Getting Correlations Using PROC CORR

Correlation analysis provides a method to measure the strength of a linear relationship between two numeric variables. PROC CORR can be used to compute Pearson product-moment correlation coefficient between variables, as well as three nonparametric measures of association, Spearman's rank-order correlation, Kendall's tau-b, and Hoeffding's measure of dependence D. PROC CORR also computes simple descriptive statistics. The general form of the PROC CORR statement is

PROC CORR options;

The simplest form

PROC CORR;

will compute pairwise Pearson correlation coefficients for all numeric variables in the most recently created SAS data set. Allowable options in the PROC CORR statement include the DATA= option, as well as options to produce an output data set. The OUTP = option will create a data set containing Pearson correlations; the OUTS = option will create a data set containing Spearman correlations; the OUTK = option will create a data set containing Kendall correlations; and the OUTH = option will create a data set containing Hoeffding statistics.

Options to select the kinds of correlations to be computed include PEARSON (these are computed by default), KENDALL, SPEARMAN, and HOEFFDING.

A VAR statement can be used with PROC CORR to specify which variables are to be analyzed. Pairwise correlation coefficients will be computed for all variables in the VAR statement. For example,

DATA example1;
INPUT name $ 1-10 sex $ 12 age 14-15 height 17-18 weight 20-22 ;
DATALINES;
Data records go here
;
RUN;
PROC CORR DATA= example1 PEARSON SPEARMAN
   VAR weight height ;
RUN;

PROC CORR also computes probabilities to test the null hypothesis
$ H_0: \rho=0 \$. If three or more numeric variables are in the input data set or are
specified in the VAR statement, the correlations and probabilities are printed
in matrix form.

The WITH statement can be used to obtain correlations for specific
combinations of variables. Suppose you wish to compute a correlation for
age with height, and another for age with weight, but you are not interested
in a correlation of weight with height. You can accomplish this by using the
WITH statement in conjunction with the VAR statement as follows:

PROC CORR DATA= example1 PEARSON SPEARMAN
   VAR weight height ;
   WITH age;
RUN;

A BY statement can be used with PROC CORR to obtain separate
analyses on observations in the groups defined by the BY statement. When a
BY statement is used, PROC CORR expects the data to be sorted in the
order specified in the BY statement. As an example suppose that you want
separate analyses for males and females. A SAS program to accomplish this
follows:

DATA example1;
INPUT name $ 1-10 sex $ 12 age 14-15 height 17-18 weight 20-22 ;
SAS programming statements go here
DATALINES;
Data records go here
; 
RUN;
PROC SORT DATA = example1;
   BY sex;
RUN;
PROC CORR DATA=example1;
RUN;
PROC CORR DATA= example1 ;
   VAR height weight age ;
   BY sex;
RUN;

The default method for handling missing values is to use all nonmissing pairs of values for each pair of variables in the VAR or WITH statements. This implies that differing numbers of observations may be used to compute correlation coefficients for different pairs of variables. If the NOMISS option is used in the PROC CORR statement, any observations with missing values for any of the variables in the VAR or WITH statements are excluded from the analysis, i.e., all pairwise correlations are computed using the same number of observations.

Exercise 8: Use one of your previous examples and compute pairwise correlations for all numeric variables. Run another analysis using a VAR and a WITH statement. Next run another analysis using a BY statement. Interpret the results.

Producing Graphs Using PROC PLOT

PROC PLOT can be used to generate a graph of the values of one variable plotted against values of another variable. The general form of the PROC PLOT statement is

PROC PLOT options;

The simplest form is

PROC PLOT;

Allowable options include the DATA= option. Another useful option is the UNIFORM option. This option requests uniform axes scaling when a BY statement is used. This allows you to easily compare plots for different levels of the BY variables.
You can ask PROC PLOT to construct and print plots for specified variables in the data set by using the PLOT statement. Any number of PLOT statements may be included in one execution of PROC PLOT. The PLOT statement has the form

```
PLOT requests / options;
```

The requests in the PLOT statement indicate the variables to be plotted and the plotting characters to be used to mark points on the plot. The first variable in a request will be represented on the vertical axis (y-axis) and the second variable in the request will be represented by the horizontal axis (x-axis). The plot requests can have the form

```
var1 * var2
```

When this form is used, SAS determines the plotting character to be used to mark points on the plot. When a point on the plot represents one observation PROC PLOT uses the plotting character A to mark that point; when a point on the plot represents two observations PROC PLOT uses the plotting character B to mark that point; and so on.

Plot requests may also have the form

```
var1 * var2 = 'character'
```

This form allows the user to specify the plotting character to be used to mark points on the plot. The character you choose can represent values from one observation, or more than one observation. As an example,

```
PROC PLOT;
   PLOT weight * height = '*' ;
```

The third form that a plot request may have is

```
var1 * var2 = var3
```

This form specifies that points on the plot of var1 * var2 will be indicated by the value of var3. As an example, suppose that the variable sex has the values 'F' or 'M'. Then the statements
PROC PLOT;
   PLOT weight * height = sex;

Will cause the points on the plot to be marked with either an F or an M.

Any number of plot requests can appear in a single PLOT statement. The plots will appear on separate graphs unless the OVERLAY option is specified in the PLOT statement. The OVERLAY tells PROC PLOT to print the plot requests specified in the PLOT statement on a single graph. This can be useful for plots in a regression analysis.

If values of any of the variables in a plot request are missing, PROC PLOT does not include the observation in the plot.

A BY statement can be used with PROC PLOT to obtain separate plots on observations in groups defined by the BY variable(s). When a BY statement appears, PROC PLOT expects the data to be sorted in the order of the BY variables.

Now for a full-blown example:

DATA example1;
INPUT name $ 1-10 sex $ 12 age 14-15 height 17-18 weight 20-22;
SAS programming statements go here
DATALINES;
Data records go here;
RUN;
PROC SORT DATA = example1;
   BY sex;
RUN;
PROC PLOT DATA = example1;
   PLOT weight * height = sex height * age weight * age;
RUN;
PROC PLOT DATA = example1;
   PLOT age * height = 'H' age * weight = 'W' / OVERLAY;
   BY sex;
RUN;