Chapter 1

Using the Cluster Analysis

Background Information

Cluster analysis is the name of a multivariate technique used to identify similar characteristics in a group of observations. In cluster analysis, you also identify and classify observations or variables so that each observation or variable in a cluster is most like others in the same cluster. If the classification is correct, the observations or variables within the clusters will be close together; objects in different clusters will be far apart.

Cluster analysis is useful in many different disciplines such as business, biology, psychology, and sociology. Also referred to as Q-analysis, classification analysis, and numerical taxonomy, the name varies from discipline to discipline.

There is, however, one commonality: classification by natural relationship, which suggests that the value of cluster analysis lies in preclassifying data. For example, you might classify large amounts of otherwise meaningless data into manageable groups; you might reduce the data into specific smaller subgroups; or you might develop hypotheses about the nature of the data. Whatever the situation, using cluster analysis becomes increasingly complex when you add more variables or include mixed datasets.

Cluster analysis involves using techniques in three stages: partition, interpret, and profile. During the partition stage decisions are made about how to measure the data, which algorithms are the best suited for classifying the data, and determining the number of clusters that will be formed.

The most commonly used algorithms are hierarchical and nonhierarchical. Hierarchical methods are used to construct either agglomerative or divisive tree-like structures. In the agglomerative method, each observation begins in its own cluster and in subsequent steps combines with new aggregate clusters, which reduces the number of clusters in each step. When the clustering process goes in the opposite direction, it is known as a divisive or K-means method. Nonhierarchical algorithms do not involve tree-like structures.

Determining how to choose the number of clusters is not a definitive process. Generally the distance between clusters at sequential steps provides guidelines. Usually it is best to first try several solutions for different clusters then make a final decision from among the solutions.

In the interpretion stage, the statements used to develop the clusters are examined and assigned a name or label that accurately describes the characteristics of the cluster.

The profile stage describes the characteristics of each cluster, which explains how each cluster may differ on relevant dimensions. This stage usually involves using discriminant analysis or other appropriate statistics, then uses the clusters as they are labeled in stage two. The analysis continues by using data that have not been previously identified to form each cluster's characteristics. In other words, this stage focuses on identifying the clusters then describing their characteristics instead of focusing on the composition of the clusters.

Clustering Methods Available in STATGRAPHICS *Plus*

Cluster Analysis in STATGRAPHICS *Plus* uses six hierarchical clustering methods: Nearest Neighbor, Furthest Neighbor, Centroid, Median, Group Average, and Ward's. K-means is the only nonhierarchical method. Each method uses different criteria to measure the distance among the clusters. The data analysis in each hierarchical method begins with an individual observation as its own cluster. Then the observations are combined into successively larger clusters until the number of specified clusters is reached. The methods are briefly described below.

Nearest Neighbor

This method finds the two observations that have the shortest distance and places them in the first cluster. Then it finds the next shortest distance and either joins a third observation to the first two to form a cluster or forms a new two-observation cluster. This process continues until all the observations are in one cluster. The Nearest Neighbor method is also known as the *Single Linkage* method.

Furthest Neighbor

This method uses the maximum distance between any two observations in a cluster. All the observations in a cluster are linked to each other at some maximum distance or by some minimum similarity. This method is also known as *Complete Linkage*.

Centroid

In this method the distance between the means of two clusters is used as the measurement. Each time observations are grouped, new centroids are formed. The cluster centroids move each time a new observation or group of observations is added to an existing cluster. Centroid methods require that you use metric data; other methods do not.

Median

This method uses the median distance from observations in one cluster to observations in another as the measurement. This approach tends to combine clusters that have small variances and may produce clusters that have the same variance.

Group Average

In this method the distance between clusters is calculated as the average distance between all the observations in one cluster and the average distance between all the observations in another cluster.

Ward's

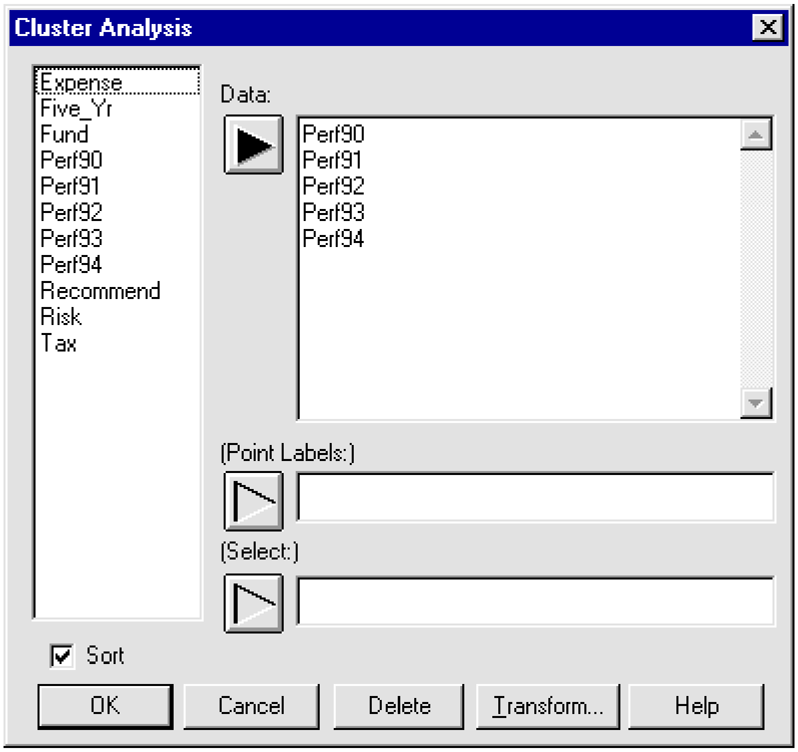
In Ward's method the distance between two clusters is the sum of squares between two clusters summed over all the variables. This method combines clusters that have a small number of observations, and tends to produce clusters that have approximately the same number of observations.

K-means

K-means clustering is a nonhierarchical method. Each cluster begins with a specified number of groups, each of which has a single random point. A sequence of points is sampled, and each point is added, in turn, to the group whose mean it is closest to. The group mean is then adjusted.

Using Cluster Analysis

To access the analysis, choose SPECIAL... MULTIVARIATE METHODS... CLUSTER ANALYSIS... from the Menu bar to display the Cluster Analysis dialog box shown in Figure 1-1.



Data:
Perf90
Perf91
Perf92
Perf93
Perf94

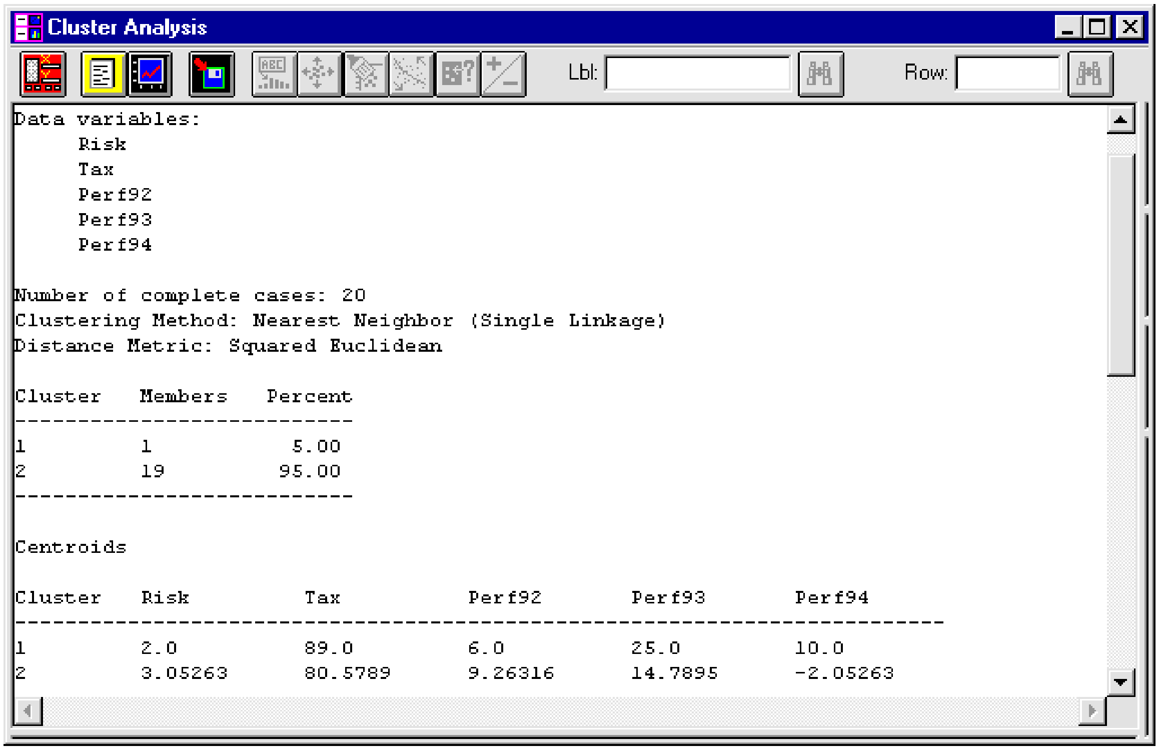
☑ Sort

Figure 1-1. Cluster Analysis Dialog Box

Tabular Options

Analysis Summary

The Analysis Summary option creates a summary of the analysis (see Figure 1-2). The summary displays the names of the variables, the number of complete cases, and the names of the clustering method and distance metric you are using. It then displays a summary of the clusters, the centroids (means) for the cluster, and for each of the variables.



```
Data variables:
     Risk
     Tax
     Perf92
     Perf93
     Perf94

Number of complete cases: 20
Clustering Method: Nearest Neighbor (Single Linkage)
Distance Metric: Squared Euclidean

Cluster   Members   Percent
---------------------------
1         1           5.00
2         19         95.00
---------------------------


Centroids

Cluster   Risk         Tax          Perf92       Perf93       Perf94
--------------------------------------------------------------------------
1         2.0          89.0         6.0          25.0         10.0
2         3.05263      80.5789      9.26316      14.7895      -2.05263
```

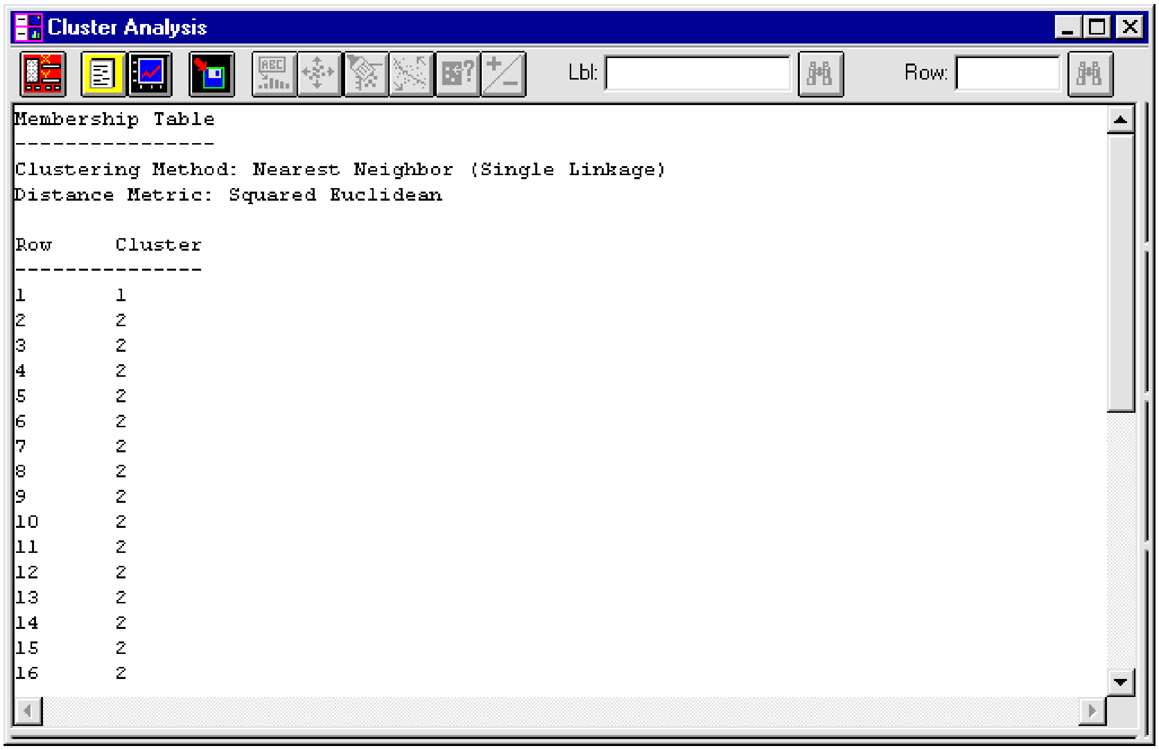
Figure 1-2. Analysis Summary

Use the *Cluster Analysis Options* dialog box to choose the method that will be used to combine the observations or variables into clusters, to enter the number of clusters that will be created, and to choose the type of measurement method that will be used to calculate the distance between clusters.

You can also use the Seeds... command to access the *Seed Options* dialog box, which you use to enter the row numbers of the observations that will be used for each corresponding cluster. This command is available only when you choose the K-means option.

Membership Table

The Membership Table option creates a report that lists each observation and the cluster in which it is placed (see Figure 1-3).



```
Membership Table
----------------
Clustering Method: Nearest Neighbor (Single Linkage)
Distance Metric: Squared Euclidean

Row      Cluster
---------------
1        1
2        2
3        2
4        2
5        2
6        2
7        2
8        2
9        2
10       2
11       2
12       2
13       2
14       2
15       2
16       2
```

Figure 1-3. Membership Table

Use the *Membership Table Options* dialog box to indicate if the clusters should be sorted and displayed together. For example, if you are using two clusters, the table with cluster 1's will display first followed by cluster 2's.

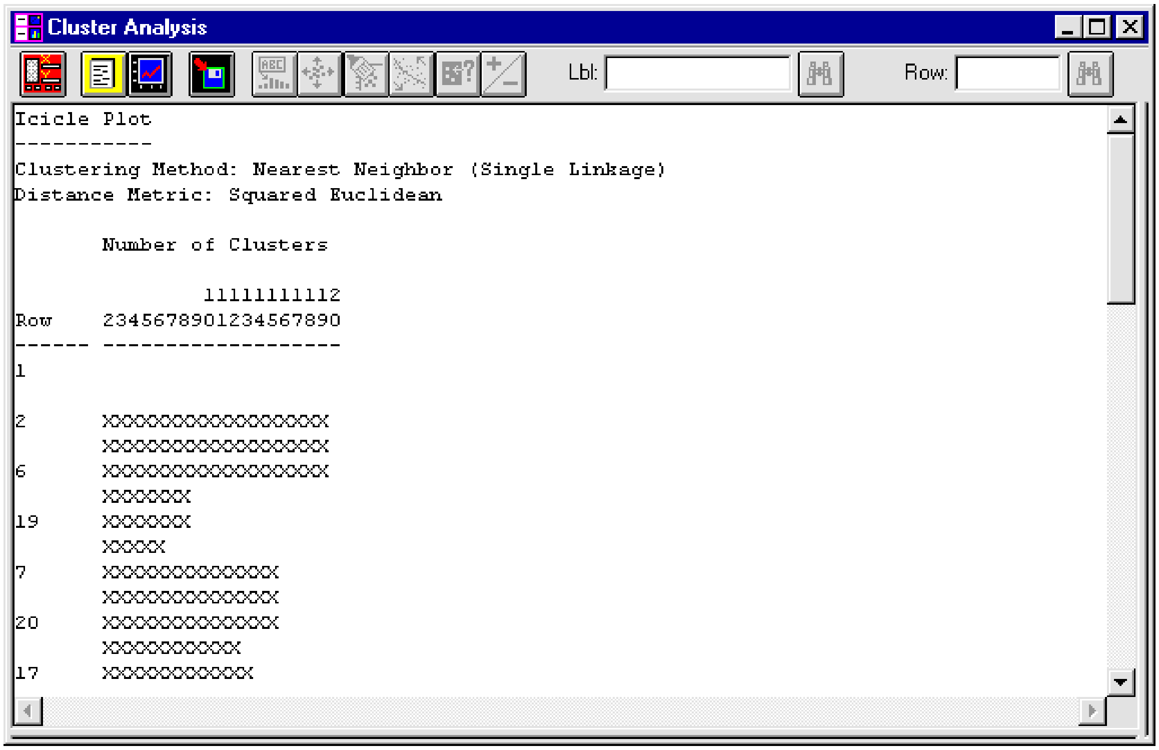
Icicle Plot

The Icicle Plot option creates an icicle plot that shows the number of clusters horizontally across the top of the plot and the observations vertically down the side (see Figure 1-4).

Use the *Icicle Plot Options* dialog box to change the width of the plot.

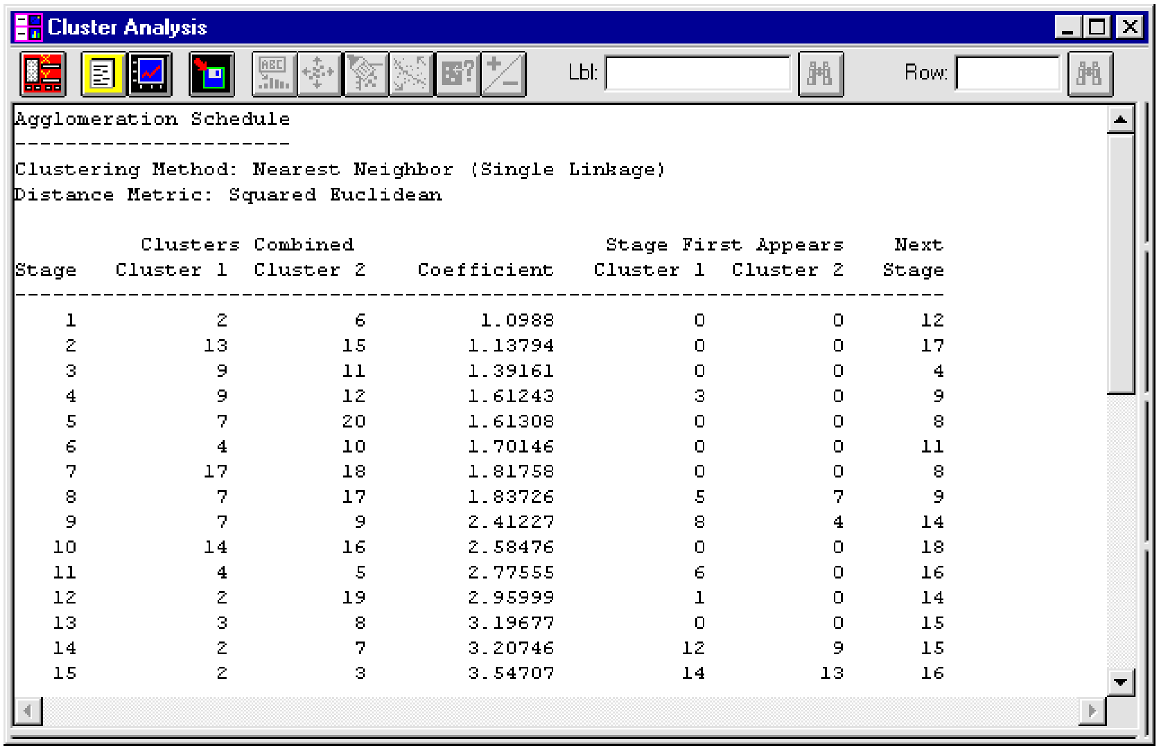
Agglomeration Schedule

The Agglomeration Schedule option creates a table that shows the clusters as they are combined at each step (see Figure 1-5).



```
Icicle Plot
-----------
Clustering Method: Nearest Neighbor (Single Linkage)
Distance Metric: Squared Euclidean

       Number of Clusters

               1111111112
Row    2345678901234567890
------ -------------------
1

2      XXXXXXXXXXXXXXXXXX
       XXXXXXXXXXXXXXXXXX
6      XXXXXXXXXXXXXXXXXX
       XXXXXXX
19     XXXXXXX
       XXXXX
7      XXXXXXXXXXXXXX
       XXXXXXXXXXXXXX
20     XXXXXXXXXXXXXX
       XXXXXXXXXXX
17     XXXXXXXXXXXX
```

Figure 1-4. Icicle Plot



```
Agglomeration Schedule
----------------------
Clustering Method: Nearest Neighbor (Single Linkage)
Distance Metric: Squared Euclidean

          Clusters Combined                     Stage First Appears      Next
Stage    Cluster 1  Cluster 2     Coefficient  Cluster 1  Cluster 2     Stage
-------------------------------------------------------------------------------
    1            2          6          1.0988          0          0        12
    2           13         15         1.13794          0          0        17
    3            9         11         1.39161          0          0         4
    4            9         12         1.61243          3          0         9
    5            7         20         1.61308          0          0         8
    6            4         10         1.70146          0          0        11
    7           17         18         1.81758          0          0         8
    8            7         17         1.83726          5          7         9
    9            7          9         2.41227          8          4        14
   10           14         16         2.58476          0          0        18
   11            4          5         2.77555          6          0        16
   12            2         19         2.95999          1          0        14
   13            3          8         3.19677          0          0        15
   14            2          7         3.20746         12          9        15
   15            2          3         3.54707         14         13        16
```

Figure 1-5. Agglomeration Schedule

Graphical Options

Dendrogram

The Dendrogram option creates a graphical representation (a tree graph) of the results (see Figure 1-6). The vertical axis represents the distance at each step; the horizontal axis represents the observations as they are combined.

