SAS/STAT® 13.1 User’s Guide
The CANCORR Procedure
Gain Greater Insight into Your SAS® Software with SAS Books.

Discover all that you need on your journey to knowledge and empowerment.

support.sas.com/bookstore
for additional books and resources.
Overview: CANCORR Procedure

The CANCORR procedure performs canonical correlation, partial canonical correlation, and canonical redundancy analysis.

Canonical correlation is a generalization of multiple correlation for analyzing the relationship between two sets of variables. In multiple correlation, you examine the relationship between a linear combination of a set of explanatory variables, X, and a single response variable, Y. In canonical correlation, you examine the relationship between linear combinations of the set of X variables and linear combinations of a set of Y variables. These linear combinations are called canonical variables or canonical variates. Either set of variables can be considered explanatory or response variables, since the statistical model is symmetric in the two sets of variables. Simple and multiple correlation are special cases of canonical correlation in which one or both sets contain a single variable.
The CANCORR procedure tests a series of hypotheses that each canonical correlation and all smaller canonical correlations are zero in the population. PROC CANCORR uses an $F$ approximation (Rao 1973; Kshirsagar 1972) that gives better small sample results than the usual $\chi^2$ approximation. At least one of the two sets of variables should have an approximate multivariate normal distribution in order for the probability levels to be valid.

Both standardized and unstandardized canonical coefficients are computed, as well as the four canonical structure matrices showing correlations between the two sets of canonical variables and the two sets of original variables. A canonical redundancy analysis (Stewart and Love 1968; Cooley and Lohnes 1971) can also be done. PROC CANCORR provides multiple regression analysis options to aid in interpreting the canonical correlation analysis. You can examine the linear regression of each variable on the opposite set of variables.

PROC CANCORR can produce a data set containing the scores of each observation on each canonical variable, and you can use the PRINT procedure to list these values. A plot of each canonical variable against its counterpart in the other group is often useful, and you can use PROC SGPLOT with the output data set to produce these plots. A second output data set contains the canonical correlations, coefficients, and most other statistics computed by the procedure.

**Background**

Canonical correlation was developed by Hotelling (1935, 1936).

The application of canonical correlation is discussed by Cooley and Lohnes (1971); Tatsuoka (1971); Mardia, Kent, and Bibby (1979). One of the best theoretical treatments is given by Kshirsagar (1972).

Given a set of $p$ $X$ variables and $q$ $Y$ variables, the CANCORR procedure finds the linear combinations

\[
\begin{align*}
  w_1 &= a_{11}x_1 + a_{21}x_2 + \cdots + a_{p1}x_p \\
  v_1 &= b_{11}y_1 + b_{21}y_2 + \cdots + b_{q1}y_q
\end{align*}
\]

such that the two canonical variables, $w_1$ and $v_1$, have the largest possible correlation. This maximized correlation between the two canonical variables is the first canonical correlation. The coefficients of the linear combinations are canonical coefficients or canonical weights. It is customary to normalize the canonical coefficients so that each canonical variable has a variance of 1.

PROC CANCORR continues by finding a second set of canonical variables, uncorrelated with the first pair, that produces the second-highest correlation coefficient. That is, the second pair of canonical variables is

\[
\begin{align*}
  w_2 &= a_{12}x_1 + a_{22}x_2 + \cdots + a_{p2}x_p \\
  v_2 &= b_{12}y_1 + b_{22}y_2 + \cdots + b_{q2}y_q
\end{align*}
\]

such that $w_2$ is uncorrelated with $w_1$ and $v_1$, $v_2$ is uncorrelated with $w_1$ and $v_1$, and $w_2$ and $v_2$ have the largest possible correlation subject to these constraints. The process of constructing canonical variables continues until the number of pairs of canonical variables is $\min(p, q)$, the number of variables in the smaller group.

Each canonical variable is uncorrelated with all the other canonical variables of either set except for the one corresponding canonical variable in the opposite set. The canonical coefficients are not generally orthogonal,
however, so the canonical variables do not represent jointly perpendicular directions through the space of the
original variables.

The first canonical correlation is at least as large as the multiple correlation between any variable and the
opposite set of variables. It is possible for the first canonical correlation to be very large while all the multiple
correlations for predicting one of the original variables from the opposite set of canonical variables are small.
Canonical redundancy analysis (Stewart and Love 1968; Cooley and Lohnes 1971; van den Wollenberg 1977)
examines how well the original variables can be predicted from the canonical variables.

PROC CANCORR can also perform partial canonical correlation, which is a multivariate generalization
of ordinary partial correlation (Cooley and Lohnes 1971; Timm 1975). Most commonly used parametric
statistical methods, ranging from $t$ tests to multivariate analysis of covariance, are special cases of partial
canonical correlation.

---

**Getting Started: CANCORR Procedure**

The following example demonstrates how you can use the CANCORR procedure to calculate and test
canonical correlations between two sets of variables.

Suppose you want to determine the degree of correspondence between a set of job characteristics and
measures of employee satisfaction. Using a survey instrument for employees, you calculate three measures
of job satisfaction. With another instrument designed for supervisors, you calculate the corresponding job
characteristics profile.

Your three variables associated with job satisfaction are as follows:

- career track satisfaction: employee satisfaction with career direction and the possibility of future
  advancement, expressed as a percent
- management and supervisor satisfaction: employee satisfaction with supervisor’s communication and
  management style, expressed as a percent
- financial satisfaction: employee satisfaction with salary and other benefits, using a scale measurement
  from 1 to 10 (1=unsatisfied, 10=satisfied)

The three variables associated with job characteristics are as follows:

- task variety: degree of variety involved in tasks, expressed as a percent
- feedback: degree of feedback required in job tasks, expressed as a percent
- autonomy: degree of autonomy required in job tasks, expressed as a percent
The following statements create the SAS data set Jobs and request a canonical correlation analysis:

```sas
data Jobs;
  input Career Supervisor Finance Variety Feedback Autonomy;
  label Career = 'Career Satisfaction' Variety = 'Task Variety'
    Supervisor = 'Supervisor Satisfaction' Feedback = 'Amount of Feedback'
    Finance = 'Financial Satisfaction' Autonomy = 'Degree of Autonomy';
  datalines;
    72 26 9 10 11 70
    63 76 7 85 22 93
    96 31 7 83 63 73
    96 98 6 82 75 97
    84 94 6 36 77 97
    66 10 5 28 24 75
    31 40 9 64 23 75
    45 14 2 19 15 50
    42 18 6 33 13 70
    79 74 4 23 14 90
    39 12 2 37 13 70
    54 35 3 23 74 53
    60 75 5 45 58 83
    63 45 5 22 67 53
; proc cancorr data=Jobs
  vprefix=Satisfaction wprefix=Characteristics
  vname='Satisfaction Areas' wname='Job Characteristics';
  var Career Supervisor Finance;
  with Variety Feedback Autonomy;
run;
```

The DATA= option in the PROC CANCORR statement specifies Jobs as the SAS data set to be analyzed. The VPREFIX and WPREFIX options specify the prefixes for naming the canonical variables from the VAR statement and the WITH statement, respectively. The VNAME option specifies 'Satisfaction Areas' to refer to the set of variables from the VAR statement. Similarly, the WNAME option specifies 'Job Characteristics' to refer to the set of variables from the WITH statement.

The VAR statement defines the first of the two sets of variables to be analyzed as Career, Supervisor, and Finance. The WITH statement defines the second set of variables to be Variety, Feedback, and Autonomy. The results of this analysis are displayed in Figure 30.1 to Figure 30.4.

Figure 30.1 displays the canonical correlation, adjusted canonical correlation, approximate standard error, and squared canonical correlation for each pair of canonical variables. The first canonical correlation (the correlation between the first pair of canonical variables) is 0.9194. This value represents the highest possible correlation between any linear combination of the job satisfaction variables and any linear combination of the job characteristics variables.

