# “What is the annotated output for Canonical Correlation Analysis in Stata?”

Authored by

**stats writer**

June 30, 2024

Canonical Correlation Analysis (CCA) is a statistical technique used to explore the relationships between two sets of variables. The annotated output for CCA in Stata provides a comprehensive summary of the analysis, including the correlation coefficients, eigenvalues, and canonical loadings for each variable in both sets. It also includes graphical representations, such as scatter plots and correlation matrices, to aid in the interpretation of the results. The annotated output also provides information on the significance of the correlations and any potential outliers or influential observations. Overall, the annotated output for CCA in Stata serves as a useful tool for understanding the underlying relationships between two sets of variables and identifying any potential patterns or trends.

## Canonical Correlation Analysis | Stata Annotated Output

**This page shows an example of canonical correlation analysis with footnotes explaining the output in Stata. 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 are necessary to understand the association between the two sets of variables.**

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, writing, math and science.
Additionally, the variable female is a zero-one indicator variable with
the one indicating a female student. The researcher is interested in the
relationship between the psychological variables and the academic variables,
with gender considered as well. Canonical correlation analysis aims to
find pairs of linear combinations of each group of variables that are highly
correlated. These linear combinations are called canonical variates. Each
canonical variate is orthogonal to the other canonical variates except for the
one with which its correlation has been maximized. The possible number of such pairs is limited to the number of variables in the

**smallest group. In our example, there are three psychological variables and more than three academic variables. Thus, a canonical correlation analysis on these sets of variables will generate three pairs of canonical variates.**

**To begin, let's read in the dataset.**

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

**We can now proceed with our analysis. In Stata, canonical correlation analysis is conducted using the canon command. Each group of variables is enclosed in parenthesis. We are also interested in multivariate tests for dimensionality, so we will add a test option that will allow us to determine how many of the three generated canonical dimensions are needed to describe the relationship between our sets of variables. Using the stderr option will give us the standard errors**

**and tests of significance for the raw coefficients of the canonical correlations.**

**canon (locus_of_control self_concept motivation)(read write math science female), test(1 2 3) stderr**

**Linear combinations for canonical correlations Number of obs = 600**

**----------**

**--**

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

**----------+----------**

**---**

**u1 |**

**locus_of_control | 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_control | -.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_control | -.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 |

---

**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**

## Linear Combination Output

**Linear combinations for canonical correlations Number of obsa = 600**

**----------**

**| Coef.b Std. Err.c td P>|t|e f**

**----------+----------**

**u1g |**

**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**

**----------+----------**

**v1h |**

**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**

**----------+----------**

**u2i |**

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

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

v2j |

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

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

u3k |

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

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

v3l |

**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 correlationsm:**
**0.4641 0.1675 0.1040**

**a. Number of obs - This is the number of observations in the dataset**
**with valid data in all of the variables listed in the canon command. In**
**this example, as we saw in the summary, we do not have any missing data in this**
**dataset.**

**b. Coef. - These are the raw canonical coefficients. They define the**
**linear relationship between the variables in a given group and the canonical**

variates (v1, u1, etc.). They can be interpreted in the same way you would interpret regression coefficients, assuming the canonical variate as the outcome variable. For example, a one
unit increase in locus_of_control leads to a 1.253834 unit increase in
u1.

c. Std. Err. - These are the standard errors associated with the raw
canonical coefficients. The standard error is used for testing whether the
coefficient is significantly different from 0 by dividing the coefficient estimate
by the standard error to obtain a t-value (see superscript d). The standard
errors can also be used to form a confidence interval for the coefficient (see
superscript f).

d. t - These are the t-values used in testing the null hypothesis that
the coefficient is 0. The t-values follow a t-distribution which is used to

test against a two-sided alternative hypothesis that the Coef. is not
equal to zero.