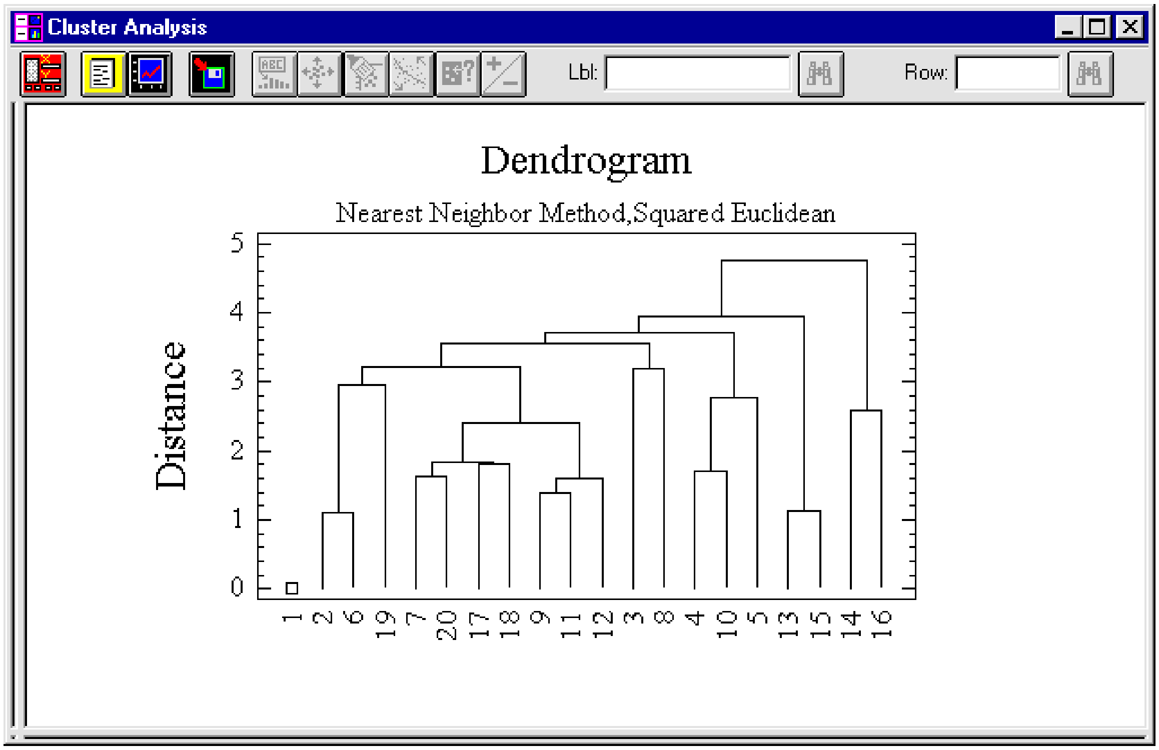


Figure 1-6. Dendrogram

Two-Dimensional Scatterplot

The Two-dimensional Scatterplot option creates a two-dimensional scatterplot, which plots clustered observations versus two variables (see Figure 1-7). A different point symbol is used for each cluster. As an option, you can circle clusters to see them more clearly.

Use the *Two-Dimensional Scatterplot Options* dialog box to choose the names of the variables that will be used and to indicate if you want the clusters circled.

Agglomeration Distance Plot

The Agglomeration Distance Plot option plots the distance as the clusters are combined at each step in the dendrogram (see Figure 1-8).

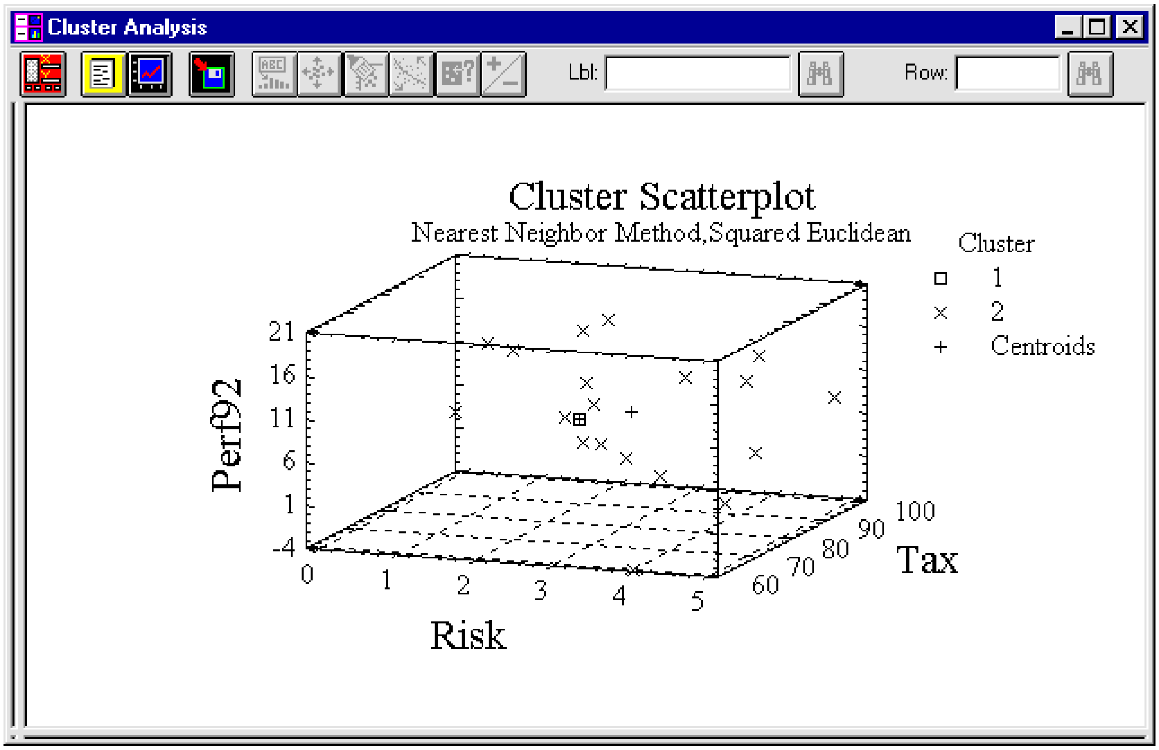


Figure 1-7. Two-Dimensional Scatterplot

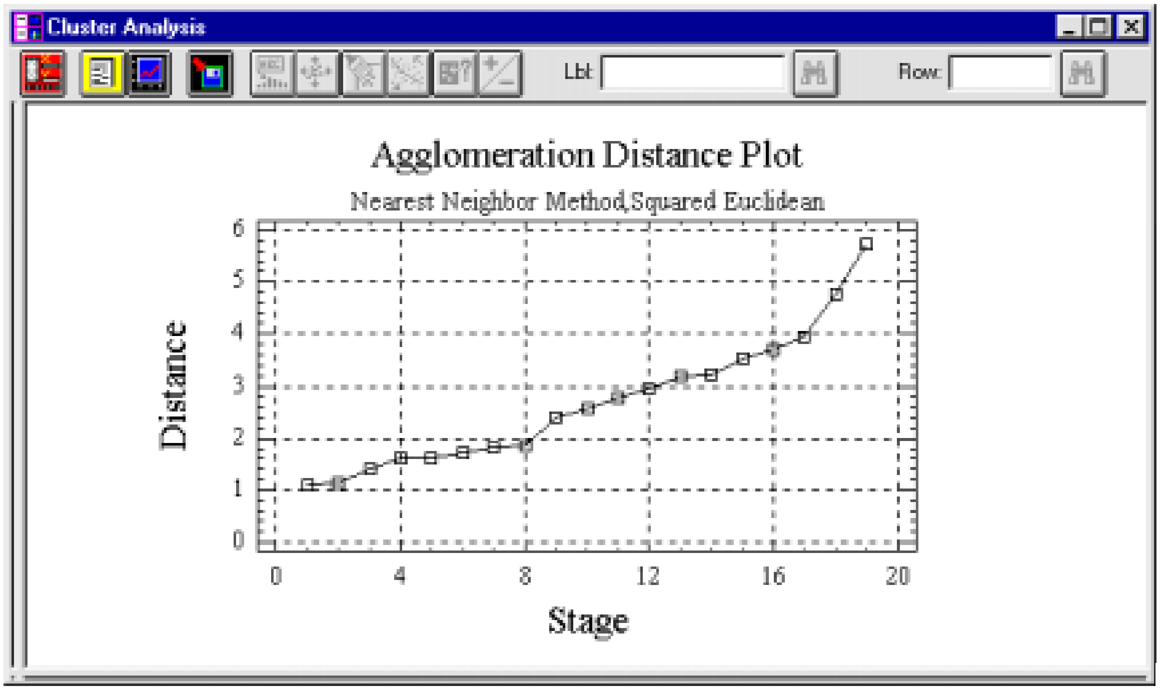

Cluster Analysis
Lbl:
Row:
Agglomeration Distance Plot
Nearest Neighbor Method,Squared Euclidean
Distance
0
1
2
3
4
5
6
0
4
8
12
16
20
Stage

Figure 1-8. Agglomeration Distance Plot

Saving the Results

The Save Results Options dialog box allows you to choose the results you want to save. There are two selections: Cluster Numbers and Distance Matrix.

You can also use the Target Variables text boxes to enter the names of the variables in which you want to save the values generated during the analysis. You can enter new names or accept the defaults.

Note: To access the Save Results Options dialog box, click the Save Results button on the Analysis toolbar (the fourth button from the left).

References

Anderson, T. W. 1984. *An Introduction to Multivariate Statistical Analysis*, second edition. New York: Wiley.

Bolch, B. W. and Huang, C. J. 1974. *Multivariate Statistical Methods for Business and Economics*. New Jersey: Prentice-Hall.

Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. 1983. *Graphical Methods for Data Analysis*. Belmont, CA: Wadsworth International Group.

Hair, J., Anderson, R., and Tatham, R. 1992. *Multivariate Data Analysis*, third edition. Englewood Cliffs, NJ: Prentice-Hall.

Johnson, R. A. and Wichern, D. W. 1988. *Applied Multivariate Statistical Analysis*, second edition. Englewood Cliffs, NJ: Prentice-Hall.

Milligan, G. W. 1980. "An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms," *Psychometrika* **45**:325-342.

Morrison, D. F. 1990. *Multivariate Statistical Methods*, third edition. New York: McGraw-Hill Publising Company.

Chapter 2

Using the Factor Analysis

Background Information

Factor analysis is a technique useful for reducing information in a large number of variables into a smaller set, while losing only a minimal amount of information. Variables in a factor analysis model represent a linear function containing a small number of common factor variables and a single specific variable. The common factors create covariances among the responses, while the specific variable adds only to the variances of specific responses.

The general purposes of factor analysis are to:

- identify a set of dimensions you cannot easily observe in a large set of variables

- devise a means for combining or condensing large numbers of observations into distinctly different groups within a larger population

- create an entirely new set of a smaller number of variables to partially or completely replace the original set of variables you can then use in regression, correlation, cluster, or discriminant analysis.

There are a number of general steps you must consider when you are deciding which factor analysis technique to use:

- your problem
- the correlation matrix
- the type of model
- how you will extract the factors
- the number of factors you will extract
- whether or not you will rotate the factors
- how you will rotate the factors
- how you will interpret the rotated factors and their scores.

Generally, you would not use factor analysis if the data contain fewer than 50 observations but preferably 100 or more.

Although there are several types of general factor models, the two that analysts use most are principal component analysis and common factor analysis.

Principal Component Analysis

Principal component analysis is the amount of variance in a variable that is shared by all the variables in the analysis. The analysis summarizes and groups nearly all of the original information into a smaller number of factors that can then be used for estimation purposes.

Common Factor Analysis

Common factor analysis is the amount of variance that can be related only to a specific variable. The analysis identifies conditions or proportions that are not easily recognizable.

Besides determining the type of factor analysis you want to use, you must also determine if you want to extract the factors orthogonally. Mathematically, orthogonal solutions are simpler; they extract factors so the axes remain at 90 degrees, and each factor remains independent of the others.

Factor rotation is another important concept in factor analysis. Rotation means that the factors are turned until they reach another position. The primary reason for rotating factors is to attain a simpler and more meaningful solution. Rotation is generally desirable because it simplifies the rows and/or columns of the matrix.

STATGRAPHICS *Plus* contains three primary approaches to rotation: Quartimax, Varimax, and Equimax.

Quartimax

This approach simplifies the rows of the factor matrix.

Varimax

This approach simplifies the columns of the factor matrix.

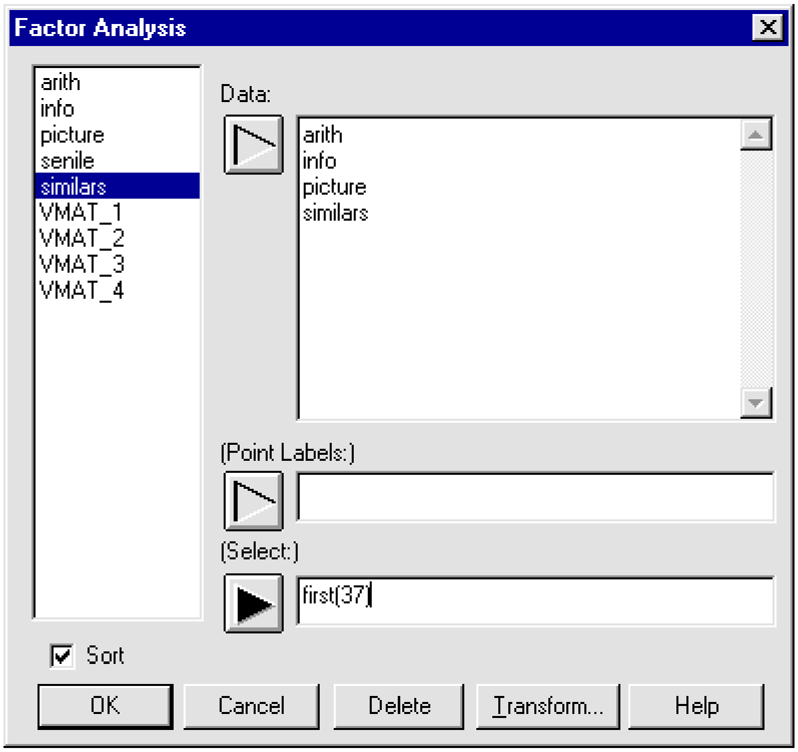
Equimax

This approach simplifies the rows or the columns by combining aspects of each of the above two approaches.

STATGRAPHICS *Plus* also contains two methods you can use to determine the number of factors to extract from an analysis: minimum eigenvalues or number of factors (see the Tabular Options section for descriptions of the two methods).

Using Factor Analysis

To access the analysis, choose SPECIAL... MULTIVARIATE METHODS... FACTOR ANALYSIS... from the Menu bar to display the Factor Analysis dialog box shown in Figure 2-1.



Factor Analysis

Data:
arith
info
picture
similars

(Point Labels:)

(Select:)
first(37)

☑ Sort

Figure 2-1. Factor Analysis Dialog Box

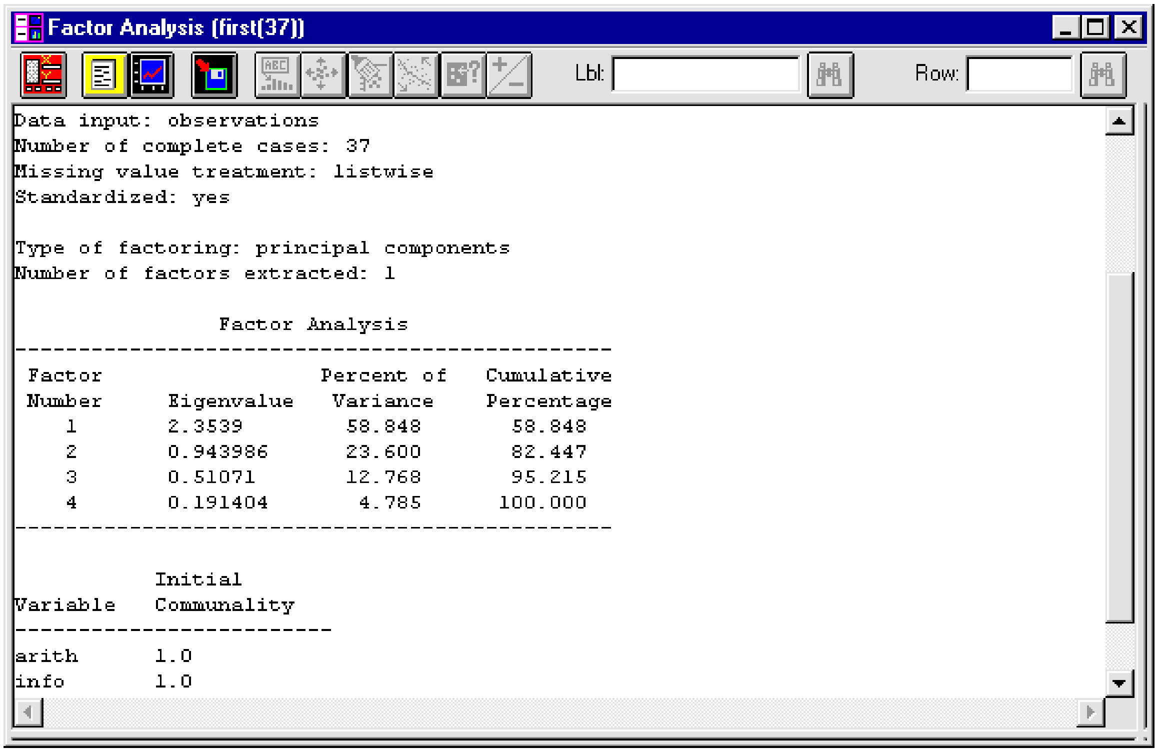
Tabular Options

Analysis Summary

The Analysis Summary option creates a summary of the analysis that displays the name of each of the variables, the type of data you entered, the number of complete cases, the name of the missing value treatment, whether or not the data are standardized, and the type of factoring used (see Figure 2-2).

The summary then displays the factor numbers, eigenvalues, percentage of variance, and cumulative percentage for each factor. The end of the summary displays the name of each variable and the initial communalities.