Figure 30.1 also lists the likelihood ratio and associated statistics for testing the hypothesis that the canonical correlations in the current row and all that follow are zero.
Figure 30.1  Canonical Correlations, Eigenvalues, and Likelihood Tests

The CANCORR Procedure

Canonical Correlation Analysis

<table>
<thead>
<tr>
<th>Canonical Correlation</th>
<th>Adjusted Canonical Correlation</th>
<th>Approximate Standard Error</th>
<th>Squared Canonical Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.919412</td>
<td>0.898444</td>
<td>0.042901</td>
</tr>
<tr>
<td>2</td>
<td>0.418649</td>
<td>0.276633</td>
<td>0.228740</td>
</tr>
<tr>
<td>3</td>
<td>0.113366</td>
<td>.</td>
<td>0.273786</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Eigenvalue</th>
<th>Difference</th>
<th>Proportion</th>
<th>Cumulative</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.4649</td>
<td>0.9604</td>
<td>0.9604</td>
</tr>
<tr>
<td>2</td>
<td>0.2125</td>
<td>0.0373</td>
<td>0.9977</td>
</tr>
<tr>
<td>3</td>
<td>0.0130</td>
<td>0.0023</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Likelihood Ratio</th>
<th>Approximate F Value</th>
<th>Num DF</th>
<th>Den DF</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.12593148</td>
<td>2.93</td>
<td>9</td>
<td>0.0223</td>
</tr>
<tr>
<td>2</td>
<td>0.81413359</td>
<td>0.49</td>
<td>4</td>
<td>0.7450</td>
</tr>
<tr>
<td>3</td>
<td>0.98714819</td>
<td>0.13</td>
<td>1</td>
<td>0.7257</td>
</tr>
</tbody>
</table>

The first approximate F value of 2.93 corresponds to the test that all three canonical correlations are zero. Since the p-value is small (0.0223), you would reject the null hypothesis at the 0.05 level. The second approximate F value of 0.49 corresponds to the test that both the second and the third canonical correlations are zero. Since the p-value is large (0.7450), you would fail to reject the hypothesis and conclude that only the first canonical correlation is significant.

Figure 30.2 lists several multivariate statistics and tests that use approximations based on the F distribution for the null hypothesis that all canonical correlations are zero. Alternatively, you can specify MSTAT=EXACT to compute exact p-values for three of the four tests (Wilk’s Lambda, the Hotelling-Lawley Trace, and Roy’s greatest root) and an improved F approximation for the fourth (Pillai’s Trace). These statistics are described in the section “Multivariate Tests” on page 94 in Chapter 4, “Introduction to Regression Procedures.”
### Figure 30.2 Multivariate Statistics and $F$ Approximations

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
<th>$F$ Value</th>
<th>Num DF</th>
<th>Den DF</th>
<th>Pr &gt; $F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wilks' Lambda</td>
<td>0.12593148</td>
<td>2.93</td>
<td>9</td>
<td>19.621</td>
<td>0.0223</td>
</tr>
<tr>
<td>Pillai's Trace</td>
<td>1.03343732</td>
<td>1.75</td>
<td>9</td>
<td>30</td>
<td>0.1204</td>
</tr>
<tr>
<td>Hotelling-Lawley Trace</td>
<td>5.69042615</td>
<td>4.76</td>
<td>9</td>
<td>9.8113</td>
<td>0.0119</td>
</tr>
<tr>
<td>Roy's Greatest Root</td>
<td>5.46489324</td>
<td>18.22</td>
<td>3</td>
<td>10</td>
<td>0.0002</td>
</tr>
</tbody>
</table>

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

The small $p$-values for these tests ($< 0.05$), except for Pillai’s trace, suggest rejecting the null hypothesis that all canonical correlations are zero in the population, confirming the results of the preceding likelihood ratio test (Figure 30.1). With only one of the tests resulting in a $p$-value larger than 0.05, you can assume that the first canonical correlation is significant. The next step is to interpret or identify the two canonical variables corresponding to this significant correlation.

Even though canonical variables are artificial, they can often be “identified” in terms of the original variables. This is done primarily by inspecting the standardized coefficients of the canonical variables and the correlations between the canonical variables and their original variables. Since only the first canonical correlation is significant, only the first pair of canonical variables (Satisfaction1 and Characteristics1) need to be identified.

PROC CANCORR calculates and displays the raw canonical coefficients for the job satisfaction variables and the job characteristic variables. However, since the original variables do not necessarily have equal variance and are not measured in the same units, the raw coefficients must be standardized to allow interpretation. The coefficients are standardized by multiplying the raw coefficients with the standard deviation of the associated variable.

The standardized canonical coefficients in Figure 30.3 show that the first canonical variable for the Satisfaction group is a weighted sum of the variables Supervisor (0.7854) and Career (0.3028), with the emphasis on Supervisor. The coefficient for the variable Finance is near 0. Thus, a person satisfied with his or her supervisor and with a large degree of career satisfaction would score high on the canonical variable Satisfaction1.

### Figure 30.3 Standardized Canonical Coefficients from the CANCORR Procedure

<table>
<thead>
<tr>
<th></th>
<th>Satisfaction1</th>
<th>Satisfaction2</th>
<th>Satisfaction3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Career</td>
<td>0.3028</td>
<td>-0.5416</td>
<td>1.0408</td>
</tr>
<tr>
<td>Supervisor</td>
<td>0.7854</td>
<td>0.1305</td>
<td>-0.9085</td>
</tr>
<tr>
<td>Finance</td>
<td>0.0538</td>
<td>0.9754</td>
<td>0.3329</td>
</tr>
</tbody>
</table>
The coefficients for the job characteristics variables show that degree of autonomy (Autonomy) and amount of feedback (Feedback) contribute heavily to the Characteristics1 canonical variable (0.8403 and 0.5520, respectively).

Figure 30.4 shows the table of correlations between the canonical variables and the original variables.
Although these univariate correlations must be interpreted with caution since they do not indicate how the original variables contribute jointly to the canonical analysis, they are often useful in the identification of the canonical variables.

Figure 30.4 shows that the supervisor satisfaction variable Supervisor is strongly associated with the Satisfaction1 canonical variable, with a correlation of 0.9644. Slightly less influential is the variable Career, which has a correlation with the canonical variable of 0.7499. Thus, the canonical variable Satisfaction1 seems to represent satisfaction with supervisor and career track.

The correlations for the job characteristics variables show that the canonical variable Characteristics1 seems to represent all three measured variables, with degree of autonomy variable (Autonomy) being the most influential (0.8459).

Hence, you can interpret these results to mean that job characteristics and job satisfaction are related—jobs that possess a high degree of autonomy and level of feedback are associated with workers who are more satisfied with their supervisor and their career. While financial satisfaction is a factor in job satisfaction, it is not as important as the other measured satisfaction-related variables.
Syntax: CANCORR Procedure

The following statements are available in the CANCORR procedure:

```
PROC CANCORR <options> ;
    WITH variables ;
    BY variables ;
    FREQ variable ;
    PARTIAL variables ;
    VAR variables ;
    WEIGHT variable ;
```

The PROC CANCORR statement and the WITH statement are required. The rest of this section provides detailed syntax information for each of the preceding statements, beginning with the PROC CANCORR statement. The remaining statements are covered in alphabetical order.

PROC CANCORR Statement

```
PROC CANCORR <options> ;
```

The PROC CANCORR statement invokes the CANCORR procedure. Optionnally, it also identifies input and output data sets, specifies the analyses performed, and controls displayed output. Table 30.1 summarizes the options available in the PROC CANCORR statement.

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Specify computational details</td>
<td></td>
</tr>
<tr>
<td>EDF=</td>
<td>Specifies error degrees of freedom if input observations are regression residuals</td>
</tr>
<tr>
<td>MSTAT=</td>
<td>Specifies the method of evaluating the multivariate test statistics</td>
</tr>
<tr>
<td>NOINT</td>
<td>Omits intercept from canonical correlation and regression models</td>
</tr>
<tr>
<td>RDF=</td>
<td>Specifies regression degrees of freedom if input observations are regression residuals</td>
</tr>
<tr>
<td>SINGULAR=</td>
<td>Specifies the singularity criterion</td>
</tr>
<tr>
<td>Specify input and output data sets</td>
<td></td>
</tr>
<tr>
<td>DATA=</td>
<td>Specifies input data set name</td>
</tr>
<tr>
<td>OUT=</td>
<td>Specifies output data set name</td>
</tr>
<tr>
<td>OUTSTAT=</td>
<td>Specifies output data set name containing various statistics</td>
</tr>
<tr>
<td>Specify labeling options</td>
<td></td>
</tr>
<tr>
<td>PARPREFIX=</td>
<td>Specifies a prefix for naming residual variables</td>
</tr>
<tr>
<td>VNAME=</td>
<td>Specifies a name to refer to VAR statement variables</td>
</tr>
<tr>
<td>VPREFIX=</td>
<td>Specifies a prefix for naming VAR statement canonical variables</td>
</tr>
<tr>
<td>WNAME=</td>
<td>Specifies a name to refer to WITH statement variables</td>
</tr>
<tr>
<td>WPREFIX=</td>
<td>Specifies a prefix for naming WITH statement canonical variables</td>
</tr>
</tbody>
</table>
### Table 30.1 continued

