The CORR Procedure

A measure of linear association between two random variables is a correlation coefficient. Correlation differs from regression analysis in that there are no independent and dependent variable associations. The two variables that are analyzed are both random variable responses. When the data are normally distributed, a Pearson correlation coefficient can be computed. When the data are not normally distributed, a Spearman correlation coefficient (rank based calculation) can be computed. In either case, the sample correlation coefficient, r, estimates the population correlation coefficient, ρ. Both $-1 \leq r \leq 1$ and $-1 \leq \rho \leq 1$. When the correlation coefficient is near zero, the two variables are not linearly related. If one variable increases as the other increases, the correlation is positive. And, if one variable decreases as the other increases the correlation is negative. For either the Pearson or Spearman measurements, it may be of interest to test whether or not there is a linear association between the two variables. That is, is the correlation non-zero.

For the Pearson correlation coefficient, the details of the test are:

Hypotheses	**Test Statistic**	**Reject H_0 if**
H_0: $\rho = 0$ (No linear association.) H_1: $\rho \neq 0$ (Linear association)	$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$	$\lvert t \rvert \geq t_{\alpha/2,n-2}$

For the Spearman correlation coefficient, the details of the test statistic are not included here, but the hypotheses and the rejection region would be the same.

There are formulas for confidence intervals for correlation coefficients, but they have been omitted here.

SAS can compute both Pearson and Spearman correlation coefficients for a pair of variables. The syntax of the CORR procedure is as follows:

PROC CORR DATA=*setname* <options>;
VAR *variablelist1;*
WITH *variablelist2;*
RUN;

Options include:

PEARSON requests that the Pearson correlation coefficient and its test of significance be computed. If no options are specified, the Pearson coefficient prints by default.

SPEARMAN requests that the Spearman correlation coefficient and its test of significance be computed. The Spearman correlation coefficient is appropriate when no distribution assumptions are placed on the two random variables.

PLOTS = NONE suppresses all plots

ODS Graphics must be enabled to produce the following plots. Many types of graphics are available. Only the MATRIX and SCATTER plots are overviewed here.

PLOTS = MATRIX<(*matrix options*) >

requests a scatter plot matrix for variables. That is, the procedure displays a symmetric matrix plot with variables in the VAR list if a WITH statement is not specified. Otherwise, the procedure displays a rectangular matrix plot with the WITH variables appearing down the side and the VAR variables appearing across the top.

Options include
HISTOGRAM produces a histogram for each variable on the diagonal of the scatterplot matrix

NVAR=ALL or NVAR= *n*
specifies the maximum number of variables in the VAR list to be displayed in the matrix plot, where. The NVAR=ALL option uses all variables in the VAR list. By default, NVAR=5. For n values greater than 5 the resulting scatterplot matrix may have plots that are too small to be useful.

PLOTS=SCATTER ***<(scatter-options)>***

requests scatter plots for pairs of variables. That is, the procedure displays a scatter plot for each applicable pair of distinct variables from the VAR list if a WITH statement is not specified. Otherwise, the procedure displays a scatter plot for each applicable pair of variables, one from the WITH list and the other from the VAR list. Initially use the scatter options ELLIPSE=NONE. That is, PLOTS=SCATTER (ELLIPSE=NONE) in the PROC CORR statement.

When a scatter plot or a scatter plot matrix is requested, the Pearson correlations will also be displayed.

If the resulting maximum number of variables in the VAR or WITH list is greater than 10, only the first 10 variables in the list are displayed in the scatter plots.

Consider the following two blocks of statements;

1
```
PROC CORR  DATA=one;
VAR  a b  ;
WITH  x y z ;
RUN;
```

2
```
PROC CORR  DATA=one;
VAR a  b  x  y  z ;
RUN;
```

In **1**, six correlation coefficients will be computed. A table of correlations and their significances will be computed. Significance levels for the test of no linear association versus some linear

association will be computed below each of the correlations. The general form of the output will look like this:

	A	B
X	corr(X,A)	corr(X,B)
Y	corr(Y,A)	corr(Y,B)
Z	corr(Z,A)	corr(Z,B)

where corr(x,y) represents the correlation coefficient for the variables X and Y.

In **2**, fifteen correlation coefficients will be computed. A correlation matrix of the following form will be computed. Significance levels will also be computed.

	A	B	X	Y	Z
A	corr(A,A)=1	corr(A,B)	corr(A,X)	corr(A,Y)	corr(A,Z)
B	corr(B,A)	corr(B,B)=1	corr(B,X)	corr(B,Y)	corr(B,Z)
X	corr(X,A)	corr(X,B)	corr(X,X)=1	corr(X,Y)	corr(X,Z)
Y	corr(Y,A)	corr(Y,B)	corr(Y,X)	corr(Y,Y)=1	corr(Y,Z)
Z	corr(Z,A)	corr(Z,B)	corr(Z,X)	corr(Z,Y)	corr(Z,Z)=1

Objective 9: Test whether or not all pairs of the variables are correlated. Produce the default scatterplot matrix.

