CHAPTER 12

The CORR Procedure

Overview

The CORR procedure is a statistical procedure for numeric random variables that computes Pearson correlation coefficients, three nonparametric measures of association, and the probabilities associated with these statistics. The correlation statistics include

- Pearson product-moment and weighted product-moment correlation
- Spearman rank-order correlation
- Kendall's tau-b

- □ Hoeffding's measure of dependence, D
- □ Pearson, Spearman, and Kendall partial correlation.

PROC CORR also computes Cronbach's coefficient alpha for estimating reliability.

The default correlation analysis includes descriptive statistics, Pearson correlation statistics, and probabilities for each analysis variable. You can save the correlation statistics in a SAS data set for use with other statistical and reporting procedures.

Output 12.1 on page 274 is the simplest form of PROC CORR output. Pearson correlation statistics are computed for all numeric variables from a study investigating the effect of exercise on physical fitness. The statements that produce the output follow:

```
options pagesize=60;
proc corr data=fitness;
run;
```

Output 12.1 Simple Correlation Analysis for a Fitness Study Using PROC CORR

```
                              The SAS System                              1

                            The CORR Procedure

               4  Variables:  Age      Weight   Runtime  Oxygen


                            Simple Statistics

Variable          N         Mean      Std Dev          Sum     Minimum     Maximum

Age              30     47.56667      5.26330         1427    38.00000    57.00000
Weight           30     77.70500      8.34152         2331    59.08000    91.63000
Runtime          29     10.61448      1.41655    307.82000     8.17000    14.03000
Oxygen           29     47.06445      5.32129         1365    37.38800    60.05500


                    Pearson Correlation Coefficients
                      Prob > |r| under H0: Rho=0
                        Number of Observations

                         Age        Weight       Runtime        Oxygen

       Age           1.00000      -0.21777       0.19528      -0.32899
                                    0.2477        0.3100        0.0814
                          30            30            29            29

       Weight       -0.21777       1.00000       0.15155      -0.19900
                      0.2477                      0.4326        0.3007
                          30            30            29            29

       Runtime       0.19528       0.15155       1.00000      -0.78346
                      0.3100        0.4326                      <.0001
                          29            29            29            28

       Oxygen       -0.32899      -0.19900      -0.78346       1.00000
                      0.0814        0.3007        <.0001
                          29            29            28            29
```

Output 12.2 on page 275 and Output 12.3 on page 276 illustrate the use of PROC CORR to calculate partial correlation statistics for the fitness study and to store the results in an output data set. The statements that produce the analysis also

- □ suppress the descriptive statistics
- □ select and label analysis variables

- □ exclude all observations with missing values
- □ calculate the partial covariance matrix
- □ calculate three types of partial correlation coefficients
- □ generate an output data set that contains Pearson correlation statistics and print the output data set.

For an explanation of the program that produces the following output, see Example 4 on page 309.

Output 12.2 Customized Correlation Analysis with Partial Covariances and Correlation Statistics

```
                  Partial Correlations for a Fitness and Exercise Study                              1

                                    The CORR Procedure

                      1 Partial Variables:  Age
                      3         Variables:  Weight   Oxygen   Runtime


                            Partial Covariance Matrix, DF = 26

                                         Weight            Oxygen           Runtime

 Weight       Wt in kg              72.43742055      -12.75113194        2.06766763
 Oxygen       O2 use               -12.75113194       27.01654904       -5.59370556
 Runtime      1.5 mi in minutes      2.06766763       -5.59370556        1.94512451


                     Pearson Partial Correlation Coefficients, N = 28
                            Prob > |r| under H0: Partial Rho=0

                                 Weight        Oxygen       Runtime

             Weight             1.00000      -0.28824       0.17419
             Wt in kg                          0.1448        0.3849

             Oxygen            -0.28824       1.00000      -0.77163
             O2 use              0.1448                      <.0001

             Runtime            0.17419      -0.77163       1.00000
             1.5 mi in minutes   0.3849        <.0001


                    Spearman Partial Correlation Coefficients, N = 28
                            Prob > |r| under H0: Partial Rho=0

                                 Weight        Oxygen       Runtime

             Weight             1.00000      -0.16407       0.08708
             Wt in kg                          0.4135        0.6658

             Oxygen            -0.16407       1.00000      -0.67112
             O2 use              0.4135                      0.0001

             Runtime            0.08708      -0.67112       1.00000
             1.5 mi in minutes   0.6658        0.0001


             Kendall Partial Tau b Correlation Coefficients, N = 28

                                 Weight        Oxygen       Runtime

             Weight             1.00000      -0.09021       0.02854
             Wt in kg

             Oxygen            -0.09021       1.00000      -0.52158
             O2 use

             Runtime            0.02854      -0.52158       1.00000
             1.5 mi in minutes
```

Output 12.3 Output Data Set with Pearson Partial Correlation Statistics

```
          Pearson Correlation Statistics Using the PARTIAL Statement         2
                       Output Data Set from PROC CORR

             _TYPE_    _NAME_       Weight      Oxygen     Runtime

              COV      Weight      72.4374    -12.7511      2.0677
              COV      Oxygen     -12.7511     27.0165     -5.5937
              COV      Runtime      2.0677     -5.5937      1.9451
              MEAN                  0.0000      0.0000      0.0000
              STD                   8.5110      5.1977      1.3947
              N                    28.0000     28.0000     28.0000
              CORR     Weight       1.0000     -0.2882      0.1742
              CORR     Oxygen      -0.2882      1.0000     -0.7716
              CORR     Runtime      0.1742     -0.7716      1.0000
```

Procedure Syntax

Tip: Supports the Output Delivery System, see "Output Delivery System" on page 19

Reminder: You can use the ATTRIB, FORMAT, LABEL, and WHERE statements. See Chapter 3, "Statements with the Same Function in Multiple Procedures," for details. You can also use any global statements as well. See Chapter 2, "Fundamental Concepts for Using Base SAS Procedures," for a list.

PROC CORR <*option(s)*>;
  BY <DESCENDING> *variable-1*<…<DESCENDING> *variable-n*>
    <NOTSORTED>;
  FREQ *frequency-variable*;
  PARTIAL *variable(s)*;
  VAR *variable(s)*;
  WEIGHT *weight-variable*;
  WITH *variable(s)*;

To do this	Use this statement
Produce separate correlation analyses for each BY group	BY
Identify a variable whose values represent the frequency of each observation	FREQ
Identify controlling variables to compute Pearson, Spearman, or Kendall partial correlation coefficients	PARTIAL
Identify variables to correlate and their order in the correlation matrix	VAR

To do this	Use this statement
Identify a variable whose values weight each observation to compute Pearson weight product-moment correlation	WEIGHT
Compute correlations for specific combinations of variables	WITH

PROC CORR Statement

PROC CORR <*option(s)*>;

To do this	Use this option
Specify the input data set	DATA=
Create output data sets	
Specify an output data set to contain Hoeffding's D statistics	OUTH=
Specify an output data set to contain Kendall correlations	OUTK=
Specify an output data set to contain Pearson correlations	OUTP=
Specify an output data set to contain Spearman correlations	OUTS=
Control statistical analysis	
Exclude observations with nonpositive weight values from the analysis	EXCLNPWGT
Request Hoeffding's measure of dependence, D	HOEFFDING
Request Kendall's tau-b	KENDALL
Request Pearson product-moment correlation	PEARSON
Request Spearman rank-order correlation	SPEARMAN
Control Pearson correlation statistics	
Compute Cronbach's coefficient alpha	ALPHA
Compute covariances	COV
Compute corrected sums of squares and crossproducts	CSSCP
Exclude missing values	NOMISS
Specify singularity criterion	SINGULAR=
Compute sums of squares and crossproducts	SSCP
Specify the divisor for variance calculations	VARDEF=
Control printed output	
Specify the number and order of correlation coefficients	BEST=
Suppress Pearson correlations	NOCORR
Suppress all printed output	NOPRINT
Suppress significance probabilities	NOPROB

To do this	Use this option
Suppress descriptive statistics	NOSIMPLE
Change the order of correlation coefficients	RANK

Options

ALPHA
calculates and prints Cronbach's coefficient alpha. PROC CORR computes separate coefficients using raw and standardized values (scaling the variables to a unit variance of 1). For each VAR statement variable, PROC CORR computes the correlation between the variable and the total of the remaining variables. It also computes Cronbach's coefficient alpha using only the remaining variables.

Main discussion: "Cronbach's Coefficient Alpha" on page 293

Restriction: If you use a WITH statement, ALPHA is invalid.

Interaction: ALPHA invokes PEARSON.

Interaction: If you specify OUTP=, the output data set also contains six observations with Cronbach's coefficient alpha.

Interaction: When you use the PARTIAL statement, PROC CORR calculates Cronbach's coefficient alpha for partialled variables.

See also: OUTP= option

Featured in: Example 3 on page 306

BEST=*n*
prints *n* correlation coefficients for each variable. Correlations are ordered from highest to lowest in absolute value. Otherwise, PROC CORR prints correlations in a rectangular table using the variable names as row and column labels.

Interaction: When you specify HOEFFDING, PROC CORR prints the D statistics in order from highest to lowest.

Range: 1 to the maximum number of variables

COV
calculates and prints covariances.

Interaction: COV invokes PEARSON.

Interaction: If you specify OUTP=, the output data set contains the covariance matrix and the _TYPE_ variable value is COV.

Interaction: When you use the PARTIAL statement, PROC CORR computes a partial covariance matrix.

