Chapter 13
Multivariate Techniques

Chapter Table of Contents

Chapter 13
Multivariate Techniques

Introduction

Multivariate analysis techniques, such as principal components analysis and canonical correlation, enable you to investigate relationships in your data. Unlike statistical modeling, you do this without designating dependent or independent variables. In principal component analysis, you examine relationships within a single set of variables. In canonical correlation analysis, you examine the relationship between two sets of variables.

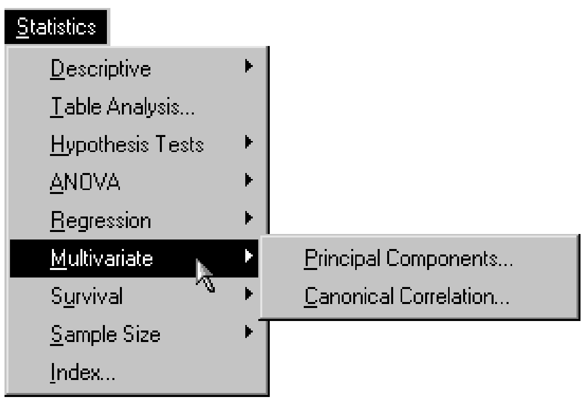


Figure 13.1. Multivariate Menu

The Analyst Application enables you to perform principal components analysis and canonical correlation. The Principal Components task enables you to compute principal components from a single set of variables. The Canonical Correlation task enables you to examine the relationship between two sets of variables.

The examples in this chapter demonstrate how you can use the Analyst Application to perform principal components and canonical correlation analyses.

Principal Components Analysis

The purpose of principal component analysis is to derive a small number of independent linear combinations (principal components) of a set of variables that retain as much of the information in the original variables as possible.

For example, suppose you are interested in examining the relationship among measures of food consumption from different sources. The sample data set Protein records the amount of protein consumed from nine food groups for each of 25 European countries. The nine food groups are red meat (RedMt), white meat (WhiteMt), eggs (Eggs), milk (Milk), fish (Fish), cereal (Cereal), starch (Starch), nuts (Nuts), and fruits and vegetables (FruVeg).

Open the Protein Data Set

The data are provided in the Analyst Sample Library. To access this Analyst sample data set, follow these steps:

1. Select **Tools** → **Sample Data** . . .
2. Select Protein.
3. Click **OK** to create the sample data set in your Sasuser directory.
4. Select **File** → **Open By SAS Name** . . .
5. Select Sasuser from the list of **Libraries**.
6. Select Protein from the list of members.
7. Click **OK** to bring the Protein data set into the data table.

Request the Principal Components Analysis

To perform a principal components analysis, follow these steps:

1. Select **Statistics** → **Multivariate** → **Principal Components** . . .
2. Highlight all of the quantitative variables (RedMt, WhiteMt, Eggs, Milk, Fish, Cereal, Starch, Nuts, and FruVeg).
3. Click on the **Variables** button.

The goal of this analysis is to determine the principal components of all protein sources. Therefore, all of the protein source variables are included in the **Variables** list, as displayed in Figure 13.2. The character variable Country is an identifier variable and is omitted from the **Variables** list.

Note that you can analyze a partial correlation or covariance matrix by specifying the variables to be partialed out in the **Partial** list. The full correlation matrix is used for this analysis.

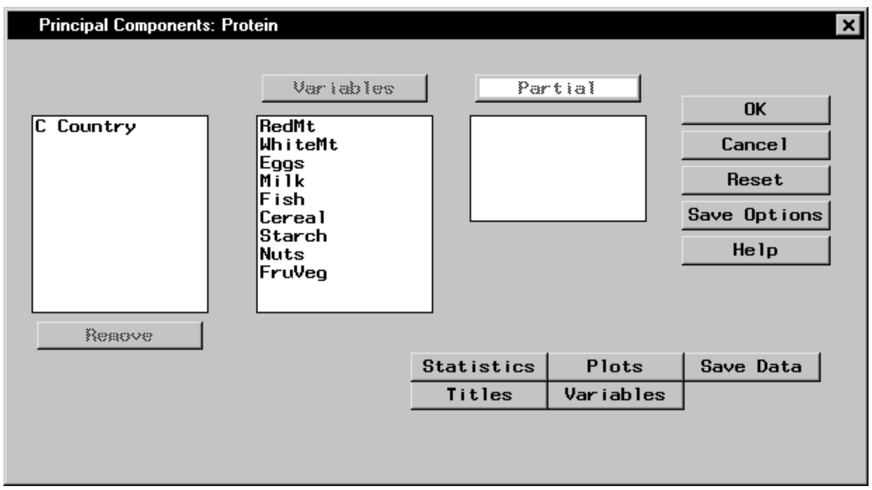


Figure 13.2. Principal Components Dialog

The default principal components analysis includes simple statistics, the correlation matrix for the analysis variables, and the associated eigenvalues and eigenvectors.

Request Principal Component Plots

You can use the Plots dialog to request a scree plot or component plots. A scree plot is useful in determining the appropriate number of components to interpret. It displays the eigenvalues on the vertical axis and the principal component number on the horizontal axis.