Communality is the amount of variance each of the original variables shares with all the other variables in the analysis. For the Principal Components method, the communalities are all ones. For the Classical method, the communalities are the multiple correlations for each variable when regressed against the other variables.



```
Data input: observations
Number of complete cases: 37
Missing value treatment: listwise
Standardized: yes

Type of factoring: principal components
Number of factors extracted: 1

                Factor Analysis
-----------------------------------------------
 Factor                 Percent of   Cumulative
 Number     Eigenvalue   Variance    Percentage
    1       2.3539        58.848       58.848
    2       0.943986      23.600       82.447
    3       0.51071       12.768       95.215
    4       0.191404       4.785      100.000
-----------------------------------------------


           Initial
Variable   Communality
-------------------------
arith      1.0
info       1.0
```

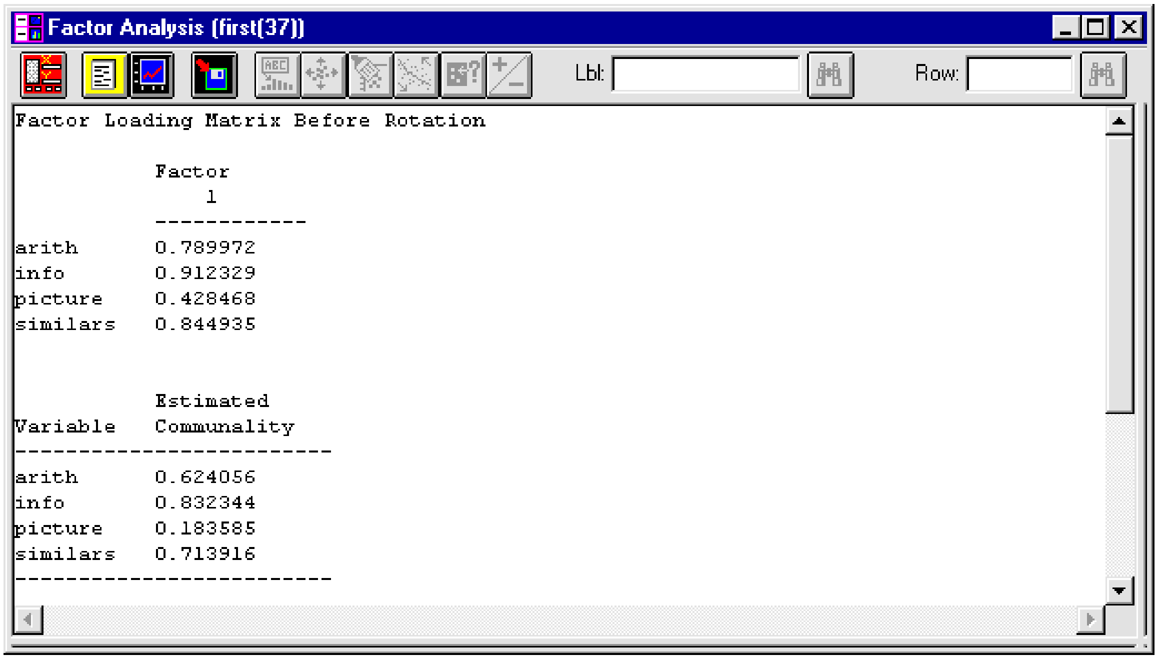
Figure 2-2. Analysis Summary

Use the *Factor Analysis Options* dialog box to indicate how missing values will be treated; to indicate if the variables will be standardized; to choose the technique that will be used to estimate the values for the factors and the type of rotation that will be used to rotate the factors; to choose how the factors will be extracted from the analysis; and to enter values for the minimum eigenvalue and the number of factors that will be included.

You can also use the Estimation... command to display the Estimation Options dialog box that you use to limit and stop the factor-rotation process, as well as the Communalities... command to display the Communalities Options dialog box that you use to choose or enter the name of the variable that contains the communalities that will be used.

Extraction Statistics

The Extraction Statistics option creates a factor-loading matrix that shows the loadings for each of the factors included in the analysis before the factor is rotated (see Figure 2-3).



```
Factor Loading Matrix Before Rotation

           Factor
              1
           ------------
arith      0.789972
info       0.912329
picture    0.428468
similars   0.844935


           Estimated
Variable   Communality
-------------------------
arith      0.624056
info       0.832344
picture    0.183585
similars   0.713916
-------------------------
```

Figure 2-3. Extraction Statistics

Rotation Statistics

The Rotation Statistics option displays the rotated factor-loading matrix (see Figure 2-4). The table also displays the name of the rotation method and the extimated communality for each of the variables.

Factor Scores

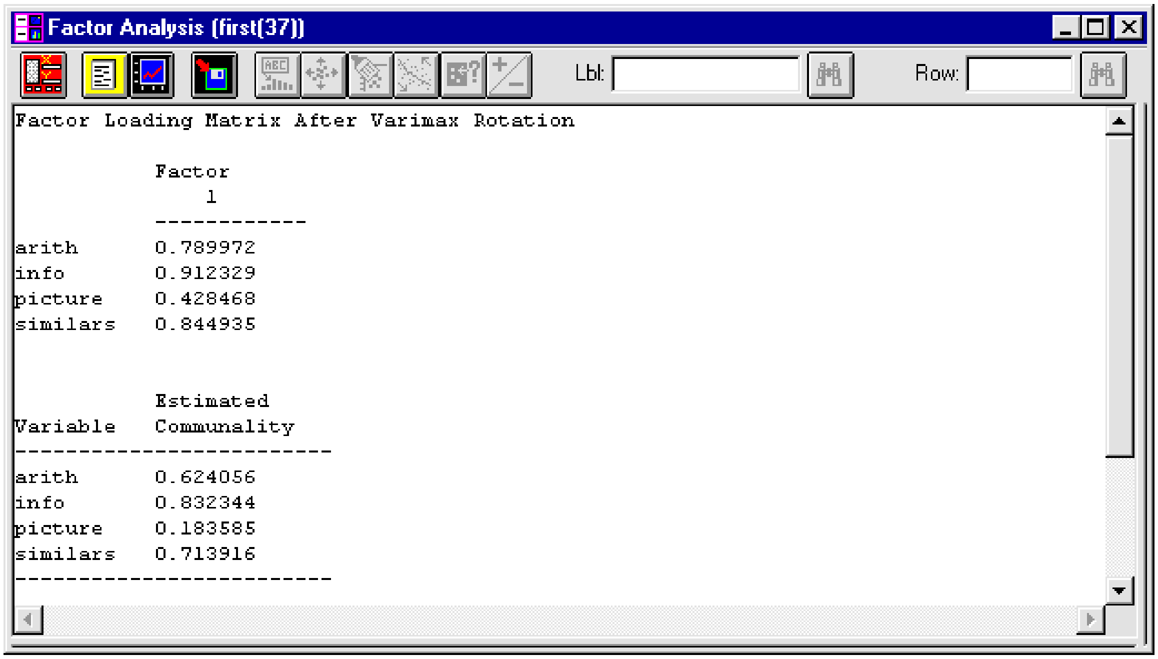
The Factor Scores option creates a table of the factor scores for each row of the data file (see Figure 2-5).

Graphical Options

Scree Plot

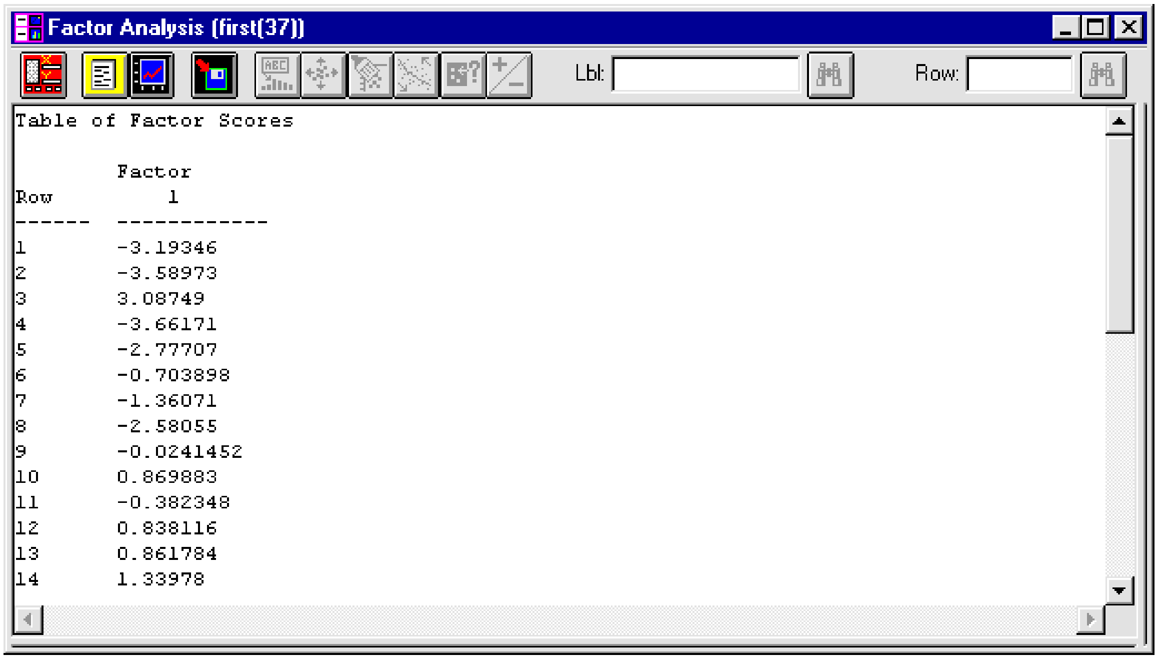
The Scree Plot option creates a plot of the eigenvalues for each of the factors in the analysis (see Figure 2-6). The eigenvalues are proportional to the percent of variability in the data that can be attributed to the factors. If you use the Minimum Eigenvalue option, a horizontal line is plotted at that value.

Use the *Scree Plot Options* dialog box to indicate if eigenvalues or the percent of variance will appear as data on the plot.



```
Factor Loading Matrix After Varimax Rotation

           Factor
              1
           ------------
arith      0.789972
info       0.912329
picture    0.428468
similars   0.844935


           Estimated
Variable   Communality
-------------------------
arith      0.624056
info       0.832344
picture    0.183585
similars   0.713916
-------------------------
```

Figure 2-4. Rotation Statistics



```
Table of Factor Scores

        Factor
Row        1
------  ------------
1       -3.19346
2       -3.58973
3       3.08749
4       -3.66171
5       -2.77707
6       -0.703898
7       -1.36071
8       -2.58055
9       -0.0241452
10      0.869883
11      -0.382348
12      0.838116
13      0.861784
14      1.33978
```

Figure 2-5. Factor Scores

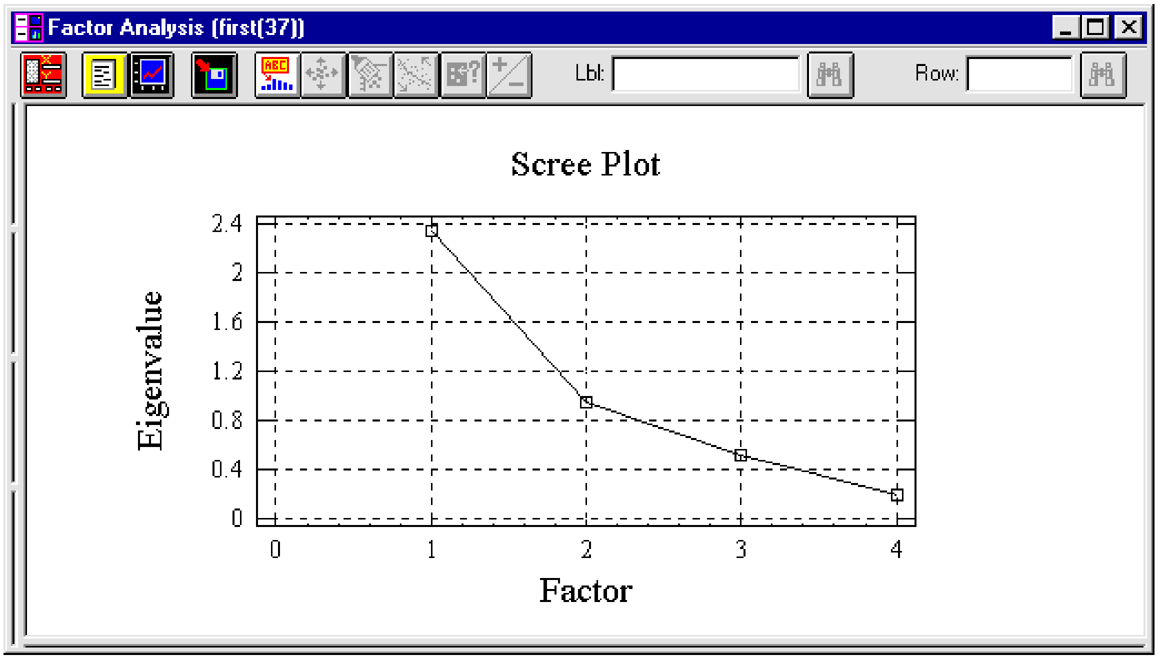


Figure 2-6. Scree Plot

2D Scatterplot

The 2D (Two-Dimensional) Scatterplot option creates a two-dimensional scatterplot of the values for two factors (see Figure 2-7). One point appears for each row in the data file. The plot is helpful in interpreting the factors and in understanding how the factors compare.

Use the *2D Scatterplot Options* dialog box to enter the number of the factors you want plotted on the X- and Y-axes.

3D Scatterplot

The 3D (Three-Dimensional) Scatterplot option creates a three-dimensional scatterplot of the values for three factors (see Figure 2-8). One point appears for each row in the data file. The plot is helpful in interpreting the factors and in understanding how the factors compare.

Use the *3D Scatterplot Options dialog box* to enter the numbers of the factors you want to appear on the X-, Y-, and Z-axes.

2D Factor Plot

The 2D (Two-Dimensional) Factor Plot option creates a two-dimensional plot of the weight loadings for each chosen factor (see Figure 2-9). Reference lines are drawn at 0.0 in each dimension. A weight close to 0.0 indicates that the variable contributes little to the factor.

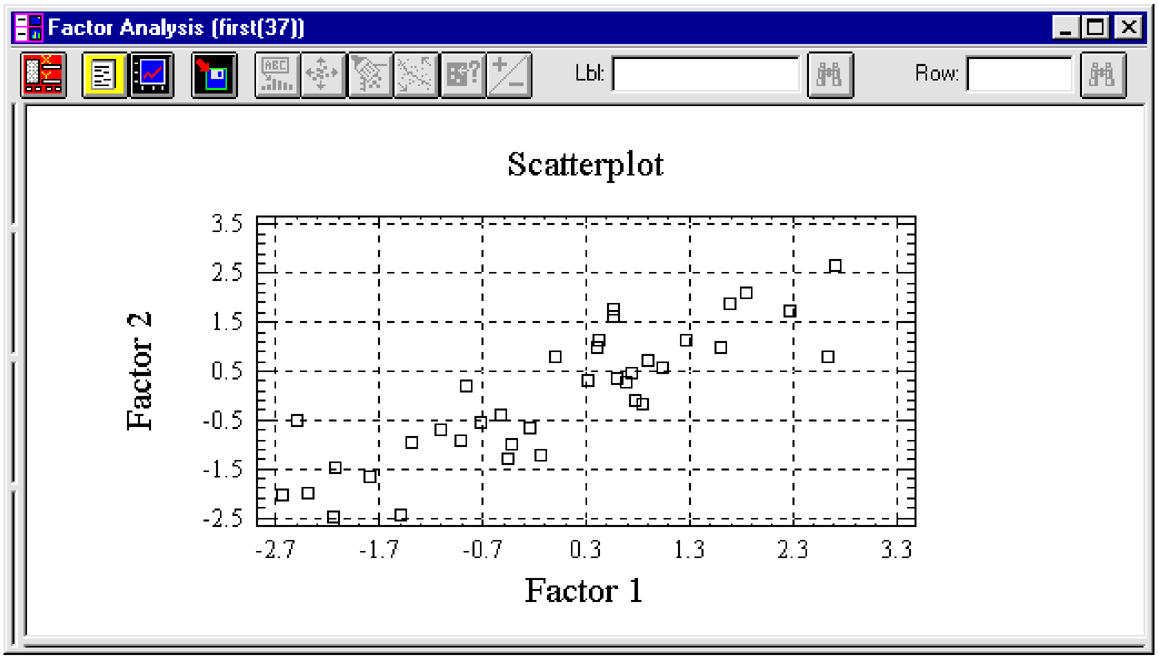


Figure 2-7. 2D Scatterplot

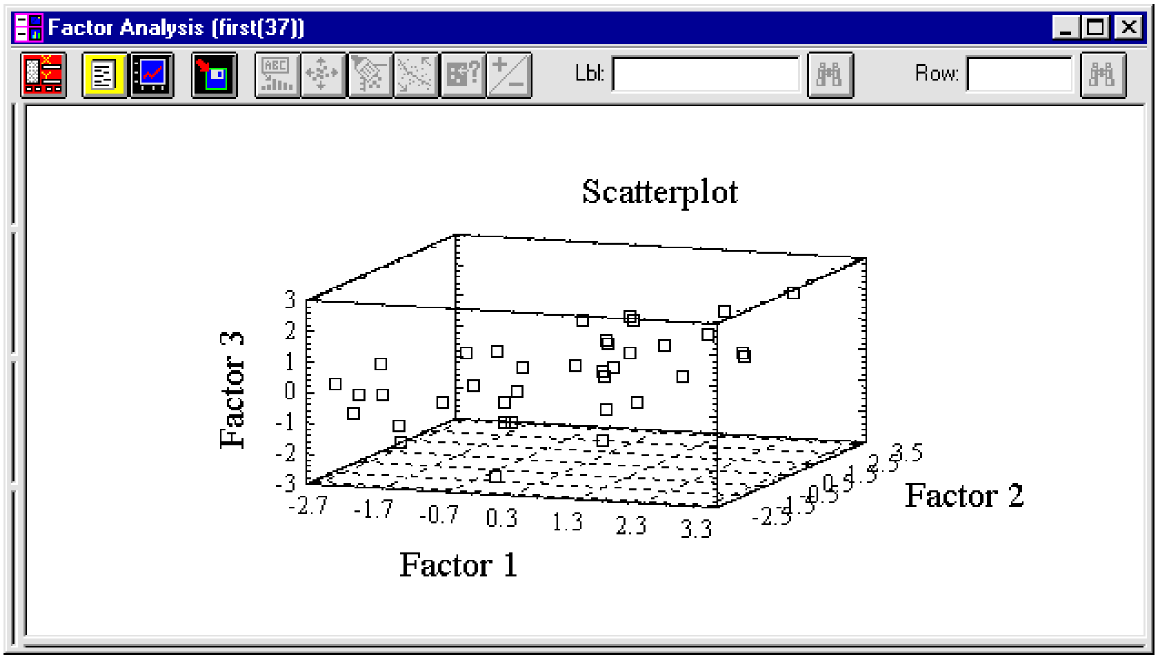


Figure 2-8. 3D Scatterplot

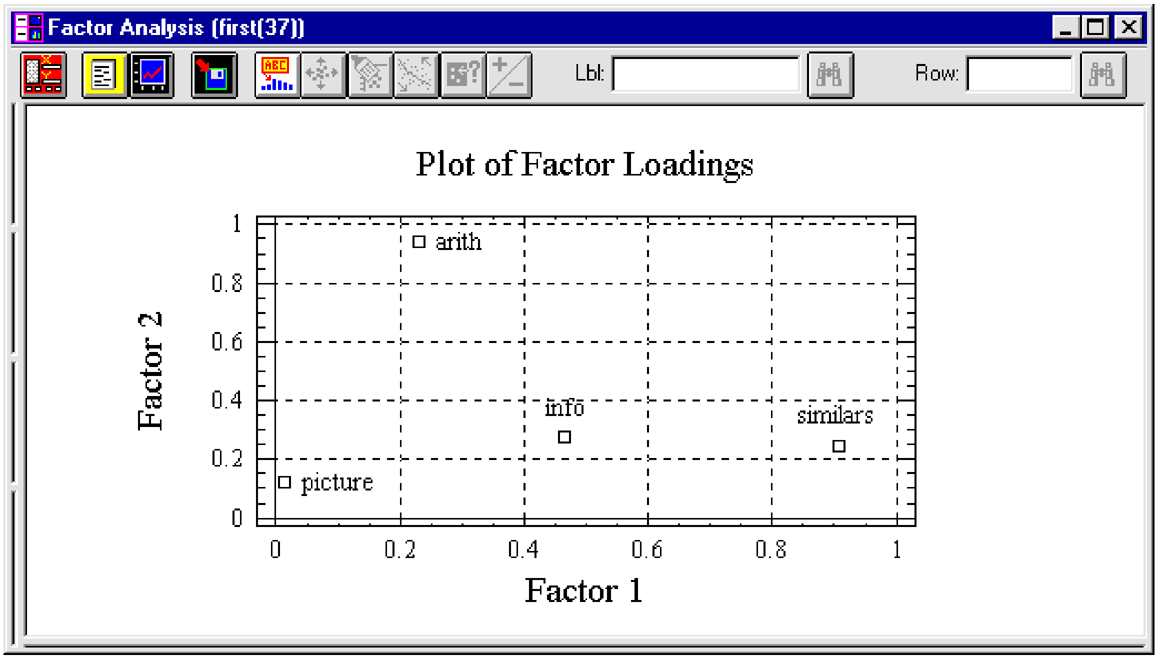


Figure 2-9. 2D Factor Plot

3D Factor Plot

The 3D (Three-Dimensional) Factor Plot option creates a three-dimensional plot of the weight loadings for each chosen variable (see Figure 2-10). One point appears for each variable. A weight close to 0.0 indicates that the variable contributes little to the factor.