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Control amount of output</strong></td>
<td></td>
</tr>
<tr>
<td>ALL</td>
<td>Produces simple statistics, input variable correlations, and canonical redundancy analysis</td>
</tr>
<tr>
<td>CORR</td>
<td>Produces input variable correlations</td>
</tr>
<tr>
<td>NCAN=</td>
<td>Specifies number of canonical variables for which full output is desired</td>
</tr>
<tr>
<td>NOPRINT</td>
<td>Suppresses all displayed output</td>
</tr>
<tr>
<td>REDUNDANCY</td>
<td>Produces canonical redundancy analysis</td>
</tr>
<tr>
<td>SHORT</td>
<td>Suppresses default output from canonical analysis</td>
</tr>
<tr>
<td>SIMPLE</td>
<td>Produces means and standard deviations</td>
</tr>
<tr>
<td><strong>Request regression analyses</strong></td>
<td></td>
</tr>
<tr>
<td>VDEP</td>
<td>Requests multiple regression analyses with the VAR variables as dependents and the WITH variables as regressors</td>
</tr>
<tr>
<td>VREG</td>
<td>Requests multiple regression analyses with the VAR variables as regressors and the WITH variables as dependents</td>
</tr>
<tr>
<td>WDEP</td>
<td>Same as VREG</td>
</tr>
<tr>
<td>WREG</td>
<td>Same as VDEP</td>
</tr>
<tr>
<td><strong>Specify regression statistics</strong></td>
<td></td>
</tr>
<tr>
<td>ALL</td>
<td>Produces all regression statistics and includes these statistics in the OUTSTAT= data set</td>
</tr>
<tr>
<td>B</td>
<td>Produces raw regression coefficients</td>
</tr>
<tr>
<td>CLB</td>
<td>Produces 95% confidence interval limits for the regression coefficients</td>
</tr>
<tr>
<td>CORRBB</td>
<td>Produces correlations among regression coefficients</td>
</tr>
<tr>
<td>INT</td>
<td>Requests statistics for the intercept when you specify the B, CLB, SEB, T, or PROBT option</td>
</tr>
<tr>
<td>PCORR</td>
<td>Displays partial correlations between regressors and dependents</td>
</tr>
<tr>
<td>PROBT</td>
<td>Displays probability levels for t statistics</td>
</tr>
<tr>
<td>SEB</td>
<td>Displays standard errors of regression coefficients</td>
</tr>
<tr>
<td>SMC</td>
<td>Displays squared multiple correlations and F tests</td>
</tr>
<tr>
<td>SPCORR</td>
<td>Displays semipartial correlations between regressors and dependents</td>
</tr>
<tr>
<td>SQPCORR</td>
<td>Displays squared partial correlations between regressors and dependents</td>
</tr>
<tr>
<td>SQSPCORR</td>
<td>Displays squared semipartial correlations between regressors and dependents</td>
</tr>
<tr>
<td>STB</td>
<td>Displays standardized regression coefficients</td>
</tr>
<tr>
<td>T</td>
<td>Displays t statistics for regression coefficients</td>
</tr>
</tbody>
</table>

Following are explanations of the options that can be used in the PROC CANCORR statement (in alphabetic order).

**ALL**

displays simple statistics, correlations among the input variables, the confidence limits for the regression coefficients, and the canonical redundancy analysis. If you specify the VDEP or WDEP option, the ALL option displays all related regression statistics (unless the NOPRINT option is specified) and includes these statistics in the OUTSTAT= data set.
PROCCANCORR Statement  ➤  1875

B
produces raw regression coefficients from the regression analyses.

CLB
produces the 95% confidence limits for the regression coefficients from the regression analyses.

CORR
produces correlations among the original variables. If you include a PARTIAL statement, the CORR option produces a correlation matrix for all variables in the analysis, the regression statistics (R square, RMSE), the standardized regression coefficients for both the VAR and WITH variables as predicted from the PARTIAL statement variables, and partial correlation matrices.

CORRB
produces correlations among the regression coefficient estimates.

DATA=SAS-data-set
names the SAS data set to be analyzed by PROC CANCORR. It can be an ordinary SAS data set or a TYPE=CORR, COV, FACTOR, SSCP, UCORR, or UCOV data set. By default, the procedure uses the most recently created SAS data set.

EDF=error-df
specifies the error degrees of freedom if the input observations are residuals from a regression analysis. The effective number of observations is the EDF= value plus one. If you have 100 observations, then specifying EDF=99 has the same effect as omitting the EDF= option.

INT
requests that statistics for the intercept be included when B, CLB, SEB, T, or PROBT is specified for the regression analyses.

MSTAT=FAPPROX | EXACT
specifies the method of evaluating the multivariate test statistics. The default is MSTAT=FAPPROX, which specifies that the multivariate tests are evaluated using the usual approximations based on the $F$ distribution, as discussed in the section “Multivariate Tests” on page 94 in Chapter 4, “Introduction to Regression Procedures.” Alternatively, you can specify MSTAT=EXACT to compute exact $p$-values for three of the four tests (Wilks’ lambda, the Hotelling-Lawley trace, and Roy’s greatest root) and an improved $F$ approximation for the fourth (Pillai’s trace). While MSTAT=EXACT provides better control of the significance probability for the tests, especially for Roy’s greatest root, computations for the exact $p$-values can be appreciably more demanding, and are in fact infeasible for large problems (many dependent variables). Thus, although MSTAT=EXACT is more accurate for most data, it is not the default method.

NCAN=number
specifies the number of canonical variables for which full output is desired. The number must be less than or equal to the number of canonical variables in the analysis.

The value of the NCAN= option specifies the number of canonical variables for which canonical coefficients and canonical redundancy statistics are displayed, and the number of variables shown in the canonical structure matrices. The NCAN= option does not affect the number of displayed canonical correlations.
If an OUTSTAT= data set is requested, the NCAN= option controls the number of canonical variables for which statistics are output. If an OUT= data set is requested, the NCAN= option controls the number of canonical variables for which scores are output.

NOINT
omits the intercept from the canonical correlation and regression models. Standard deviations, variances, covariances, and correlations are not corrected for the mean. If you use a TYPE=SSCP data set as input to the CANCORR procedure and list the variable Intercept in the VAR or WITH statement, the procedure runs as if you also specified the NOINT option. If you use NOINT and also create an OUTSTAT= data set, the data set is TYPE=UCORR.

NOPRINT
suppresses the display of all output. Note that this option temporarily disables the Output Delivery System (ODS). For more information, see Chapter 20, “Using the Output Delivery System.”

OUT=SAS-data-set
creates an output SAS data set to contain all the original data plus scores on the canonical variables. The OUT= option cannot be used when the DATA= data set is TYPE=CORR, COV, FACTOR, SSCP, UCORR, or UCOV. For details about OUT= data sets, see the section “Output Data Sets” on page 1881. If you want to create a SAS data set in a permanent library, you must specify a two-level name. For more information about permanent libraries and SAS data sets, see SAS Language Reference: Concepts.

OUTSTAT=SAS-data-set
creates an output SAS data set containing various statistics, including the canonical correlations and coefficients and the multiple regression statistics you request. For details about OUTSTAT= data sets, see the section “Output Data Sets” on page 1881. If you want to create a SAS data set in a permanent library, you must specify a two-level name. For more information about permanent libraries and SAS data sets, see SAS Language Reference: Concepts.

PCORR
produces partial correlations between regressors and dependent variables, removing from each dependent variable and regressor the effects of all other regressors.

PROBT
produces probability levels for the t statistics in the regression analyses.

RDF=regression-df
specifies the regression degrees of freedom if the input observations are residuals from a regression analysis. The effective number of observations is the actual number minus the RDF= value. The degrees of freedom for the intercept should not be included in the RDF= option.

REDUNDANCY
produces canonical redundancy statistics.
PARPREFIX=\textit{name}

\textbf{PPREFIX=\textit{name}}

specifies a prefix for naming the residual variables in the OUT= data set and the OUTSTAT= data set. By default, the prefix is R_. The number of characters in the prefix plus the maximum length of the variable names should not exceed the current name length defined by the VALIDVARNAME= system option.

\textbf{SEB}

produces standard errors of the regression coefficients.

\textbf{SHORT}

suppresses all default output from the canonical analysis except the tables of canonical correlations and multivariate statistics.

\textbf{SIMPLE}

\textbf{S}

produces means and standard deviations.