To request a scree plot, follow these steps:

1. Click on the **Plots** button in the main dialog.
2. Select **Create scree plot**.

Figure 13.3 displays the **Scree Plot** tab, in which a scree plot of the positive eigenvalues is requested.

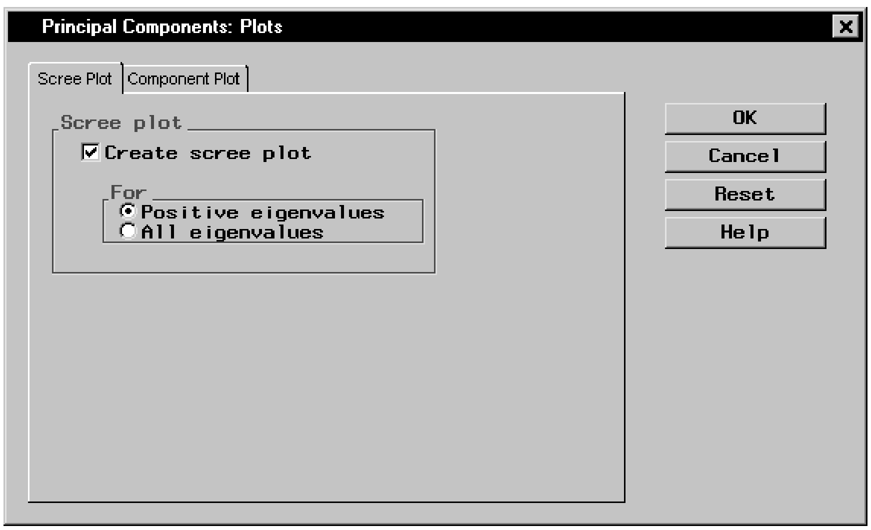


Figure 13.3. Principal Components: Plots Dialog, Scree Plot Tab

A component plot displays the component score of each observation for a pair of components. When you specify an Id variable, the values of that variable are also displayed in the plot.

To request a component plot in addition to the scree plot, follow these steps.

1. Click on the **Component Plot** tab in the Plots dialog.
2. Select **Create component plots**.
3. Click on the down arrow in the box labeled **Type:**
4. Select **Enhanced**. An enhanced component plot displays the variable names and values of the Id variable in the plot.
5. Select the variable Country in the **Id variable** list.
6. Click on the **Id** button to select the variable Country as an Id variable.

You can also enter the **Dimensions** for which you want plots. For example, to request plots of the first versus second, first versus third, and second versus third principal components, you type the values 1 and 3.

7. Click **OK**.

Figure 13.4 displays the **Component Plot** tab, which requests an enhanced component plot.

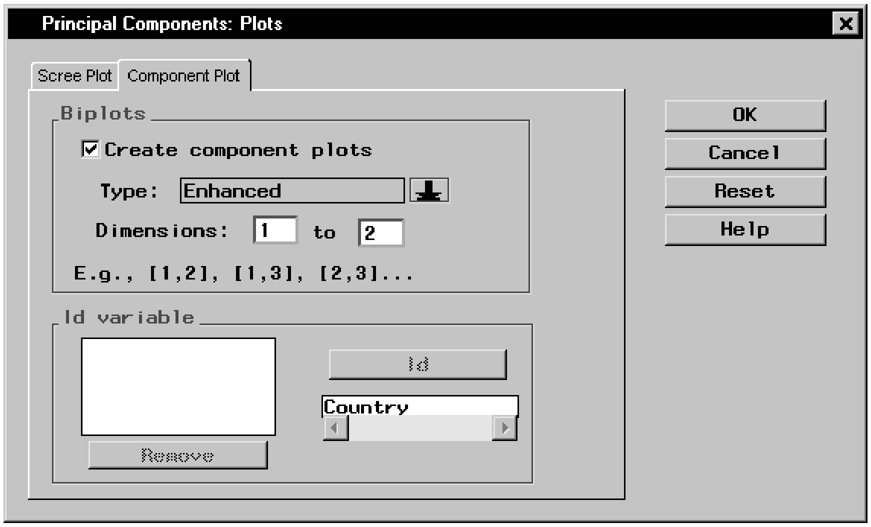


Figure 13.4. Principal Components: Plots Dialog, Component Plot Tab

Click **OK** in the Principal Components dialog to perform the analysis.

Review the Results

Figure 13.5 displays simple statistics and correlations among the variables.