Use the *3D Factor Weights Plot Options* dialog box to enter the number of the factor that will be plotted on the X-, Y-, and Z-axes. Reference lines are drawn at 0.0 in each dimension. A weight close to 0.0 indicates that the variable contributes little to the factor.

Saving the Results

The Save Results Options dialog box allows you to choose the results you want to save. There are six selections: Eigenvalues, Factor Matrix, Rotated Factor Matrix, Transition Matrix, Communalities, and Factor Scores.

You can also use the Target Variables text boxes to enter the names of the variables in which you want to save the values generated during the analysis. You can enter new names or accept the defaults.

Note: To access the Save Results Options dialog box, click the Save Results button on the Analysis toolbar (the fourth button from the left).

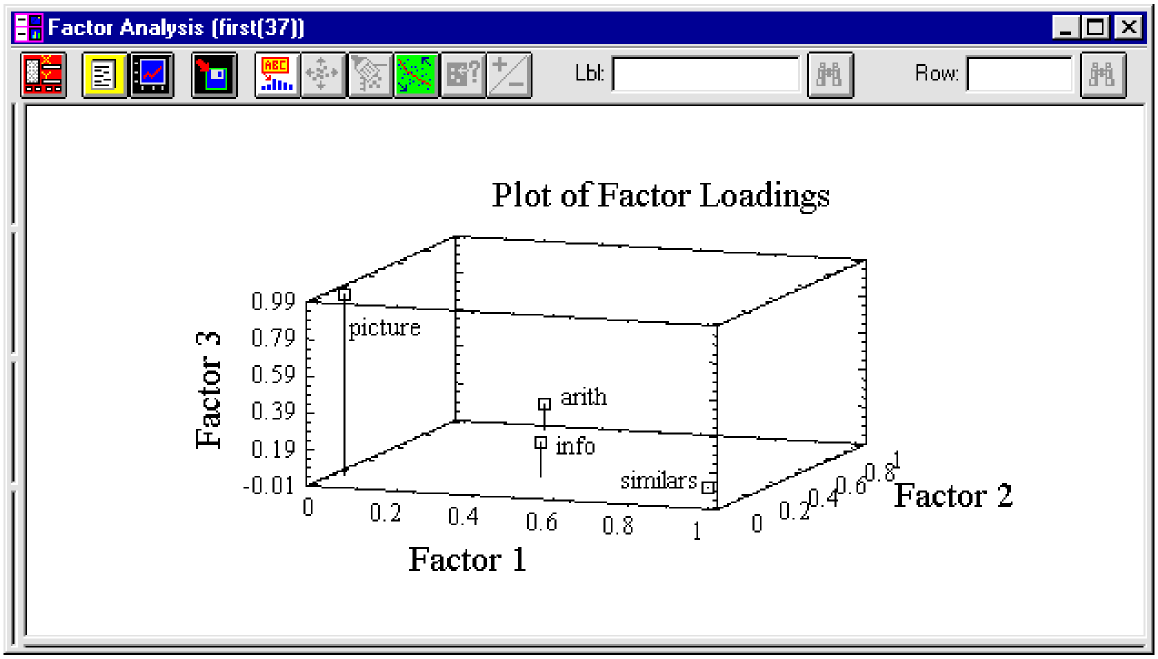


Figure 2-10. 3D Factor Plot

References

Anderson, T. W. 1984. *An Introduction to Multivariate Statistical Analysis*, second edition. New York: Wiley.

Cooper, C. B. J. 1983. “Factor Analysis: An Overview,” *The American Statistician*, **38**:141-148.

Hair, J., Anderson, R., and Tatham, R. 1992. *Multivariate Data Analysis*, third edition. Englewood Cliffs, New Jersey: Prentice-Hall.

Johnson, R. A. and Wichern, D. W. 1988. *Applied Multivariate Statistical Analysis*, second edition. Englewood Cliffs, New Jersey: Prentice-Hall.

Morrison, D. F. 1990. *Multivariate Statistical Methods*, third edition. New York: McGraw-Hill.

Tatsuoka, M. M. 1971. *Multivariate Analysis*. New York: Wiley.

Chapter 3

Using the Principal Components Analysis

Background Information

Principal components analysis differs from factor analysis in that you use it whenever you want to use uncorrelated linear combinations of variables. That is, principal components analysis is a factor-analysis technique that reduces the dimensions of a set of variables by reconstructing them into uncorrelated combinations.

The analysis combines the variables that account for the largest amount of variance to form the first principal component. The second principal component accounts for the next largest amount of variance, and so on, until the total sample variance is combined into component groups. Each successive component explains progressively smaller portions of the variance in the total sample. All of the components are uncorrelated with each other.

Often, a few components will account for 75 to 90 percent of the variance in an analysis. These components are then the ones you use to plot the data. Other uses for principal components analysis include various forms of regression analysis and classification and discrimination problems.

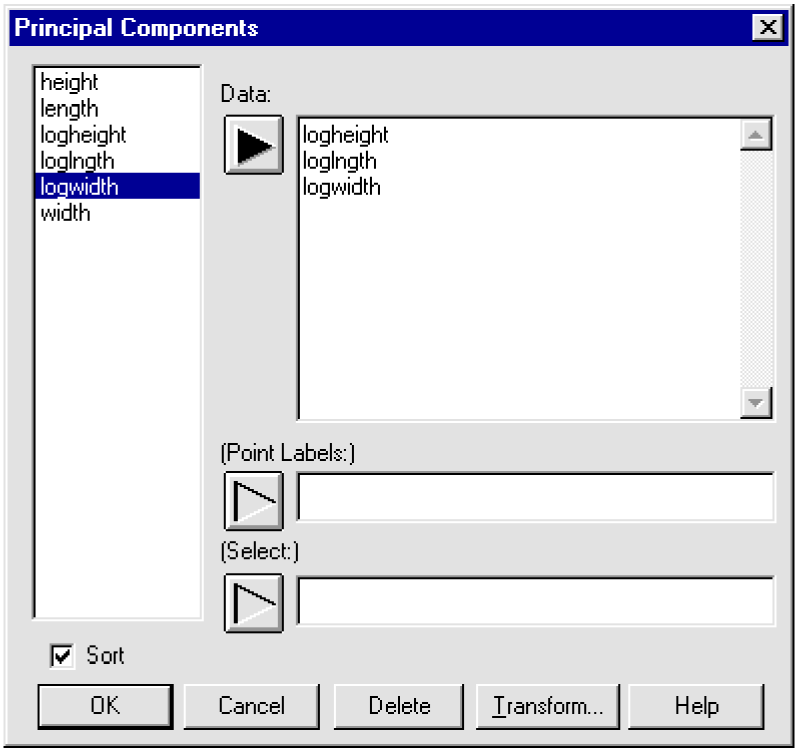
Using Principal Components Analysis

To access the analysis, choose SPECIAL... MULTIVARIATE METHODS... PRINCIPAL COMPONENTS... from the Menu bar to display the Principal Components Analysis dialog box shown in Figure 3-1.

Tabular Options

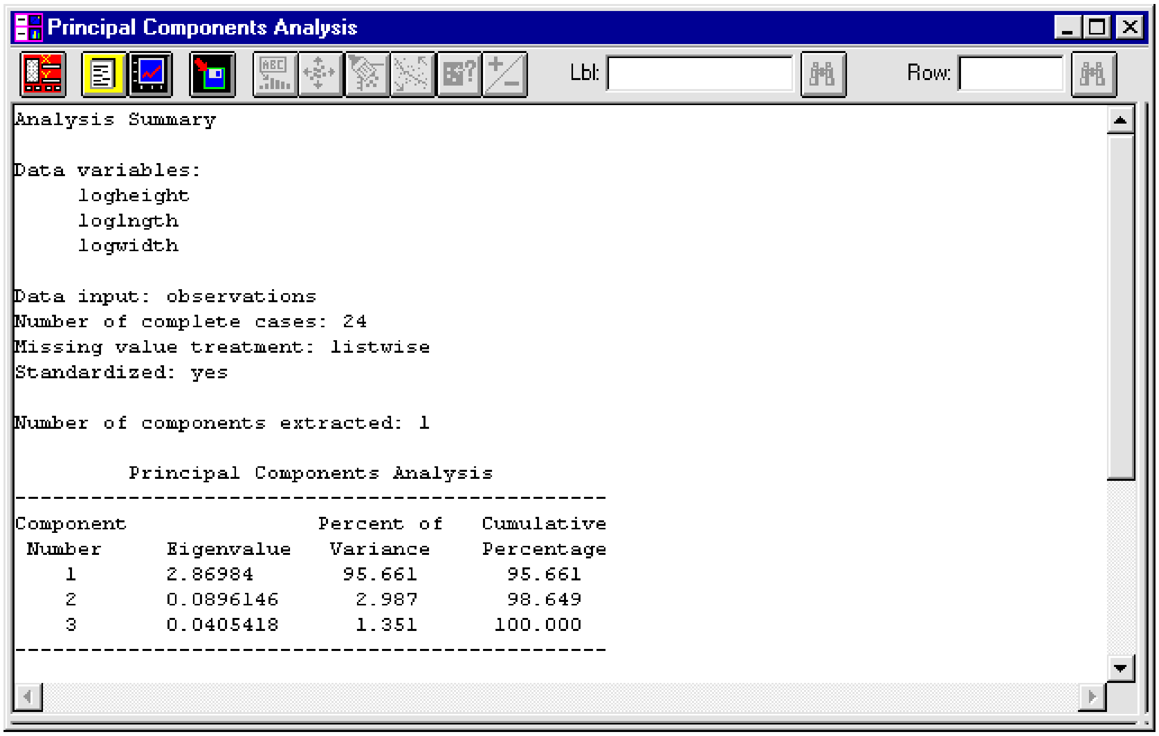
Analysis Summary

The Analysis Summary option creates a summary of the analysis that first displays the names of the chosen variables, the type of data entered, the number of complete cases, the type of treatment used for missing values, indicates if the data were standardized, and displays the number of components extracted from the analysis (see Figure 3-2).



Data:
```
logheight
loglngth
logwidth
```

Figure 3-1. Principal Components Analysis Dialog Box



```
Analysis Summary

Data variables:
     logheight
     loglngth
     logwidth

Data input: observations
Number of complete cases: 24
Missing value treatment: listwise
Standardized: yes

Number of components extracted: 1

         Principal Components Analysis
-----------------------------------------------
Component              Percent of   Cumulative
 Number     Eigenvalue  Variance    Percentage
    1       2.86984      95.661       95.661
    2       0.0896146     2.987       98.649
    3       0.0405418     1.351      100.000
-----------------------------------------------
```

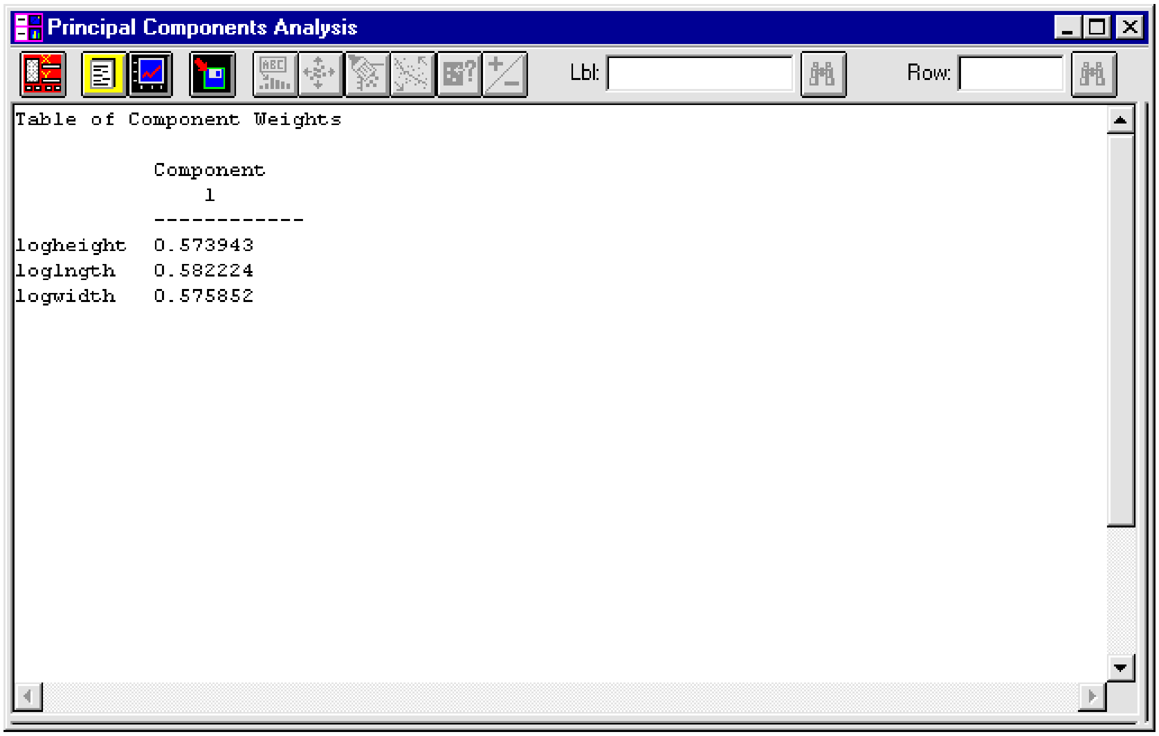
Figure 3-2. Analysis Summary

The summary then displays a table of statistics run for the analysis, including the component number, the eigenvalue, the percent of variance, and the cumulative percentage.

Use the *Principal Components Options* dialog box to choose the missing value treatment, to indicate if the data will be standardized, to indicate the method that will be used to extract components from the analysis, and to enter values for the minimum eigenvalue and number of components.

Component Weights

The Component Weights option creates the values that are used in the equations for the principal components (see Figure 3-3). The values for the variables in the equation are standardized by subtracting the means (0.573943*logheight + 0.582224*loglngth + 0.575852 *logwidth), and dividing by the standard deviations.

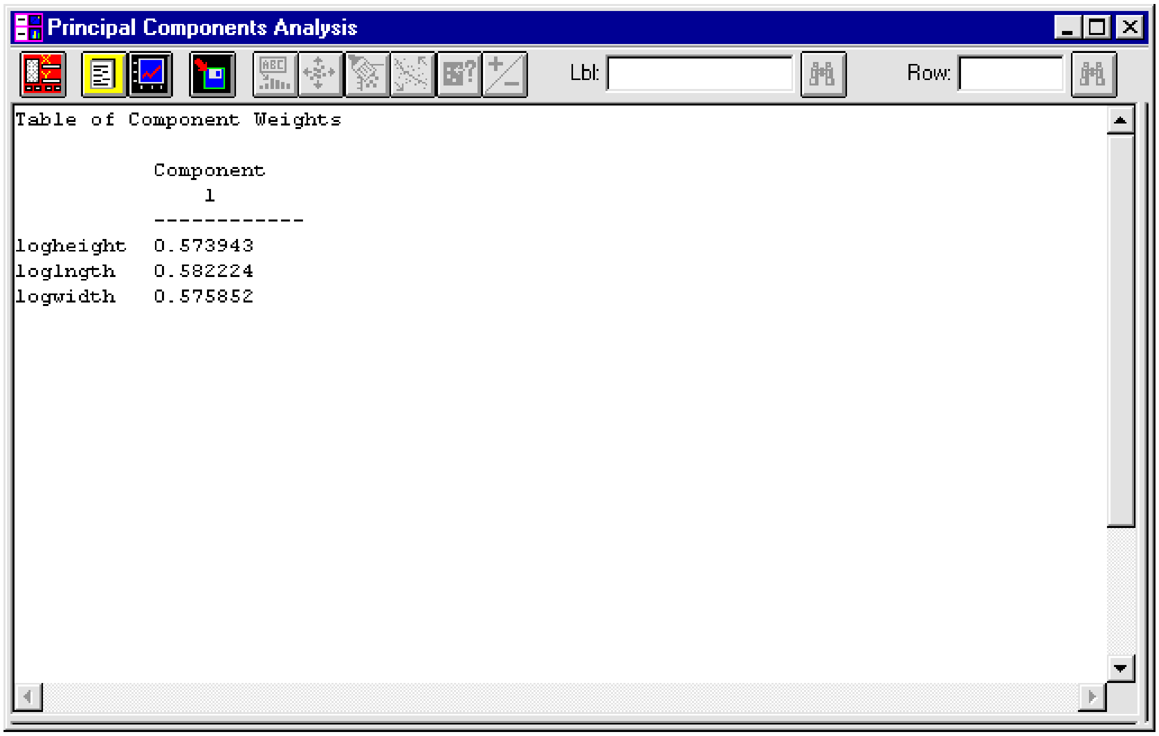


```
Table of Component Weights

             Component
                 1
             ------------
logheight    0.573943
loglngth     0.582224
logwidth     0.575852
```

Figure 3-3. Component Weights

Data Table

The Data Table option creates a table that shows the values for each principal component for each row of the data table (see Figure 3-4).



```
Table of Component Weights

             Component
                 1
             ------------
logheight    0.573943
loglngth     0.582224
logwidth     0.575852
```

Figure 3-4. Data Table

Graphical Options

Scree Plot

The Scree Plot option creates a plot of the eigenvalues for each of the principal components (the total variance contributed by each component) (see Figure 3-5). The eigenvalues are proportional to the percentage of variability in the data that can be attributed to the components.

Use the plot to see the gap between the steepest slope of the large components and a progressive downward path (the scree). The plot identifies at least one and sometimes two or three of the most significant components.

Use the *Scree Plot Options* dialog box to indicate whether eigenvalues or the percent of variance will be displayed on the plot.

2D Scatterplot

The 2D (Two-Dimensional) Scatterplot option creates a plot of the values for two principal components (see Figure 3-6).

Use the *2D Scatterplot Options* dialog box to enter the number of the components that will be plotted on the X- and Y-axes.

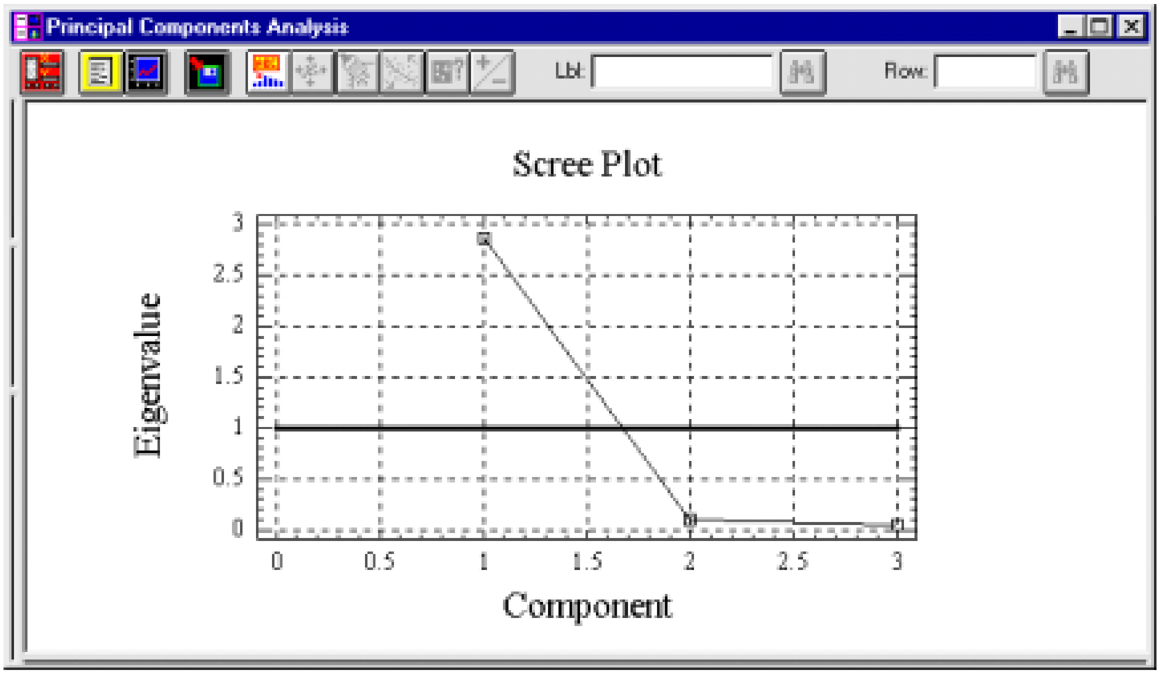


Figure 3-5. Scree Plot

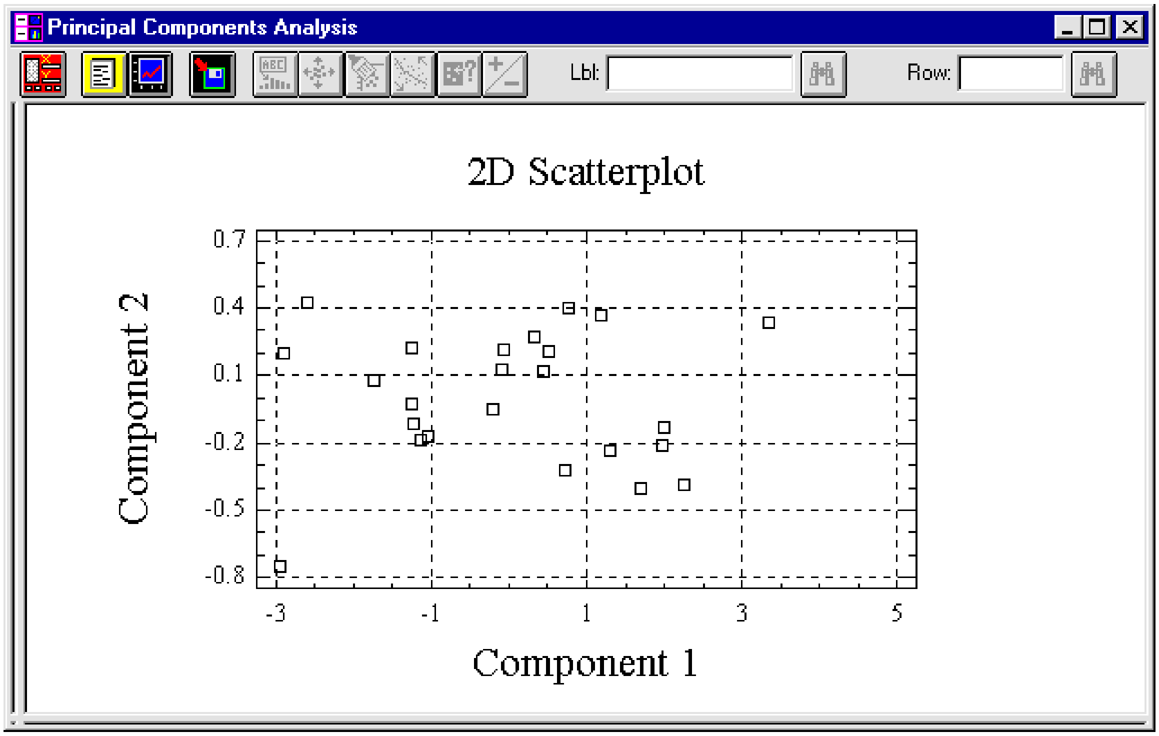


Figure 3-6. 2D Scatterplot

3D Scatterplot

The 3D (Three-Dimensional) Scatterplot option creates a plot of the values for three principal components (see Figure 3-7). One point displays for each row in the data file.