\textbf{SINGULAR=\textit{p}}

\textbf{SING=}\textit{p}

specifies the singularity criterion, where \(0 < \textit{p} < 1\). If a variable in the PARTIAL statement has an \(R^2\) square as large as \(1 - \textit{p}\) (where \textit{p} is the value of the SINGULAR= option) when predicted from the variables listed before it in the statement, the variable is assigned a standardized regression coefficient of 0, and the SAS log generates a linear dependency warning message. By default, SINGULAR=1E–8.

\textbf{SMC}

produces squared multiple correlations and \(F\) tests for the regression analyses.

\textbf{SPCORR}

produces semipartial correlations between regressors and dependent variables, removing from each regressor the effects of all other regressors.

\textbf{SQPCORR}

produces squared partial correlations between regressors and dependent variables, removing from each dependent variable and regressor the effects of all other regressors.

\textbf{SQSPCORR}

produces squared semipartial correlations between regressors and dependent variables, removing from each regressor the effects of all other regressors.

\textbf{STB}

produces standardized regression coefficients.

\textbf{T}

produces \(t\) statistics for the regression coefficients.
requests multiple regression analyses with the VAR variables as dependent variables and the WITH variables as regressors.

specifies a character constant to refer to variables from the VAR statement in the output. Enclose the constant in single or double quotes. If you omit the VNAME= option, these variables are referred to as the VAR variables. The number of characters in the label should not exceed the label length defined by the VALIDVARNAME= system option. For more information about the VALIDVARNAME= system option, see SAS System Options: Reference.

specifies a prefix for naming canonical variables from the VAR statement. By default, these canonical variables are given the names V1, V2, and so on. If you specify VPREFIX=ABC, the names are ABC1, ABC2, and so on. The number of characters in the prefix plus the number of digits required to designate the variables should not exceed the name length defined by the VALIDVARNAME= system option. For more information about the VALIDVARNAME= system option, see SAS System Options: Reference.

requests multiple regression analyses with the WITH variables as dependent variables and the VAR variables as regressors.

specifies a character constant to refer to variables in the WITH statement in the output. Enclose the constant in single or double quotes. If you omit the WNAME= option, these variables are referred to as the WITH variables. The number of characters in the label should not exceed the label length defined by the VALIDVARNAME= system option. For more information about the VALIDVARNAME= system option, see SAS System Options: Reference.

specifies a prefix for naming canonical variables from the WITH statement. By default, these canonical variables are given the names W1, W2, and so on. If you specify WPREFIX=XYZ, the names are XYZ1, XYZ2, and so on. The number of characters in the prefix plus the number of digits required to designate the variables should not exceed the label length defined by the VALIDVARNAME= system option. For more information about the VALIDVARNAME= system option, see SAS System Options: Reference.
BY Statement

BY variables;

You can specify a BY statement with PROC CANCORR to obtain separate analyses of observations in groups that are defined by the BY variables. When a BY statement appears, the procedure expects the input data set to be sorted in order of the BY variables. If you specify more than one BY statement, only the last one specified is used.

If your input data set is not sorted in ascending order, use one of the following alternatives:

- Sort the data by using the SORT procedure with a similar BY statement.
- Specify the NOTSORTED or DESCENDING option in the BY statement for the CANCORR procedure. The NOTSORTED option does not mean that the data are unsorted but rather that the data are arranged in groups (according to values of the BY variables) and that these groups are not necessarily in alphabetical or increasing numeric order.
- Create an index on the BY variables by using the DATASETS procedure (in Base SAS software).

For more information about BY-group processing, see the discussion in SAS Language Reference: Concepts. For more information about the DATASETS procedure, see the discussion in the Base SAS Procedures Guide.

FREQ Statement

FREQ variable;

If one variable in your input data set represents the frequency of occurrence for other values in the observation, specify the variable’s name in a FREQ statement. PROC CANCORR then treats the data set as if each observation appeared \( n \) times, where \( n \) is the value of the FREQ variable for the observation. If the value of the FREQ variable is less than one, the observation is not used in the analysis. Only the integer portion of the value is used. The total number of observations is considered to be equal to the sum of the FREQ variable when PROC CANCORR calculates significance probabilities.

PARTIAL Statement

PARTIAL variables;

You can use the PARTIAL statement to base the canonical analysis on partial correlations. The variables in the PARTIAL statement are partialed out of the VAR and WITH variables. If you request an OUT= or OUTSTAT= data set, the residual variables are named by prefixing the characters R_ by default or the string specified in the RPREFIX= option to the VAR variables.
**VAR Statement**

```
VAR variables;
```

The VAR statement lists the variables in the first of the two sets of variables to be analyzed. The variables must be numeric. If you omit the VAR statement, all numeric variables not mentioned in other statements make up the first set of variables. If, however, the DATA= data set is TYPE=SSCP, the default set of variables used as VAR variables does not include the variable Intercept.

**WEIGHT Statement**

```
WEIGHT variable;
```

If you want to compute weighted product-moment correlation coefficients, specify the name of the weighting variable in a WEIGHT statement. The WEIGHT and FREQ statements have a similar effect, except the WEIGHT statement does not alter the degrees of freedom or number of observations. An observation is used in the analysis only if the WEIGHT variable is greater than zero.

**WITH Statement**

```
WITH variables;
```

The WITH statement lists the variables in the second set of variables to be analyzed. The variables must be numeric. The WITH statement is required.

**Details: CANCORR Procedure**

**Missing Values**

If an observation has a missing value for any of the variables in the analysis, that observation is omitted from the analysis.

**Formulas**

Assume without loss of generality that the two sets of variables, \( X \) with \( p \) variables and \( Y \) with \( q \) variables, have means of zero. Let \( n \) be the number of observations, and let \( m = n - 1 \).

Note that the scales of eigenvectors and canonical coefficients are arbitrary. PROC CANCORR follows the usual procedure of rescaling the canonical coefficients so that each canonical variable has a variance of one.

There are several different sets of formulas that can be used to compute the canonical correlations, \( \rho_i \), \( i = 1, \ldots, \min(p, q) \), and unscaled canonical coefficients:
1. Let $S_{XX} = X'X/m$ be the covariance matrix of $X$, $S_{YY} = Y'Y/m$ be the covariance matrix of $Y$, and $S_{XY} = X'Y/m$ be the covariance matrix between $X$ and $Y$. Then the eigenvalues of $S_{YY}^{-1}S_{XY}^{-1}S_{XX}^{-1}S_{XY}$ are the squared canonical correlations, and the right eigenvectors are raw canonical coefficients for the $Y$ variables. The eigenvalues of $S_{XX}^{-1}S_{XY}S_{YY}^{-1}S_{XY}^{-1}$ are the squared canonical correlations, and the right eigenvectors are raw canonical coefficients for the $X$ variables.

2. Let $T = Y'Y$ and $H = Y'X(X'X)^{-1}X'Y$. The eigenvalues $\xi_i$ of $T^{-1}H$ are the squared canonical correlations, $\rho_i^2$, and the right eigenvectors are raw canonical coefficients for the $Y$ variables. Interchange $X$ and $Y$ in the preceding formulas, and the eigenvalues remain the same, but the right eigenvectors are raw canonical coefficients for the $X$ variables.

3. Let $E = T - H$. The eigenvalues of $E^{-1}H$ are $\lambda_i = \rho_i^2/(1 - \rho_i^2)$. The right eigenvectors of $E^{-1}H$ are the same as the right eigenvectors of $T^{-1}H$.

4. Canonical correlation can be viewed as a principal component analysis of the predicted values of one set of variables from a regression on the other set of variables, in the metric of the error covariance matrix. For example, regress the $Y$ variables on the $X$ variables. Call the predicted values $P = X(X'X)^{-1}X'Y$ and the residuals $R = Y - P = (I - X(X'X)^{-1}X')Y$. The error covariance matrix is $R'R/m$. Choose a transformation $Q$ that converts the error covariance matrix to an identity—that is, $(RQ)'(RQ) = Q'R'RQ = mI$. Apply the same transformation to the predicted values to yield, say, $Z = PQ$. Now do a principal component analysis on the covariance matrix of $Z$, and you get the eigenvalues of $E^{-1}H$. Repeat with $X$ and $Y$ variables interchanged, and you get the same eigenvalues.