Analysis

The PRINCOMP Procedure

Observations	25
Variables	9

Simple Statistics

	RedMt	WhiteMt	Eggs	Milk	Fish
Mean	9.828000000	7.896000000	2.936000000	16.31200000	4.284000000
StD	3.347078328	3.694080851	1.117616511	7.33759270	3.402533370

Simple Statistics

	Cereal	Starch	Nuts	FruVeg
Mean	32.24800000	4.276000000	3.072000000	4.136000000
StD	10.97478625	1.634084861	1.985682083	1.803903176

Correlation Matrix

		RedMt	White Mt	Eggs
RedMt	Red Meat	1.0000	0.1530	0.5856
WhiteMt	White Meat	0.1530	1.0000	0.6204
Eggs	Eggs	0.5856	0.6204	1.0000
Milk	Milk	0.4846	0.2755	0.5000
Fish	Fish	0.0610	-.2340	0.0656
Cereal	Cereal	-.4999	-.4138	-.7124
Starch	Starch	0.1354	0.3138	0.4522
Nuts	Nuts	-.3494	-.6350	-.5598
FruVeg	Fruits and Vegetables	-.0742	-.0613	-.0455

Figure 13.5. Principal Components: Simple Statistics and Correlations

Figure 13.6 displays the eigenvalues and eigenvectors of the correlation matrix for the nine variables. The eigenvalues indicate that four components provide a reasonable summary of the data, accounting for about 84% of the total variance. Subsequent components each contribute 5% or less.

Analysis

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	3.88155846	2.24699126	0.4313	0.4313
2	1.63456720	0.57436630	0.1816	0.6129
3	1.06020090	0.10480537	0.1178	0.7307

The PRINCOMP Procedure

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
4	0.95539554	0.42415968	0.1062	0.8369
5	0.53123586	0.10477119	0.0590	0.8959
6	0.42646467	0.13882496	0.0474	0.9433
7	0.28763971	0.16930381	0.0320	0.9752
8	0.11833590	0.01373414	0.0131	0.9884
9	0.10460176		0.0116	1.0000

Eigenvectors

		Prin1	Prin2	Prin3	Prin4
RedMt	Red Meat	0.309748	-.065972	-.515118	0.463822
WhiteMt	White Meat	0.324237	-.260235	0.606207	0.141962
Eggs	Eggs	0.435600	-.049211	0.078644	0.326051
Milk	Milk	0.337594	-.191821	-.367087	0.013510
Fish	Fish	0.130009	0.667080	-.212505	-.283382
Cereal	Cereal	-.438556	-.240522	0.080827	0.049592
Starch	Starch	0.314340	0.334045	0.281898	-.241259
Nuts	Nuts	-.422473	0.140097	-.126266	0.337759
FruVeg	Fruits and Vegetables	-.093176	0.504628	0.281296	0.634959

Figure 13.6. Principal Components: Eigenvectors and Eigenvalues

The table of eigenvectors in Figure 13.6 reveals that the first eigenvector has equally large loadings on all of the animal-protein variables. This suggests that the first component is primarily a measure of animal-protein consumption. This eigenvector also has a large loading on the variable Starch and negative loadings on the variables Cereal and Nuts.

The second eigenvector has high positive loadings on the variables Fish, Starch, and FruVeg. This component seems to account for diets in coastal regions or warmer climates. The remaining components are not as easily identified.

The scree plot displayed in Figure 13.7 shows a gradual decrease in eigenvalues. However, the contributions are relatively low after the fourth component, which agrees with the preceding conclusion that four principal components provide a reasonable summary of the data.