See also: OUTP= option

Featured in: Example 2 on page 303 and Example 4 on page 309

CSSCP
prints the corrected sums of squares and crossproducts.

Interaction: CSSCP invokes PEARSON.

Interaction: If you specify OUTP=, the output data set contains a CSSCP matrix and the _TYPE_ variable value is CSSCP. If you use a PARTIAL statement, the output data set contains a partial CSSCP matrix.

Interaction: When you use a PARTIAL statement, PROC CORR prints both an unpartial and a partial CSSCP matrix.

See also: OUTP= option

DATA=*SAS-data-set*
specifies the input SAS data set.

Main discussion: "Input Data Sets" on page 18

EXCLNPWGT
excludes observations with nonpositive weight values (zero or negative) from the analysis. By default, PROC CORR treats observations with negative weights like those with zero weights and counts them in the total number of observations.

Requirement: You must use a WEIGHT statement.

See also: "WEIGHT Statement" on page 285

HOEFFDING
calculates and prints Hoeffding's D statistics. This D statistic is 30 times larger than the usual definition and scales the range between -0.5 and 1 so that only large positive values indicate dependence.

Main discussion: "Hoeffding's Measure of Dependence, D" on page 291

Restriction: When you use a WEIGHT or PARTIAL statement, HOEFFDING is invalid.

Featured in: Example 1 on page 300

KENDALL
calculates and prints Kendall tau-b coefficients based on the number of concordant and discordant pairs of observations. Kendall's tau-b ranges from -1 to 1.

Main discussion: "Kendall's tau-b" on page 290

Restriction: When you use a WEIGHT statement, KENDALL is invalid.

Interactions: When you use a PARTIAL statement, probability values for Kendall's partial tau-b are not available.

Featured in: Example 4 on page 309

NOCORR
suppresses calculating and printing of Pearson correlations.

Interaction: If you specify OUTP=, the data set type remains CORR. To change the data set type to COV, CSSCP, or SSCP, use the TYPE= data set option.

See also: "Output Data Sets" on page 298

Featured in: Example 3 on page 306

NOMISS
excludes observations with missing values from the analysis. Otherwise, PROC CORR computes correlation statistics using all the nonmissing pairs of variables.

Main discussion: "Missing Values" on page 297

Tip: Using NOMISS is computationally more efficient.

Featured in: Example 3 on page 306

NOPRINT
suppresses all printed output.

Tip: Use NOPRINT when you want to create an output data set only.

NOPROB
suppresses printing the probabilities associated with each correlation coefficient.

NOSIMPLE
suppresses printing simple descriptive statistics for each variable. However, if you request an output data set, the output data set still contains simple descriptive statistics for the variables.

Featured in: Example 2 on page 303

OUTH=*output-data-set*
creates an output data set containing Hoeffding's D statistics. The contents of the output data set are similar to the OUTP= data set.

Main discussion: "Output Data Sets" on page 298

Interaction: OUTH= invokes HOEFFDING.

OUTK=*output-data-set*
creates an output data set containing Kendall correlation statistics. The contents of the output data set are similar to the OUTP= data set.

Main discussion: "Output Data Sets" on page 298

Interaction: OUTK= option invokes KENDALL.

OUTP=*output-data-set*
creates an output data set containing Pearson correlation statistics. This data set also includes means, standard deviations, and the number of observations. The value of the _TYPE_ variable is CORR.

Main discussion: "Output Data Sets" on page 298

Interaction: OUTP= invokes PEARSON.

Interaction: If you specify ALPHA, the output data set also contains six observations with Cronbach's coefficient alpha.

Featured in: Example 4 on page 309

OUTS=*SAS-data-set*
creates an output data set containing Spearman correlation statistics. The contents of the output data set are similar to the OUTP= data set.

Main discussion: "Output Data Sets" on page 298

Interaction: OUTS= invokes SPEARMAN.

PEARSON
calculates and prints Pearson product-moment correlations when you use the HOEFFDING, KENDALL, or SPEARMAN option. If you omit the correlation type, PROC CORR automatically produces Pearson correlations. The correlations range from -1 to 1.

Main discussion: "Pearson Product-Moment Correlation" on page 289

Featured in: Example 1 on page 300

RANK
prints the correlation coefficients for each variable. Correlations are ordered from highest to lowest in absolute value. Otherwise, PROC CORR prints correlations in a rectangular table using the variable names as row and column labels.

Interaction: If you use HOEFFDING, PROC CORR prints the D statistics in order from highest to lowest.

SINGULAR=*p*
specifies the criterion for determining the singularity of a variable when you use a PARTIAL statement. A variable is considered singular if its corresponding diagonal element after Cholesky decomposition has a value less than p times the original unpartialled corrected sum of squares of that variable.

Main discussion: "Partial Correlation" on page 291

Default: 1E-8

Range: between 0 and 1

SPEARMAN
calculates and prints Spearman correlation coefficients based on the ranks of the variables. The correlations range from -1 to 1.

Main discussion: "Spearman Rank-Order Correlation" on page 290

Restriction: When you specify a WEIGHT statement, SPEARMAN is invalid.
Example 1 on page 300

Featured in:

SSCP

prints the sums of squares and crossproducts.

Interaction: SSCP invokes PEARSON.

Interaction: When you specify OUTP=, the output data set contains a SSCP matrix and the _TYPE_ variable value is SSCP. If you use a PARTIAL statement, the output data set does not contain an SSCP matrix.

Interaction: When you use a PARTIAL statement, PROC CORR prints the unpartial SSCP matrix.

Featured in: Example 2 on page 303

VARDEF=*divisor*

specifies the divisor to use in the calculation of variances, standard deviations, and covariances.

Table 12.1 on page 282 shows the possible values for *divisor* and associated divisors where *k* is the number of PARTIAL statement variables.

Table 12.1 Possible Values for VARDEF=

Value	Divisor	Formula
DF	degrees of freedom	n - k - 1
N	number of observations	n
WDF	sum of weights minus one	$(\Sigma_i w_i)$ - k - 1
WEIGHT | WGT	sum of weights	$\Sigma_i w_i$

The procedure computes the variance as $CSS/divisor$, where CSS is the corrected sums of squares and equals $\sum (x_i - \overline{x})^2$. When you weight the analysis variables, CSS equals $\sum w_i (x_i - \overline{x}_w)^2$, where $\overline{x}_w$ is the weighted mean.

Default: DF

Tip: When you use the WEIGHT statement and VARDEF=DF, the variance is an estimate of σ^2, where the variance of the ith observation is $var(x_i) = \sigma^2/w_i$ and w_i is the weight for the ith observation. This yields an estimate of the variance of an observation with unit weight.

Tip: When you use the WEIGHT statement and VARDEF=WGT, the computed variance is asymptotically (for large n) an estimate of $\sigma^2/\overline{w}$, where $\overline{w}$ is the average weight. This yields an asymptotic estimate of the variance of an observation with average weight.

Main discussion: Weighted statistics "Example" on page 74.

BY Statement

Calculates separate correlation statistics for each BY group.

Main discussion: "BY" on page 68

BY <DESCENDING> *variable-1* <…<DESCENDING> *variable-n*><NOTSORTED>;

Required Arguments

variable
specifies the variable that the procedure uses to form BY groups. You can specify more than one variable. If you do not use the NOTSORTED option in the BY statement, the observations in the data set must either be sorted by all the variables that you specify, or they must be indexed appropriately. Variables in a BY statement are called *BY variables*.

Options

DESCENDING
specifies that the observations are sorted in descending order by the variable that immediately follows the word DESCENDING in the BY statement.

NOTSORTED
specifies that observations are not necessarily sorted in alphabetic or numeric order. The observations are grouped in another way, for example, chronological order.

The requirement for ordering or indexing observations according to the values of BY variables is suspended for BY-group processing when you use the NOTSORTED option. In fact, the procedure does not use an index if you specify NOTSORTED. The procedure defines a BY group as a set of contiguous observations that have the same values for all BY variables. If observations with the same values for the BY variables are not contiguous, the procedure treats each contiguous set as a separate BY group.

FREQ Statement

Treats observations as if they appear multiple times in the input data set.

Tip: The effects of the FREQ and WEIGHT statements are similar except when calculating degrees of freedom.

See also: For an example that uses the FREQ statement, see "FREQ" on page 70

FREQ *variable*;

Required Arguments

variable
specifies a numeric variable whose value represents the frequency of the observation. If you use the FREQ statement, the procedure assumes that each observation represents n observations, where n is the value of *variable*. If n is not an integer, the SAS System truncates it. If n is less than 1 or is missing, the procedure does not use that observation to calculate statistics.

The sum of the frequency variable represents the total number of observations.

PARTIAL Statement

Computes Pearson partial correlation, Spearman partial rank-order correlation, or Kendall's partial tau-b.

Restriction: Not valid with the HOEFFDING option.

Interaction: Invokes the NOMISS option to exclude all observations with missing values.

Main discussion: "Partial Correlation" on page 291

Featured in: Example 4 on page 309

PARTIAL *variable(s)*;

Required Arguments

variable(s)
identifies one or more variables to use in the calculation of partial correlation statistics.

Details

- If you use the PEARSON option, PROC CORR also prints the partial variance and standard deviation for each VAR or WITH statement variable.
- If you use the KENDALL option, PROC CORR cannot compute probability values for Kendall's partial tau-b.

VAR Statement

Specifies the variables to use to calculate correlation statistics.

Default: If you omit this statement, PROC CORR computes correlations for all numeric variables not listed in the other statements.