A	12	15	14	19	19	21	12	15
B	24	27	26	33	34	41	22	25
C	8	9	7	4	2	3	10	6

```
DM 'LOG; CLEAR; ODSRESULTS; CLEAR; ';

DATA  one;
INPUT  a b c;
DATALINES;
12 24 8
15 27 9
14 26 7
19 33 4
19 34 2
21 41 3
12 22 10
15 25 6
;
PROC CORR DATA=one  PLOTS=MATRIX;
VAR  a b c;
TITLE 'Correlation Example' ;
TITLE2 'Objective 9 - Default Scatterplot Matrix';

PROC CORR DATA=one SPEARMAN PEARSON PLOTS=MATRIX;
VAR a b ;
WITH c;
RUN;

QUIT;
```

Modify this program so that the histograms of the distributions appear for each variable.

```
PROC CORR DATA=one  PLOTS=MATRIX(HISTOGRAM) ;
VAR  a b c;
TITLE 'Correlation Example' ;
TITLE2 'Objective 9 - Scatterplot Matrix with Histograms';
```

The HISTOGRAM option does not "work" when a WITH statement is used in the procedure.

Objective 10: Compute the Pearson and Spearman correlation coefficients between A and C, and B and C only. Produce the default scatterplot matrix.

Correlation Example
Objective 9 - Default Scatterplot Matrix

The CORR Procedure

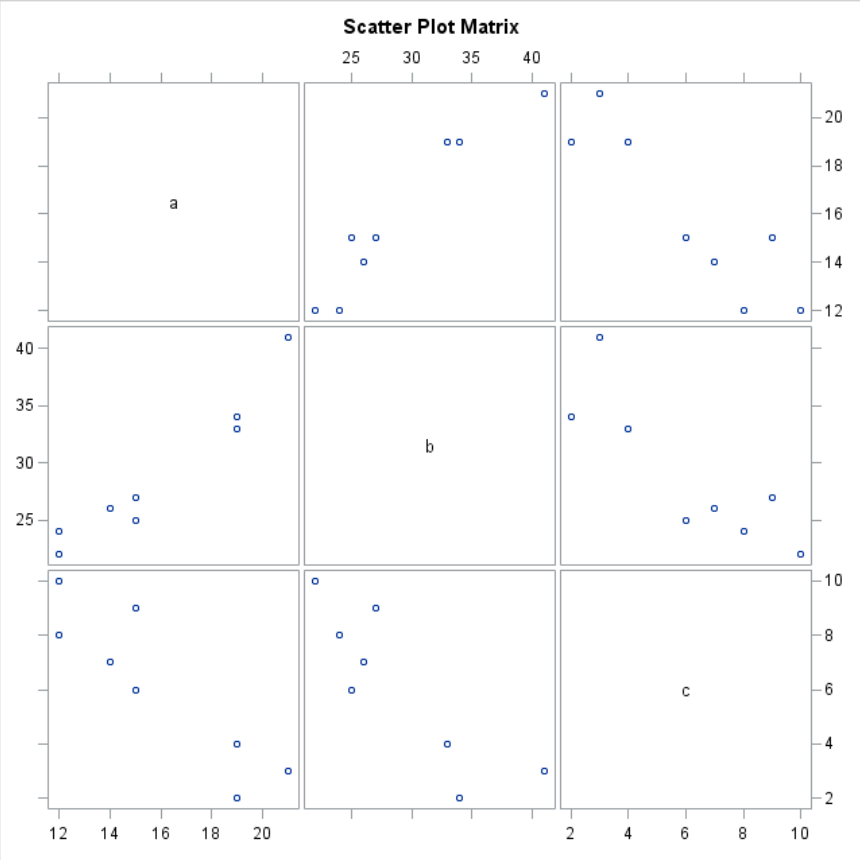
3 Variables: a b c

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
a	8	15.87500	3.39905	127.00000	12.00000	21.00000
b	8	29.00000	6.41427	232.00000	22.00000	41.00000
c	8	6.12500	2.90012	49.00000	2.00000	10.00000

Pearson Correlation Coefficients, N = 8
Prob > |r| under H0: Rho=0

	a	b	c
a	1.00000	0.96319	-0.89669
		0.0001	0.0025
b	0.96319	1.00000	-0.85244
	0.0001		0.0072
c	-0.89669	-0.85244	1.00000
	0.0025	0.0072	



Correlation Example
Objective 9 - Default Scatterplot Matrix
Using a WITH statement

The CORR Procedure

1 With Variables:	c
2 Variables:	a b

Simple Statistics

Variable	N	Mean	Std Dev	Median	Minimum	Maximum
c	8	6.12500	2.90012	6.50000	2.00000	10.00000
a	8	15.87500	3.39905	15.00000	12.00000	21.00000
b	8	29.00000	6.41427	26.50000	22.00000	41.00000

Pearson Correlation Coefficients, N = 8
Prob > |r| under H0: Rho=0

	a	b
c	-0.89669	-0.85244
	0.0025	0.0072

Spearman Correlation Coefficients, N = 8
Prob > |r| under H0: Rho=0

	a	b
c	-0.83650	-0.80952
	0.0096	0.0149