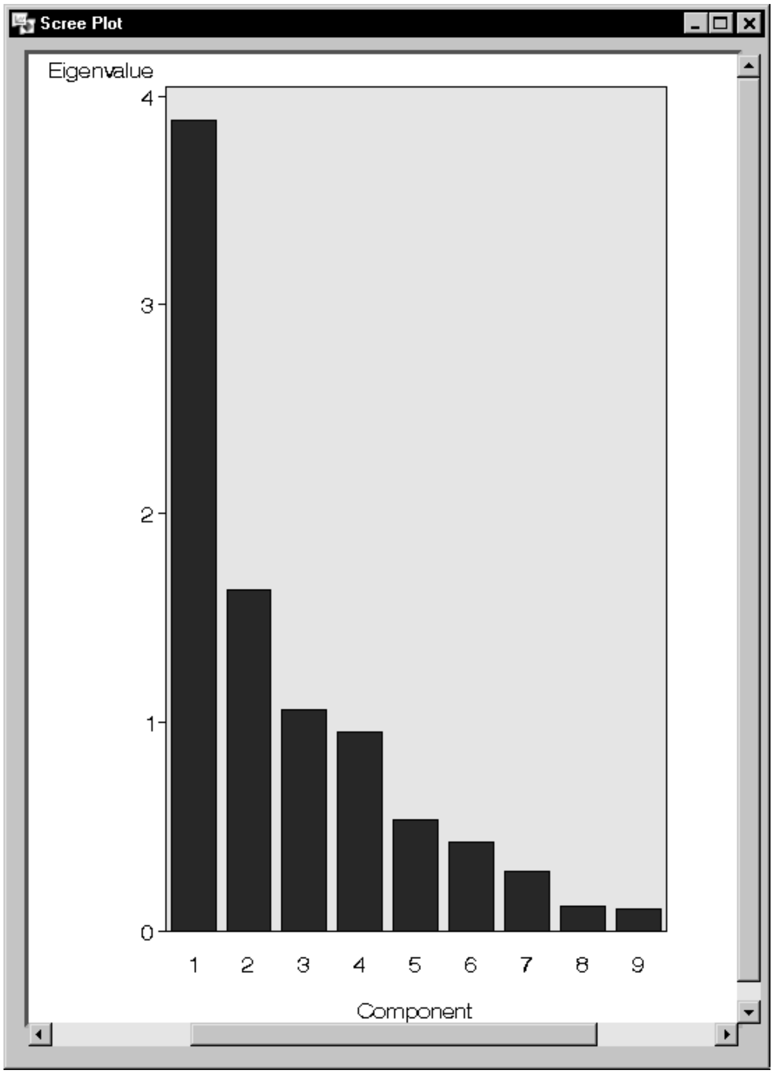


Figure 13.7. Principal Components: Scree Plot

The following enhanced component plot (Figure 13.8) displays the relationship between the first two components; each observation is identified by country.

In addition, the plot is enhanced to depict the correlations between the variables and the components. This correlation is often called the *component loading*. The amount by which each variable "loads" on a component is measured by its correlation with the component.

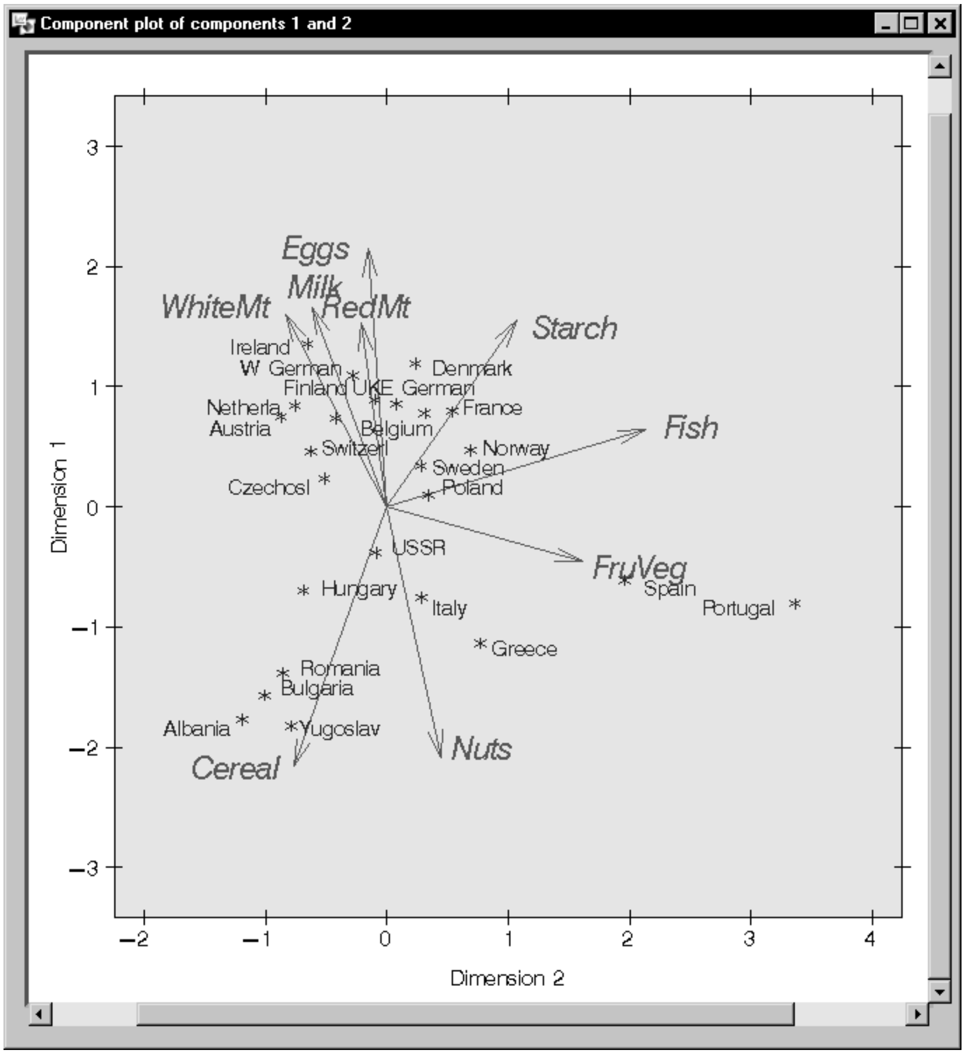


Figure 13.8. Principal Components: Scores and Component Loading Plot

In Figure 13.8, each vector corresponds to one of the analysis variables and is proportional to its component loading. For example, the variables Eggs, Milk, and RedMt all load heavily on the first component. The variables Fish and FruVeg load heavily on the second component but load very little on the first component.

The information provided by the variable Country reveals that western European countries tend to consume protein from more expensive sources (that is, meat, eggs, and milk), while countries near the Mediterranean Sea rely more heavily on fruits, vegetables, nuts, and fish for their protein sources. Eastern European countries rely more on cereal crops and nuts to supply their protein.