Featured in: Example 1 on page 300 and Example 2 on page 303

VAR *variable(s)*;

Required Arguments

variable(s)
identifies one or more variables to use in the calculation of correlation coefficients.

WEIGHT Statement

Specifies weights for the analysis variables in the calculation of Pearson weighted product-moment correlation.

Restriction: Not valid with the HOEFFDING, KENDALL, or SPEARMAN option.

See also: For information on calculating weighted correlations, see "Pearson Product-Moment Correlation" on page 289.

WEIGHT *variable*;

Required Arguments

variable
specifies a numeric variable to use to compute weighted product-moment correlation coefficients. The variable does not have to be an integer. If the value of the weight variable is

Weight value…	PROC CORR…
0	counts the observation in the total number of observations
less than 0	converts the value to zero and counts the observation in the total number of observations
missing	excludes the observation

To exclude observations that contain negative and zero weights from the analysis, use EXCLNPWGT. Note that most SAS/STAT procedures, such as PROC GLM, exclude negative and zero weights by default.

Tip: When you use the WEIGHT statement, consider which value of the VARDEF= option is appropriate. See the discussion of the VARDEF= option on page 282 for more information.

Note: Prior to Version 8 of the SAS System, the procedure did not exclude the observations with missing weights from the count of observations. △

WITH Statement

Determines the variables to use in conjunction with the VAR statement variables to calculate limited combinations of correlation coefficients.

Restriction: Not valid with the ALPHA option.

Featured in: Example 2 on page 303

WITH *variable(s)*;

Required Argument

variable(s)

lists one or more variables to obtain correlations for specific combinations of variables. The WITH statement variables appear down the side of the correlation matrix and the VAR statement variables appear across the top of the correlation matrix. PROC CORR computes the following correlations for the VAR statement variables A and B and the WITH statement variables X, Y, and Z:

X and A	X and B
Y and A	Y and B
Z and A	Z and B

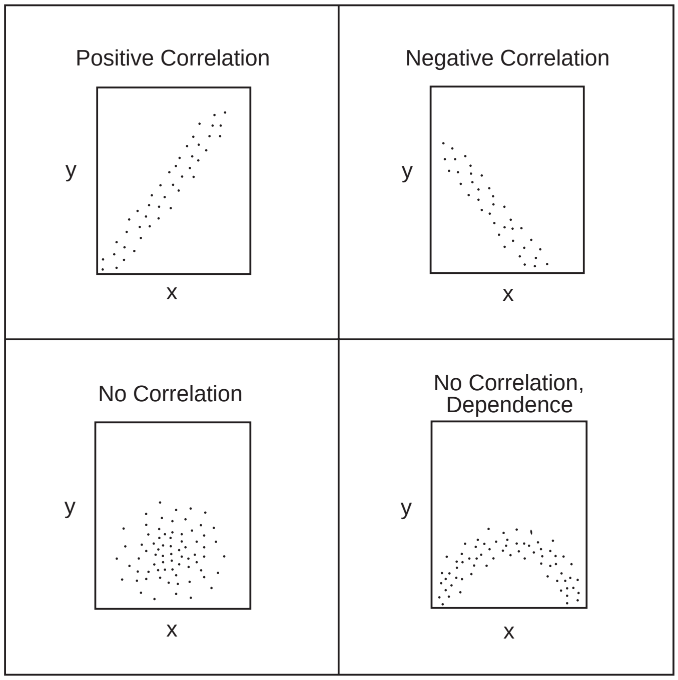
Concepts

Interpreting Correlation Coefficients

Correlation coefficients contain information on both the strength and direction of a linear relationship between two numeric random variables. If one variable x is an exact linear function of another variable y, a positive relationship exists when the correlation is 1 and an inverse relationship exists when the correlation is -1. If there is no linear predictability between the two variables, the correlation is 0. If the variables are normal and correlation is 0, the two variables are independent. However, correlation does not imply causality because, in some cases, an underlying causal relationship may exist.

The scatterplots in Figure 12.1 on page 287 depict the relationship between two numeric random variables.

Figure 12.1 Examining Correlations Using Scatterplots



When the relationship between two variables is nonlinear or when outliers are present, the correlation coefficient incorrectly estimates the strength of the relationship. Plotting the data before computing a correlation coefficient enables you to verify the linear relationship and to identify the potential outliers.

Determining Computer Resources

The only factor limiting the number of variables that you can analyze is the amount of available memory. The computer resources that PROC CORR requires depend on which statements and options you specify. To determine the computer resources that you need, use

N	number of observations in the data set.
C	number of correlation types (1 to 4).
V	number of VAR statement variables.
W	number of WITH statement variables.
P	number of PARTIAL statement variables.

so that

T=	V+W+P	
K=	V*W	when W>0
	V*(V+1)/2	when W=0
L=	K	when P=0
	T*(T+1)/2	when P>0

For small N and large K, the CPU time varies as K for all types of correlations. For large N, the CPU time depends on the type of correlation. To calculate CPU time use

K*N	with PEARSON (default)
T*N*log N	with SPEARMAN
K*N*log N	with HOEFFDING or KENDALL

You can reduce CPU time by specifying NOMISS. Without NOMISS, processing is much faster when most observations do not contain missing values.

The options and statements you use in the procedure require different amounts of storage to process the data. For Pearson correlations, the amount of temporary storage in bytes (M) is

40T+16L	with NOMISS and NOSIMPLE
40T+16L+56T	with NOMISS
40T+16L+56K	with NOSIMPLE
40T+16L+56K+56T	with no options

Using a PARTIAL statement increases the amount of temporary storage by 12T bytes. Using the ALPHA option increases the amount of temporary storage by 32V+16 bytes.

The following example uses a PARTIAL statement, which invokes NOMISS.

```
proc corr;
   var x1 x2;
   with y1 y2 y3;
   partial z1;
```

Therefore, using 40T+16L+56T+12T, the minimum temporary storage equals 984 bytes (T=2+3+1 and L=T(T+1)/2).

Using the SPEARMAN, KENDALL, or HOEFFDING option requires additional temporary storage for each observation. For the most time-efficient processing, the amount of temporary storage in bytes is

40T+8K+8L*C+12T*N+28N+QS+QP+QK

where

QS=	0	with NOSIMPLE
	68T	otherwise
QP=	56K	with PEARSON and without NOMISS
	0	otherwise
QK =	32N	with KENDALL or HOEFFDING
	0	otherwise.

The following example uses KENDALL:

```
proc corr kendall;
   var x1 x2 x3;
```

Therefore, the minimum temporary storage in bytes is

40*3+8*6+8*6*1+12*3N+28N+3*68+32N = 420+96N

where N is the number of observations.

If M bytes are not available, PROC CORR must process the data multiple times to compute all the statistics. This reduces the minimum temporary storage you need by

12(T–2)N bytes. When this occurs, PROC CORR prints a note suggesting a larger memory region.

Statistical Computations

PROC CORR computes several parametric and nonparametric correlation statistics as measures of association. The formulas for computing these measures and the associated probabilities follow.

Pearson Product-Moment Correlation

The Pearson product-moment correlation is a parametric measure of association for two continuous random variables. The formula for the true Pearson product-moment correlation, denoted ρ_{xy}, is

$$\begin{aligned}\rho_{xy} &= \frac{\mathrm{cov}\left(x,y\right)}{\sqrt{\mathrm{var}\left(x\right)\mathrm{var}\left(y\right)}} \\ &= \frac{\mathrm{E}\left(\left(x-\mathrm{E}x\right)\left(y-\mathrm{E}y\right)\right)}{\sqrt{\mathrm{E}\left(x-\mathrm{E}x\right)^2\mathrm{E}\left(y-\mathrm{E}y\right)^2}}\end{aligned}$$

The sample correlation, such as a Pearson product-moment correlation or weighted product-moment correlation, estimates the true correlation. The formula for the Pearson product-moment correlation is

$$r_{xy} = \frac{\sum\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sqrt{\sum\left(x_i-\bar{x}\right)^2\sum\left(y_i-\bar{y}\right)^2}}$$

where $\bar{x}$ is the sample mean of x and $\bar{y}$ is the sample mean of y.

The formula for a weighted Pearson product-moment correlation is

$$r_{xy} = \frac{\sum w_i\left(x_i-\bar{x}_w\right)\left(y_i-\bar{y}_w\right)}{\sqrt{\sum w_i\left(x_i-\bar{x}_w\right)^2\sum w_i\left(y_i-\bar{y}_w\right)^2}}$$

where

$$\begin{aligned}\bar{x}_w &= \sum w_i x_i / \sum w_i \\ \bar{y}_w &= \sum w_i y_i / \sum w_i\end{aligned}$$

Note that $\bar{x}_w$ is the weighted mean of x, $\bar{y}_w$ is the weighted mean of y, and w_i is the weight.

When one variable is dichotomous (0,1) and the other variable is continuous, a Pearson correlation is equivalent to a point biserial correlation. When both variables are dichotomous, a Pearson correlation coefficient is equivalent to the phi coefficient.

Spearman Rank-Order Correlation

Spearman rank-order correlation is a nonparametric measure of association based on the rank of the data values. The formula is

$$\theta = \frac{\sum \left(\mathrm{R}_i - \bar{\mathrm{R}}\right)\left(\mathrm{S}_i - \bar{\mathrm{S}}\right)}{\sqrt{\sum \left(\mathrm{R}_i - \bar{\mathrm{R}}\right)^2 \sum \left(\mathrm{S}_i - \bar{\mathrm{S}}\right)^2}}$$

where R_i is the rank of the $i\mathrm{th}$ x value, S_i is the rank of the $i\mathrm{th}$ y value, $\bar{\mathrm{R}}$ is the mean of the R_i values, and $\bar{\mathrm{S}}$ is the mean of the S_i values.

