What is the SAS Annotated Output for Discriminant Analysis?

Authored by

stats writer

June 30, 2024

The SAS annotated output for discriminant analysis is a statistical tool that provides a detailed summary of the results obtained from conducting a discriminant analysis using SAS software. This output includes various tables and graphs that display the key statistical measures, such as discriminant function coefficients, group means, and classification results. It also includes annotations and explanations to help interpret the findings and understand the significance of the results. The SAS annotated output for discriminant analysis is a valuable resource for researchers and analysts, as it allows for a comprehensive and thorough understanding of the discriminant analysis results.

Discriminant Analysis | SAS Annotated Output

This page shows an example of a discriminant analysis in SAS with footnotes explaining the output. The data used in this example are from a data file, https://stats.idre.ucla.edu/wp-content/uploads/2016/02/discrim.sas7bdat, with 244 observations on four variables. The variables include three continuous, numeric variables (outdoor, social and conservative) and one categorical variable (job) with three levels: 1) customer service, 2) mechanic and 3) dispatcher. We will use outdoor, social and conservative as our predictors or "discriminating variables" and job as the grouping variable of interest.

**We are interested in the relationship between the three predictors
and our grouping variable. Specifically, we would like to know how many
dimensions we would need to express this relationship. Using this relationship,
we can predict a classification based on the predictors or assess how
well the predictors separate the groups in the classification. We
will be discussing the degree to which the predictors 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.**

To start, we can examine the overall means of the predictors.

```
proc sort data = 'd:datadiscrim';
by job;
run;
```

proc means mean std min max;
var outdoor social conservative;
run;

Variable N Mean Std Dev Minimum Maximum

OUTDOOR 244 15.6393443 4.8399326 0 28.0000000
SOCIAL 244 20.6762295 5.4792621 7.0000000 35.0000000
CONSERVATIVE 244 10.5901639 3.7267890 0 20.0000000

We are interested in how job relates to outdoor, social and
conservative. Let's look at summary statistics of these three continuous variables for each job category.

proc means mean std min max;
by job;
var outdoor social conservative;
run;

JOB=1

Variable N Mean Std Dev Minimum Maximum

OUTDOOR 85 12.5176471 4.6486346 0 22.0000000

SOCIAL 85 24.2235294 4.3352829 12.0000000 35.0000000

CONSERVATIVE 85 9.0235294 3.1433091 2.0000000 17.0000000

JOB=2

Variable N Mean Std Dev Minimum Maximum

OUTDOOR 93 18.5376344 3.5648012 11.0000000 28.0000000

SOCIAL 93 21.1397849 4.5506602 9.0000000 29.0000000

CONSERVATIVE 93 10.1397849 3.2423535 0 17.0000000

JOB=3

Variable N Mean Std Dev Minimum Maximum

OUTDOOR 66 15.5757576 4.1102521 4.0000000 25.0000000

SOCIAL 66 15.4545455 3.7669895 7.0000000 26.0000000
CONSERVATIVE 66 13.2424242 3.6922397 4.0000000 20.0000000

From this output, we can see that some of the means of outdoor, social
and conservative differ noticeably from group to group in job.
These differences will hopefully allow us to use these predictors to distinguish
observations in one job group from observations in another job
group. Next, we can look at the
correlations between these three predictors. These correlations will give
us some indication of how much unique information each predictor will contribute
to the analysis. If two predictor variables are very highly correlated, then they will be contributing shared information to the analysis.
Uncorrelated variables are likely preferable in this respect. We will also look at the

frequency of each job group.

proc corr;
var outdoor social conservative;
run;

Pearson Correlation Coefficients, N = 244
Prob > |r| under H0: Rho=0

OUTDOOR SOCIAL CONSERVATIVE

OUTDOOR 1.00000 -0.07130 0.07938
0.2672 0.2166

SOCIAL -0.07130 1.00000 -0.23586
0.2672 0.0002

CONSERVATIVE 0.07938 -0.23586 1.00000
0.2166 0.0002

proc freq;
table job;
run;