Canonical Correlation

Canonical correlation analysis is a variation on the concept of multiple regression and correlation analysis. In multiple regression and correlation analysis, you examine the relationship between a single Y variable and a linear combination of a set of X variables. In canonical correlation analysis, you examine the relationship between a linear combination of the set of Y variables and a linear combination of the set of X variables.

For example, suppose that you want to determine the degree of correspondence between a set of job characteristics and measures of employee satisfaction. The sample data set Jobs contains the task characteristics and satisfaction profiles for 14 jobs. The three variables associated with job satisfaction are career track satisfaction (Career), management and supervisor satisfaction (Supervis), and financial satisfaction (Finance). The three variables associated with job characteristics are task variety (Variety), supervisor feedback (Feedback), and autonomy (Autonomy).

In this task, the canonical correlation analysis is performed, labels are specified to identify each set of canonical variables, and a plot of the canonical variables is requested.

Open the Jobs Data Set

The data are provided in the Analyst Sample Library. To access this Analyst sample data set, follow these steps:

1. Select **Tools** → **Sample Data** . . .
2. Select Jobs.
3. Click **OK** to create the sample data set in your Sasuser directory.
4. Select **File** → **Open By SAS Name** . . .
5. Select Sasuser from the list of **Libraries**.
6. Select Jobs from the list of members.
7. Click **OK** to bring the Jobs data set into the data table.

Request the Canonical Correlation Analysis

To perform a canonical correlation analysis, follow these steps:

1. Select **Statistics**→**Multivariate** → **Canonical Correlation**. . .
2. Select the job satisfaction variables (Career, Supervis, and Finance) as the variables in **Set 1**.
3. Select the job characteristic variables (Variety, Feedback, and Autonomy) as the variables in **Set 2**.

Figure 13.9 displays the Canonical Correlation dialog, with each of the two sets of variables defined.

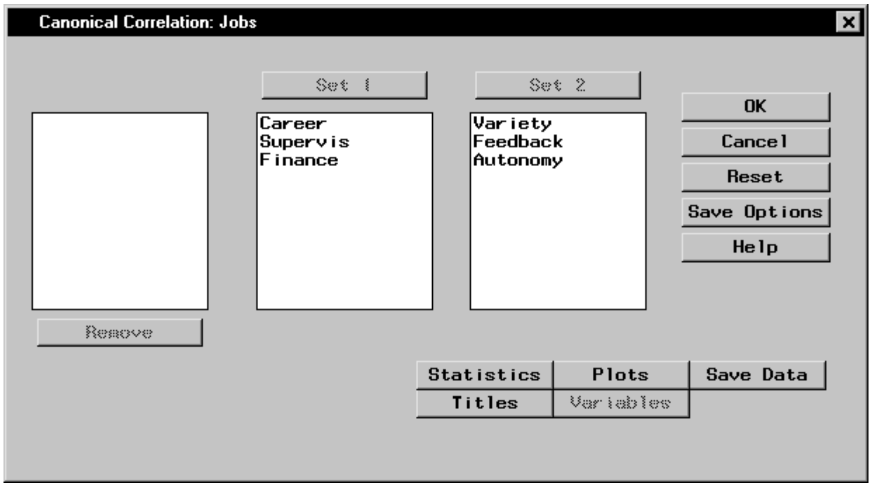


Figure 13.9. Canonical Correlation Dialog

The default analysis includes the canonical correlations, eigenvalues, likelihood ratios, and tests of significance.

Specify Identifying Labels

You can optionally specify labels and prefixes to identify the two groups of calculated canonical variables. To specify labels and prefixes, follow these steps:

1. Click on the **Statistics** button in the main dialog.
2. Enter a label for each of the two sets of canonical variables.
3. Enter a prefix for each set of canonical variables. The prefix is used to assign names to the canonical variables.
4. Click **OK**.

Figure 13.10 displays the **Canonical Analysis** tab with labels and prefixes specified.

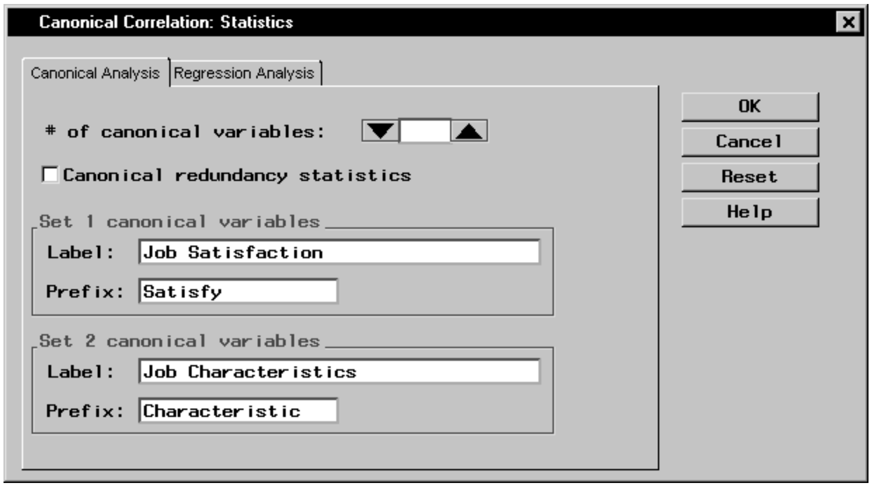


Figure 13.10. Canonical Correlation: Statistics Dialog, Canonical Analysis Tab

Request Canonical Variate Plots

To request plots of the canonical variables, follow these steps:

1. Click on the **Plots** button in the main dialog.
2. Select **Create canonical variable plots**.

You can also enter the **Canonical variables** for which you want plots. For example, to request plots of the first, second, and third canonical variable pairs, you would type the values 1 and 3.

