Title

canon — Canonical correlations

Syntax

`canon` (*varlist*$_1$) (*varlist*$_2$) [*if*] [*in*] [*weight*] [, *options*]

options	Description
Model	
`lc(#)`	calculate the linear combinations for canonical correlation #
`first(#)`	calculate the linear combinations for the first # canonical correlations
`noconstant`	do not subtract means when calculating correlations
Reporting	
`stdcoef`	output matrices of standardized coefficients
`stderr`	display raw coefficients and conditionally estimated standard errors
`level(#)`	set confidence level; default is `level(95)`
`test(`*numlist*`)`	display significance tests for the specified canonical correlations
`notests`	do not display tests
`format(`*%fmt*`)`	numerical format for coefficient matrices; default is `format(%8.4f)`

`by` and `statsby` are allowed; see [U] **11.1.10 Prefix commands**.

`aweights` and `fweights` are allowed; see [U] **11.1.6 weight**.

See [U] **20 Estimation and postestimation commands** for more capabilities of estimation commands.

Menu

Statistics > Multivariate analysis > MANOVA, multivariate regression, and related > Canonical correlations

Description

`canon` estimates canonical correlations and provides the coefficients for calculating the appropriate linear combinations corresponding to those correlations.

`canon` typed without arguments redisplays previous estimation results.

Options

Model

`lc(#)` specifies that linear combinations for canonical correlation # be calculated. By default, all are calculated.

`first(#)` specifies that linear combinations for the first # canonical correlations be calculated. By default, all are calculated.

`noconstant` specifies that means not be subtracted when calculating correlations.

Reporting

`stdcoef` specifies that the first part of the output contain the standard coefficients of the canonical correlations in matrix form. The default is to present the raw coefficients of the canonical correlations in matrix form.

`stderr` specifies that the first part of the output contains the raw coefficients of the canonical correlations, the conditionally estimated standard errors, and the conditionally estimated confidence intervals in the standard estimation table. The default is to present the raw coefficients of the canonical correlations in matrix form.

`level(#)` specifies the confidence level, as a percentage, for confidence intervals of the coefficients. The default is `level(95)` or as set by `set level`; see [U] **20.7 Specifying the width of confidence intervals**. These "confidence intervals" are the result of an approximate calculation; see the technical note later in this entry.

`test(`*numlist*`)` specifies that significance tests of the canonical correlations in the *numlist* be displayed. Because of the nature of significance testing, if there are three canonical correlations, `test(1)` will test the significance of all three correlations, `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.

`notests` specifies that significance tests of the canonical correlation not be displayed.

`format(%`*fmt*`)` specifies the display format for numbers in coefficient matrices; see [D] **format**. `format(%8.4f)` is the default. `format()` may not be specified with `stderr`.

Remarks and examples

stata.com

Canonical correlations attempt to describe the relationships between two sets of variables. Given two sets of variables, $\mathbf{X} = (x_1, x_2, \ldots, x_K)$ and $\mathbf{Y} = (y_1, y_2, \ldots, y_L)$, the goal is to find linear combinations of $\mathbf{X}$ and $\mathbf{Y}$ so that the correlation between the linear combinations is as high as possible. That is, letting $\widehat{x}_1$ and $\widehat{y}_1$ be the linear combinations,

$$\widehat{x}_1 = \beta_{11}x_1 + \beta_{12}x_2 + \cdots + \beta_{1K}x_K$$
$$\widehat{y}_1 = \gamma_{11}y_1 + \gamma_{12}y_2 + \cdots + \gamma_{1L}y_L$$

you wish to find the maximum correlation between $\widehat{x}_1$ and $\widehat{y}_1$ as functions of the β's and the γ's. The second canonical correlation coefficient is defined as the ordinary correlation between

$$\widehat{x}_2 = \beta_{21}x_1 + \beta_{22}x_2 + \cdots + \beta_{2K}x_K \quad \text{and}$$
$$\widehat{y}_2 = \gamma_{21}y_1 + \gamma_{22}y_2 + \cdots + \gamma_{2L}y_L$$

This correlation is maximized subject to the constraints that $\widehat{x}_1$ and $\widehat{x}_2$, along with $\widehat{y}_1$ and $\widehat{y}_2$, are orthogonal and that $\widehat{x}_1$ and $\widehat{y}_2$, along with $\widehat{x}_2$ and $\widehat{y}_1$, are also orthogonal. The third and further correlations are defined similarly. There are $m = \min(K, L)$ such correlations.