Cumulative Cumulative
JOB Frequency Percent Frequency Percent

----

1 85 34.84 85 34.84

2 93 38.11 178 72.95

3 66 27.05 244 100.00

SAS has several commands that can be used for discriminant analysis.
The candisc procedure performs canonical linear discriminant analysis which is the
classical form of discriminant analysis.

proc candisc;
class job;
var outdoor social conservative;
run;

Observations 244 DF Total 243
Variables 3 DF Within Classes 241
Classes 3 DF Between Classes 2

Class Level Information

Variable
JOB Name Frequency Weight Proportion

1 _1 85 85.0000 0.348361

2 _2 93 93.0000 0.381148

3 _3 66 66.0000 0.270492

Multivariate Statistics and F Approximations

S=2 M=0 N=118.5

Statistic Value F Value Num DF Den DF Pr > F

Wilks' Lambda 0.36398797 52.38 6 478 <.0001

Pillai's Trace 0.76206574 49.25 6 480 <.0001

Hotelling-Lawley Trace 1.40103067 55.69 6 316.9 <.0001

Roy's Greatest Root 1.08052702 86.44 3 240 <.0001

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

NOTE: F Statistic for Wilks' Lambda is exact.

Adjusted Approximate Squared

Canonical Canonical Standard Canonical

Correlation Correlation Error Correlation

1 0.720661 0.716099 0.030834 0.519353

2 0.492659 . 0.048580 0.242713

Eigenvalues of Inv(E)*H
= CanRsq/(1-CanRsq)

Eigenvalue Difference Proportion Cumulative

1 1.0805 0.7600 0.7712 0.7712
2 0.3205 0.2288 1.0000

Test of H0: The canonical correlations in the current row and all that follow are zero

Likelihood Approximate
Ratio F Value Num DF Den DF Pr > F

1 0.36398797 52.38 6 478 <.0001
2 0.75728681 38.46 2 240 <.0001

.....

Data Summary

Observationsa 244 DF Totald 243
Variablesb 3 DF Within Classese 241
Classesc 3 DF Between Classesf 2

Class Level Information

Variable
JOBg Name Frequencyh Weighti Proportionj

1 _1 85 85.0000 0.348361
2 _2 93 93.0000 0.381148
3 _3 66 66.0000 0.270492

a. Observations - This is the number of observations in the analysis.

b. Variables - This is the number of discriminating continuous variables, or predictors, used in the discriminant analysis. In this example, the discriminating variables are outdoor, social and conservative.

c. Classes - This is the number of levels found in the grouping variable of interest. In this example, the grouping variable job has three values.

d. DF Total - This is the total degrees of freedom. It is equal to (number of observations - 1).

e.

DF Within Classes -

This is the number of degrees of freedom within classes. This is equal to

(number of observations - number of classes).

f.

DF Between Classes -

This is the number of degrees of freedom between classes. This is equal to

(number of classes - 1).

g. JOB - This is the grouping variable of interest. The values

of job are found in this column (1, 2 and 3 representing various job types).

h. Frequency - This is the number of times a given value of the

grouping variable appears in the data. It indicates how the observations

are distributed among the groups.

i.

Weight -

**This is the weight given to each group. In this analysis, each observation
has a weight of 1, so each group's weight is equal to the number of observations
in the group.**

**j. Proportion - This is the proportion of the records that fall into a
given job category. In this example, we see that 35% fall into job
category 1, 38% fall into job category 2, and the remaining 27% fall into job
category 3.**

Multivariate Tests, Canonical Correlations, and Eigenvalues

Statistic Value F Valueo Num DFp Den DFp Pr > Fq

**Wilks' Lambdak 0.36398797 52.38 6 478 <.0001
Pillai's Tracel 0.76206574 49.25 6 480 <.0001
Hotelling-Lawley Tracem 1.40103067 55.69 6 316.9 <.0001
Roy's Greatest Rootn 1.08052702 86.44 3 240 <.0001**

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

NOTE: F Statistic for Wilks' Lambda is exact.