3. Click **OK**.

Figure 13.11 displays the Plots dialog, in which plots of the first two canonical variables are requested.

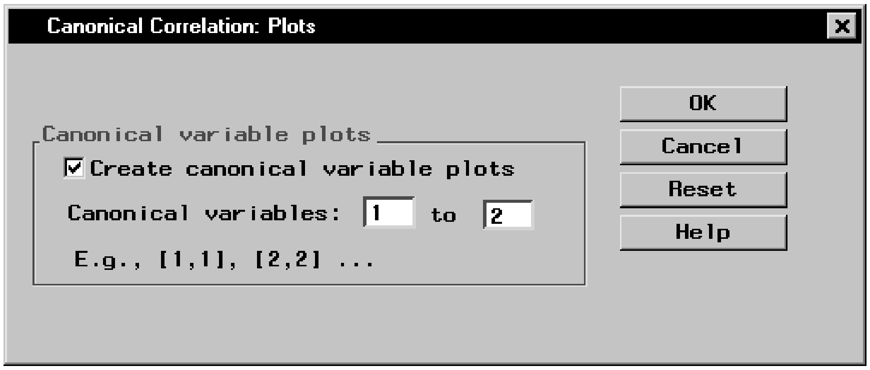
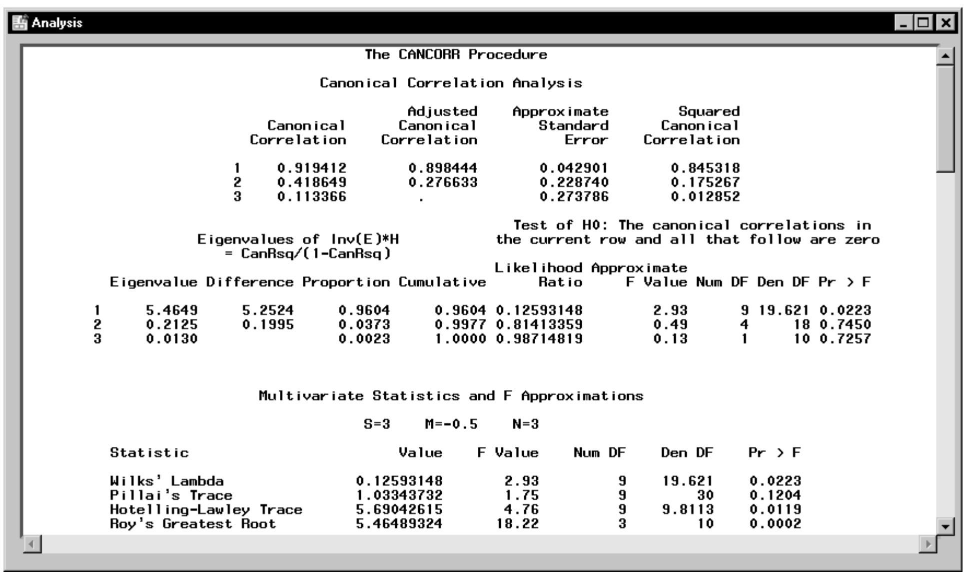


Figure 13.11. Canonical Correlation: Plots Dialog

Click **OK** in the Canonical Correlation dialog to perform the analysis.

Review the Results

Figure 13.12 displays the canonical correlation, adjusted canonical correlation, approximate standard error, and squared canonical correlation for each pair of canonical variables.



The CANCORR Procedure

Canonical Correlation Analysis

	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation
1	0.919412	0.898444	0.042901	0.845318
2	0.418649	0.276633	0.228740	0.175267
3	0.113366	.	0.273786	0.012852

Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)

Test of H0: The canonical correlations in the current row and all that follow are zero

	Eigenvalue	Difference	Proportion	Cumulative	Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
1	5.4649	5.2524	0.9604	0.9604	0.12593148	2.93	9	19.621	0.0223
2	0.2125	0.1995	0.0373	0.9977	0.81413359	0.49	4	18	0.7450
3	0.0130		0.0023	1.0000	0.98714819	0.13	1	10	0.7257

Multivariate Statistics and F Approximations

S=3 M=-0.5 N=3

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.12593148	2.93	9	19.621	0.0223
Pillai's Trace	1.03343732	1.75	9	30	0.1204
Hotelling-Lawley Trace	5.69042615	4.76	9	9.8113	0.0119
Roy's Greatest Root	5.46489324	18.22	3	10	0.0002

Figure 13.12. Canonical Correlation: Correlations and Eigenvalues

The first canonical correlation (the correlation between the first pair of canonical variables) is 0.9194. This value represents the highest possible correlation between any linear combination of the job satisfaction variables and any linear combination of the job characteristics variables.

