# Chapter 401

# Correlation Matrix

## Introduction

This program calculates Pearsonian and Spearman-rank correlation matrices. It allows missing values to be deleted in a pair-wise or row-wise fashion.

When someone speaks of a correlation matrix, they usually mean a matrix of Pearson-type correlations. Unfortunately, these correlations are unduly influenced by outliers, unequal variances, nonnormality, and nonlinearities. One of the chief competitors of the Pearson correlation coefficient is the Spearman-rank correlation coefficient. This latter correlation is calculated by applying the Pearson correlation formulas to the ranks of the data rather than to the actual data values themselves. In so doing, many of the distortions that infect the Pearson correlation are reduced considerably.

To allow you to compare the two types of correlation matrices, a matrix of differences can be displayed. This allows you to determine which pairs of variables require further investigation.

This program lets you specify a set of partial variables. The linear influence of these variables is removed by sweeping them out of the matrix. This provides a statistical adjustment to the remaining variables using multiple regression. Note that in the case of Spearman correlations, this sweeping occurs after the complete correlation matrix has been formed.

## Discussion

When there is more than one independent variable, the collection of all pair-wise correlations are succinctly represented in a correlation form. In regression analysis, the purpose of examining these correlations is two-fold: to find outliers and to identify collinearity. In the case of outliers, there should be major differences between the parametric measure, the Pearson correlation coefficient, and the nonparametric measure, the Spearman rank correlation coefficient. In the case of collinearity, high pair-wise correlations could be the first indicators of collinearity problems.

The Pearson correlation coefficient is unduly influenced by outliers, unequal variances, nonnormality, and nonlinearities. As a result of these problems, the Spearman correlation coefficient, which is based on the ranks of the data rather than the actual data, may be a better choice for examining the relationships between variables.

Finally, the patterns of missingness in multiple regression and correlation analysis can be very complex. As a result, missing values can be deleted in a pair-wise or a row-wise fashion. If there are only a few observations with missing values, it might be preferable to use the row-wise deletion, especially for large data sets. The row-wise deletion procedure omits the entire observation from the analysis. On the other hand, if the pattern of missingness is randomly dispersed throughout the data and the use of the row-wise deletion would omit at least 25% of the observations, the pair-wise deletion procedure for missing values would be a safer way to capture the essence of the relationships between variables. While this method appears to make full use of all your data, the resulting correlation matrix may have mathematical and interpretation difficulties. Mathematically, this correlation matrix may not have a positive determinant. Since each correlation may be based on a different set of rows, practical interpretations could be difficult, if not illogical.

The Spearman correlation coefficient measures the monotonic association between two variables in terms of ranks. It measures whether one variable increases or decreases with another even when the relationship between the two variables is not linear or bivariate normal. Computationally, each of the two variables is ranked separately, and the

ordinary Pearson correlation coefficient is computed on the ranks. This nonparametric correlation coefficient is a good measure of the association between two variables when outliers, nonnormality, nonconstant variance, and nonlinearity may exist between the two variables being investigated.

# Data Structure

The data are entered in two or more variables. An example of data appropriate for this procedure is shown in the table below. These data are contained in the Yield dataset. We suggest that you open this dataset now so that you can follow along with the example.

**Yield dataset (subset)**

| YldA | YldB | YldC |
|---|---|---|
| 452 | 546 | 785 |
| 874 | 547 | 458 |
| 554 | 774 | 886 |
| 447 | 465 | 536 |
| 356 | 459 | |
| 754 | 665 | 669 |
| 558 | 467 | 857 |
| 574 | 365 | 821 |
| 664 | 589 | 772 |
| 682 | 534 | 732 |
| | 456 | 689 |
| 547 | 651 | 654 |
| | 654 | |
| 435 | 665 | 297 |
| | 546 | 830 |
| 245 | 537 | 827 |

# Procedure Options

This section describes the options available in this procedure.

## Variables Tab

Specify the variables on which to run the analysis.

### Data Variables

#### Correlation Variables

Specify the variables whose correlations are to be formed. Only numeric data are analyzed.

#### Partial Variables

An optional set of variables that are to be "partialled out" of the correlation matrix. The influence of these variables is removed by sweeping them from the remaining variables. For the Pearson-type correlations, the resulting matrix is the same that would be formed if the regular variables were regressed on the partial variables, the residuals were stored, and the correlation matrix of these residuals was formed. The correlations that are formed are the partial correlations.

## Options

### Correlation Type

Specify the type of correlation to be computed

- **Pearson Product-Moment**
  Display the Pearson product-moment correlation matrix.
- **Spearman Rank**
  Display the Spearman-Rank correlation matrix.
- **Both**
  Display both the Pearson product-moment and the Spearman-Rank correlation matrices.