Canonical correlation analysis originated with the work of Hotelling (1935, 1936). For an introduction, see Rencher and Christensen (2012, chap. 11), Johnson and Wichern (2007), or Afifi, May, and Clark (2012).

▷ Example 1

Consider two scientists trying to describe how "big" a car is. The first scientist takes physical measurements—the length, weight, headroom, and trunk space—whereas the second takes mechanical measurements—the engine displacement, mileage rating, gear ratio, and turning circle. Can they agree on a conceptual framework?

```
. use http://www.stata-press.com/data/r13/auto
(1978 Automobile Data)

. canon (length weight headroom trunk) (displ mpg gear_ratio turn)
Canonical correlation analysis                        Number of obs =      74

Raw coefficients for the first variable set
                 |         1         2         3         4
-----------------+------------------------------------------
          length |    0.0095    0.1441    0.0329    0.0212
          weight |    0.0010   -0.0037   -0.0010    0.0007
        headroom |    0.0351   -0.3701    1.5361   -0.0440
           trunk |   -0.0023   -0.0343   -0.2135   -0.3253
-----------------------------------------------------------

Raw coefficients for the second variable set
                 |         1         2         3         4
-----------------+------------------------------------------
    displacement |    0.0054   -0.0125    0.0191   -0.0005
             mpg |   -0.0461   -0.0413    0.0683    0.2478
      gear_ratio |    0.0330    1.0280    3.6596   -1.0311
            turn |    0.0794    0.3113    0.0033    0.2240
-----------------------------------------------------------

--------------------------------------------------------------------------
Canonical correlations:
  0.9476  0.3400  0.0634  0.0447
--------------------------------------------------------------------------
Tests of significance of all canonical correlations
                         Statistic      df1      df2            F     Prob>F
         Wilks' lambda    .0897314       16  202.271      15.1900     0.0000 a
        Pillai's trace     1.01956       16      276       5.9009     0.0000 a
Lawley-Hotelling trace     8.93344       16      258      36.0129     0.0000 a
    Roy's largest root     8.79667        4       69     151.7426     0.0000 u
--------------------------------------------------------------------------
                            e = exact, a = approximate, u = upper bound on F
```

By default, `canon` presents the raw coefficients of the canonical correlations in matrix form, reports the canonical correlations, and finally reports the tests of significance of all canonical correlations. The two views on car size are closely related: the best linear combination of the physical measurements is correlated at almost 0.95 with the best linear combination of the mechanical measurements. All the tests are significant.

To see the standardized coefficients instead of the raw coefficients, we can use the `stdcoef` option on replay, which gives the standardized coefficients in matrix form. We specify the `notests` option to suppress the display of tests this time.

```
. canon, stdcoef notests

Canonical correlation analysis                        Number of obs =      74

Standardized coefficients for the first variable set
```

	1	2	3	4
length	0.2110	3.2095	0.7334	0.4714
weight	0.7898	-2.8469	-0.7448	0.5308
headroom	0.0297	-0.3131	1.2995	-0.0373
trunk	-0.0098	-0.1466	-0.9134	-1.3914

```
Standardized coefficients for the second variable set
```

	1	2	3	4
displacement	0.4932	-1.5125	1.7568	-0.0493
mpg	-0.2670	-0.2388	0.3954	1.4337
gear_ratio	0.0150	0.4691	1.6698	-0.4705
turn	0.3493	1.3694	0.0145	0.9857

```
Canonical correlations:
  0.9476  0.3400  0.0634  0.0447
```

◁

❑ Technical note

canon, with the stderr option, reports standard errors for the coefficients in the linear combinations; most other software does not. You should view these standard errors as lower bounds for the true standard errors. It is based on the assumption that the coefficients for one set of measurements are correct for calculating the coefficients and standard errors of the other relationship on the basis of a linear regression.

After canon, if you predict a canonical variate and regress it on the other variable set, the variance you get from the regression will be the variance you get from canon multiplied by the square of the corresponding canonical correlation.