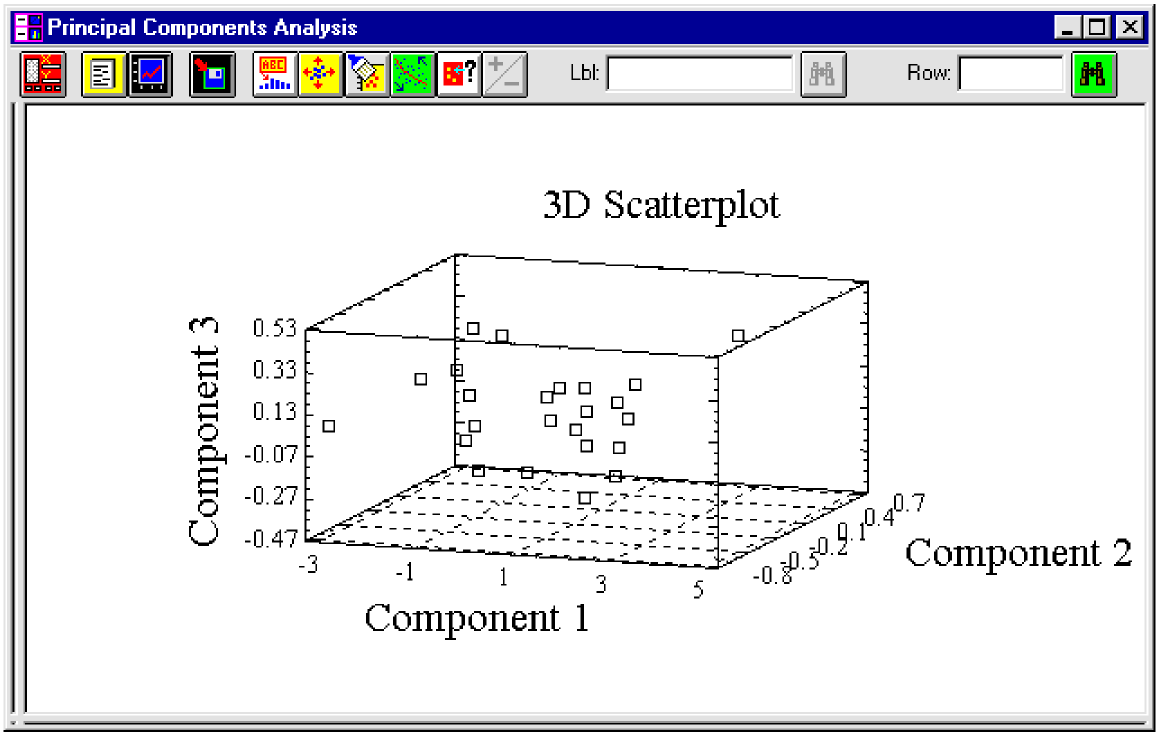


Figure 3-7. 3D Scatterplot

Use the *3D Scatterplot Options* dialog box to enter the number of the components that will be plotted on the X-, Y-, and Z-axes.

2D Component Plot

The 2D (Two-Dimensional) Component Plot option creates a plot that shows the weights for the chosen principal components (see Figure 3-8). One point appears on the plot for each variable in the analysis. Reference lines are drawn at 0.0 for each dimension. A weight close to 0.0 indicates that the variable contributes little to the component.

Use the *2D Component Plot Options* dialog box to enter the number of the components that will be plotted on the X- and Y-axes.

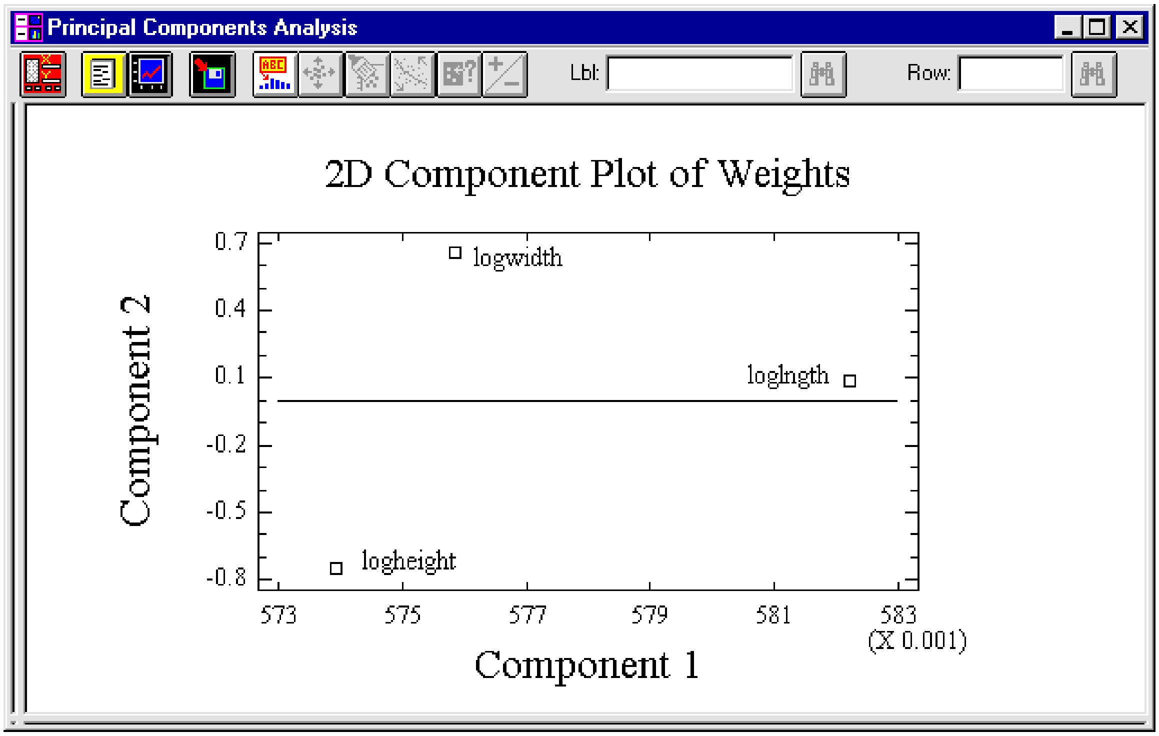


Figure 3-8. 2D Component Plot

3D Component Plot

The 3D (Three-Dimensional) Component Plot option creates a plot that shows the weights for the chosen principal components (see Figure 3-9). One point appears on the plot for each variable in the analysis. Reference lines are drawn at 0.0 for each dimension. A weight close to 0.0 indicates that the variable contributes little to the component.

Use the *3D Component Plot Options* dialog box to enter the numbers that will be plotted on the X-, Y-, and Z-axes.

2D Biplot

The 2D Biplot option creates a plot of the chosen principal components (see Figure 3-10). A point appears on the plot for each row in the data file. Reference lines are drawn for each of the variables that represent the location of the variable in the location in the space of the component. A weight close to 0.0 indicates that the variable contributes little to that component.

Use the *2D Biplot Options* dialog box to enter the number of the components that will be plotted on the X- and Y-axes.

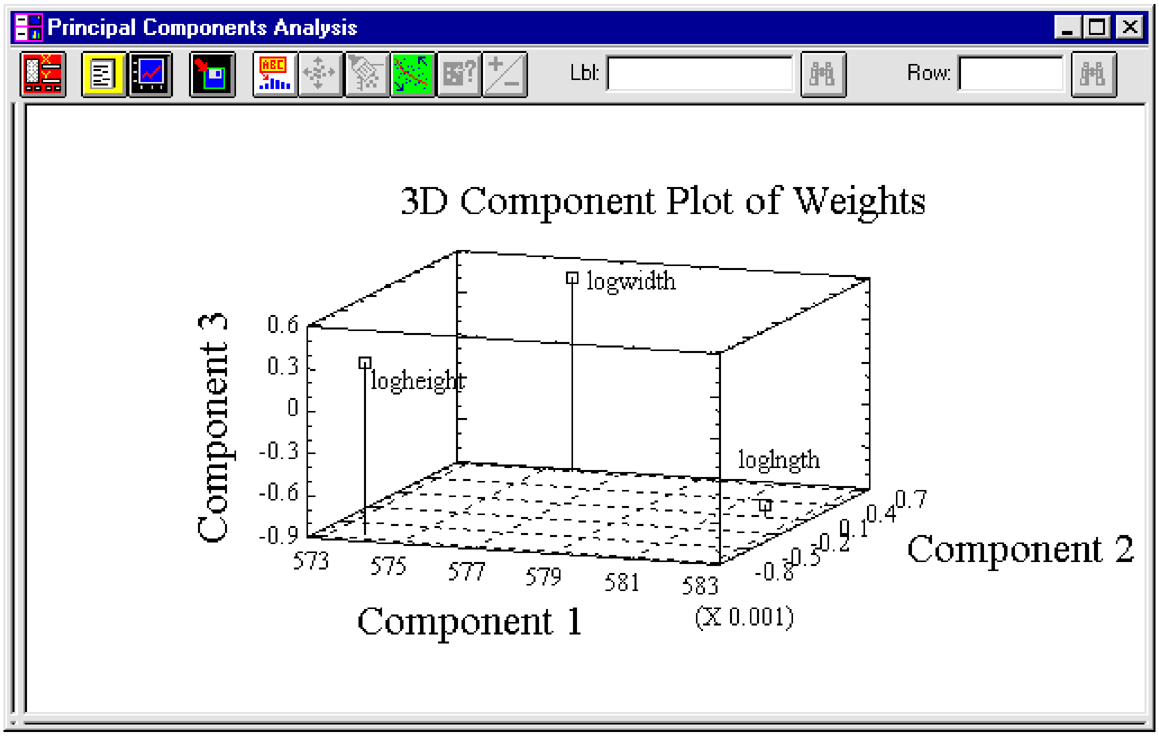


Figure 3-9. 3D Component Plot

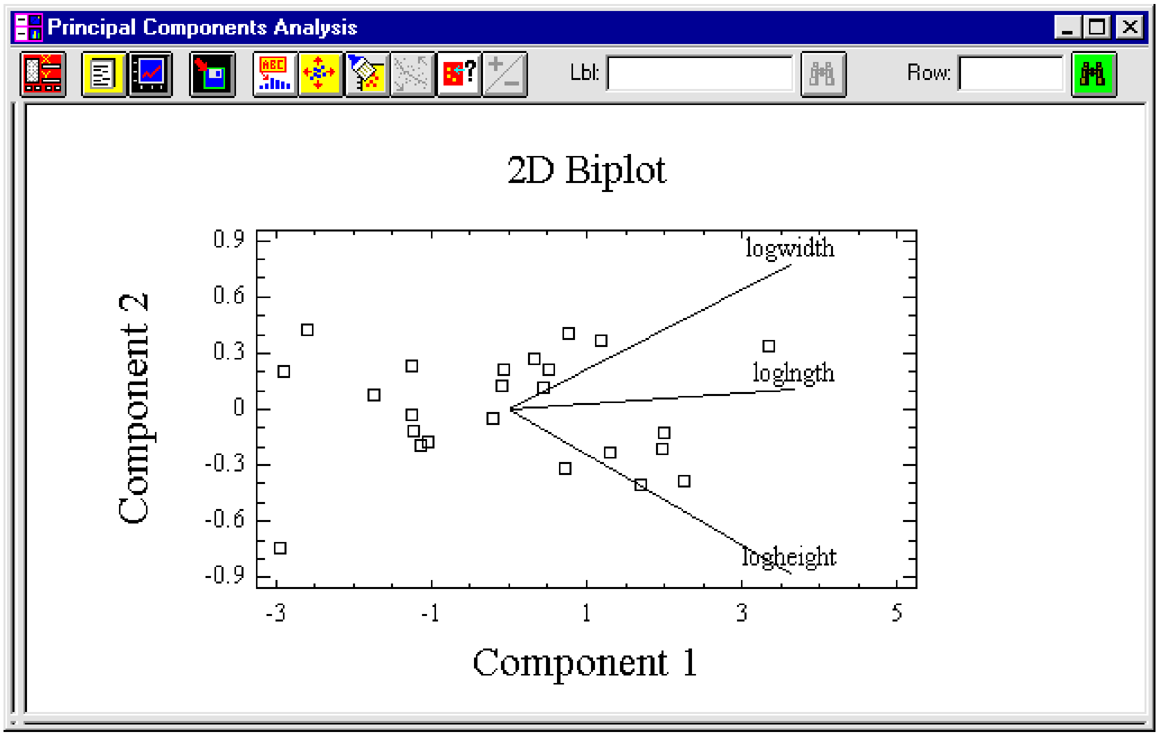


Figure 3-10. 2D Biplot

3D Biplot

The 3D Biplot option creates a three-dimensional biplot that has rays representing the magnitude and direction for each variable (see Figure 3-11).

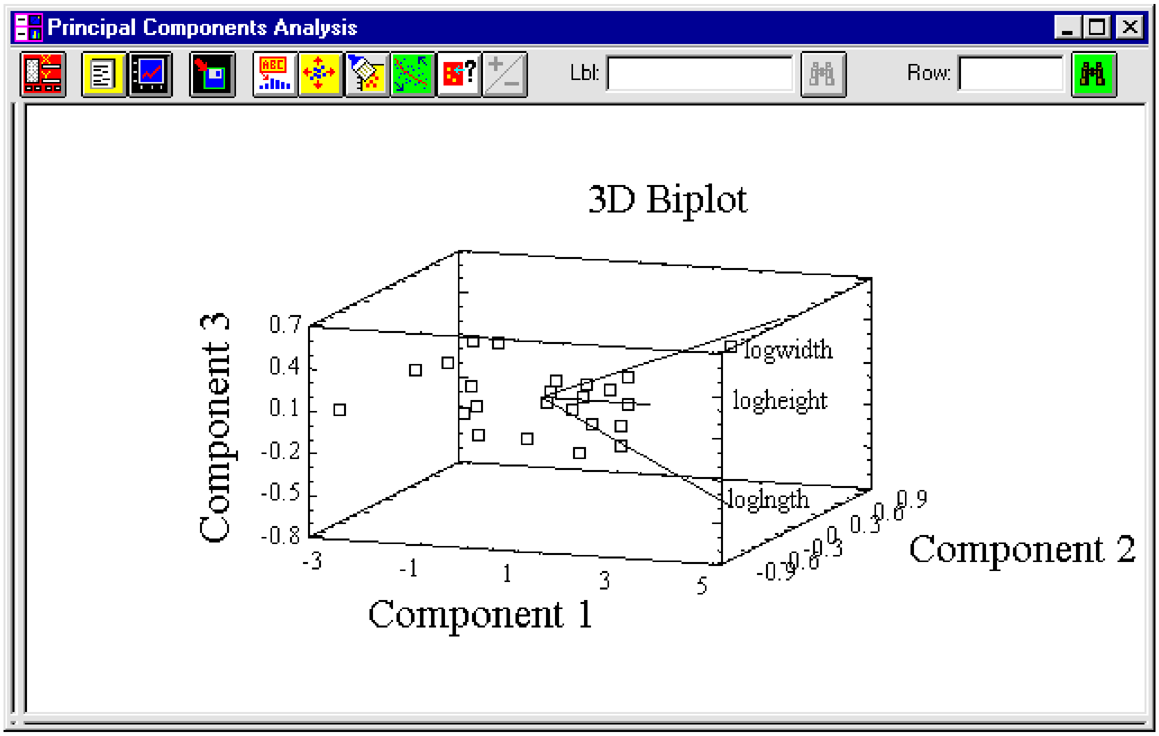


Figure 3-11. 3D Biplot

Use the *3D Biplot Options* dialog box to enter the number of the components that will be plotted on the X-, Y-, and Z-axes.

Saving the Results

The Save Results Options dialog box allows you to choose the results you want to save. There are three selections: Eigenvalues, Component Weights, and Principal Components.

You can also use the Target Variables text boxes to enter the names of the variables in which you want to save the values generated during the analysis. You can enter new names or accept the defaults.

Note: To access the Save Results Options dialog box, click the Save Results button on the Analysis toolbar (the fourth button from the left).

References

Anderson, T. W. 1958. *An Introduction to Multivariate Statistical Analysis.* New York: Wiley.

Hair, J., Anderson, R., and Tatham, R. 1992. *Multivariate Data Analysis*, third edition. Englewood Cliffs, NJ: Prentice-Hall.

Johnson, R. A. and Wichern, D. W. 1988. *Applied Multivariate Statistical Analysis*, second edition. Englewood Cliffs, NJ: Prentice-Hall.

Morrison, D. F. 1990. *Multivariate Statistical Methods*, third edition. New York: Wiley.

Chapter 4

Using the Discriminant Analysis

Background Information

Complex problems and the results of bad decisions frequently force researchers to look for more objective ways to predict outcomes. A classic example often cited in statistical literature is creditworthiness where, based on a collection of variables, potential borrowers are identified as either good or bad credit risks.

Discriminant analysis is the statistical technique that is most commonly used to solve these types of problems. Its use is appropriate when you can classify data into two or more groups, and when you want to find one or more functions of quantitative measurements that can help you discriminate among the known groups. The objective of the analysis is to provide a method for predicting which group a new case is most likely to fall into, or to obtain a small number of useful predictor variables.

Discriminant analysis is capable of handling either two groups or multiple groups (three or more). When two classifications are involved, it is known as two-group discriminant analysis. When three or more classifications are identified, it is known as multiple discriminant analysis. The concept of discriminant analysis involves forming linear combinations of independent (predictor) variables, which become the basis for group classifications.

Discriminant analysis is appropriate for testing the hypothesis that the group means of two or more groups are equal. This is done by multiplying each independent variable by its corresponding weight and adding the products together, which results in a single composite discriminant score for each individual in the analysis. Averaging the scores derives a group centroid. If the analysis involves two groups there are two centroids; in three groups there are three centroids, and so on. Comparing the centroids shows how far apart the groups are along the dimension you are testing.