### Missing Value Removal

This option indicates how you want the program to handle missing values.

- **Pair-wise**
  Pair-wise removal of missing values. Each correlation is based on all pairs of data values in which no missing values occur. Missing values occurring in other variables do not influence the calculations. Note that although this method appears to make full use of all your data, the resulting correlation matrix is difficult to analyze. Mathematically, it may not have a positive determinant. Practically, each correlation may be based on a different set of rows, making it difficult to interpret.
- **Row-wise**
  Row-wise removal of missing values. If a missing value occurs in any of the variables specified, the row of data is ignored in the calculation of all correlations.

# Reports Tab

These options specify the reports.

## Select Reports

### Difference Report

Specify whether to display the matrix of differences between the Pearson and the Spearman correlations.

## Report Options

### Report Format

Specifies the length and format of the correlation matrix.

- **Short**
  Display only the correlation matrix.
- **Full**
  Display the full report with sample sizes and significance levels.

### Variable Names

This option lets you select whether to display variable names, variable labels, or both.

## Storage Tab

Specify if and where the correlation matrices are to be stored.

### Data Storage Variables

#### Pearson Correlations Storage Variables

A list of variables into which the Pearson correlation matrix is stored.

#### Spearman Correlations Storage Variables

A list of variables into which the Spearman correlation matrix is stored.

# Example 1 – Creating a Correlation Matrix

This section presents an example of how to run an analysis of the data contained in the Yield dataset.

You may follow along here by making the appropriate entries or load the completed template **Example 1** by clicking on Open Example Template from the File menu of the Correlation Matrix window.

**1 Open the Yield dataset.**

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Click on the file **Yield.NCSS**.
- Click **Open**.

**2 Open the Correlation Matrix window.**

- Using the Analysis menu or the Procedure Navigator, find and select the **Correlation Matrix** procedure.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

**3 Specify the variables.**

- On the Correlation Matrix window, select the **Variables tab**.
- Double-click in the **Correlation Variables** text box. This will bring up the variable selection window.
- Select **YldA, YldB, YldC** from the list of variables and then click **Ok**. "YldA-YldC" will appear in the Correlation Variables box.
- Enter **Both** in the Correlation Type box.
- Enter **Pair Wise** in the Missing Value Removal box.

**4 Specify the reports.**

- On the Correlation Matrix window, select the **Reports tab**.
- Check the **Different Report** box.
- Enter **Full** in the Report Format box.

**5 Run the procedure.**

- From the Run menu, select **Run Procedure**. Alternatively, just click the green Run button.

# Pearson Correlations Section

## Full Report Format

**Pearson Correlations Section (Pair-Wise Deletion)**

| | YldA | YldB | YldC |
|---|---|---|---|
| YldA | 1.000000 | .170692 | -.361414 |
| | .000000 | .577154 | .248377 |
| | 12.000000 | 13.000000 | 12.000000 |
| YldB | .170692 | 1.000000 | -.004071 |
| | .577154 | .000000 | .988980 |
| | 13.000000 | 14.000000 | 14.000000 |
| YldC | -.361414 | -.004071 | 1.000000 |
| | .248377 | .988980 | .000000 |
| | 12.000000 | 14.000000 | 14.000000 |

Cronbachs Alpha = -0.250337 Standardized Cronbachs Alpha = -0.223864

The above report displays the correlations, significance level, and sample size of each pair of variables. This format is obtained when Report Format is set to "Full."

### Reliability

Because of the central role of measurement in science, scientists of all disciplines are concerned with the accuracy of their measurements. Item analysis is a methodology for assessing the accuracy of measurements that are obtained in the social sciences where precise measurements are often hard to secure. The accuracy of a measurement may be broken down into two main categories: validity and reliability. The validity of an instrument refers to whether it accurately measures the attribute of interest. The reliability of an instrument concerns whether it produces identical results in repeated applications. An instrument may be reliable but not valid. However, it cannot be valid without being reliable.

The methods described here assess the reliability of an instrument. They do not assess its validity. This should be kept in mind when using the techniques of item analysis since they address reliability, not validity.

An instrument may be valid for one attribute but not for another. For example, a driver's license exam may accurately measure an individual's ability to drive. However, it does not accurately measure that individual's ability to do well in college. Hence the exam is reliable and valid for measuring driving ability. It is reliable and invalid for measuring success in college.

Several methods have been proposed for assessing the reliability of an instrument. These include the retest method, alternative-form method, split-halves method, and the internal consistency method. We will focus on internal consistency here.