Figure 13.12 also displays the likelihood ratios and associated statistics for testing the hypothesis that the canonical correlations in the current row and all that follow are zero. The first approximate *F* value of 2.93 corresponds to the test that all three canonical correlations are zero. Since the *p*-value is small (0.0223), you can reject the null hypothesis at the $\alpha = 0.05$ level. The second approximate *F* value of 0.49 corresponds to the test that both the second and the third canonical correlations are zero. Since the *p*-value is large (0.7450), you fail to reject the hypothesis and conclude that only the first canonical correlation is significant at the $\alpha = 0.05$ level.

Several multivariate statistics and *F* test approximations are also provided. These statistics test the null hypothesis that all canonical correlations are zero. The small *p*-values for these tests (< 0.05), except for Pillai's Trace, suggest rejecting the null hypothesis that all canonical correlations are zero.

Analysis

The CANCORR Procedure

Canonical Correlation Analysis

Raw Canonical Coefficients for the Job Satisfaction

		Satisfy1	Satisfy2	Satisfy3
Career	Career Satisfaction	0.0148378305	-0.026536591	0.0509931964
Supervis	Supervisor Satisfaction	0.0252519157	0.0041970746	-0.02920936
Finance	Financial Satisfaction	0.0243430387	0.4415920204	0.1507204075

Raw Canonical Coefficients for the Job Characteristics

		Characteristic1	Characteristic2	Characteristic3
Variety	Task Variety	-0.004300092	0.031408316	0.0351951723
Feedback	Amount of Feedback	0.0201108856	-0.028134366	0.0152825384
Autonomy	Degree of Autonomy	0.0531209636	0.0064473365	-0.052450863

The CANCORR Procedure

Canonical Correlation Analysis

Standardized Canonical Coefficients for the Job Satisfaction

		Satisfy1	Satisfy2	Satisfy3
Career	Career Satisfaction	0.3028	-0.5416	1.0408
Supervis	Supervisor Satisfaction	0.7854	0.1305	-0.9085
Finance	Financial Satisfaction	0.0538	0.9754	0.3329

Standardized Canonical Coefficients for the Job Characteristics

		Characteristic1	Characteristic2	Characteristic3
Variety	Task Variety	-0.1108	0.8095	0.9071
Feedback	Amount of Feedback	0.5520	-0.7722	0.4194
Autonomy	Degree of Autonomy	0.8403	0.1020	-0.8297

Figure 13.13. Canonical Correlation: Correlation Coefficients

Even though canonical variables are artificial, they can often be identified in terms of the original variables. To identify the variables, inspect the standardized coefficients of the canonical variables and the correlations between the canonical variables and their original variables. Based on the results displayed in Figure 13.12, only the first canonical correlation is significant. Thus, only the first pair of canonical variables (Satisfy1 and Characteristic1) need to be identified.

The standardized canonical coefficients in Figure 13.13 show that the first canonical variable for the Job Satisfaction group is a weighted sum of the variables Supervis (0.7854) and Career (0.3028), with the emphasis on Supervis. The coefficient for the variable Finance is near 0. Therefore, a person satisfied with his or her supervisor and with a large degree of career satisfaction would score high on the canonical variable Satisfaction1.

The coefficients for the Job Characteristics variables show that degree of autonomy (Autonomy) and amount of feedback (Feedback) contribute heavily to the Characteristic1 canonical variable (0.8403 and 0.5520, respectively).

Figure 13.14 displays the table of correlations between the canonical variables and the original variables. Although these univariate correlations must be interpreted with caution, since they do not indicate how the original variables contribute jointly to the canonical analysis, they are often useful in the identification of the canonical variables.

Analysis

Canonical Structure

Correlations Between the Job Satisfaction and Their Canonical Variables

		Satisfy1	Satisfy2	Satisfy3
Career	Career Satisfaction	0.7499	-0.2503	0.6123
Supervis	Supervisor Satisfaction	0.9644	0.0362	-0.2618
Finance	Financial Satisfaction	0.2873	0.8814	0.3750

Correlations Between the Job Characteristics and Their Canonical Variables

		Characteristic1	Characteristic2	Characteristic3
Variety	Task Variety	0.4863	0.6592	0.5736
Feedback	Amount of Feedback	0.6216	-0.5452	0.5625
Autonomy	Degree of Autonomy	0.8459	0.4451	-0.2938

Correlations Between the Job Satisfaction and the Canonical Variables of the Job Characteristics

		Characteristic1	Characteristic2	Characteristic3
Career	Career Satisfaction	0.6895	-0.1048	0.0694
Supervis	Supervisor Satisfaction	0.8867	0.0152	-0.0297
Finance	Financial Satisfaction	0.2642	0.3690	0.0425