❑

Stored results

canon stores the following in e():

Scalars

e(N)	number of observations
e(df_r)	residual degrees of freedom
e(df)	degrees of freedom
e(df1)	numerator degrees of freedom for significance tests
e(df2)	denominator degrees of freedom for significance tests
e(n_lc)	the linear combination calculated
e(n_cc)	number of canonical correlations calculated
e(rank)	rank of e(V)

Macros

e(cmd)	canon
e(cmdline)	command as typed
e(wtype)	weight type
e(wexp)	weight expression
e(properties)	b V
e(estat_cmd)	program used to implement estat
e(predict)	program used to implement predict
e(marginsnotok)	predictions disallowed by margins

Matrices

e(b)	coefficient vector
e(stat_#)	statistics for canonical correlation #
e(stat_m)	statistics for overall model
e(canload11)	canonical loadings for *varlist*$_1$
e(canload22)	canonical loadings for *varlist*$_2$
e(canload12)	correlation between *varlist*$_1$ and the canonical variates from *varlist*$_2$
e(canload21)	correlation between *varlist*$_2$ and the canonical variates from *varlist*$_1$
e(rawcoef_var1)	raw coefficients for *varlist*$_1$
e(rawcoef_var2)	raw coefficients for *varlist*$_2$
e(stdcoef_var1)	standardized coefficients for *varlist*$_1$
e(stdcoef_var2)	standardized coefficients for *varlist*$_2$
e(ccorr)	canonical correlation coefficients
e(corr_var1)	correlation matrix for *varlist*$_1$
e(corr_var2)	correlation matrix for *varlist*$_2$
e(corr_mixed)	correlation matrix between *varlist*$_1$ and *varlist*$_2$
e(V)	variance–covariance matrix of the estimators

Functions

e(sample)	marks estimation sample

Methods and formulas

Let the covariance matrix between the two sets of variables be

$$\begin{pmatrix} \mathbf{S_{yy}} & \mathbf{S_{yx}} \\ \mathbf{S_{xy}} & \mathbf{S_{xx}} \end{pmatrix}$$

Here $\mathbf{y}$ indicates the first variable set and $\mathbf{x}$ indicates the second variable set.

The squared canonical correlations are the eigenvalues of $\mathbf{V} = \mathbf{S_{yy}^{-1}S_{yx}S_{xx}^{-1}S_{xy}}$ or $\mathbf{W} = \mathbf{S_{xx}^{-1}S_{xy}S_{yy}^{-1}S_{yx}}$ (either will work), which are both nonsymmetric matrices (Rencher 1998, 312–317; Rencher and Christensen 2012, 385–389). Let the eigenvalues of $\mathbf{V}$ (and $\mathbf{W}$) be called r_k, the eigenvectors of $\mathbf{V}$ be called $\mathbf{a}_k$, and the eigenvectors of $\mathbf{W}$ be called $\mathbf{b}_k$. These eigenvectors are the raw coefficients for calculating the canonical variates, which are the linear combinations for the two sets of variables with maximal correlation. The eigenvalue equation for $\mathbf{V}$ is

$$\mathbf{S_{yy}^{-1}S_{yx}S_{xx}^{-1}S_{xy}a}_k - r_k^2\mathbf{a}_k = 0$$

Premultiplying by $\mathbf{S_{xx}^{-1}S_{xy}}$, we see that

$$(\mathbf{S_{xx}^{-1}S_{xy}S_{yy}^{-1}S_{yx}})(\mathbf{S_{xx}^{-1}S_{xy}a}_k) - r_k^2\mathbf{S_{xx}^{-1}S_{xy}a}_k = 0$$

so the $\mathbf{b}_k$ are proportional to $\mathbf{S_{xx}^{-1}S_{xy}a}_k$. Eigenvectors are determined up to a scale factor, and we choose the eigenvectors to give canonical variates with variance one. The canonical variates with correlation r_k are given by

$$\mathbf{u}_k = \mathbf{a}_k\mathbf{x} \qquad \text{and} \qquad \mathbf{v}_k = \mathbf{b}_k\mathbf{y}$$

In fact

$$\mathbf{b}_k = \frac{1}{r_k}\mathbf{S_{xx}^{-1}S_{xy}a}_k$$

To calculate lower bounds for the standard errors in this form, assume that the eigenvectors $\mathbf{a}_k$ are fixed. The formula relating $\mathbf{a}_k$ and $\mathbf{b}_k$ is given above. The coefficients given by $\mathbf{b}_k$ have covariance matrix

$$\frac{1 - r_k^2}{r_k^2(n-k-1)}\mathbf{S_{xx}^{-1}}$$

Here n is the number of observations and k is the number of variables in the set $\mathbf{x}$.

