What is the Stata annotated output for Discriminant Analysis?

Authored by

stats writer

June 30, 2024

The Stata annotated output for Discriminant Analysis is a comprehensive summary of the results from a statistical analysis technique used to classify observations into distinct groups based on a set of predictor variables. This output includes detailed information on the discriminant function coefficients, classification accuracy, and significance tests for each variable, as well as graphical representations of the data and model diagnostics. It serves as a comprehensive and organized record of the analysis, providing researchers with valuable insights and interpretations of the discriminant analysis results.

Discriminant Analysis | Stata Annotated Output

This page shows an example of a discriminant analysis in Stata with footnotes explaining the output. The data used in this example are from a data file, discrim.dta, with 244 observations on four variables. The variables include three continuous, numeric variables (outdoor, social and conservative) and one categorical variable (job type) with three levels: 1) customer service, 2) mechanic, and 3) dispatcher.

We are interested in the relationship between the three continuous variables and our categorical variable. Specifically, we would like to know how many

dimensions we would need to express the relationship. Using this relationship,
we can predict a classification based on the continuous variables or assess how
well the continuous variables separate the categories in the classification. We
will be discussing the
degree to which the continuous variables can be used to discriminate
between the groups. Some options for visualizing what occurs in discriminant analysis can be found in the
Discriminant Analysis
Data Analysis
Example.

First, let's read in our data and look at them.

use https://stats.idre.ucla.edu/stat/stata/dae/discrim, clear

Stata has several commands that can be used for discriminant analysis.

Candisc performs canonical linear discriminant

**analysis which is the
classical form of discriminant analysis. We have opted to use candisc, but you could also use discrim
lda which performs the same analysis with a slightly different set of
output. We first list the continuous variables (the "discriminating" variables), and then indicate with group() the categorical variable of interest.**

candisc outdoor social conservative, group(job)

Canonical linear discriminant analysis

**| | Like-
| Canon. Eigen- Variance | lihood
Fcn | Corr. value Prop. Cumul. | Ratio F df1 df2 Prob>F
----+--------------------------------+-----------------------------------
1 | 0.7207 1.08053 0.7712 0.7712 | 0.3640 52.382 6 478 0.0000 e
2 | 0.4927 .320504 0.2288 1.0000 | 0.7573 38.46 2 240 0.0000 e
----
Ho: this and smaller canon. corr. are zero; e = exact F**

Standardized canonical discriminant function

coefficients

| function1 function2
----------+----------
outdoor | .3785725 .9261104
social | -.8306986 .2128593
conservative | .5171682 -.2914406

Canonical structure

| function1 function2
----------+----------
outdoor | .3230982 .9372155
social | -.7653907 .2660298
conservative | .467691 -.2587426

Group means on canonical variables

| job
--------+---------
group1 | customer service
group2 | mechanic
group3 | dispatch

| function1 function2
----------+----------

group1 | -1.2191 -.3890039

group2 | .1067246 .7145704

group3 | 1.419669 -.5059049

Resubstitution classification summary

+---------+

| Key |

|---------|

| Number |

| Percent |

+---------+

| Classified

True | group1 group2 group3 | Total

-----------+---------------------+-------

group1 | 70 11 4 | 85

| 82.35 12.94 4.71 | 100.00

| |

group2 | 16 62 15 | 93

| 17.20 66.67 16.13 | 100.00

| |

group3 | 3 12 51 | 66

| 4.55 18.18 77.27 | 100.00

-----------+---------------------+-------

Total | 89 85 70 | 244

| 36.48 34.84 28.69 | 100.00

| |

Priors | 0.3333 0.3333 0.3333 |

Linear Discriminant Analysis and Coefficients

Canonical linear discriminant analysis

| | Like-

| Canon. Eigen- Variance | lihood

Fcna | Corr.b valuec Prop.d Cumul.e | Ratiof Fg df1h df2i Prob>Fj

-----+--------------------------------+-----------------------------------

1 | 0.7207 1.08053 0.7712 0.7712 | 0.3640 52.382 6 478 0.0000 e

2 | 0.4927 .320504 0.2288 1.0000 | 0.7573 38.46 2 240 0.0000 e

--

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

a. Fcn -

This indicates the first or second canonical linear discriminant function. The number of functions

is equal to 1 less than the number of levels in the group variable or the number of discriminating variables, if there are more groups than variables. In this example, job has three levels and three discriminating variables were used, so two functions are calculated. Each function acts as projections of the data onto a dimension that best separates or discriminates between the groups.