PROC CORR computes the Spearman's correlation by ranking the data and using the ranks in the Pearson product-moment correlation formula. In case of ties, the averaged ranks are used.

Kendall's tau-b

Kendall's tau-b is a nonparametric measure of association based on the number of concordances and discordances in paired observations. Concordance occurs when paired observations vary together, and discordance occurs when paired observations vary differently. The formula for Kendall's tau-b is

$$\tau = \frac{\sum_{i<j} \mathrm{sgn}\left(x_i - x_j\right) \mathrm{sgn}\left(y_i - y_j\right)}{\sqrt{\left(\mathrm{T}_0 - \mathrm{T}_1\right)\left(\mathrm{T}_0 - \mathrm{T}_2\right)}}$$

where

$$\begin{aligned} \mathrm{T}_0 &= n\left(n-1\right)/2 \\ \mathrm{T}_1 &= \sum t_i\left(t_i - 1\right)/2 \\ \mathrm{T}_2 &= \sum u_i\left(u_i - 1\right)/2 \end{aligned}$$

and where t_i is the number of tied x values in the $i\mathrm{th}$ group of tied x values, u_i is the number of tied y values in the $i\mathrm{th}$ group of tied y values, f is the number of observations, and sgn(z) is defined as

$$\mathrm{sgn}\left(z\right) = \begin{cases} 1 & \text{if } z > 0 \\ 0 & \text{if } z = 0 \\ -1 & \text{if } z < 0 \end{cases}$$

PROC CORR computes Kendall's correlation by ranking the data and using a method similar to Knight (1966). The data are double sorted by ranking observations according to values of the first variable and reranking the observations according to values of the second variable. PROC CORR computes Kendall's tau-b from the number of interchanges of the first variable and corrects for tied pairs (pairs of observations with equal values of X or equal values of Y).

Hoeffding's Measure of Dependence, D

Hoeffding's measure of dependence, D, is a nonparametric measure of association that detects more general departures from independence. The statistic approximates a weighted sum over observations of chi-square statistics for two-by-two classification tables (Hoeffding 1948). Each set of (x, y) values are cut points for the classification. The formula for Hoeffding's D is

$$\mathrm{D} = 30\,\frac{(n-2)(n-3)\,\mathrm{D}_1 + \mathrm{D}_2 - 2(n-2)\,\mathrm{D}_3}{n(n-1)(n-2)(n-3)(n-4)}$$

where

$$\mathrm{D}_1 = \sum_i (\mathrm{Q}_i - 1)(\mathrm{Q}_i - 2)$$
$$\mathrm{D}_2 = \sum_i (\mathrm{R}_i - 1)(\mathrm{R}_i - 2)(\mathrm{S}_i - 1)(\mathrm{S}_i - 2)$$
$$\mathrm{D}_3 = \sum_i (\mathrm{R}_i - 2)(\mathrm{S}_i - 2)(\mathrm{Q}_i - 1)$$

R_i is the rank of x_i, S_i is the rank of y_i, and Q_i (also called the bivariate rank) is 1 plus the number of points with both x and y values less than the $i\mathrm{th}$ point. A point that is tied on only the x value or y value contributes 1/2 to Q_i if the other value is less than the corresponding value for the $i\mathrm{th}$ point. A point that is tied on both x and y contributes 1/4 to Q_i.

PROC CORR obtains the Q_i values by first ranking the data. The data are then double sorted by ranking observations according to values of the first variable and reranking the observations according to values of the second variable. Hoeffding's D statistic is computed using the number of interchanges of the first variable.

When no ties occur among data set observations, the D statistic values are between -0.5 and 1, with 1 indicating complete dependence. However, when ties occur, the D statistic may result in a smaller value. That is, for a pair of variables with identical values, the Hoeffding's D statistic may be less than 1. With a large number of ties in a small data set, the D statistic may be less than -0.5 . For more information on Hoeffding's D, see Hollander and Wolfe (1973, p. 228).

Partial Correlation

A partial correlation measures the strength of a relationship between two variables, while controlling the effect of one or more additional variables. The Pearson partial correlation for a pair of variables may be defined as the correlation of errors after regression on the controlling variables. Let $\mathbf{y} = (y_1, y_2, \ldots, y_v)$ be the set of variables to correlate. Also let $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ be sets of regression parameters and $\mathbf{z}$ be the set of controlling variables, where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_v)$, $\boldsymbol{\beta}$ is the slope, and $\mathbf{z} = (z_1, z_2, \ldots, z_p)$. Suppose

$$\mathrm{E}(\mathbf{y}) = \boldsymbol{\alpha} + \mathbf{z}\boldsymbol{\beta}$$

is a regression model for $\mathbf{y}$ given $\mathbf{z}$. The population Pearson partial correlation between the $i\mathrm{th}$ and the $j\mathrm{th}$ variables of $\mathbf{y}$ given $\mathbf{z}$ is defined as the correlation between errors $(y_i - \mathrm{E}(y_i))$ and $(y_j - \mathrm{E}(y_j))$.

If the exact values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are unknown, you can use a sample Pearson partial correlation to estimate the population Pearson partial correlation. For a given sample of observations, you estimate the sets of unknown parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ using the least-squares estimators $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$. Then the fitted least-squares regression model is

$$\widehat{\mathbf{y}} = \widehat{\boldsymbol{\alpha}} + \mathbf{z}\widehat{\boldsymbol{\beta}}$$

The partial corrected sums of squares and crossproducts (CSSCP) of $\mathbf{y}$ given $\mathbf{z}$ are the corrected sums of squares and crossproducts of the residuals $\mathbf{y} - \widehat{\mathbf{y}}$. Using these partial corrected sums of squares and crossproducts, you can calculate the partial variances, partial covariances, and partial correlations.

PROC CORR derives the partial corrected sums of squares and crossproducts matrix by applying the Cholesky decomposition algorithm to the CSSCP matrix. For Pearson partial correlations, let $\mathbf{S}$ be the partitioned CSSCP matrix between two sets of variables, $\mathbf{z}$ and $\mathbf{y}$:

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_{\mathbf{zz}} & \mathbf{S}_{\mathbf{zy}} \\ \mathbf{S}'_{\mathbf{zy}} & \mathbf{S}_{\mathbf{yy}} \end{bmatrix}$$

PROC CORR calculates $\mathbf{S}_{\mathbf{yy}\cdot\mathbf{z}}$, the partial CSSCP matrix of $\mathbf{y}$ after controlling for $\mathbf{z}$, by applying the Cholesky decomposition algorithm sequentially on the rows associated with $\mathbf{z}$, the variables being partialled out.

After applying the Cholesky decomposition algorithm to each row associated with variables $\mathbf{z}$, PROC CORR checks all higher numbered diagonal elements associated with $\mathbf{z}$ for singularity. After the Cholesky decomposition, a variable is considered singular if the value of the corresponding diagonal element is less than p times the original unpartialled corrected sum of squares of that variable. You can specify the singularity criterion p using the SINGULAR= option. For Pearson partial correlations, a controlling variable $\mathbf{z}$ is considered singular if the R^2 for predicting this variable from the variables that are already partialled out exceeds $1 - p$. When this happens, PROC CORR excludes the variable from the analysis. Similarly, a variable is considered singular if the R^2 for predicting this variable from the controlling variables exceeds $1 - p$. When this happens, its associated diagonal element and all higher numbered elements in this row or column are set to zero.

After the Cholesky decomposition algorithm is performed on all rows associated with $\mathbf{z}$, the resulting matrix has the form

$$\begin{bmatrix} \mathbf{T}_{\mathbf{zz}} & \mathbf{T}_{\mathbf{zy}} \\ \mathbf{0} & \mathbf{S}_{\mathbf{yy}\cdot\mathbf{z}} \end{bmatrix}$$

where $\mathbf{T}_{\mathbf{zz}}$ is an upper triangular matrix with

$$\mathbf{T}'_{\mathbf{zz}}\mathbf{T}_{\mathbf{zz}} = \mathbf{S}_{\mathbf{zz}'}$$
$$\mathbf{T}'_{\mathbf{zz}}\mathbf{T}_{\mathbf{zy}} = \mathbf{S}_{\mathbf{zy}'}$$
$$\mathbf{S}_{\mathbf{yy}\cdot\mathbf{z}} = \mathbf{S}_{\mathbf{yy}} - \mathbf{T}'_{\mathbf{zy}}\mathbf{T}_{\mathbf{zy}}.$$

If $\mathbf{S}_{\mathbf{zz}}$ is positive definite, then the partial CSSCP matrix $\mathbf{S}_{\mathbf{yy}\cdot\mathbf{z}}$ is identical to the matrix derived from the formula

$$\mathbf{S}_{\mathbf{yy}\cdot\mathbf{z}} = \mathbf{S}_{\mathbf{yy}} - \mathbf{S}'_{\mathbf{zy}}\ \mathbf{S}^{-1}_{\mathbf{zz}}\ \mathbf{S}_{\mathbf{zy}}$$

The partial variance-covariance matrix is calculated with the variance divisor (VARDEF= option). PROC CORR can then use the standard Pearson correlation formula on the partial variance-covariance matrix to calculate the Pearson partial correlation matrix. Another way to calculate Pearson partial correlation is by applying the Cholesky decomposition algorithm directly to the correlation matrix and by using the correlation formula on the resulting matrix.

