Chapter 19
Calculating Principal Components

Chapter Table of Contents

Chapter 19
Calculating Principal Components

Principal component analysis is a technique for reducing the complexity of high dimensional data. You can use principal component analysis to approximate high dimensional data with a few dimensions so you can examine them visually. In SAS/INSIGHT software you can calculate principal components, store them, and plot them in two and three dimensions.

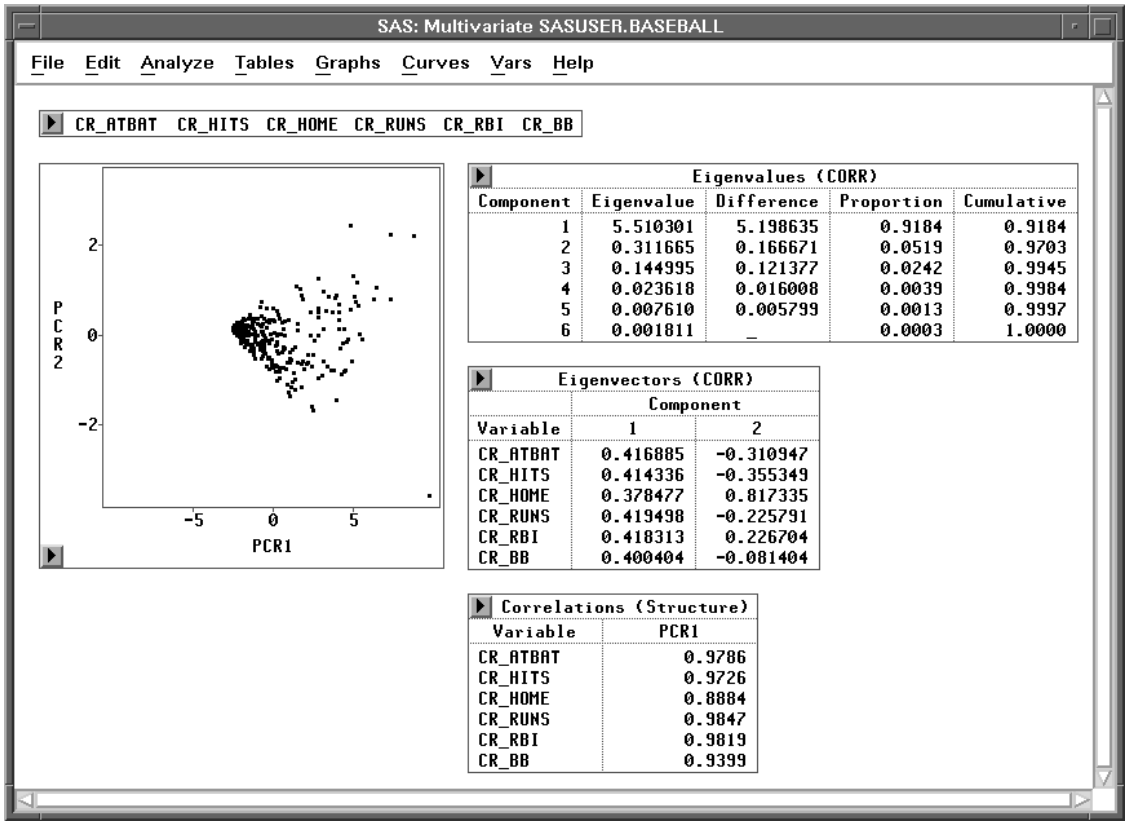


Figure 19.1. Principal Component Analysis

Calculating Principal Components

Principal component analysis summarizes high dimensional data into a few dimensions. Each dimension is called a *principal component* and represents a linear combination of the variables. The first principal component accounts for as much variation in the data as possible. Each succeeding principal component accounts for as much of the variation unaccounted for by preceding principal components as possible.

Consider the **BASEBALL** data set. These data contain performance measures and salary levels for regular hitters and leading substitute hitters in the major leagues in 1986. Suppose you are interested in exploring the relationship between players' performances and their salaries.

If you can first reduce the six career hitting and fielding variables into two or three dimensions—that is, two or three linear combinations of these variables—then graphing these against the **SALARY** variable would be useful. You can then look for relationships between performance and salary.

To create the principal component analysis, follow these steps.

