What is Canonical Correlation Analysis and how can it be used for Stata data analysis?

Authored by

stats writer

June 29, 2024

Canonical Correlation Analysis (CCA) is a statistical technique used to explore the relationship between two sets of variables. It seeks to identify linear combinations of variables from each set that are maximally correlated with each other. CCA is commonly used in Stata data analysis to identify and measure the strength of associations between different sets of variables. This technique can be particularly useful when dealing with large and complex datasets, as it allows for the identification of meaningful patterns and relationships that may not be apparent through traditional methods. CCA can also be used for predictive modeling, as it can help identify which variables are most strongly associated with a particular outcome. Overall, CCA is a powerful tool that can provide valuable insights into the relationships between sets of variables in Stata data analysis.

Canonical Correlation Analysis | Stata Data Analysis Examples

Version info: Code for this page was tested in Stata 12.

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

For our analysis example, we are going to expand example 1 about investigating the associations between psychological measures and academic achievement measures.

We have a data file, mmreg.dta, with 600 observations on eight variables.
The psychological variables are locus of control, self-concept and motivation. The academic variables are standardized tests in reading(read), writing (write),math (math) and science (science). Additionally, the variable female is a zero-one indicator variable with the one indicating a female student.

Let's look at the data.

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

summarize locus_of_control self_concept motivation

Variable | Obs Mean Std. Dev. Min Max
----------+----------
locus_of_c~l | 600 .0965333 .6702799 -2.23 1.36
self_concept | 600 .0049167 .7055125 -2.62 1.19
motivation | 600 .6608333 .3427294 0 1

summarize read write math science female

Variable | Obs Mean Std. Dev. Min Max
----------+----------
read | 600 51.90183 10.10298 28.3 76
write | 600 52.38483 9.726455 25.5 67.1
math | 600 51.849 9.414736 31.8 75.5
science | 600 51.76333 9.706179 26 74.2
female | 600 .545 .4983864 0 1

Analysis methods you might consider

Below is a list of some analysis methods you may have encountered.
Some of the methods listed are quite reasonable while others have either
fallen out of favor or have limitations.

Canonical correlation analysis

Below we use the canon command to conduct a canonical correlation
analysis. It requires two sets of variables enclosed with a pair of
parentheses. We specify our psychological variables as the first set of
variables and our academic variables plus gender as the second set. For
convenience, the variables in the first set are called "u" variables and the
variables in the second set are called "v" variables.

canon (locus_of_control self_concept motivation)(read write math science female)

Canonical correlation analysis Number of obs = 600

Raw coefficients for the first variable set

| 1 2 3
---------+----------------------------
locus_of_c~l | 1.2538 -0.6215 -0.6617
self_concept | -0.3513 -1.1877 0.8267
motivation | 1.2624 2.0273 2.0002

Raw coefficients for the second variable set

| 1 2 3
---------+----------------------------
read | 0.0446 -0.0049 0.0214
write | 0.0359 0.0421 0.0913
math | 0.0234 0.0042 0.0094
science | 0.0050 -0.0852 -0.1098
female | 0.6321 1.0846 -1.7946

Canonical correlations:

0.4641 0.1675 0.1040

Tests of significance of all canonical correlations

Statistic	df1	df2	F	Prob>F	
Wilks' lambda	.754361	15	1634.65	11.7157	0.0000 a
Pillai's trace	.254249	15	1782	11.0006	0.0000 a
Lawley-Hotelling trace	.314297	15	1772	12.3763	0.0000 a
Roy's largest root	.274496	5	594	32.6101	0.0000 u

---

e = exact, a = approximate, u = upper bound on F

The output for canonical correlation analysis is made up of
two parts. First is the raw canonical coefficients. The second part begins with
the canonical correlations and includes the overall multivariate tests for
dimensionality.

The raw canonical coefficients can be used to generate the canonical variates,
represented by the columns (1 2 3) in the coefficient tables,
for each set. They 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 the "v" set when all of the
other variables are held constant. Here is another example: being female leads
to a .6321 increase in the dimension 1 for the "v" set with the other predictors held constant.