e. P>|t| - This is the probability the t test statistic (or a more extreme test statistic) would be observed under the null hypothesis that
a particular variable's canonical coefficient is zero, given that the rest of
the variables in the set. For a given alpha level, P>|t| determines whether or not the null hypothesis can be rejected. If P>|t|
is less than alpha, then the null hypothesis can be rejected and the canonical
coefficient estimate is considered statistically significant at that alpha level.

f. - This shows
a 95% confidence interval for the canonical coefficient. This is very useful as it helps
you understand how high and how low the actual population value of the
canonical coefficient might be. The confidence intervals are related to the p-values such that the

coefficient will not be statistically significant if the confidence interval
includes 0.

g. u1 - This is the first canonical variate, or first dimension, for
variable set 1. It is a linear combination of the psychological variables: locus of control,
self-concept and motivation. It is calculated to be maximally
correlated with v1 and orthogonal to all of the other canonical variates.

h. v1 - This is the first canonical variate, or first dimension, for
variable set 2. It is a linear combination of the academic variables (reading, writing, math and science)
and female. It is calculated to be maximally correlated with u1
and orthogonal to all of the other canonical variates.

i. u2 - This is the second canonical variate, or second dimension, for
variable set 1. It is a linear combination of the psychological variables: locus of control,

self-concept and motivation. It is calculated to be maximally
correlated with v2 and orthogonal to all of the other canonical variates.

j. v2 - This is the second canonical variate, or second dimension, for
variable set 2. It is a linear combination of the academic variables (reading, writing, math and science)
and female. It is calculated to be maximally correlated with u2
and orthogonal to all of the other canonical variates.

k. u3 - This is the third canonical variate, or third dimension, for
variable set 1. It is a linear combination of the psychological variables: locus of control,
self-concept and motivation. It is calculated to be maximally
correlated with v3 and orthogonal to all of the other canonical variates.

l. v3 - This is the third canonical variate, or third dimension, for
variable set 2. It is a linear combination of the academic

**variables (reading, writing, math and science)**
**and female. It is calculated to be maximally correlated with u3**
**and orthogonal to all of the other canonical variates.**

**m. Canonical correlation - These are the Pearson correlation coefficients of**
**u1 and v1, u2 and v2, and u3 and v3,**
**respectively. We can use the predict command after running canon**
**to generate the canonical variates, and then examine the correlation matrix of**
**these six variables to see these correlations.**

```
predict u1, u corr(1)
predict v1, v corr(1)
predict u2, u corr(2)
predict v2, v corr(2)
predict u3, u corr(3)
predict v3, v corr(3)

corr u1 v1 u2 v2 u3 v3

| u1 v1 u2 v2 u3 v3
-------------+------------------------------------------------
```

**u1 | 1.0000**

**v1 | 0.4641 1.0000**

**u2 | -0.0000 0.0000 1.0000**

**v2 | -0.0000 0.0000 0.1675 1.0000**

**u3 | -0.0000 0.0000 -0.0000 -0.0000 1.0000**

**v3 | -0.0000 0.0000 -0.0000 0.0000 0.1040 1.0000**

**From this, we can**
**see the non-zero correlations of the canonical variate pairs (u1,v1), (u2,v2)**
**and (u3,v3) and the zero correlations of all other combinations of the canonical**
**variates.**

### Significance Test Output

**Tests of significance of all canonical correlationsn**

**Statistico df1 df2p Fq Prob>Fr**
**Wilks' lambdas .754361 15 1634.65 11.7157 0.0000 a**
**Pillai's tracet .254249 15 1782 11.0006 0.0000 a**
**Lawley-Hotelling traceu .314297 15 1772 12.3763 0.0000 a**
**Roy's largest rootv .274496 5 594 32.6101 0.0000 u**

**----------**

**Test of significance of canonical correlations 1-3w**

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

---

**Test of significance of canonical correlations 2-3x**

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

---

**Test of significance of canonical correlation 3y**

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

---

**e = exact, a = approximate, u = upper bound on Fz**

**n. Tests of significance of all canonical correlations - By default,**
**Stata tests all of the canonical dimensions together, listing four multivariate test**
**statistics and their significance levels. Our null hypothesis is that our two**
**sets of variables are not linearly related. We evaluate this hypothesis**

based on the p-values for the multivariate tests.