Applying and interpreting discriminant analysis is similar to regression analysis, where a linear combination of measurements for two or more independent variables describes or predicts the behavior of a single dependent variable. The most significant difference is that you use discriminant analysis for problems where the dependent variable is categorical versus regression where the dependent variable is metric.

The objectives for applying discriminant analysis include:

- determining if there are statistically significant differences among two or more groups
- establishing procedures for classifying units into groups
- determining which independent variables account for most of the differences of two or more groups.

Discriminant analysis involves three steps: deriviation, validation, and interpretion. In the first step, you must first select the variables, test the validity of the discriminant function, determine a computational method, and assess the level of significance.

The second step involves determining the reason for developing classification matrices, deciding how well the groups are classified into statistical groups, determining the criterion against which each individual score is judged, constructing the classification matrices, and interpreting the discriminant functions to determine the accuracy of their classification.

The last step, interpretation, involves examining the discriminant functions to determine the importance of each independent variable in discriminating between the groups, then examining the group means for each important variable to outline the differences in the groups.

Using Discriminant Analysis

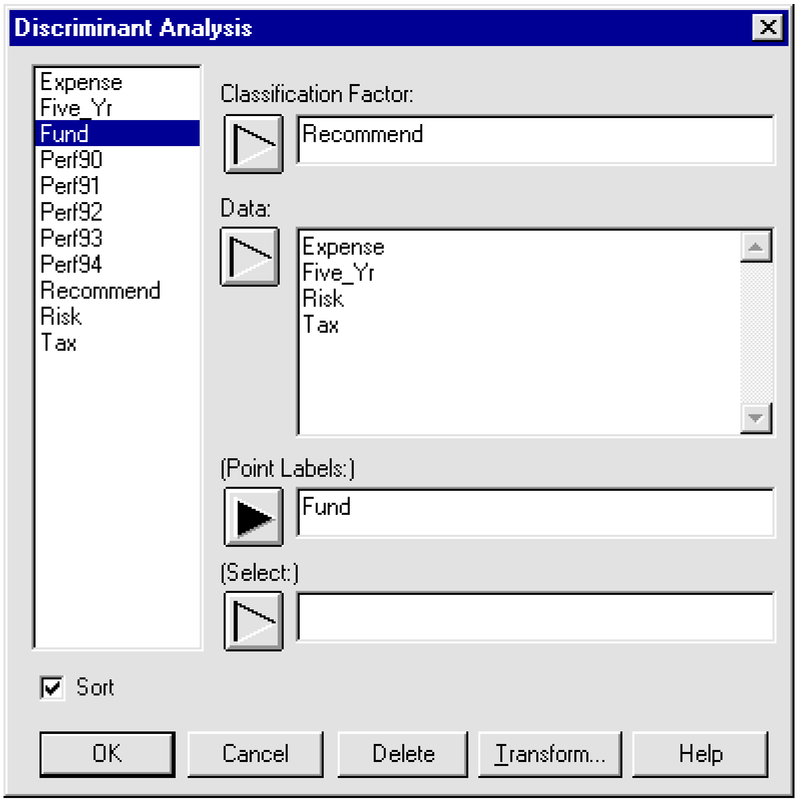
The Discriminant Analysis in STATGRAPHICS *Plus* allows you to generate discriminant functions from the variables in a dataset, and return values for the discriminant functions in each case. The analysis assumes that the variables are drawn from populations that have multivariate normal distributions and that the variables have equal variances.

To access the analysis, choose SPECIAL... MULTIVARIATE METHODS... DISCRIMINANT ANALYSIS... from the Menu bar to display the Discriminant Analysis dialog box shown in Figure 4-1.

Tabular Options

Analysis Summary

The Analysis Summary option creates a summary of the analysis that shows the name of the classification variable, the names of the independent variables, the number of complete cases, and the number of groups in the study (see Figure 4-2). It then displays the results, eigenvalues, relative percentages, canonical



Classification Factor: Recommend

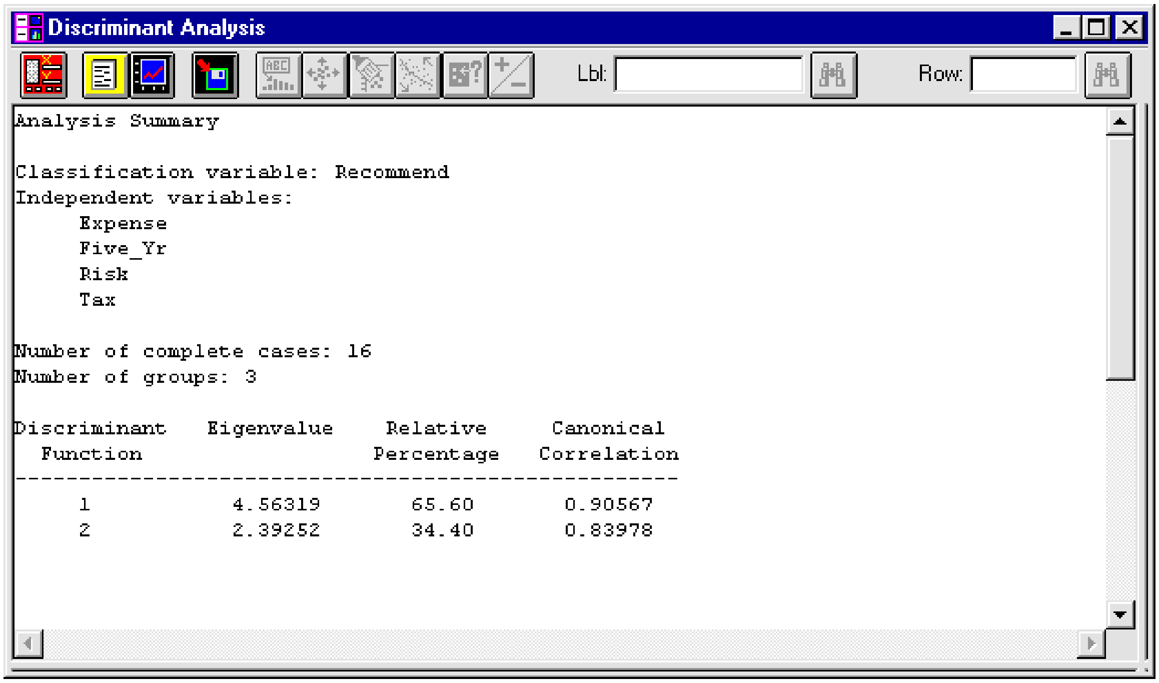
Data:
Expense
Five_Yr
Risk
Tax

(Point Labels:) Fund

(Select:)

☑ Sort

Figure 4-1. Discriminant Analysis Dialog Box



```
Analysis Summary

Classification variable: Recommend
Independent variables:
     Expense
     Five_Yr
     Risk
     Tax

Number of complete cases: 16
Number of groups: 3

Discriminant   Eigenvalue    Relative     Canonical
  Function                  Percentage   Correlation
----------------------------------------------------
     1           4.56319       65.60        0.90567
     2           2.39252       34.40        0.83978
```

Figure 4-2. Analysis Summary

correlations, Wilks' lambda, chi-square, degrees of freedom, and the *p*-values for one less than the number of groups.

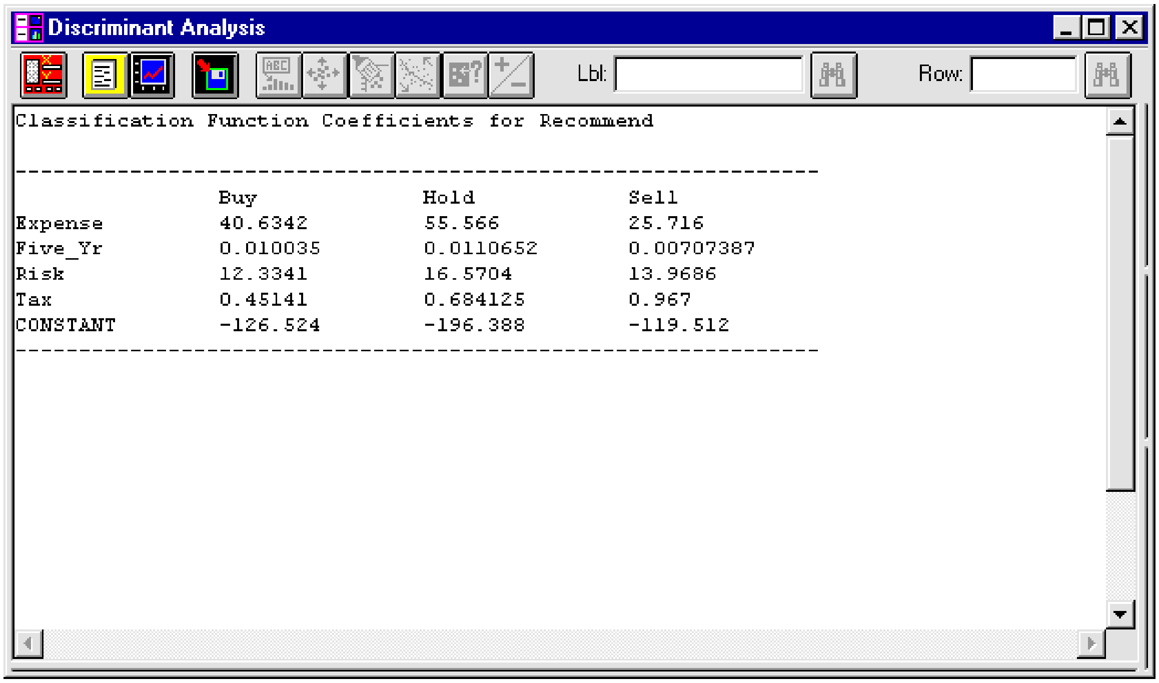
Use the *Discriminant Analysis Options* dialog box to choose the way the variables will be entered into the discriminant model, to enter values for the F-ratio at or above which the variables will be entered into the model, and to enter the maximum number of steps that will be performed before the selection process ends.

Classification Functions

The Classification Functions option creates a table that displays Fisher's linear discriminant function coefficients for each group (see Figure 4-3). The functions are used to classify the observations into groups. For example, the function for the first level of the variable is:

-126.524 + 40.6342*Expense + 0.010035*Five Yr + 12.3341*Risk + 0.45141*Tax

The function that yields the largest value for an observation represents the predicted group.



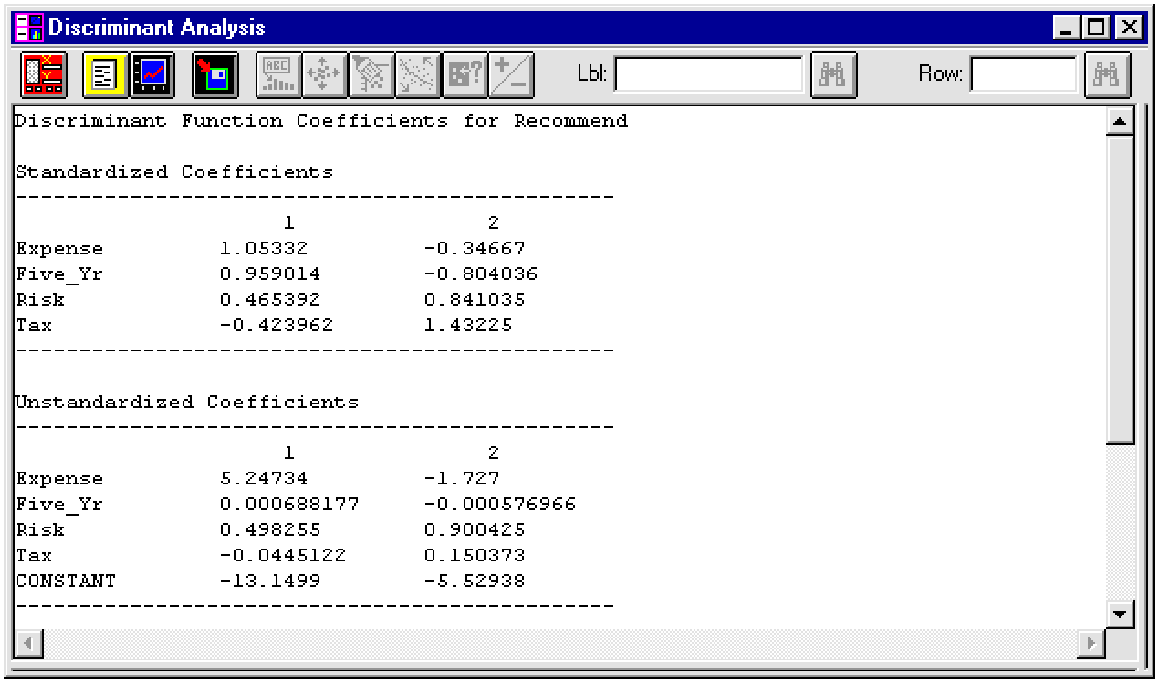
```
Classification Function Coefficients for Recommend

---------------------------------------------------------------
                Buy             Hold            Sell
Expense         40.6342         55.566          25.716
Five_Yr         0.010035        0.0110652       0.00707387
Risk            12.3341         16.5704         13.9686
Tax             0.45141         0.684125        0.967
CONSTANT        -126.524        -196.388        -119.512
---------------------------------------------------------------
```

Figure 4-3. Classification Functions

Discriminant Functions

The Discriminant Functions options creates a table of the standardized and unstandardized canonical discriminant function coefficients for each discriminant

function (see Figure 4-4). The StatAdvisor displays the equation for the first correlations, Wilks' lambda, chi-square, degrees of freedom, and the *p*-values for one less than the number of groups.



```
Discriminant Function Coefficients for Recommend

Standardized Coefficients
-----------------------------------------------
                     1                2
Expense          1.05332          -0.34667
Five_Yr          0.959014         -0.804036
Risk             0.465392         0.841035
Tax              -0.423962        1.43225
-----------------------------------------------

Unstandardized Coefficients
-----------------------------------------------
                     1                2
Expense          5.24734          -1.727
Five_Yr          0.000688177      -0.000576966
Risk             0.498255         0.900425
Tax              -0.0445122       0.150373
CONSTANT         -13.1499         -5.52938
-----------------------------------------------
```

Figure 4-4. Discriminant Functions

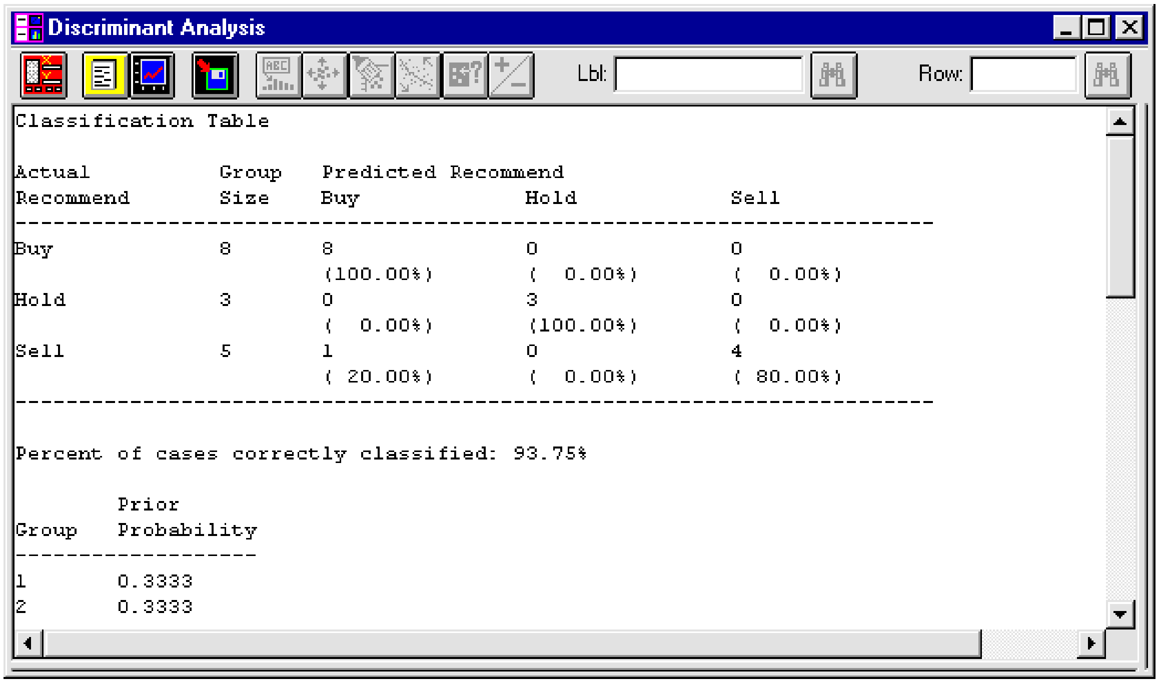
Classification Table

The Classification Table option creates a table of the actual and predicted results for the classifications (see Figure 4-5). The program tabulates the number of observations in each group that were correctly predicted as being members of that group. It also tabulates the number of observations actually belonging in other groups that were incorrectly predicted as being members of that group. The counts and percentages for each group are also displayed.

Use the *Classification Table Options* dialog box to indicate how the prior probabilities will be assigned to groups and to indicate how the observations will be displayed.

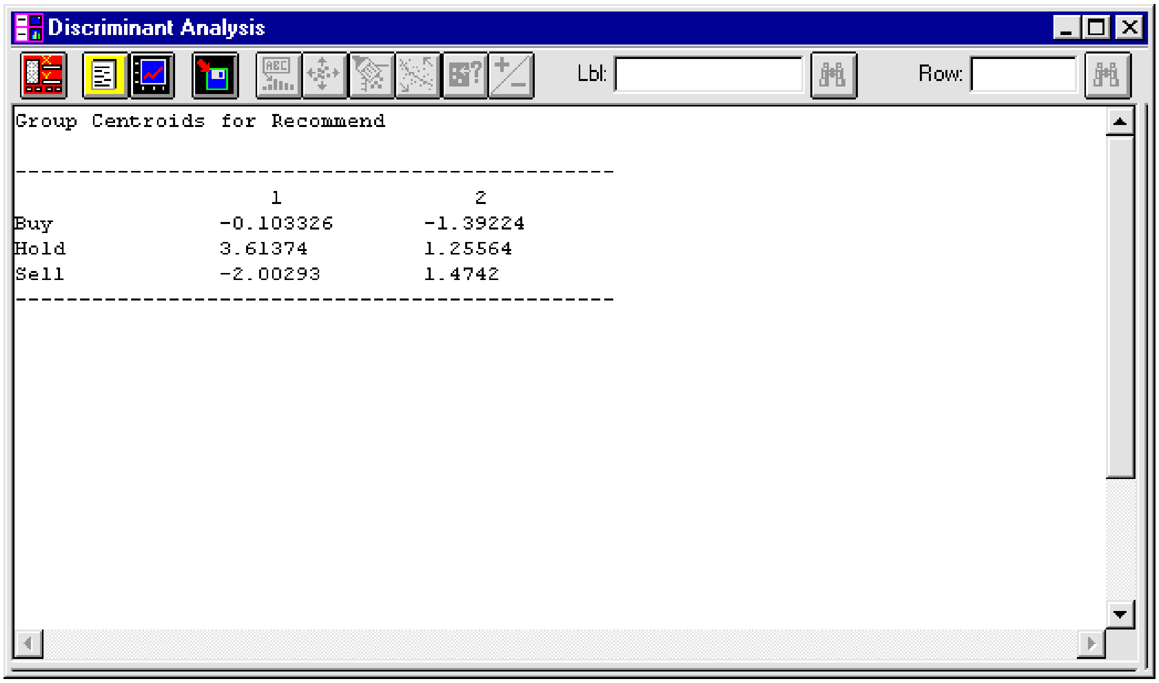
Group Centroids

The Group Centroids option creates a table of statistics of the location of the centroids (means) for the unstandardized discriminant functions (see Figure 4-6). Each row in the table represents a group. Each column contains the centroids for a single canonical discriminant function.



```
Classification Table

Actual          Group   Predicted Recommend
Recommend       Size    Buy             Hold            Sell
------------------------------------------------------------------------
Buy             8       8               0               0
                        (100.00%)       (  0.00%)       (  0.00%)
Hold            3       0               3               0
                        (  0.00%)       (100.00%)       (  0.00%)
Sell            5       1               0               4
                        ( 20.00%)       (  0.00%)       ( 80.00%)
------------------------------------------------------------------------

Percent of cases correctly classified: 93.75%

        Prior
Group   Probability
-------------------
1       0.3333
2       0.3333
```