Adjusted Approximate Squared
Canonical Canonical Standard Canonical
Correlationr Correlations Errort Correlationu

1 0.720661 0.716099 0.030834 0.519353
2 0.492659 . 0.048580 0.242713

Eigenvalues of Inv(E)*H
= CanRsq/(1-CanRsq)

Eigenvaluev Differencew Proportionx Cumulativey

1 1.0805 0.7600 0.7712 0.7712
2 0.3205 0.2288 1.0000

Test of H0: The canonical correlations in the current row and all that follow are zero

Likelihood Approximate
Ratioz F Valueo Num DFp Den DFp Pr > Fq

1 0.36398797 52.38 6 478 <.0001
2 0.75728681 38.46 2 240 <.0001

k. Wilks' Lambda -

This is one of the four multivariate statistics calculated by SAS to test the null hypothesis that the canonical correlations are zero (which, in turn, means that there is no linear relationship between the predictors and the grouping variable). Wilks' lambda is the product of the values of (1-canonical correlation2). In this example, our canonical correlations are 0.720661 and 0.492659 so the Wilks' Lambda testing all three of the correlations is (1- 0.7206612)*(1-0.4926592) = 0.36398797. This test statistic is equal to the likelihood ratio (see superscript z).

l. Pillai's Trace -

Pillai's trace is another of the four multivariate statistics calculated by SAS. Pillai's trace is the sum of the squared canonical correlations: 0.7206612 + 0.4926592 = 0.76206574.

m. Hotelling-Lawley Trace -

This is very similar to Pillai's trace. It is the sum of the values of
(canonical correlation2/(1-canonical correlation2)). We can calculate 0.7206612 /(1- 0.7206612) + 0.4926592/(1-0.4926592)
= 1.40103067.

n. Roy's Greatest Root -

This is the largest eigenvalue. Because it is based on a maximum, it can behave
differently from the other three test statistics. In instances where the other
three are not significant and Roy's is significant, the effect should be
considered not significant.

o. (Approximate) F Value -

These are the F values associated with the various tests (likelihood ratio or
one of the four multivariate tests) that are included, by default, in SAS
output. For the likelihood ratio tests, the F values are approximate. For

Roy's Greatest Root, the F value is an upper bound. In the likelihood tests, the F values are testing the hypotheses that the given canonical correlation and all smaller ones are equal to zero in the population. For the multivariate tests, the F values are testing the hypothesis that both canonical correlations are equal to zero in the population.

p. Num DF, Den DF -
These are the degrees of freedom used in determining the F values. Note that there are instances in which the degrees of freedom may be a non-integer (here, the Den DF associated with Hotelling-Lawley Trace is a non-integer) because these degrees of freedom are calculated using the mean squared errors, which are often non-integers.

q. Pr > F -
This is the p-value associated with the F value of a given test statistic. The

null hypothesis the specified canonical correlations are equal to zero is
evaluated with regard to this p-value. The null hypothesis is rejected if the
p-value is less than the specified alpha level (often 0.05). If not, then we
fail to reject the null hypothesis. In this example, we reject the null
hypothesis that both canonical correlations are equal to zero at alpha level
0.05 because the p-values for all tests of this hypothesis are less than 0.05 (Wilks'
Lambda, Pillai's Trace, Hotelling-Lawley Trace, Roy's Greatest
Root and the first Likelihood Ratio). The p-value associated with
the likelihood ratio test of the second canonical correlation suggests that
they we can also reject the hypothesis that the second canonical correlation is
zero.

r. Canonical Correlation -
These are the canonical correlations of our predictor

variables (outdoor, social
and conservative) and the groupings in job. 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. From this analysis, we would arrive at these
canonical correlations.

