What is the process of conducting Canonical Correlation Analysis in SAS for data analysis?

Authored by

stats writer

June 29, 2024

Canonical Correlation Analysis (CCA) is a statistical method used for analyzing the relationship between two sets of continuous variables. In SAS, the process of conducting CCA involves several steps. First, the data must be organized into two separate datasets, one for each set of variables. Next, the PROC CANCORR procedure is used to perform the analysis and generate the canonical correlation coefficients and canonical variates. The results can be interpreted using various output options, such as scatter plots, correlation matrices, and significance tests. It is important to check for assumptions and potential issues, such as multicollinearity, before drawing conclusions from the results. Overall, the process of conducting CCA in SAS involves proper data preparation, using the appropriate procedure, and careful interpretation of the output to understand the relationship between the two sets of variables.

Canonical Correlation Analysis | SAS Data Analysis Examples

Examples of Canonical Correlation Analysis

Version info: Code for this page was tested in SAS 9.3.

Canonical correlation analysis is used to identify and measure the associations among two sets of variables. Canonical correlation is appropriate in the same situations where multiple regression would be, but where are there are multiple intercorrelated outcome variables. Canonical correlation analysis determines a set of canonical variates, orthogonal linear combinations of the variables within each set that best

explain the variability both within and between sets.

Please Note: The purpose of this page is to show how to use various data analysis commands.
It does not cover all aspects of the research process which researchers are expected to do. In
particular, it does not cover data cleaning and checking, verification of assumptions, model
diagnostics and potential follow-up analyses.

Examples of canonical correlation analysis

Example 1. A researcher has collected data on three psychological variables, four academic variables (standardized test scores) and gender for 600 college freshman. She is interested in
how the set of psychological variables relates to the academic variables and gender. In
particular, the researcher is interested in how many dimensions (canonical
variables) are necessary to understand
the association between the two sets of variables.

Example 2. A researcher is interested in exploring associations among factors from two multidimensional

personality tests, the MMPI and the NEO. She is interested in what dimensions
are common between the tests and how much
variance is shared between them. She is specifically interested in finding
whether the neuroticism dimension from the NEO can account for a substantial amount of shared variance between the two tests.

Description of the Data

Let's pursue Example 1 from above.

We have included the data file, which can be obtained by clicking on
mmreg.sas7bdat.
The dataset has 600 observations on eight variables.
The psychological variables are locus of control, self-concept and
motivation. The academic variables are standardized tests in
reading, writing, math and science. Additionally,
the variable female is a zero-one indicator variable
with the one indicating a female student.

Let's look at the data.

proc means data=mylib.mmreg;
run;

The MEANS Procedure

Variable Label N Mean Std Dev Minimum Maximum

ID 600 300.5000000 173.3493582 1.0000000 600.0000000
LOCUS_OF_CONTROL locus of control 600 0.0965333 0.6702799 -2.2300000 1.3600000
SELF_CONCEPT self-concept 600 0.0049167 0.7055125 -2.6199999 1.1900001
MOTIVATION motivation 600 0.6608333 0.3427294 0 1.0000000
READ reading score 600 51.9018334 10.1029830 28.2999992 76.0000000
WRITE writing score 600 52.3848333 9.7264550 25.5000000 67.0999985
MATH math score 600 51.8490000 9.4147363 31.7999992 75.5000000
SCIENCE science score 600 51.7633332 9.7061789

26.0000000 74.1999969

FEMALE 600 0.5450000 0.4983864 0 1.0000000

----

proc freq data=mylib.mmreg;

table female;

run;

The FREQ Procedure

Cumulative Cumulative

FEMALE Frequency Percent Frequency Percent

----

0 273 45.50 273 45.50

1 327 54.50 600 100.00

We did not include correlations among the variables at this point because we will get them later as part of the canonical correlation analysis.

Analysis methods you might consider

Before we show how you can analyze this with a canonical correlation analysis, let's

consider some other methods that you might use.

Canonical correlation analysis

Due to the length of the output, we will be making comments in several places along
the way.

```
proc cancorr corr data=mylib.mmreg;
var locus_of_control self_concept motivation;
with read write math science female;
run;
```

The corr option on the proc cancorr statement produces correlations within and between the two sets of
variables are given below.

The CANCORR Procedure

Correlations Among the Original Variables

Correlations Among the VAR Variables

LOCUS_OF_
CONTROL SELF_CONCEPT MOTIVATION

LOCUS_OF_CONTROL 1.0000 0.1712 0.2451
SELF_CONCEPT 0.1712 1.0000 0.2886
MOTIVATION 0.2451 0.2886 1.0000

Correlations Among the WITH Variables

READ WRITE MATH SCIENCE FEMALE

READ 1.0000 0.6286 0.6793 0.6907 -0.0417
WRITE 0.6286 1.0000 0.6327 0.5691 0.2443
MATH 0.6793 0.6327 1.0000 0.6495 -0.0482
SCIENCE 0.6907 0.5691 0.6495 1.0000 -0.1382
FEMALE -0.0417 0.2443 -0.0482 -0.1382 1.0000

Correlations Between the VAR Variables and the WITH Variables

READ WRITE MATH SCIENCE FEMALE

LOCUS_OF_CONTROL 0.3736 0.3589 0.3373 0.3246 0.1134
SELF_CONCEPT 0.0607 0.0194 0.0536 0.0698 -0.1260
MOTIVATION 0.2106 0.2542 0.1950 0.1157 0.0981

The SAS System 10:43 Tuesday,

The output below gives the three canonical correlations and the multivariate tests of the dimensions. These results show that the first two of the three canonical correlations are statistically significant. The output also includes the four multivariate criteria and the F approximations.