To derive the corresponding Spearman partial rank-order correlations and Kendall partial tau-b correlations, PROC CORR applies the Cholesky decomposition algorithm to the Spearman rank-order correlation matrix and Kendall tau-b correlation matrix and uses the correlation formula. The singularity criterion for nonparametric partial correlations is identical to Pearson partial correlation except that PROC CORR uses a matrix of nonparametric correlations and sets a singular variable's associated correlations to missing. The partial tau-b correlations range from –1 to 1. However, the sampling distribution of this partial tau-b is unknown; therefore, the probability values are not available.

When a correlation matrix (Pearson, Spearman, or Kendall tau-b correlation matrix) is positive definite, the resulting partial correlation between variables x and y after adjusting for a single variable z is identical to that obtained from the first-order partial correlation formula

$$r_{xy\cdot z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{\left(1 - r^2_{xz}\right)\left(1 - r^2_{yz}\right)}}$$

where r_{xy}, r_{xz}, and r_{yz} are the appropriate correlations.

The formula for higher-order partial correlations is a straightforward extension of the above first-order formula. For example, when the correlation matrix is positive definite, the partial correlation between x and y controlling for both z_1 and z_2 is identical to the second-order partial correlation formula

$$r_{xy\cdot z_1 z_2} = \frac{r_{xy\cdot z_1} - r_{xz_2\cdot z_1}r_{yz_2\cdot z_1}}{\sqrt{\left(1 - r^2_{xz_2\cdot z_1}\right)\left(1 - r^2_{yz_2\cdot z_1}\right)}}$$

where $r_{xy\cdot z_1}$, $r_{xz_2\cdot z_1}$, and $r_{yz_2\cdot z_1}$ are first-order partial correlations among variables x, y, and z_2 given z_1.

Cronbach's Coefficient Alpha

Analyzing latent constructs such as job satisfaction, motor ability, sensory recognition, or customer satisfaction requires instruments to accurately measure the constructs. Interrelated items may be summed to obtain an overall score for each participant. Cronbach's coefficient alpha estimates the reliability of this type of scale by determining the internal consistency of the test or the average correlation of items within the test (Cronbach 1951).

When a value is recorded, the observed value contains some degree of measurement error. Two sets of measurements on the same variable for the same individual may not have identical values. However, repeated measurements for a series of individuals will show some consistency. Reliability measures internal consistency from one set of measurements to another. The observed value Y is divided into two components, a true value T and a measurement error E. The measurement error is assumed to be independent of the true value, that is,

$$\mathrm{Y} = \mathrm{T} + \mathrm{E}\ , \quad \mathrm{cov}\,(\mathrm{T},\mathrm{E}) = 0$$

The reliability coefficient of a measurement test is defined as the squared correlation between the observed value Y and the true value T, that is,

$$\begin{aligned}\rho^2\,(\mathrm{Y},\mathrm{T}) &= \frac{\mathrm{cov}\,(\mathrm{Y},\mathrm{T})^2}{\mathrm{var}\,(\mathrm{Y})\,\mathrm{var}\,(\mathrm{T})} \\ &= \frac{\mathrm{var}\,(\mathrm{T})^2}{\mathrm{var}\,(\mathrm{Y})\,\mathrm{var}\,(\mathrm{T})} \\ &= \frac{\mathrm{var}\,(\mathrm{T})}{\mathrm{var}\,(\mathrm{Y})}\end{aligned}$$

which is the proportion of the observed variance due to true differences among individuals in the sample. If Y is the sum of several observed variables measuring the same feature, you can estimate var(T). Cronbach's coefficient alpha, based on a lower bound for var(T), is an estimate of the reliability coefficient.

Suppose p variables are used with $\mathrm{Y}_j = \mathrm{T}_j + \mathrm{E}_j$ for $j = 1, 2, \ldots, p$, where Y_j is the observed value, T_j is the true value, and E_j is the measurement error. The measurement errors (E_j) are independent of the true values (T_j) and are also independent of each other. Let $\mathrm{Y}_0 = \sum \mathrm{Y}_j$ be the total observed score and $\mathrm{T}_0 = \sum \mathrm{T}_j$ be the total true score. Because

$$(p-1)\sum \mathrm{var}\,(\mathrm{T}_j) \geq \sum_{i \neq j} \mathrm{cov}\,(\mathrm{T}_i, \mathrm{T}_j)\ ,$$

a lower bound for $\mathrm{var}\,(\mathrm{T}_0)$ is given by

$$\frac{p}{p-1} \sum_{i \neq j} \mathrm{cov}\,(\mathrm{T}_i, \mathrm{T}_j)$$

With $\mathrm{cov}\,(\mathrm{Y}_i, \mathrm{Y}_j) = \mathrm{cov}\,(\mathrm{T}_i, \mathrm{T}_j)$ for $i \neq j$, a lower bound for the reliability coefficient is then given by the Cronbach's coefficient alpha:

$$\boldsymbol{\alpha} = \left(\frac{p}{p-1}\right)\frac{\sum_{i\neq j}\mathrm{cov}\left(\mathrm{Y}_i, \mathrm{Y}_j\right)}{\mathrm{var}\left(\mathrm{Y}_0\right)}$$

$$= \left(\frac{p}{p-1}\right)\left(1 - \frac{\sum_{j}\mathrm{var}\left(\mathrm{Y}_j\right)}{\mathrm{var}\left(\mathrm{Y}_0\right)}\right)$$

If the variances of the items vary widely, you can standardize the items to a standard deviation of 1 before computing the coefficient alpha. If the variables are dichotomous (0,1), the coefficient alpha is equivalent to the Kuder-Richardson 20 (KR-20) reliability measure.

When the correlation between each pair of variables is 1, the coefficient alpha has a maximum value of 1. With negative correlations between some variables, the coefficient alpha can have a value less than zero. The larger the overall alpha coefficient, the more likely that items contribute to a reliable scale. Nunnally (1978) suggests .70 as an acceptable reliability coefficient; smaller reliability coefficients are seen as inadequate. However, this varies by discipline.

To determine how each item reflects the reliability of the scale, you calculate a coefficient alpha after deleting each variable independently from the scale. The Cronbach's coefficient alpha from all variables except the $k\mathrm{th}$ variable is given by

$$\boldsymbol{\alpha}_k = \left(\frac{p-1}{p-2}\right)\left(1 - \frac{\sum_{i\neq k}\mathrm{var}\left(\mathrm{Y}_i\right)}{\mathrm{var}\left(\sum_{i\neq k}\mathrm{Y}_i\right)}\right)$$

If the reliability coefficient increases after deleting an item from the scale, you can assume that the item is not correlated highly with other items in the scale. Conversely, if the reliability coefficient decreases you can assume that the item is highly correlated with other items in the scale. See *SAS Communications*, 4th Quarter 1994, for more information on how to interpret Cronbach's coefficient alpha.

Listwise deletion of observations with missing values is necessary to correctly calculate Cronbach's coefficient alpha. PROC CORR does not automatically use listwise deletion when you specify ALPHA. Therefore, use the NOMISS option if the data set contains missing values. Otherwise, PROC FREQ prints a warning message in the SAS log indicating the need to use NOMISS with ALPHA.

Probability Values

Probability values for the Pearson and Spearman correlations are computed by treating

$$\frac{(n-2)^{1/2}\, r}{\left(1-r^2\right)^{1/2}}$$

as coming from a t distribution with $n-2$ degrees of freedom, where r is the appropriate correlation.

Probability values for the Pearson and Spearman partial correlations are computed by treating

$$\frac{(n-k-2)^{1/2}\,r}{(1-r^2)^{1/2}}$$

as coming from a t distribution with $n-k-2$ degrees of freedom, where r is the appropriate partial correlation and k is the number of variables being partialled out.

Probability values for Kendall correlations are computed by treating

$$\frac{\mathrm{s}}{\sqrt{\mathrm{var}\,(\mathrm{s})}}$$

as coming from a normal distribution when

$$\mathrm{s}=\sum_{i<j}\mathrm{sgn}\,(x_i-x_j)\,\mathrm{sgn}\,(y_i-y_j)$$

and where x_i are the values of the first variable, y_i are the values of the second variable, and the function sgn(z) is defined as

$$\mathrm{sgn}\,(z)=\begin{cases}1 & \text{if } z>0\\ 0 & \text{if } z=0\\ -1 & \text{if } z<0\end{cases}$$

The formula for the variance of s, var(s), is computed as

$$\mathrm{var}\,(\mathrm{s})=\frac{\mathrm{v}_0-\mathrm{v}_t-\mathrm{v}_u}{18}+\frac{\mathrm{v}_1}{2n\,(n-1)}+\frac{\mathrm{v}_2}{9n\,(n-1)\,(n-2)}$$

where

$$\begin{aligned}\mathrm{v}_0&=n\,(n-1)\,(2n+5)\\ \mathrm{v}_t&=\textstyle\sum t_i\,(t_i-1)\,(2t_i+5)\\ \mathrm{v}_u&=\textstyle\sum u_i\,(u_i-1)\,(2u_i+5)\\ \mathrm{v}_1&=\left(\textstyle\sum t_i\,(t_i-1)\right)\left(\textstyle\sum u_i\,(u_i-1)\right)\\ \mathrm{v}_2&=\left(\textstyle\sum t_i\,(t_i-1)\,(t_i-2)\right)\left(\textstyle\sum u_i\,(u_i-1)\,(u_i-2)\right)\end{aligned}$$

The sums are over tied groups of values where t_i is the number of tied x values and u_i is the number of tied y values (Noether 1967). The sampling distribution of Kendall's partial tau-b is unknown; therefore, the probability values are not available.