To show this relationship between canonical correlation and principal components, note that $P'P = H$, $R'R = E$, and $QQ' = mE^{-1}$. Let the covariance matrix of $Z$ be $G$. Then $G = Z'Z/m = (PQ)'PQ/m = Q'P'PQ/m = Q'HQ/m$. Let $u$ be an eigenvector of $G$ and $\kappa$ be the corresponding eigenvalue. Then by definition, $Gu = \kappa u$; hence $Q'HQu/m = \kappa u$. Premultiplying both sides by $Q$ yields $QQ'HQu/m = \kappa Qu$ and thus $E^{-1}HQu = \kappa Qu$. Hence $Qu$ is an eigenvector of $E^{-1}H$ and $\kappa$ is also an eigenvalue of $E^{-1}H$.

5. If the covariance matrices are replaced by correlation matrices, the preceding formulas yield standardized canonical coefficients instead of raw canonical coefficients.

The formulas for multivariate test statistics are shown in the section “Multivariate Tests” on page 94 in Chapter 4, “Introduction to Regression Procedures.” Formulas for linear regression are provided in other sections of that chapter.

---

**Output Data Sets**

**OUT= Data Set**

The OUT= data set contains all the variables in the original data set plus new variables containing the canonical variable scores. The number of new variables is twice that specified by the NCAN= option. The names of the new variables are formed by concatenating the values given by the VPREFIX= and WPREFIX= options (the defaults are V and W) with the numbers 1, 2, 3, and so on. The new variables have mean 0 and variance equal to 1. An OUT= data set cannot be created if the DATA= data set is TYPE=CORR, COV, FACTOR, SSCP, UCORR, or UCOV.
If you use a PARTIAL statement, the OUT= data set also contains the residuals from predicting the VAR variables from the PARTIAL variables. The names of the residual variables are formed by concatenating the values given by the PARPREFIX= option (the default is R_) with the numbers 1, 2, 3, and so on.

**OUTSTAT= Data Set**

The OUTSTAT= data set is similar to the TYPE=CORR or TYPE=UCORR data set produced by the CORR procedure, but it contains several results in addition to those produced by PROC CORR.

The new data set contains the following variables:

- the BY variables, if any
- two new character variables, _TYPE_ and _NAME_
- Intercept, if the INT option is used
- the variables analyzed (those in the VAR statement and the WITH statement)

Each observation in the new data set contains some type of statistic as indicated by the _TYPE_ variable. The values of the _TYPE_ variable are as follows.

<table>
<thead>
<tr>
<th><em>TYPE</em></th>
<th>Contents</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEAN</td>
<td>means</td>
</tr>
<tr>
<td>STD</td>
<td>standard deviations</td>
</tr>
<tr>
<td>USTD</td>
<td>uncorrected standard deviations. When you specify the NOINT option in the PROC CANCORR statement, the OUTSTAT= data set contains standard deviations not corrected for the mean (<em>TYPE</em>='USTD').</td>
</tr>
<tr>
<td>N</td>
<td>number of observations on which the analysis is based. This value is the same for each variable.</td>
</tr>
<tr>
<td>SUMWGT</td>
<td>sum of the weights if a WEIGHT statement is used. This value is the same for each variable.</td>
</tr>
<tr>
<td>CORR</td>
<td>correlations. The <em>NAME</em> variable contains the name of the variable corresponding to each row of the correlation matrix.</td>
</tr>
<tr>
<td>UCORR</td>
<td>uncorrected correlation matrix. When you specify the NOINT option in the PROC CANCORR statement, the OUTSTAT= data set contains a matrix of correlations not corrected for the means.</td>
</tr>
<tr>
<td>CORRB</td>
<td>correlations among the regression coefficient estimates</td>
</tr>
<tr>
<td>STB</td>
<td>standardized regression coefficients. The <em>NAME</em> variable contains the name of the dependent variable.</td>
</tr>
</tbody>
</table>
Output Data Sets

B  raw regression coefficients
SEB  standard errors of the regression coefficients
LCLB  95% lower confidence limits for the regression coefficients
UCLB  95% upper confidence limits for the regression coefficients
T  $t$ statistics for the regression coefficients
PROBT  probability levels for the $t$ statistics
SPCORR  semipartial correlations between regressors and dependent variables
SQSPCORR  squared semipartial correlations between regressors and dependent variables
PCORR  partial correlations between regressors and dependent variables
SQPCORR  squared partial correlations between regressors and dependent variables
RSQUARED  $R$ squares for the multiple regression analyses
ADJRSQ  adjusted $R$ squares
LCLRSQ  approximate 95% lower confidence limits for the $R$ squares
UCLRSQ  approximate 95% upper confidence limits for the $R$ squares
F  $F$ statistics for the multiple regression analyses
PROBF  probability levels for the $F$ statistics
CANCORR  canonical correlations
SCORE  standardized canonical coefficients. The _NAME_ variable contains the name of the canonical variable.
  
  To obtain the canonical variable scores, these coefficients should be multiplied by the standardized data, using means obtained from the observation with _TYPE_='MEAN' and standard deviations obtained from the observation with _TYPE_='STD'.

RAWSCORE  raw canonical coefficients.
  
  To obtain the canonical variable scores, these coefficients should be multiplied by the raw data centered by means obtained from the observation with _TYPE_='MEAN'.

USCORE  scoring coefficients to be applied without subtracting the mean from the raw variables. These are standardized canonical coefficients computed under a NOINT model.
  
  To obtain the canonical variable scores, these coefficients should be multiplied by the data that are standardized by the uncorrected standard deviations obtained from the observation with _TYPE_='USTD'.

STRUCTUR  canonical structure.
Computational Resources

Notation

Let

\[ \begin{align*}
  n & = \text{number of observations} \\
  v & = \text{number of variables} \\
  w & = \text{number of WITH variables} \\
  p & = \max(v, w) \\
  q & = \min(v, w) \\
  b & = v + w \\
  t & = \text{total number of variables (VAR, WITH, and PARTIAL)}
\end{align*} \]

Time Requirements

The time required to compute the correlation matrix is roughly proportional to

\[ n(p + q)^2 \]

The time required for the canonical analysis is roughly proportional to

\[ \frac{1}{6} p^3 + p^2q + \frac{3}{2} pq^2 + 5q^3 \]

but the coefficient for \( q^3 \) varies depending on the number of QR iterations in the singular value decomposition.

Memory Requirements

The minimum memory required is approximately

\[ 4(v^2 + w^2 + t^2) \]

bytes. Additional memory is required if you request the VDEP or WDEP option.
Displayed Output

If the SIMPLE option is specified, PROC CANCORR produces means and standard deviations for each input variable. If the CORR option is specified, PROC CANCORR produces correlations among the input variables. Unless the NOPRINT option is specified, PROC CANCORR displays a table of canonical correlations containing the following:

- Canonical Correlations. These are always nonnegative.

- Adjusted Canonical Correlations (Lawley 1959), which are asymptotically less biased than the raw correlations and can be negative. The adjusted canonical correlations might not be computable, and they are displayed as missing values if two canonical correlations are nearly equal or if some are close to zero. A missing value is also displayed if an adjusted canonical correlation is larger than a previous adjusted canonical correlation.

- Approx Standard Errors, which are the approximate standard errors of the canonical correlations

- Squared Canonical Correlations

- Eigenvalues of INV(E)*H, which are equal to CanRsq/(1–CanRsq), where CanRsq is the corresponding squared canonical correlation. Also displayed for each eigenvalue is the Difference from the next eigenvalue, the Proportion of the sum of the eigenvalues, and the Cumulative proportion.

- Likelihood Ratio for the hypothesis that the current canonical correlation and all smaller ones are zero in the population. The likelihood ratio for all canonical correlations equals Wilks’ lambda.

- Approx F statistic based on Rao’s approximation to the distribution of the likelihood ratio (Rao 1973, p. 556; Kshirsagar 1972, p. 326)

- Num DF and Den DF (numerator and denominator degrees of freedom) and Pr > F (probability level) associated with the F statistic

Unless you specify the NOPRINT option, PROC CANCORR produces a table of multivariate statistics for the null hypothesis that all canonical correlations are zero in the population. These statistics, as described in the section “Multivariate Tests” on page 94 in Chapter 4, “Introduction to Regression Procedures,” are as follows:

- Wilks’ lambda
- Pillai’s trace
- Hotelling-Lawley trace
- Roy’s greatest root

For each of the preceding statistics, PROC CANCORR displays the following, depending on the specification of the MSTAT= option.