o. Statistic - This is the test statistic based on the multivariate
statistic indicated in the prior column. We wish to test the hypothesis that our
two sets of variables are not linearly related. Since our canonical variates
are linear combinations of the sets, correlations between our canonical variates
imply linear relationships between our sets of variables. Thus, these
statistics are calculated from the canonical correlations.

p. df1 & df2 - These are the degrees of freedom used in determining the F statistic.
Note that there are instances in manova when the degrees of freedom may be a
non-integer (here, the df2 associated with Wilks' lambda is a
non-integer) because these degrees of freedom are calculated using the mean
squared errors, which are often non-integers.

q. F - This is the F statistic for the given multivariate

test. Note
that the F statistic for Roy's largest root is quite different that the other F
statistics. For more details about each of the tests, see superscripts s,
t, u, and v.

r. Prob>F - This is the p-value associated with the F statistic of a given test statistic. The
null hypothesis that our two sets of variables are not linearly related is evaluated with regard to this p-value. For a given alpha level, if the p-value is less than alpha, the null hypothesis
is rejected. If not, then we fail to reject the null hypothesis. In
this example, we reject the null hypothesis that our two sets of variables are
not linearly related at alpha level .05 because the p-values are
all less than .05.

s. Wilks' lambda - This is one of the four multivariate statistics
calculated by Stata. Wilks' lambda is the product of the

values of
(1-canonical correlation2). In this example, our canonical
correlations are 0.4641, 0.1675, and 0.1040 so the Wilks' Lambda testing
all three of the correlations is (1-0.46412)*(1-0.16752)*(1-0.10402)
= 0.754361. For the dimensions 2-3, Wilks' lambda is calculated using just 0.1675
and 0.1040: (1-0.16752)*(1-0.10402) = 0.96143. For dimension 3, Wilks' lambda is (1-0.10402) = 0.989186.

t. Pillai's trace - Pillai's trace is another of the four multivariate
statistics calculated by Stata. Pillai's trace is the sum of the squared
canonical correlations. It is only presented in the test of all three
correlations: 0.46412 + 0.16752 + 0.10402 = .254249.

u. Lawley-Hotelling 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 0.46412 /(1- 0.46412) + 16752/(1-0.16752)**
**+ 0.10402/(1-0.10402) = 0.314297.**

**v. Roy's largest root - This is the square of the largest canonical**
**correlation. 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.**

**w. Test of significance of canonical correlations 1-3 - In addition to the tests given in the output produced without options, we specified additional tests in our command. The first test**
**of dimensions again tests whether all three dimensions are significant. We see**
**that, for this test, only one multivariate test statistic is listed (Wilks'**
**lambda), and it is redundant given the default test printed above. In**
**this example, we reject the null hypothesis that our two**

sets of variables are
not linearly related at alpha level .05 because the p-values are
all less than .05.

x. Test of significance of canonical correlations 2-3 - Here, we
test whether dimensions 2 and 3 combined are significant. We reject the null hypothesis at alpha level 0.05 because the p-value
is less than .05.

y. Test of significance of canonical correlation 3 - Here, we
test whether dimension 3, by itself, is significant. We fail to reject
the null hypothesis at alpha level 0.05 because the p-value is greater than 0.05.
Based on the results of the three tests, we can conclude that dimensions 1 and 2
must each be significant. This answers the researcher's original question of how
many dimensions are needed to describe the relationship between the two sets of

**variables.**

**z. e = exact, a = approximate, u = upper bound on F - This indicates how the F statistic was calculated (whether it was an exact calculation, an approximation, or an upper bound) for each of the multivariate tests.**

**For a further exploration of the options available in canon, see the corresponding Data Analysis Example page.**