b. Canon. Corr. - These are the canonical correlations of the functions. If we consider our discriminating variables to be one set of variables and the set of dummies generated from our grouping variable to be another set of variables, we can perform a canonical correlation analysis on these two sets.

xi: canon (outdoor social conservative) (i.job)

**This analysis determines how the sets of variables relate to each other using
pairs of linear combinations of the variables from each set ("canonical variates").
Canonical correlations are the Pearson correlations of these pairs of canonical
variates. So if we run the above command, the Stata output will include the canonical correlations we see in our
candisc output:**

**Canonical correlations:
0.7207 0.4927**

**In canonical correlation, each pair of linear combinations is generated to be
maximally correlated, (i.e. best relate the sets of variables to each other).
It makes sense that finding the ways in which the discriminating variables can
be most predictive of the grouping variable would be part of discriminant
analysis. These correlations are closely associated with the eigenvalues of the functions and can**

be calculated as the square root of (eigenvalue)/(1+eigenvalue). They are indicative of how much discriminating power the functions possess. For more on information on canonical correlation, see Stata Annotated Output: CCA.

c. Eigenvalue -

These are the eigenvalues of the matrix product of the inverse of the within-group sums-of-squares and cross-product matrix and the between-groups sums-of-squares and cross-product matrix. These eigenvalues are related to the canonical correlations and describe how much discriminating power a function possesses.

d.

Prop. - This is the proportion of discriminating power of the three continuous variables found in a given function. This proportion is calculated as the proportion of the function's eigenvalue to the sum of all the

eigenvalues. In this analysis, the first function accounts for 77% of the discriminating power of the discriminating variables and the second function accounts for 23%. We can verify this by noting that the sum of the eigenvalues is 1.08053+.320504 = 1.401034. Then (1.08053/1.401034) = 0.7712 and (0.320504/1.401034) = 0.2288.

e. Cumul. -

This is the cumulative proportion of discriminating power. For any analysis, the proportions of discriminating power will sum to one. Thus, the last entry in the cumulative column will also be one.

f. Likelihood Ratio -

This is the likelihood ratio of a given function. It can be used as a test statistic to evaluate the hypothesis that the current canonical correlation and all smaller ones are zero in the population. This is

**equivalent to Wilks' lambda
and is calculated as the product of (1/(1+eigenvalue)) for all functions
included in a given test. For example, the likelihood ratio associated
with the first function is based on the eigenvalues of both the first and second
functions and is equal to (1/(1+1.08053))*(1/(1+.320504)) = 0.3640. The test
associated with the second function is based only on the second eigenvalue and
has a likelihood ratio of (1/(1+.320504)) = 0.7573.**

**g. F - This is the F statistic testing that the canonical correlation
of the given function is equal to zero. In other words, the null
hypothesis is that the function, and all functions that follow, have no
discriminating power. This hypothesis is tested using the F statistic, which is
generated from the likelihood ratio.**

h. df1 -

This is the effect degrees of freedom for the given function. It is based on the number of groups present in the categorical variable and the number of continuous discriminant variables.

i. df2 -

This is the error degrees of freedom for the given function. It is based on the number of groups present in the categorical variable, the number of continuous discriminant variables, and the number of observations in the analysis.

j. Prob>F -

This is the p-value associated with the F statistic of a given function. The null hypothesis that a given function's canonical correlation and all smaller canonical correlations are equal to zero is evaluated with regard to this p-value. If the p-value is less than

the specified alpha (say 0.05), the null hypothesis is rejected. If not, then
we fail to reject the null hypothesis. In this example, we reject both null hypotheses that
the canonical correlations of functions 1 and 2 are zero at alpha level 0.05 because the p-values are
both less than 0.05. Thus, both functions are helpful in discriminating between
the groups found in job based on the discriminant variables in the model.