If you specify MSTAT=FAPPROX (also the default value), the following statistics are displayed:

- an F approximation or upper bound
- Num DF, the numerator degrees of freedom
• Den DF, the denominator degrees of freedom
• Pr > F, the probability level

If you specify MSTAT=EXACT, the following statistic is displayed:

• a t value

Unless you specify the SHORT or NOPRINT option, PROC CANCORR displays the following:

• both Raw (unstandardized) and Standardized Canonical Coefficients normalized to give canonical variables with unit variance. Standardized coefficients can be used to compute canonical variable scores from the standardized (zero mean and unit variance) input variables. Raw coefficients can be used to compute canonical variable scores from the input variables without standardizing them.
• all four Canonical Structure matrices, giving Correlations Between the canonical variables and the original variables

If you specify the REDUNDANCY option, PROC CANCORR displays the following:

• the Canonical Redundancy Analysis (Stewart and Love 1968; Cooley and Lohnes 1971), including Raw (unstandardized) and Standardized Variance and Cumulative Proportion of the Variance of each set of variables Explained by Their Own Canonical Variables and Explained by The Opposite Canonical Variables
• the Squared Multiple Correlations of each variable with the first \( m \) canonical variables of the opposite set, where \( m \) varies from 1 to the number of canonical correlations

If you specify the VDEP option, PROC CANCORR performs multiple regression analyses with the VAR variables as dependent variables and the WITH variables as regressors. If you specify the WDEP option, PROC CANCORR performs multiple regression analyses with the WITH variables as dependent variables and the VAR variables as regressors. If you specify the VDEP or WDEP option and also specify the ALL option, PROC CANCORR displays the following items. You can also specify individual options to request a subset of the output generated by the ALL option; or you can suppress the output by specifying the NOPRINT option.

SMC  Squared Multiple Correlations and F Tests. For each regression model, identified by its dependent variable name, PROC CANCORR displays the R square, Adjusted R square (Wherry 1931), F Statistic, and Pr > F. Also for each regression model, PROC CANCORR displays an Approximate 95% Confidence Interval for the population R square Helland (1987). These confidence limits are valid only when the regressors are random and when the regressors and dependent variables are approximately distributed according to a multivariate normal distribution. The average R squares for the models considered, unweighted and weighted by variance, are also given.

CORRB  Correlations Among the Regression Coefficient Estimates

STB  Standardized Regression Coefficients

B  Raw Regression Coefficients
PROC CANCORR assigns a name to each table it creates. You can use these names to reference the table when using the Output Delivery System (ODS) to select tables and create output data sets. These names are listed in Table 30.2.

For more information about ODS, see Chapter 20, “Using the Output Delivery System.”

All of the tables in Table 30.2 are created with the specification of the PROC CANCORR statement; a few tables need an additional PARTIAL statement.

<table>
<thead>
<tr>
<th>ODS Table Name</th>
<th>Description</th>
<th>Statement and Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>AvgRSquare</td>
<td>Average R squares (weighted and unweighted)</td>
<td>VDEP, WDEP, SMC, or ALL</td>
</tr>
<tr>
<td>CanCorr</td>
<td>Canonical correlations</td>
<td>default</td>
</tr>
<tr>
<td>CanStructureVCan</td>
<td>Correlations between the VAR canonical variables and the VAR and WITH variables</td>
<td>default (if SHORT is not specified)</td>
</tr>
<tr>
<td>CanStructureWCAn</td>
<td>Correlations between the WITH canonical variables and the WITH and VAR variables</td>
<td>default (if SHORT is not specified)</td>
</tr>
<tr>
<td>ConfidenceLimits</td>
<td>95% confidence limits for the regression coefficients</td>
<td>VDEP, WDEP, CLB, or ALL</td>
</tr>
<tr>
<td>Corr</td>
<td>Correlations among the original variables</td>
<td>ALL, CORR or ALL</td>
</tr>
<tr>
<td>CorrOnPartial</td>
<td>Partial correlations</td>
<td>PARTIAL statement with CORR or ALL</td>
</tr>
<tr>
<td>CorrRegCoeffEst</td>
<td>Correlations among the regression coefficient estimates</td>
<td>VDEP, WDEP, CORRB, or ALL</td>
</tr>
<tr>
<td>MultStat</td>
<td>Multivariate statistics</td>
<td>default</td>
</tr>
</tbody>
</table>
### Table 30.2  continued

<table>
<thead>
<tr>
<th>ODS Table Name</th>
<th>Description</th>
<th>Statement and Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>NObsNVar</td>
<td>Number of observations and variables</td>
<td>SIMPLE or ALL VDEP, WDEP, PCORR, or ALL</td>
</tr>
<tr>
<td>ParCorr</td>
<td>Partial correlations</td>
<td>VDEP, WDEP, PCORR, or ALL</td>
</tr>
<tr>
<td>ProbtRegCoef</td>
<td>Prob &gt;</td>
<td>t</td>
</tr>
<tr>
<td>RawCanCoefV</td>
<td>Raw canonical coefficients for the VAR variables</td>
<td>default (if SHORT is not specified)</td>
</tr>
<tr>
<td>RawCanCoefW</td>
<td>Raw canonical coefficients for the WITH variables</td>
<td>default (if SHORT is not specified)</td>
</tr>
<tr>
<td>RawRegCoef</td>
<td>Raw regression coefficients</td>
<td>VDEP, WDEP, B, or ALL</td>
</tr>
<tr>
<td>Redundancy</td>
<td>Canonical redundancy analysis</td>
<td>REDUNDANCY or ALL</td>
</tr>
<tr>
<td>Regression</td>
<td>Squared multiple correlations and</td>
<td>VDEP, WDEP, SMC, or ALL</td>
</tr>
<tr>
<td>RSquareRMSEOnPartial</td>
<td>R squares and RMSEs on PARTIAL variables</td>
<td>PARTIAL statement with CORR or ALL</td>
</tr>
<tr>
<td>SemiParCorr</td>
<td>Semipartial correlations</td>
<td>VDEP, WDEP, SPCORR, or ALL</td>
</tr>
<tr>
<td>SimpleStatistics</td>
<td>Simple statistics</td>
<td>SIMPLE or ALL</td>
</tr>
<tr>
<td>SqMultCorr</td>
<td>Canonical redundancy analysis: squared multiple correlations</td>
<td>REDUNDANCY or ALL</td>
</tr>
<tr>
<td>SqParCorr</td>
<td>Squared partial correlations</td>
<td>VDEP, WDEP, SQPCORR, or ALL</td>
</tr>
<tr>
<td>SqSemiParCorr</td>
<td>Squared semipartial correlations</td>
<td>VDEP, WDEP, SQSPCORR, or ALL</td>
</tr>
<tr>
<td>StdCanCoefV</td>
<td>Standardized canonical coefficients for the VAR variables</td>
<td>default (if SHORT is not specified)</td>
</tr>
<tr>
<td>StdCanCoefW</td>
<td>Standardized canonical coefficients for the WITH variables</td>
<td>default (if SHORT is not specified)</td>
</tr>
<tr>
<td>StdErrRawRegCoef</td>
<td>Standard errors of the raw regression coefficients</td>
<td>VDEP, WDEP, SEB, or ALL</td>
</tr>
<tr>
<td>StdRegCoef</td>
<td>Standardized regression coefficients</td>
<td>VDEP, WDEP, STB, or ALL</td>
</tr>
<tr>
<td>StdRegCoefOnPartial</td>
<td>Standardized regression coefficients on PARTIAL variables</td>
<td>PARTIAL statement with CORR or ALL</td>
</tr>
<tr>
<td>tValueRegCoef</td>
<td>t values for the regression coefficients</td>
<td>VDEP, WDEP, T, or ALL</td>
</tr>
</tbody>
</table>
Example: CANCORR Procedure

Example 30.1: Canonical Correlation Analysis of Fitness Club Data

Three physiological and three exercise variables are measured on 20 middle-aged men in a fitness club. You can use the CANCORR procedure to determine whether the physiological variables are related in any way to the exercise variables. The following statements create the SAS data set Fit and produce Output 30.1.1 through Output 30.1.5:

```sas
   data Fit;
     input Weight Waist Pulse Chins Situps Jumps;
   datalines;
   191 36 50 5 162 60
   189 37 52 2 110 60
   193 38 58 12 101 101
   162 35 62 12 105 37
   189 35 46 13 155 58
   182 36 56 4 101 42
   211 38 56 8 101 38
   167 34 60 6 125 40
   176 31 74 15 200 40
   154 33 56 17 251 250
   169 34 50 17 120 38
   166 33 52 13 210 115
   154 34 64 14 215 105
   247 46 50 1 50 50
   193 36 46 6 70 31
   202 37 62 12 210 120
   176 37 54 4 60 25
   157 32 52 11 230 80
   156 33 54 15 225 73
   138 33 68 2 110 43
   ;

   proc cancorr data=Fit all
      vprefix=Physiological vname='Physiological Measurements'
      wprefix=Exercises wname='Exercises';
      var Weight Waist Pulse;
      with Chins Situps Jumps;
      title 'Middle-Aged Men in a Health Fitness Club';
      title2 'Data Courtesy of Dr. A. C. Linnerud, NC State Univ';
   run;
```
Output 30.1.1 Correlations among the Original Variables