s. Adjusted Canonical Correlation -
These are adjusted canonical correlations, which are less biased than the raw
correlations. These adjusted values may be negative. If an adjusted canonical
correlation is close to zero or if it is greater than the previous adjusted
canonical correlation, then it is reported as missing.

t. Approximate Standard Error -
These are the approximate standard errors for the canonical correlations.

u. Squared Canonical Correlation -

**These are the squares of the canonical correlations. For example, (0.720661*0.720661)
= 0.519353. These values can be interpreted similarly to R-squared values in OLS
regression: they are the proportion of the variance in the canonical variate of
one set of variables explained by the canonical variate of the other set of
variables.**

v. Eigenvalue -

**These are the eigenvalues of the product of the model matrix and the inverse of
the error matrix from the canonical correlation analysis described in
superscript r. These eigenvalues can also be calculated using the
squared canonical correlations. The largest eigenvalue is equal to largest
squared correlation /(1- largest squared correlation). So 0.519353/(1-0.519353)
= 1.0805. These calculations can be completed for each correlation to find the
corresponding eigenvalue. The magnitudes of the**

eigenvalues are related to
the tests of the correlations. The larger eigenvalues are associated with lower
p-values. If we think about the relationship between the canonical correlations
and the eigenvalues, it makes sense that the larger correlations are more likely
to be significantly different from zero.

w. Difference -
This is the difference between the given eigenvalue and the next-largest
eigenvalue: 1.0805-0.3205 = 0.7600.

x. Proportion -
This is the proportion of the sum of the eigenvalues represented by a given
eigenvalue. The sum of the three eigenvalues is (1.0805+0.3205) = 1.401. Then,
the proportions can be calculated: 1.0805/1.401 = 0.7712 and 0.3205/1.401 = 0.2288.

y. Cumulative -
This is the cumulative sum of the proportions.

z. Likelihood Ratio -

This is the likelihood ratio for testing the hypothesis that the given canonical correlation and all smaller ones are equal to zero in the population. It is equivalent to Wilks' lambda (see superscript k) and can be calculated as the product of the values of (1-canonical correlation2). In this example, our canonical correlations are 0.720661 and 0.492659. Hence the likelihood ratio for testing that both of the correlations are zero is (1- 0.7206612)*(1-0.4926592) = 0.36398797. To test if the smaller canonical correlation, 0.492659, is zero in the population, the likelihood is (1-0.4926592) = 0.75728681.

Canonical Structures

Total Canonical Structureaa

Variable Can1 Can2

OUTDOOR -0.394675 0.912070

SOCIAL 0.857989 0.237581
CONSERVATIVE -0.601504 -0.265113

Between Canonical Structurebb

Variable Can1 Can2

OUTDOOR -0.534845 0.844950
SOCIAL 0.982551 0.185995
CONSERVATIVE -0.957481 -0.288495

Pooled Within Canonical Structurecc

Variable Can1 Can2

OUTDOOR -0.323098 0.937215
SOCIAL 0.765391 0.266030
CONSERVATIVE -0.467691 -0.258743

aa.
Total Canonical Structure
- These are the correlations between the continuous variables and the two
discriminant functions. From this output, we can see that the first
discriminant function is negatively correlated with

outdoor and
conservative and positively correlated with social. The second
discriminant function is positively correlated with outdoor and social
and negatively correlated with conservative. Note that these
correlations do not control for group membership.

bb.
Between Canonical Structure
- These are the correlations between the canonical variates and the continuous
variables between the groups. As in the total canonical structure, the first
discriminant function is negatively correlated with outdoor and
conservative and positively correlated with social; and the second
discriminant function is positively correlated with outdoor and social
and negatively correlated with conservative.

cc.

Pooled Within Canonical Structure

- These are the correlations between the continuous variables and the discriminant functions after controlling for group membership. Note that after controlling for group membership, the signs of the correlations (positive or negative) are unchanged from the total canonical structure, but the magnitudes of the correlations have changed.