The number of possible canonical variates, also known as canonical dimensions, is
equal to the number of variables in the smaller set. In our example, the "u"
set (the first set) has three variables and the "v" set (the second set)
has five. This leads to three possible canonical variates for each set,
which corresponds to the three columns for each set and three canonical
correlation coefficients in the output. Canonical dimensions are latent variables that are analogous to factors obtained in factor analysis,
except that canonical variates also maximize the correlation between the two

**sets of variables. In general,
not all the canonical dimensions would be statistically significant. A
significant dimension corresponds to a significant canonical correlation and
vice versa. To test if a canonical correlation is statistically different from zero, we
can use the test option in canon command as shown below. We don't
need to rerun the model, instead we just ask Stata to redisplay the model with
additional information on the requested tests. In order to test all the
canonical dimensions, we need to specify test(1 2 3). Essentially
test(1) is the overall test on three dimensions, test(2) will test
the significance of canonical correlations 2 and 3, and test(3) will test
the significance of the third canonical correlation alone.**

canon, test(1 2 3)

(some output is omitted)

Tests of significance of all canonical correlations

Statistic	df1	df2	F	Prob>F	
Wilks' lambda	.754361	15	1634.65	11.7157	0.0000 a
Pillai's trace	.254249	15	1782	11.0006	0.0000 a
Lawley-Hotelling trace	.314297	15	1772	12.3763	0.0000 a
Roy's largest root	.274496	5	594	32.6101	0.0000 u

Test of significance of canonical correlations 1-3

Statistic	df1	df2	F	Prob>F	
Wilks' lambda	.754361	15	1634.65	11.7157	0.0000 a

Test of significance of canonical correlations 2-3

Statistic	df1	df2	F	Prob>F	
Wilks' lambda	.96143	8	1186	2.9445	0.0029 e

Test of significance of canonical correlation 3

Statistic	df1	df2	F	Prob>F	
Wilks' lambda	.989186	3	594	2.1646	0.0911 e

---

e = exact, a = approximate, u = upper bound on F

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 combined are significant (they are), the next test tests whether dimensions 2 and 3 combined are significant (they are). Finally, the last test tests whether dimension 3, by itself, is significant (it is not). Therefore dimensions 1 and 2 must each be significant.

Now, we might want to inspect what raw coefficients for each of the canonical variates are significant. We can request the standard errors and significant tests via stderr option.

canon, stderr

Linear combinations for canonical correlations Number of obs = 600

| Coef. Std. Err. t P>|t|

```
-------------+----------------------------------------------------------------
u1 |
locus_of_c~l | 1.253834 .1210229 10.36 0.000 1.016153 1.491515
self_concept | -.3513499 .116424 -3.02 0.003 -.5799987 -.1227012
motivation | 1.26242 .2435532 5.18 0.000 .7840983 1.740742
-------------+----------------------------------------------------------------
v1 |
read | .0446206 .0122741 3.64 0.000 .0205152 .068726
write | .0358771 .0122944 2.92 0.004 .0117318 .0600224
math | .0234172 .0127339 1.84 0.066 -.0015914 .0484258
science | .0050252 .0122762 0.41 0.682 -.0190845 .0291348
female | .6321192 .1747222 3.62 0.000 .2889767 .9752618
-------------+----------------------------------------------------------------
u2 |
locus_of_c~l | -.6214775 .3731786 -1.67 0.096 -1.354375 .11142
self_concept | -1.187687 .3589975 -3.31 0.001 -1.892733 -.4826399
motivation | 2.027264 .7510053 2.70 0.007 .5523406 3.502187
```

```
-------------+----------------------------------------------------------------
v2 |
read | -.00491 .0378475 -0.13 0.897 -.07924 .0694199
write | .0420715 .0379101 1.11 0.268 -.0323814 .1165244
math | .0042295 .0392656 0.11 0.914 -.0728854 .0813444
science | -.0851622 .0378541 -2.25 0.025 -.1595052 -
.0108192
female | 1.084642 .5387622 2.01 0.045 .02655 2.142735
-------------+----------------------------------------------------------------
u3 |
locus_of_c~l | -.6616896 .6064262 -1.09 0.276 -1.85267
.5292904
self_concept | .8267209 .5833814 1.42 0.157 -.3190007
1.972443
motivation | 2.000228 1.220406 1.64 0.102 -.3965655
4.397022
-------------+----------------------------------------------------------------
v3 |
read | .0213806 .0615033 0.35 0.728 -.0994078 .1421689
write | .0913073 .0616051 1.48 0.139 -.0296808 .2122955
math | .0093982 .0638077 0.15 0.883 -.1159158 .1347122
science | -.109835 .0615141 -1.79 0.075 -.2306445
.0109745
female | -1.794647 .8755045 -2.05 0.041 -3.514078 -
```