<table>
<thead>
<tr>
<th></th>
<th>Weight</th>
<th>Waist</th>
<th>Pulse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight</td>
<td>1.0000</td>
<td>0.8702</td>
<td>-0.3658</td>
</tr>
<tr>
<td>Waist</td>
<td>0.8702</td>
<td>1.0000</td>
<td>-0.3529</td>
</tr>
<tr>
<td>Pulse</td>
<td>-0.3658</td>
<td>-0.3529</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Correlations Among the Exercises

<table>
<thead>
<tr>
<th></th>
<th>Chins</th>
<th>Situps</th>
<th>Jumps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chins</td>
<td>1.0000</td>
<td>0.6957</td>
<td>0.4958</td>
</tr>
<tr>
<td>Situps</td>
<td>0.6957</td>
<td>1.0000</td>
<td>0.6692</td>
</tr>
<tr>
<td>Jumps</td>
<td>0.4958</td>
<td>0.6692</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Correlations Between the Physiological Measurements and the Exercises

<table>
<thead>
<tr>
<th></th>
<th>Chins</th>
<th>Situps</th>
<th>Jumps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight</td>
<td>-0.3897</td>
<td>-0.4931</td>
<td>-0.2263</td>
</tr>
<tr>
<td>Waist</td>
<td>-0.5522</td>
<td>-0.6456</td>
<td>-0.1915</td>
</tr>
<tr>
<td>Pulse</td>
<td>0.1506</td>
<td>0.2250</td>
<td>0.0349</td>
</tr>
</tbody>
</table>

Output 30.1.1 displays the correlations among the original variables. The correlations between the physiological and exercise variables are moderate, the largest being –0.6456 between Waist and Situps. There are larger within-set correlations: 0.8702 between Weight and Waist, 0.6957 between Chins and Situps, and 0.6692 between Situps and Jumps.
Example 30.1: Canonical Correlation Analysis of Fitness Club Data

Output 30.1.2 Canonical Correlations and Multivariate Statistics

Middle-Aged Men in a Health Fitness Club
Data Courtesy of Dr. A. C. Linnerud, NC State Univ

The CANCORR Procedure

Canonical Correlation Analysis

<table>
<thead>
<tr>
<th>Canonical Correlation</th>
<th>Adjusted</th>
<th>Approximate</th>
<th>Squared Canonical Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.795608</td>
<td>0.754056</td>
<td>0.084197</td>
</tr>
<tr>
<td>2</td>
<td>0.200556</td>
<td>-0.076399</td>
<td>0.220188</td>
</tr>
<tr>
<td>3</td>
<td>0.072570</td>
<td></td>
<td>0.228208</td>
</tr>
</tbody>
</table>

Eigenvalues of $\text{Inv}(E) \ast H = \frac{\text{CanRsq}}{1-\text{CanRsq}}$

<table>
<thead>
<tr>
<th>Eigenvalue</th>
<th>Difference</th>
<th>Proportion</th>
<th>Cumulative</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.7247</td>
<td>0.9734</td>
<td>0.9734</td>
</tr>
<tr>
<td>2</td>
<td>0.0419</td>
<td>0.0237</td>
<td>0.9970</td>
</tr>
<tr>
<td>3</td>
<td>0.0053</td>
<td>0.0030</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Test of $H_0$: The canonical correlations in the current row and all that follow are zero

<table>
<thead>
<tr>
<th>Likelihood Ratio</th>
<th>Approximate</th>
<th>F Value</th>
<th>Num DF</th>
<th>Den DF</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.35039053</td>
<td>2.05</td>
<td>9</td>
<td>34.223</td>
<td>0.0635</td>
</tr>
<tr>
<td>2</td>
<td>0.95472266</td>
<td>0.18</td>
<td>4</td>
<td>30</td>
<td>0.9491</td>
</tr>
<tr>
<td>3</td>
<td>0.99473355</td>
<td>0.08</td>
<td>1</td>
<td>16</td>
<td>0.7748</td>
</tr>
</tbody>
</table>

Multivariate Statistics and F Approximations

$S=3$ $M=-0.5$ $N=6$

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
<th>F Value</th>
<th>Num DF</th>
<th>Den DF</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wilks' Lambda</td>
<td>0.35039053</td>
<td>2.05</td>
<td>9</td>
<td>34.223</td>
<td>0.0635</td>
</tr>
<tr>
<td>Pillai's Trace</td>
<td>0.67848151</td>
<td>1.56</td>
<td>9</td>
<td>48</td>
<td>0.1551</td>
</tr>
<tr>
<td>Hotelling-Lawley Trace</td>
<td>1.77194146</td>
<td>2.64</td>
<td>9</td>
<td>19.053</td>
<td>0.0357</td>
</tr>
<tr>
<td>Roy's Greatest Root</td>
<td>1.72473874</td>
<td>9.20</td>
<td>3</td>
<td>16</td>
<td>0.0009</td>
</tr>
</tbody>
</table>

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

As Output 30.1.2 shows, the first canonical correlation is 0.7956, which would appear to be substantially larger than any of the between-set correlations. The probability level for the null hypothesis that all the canonical correlations are zero in the population is only 0.0635, so no firm conclusions can be drawn. The remaining canonical correlations are not worthy of consideration, as can be seen from the probability levels and especially from the negative adjusted canonical correlations.
Because the variables are not measured in the same units, the standardized coefficients rather than the raw coefficients should be interpreted. The correlations given in the canonical structure matrices should also be examined.

**Output 30.1.3 Raw and Standardized Canonical Coefficients**

<table>
<thead>
<tr>
<th>Raw Canonical Coefficients for the Physiological Measurements</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Weight</td>
</tr>
<tr>
<td>Waist</td>
</tr>
<tr>
<td>Pulse</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Raw Canonical Coefficients for the Exercises</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Chins</td>
</tr>
<tr>
<td>Situps</td>
</tr>
<tr>
<td>Jumps</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Standardized Canonical Coefficients for the Physiological Measurements</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Weight</td>
</tr>
<tr>
<td>Waist</td>
</tr>
<tr>
<td>Pulse</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Standardized Canonical Coefficients for the Exercises</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Chins</td>
</tr>
<tr>
<td>Situps</td>
</tr>
<tr>
<td>Jumps</td>
</tr>
</tbody>
</table>

The first canonical variable for the physiological variables, displayed in **Output 30.1.3**, is a weighted difference of Waist (1.5793) and Weight (–0.7754), with more emphasis on Waist. The coefficient for Pulse is near 0. The correlations between Waist and Weight and the first canonical variable are both positive, 0.9254 for Waist and 0.6206 for Weight. Weight is therefore a suppressor variable, meaning that its coefficient and its correlation have opposite signs.

The first canonical variable for the exercise variables also shows a mixture of signs, subtracting Situps (–1.0540) and Chins (–0.3495) from Jumps (0.7164), with the most weight on Situps. All the correlations are negative, indicating that Jumps is also a suppressor variable.

It might seem contradictory that a variable should have a coefficient of opposite sign from that of its correlation with the canonical variable. In order to understand how this can happen, consider a simplified situation: predicting Situps from Waist and Weight by multiple regression. In informal terms, it seems plausible that obese people should do fewer sit-ups than skinny people. Assume that the men in the sample
do not vary much in height, so there is a strong correlation between Waist and Weight (0.8702). Examine the relationships between obesity and the independent variables:

- People with large waists tend to be more obese than people with small waists. Hence, the correlation between Waist and Situps should be negative.
- People with high weights tend to be more obese than people with low weights. Therefore, Weight should correlate negatively with Situps.
- For a fixed value of Weight, people with large waists tend to be shorter and more obese. Thus, the multiple regression coefficient for Waist should be negative.
- For a fixed value of Waist, people with higher weights tend to be taller and skinnier. The multiple regression coefficient for Weight should therefore be positive, of opposite sign from the correlation between Weight and Situps.

Therefore, the general interpretation of the first canonical correlation is that Weight and Jumps act as suppressor variables to enhance the correlation between Waist and Situps. This canonical correlation might be strong enough to be of practical interest, but the sample size is not large enough to draw definite conclusions.