Correlations Between the Job Characteristics and the Canonical Variables of the Job Satisfaction

		Satisfy1	Satisfy2	Satisfy3
Variety	Task Variety	0.4471	0.2760	0.0650
Feedback	Amount of Feedback	0.5715	-0.2283	0.0638
Autonomy	Degree of Autonomy	0.7777	0.1863	-0.0333

Figure 13.14. Canonical Correlation: Canonical Structure

As displayed in Figure 13.14, the supervisor satisfaction variable, Supervis, is strongly associated with the Satisfy1 canonical variable ($r = 0.9644$). Slightly less influential is the variable Career, which has a correlation with the canonical variable of 0.7499. Thus, the canonical variable Satisfy1 seems to represent satisfaction with supervisor and career track.

The correlations for the job characteristics variables show that the canonical variable Characteristic1 seems to represent all three measured variables, with the degree of autonomy variable (Autonomy) being the most influential (0.8459).

Hence, you can interpret these results to mean that job characteristics and job satisfaction are related. Jobs that possess a high degree of autonomy and level of feedback are associated with workers who are more satisfied with their supervisors and their careers. Additionally, the analysis suggests that, although the financial component is a factor in job satisfaction, it is not as important as the other satisfaction-related variables.

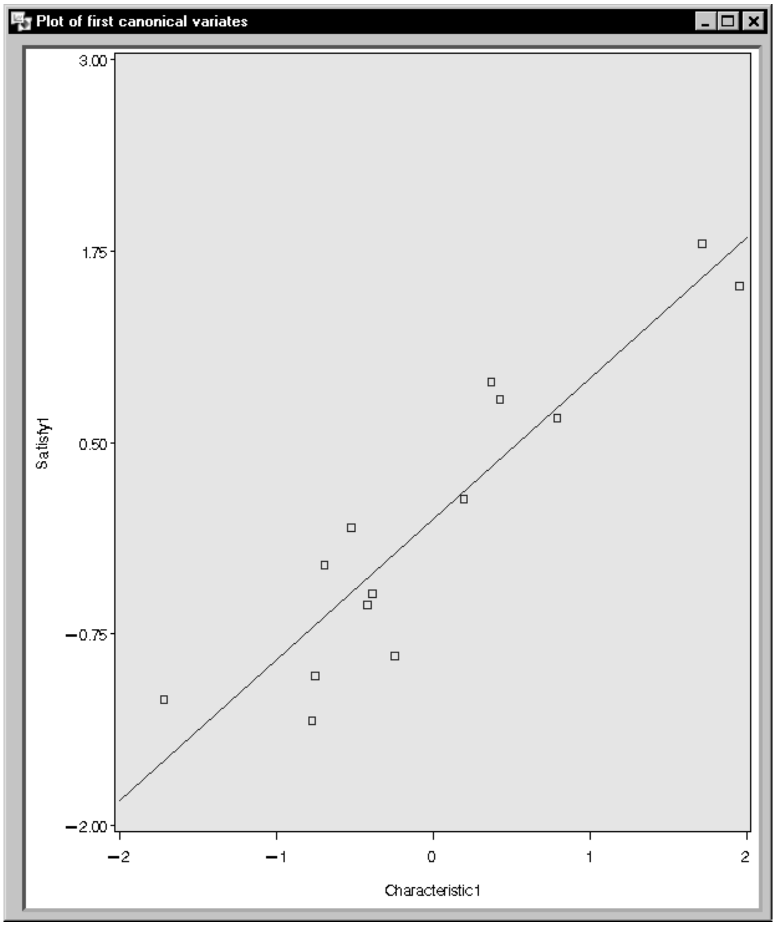


Figure 13.15. Canonical Correlation: Plot of the First Canonical Variables

The plot of the first canonical variables, Satisfy1 and Characteristic1, is displayed in Figure 13.15. The plot depicts the strength of the relationship between the set of job satisfaction variables and the set of job characteristic variables.

References

SAS Institute Inc. (1999), *SAS/STAT User's Guide, Version 7-1*, Cary, NC: SAS Institute Inc.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc., *The Analyst Application, First Edition*, Cary, NC: SAS Institute Inc., 1999. 476 pp.

The Analyst Application, First Edition

Copyright © 1999 SAS Institute Inc., Cary, NC, USA.

ISBN 1–58025–446–2

All rights reserved. Printed in the United States of America. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, by any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher, SAS Institute, Inc.

U.S. Government Restricted Rights Notice. Use, duplication, or disclosure of the software by the government is subject to restrictions as set forth in FAR 52.227–19 Commercial Computer Software-Restricted Rights (June 1987).

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

1st printing, October 1999

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries.® indicates USA registration.

IBM®, ACF/VTAM®, AIX®, APPN®, MVS/ESA®, OS/2®, OS/390®, VM/ESA®, and VTAM® are registered trademarks or trademarks of International Business Machines Corporation. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.

The Institute is a private company devoted to the support and further development of its software and related services.