.0752155

(Standard errors estimated conditionally)

Canonical correlations:

0.4641 0.1675 0.1040

Tests of significance of all canonical correlations

Statistic df1 df2 F Prob>F

Wilks' lambda .754361 15 1634.65 11.7157 0.0000 a

Pillai's trace .254249 15 1782 11.0006 0.0000 a

Lawley-Hotelling trace .314297 15 1772 12.3763 0.0000 a

Roy's largest root .274496 5 594 32.6101 0.0000 u

---

e = exact, a = approximate, u = upper bound on F

Note that for the first dimension all of the variables except for math and

science

are statistically significant along with the dimension as a whole. Thus, locus

of control, self-concept, and motivation share some variability with one another, as well as with read, write,

and
female, which also share variablity among each other. For the second
dimension only self-concept, motivation, science and female are significant. The third dimension is not significant and no attention will be paid to its coefficients or
to the Wald tests.

When the variables in the model have very different standard deviations,
the standardized coefficients allow for easier comparisons among the variables. Next we'll display the standardized canonical coefficients for the first two (significant)
dimensions.

canon (locus_of_control self_concept motivation)(read write math science female), first(2) stdcoef notest
Canonical correlation analysis Number of obs = 600

Standardized coefficients for the first variable set

| 1 2
----------+----------

locus_of_c~l | 0.8404 -0.4166
self_concept | -0.2479 -0.8379
motivation | 0.4327 0.6948

Standardized coefficients for the second variable set

| 1 2
----------+----------
read | 0.4508 -0.0496
write | 0.3490 0.4092
math | 0.2205 0.0398
science | 0.0488 -0.8266
female | 0.3150 0.5406

Canonical correlations:
0.4641 0.1675 0.1040

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.**

**Next, we'll use the estat correlations command to look at all of the correlations
within and between sets of variables.**

```
estat correlations

Correlations for variable list 1

| locus_~l self_c~t motiva~n
-------------+---------------------------
locus_of_c~l | 1.0000
self_concept | 0.1712 1.0000
motivation | 0.2451 0.2886 1.0000
-----------------------------------------

Correlations for variable list 2

| read write math sci female
-------------+-------------------------------------------------
read | 1.0000
write | 0.6286 1.0000
```

math | 0.6793 0.6327 1.0000
science | 0.6907 0.5691 0.6495 1.0000
female | -0.0417 0.2443 -0.0482 -0.1382 1.0000

Correlations between variable lists 1 and 2

| locus_~l self_c~t motiva~n
-------------+------------------------------
read | 0.3736 0.0607 0.2106
write | 0.3589 0.0194 0.2542
math | 0.3373 0.0536 0.1950
science | 0.3246 0.0698 0.1157
female | 0.1134 -0.1260 0.0981

Finally, we'll use the estat loadings command to display the
loadings of the variables on the canonical dimensions (variates). These
loadings are correlations between variables and the canonical variates.

estat loadings

Canonical loadings for variable list 1

| 1 2
----------+----------
locus_of_c~l | 0.9040 -0.3897
self_concept | 0.0208 -0.7087
motivation | 0.5672 0.3509

Canonical loadings for variable list 2

| 1 2
----------+----------
read | 0.8404 -0.3588
write | 0.8765 0.0648
math | 0.7639 -0.2979
science | 0.6584 -0.6768
female | 0.3641 0.7549

Correlation between variable list 1 and canonical variates from list 2

| 1 2
----------+----------

locus_of_c~l | 0.4196 -0.0653

self_concept | 0.0097 -0.1187

motivation | 0.2632 0.0588

Correlation between variable list 2 and canonical variates from list 1

| 1 2

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

read | 0.3900 -0.0601

write | 0.4068 0.0109

math | 0.3545 -0.0499

science | 0.3056 -0.1134

female | 0.1690 0.1265

Things to consider

See also

References