The canonical redundancy analysis (Output 30.1.4) shows that neither of the first pair of canonical variables is a good overall predictor of the opposite set of variables, the proportions of variance explained being 0.2854 and 0.2584. The second and third canonical variables add virtually nothing, with cumulative proportions for all three canonical variables being 0.2969 and 0.2767.

**Output 30.1.4 Canonical Redundancy Analysis**

<table>
<thead>
<tr>
<th>Canonical Variable Number</th>
<th>Proportion</th>
<th>Cumulative Proportion</th>
<th>Canonical R-Square</th>
<th>Proportion</th>
<th>Cumulative Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.4508</td>
<td>0.4508</td>
<td>0.6330</td>
<td>0.2854</td>
<td>0.2854</td>
</tr>
<tr>
<td>2</td>
<td>0.2470</td>
<td>0.6978</td>
<td>0.0402</td>
<td>0.0099</td>
<td>0.2953</td>
</tr>
<tr>
<td>3</td>
<td>0.3022</td>
<td>1.0000</td>
<td>0.0053</td>
<td>0.0016</td>
<td>0.2969</td>
</tr>
</tbody>
</table>
The squared multiple correlations (Output 30.1.5) indicate that the first canonical variable of the physiological measurements has some predictive power for Chins (0.3351) and Situps (0.4233) but almost none for Jumps (0.0167). The first canonical variable of the exercises is a fairly good predictor of Waist (0.5421), a poorer predictor of Weight (0.2438), and nearly useless for predicting Pulse (0.0701).


Subject Index

CANCORR procedure
  canonical coefficients, 1866
  canonical redundancy analysis, 1866, 1876
  computational resources, 1884
  correction for means, 1876
  correlation, 1875
  eigenvalues, 1881
  eigenvalues and eigenvectors, 1869, 1885
  examples, 1867, 1889
  formulas, 1880
  input data set, 1875
  missing values, 1880
  OUT= data sets, 1881
  output data sets, 1876, 1881
  output table names, 1887
  OUTSTAT= data sets, 1876, 1882
  partial correlation, 1876, 1877, 1879
  principal components, relation to, 1881
  regression coefficients, 1875
  semipartial correlation, 1877
  singularity checking, 1877
  squared multiple correlation, 1877
  squared partial correlation, 1877
  squared semipartial correlation, 1877
  statistical methods used, 1866
  statistics computed, 1866
  suppressing output, 1876
  weighted product-moment correlation coefficients, 1880
  canonical correlation
    CANCORR procedure, 1865
    definition, 1866
    hypothesis tests, 1865
  canonical redundancy analysis
    CANCORR procedure, 1866, 1876
  canonical weights, 1866
  computational resources
    CANCORR procedure, 1884
  correction for means
    CANCORR procedure, 1876
  correlation
    CANCORR procedure, 1875
  eigenvalues and eigenvectors
    CANCORR procedure, 1869, 1885
  formulas
    CANCORR procedure, 1880

missing values
  CANCORR procedure, 1880

OUT= data sets
  CANCORR procedure, 1881
output data sets
  CANCORR procedure, 1876, 1881
output table names
  CANCORR procedure, 1887
OUTSTAT= data sets
  CANCORR procedure, 1876, 1882

partial canonical correlation, 1867
partial correlation
  CANCORR procedure, 1876, 1877, 1879

redundancy analysis
  CANCORR procedure, 1867
regression coefficients
  CANCORR procedure, 1875

semipartial correlation
  CANCORR procedure, 1877
  singularity checking
    CANCORR procedure, 1877
  squared multiple correlation
    CANCORR procedure, 1877
  squared partial correlation
    CANCORR procedure, 1877
  squared semipartial correlation
    CANCORR procedure, 1877
  suppressing output
    CANCORR procedure, 1876

weighted product-moment correlation coefficients
  CANCORR procedure, 1880
Syntax Index

ALL option
  PROC CANCORR statement, 1874
B option
  PROC CANCORR statement, 1875
BY statement
  CANCORR procedure, 1879
C option
  PROC CANCORR statement, 1875
CANCORR procedure
  syntax, 1873
CANCORR procedure, BY statement, 1879
CANCORR procedure, FREQ statement, 1879
CANCORR procedure, PARTIAL statement, 1879
CANCORR procedure, PROC CANCORR statement, 1873
  ALL option, 1874
  B option, 1875
  C option, 1875
  CLB option, 1875
  CORR option, 1875
  CORRB option, 1875
  DATA= option, 1875
  EDF= option, 1875
  INT option, 1875
  MSTAT= option, 1875
  NCAN= option, 1875
  NOINT option, 1876
  NOPRINT option, 1876
  OUT= option, 1876
  OUTSTAT= option, 1876
  PARPREFIX= option, 1877
  PCORR option, 1876
  PPREFIX= option, 1877
  PROBT option, 1876
  RDF= option, 1876
  RED option, 1876
  REDUNDANCY option, 1876
  S option, 1877
  SEB option, 1877
  SHORT option, 1877
  SIMPLE option, 1877
  SING= option, 1877
  SINGULAR= option, 1877
  SMC option, 1877
  SPCORR option, 1877
  SQPCORR option, 1877
  SQSPCORR option, 1877
  STB option, 1877
  T option, 1877
  VDEP option, 1878
  VN= option, 1878
  VNAME= option, 1878
  VP= option, 1878
  VPREFIX= option, 1878
  VREG option, 1878
  WDEP option, 1878
  WN= option, 1878
  WNAME= option, 1878
  WP= option, 1878
  WPREFIX= option, 1878
  WREG option, 1878
CANCORR procedure, VAR statement, 1880
CANCORR procedure, WEIGHT statement, 1880
CANCORR procedure, WITH statement, 1880
CLB option
  PROC CANCORR statement, 1875
CORR option
  PROC CANCORR statement, 1875
CORRB option
  PROC CANCORR statement, 1875
DATA= option
  PROC CANCORR statement, 1875
EDF= option
  PROC CANCORR statement, 1875
INT option
  PROC CANCORR statement, 1875
NCAN= option
  PROC CANCORR statement, 1875
NOINT option
  PROC CANCORR statement, 1876
NOPRINT option
  PROC CANCORR statement, 1876
OUT= option
  PROC CANCORR statement, 1876
OUTSTAT= option, 1876
PARPREFIX= option, 1877
PCORR option, 1876
PPREFIX= option, 1877
PROBT option, 1876
RDF= option, 1876
RED option, 1876
REDUNDANCY option, 1876
S option, 1877
SEB option, 1877
SHORT option, 1877
SIMPLE option, 1877
SING= option, 1877
SINGULAR= option, 1877
SMC option, 1877
SPCORR option, 1877
SQPCORR option, 1877
SQSPCORR option, 1877
VNAME= option, 1878
VPREFIX= option, 1878
VREG option, 1878
WNAME= option, 1878
WPREFIX= option, 1878
WREG option, 1878
PROBT option
  PROC CANCORR statement, 1876
PROC CANCORR statement, see CANCORR procedure

RDF= option
  PROC CANCORR statement, 1876
RED option
  PROC CANCORR statement, 1876
REDUNDANCY option
  PROC CANCORR statement, 1876
RPREFIX= option
  PROC CANCORR statement, 1877

S option
  PROC CANCORR statement, 1877
SEB option
  PROC CANCORR statement, 1877
SHORT option
  PROC CANCORR statement, 1877
SIMPLE option
  PROC CANCORR statement, 1877
SING= option
  PROC CANCORR statement, 1877
SINGULAR= option
  PROC CANCORR statement, 1877
SMC option
  PROC CANCORR statement, 1877
SPCORR option
  PROC CANCORR statement, 1877
SQPCORR option
  PROC CANCORR statement, 1877
SQSPCORR option
  PROC CANCORR statement, 1877
STB option
  PROC CANCORR statement, 1877

T option
  PROC CANCORR statement, 1877
VDEP option
  PROC CANCORR statement, 1878
VN= option
  PROC CANCORR statement, 1878
VNAME= option
  PROC CANCORR statement, 1878
VP= option
  PROC CANCORR statement, 1878
VPREFIX= option
  PROC CANCORR statement, 1878
VREG option
  PROC CANCORR statement, 1878

WDEP option
  PROC CANCORR statement, 1878
WN= option
  PROC CANCORR statement, 1878
WNAME= option
  PROC CANCORR statement, 1878
WP= option
  PROC CANCORR statement, 1878
WPREFIX= option
  PROC CANCORR statement, 1878
WREG option
  PROC CANCORR statement, 1878