The probability values for Hoeffding's D statistic are computed using the asymptotic distribution computed by Blum, Kiefer, and Rosenblatt (1961). The formula is

$$\frac{(n-1)\,\pi^4}{60}\mathrm{D}+\frac{\pi^4}{72}$$

which comes from the asymptotic distribution. When the sample size is less than 10, see the tables for the distribution of D in Hollander and Wolfe (1973).

Results

Missing Values

By default, PROC CORR uses *pairwise deletion* when observations contain missing values. PROC CORR includes all nonmissing pairs of values for each pair of variables in the statistical computations. Therefore, the correlations statistics may be based on different numbers of observations.

If you specify the NOMISS option, PROC CORR uses *listwise deletion* when a value of the BY, FREQ, VAR, WEIGHT, or WITH statement variable is missing. PROC CORR excludes all observations with missing values from the analysis. Therefore, the number of observations for each pair of variables is identical. The PARTIAL statement always excludes the observations with missing values by automatically invoking NOMISS. Listwise deletion is needed to correctly calculate Cronbach's coefficient alpha when data are missing. If a data set contains missing values, when you specify ALPHA use the NOMISS option

There are two reasons to specify NOMISS and, thus, to avoid pairwise deletion. First, NOMISS is computationally more efficient, so you use fewer computer resources. Second, if you use the correlations as input to regression or other statistical procedures, a pairwise-missing correlation matrix leads to several statistical difficulties. Pairwise correlation matrices may not be nonnegative definite, and the pattern of missing values may bias the results.

Procedure Output

By default, PROC CORR prints a report that includes descriptive statistics and correlation statistics for each variable.The descriptive statistics include the number of observations with nonmissing values, the mean, the standard deviation, the minimum, and the maximum. PROC CORR reports the following additional descriptive statistics when you request various correlation statistics:

sum
: for Pearson correlation only

median
: for nonparametric measures of association

partial variance
: for Pearson partial correlation

partial standard deviation
: for Pearson partial correlation.

If variable labels are available, PROC CORR labels the variables.

When you specify the CSSCP, SSCP, or COV option, the appropriate sum-of-squares and crossproducts and covariance matrix appears at the top of the correlation report. If the data set contains missing values, PROC CORR prints additional statistics for each pair of variables. These statistics, calculated from the observations with nonmissing row and column variable values, may include

SSCP(W','V')
: uncorrected sum-of-squares and crossproducts

USS(W')
: uncorrected sum-of-squares for the row variable

USS(V')
: uncorrected sum-of-squares for the column variable

CSSCP(W','V')
: corrected sum-of-squares and crossproducts

CSS(W')
: corrected sum-of-squares for the row variable

CSS(V')
: corrected sum-of-squares for the column variable

COV (W','V')
: covariance

VAR (W')
: variance for the row variable

VAR (V')
: variance for the column variable

DF(W',V')
: divisor for calculating covariance and variances.

For each pair of variables, PROC CORR always prints the correlation coefficients, the number of observations used to calculate the coefficient, and the significance probability. When you specify the ALPHA option, PROC CORR prints Cronbach's coefficient alpha, the correlation between the variable and the total of the remaining variables, and Cronbach's coefficient alpha using the remaining variables for the raw variables and the standardized variables.

Output Data Sets

When you specify the OUTP=, OUTS=, OUTK=, or OUTH= option, PROC CORR creates an output data set containing statistics for Pearson correlation, Spearman correlation, Kendall correlation, or Hoeffding's D, respectively. By default, the output data set is a special data set type (TYPE=CORR) that many SAS/STAT procedures recognize, including PROC REG and PROC FACTOR. When you specify the NOCORR option and the COV, CSSCP, or SSCP option, use the TYPE= data set option to change the data set type to COV, CSSCP, or SSCP. For example, the following statement

```
proc corr nocorr cov outp=b(type=cov);
```

specifies the output data set type as COV.

PROC CORR does not print the output data set. Use PROC PRINT, PROC REPORT, or another SAS reporting tool to print the output data set.

The output data set includes the following variables

BY variables
: identifies the BY group when using a BY statement.

TYPE variable
: identifies the type of observation.

NAME variable
: identifies the variable that corresponds to a given row of the correlation matrix.

INTERCEP variable
: identifies variable sums when specifying the SSCP option.

VAR variables
: identifies the variables listed in the VAR statement.

You can use a combination of the _TYPE_ and _NAME_ variables to identify the contents of an observation. The _NAME_ variable indicates which row of the correlation matrix the observation corresponds to. The values of the _TYPE_ variable are

SSCP
: uncorrected sums of squares and crossproducts

CSSCP
: corrected sums of squares and crossproducts

COV
: covariances

MEAN
: mean of each variable

STD
: standard deviation of each variable

N
: number of nonmissing observations for each variable

SUMWGT
: sum of the weights for each variable when using a WEIGHT statement

CORR
: correlation statistics for each variable.

When you specify the SSCP option, the OUTP= data set includes an additional observation that contains intercept values. When you specify the ALPHA option, the OUTP= data set also includes observations with the following _TYPE_ values:

RAWALPHA
: Cronbach's coefficient alpha for raw variables

STDALPHA
: Cronbach's coefficient alpha for standardized variables

RAWALDEL
: Cronbach's coefficient alpha for raw variables after deleting one variable

STDALDEL
: Cronbach's coefficient alpha for standardized variables after deleting one variable

RAWCTDEL
: correlation between a raw variable and the total of the remaining raw variables

STDCTDEL
: correlation between a standardized variable and the total of the remaining standardized variables.

When you use a PARTIAL statement, the previous statistics are calculated for the variables after partialling. If PROC CORR computes Pearson correlation statistics, MEAN equals zero and STD equals the partial standard deviation associated with the partial variance for the OUTP=, OUTK=, or OUTS= data set. Otherwise, PROC CORR assigns missing values to MEAN and STD. Output 12.4 on page 299 lists the observations in an OUTP= data set when the COV option and PARTIAL statement are used to compute Pearson partial correlations. The _TYPE_ variable identifies COV, MEAN, STD, N, and CORR as the statistical values for the variables Weight, Oxygen, and Runtime. MEAN always equals 0, while STD is a partial standard deviation.

Output 12.4 OUTP= Data Set with Pearson Partial Correlations

```
   Pearson Correlation Statistics Using the PARTIAL Statement  1
                Output Data Set from PROC CORR

     _TYPE_     _NAME_       Weight      Oxygen     Runtime

      COV       Weight      72.4374    -12.7511      2.0677
      COV       Oxygen     -12.7511     27.0165     -5.5937
      COV       Runtime      2.0677     -5.5937      1.9451
      MEAN                   0.0000      0.0000      0.0000
      STD                    8.5110      5.1977      1.3947
      N                     28.0000     28.0000     28.0000
      CORR      Weight       1.0000     -0.2882      0.1742
      CORR      Oxygen      -0.2882      1.0000     -0.7716
      CORR      Runtime      0.1742     -0.7716      1.0000
```

Examples

Example 1: Computing Pearson Correlations and Other Measures of Association

Procedure features:

PROC CORR statement options:
- HOEFFDING
- PEARSON
- SPEARMAN

VAR statement

This example

- produces a correlation analysis with descriptive statistics, Pearson product-moment correlation, Spearman rank-order correlation, and Hoeffding's measure of dependence, D
- selects the analysis variables.

Program

```
options nodate pageno=1 linesize=80 pagesize=60;
```

The data set FITNESS contains measurements from a study of physical fitness for 30 participants between the ages 38 and 57. Each observation represents one person. Two observations contain missing values.

```
data fitness;
   input Age Weight Runtime Oxygen @@;
   datalines;
57 73.37 12.63 39.407   54 79.38 11.17 46.080
52 76.32 9.63  45.441   50 70.87 8.92    .
51 67.25 11.08 45.118   54 91.63 12.88 39.203
51 73.71 10.47 45.790   57 59.08 9.93  50.545
49 76.32  .    48.673   48 61.24 11.5  47.920
52 82.78 10.5  47.467   44 73.03 10.13 50.541
45 87.66 14.03 37.388   45 66.45 11.12 44.754
47 79.15 10.6  47.273   54 83.12 10.33 51.855
49 81.42 8.95  40.836   51 77.91 10.00 46.672
48 91.63 10.25 46.774   49 73.37 10.08 50.388
44 89.47 11.37 44.609   40 75.07 10.07 45.313
44 85.84 8.65  54.297   42 68.15 8.17  59.571
38 89.02 9.22  49.874   47 77.45 11.63 44.811
40 75.98 11.95 45.681   43 81.19 10.85 49.091
44 81.42 13.08 39.442   38 81.87 8.63  60.055
;
```

PEARSON, SPEARMAN, and HOEFFDING compute correlation statistics. When you request nonparametric correlations, specify PEARSON to compute Pearson correlations.

```
proc corr data=fitness pearson spearman hoeffding;
```

The VAR statement specifies the analysis variables and the order to print them.

```
   var weight oxygen runtime;
```

The TITLE statement specifies a title for the report.

```
   title 'Measures of Association for';
   title2 'a Physical Fitness Study';
run;
```

Output

The correlation report includes descriptive statistics, Pearson's rho, Spearman's rho, and Hoeffding's D. The report uses the median, instead of the sum, as a descriptive measure when PROC CORR computes nonparametric measures of association.