⟹ **Open the BASEBALL data set.**

⟹ **Choose Analyze:Multivariate (Y's).**

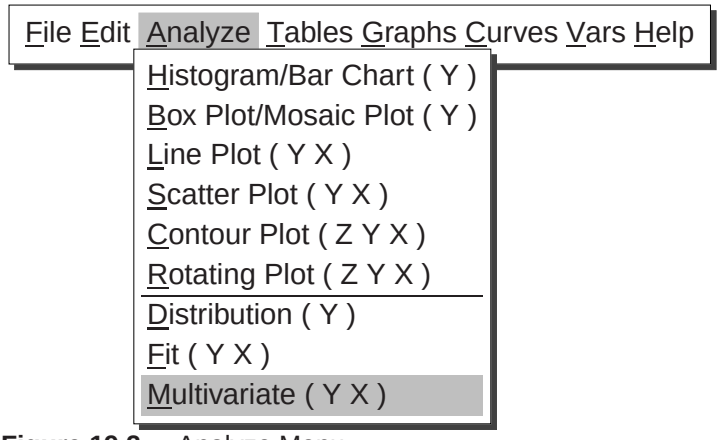


Figure 19.2. Analyze Menu

⟹ **Select the fifteen hitting and fielding variables in the list at the left.**
These are **CR_ATBAT**, **CR_HITS**, **CR_HOME**, **CR_RUNS**, **CR_RBI**, and **CR_BB**. Then Click the **Y** button. The selected variables appear in the **Y** variables list.

⟹ **Select NAME in the list at the left, then click the Label button.**
NAME appears in the **Label** variables list. Your variables dialog should now appear as shown in Figure 19.3.

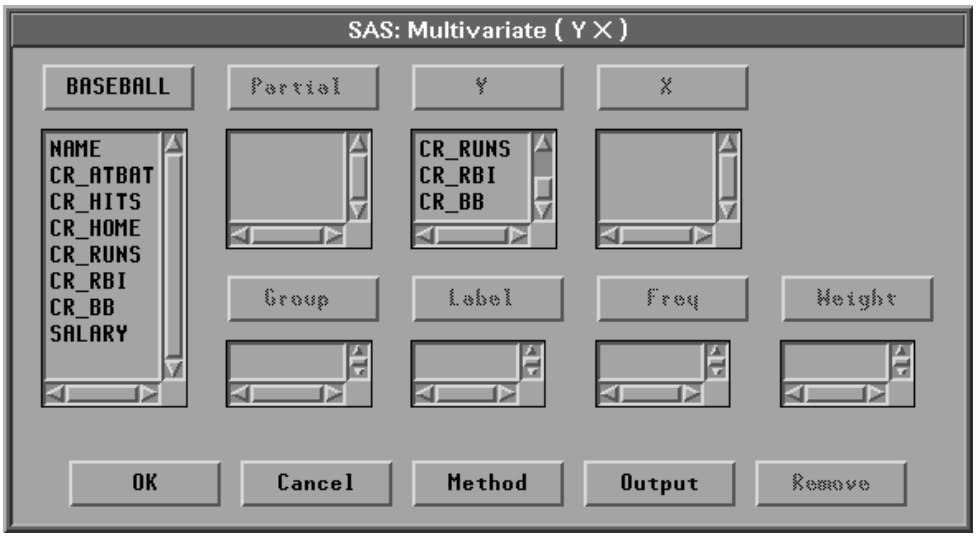


Figure 19.3. Variables Dialog with Variable Roles Assigned

⟹ **Click the Output button.**
The output options dialog appears.

⟹ **Click the Principal Component Analysis check box in the output options dialog.**
This requests a principal component analysis. Your output options dialog should now appear as shown in Figure 19.4.