Standardized canonical discriminant function coefficientsk

| function1 function2
----------+----------
outdoor | .3785725 .9261104
social | -.8306986 .2128593
conservative | .5171682 -.2914406

Canonical structurel

| function1 function2
----------+----------
outdoor | .3230982 .9372155

social | -.7653907 .2660298

conservative | .467691 -.2587426

Group means on canonical variablesm

| job

--------+------------------

group1 | customer service

group2 | mechanic

group3 | dispatch

| function1 function2

-------------+----------------------

group1 | -1.2191 -.3890039

group2 | .1067246 .7145704

group3 | 1.419669 -.5059049

k. Standardized canonical discriminant function coefficients -

These coefficients can be used

to calculate the discriminant score for a given record. The score is calculated

in the same manner as a predicted value from a linear regression, using the

standardized coefficients and the standardized variables. For example, let
zoutdoor, zsocial, and zconservative be the variables created by
standardizing our discriminating variables. Then, for each record, the
function scores would be calculated using the following equations:

Score1 = .3785725*zoutdoor - .8306986*zsocial + .5171682*zconservative

Score2 = .9261104 *zoutdoor + .2128593*zsocial - .2914406*zconservative

The distribution of the scores from each function is standardized
to have a mean of zero and standard deviation of one. The magnitudes of these
coefficients indicate how strongly the discriminating variables effect the
score. For example, we can see that the standardized coefficient for zsocial
in the first function is greater in magnitude than the coefficients for the other

two variables. Thus, social will have the greatest impact of the three on the first discriminant
score.

l.
Canonical structure -

This is the canonical structure, also known as canonical loading or discriminant
loadings, of the discriminant functions. It represents the correlations between
the observed variables (the three continuous discriminating variables) and the
dimensions created with the unobserved discriminant functions (dimensions).

m. Group means on canonical variables -

These are the means of the discriminant function scores by group for each
function calculated. If we calculated the scores of the first function for
each record in our dataset, and then looked at the means of the scores by group,
we would find that group 1 has a mean of -1.2191, group

2 has a mean of .1067246,
and group 3 has a mean of 1.419669. We know that the function scores have a mean
of zero, and we can check this by looking at the sum of the group means multiplied
by the number of records in each group: (85*-1.2191)+(93*.1067246)+(66*1.419669)
= 0.

Resubstitution classification summary

+---------+
| Key |
|---------|
| Number |
| Percent |
+---------+
| Classifiedo
Truen | group1 group2 group3 | Total
-----------+----------------------+-------
group1 | 70 11 4 | 85
| 82.35 12.94 4.71 | 100.00
| |
group2 | 16 62 15 | 93

| 17.20 66.67 16.13 | 100.00
| |
group3 | 3 12 51 | 66
| 4.55 18.18 77.27 | 100.00
----------+----------------------+-------
Totalp| 89 85 70 | 244
| 36.48 34.84 28.69 | 100.00
| |
Priorsq| 0.3333 0.3333 0.3333 |

n.
True -
These are the frequencies of groups found in the data. We can see from the row
totals that 85 records fall into group 1, 93 fall into group 2, and 66 fall into
group 3. These match the results we saw earlier when we looked at the
output for the command tabulate job. Across each row, we see how many of the records in the group are classified by
our analysis into each of the different groups. For example, of the 85 records
that are in group 1, 70 are classified correctly by the

**analysis as belonging to
group 1 and 15 are classified incorrectly as not belonging to group 1 (11 in
group 2 and 4 in group 1).**

**o. Classified -
These are the predicted frequencies of groups from the analysis. The
column totals at the bottom indicate how many total records were predicted to be
in each group. The numbers going down each column indicate how many were
correctly and incorrectly classified. For example, of the 89 records that were
predicted to be in group 1, 70 were correctly predicted, and 19 were incorrectly
predicted (16 group 2 records and 3 group 3 records were predicted to be in
group 1).**

**p. Total -
These are the sums of the counts in a given row or column (and, in the bottom
right-hand corner, the table). The**

row sums are the total number of observations in each group. The column sums are the total numbers of observations predicted to be in each group. The row percents sum to 100%, as displayed in the Total column. The column sums do not sum to 100%, nor do they sum to the percents shown in the Total row. The percents listed in the total row (36.48, 34.84 and 29.69) are the percents of the total records predicted to be in each group. These do sum to 100%, as shown in the square at the bottom right of the table.

q. Priors -

These are the prior proportions assumed for the distribution of records into the groups. By default, the records are assumed to be equally distributed among the categories. Here, we have three groups into which we are classifying records, so the priors proportions are all one third. Stata allows for different priors to be specified using the priors

option.