The CANCORR Procedure

Canonical Correlation Analysis

Adjusted Approximate Squared Canonical Canonical Standard Canonical Correlation Correlation Error Correlation

1 0.464086 0.455474 0.032059 0.215376
2 0.167509 . 0.039712 0.028059
3 0.103991 . 0.040417 0.010814

Test of H0: The canonical correlations in Eigenvalues of Inv(E)*H the current row and all that follow are zero

= CanRsq/(1-CanRsq)

Likelihood Approximate

Eigenvalue Difference Proportion Cumulative Ratio F Value Num DF Den DF Pr > F

1 0.2745 0.2456 0.8734 0.8734 0.75436113 11.72 15 1634.7 F

Wilks' Lambda 0.75436113 11.72 15 1634.7

In general, the number of canonical dimensions is equal to the number of variables in the smaller set; however, the number of significant dimensions may be even smaller. Canonical dimensions, also known as canonical variates, are similar to latent variables that are found in factor analysis, except that canonical variates also maximize the correlation between the two sets of variables. For this particular model there are three canonical dimensions of which only the first two are statistically significant. The first test of dimensions tests whether all three dimensions are significant (F = 11.72), the next test

**tests whether dimensions 2 and 3
combined are significant (F = 2.94). Finally, the last test tests whether dimension
3, by itself, is significant (F = 2.16). Therefore dimensions 1 and 2 are each
significant while the third dimension is not.**

**Next, the raw canonical coefficients are shown below. The raw canonical coefficients are interpreted in a manner analogous to interpreting
regression coefficients i.e., for the variable read, a one unit increase in reading leads to a
.0446 increase in the first canonical variate of set 2 when all of
the other variables are held constant. Here is another example: being female leads to
a .6321 increase in dimension 1 for set 2 with the other predictors held constant.**

Raw Canonical Coefficients for the VAR Variables

V1 V2 V3

LOCUS_OF_CONTROL locus of control 1.2538339076 0.6214775237 -0.661689607

SELF_CONCEPT self-concept -0.35134993 1.1876866562 0.8267209411
MOTIVATION motivation 1.2624203286 -2.027264053 2.0002284379

Raw Canonical Coefficients for the WITH Variables

W1 W2 W3

READ reading score 0.0446205959 0.0049100176 0.0213805581
WRITE writing score 0.0358771125 -0.042071471 0.0913073288
MATH math score 0.0234171847 -0.004229472 0.0093982096
SCIENCE science score 0.0050251567 0.0851621751 -0.109835018
FEMALE 0.6321192387 -1.084642482 -1.794646917

The raw coefficients are followed by the standardized canonical coefficients
shown below. When the variables in the model have very different standard
deviations, the standardized coefficients allow for easier comparisons among the

variables. The standardized canonical coefficients are interpreted in a manner analogous to
interpreting standardized regression coefficients. For example, consider the
variable read, a one
standard deviation increase in reading leads to a 0.45 standard deviation increase in the
score on the first canonical variate for set 2 when the other variables in the model are
held constant.

Standardized Canonical Coefficients for the VAR Variables

V1 V2 V3

LOCUS_OF_CONTROL locus of control 0.8404 0.4166 -0.4435
SELF_CONCEPT self-concept -0.2479 0.8379 0.5833
MOTIVATION motivation 0.4327 -0.6948 0.6855

Standardized Canonical Coefficients for the WITH Variables

W1 W2 W3

READ reading score 0.4508 0.0496 0.2160
WRITE writing score 0.3490 -0.4092 0.8881
MATH math score 0.2205 -0.0398 0.0885
SCIENCE science score 0.0488 0.8266 -1.0661
FEMALE 0.3150 -0.5406 -0.8944

Below are correlations between observed variables and canonical variables
which are known as the canonical loadings, which SAS labels as the canonical structure.

Canonical Structure

Correlations Between the VAR Variables and Their Canonical Variables

V1 V2 V3

LOCUS_OF_CONTROL locus of control 0.9040 0.3897 -0.1756
SELF_CONCEPT self-concept 0.0208 0.7087 0.7052
MOTIVATION motivation 0.5672 -0.3509 0.7451

Correlations Between the WITH Variables and Their Canonical Variables

W1 W2 W3

READ reading score 0.8404 0.3588 0.1354
WRITE writing score 0.8765 -0.0648 0.2546
MATH math score 0.7639 0.2979 0.1478
SCIENCE science score 0.6584 0.6768 -0.2304
FEMALE 0.3641 -0.7549 -0.5434

Correlations Between the VAR Variables and the Canonical Variables of the WITH Variables

W1 W2 W3

LOCUS_OF_CONTROL locus of control 0.4196 0.0653 -0.0183
SELF_CONCEPT self-concept 0.0097 0.1187 0.0733
MOTIVATION motivation 0.2632 -0.0588 0.0775

Correlations Between the WITH Variables and the Canonical Variables of the VAR Variables

V1 V2 V3

READ reading score 0.3900 0.0601 0.0141
WRITE writing score 0.4068 -0.0109 0.0265
MATH math score 0.3545 0.0499 0.0154

SCIENCE science score 0.3056 0.1134 -0.0240

FEMALE 0.1690 -0.1265 -0.0565

Things to consider

As in the case of multivariate regression, MANOVA and so on, for valid inference, canonical correlation analysis requires the multivariate normal and homogeneity of variance assumption.

Canonical correlation analysis assumes a linear relationship between the canonical variates and each set of variables.

Similar to multivariate regression, canonical correlation analysis requires a large sample size.

See Also

References