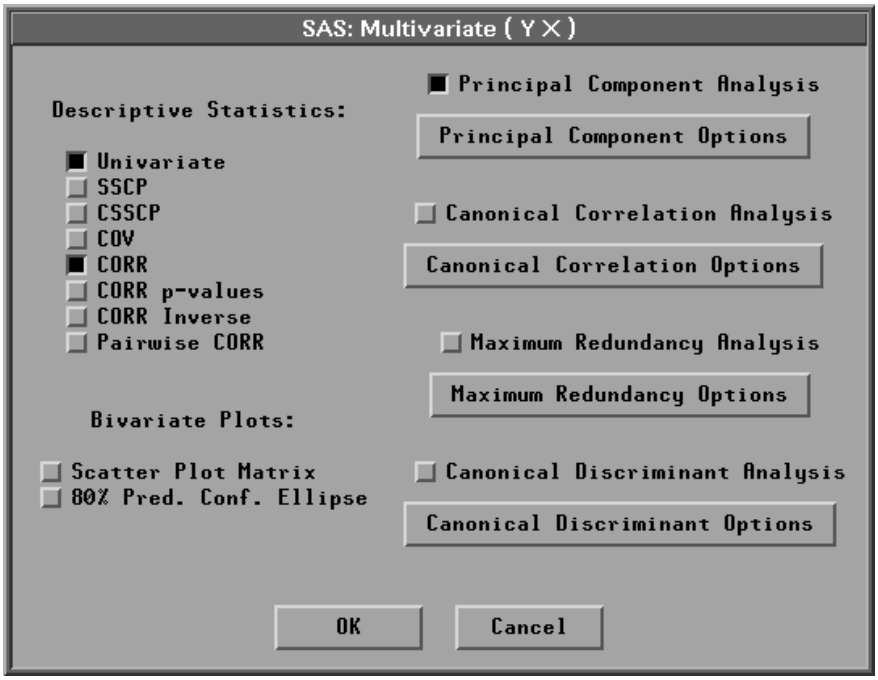


Figure 19.4. Multivariate Output Options Dialog

⟹ **Click the Principal Component Options button in the output options dialog.**
A principal component options dialog should now appear as shown in Figure 19.5.