Likewise, we can let the correlation matrix between the two sets of variables be

$$\begin{pmatrix} \mathbf{R_{yy}} & \mathbf{R_{yx}} \\ \mathbf{R_{xy}} & \mathbf{R_{xx}} \end{pmatrix}$$

That is, $\mathbf{R_{yy}}$ is the correlation matrix of the first set of variables with themselves, $\mathbf{R_{xx}}$ is the correlation matrix of the second set of variables with themselves, and $\mathbf{R_{yx}}$ (and $\mathbf{R_{xy}}$) contains the cross-correlations.

Using correlation matrices, the squared canonical correlations are the eigenvalues of $\widetilde{\mathbf{V}} = \mathbf{R_{yy}^{-1}R_{yx}R_{xx}^{-1}R_{xy}}$ or $\widetilde{\mathbf{W}} = \mathbf{R_{xx}^{-1}R_{xy}R_{yy}^{-1}R_{yx}}$ (Rencher 1998, 318–319; Rencher and Christensen 2012, 389). The corresponding eigenvectors are the standardized coefficients for determining the canonical variates from the centered and standardized original variables (mean 0 and variance 1). Eigenvectors are determined only up to a scale factor; we choose the scale to give the canonical variates in standardized (variance 1) form.

If the eigenvalues are $r_1, r_2, \ldots, r_m$ where m is the number of canonical correlations, we test the hypothesis that there is no (linear) relationship between the two variable sets. This is equivalent to the statement that none of the correlations $r_1, r_2, \ldots, r_m$ is significant.

Wilks' (1932) lambda statistic is

$$\Lambda_1 = \prod_{i=1}^{m}(1 - r_i^2)$$

and is a likelihood-ratio statistic. This statistic is distributed as the Wilks Λ-distribution. Rejection of the null hypothesis is for small values of Λ_1.

Pillai's (1955) trace for canonical correlations is

$$V^{(m)} = \sum_{i=1}^{m} r_i^2$$

and the Lawley–Hotelling trace (Lawley 1938 and Hotelling 1951) is

$$U^{(m)} = \sum_{i=1}^{m} \frac{r_i^2}{1 - r_i^2}$$

Roy's (1939) largest root is given by

$$\theta = r_1^2$$

Rencher and Christensen (2012, 391–395) has tables providing critical values for these statistics and discussion on significance testing for canonical correlations.

Canonical loadings, the correlation between a variable set and its corresponding canonical variate set, are calculated by `canon` and used in [MV] **canon postestimation**.

For a note about Harold Hotelling, see [MV] **hotelling**.

Acknowledgment

Significance testing of canonical correlations is based on the `cancor` package originally written by Philip B. Ender of UCLA Academic Technology Services.

References

Afifi, A. A., S. May, and V. A. Clark. 2012. *Practical Multivariate Analysis*. 5th ed. Boca Raton, FL: CRC Press.

Hotelling, H. 1935. The most predictable criterion. *Journal of Educational Psychology* 26: 139–142.

——. 1936. Relations between two sets of variates. *Biometrika* 28: 321–377.

——. 1951. A generalized t^2 test and measurement of multivariate dispersion. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability* 1: 23–41.

Johnson, R. A., and D. W. Wichern. 2007. *Applied Multivariate Statistical Analysis*. 6th ed. Englewood Cliffs, NJ: Prentice Hall.

Lawley, D. N. 1938. A generalization of Fisher's z-test. *Biometrika* 30: 180–187.

Pillai, K. C. S. 1955. Some new test criteria in multivariate analysis. *Annals of Mathematical Statistics* 26: 117–121.

Rencher, A. C. 1998. *Multivariate Statistical Inference and Applications*. New York: Wiley.

Rencher, A. C., and W. F. Christensen. 2012. *Methods of Multivariate Analysis*. 3rd ed. Hoboken, NJ: Wiley.

Roy, S. N. 1939. p-statistics or some generalizations in analysis of variance appropriate to multivariate problems. *Sankhyā* 4: 381–396.

Wilks, S. S. 1932. Certain generalizations in the analysis of variance. *Biometrika* 24: 471–494.

——. 1962. *Mathematical Statistics*. New York: Wiley.

Also see

[MV] **canon postestimation** — Postestimation tools for canon

[MV] **factor** — Factor analysis

[MV] **mvreg** — Multivariate regression

[MV] **pca** — Principal component analysis

[R] **correlate** — Correlations (covariances) of variables or coefficients

[R] **pcorr** — Partial and semipartial correlation coefficients

[R] **regress** — Linear regression

[U] 20 Estimation and postestimation commands