Figure 4-5. Classification Table

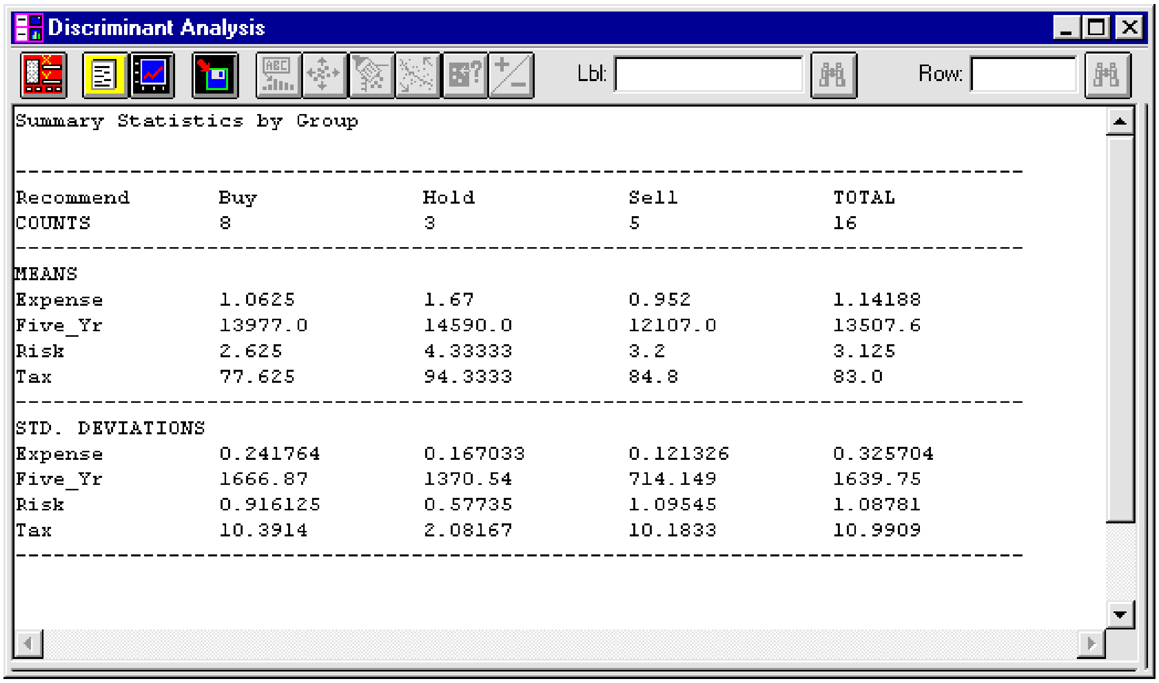


```
Group Centroids for Recommend

-----------------------------------------------
                    1               2
Buy             -0.103326       -1.39224
Hold            3.61374         1.25564
Sell            -2.00293        1.4742
-----------------------------------------------
```

Figure 4-6. Group Centroids

Group Statistics

The Group Statistics option creates a table of the counts, means, and standard deviations for each variable (see Figure 4-7). The statistics are shown by actual group and by all groups combined.



```
Summary Statistics by Group

-------------------------------------------------------------------------------
Recommend       Buy             Hold            Sell            TOTAL
COUNTS          8               3               5               16
-------------------------------------------------------------------------------
MEANS
Expense         1.0625          1.67            0.952           1.14188
Five_Yr         13977.0         14590.0         12107.0         13507.6
Risk            2.625           4.33333         3.2             3.125
Tax             77.625          94.3333         84.8            83.0
-------------------------------------------------------------------------------
STD. DEVIATIONS
Expense         0.241764        0.167033        0.121326        0.325704
Five_Yr         1666.87         1370.54         714.149         1639.75
Risk            0.916125        0.57735         1.09545         1.08781
Tax             10.3914         2.08167         10.1833         10.9909
-------------------------------------------------------------------------------
```

Figure 4-7. Group Statistics

Group Correlations

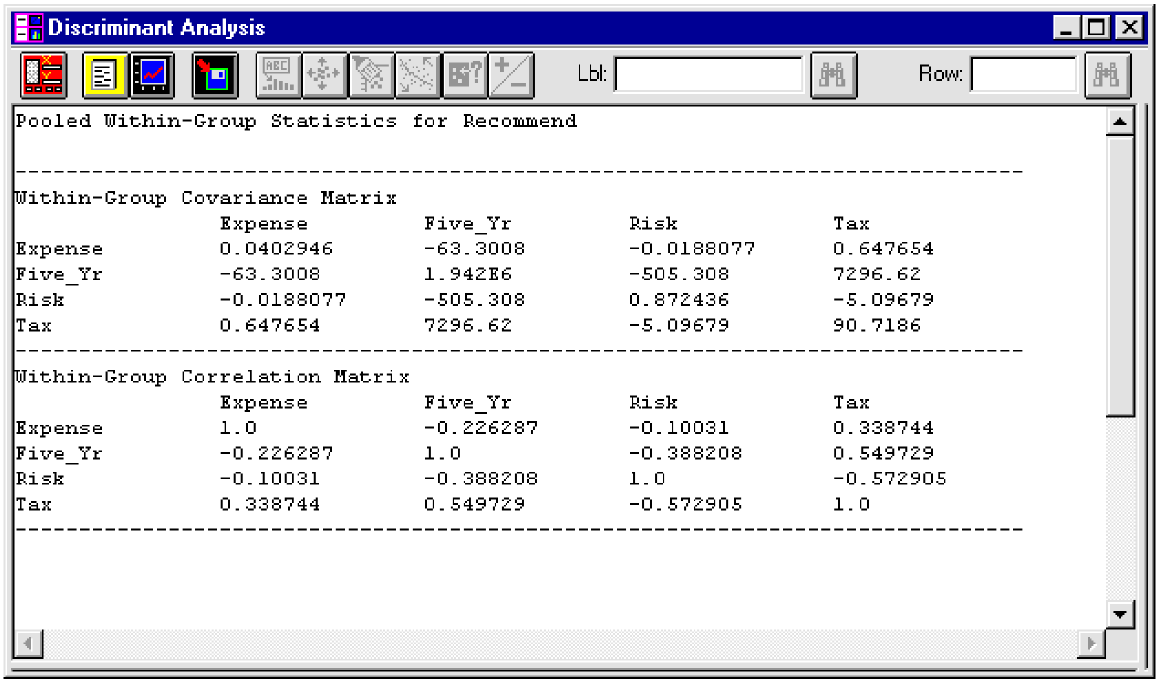
The Group Correlations option creates a table of the pooled within-group covariance and correlation matrices for all the independent variables (see Figure 4-8).

Graphical Options

2D Scatterplot

The 2D (Two-Dimensional) Scatterplot option creates a two-dimensional scatterplot of the observations by two variables (see Figure 4-9). A different point symbol is used for each group.

Use the *2D Scatterplot Options* dialog box to choose the names of the variables that will be plotted on the X- and Y-axes.



```
Pooled Within-Group Statistics for Recommend

--------------------------------------------------------------------------------
Within-Group Covariance Matrix
                Expense         Five_Yr         Risk            Tax
Expense         0.0402946       -63.3008        -0.0188077      0.647654
Five_Yr         -63.3008        1.942E6         -505.308        7296.62
Risk            -0.0188077      -505.308        0.872436        -5.09679
Tax             0.647654        7296.62         -5.09679        90.7186
--------------------------------------------------------------------------------
Within-Group Correlation Matrix
                Expense         Five_Yr         Risk            Tax
Expense         1.0             -0.226287       -0.10031        0.338744
Five_Yr         -0.226287       1.0             -0.388208       0.549729
Risk            -0.10031        -0.388208       1.0             -0.572905
Tax             0.338744        0.549729        -0.572905       1.0
--------------------------------------------------------------------------------
```

Figure 4-8. Group Correlations

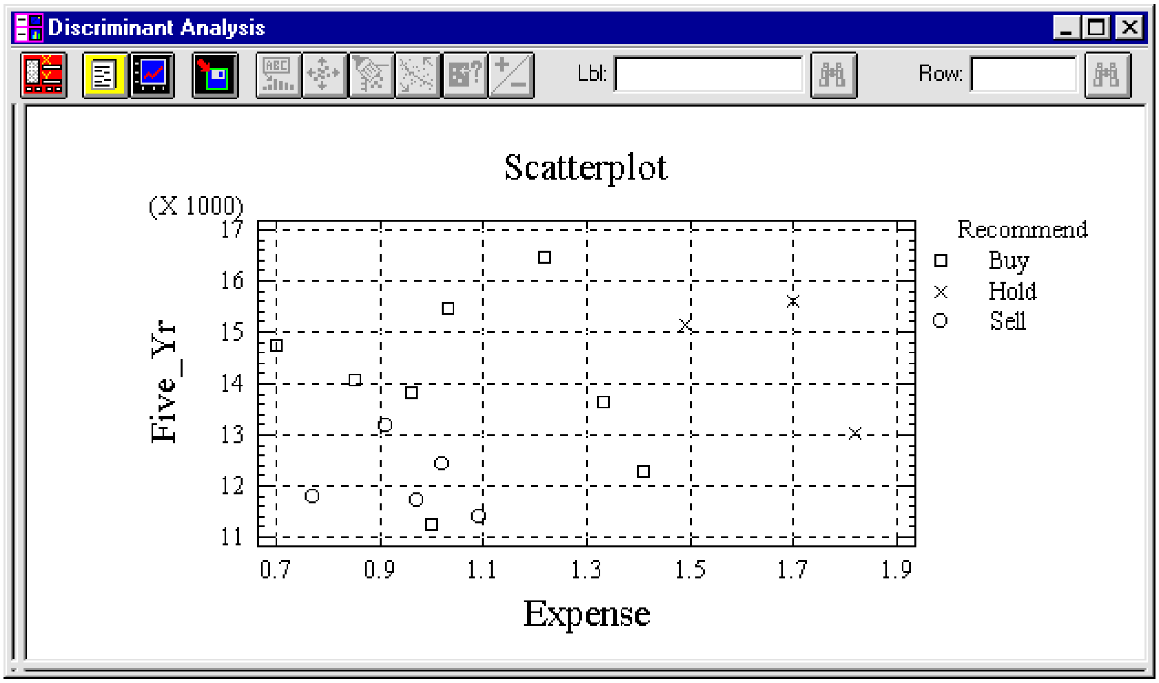


Figure 4-9. 2D Scatterplot

3D Scatterplot

The 3D (Three-Dimensional) Scatterplot option creates a three-dimensional scatterplot of the observations by three independent variables (see Figure 4-10).

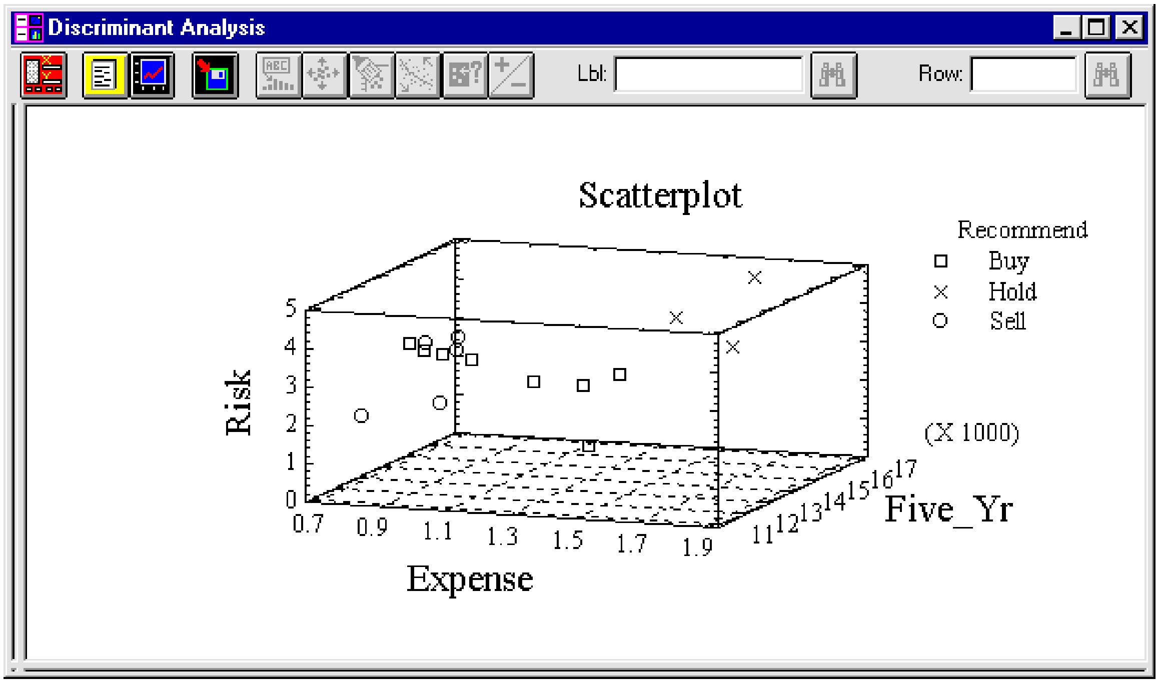


Figure 4-10. 3D Scatterplot

Use the *3D Scatterplot Options* dialog box to choose the names of the variables that will be plotted on the X-, Y-, and Z-axes.

Discriminant Functions

The Discriminant Functions option creates a plot of the values for two discriminant functions (see Figure 4-11). You can plot the points using the classification codes you chose on the Classification Factor text box on the Discriminant Analysis dialog box or you can plot the predictions for the classification group for each observation or plot using the 2D Discriminant Function Options dialog box. Different symbols identify the points in each group.

Use the *Discriminant Function Plot Options* dialog box to enter the number of the functions that will be plotted on the X- and Y-axes.

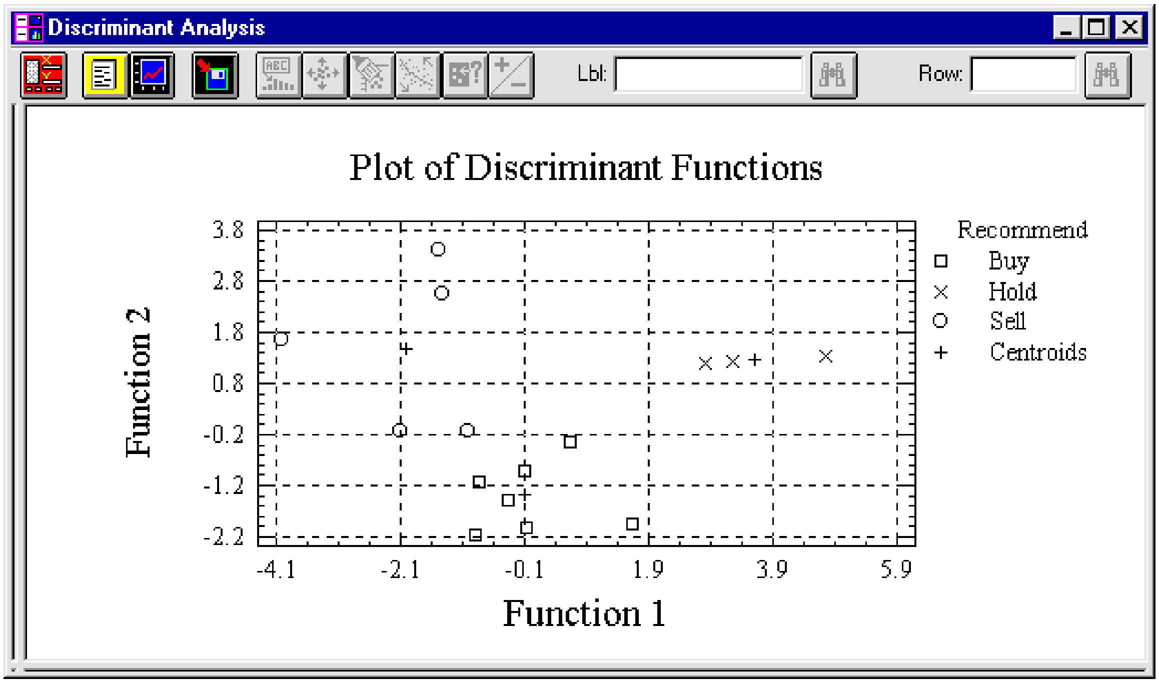


Figure 4-11. Discriminant Functions Plot

Saving the Results

The Save Results Options dialog box allows you to choose the results you want to save. There are five selections: Discriminant Function Values, Classification Function Coefficients, Standardized Coefficients, Unstandardized Coefficients, and Prior Probabilities.

You can also use the Target Variables text boxes to enter the names of the variables in which you want to save the values generated during the analysis. You can enter new names or accept the defaults.

Note: To access the Save Results Options dialog box, click the Save Results button on the Analysis toolbar (the fourth button from the left).

References

Anderson, T. W. 1958. *Introduction to Multivariate Statistical Analysis.* New York: Wiley

Bolch, B. W. and Huang, C. J. 1974. *Multivariate Statistical Methods for Business and Economics.* Englewood Cliffs, NJ: Prentice-Hall.

Hair, J., Anderson, R. and Tatham, R. 1992. *Multivariate Data Analysis*, third edition. Englewood Cliffs, NJ: Prentice-Hall.
Johnson, R. A. and Wichern, D. W. 1988. *Applied Multivariate Statistical Analysis*, second edition. Englewood Cliffs, NJ: Prentice-Hall.

Chapter 5

Using the Canonical Correlations Analysis

Background Information

Developed by Hotelling in 1936, canonical correlation analysis is a technique you can use to study the relationship between two sets of variables, each of which might contain several variables. Its purpose is to summarize or explain the relationship between two sets of variables by finding a small number of linear combinations from each set of variables that have the highest correlation possible between the sets. This is known as the first canonical correlation. The coefficients of these combinations are canonical coefficients or canonical weights. Usually the canonical
variables are normalized so each canonical variable has a variance of 1.

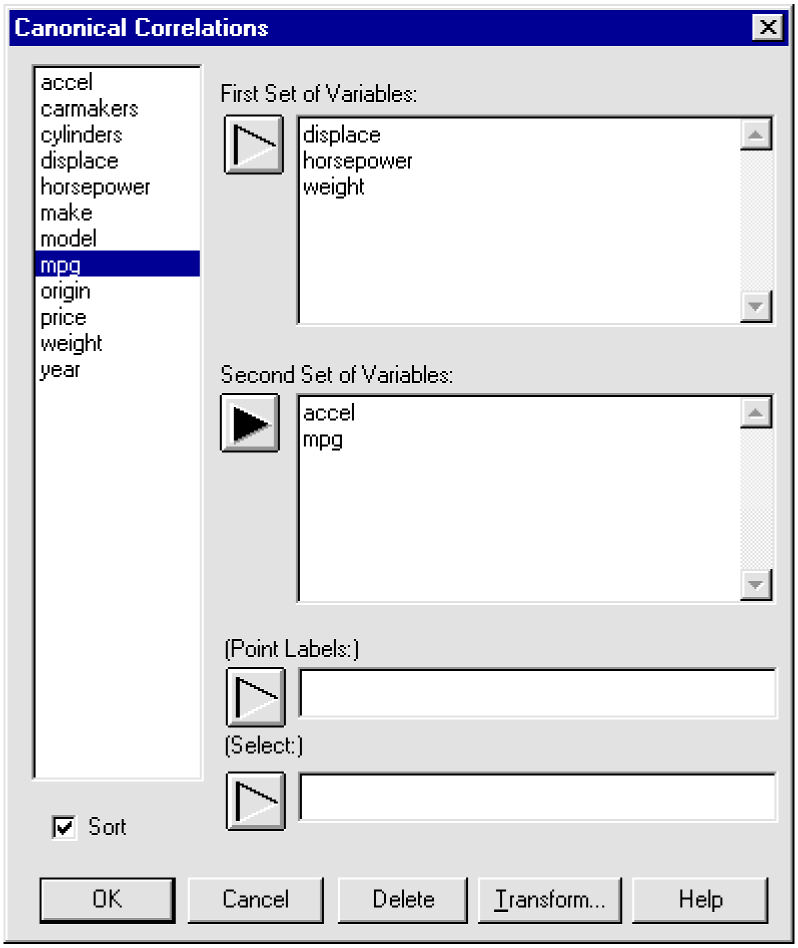
A second set of canonical variables is then found, which produces a second highest correlation coefficient. This process continues until the number of pairs equals the number of variables in the smallest group in the study.

Canonical correlation analysis is helpful in many disciplines. For example, in an educational study, the first set of variables could be measures of word attack skills (sounding-out words); the second might be measures of comprehension. In market research, the first set of variables could be measures of product characteristics (number of rolls in a package, size of each roll, price of the package, price of each roll); the second might be measures of customer reaction or buying trends.