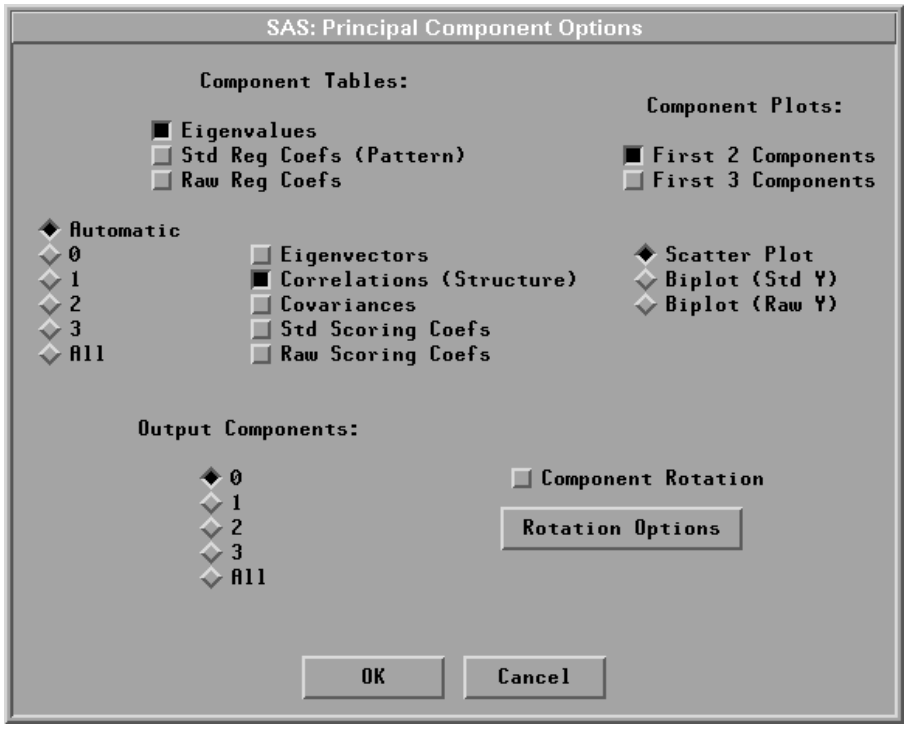


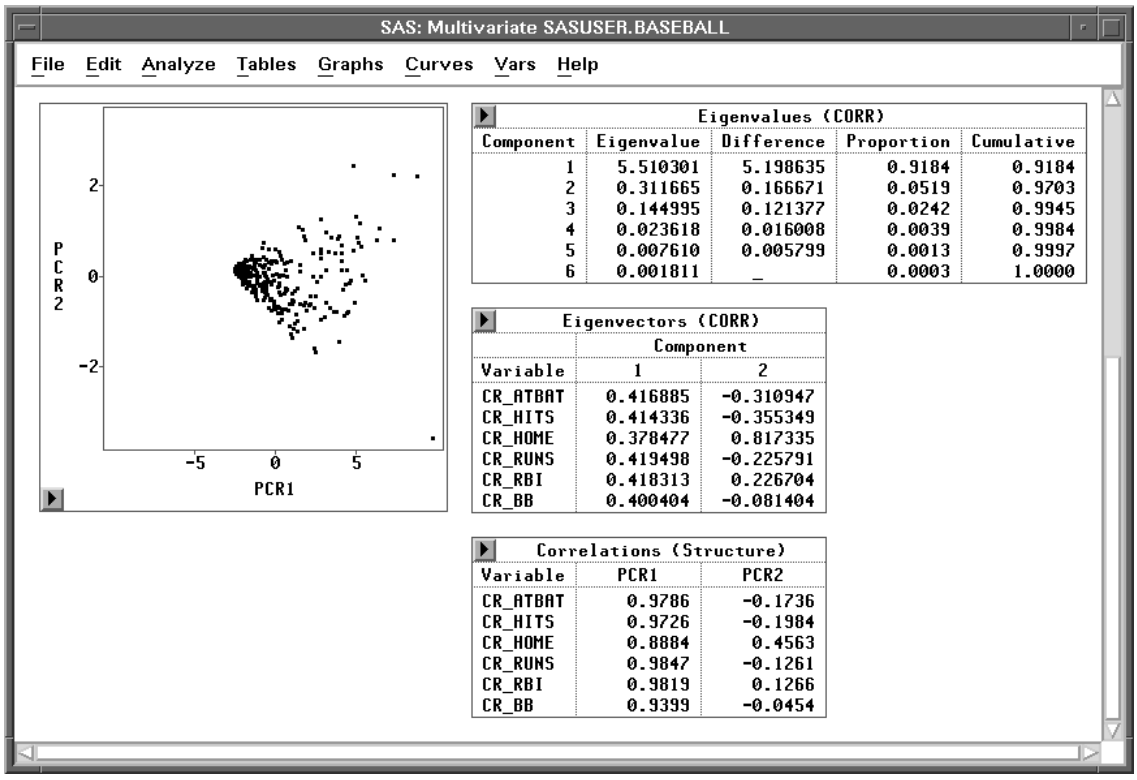
Figure 19.5. Principal Component Options Dialog

⟹ **Click the Eigenvectors check box in the principal component options dialog.**

⟹ **Click the radio mark 2 in the options dialog.**
This requests that the first two principal components are used for tables of eigenvectors and correlations.

† **Note:** By default, the analysis is carried out on the correlation matrix. You can use the covariance matrix instead by setting options with the **Method** button in the Multivariate variables dialog. The covariance matrix is recommended only when all the variables are measured in comparable units.

⟹ **Click OK in all dialogs.**
A multivariate window appears. At the bottom of the window is the principal component analysis, as shown in Figure 19.6.



Eigenvalues (CORR)

Component	Eigenvalue	Difference	Proportion	Cumulative
1	5.510301	5.198635	0.9184	0.9184
2	0.311665	0.166671	0.0519	0.9703
3	0.144995	0.121377	0.0242	0.9945
4	0.023618	0.016008	0.0039	0.9984
5	0.007610	0.005799	0.0013	0.9997
6	0.001811	_	0.0003	1.0000

Eigenvectors (CORR)

	Component	
Variable	1	2
CR_ATBAT	0.416885	-0.310947
CR_HITS	0.414336	-0.355349
CR_HOME	0.378477	0.817335
CR_RUNS	0.419498	-0.225791
CR_RBI	0.418313	0.226704
CR_BB	0.400404	-0.081404

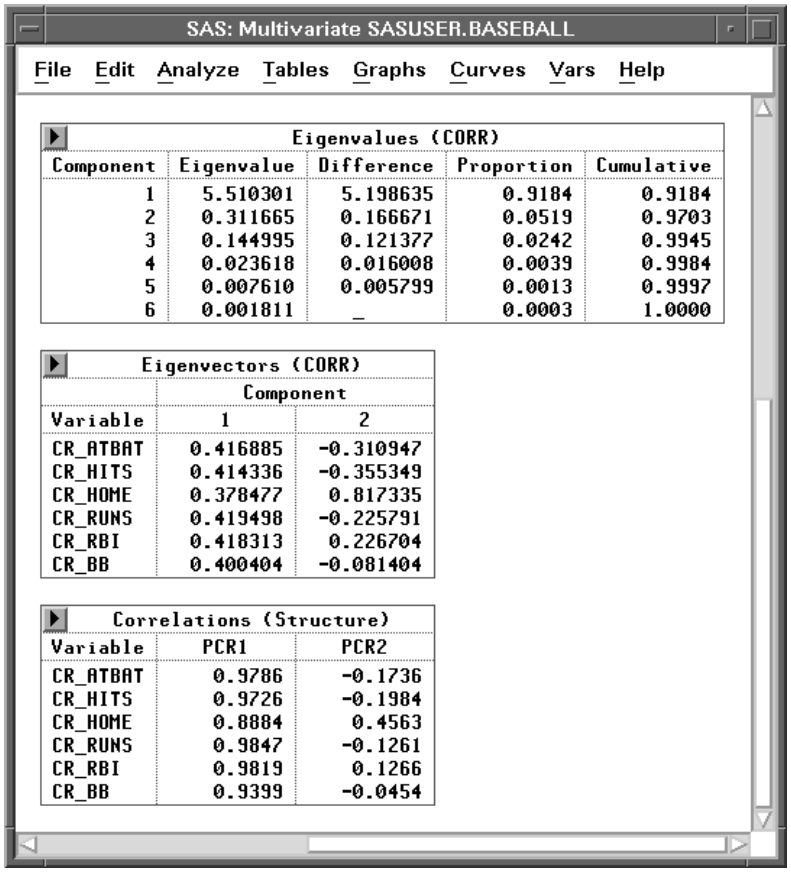
Correlations (Structure)

