maximums (`max` only)

Methods and formulas

Let $\mathbf{C}$ denote the classification table (also known as the confusion matrix), with rows corresponding to the true groups and columns corresponding to the assigned groups. Let C_{ij} denote the element from row i and column j of $\mathbf{C}$. C_{ij} represents the number of observations from group i assigned to group j. n_i is the number of observations from group i and $N = \sum_{i=1}^{g} n_i$ is the total sample size. $\mathcal{N}_i = \sum_{j=1}^{g} C_{ij}$ is the number of observations from group i that were classified into one of the g groups. If some observations from group i are unclassified (because of ties), $\mathcal{N}_i \neq n_i$ and $\mathcal{N} \neq N$ (where $\mathcal{N} = \sum \mathcal{N}_i$). Let q_i be the prior probability of group i.

`estat classtable` displays $\mathbf{C}$, with options controlling the display of cell percentages by row, average posterior probabilities, prior probabilities, row totals, and column totals.

McLachlan (2004, chap. 10) devotes a chapter to classification error-rate estimation. The `estat errorrate` command provides several popular error-rate estimates. Smith (1947) introduced the count-based apparent error-rate estimate. The count-based error-rate estimate for group i is

$$\widehat{E}_i^{(C)} = 1 - C_{ii}/\mathcal{N}_i$$

The overall (total) count-based error-rate estimate is

$$\widehat{E}^{(C)} = \sum_{i=1}^{g} q_i \widehat{E}_i^{(C)}$$

In general, $\widehat{E}^{(C)} \neq 1 - \sum_{i=1}^{g} C_{ii}/\mathcal{N}$, though some sources, Rencher and Christensen (2012, 319), appear to report this latter quantity.

If $\mathbf{C}$ is based on the same data used in the estimation of the discriminant analysis model, the error rates are called apparent error rates. Leave-one-out (LOO) error rates are obtained if $\mathbf{C}$ is based on a leave-one-out analysis where each observation to be classified is classified based on the discriminant model built excluding that observation; see Lachenbruch and Mickey (1968) and McLachlan (2004, 342).

Error rates can also be estimated from the posterior probabilities. Huberty (1994, 90–91) discusses hit rates (one minus the error rates) based on posterior probabilities and shows two versions of the posterior-probability based estimate—stratified and unstratified.

Let $\mathcal{P}_{ji}$ be the sum of the posterior probabilities for all observations from group j assigned to group i. The posterior-probability-based unstratified error-rate estimate for group i is

$$\widehat{E}_i^{(Pu)} = 1 - \frac{1}{\mathcal{N} q_i} \sum_{j=1}^{g} \mathcal{P}_{ji}$$

The overall (total) posterior-probability-based unstratified error-rate estimate is

$$\widehat{E}^{(Pu)} = \sum_{i=1}^{g} q_i \widehat{E}_i^{(Pu)}$$

The posterior-probability-based stratified error-rate estimate for group i is

$$\widehat{E}_i^{(Ps)} = 1 - \frac{1}{q_i} \sum_{j=1}^{g} \frac{q_j}{\mathcal{N}_j} \mathcal{P}_{ji}$$

The overall (total) posterior-probability-based stratified error-rate estimate is

$$\widehat{E}^{(Ps)} = \sum_{i=1}^{g} q_i \widehat{E}_i^{(Ps)}$$

References

Anderson, E. 1935. The irises of the Gaspe Peninsula. *Bulletin of the American Iris Society* 59: 2–5.

Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7: 179–188.

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.

Lachenbruch, P. A., and M. R. Mickey. 1968. Estimation of error rates in discriminant analysis. *Technometrics* 10: 1–11.

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

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

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

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

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

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

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