### Cronbach's Alpha

Cronbach's alpha (or *coefficient alpha*) is the most popular of the internal consistency coefficients. It is calculated as follows:

$$\alpha = \frac{K}{K-1}\left[1 - \frac{\sum_{i=1}^{K}\sigma_{ii}}{\sum_{i=1}^{K}\sum_{j=1}^{K}\sigma_{ij}}\right]$$

where $K$ is the number of items (questions) and $\sigma_{ij}$ is the estimated covariance between items $i$ and $j$. Note the $\sigma_{ii}$ is the variance (not standard deviation) of item $i$.

## Correlation Matrix

If the data are standardized by subtracting the item means and dividing by the item standard deviations before the above formula is used, we get the standardized version of Cronbach's alpha. A little algebra will show that this is equivalent to the following calculations based directly on the correlation matrix of the items:

$$\alpha = \frac{K\overline{\rho}}{1+\overline{\rho}(K-1)}$$

where $K$ is the number of items (variables) and $\overline{\rho}$ is the average of all the correlations among the $K$ items.

Cronbach's alpha has several interpretations. It is equal to the average value of alpha coefficients obtained for all possible combinations of dividing $2K$ items into two groups of $K$ items each and calculating the two-half tests. Also, alpha estimates the expected correlation of one instrument with an alternative form containing the same number of items. Furthermore, alpha estimates the expected correlation between an actual test and a hypothetical test which may never be written.

Since Cronbach's alpha is supposed to be a correlation, it should range between -1 and 1. However, it is possible for alpha to be less than -1 when several of the covariances are relatively large, negative numbers. In most cases, alpha is positive, although negative values arise occasionally. What value of alpha should be achieved? Carmines (1990) stipulates that as a rule, a value of at least 0.8 should be achieved for widely used instruments. An instrument's alpha value may be improved by either adding more items or by increasing the average correlation among the items.

### Short Report Format

**Pearson Correlations Section (Pair-Wise Deletion)**

| | YldA | YldB | YldC |
|---|---|---|---|
| YldA | 1.000000 | 0.170692 | -0.361414 |
| YldB | 0.170692 | 1.000000 | -0.004071 |
| YldC | -0.361414 | -0.004071 | 1.000000 |

Cronbachs Alpha = 0.219908 Standardized Cronbachs Alpha = 0.311396

The above report displays the correlation matrix only. This format is obtained when Report Format is set to "Short."

## Spearman Correlations Section

### Full Report Format

**Spearman Correlations Section (Pair-Wise Deletion)**

| | YldA | YldB | YldC |
|---|---|---|---|
| YldA | 1.000000 | 0.184319 | -0.153846 |
| | 0.000000 | 0.546634 | 0.633091 |
| | 12.000000 | 13.000000 | 12.000000 |
| YldB | 0.184319 | 1.000000 | -0.088106 |
| | 0.546634 | 0.000000 | 0.764552 |
| | 13.000000 | 14.000000 | 14.000000 |
| YldC | -0.153846 | -0.088106 | 1.000000 |
| | 0.633091 | 0.764552 | 0.000000 |
| | 12.000000 | 14.000000 | 14.000000 |

The above report displays the correlations, significance level, and sample size of each pair of variables. This format is obtained when Report Format is set to "Full."

## Correlation Matrix

### Short Report Format

**Spearman Correlations Section (Pair-Wise Deletion)**

| | YldA | YldB | YldC |
|---|---|---|---|
| YldA | 1.000000 | .184319 | -.153846 |
| YldB | .184319 | 1.000000 | -.088106 |
| YldC | -.153846 | -.088106 | 1.000000 |

The above report displays the correlation matrix only. This format is obtained when Report Format is set to "Short."

## Difference Between Correlations Section

**Difference Between Pearson and Spearman Correlations Section (Pair-Wise Deletion)**

| | YldA | YldB | YldC |
|---|---|---|---|
| YldA | 0.000000 | -0.013627 | -0.207568 |
| YldB | -0.013627 | 0.000000 | 0.084035 |
| YldC | -0.207568 | 0.084035 | 0.000000 |

The above report displays the difference between the Pearson and the Spearman correlation coefficients. The report lets you find those pairs of variables for which these two correlation coefficients are very different. A large difference here indicates the presence of outliers, nonlinearity, nonnormality, and the like. You should investigate scatter plots of pairs of variables with large differences.

## Storing the Correlations on the Database

When you specify variables in either the Pearson Correlations or the Spearman Correlations boxes, the correlation matrix will be stored in those variables during the execution of the program.