Because missing data are excluded pairwise, the number of observations PROC CORR uses to calculate the correlation coefficients varies.

```
                         Measures of Association for                          1
                           a Physical Fitness Study

                              The CORR Procedure

                    3  Variables:  Weight   Oxygen   Runtime


                              Simple Statistics

Variable         N        Mean     Std Dev      Median     Minimum     Maximum

Weight          30    77.70500     8.34152    77.68000    59.08000    91.63000
Oxygen          29    47.06445     5.32129    46.67200    37.38800    60.05500
Runtime         29    10.61448     1.41655    10.47000     8.17000    14.03000


                     Pearson Correlation Coefficients
                        Prob > |r| under H0: Rho=0
                          Number of Observations

                            Weight        Oxygen       Runtime

              Weight       1.00000      -0.19900       0.15155
                                          0.3007        0.4326
                                30            29            29

              Oxygen      -0.19900       1.00000      -0.78346
                            0.3007                      <.0001
                                29            29            28

              Runtime      0.15155      -0.78346       1.00000
                            0.4326        <.0001
                                29            28            29


                    Spearman Correlation Coefficients
                        Prob > |r| under H0: Rho=0
                          Number of Observations

                            Weight        Oxygen       Runtime

              Weight       1.00000      -0.13110       0.10546
                                          0.4979        0.5861
                                30            29            29

              Oxygen      -0.13110       1.00000      -0.68363
                            0.4979                      <.0001
                                29            29            28

              Runtime      0.10546      -0.68363       1.00000
                            0.5861        <.0001
                                29            28            29
```

```
                    Measures of Association for                          2
                     a Physical Fitness Study

                        The CORR Procedure

                Hoeffding Dependence Coefficients
                      Prob > D under H0: D=0
                      Number of Observations

                        Weight        Oxygen       Runtime

        Weight         0.97559      -0.01789      -0.02418
                        <.0001        0.9775        1.0000
                            30            29            29

        Oxygen        -0.01789       1.00000       0.16554
                        0.9775                      <.0001
                            29            29            28

        Runtime       -0.02418       0.16554       1.00000
                        1.0000        <.0001
                            29            28            29
```

Example 2: Computing Rectangular Correlation Statistics with Missing Data

Procedure features:

PROC CORR statement options:

- COV
- NOSIMPLE
- SSCP

VAR statement

WITH statement

This example

- suppresses descriptive statistics
- prints uncorrected sum-of-squares and crossproducts
- calculates a rectangular covariance matrix
- calculates a rectangular correlation matrix
- excludes observations with missing values using pairwise deletion (default method).

Program

```
options nodate pageno=1 linesize=80 pagesize=60;
```

The data set SETOSA contains measurements for four iris parts: sepal length, sepal width, petal length, and petal width based on Fisher's iris data (1936). Fifty iris specimens from the species **Iris setosa** are used. Each observation represents one specimen. Three observations contain missing values. The LABEL statement associates a label with each variable.

```
data setosa;
   input SepalLength SepalWidth PetalLength PetalWidth @@;
   label sepallength='Sepal Length in mm.'
         sepalwidth='Sepal Width in mm.'
         petallength='Petal Length in mm.'
         petalwidth='Petal Width in mm.';
   datalines;
50 33 14 02   46 34 14 03   46 36 .  02
51 33 17 05   55 35 13 02   48 31 16 02
52 34 14 02   49 36 14 01   44 32 13 02
50 35 16 06   44 30 13 02   47 32 16 02
48 30 14 03   51 38 16 02   48 34 19 02
50 30 16 02   50 32 12 02   43 30 11 .
58 40 12 02   51 38 19 04   49 30 14 02
51 35 14 02   50 34 16 04   46 32 14 02
57 44 15 04   50 36 14 02   54 34 15 04
52 41 15 .    55 42 14 02   49 31 15 02
54 39 17 04   50 34 15 02   44 29 14 02
47 32 13 02   46 31 15 02   51 34 15 02
50 35 13 03   49 31 15 01   54 37 15 02
54 39 13 04   51 35 14 03   48 34 16 02
48 30 14 01   45 23 13 03   57 38 17 03
51 38 15 03   54 34 17 02   51 37 15 04
52 35 15 02   53 37 15 02
;
```

SSCP displays the uncorrected sum-of-squares and crossproducts matrix and invokes PEARSON. COV calculates the covariance matrix. NOSIMPLE suppresses descriptive statistics.

```
proc corr data=setosa sscp cov nosimple;
```

The WITH statement together with the VAR statement produces a rectangular correlation matrix. The matrix rows are PetalLength and PetalWidth while the matrix columns are SepalLength and SepalWidth.

```
   var sepallength sepalwidth;
   with petallength petalwidth;
```

The TITLE statement specifies a title for the report.

```
   title 'Fisher (1936) Iris Setosa Data';
run;
```

Output

The correlation report includes rectangular sum-of-squares and crossproducts, covariances, and the correlation matrix using the two WITH variables and two VAR variables. The descriptive statistics do not appear. PROC CORR uses variable labels to label matrix rows (WITH variables).

PROC CORR calculates sum-of-squares and crossproducts and covariances statistics for each pair of variables by using observations with nonmissing row and column variable values.

Because missing data are excluded pairwise, the number of observations PROC CORR uses to calculate the correlation coefficients changes.

```
                      Fisher (1936) Iris Setosa Data                      1

                            The CORR Procedure

                 2 With Variables:  PetalLength PetalWidth
                 2      Variables:  SepalLength SepalWidth


                   Sums of Squares and Crossproducts
                     SSCP / Row Var SS / Col Var SS

                                 SepalLength        SepalWidth

        PetalLength              36214.00000       24756.00000
        Petal Length in mm.      10735.00000       10735.00000
                                 123793.0000        58164.0000

        PetalWidth                6113.00000        4191.00000
        Petal Width in mm.         355.00000         355.00000
                                 121356.0000        56879.0000


                       Variances and Covariances
         Covariance / Row Var Variance / Col Var Variance / DF

                                 SepalLength        SepalWidth

        PetalLength              1.270833333       1.363095238
        Petal Length in mm.      2.625000000       2.625000000
                                 12.33333333       14.60544218
                                          48                48

        PetalWidth               0.911347518       1.048315603
        Petal Width in mm.       1.063386525       1.063386525
                                 11.80141844       13.62721631
                                          47                47


                    Pearson Correlation Coefficients
                       Prob > |r| under H0: Rho=0
                         Number of Observations

                                       Sepal         Sepal
                                      Length         Width

            PetalLength              0.22335       0.22014
            Petal Length in mm.       0.1229        0.1285
                                          49            49

            PetalWidth               0.25726       0.27539
            Petal Width in mm.        0.0775        0.0582
                                          48            48
```

Example 3: Computing Cronbach's Coefficient Alpha

Procedure features:

PROC CORR statement options:

ALPHA
NOCORR
NOMISS

This example

- computes Cronbach's coefficient alpha for a multiple-item mixed-rating scale
- suppresses Pearson correlation statistics
- excludes observations with missing values using listwise deletion.

This example does not examine the correlation matrix but assumes that all items are positively correlated. Normally, you want to examine the correlation and covariance matrices to make sure that all variables are positively correlated. Positive correlation is needed because items measure a common entity. You exclude negatively correlated items from the analysis because they do not measure the same construct.

Program

```
options nodate pageno=1 linesize=80 pagesize=60;
```

The data set PYSCHDAT contains responses to a questionnaire assessing the mental stability of 30 randomly selected female psychiatric patients.* Each observation represents one patient. The scale includes seven items. The LABEL statement provides a label for each item. Seven observations contain missing values.

```
data psychdat;
   input Age Anxiety Depression Sleep Sex Life WeightChange @@;
   label age            = 'age in years'
         anxiety        = 'anxiety level'
         depression     = 'depression level'
         sleep          = 'normal sleep (1=y 2=n)'
         sex            = 'sexual (1=n 2=y)'
         life           = 'suicidal (1=n 2=y)'
         weightchange   = 'recent weight change';
    datalines;
39  2   2   2   2   2  4.9 41  2   2   2   2   2  2.2
42  3   3   .   2   2  4.0 30  2   2   2   2   2 -2.6
35  2   1   1   2   1 -0.3 44  .   1   2   1   1  0.9
31  2   2   .   2   2 -1.5 39  3   2   2   2   1  3.5
35  3   2   2   2   2 -1.2 33  2   2   2   2   2  0.8
38  2   1   1   1   1 -1.9 31  2   2   2   .   1  5.5
```

*
Data are from *Assignments in Applied Statistics* by Simon Conrad. Copyright ©1989, by John Wiley & Sons, Inc. Reprinted with permission from the publisher.

```
40  3   2   2   2   1  2.7 44  2   2   2   2   2  4.4
43  3   2   2   2   2  3.2 32  1   1   1   2   1 -1.5
32  1   2   2   .   1 -1.9 43  4   3   2   2   2  8.3
46  3   2   2   2   2  3.6 30  2   2   2   2   1  1.4
34  3   3   .   2   2   .  37  3   2   2   2   1   .
35  2   1   2   2   1 -1.0 45  2   2   2   2   2  6.5
35  2   2   2   2   1 -2.1 31  2   2   2   2   1 -0.4
32  2   2   2   2   1 -1.9 44  2   2   2   2   2  3.7
40  3   3   2   2   2  4.5 42  3   3   2   2   2  4.2
;
```

ALPHA computes Cronbach's alpha and invokes PEARSON. NOCORR suppresses Pearson correlation statistics. NOMISS excludes observations with missing values. Omitting a VAR statement causes PROC CORR to use all numeric variables.

```
proc corr data=psychdat alpha nocorr nomiss;
```

The TITLE statement specifies a title for the report.

```
   title1 'Mental Stability Scale for Female Psychiatric Patients';
run;
```

Output

The correlation report includes descriptive statistics and Cronbach's coefficient alpha, the correlation between the variable and the total of the remaining variables, and Cronbach's coefficient alpha using the remaining variables for both the raw variables and the standardized variables. These calculations use the 23 observations without missing values.