In STATGRAPHICS *Plus*, it is assumed that the data are drawn from a population that is multivariate normal. You can test this assumption for pairs of variables by producing an X-Y Plot to verify that each pair of variables is roughly elliptical and that their points are widely scattered. Or you can produce an X-Y-Z Plot to test this assumption for sets of three variables. However, even if sets of two and three variables satisfy the multivariate normal assumptions, a full set of data may not. If that is the case, you may want to try using square root, logarithmic, or some other transformation.

Using Canonical Correlations Analysis

To access the analysis, choose SPECIAL... MULTIVARIATE METHODS... CANONICAL CORRELATIONS... from the Menu bar to display the Canonical Correlations Analysis dialog box shown in Figure 5-1.



First Set of Variables: displace, horsepower, weight

Second Set of Variables: accel, mpg

☑ Sort

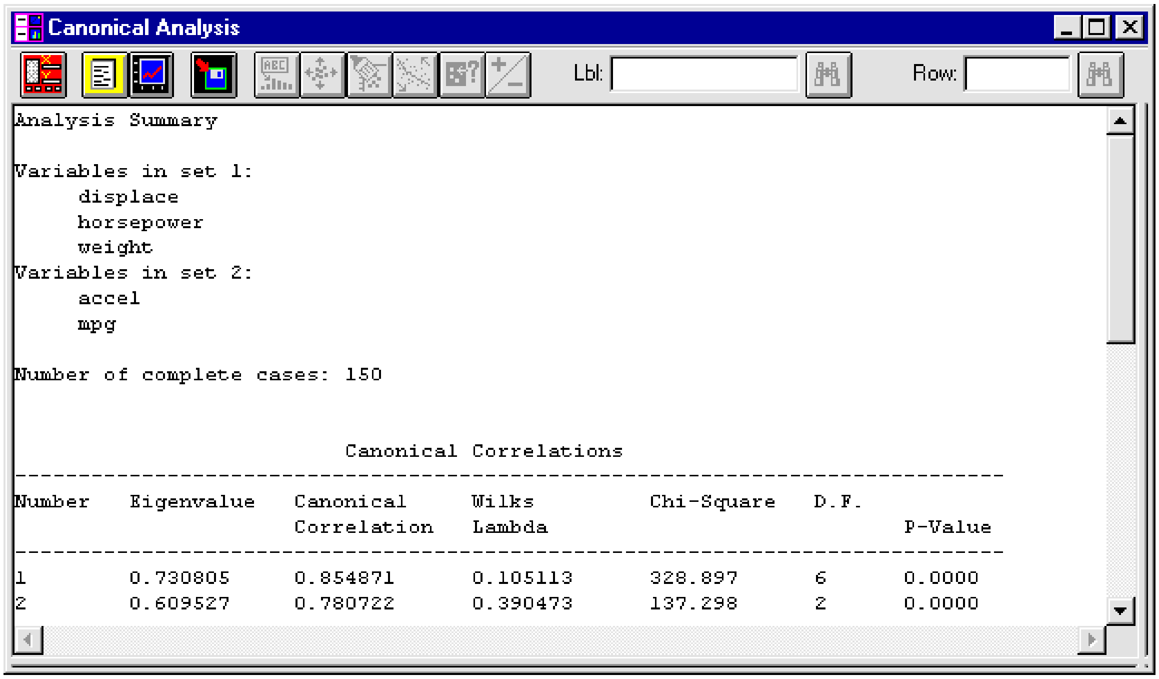
Figure 5-1. Canonical Correlations Analysis Dialog Box

Analysis Summary

The Analysis Summary option creates a summary of the analysis that includes the names of the variables in both the first and second sets, and the number of complete cases in the analysis (see Figure 5-2). The summary then displays the results of the analysis — information about the canonical correlations and the coefficients for the canonical variables in both the first and second sets. Canonical correlations with small *p*-values (less than .05) are significant. Large correlations are associated with large eigenvalues and chi-square values.

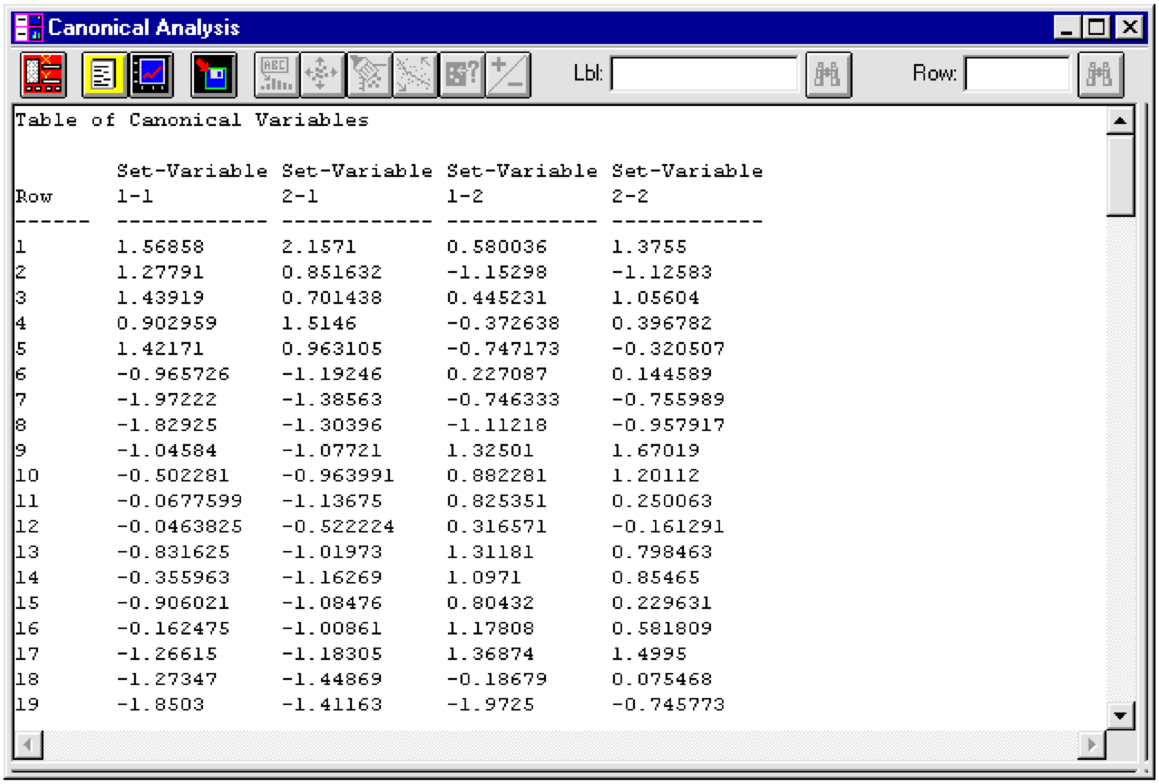
Data Table

The Data Table option creates a table of the values for the canonical variables in both the first and second sets of variables (see Figure 5-3).



```
Analysis Summary

Variables in set 1:
     displace
     horsepower
     weight
Variables in set 2:
     accel
     mpg

Number of complete cases: 150


                          Canonical Correlations
------------------------------------------------------------------------------
Number    Eigenvalue   Canonical   Wilks        Chi-Square   D.F.
                       Correlation Lambda                           P-Value
------------------------------------------------------------------------------
1         0.730805     0.854871    0.105113     328.897      6      0.0000
2         0.609527     0.780722    0.390473     137.298      2      0.0000
```

Figure 5-2. Analysis Summary



```
Table of Canonical Variables

        Set-Variable Set-Variable Set-Variable Set-Variable
Row     1-1          2-1          1-2          2-2
------  ------------ ------------ ------------ ------------
1       1.56858      2.1571       0.580036     1.3755
2       1.27791      0.851632     -1.15298     -1.12583
3       1.43919      0.701438     0.445231     1.05604
4       0.902959     1.5146       -0.372638    0.396782
5       1.42171      0.963105     -0.747173    -0.320507
6       -0.965726    -1.19246     0.227087     0.144589
7       -1.97222     -1.38563     -0.746333    -0.755989
8       -1.82925     -1.30396     -1.11218     -0.957917
9       -1.04584     -1.07721     1.32501      1.67019
10      -0.502281    -0.963991    0.882281     1.20112
11      -0.0677599   -1.13675     0.825351     0.250063
12      -0.0463825   -0.522224    0.316571     -0.161291
13      -0.831625    -1.01973     1.31181      0.798463
14      -0.355963    -1.16269     1.0971       0.85465
15      -0.906021    -1.08476     0.80432      0.229631
16      -0.162475    -1.00861     1.17808      0.581809
17      -1.26615     -1.18305     1.36874      1.4995
18      -1.27347     -1.44869     -0.18679     0.075468
19      -1.8503      -1.41163     -1.9725      -0.745773
```

Figure 5-3. Data Table

Graphical Options

Canonical Variables Plot

The Canonical Variables Plot option creates a plot of the values for the canonical variables (see Figure 5-4). One point for each row of the data file appears on the plot, which is helpful in identifying the strength of the canonical correlations. Strong correlations appear linear.

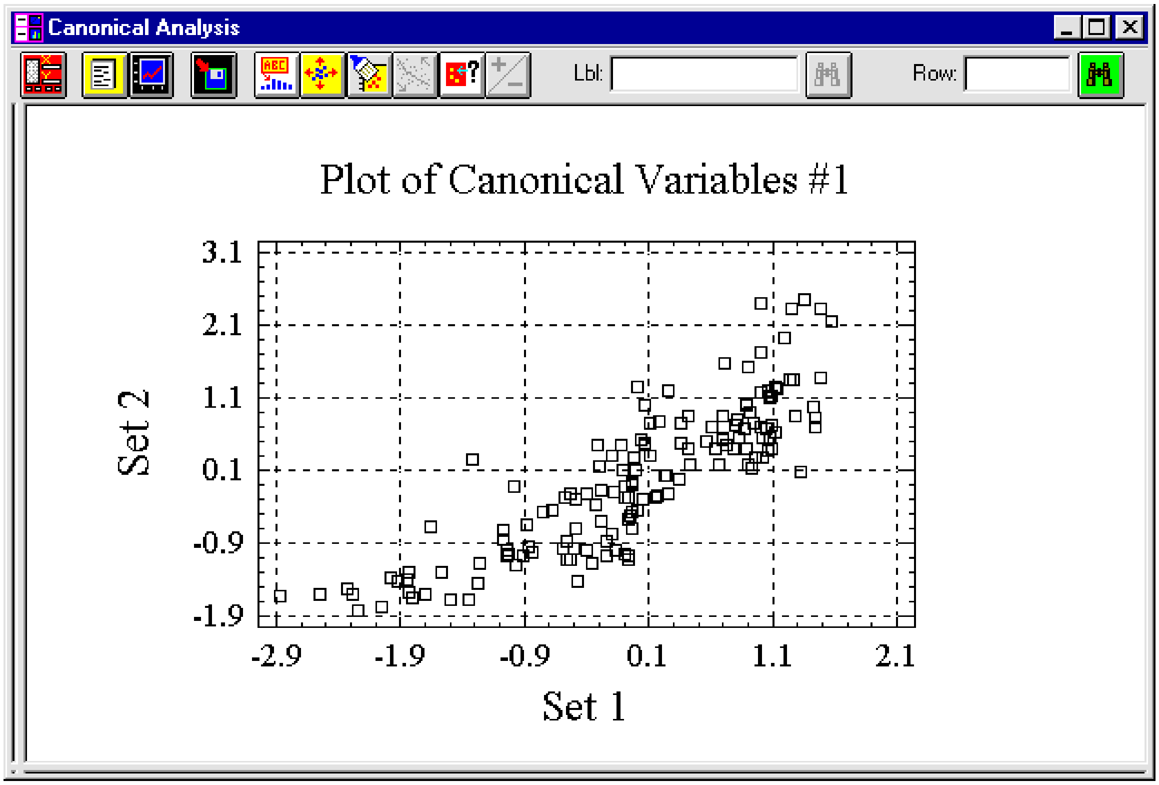


Figure 5-4. Canonical Variables Plot

Use the *Canonical Variables Plot Options* dialog box to enter the number of the variable that will appear on the plot.

Saving the Results

The Save Results Options dialog box allows you to choose the results you want to save. There are four selections: Coefficients - First Set, Coefficients - Second Set, Canonical Correlations - First Set, Canonical Correlations - Second Set.

You can also use the Target Variables text boxes to enter the names of the variables in which you want to save the values generated during the analysis. You can enter new names or accept the defaults.

Note: To access the Save Results Options dialog box, click the Save Results button on the Analysis toolbar (the fourth button from the left).

References

Anderson, T. W. 1984. *An Introduction to Multivariate Statistical Analysis*, second edition. New York: Wiley.

Hair, J., Anderson, R., and Tatham, R. 1992. *Multivariate Data Analysis*, third edition. Englewood Cliffs, NJ: Prentice-Hall.

Hotelling, H. 1936. “Relations between Two Sets of Variables,” *Biometrika*, **28**:321-77.

Johnson, R. A. and Wichern, D. W. 1988. *Applied Multivariate Statistical Analysis,* second edition. Englewood Cliffs, NJ: Prentice-Hall.

Morrison, D. F. 1990. *Multivariate Statistical Methods*, third edition. New York: McGraw-Hill.

Chapter 6

Creating, Saving, and Using Matrix Data

STATGRAPHICS *Plus* allows you to create and save correlation and covariance matrices. You can use matrix data in the Multivariate Methods product by entering the data into any of these analyses: Principal Components, Factor Analysis, and Cluster Analysis.

This chapter provides instructions for creating and saving matrix data. The *Tutorial Manual* for Multivariate Methods contains a tutorial that shows how to create a covariance matrix then use it to perform a Factor analysis.

Creating and Saving Matrix Data

To Open the Data File

Before you create the matrix, open STATGRAPHICS *Plus*, then open a data file. For this example, use the ***Obesity.sf*** data file.

1. Open STATGRAPHICS, then choose FILE... OPEN... OPEN DATA FILE... from the File menu to display the Open Data File dialog box.

2. Enter the **Obesity.sf** file into the File Name text box, then click Open to open the file.

To Enter Variables for the Matrix

1. To access the analysis, choose DESCRIBE... NUMERIC DATA... MULTIPLE-VARIABLE ANALYSIS... from the Menu bar to display the Multiple-Variable Analysis dialog box shown in Figure 6-1.

2. Choose and enter the following variable names into the Data text box: ***x2 coeff, x3 pgmntcr, x4 phospht, x5 calcium,*** and ***x6 phosphs***.

3. Type ***first(50)*** into the Select text box to use only the first 50 variables in the data set.

4. Click OK to create the matrix and display the Analysis Summary and the Scatterplot Matrix in the Analysis window.

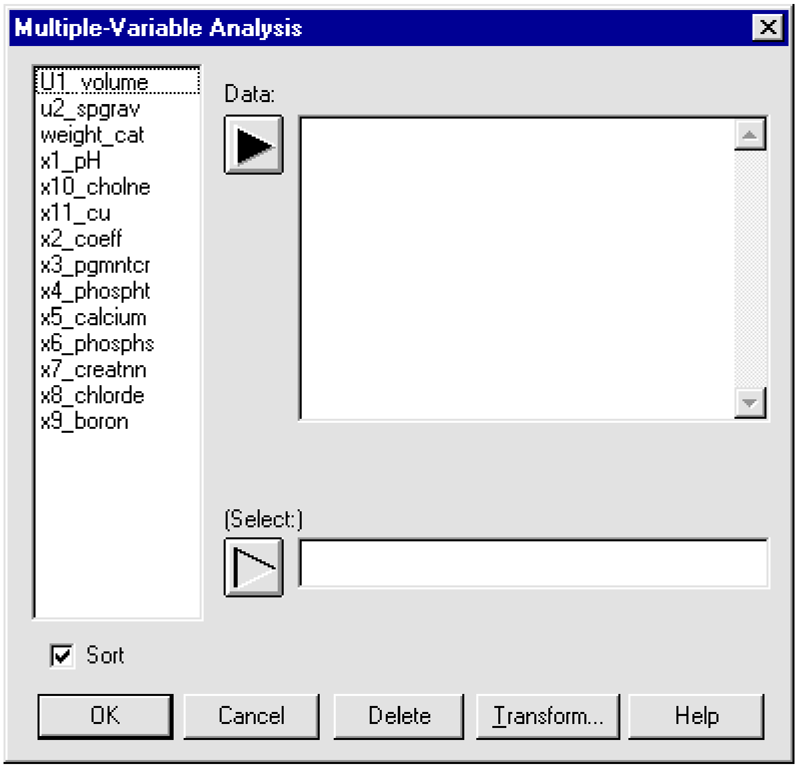


Figure 6-1. Multiple-Variable Analysis Dialog Box

To Save the Matrix

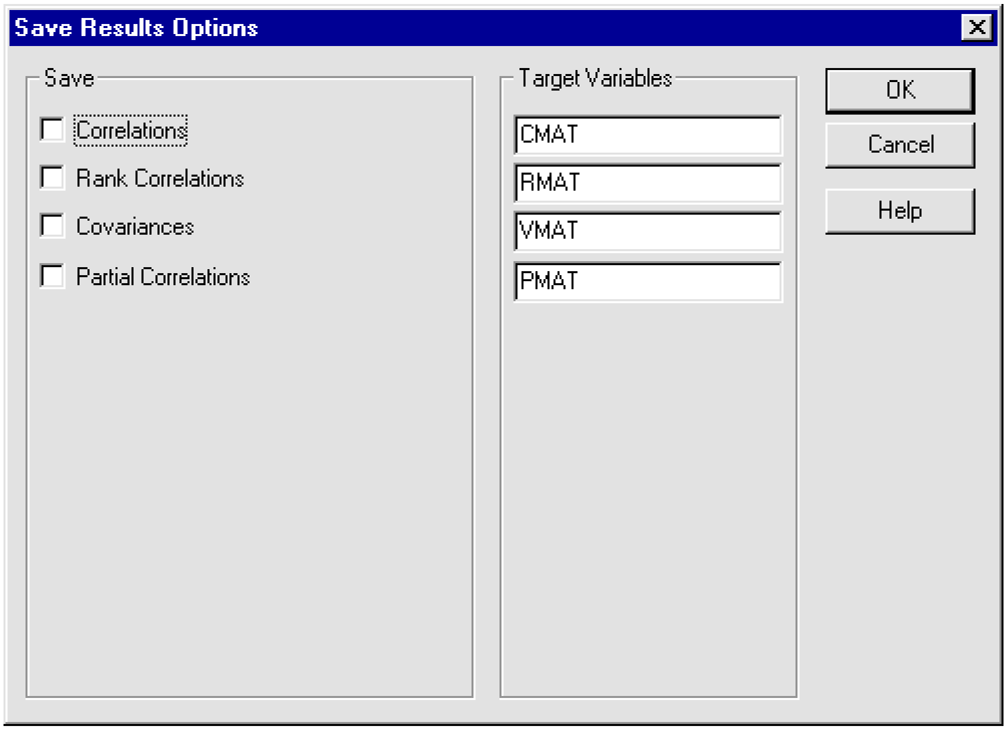
1. Click the Save Results button on the Analysis toolbar to display the Save Results Options dialog box with save options shown in Figure 6-2.

2. Click the Correlations check box. Notice the names of the variables in the Target Variables text boxes in Figure 6-2. The program uses the following convention to save the matrices:

CMAT – correlation matrix
RMAT – rank correlation matrix
VMAT – covariance matrix
PMAT – partial correlation matrix.

3. Click OK to save the correlation matrix as a set of variables and redisplay the Analysis Summary and Scatterplot Matrix.

As you can see, using a correlation or covariance matrix as data in the Multivariate Methods product is simple — just complete the Analysis dialog box, being sure to enter the variables in the order they were placed in the



Save Results Options

Save

□ Correlations

□ Rank Correlations

□ Covariances

□ Partial Correlations

Target Variables

CMAT

RMAT

VMAT

PMAT

Figure 6-2. Save Results Options Dialog Box

DataSheet. For example, if the correlation matrix has five columns, enter: ***CMAT 1***, ***CMAT 2***, ***CMAT 3***, ***CMAT 4***, and ***CMAT 5***. If the covariance matrix has five columns, enter: ***VMAT 1***, ***VMAT 2***, ***VMAT 3***, ***VMAT 4***, and ***VMAT 5***.

After you save the matrix, the variable names appear in the list box on the Multiple-Variable Analysis dialog box.

4. Close the data file.