Because the variances of some variables vary widely, you use the standardized scores to estimate reliability. The overall standardized alpha of .85 is an acceptable reliability coefficient. This is greater than Nunnally's suggested value of .70.

The standardized alpha provides information on how each item reflects the reliability of the scale. Notice that the standardized alpha decreases after removing Depression from the construct. Therefore, this variable appears strongly correlated with other items in the scale. The standardized alpha increases slightly after removing Sex from the construct. Thus, removing this variable from the scale makes the construct more reliable.

```
              Mental Stability Scale for Female Psychiatric Patients           1

                                  The CORR Procedure

    7  Variables:    Age          Anxiety      Depression   Sleep        Sex
                     Life         WeightChange


                                  Simple Statistics

     Variable                N           Mean       Std Dev           Sum

     Age                    23       37.91304       5.13378     872.00000
     Anxiety                23        2.34783       0.64728      54.00000
     Depression             23        1.95652       0.56232      45.00000
     Sleep                  23        1.86957       0.34435      43.00000
     Sex                    23        1.95652       0.20851      45.00000
     Life                   23        1.56522       0.50687      36.00000
     WeightChange           23        1.78261       3.06381      41.00000

                                  Simple Statistics

     Variable           Minimum       Maximum    Label

     Age               30.00000      46.00000    age in years
     Anxiety            1.00000       4.00000    anxiety level
     Depression         1.00000       3.00000    depression level
     Sleep              1.00000       2.00000    normal sleep (1=y 2=n)
     Sex                1.00000       2.00000    sexual (1=n 2=y)
     Life               1.00000       2.00000    suicidal (1=n 2=y)
     WeightChange      -2.60000       8.30000    recent weight change


                              Cronbach Coefficient Alpha

                             Variables              Alpha
                             ----------------------------
                             Raw                 0.627754
                             Standardized        0.845339
```

```
             Mental Stability Scale for Female Psychiatric Patients           2

                                The CORR Procedure


                Cronbach Coefficient Alpha with Deleted Variable

                          Raw Variables              Standardized Variables

    Deleted          Correlation                        Correlation
    Variable          with Total            Alpha        with Total           Alpha
    ----------------------------------------------------------------------------
    Age                 0.742614         0.557515          0.546856        0.832207



                Cronbach Coefficient Alpha with Deleted Variable

                     Deleted
                     Variable        Label
                     --------------------------------------
                     Age             age in years

                Cronbach Coefficient Alpha with Deleted Variable

                          Raw Variables              Standardized Variables

    Deleted          Correlation                        Correlation
    Variable          with Total            Alpha        with Total           Alpha
    ----------------------------------------------------------------------------
    Anxiety             0.577129         0.600944          0.590851        0.825643
    Depression          0.554983         0.608273          0.770956        0.797610
    Sleep               0.378930         0.630242          0.618367        0.821482
    Sex                 0.155115         0.642017          0.333368        0.862537
    Life                0.622207         0.607333          0.625338        0.820421
    WeightChange        0.843939         0.341006          0.749261        0.801087



                Cronbach Coefficient Alpha with Deleted Variable

                     Deleted
                     Variable        Label
                     --------------------------------------
                     Anxiety         anxiety level
                     Depression      depression level
                     Sleep           normal sleep (1=y 2=n)
                     Sex             sexual (1=n 2=y)
                     Life            suicidal (1=n 2=y)
                     WeightChange    recent weight change
```

Example 4: Storing Partial Correlations in an Output Data Set

Procedure features:

PROC CORR statement options:

- COV
- KENDALL
- NOSIMPLE
- OUTP=
- SPEARMAN

PARTIAL statement

VAR statement

Data set: FITNESS on page 301

This example

- □ suppresses descriptive statistics
- □ calculates three types of partial correlation coefficients
- □ calculates a partial covariance matrix
- □ excludes observations with missing values using listwise deletion
- □ selects the analysis variables
- □ creates an output data set with Pearson correlation statistics.

See Output 12.4 on page 299 for a listing of the output data set.

Program

```
options nodate pageno=1 linesize=120 pagesize=60;
```

SPEARMAN and KENDALL request correlation statistics. COV calculates the covariance matrix and invokes PEARSON. NOSIMPLE suppresses descriptive statistics. OUT= creates the FITCORR data set that contains Pearson correlation statistics.

```
proc corr data=fitness spearman kendall cov nosimple
          outp=fitcorr;
```

The VAR statement specifies the analysis variables and the order to print them.

```
   var weight oxygen runtime;
```

The PARTIAL statement calculates partial correlations using Age as the controlling variable.

```
   partial age;
```

The LABEL statement associates a label with each variable for the duration of the PROC step.

```
   label age      = 'Age of subject'
         weight   = 'Wt in kg'
         runtime  = '1.5 mi in minutes'
         oxygen   = 'O2 use';
```

The TITLE statement specifies a title for the report.

```
   title1 'Partial Correlations for a Fitness and Exercise Study';
run;
```

Output

The report includes a partial covariance matrix and partial correlations for Pearson's rho, Spearman's rho, and Kendall's tau-b. The *p*-values for Kendall's tau-b are not available. Because observations with missing data are excluded, PROC CORR uses 28 observations to calculate correlation coefficients.

```
                     Partial Correlations for a Fitness and Exercise Study                              1

                                       The CORR Procedure

                        1 Partial Variables:    Age
                        3         Variables:    Weight   Oxygen   Runtime


                               Partial Covariance Matrix, DF = 26

                                               Weight            Oxygen           Runtime

  Weight       Wt in kg               72.43742055      -12.75113194        2.06766763
  Oxygen       O2 use                -12.75113194       27.01654904       -5.59370556
  Runtime      1.5 mi in minutes       2.06766763       -5.59370556        1.94512451


                       Pearson Partial Correlation Coefficients, N = 28
                              Prob > |r| under H0: Partial Rho=0

                                          Weight        Oxygen       Runtime

              Weight                     1.00000      -0.28824       0.17419
              Wt in kg                                  0.1448        0.3849

              Oxygen                    -0.28824       1.00000      -0.77163
              O2 use                      0.1448                      <.0001

              Runtime                    0.17419      -0.77163       1.00000
              1.5 mi in minutes           0.3849        <.0001


                      Spearman Partial Correlation Coefficients, N = 28
                              Prob > |r| under H0: Partial Rho=0

                                          Weight        Oxygen       Runtime

              Weight                     1.00000      -0.16407       0.08708
              Wt in kg                                  0.4135        0.6658

              Oxygen                    -0.16407       1.00000      -0.67112
              O2 use                      0.4135                      0.0001

              Runtime                    0.08708      -0.67112       1.00000
              1.5 mi in minutes           0.6658        0.0001


              Kendall Partial Tau b Correlation Coefficients, N = 28

                                          Weight        Oxygen       Runtime

              Weight                     1.00000      -0.09021       0.02854
              Wt in kg

              Oxygen                    -0.09021       1.00000      -0.52158
              O2 use

              Runtime                    0.02854      -0.52158       1.00000
              1.5 mi in minutes
```

References

Blum, J.R., Kiefer, J., and Rosenblatt, M. (1961), "Distribution Free Tests of Independence Based on the Sample Distribution Function," *Annals of Mathematical Statistics*, 32, 485–498.

Conover, W.J. (1980), *Practical Nonparametric Statistics*, Second Edition, New York: John Wiley & Sons, Inc.

Cronbach, L.J. (1951), "Coefficient Alpha and the Internal Structure of Tests," *Psychometrika*, 16, 297–334.

Fisher, R.A. (1936), "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, 7, 179–188.

Hoeffding, W. (1948), "A Non-Parametric Test of Independence," *Annals of Mathematical Statistics*, 19, 546–557.

Hollander, M. and Wolfe, D. (1973), *Nonparametric Statistical Methods*, New York: John Wiley & Sons, Inc.

Knight, W.E. (1966), "A Computer Method for Calculating Kendall's Tau with Ungrouped Data," *Journal of the American Statistical Association*, 61, 436–439.

Moore, D.S. (1995), *Statistics: Concepts and Controversies, 3rd Edition*, New York: W.H. Freeman & Company.

Noether, G.E. (1967), *Elements of Nonparametric Statistics*, New York: John Wiley & Sons, Inc.

Novick, M.R. (1967), "Coefficient Alpha and the Reliability of Composite Measurements," *Psychometrika*, 32, 1–13.

Nunnally, J. (1978), *Psychometric theory*, New York: McGraw-Hill Companies.

Ott, L. (1993), *An Introduction to Statistical Methods and Data Analysis, 4th Edition*, Belmont: Wadsworth Publishing Company.

SAS Institute Inc., "Measuring the Internal Consistency of a Test, Using PROC CORR to Compute Cronbach's Coefficient Alpha," *SAS Communications*, 20:4, TT2–TT5.

Spector, P.E. (1992). *Summated Rating Scale Construction: An Introduction*, Newbury Park: Sage.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc., *SAS® Procedures Guide, Version 8,* Cary, NC: SAS Institute Inc., 1999. 1729 pp.

SAS® Procedures Guide, Version 8

ISBN 1–58025–482–9

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

1st printing, October 1999