Variable	PCR1	PCR2
CR_ATBAT	0.9786	-0.1736
CR_HITS	0.9726	-0.1984
CR_HOME	0.8884	0.4563
CR_RUNS	0.9847	-0.1261
CR_RBI	0.9819	0.1266
CR_BB	0.9399	-0.0454

Figure 19.6. Multivariate Window

Principal Component Tables

The **Eigenvalues (CORR)** table illustrated in Figure 19.7 contains all the eigenvalues of the correlation matrix, differences between successive eigenvalues, the proportion of variance explained by each eigenvalue, and the cumulative proportion of the variance explained. Eigenvalues correspond to each of the principal components and represent a partitioning of the total variation in the sample. Because correlations are used, the sum of all the eigenvalues is equal to the number of variables. The first row of the table corresponds to the first principal component, the second row to the second principal component, and so on. In this example, the first two principal components account for over 97% of the variation.



Eigenvalues (CORR)

Component	Eigenvalue	Difference	Proportion	Cumulative
1	5.510301	5.198635	0.9184	0.9184
2	0.311665	0.166671	0.0519	0.9703
3	0.144995	0.121377	0.0242	0.9945
4	0.023618	0.016008	0.0039	0.9984
5	0.007610	0.005799	0.0013	0.9997
6	0.001811	_	0.0003	1.0000

Eigenvectors (CORR)

	Component	
Variable	1	2
CR_ATBAT	0.416885	-0.310947
CR_HITS	0.414336	-0.355349
CR_HOME	0.378477	0.817335
CR_RUNS	0.419498	-0.225791
CR_RBI	0.418313	0.226704
CR_BB	0.400404	-0.081404

Correlations (Structure)

Variable	PCR1	PCR2
CR_ATBAT	0.9786	-0.1736
CR_HITS	0.9726	-0.1984
CR_HOME	0.8884	0.4563
CR_RUNS	0.9847	-0.1261
CR_RBI	0.9819	0.1266
CR_BB	0.9399	-0.0454

Figure 19.7. Principal Component Tables

The **Eigenvectors (CORR)** table illustrated in Figure 19.7 contains the first two eigenvectors of the correlation matrix. Eigenvectors correspond to each of the eigenvalues and associated principal components and are used to form linear combinations of the Y variables. The first column of the table corresponds to the first principal component, and the second column to the second principal component.

Now examine the coefficients making up the eigenvectors. The first component (**PCR1**) appears to be a measure of the player's overall performance as is evidenced by approximately the same magnitude of the coefficients corresponding to all six variables.

Next examine the coefficients making up the eigenvector for the second principal component (**PCR2**). Only the coefficients associated with the variables **CR_HOME** and **CR_RBI** are positive, and the remaining coefficients are negative. The coefficient with the variable **CR_HOME** is considerably larger than any of the other coefficients. This indicates a measure of career home runs performance versus other performance for 1986.

One way to quantify the strength of the linear relationship between the original Y variables and principal components is through the **Correlations (Structure)** table, as shown in Figure 19.7. This correlation matrix contains the correlations between the Y variables and the principal components.

Eigenvector coefficients of a relatively large magnitude translate into larger correlations and vice versa. For example, **PCR2** has one coefficient substantially larger than other coefficients in the same eigenvector, **CR_HOME**. The correlation of the variable with this **PCR2** is also large.

Principal Component Plots

Examine the scatter plot of the first two principal components shown in Figure 19.6. Each marker on the plot represents two principal component scores. The output component scores are a linear combination of the standardized Y variables with coefficients equal to the eigenvectors of the correlation matrix.

⟹ **Click on the observations with the four highest values for PCR1.**
The resulting scatter plot should now appear as shown in Figure 19.8.

These four observations correspond to Mike Schmidt, Reggie Jackson, Tony Perez, and Pete Rose. The label for Mike Schmidt is not shown because the observation is too close to Reggie Jackson. This is not unexpected since the first principal component is a measure of the player's overall career performance.

Now examine observations in the second principal component direction on the scatter plot. Recall that the second component appeared to be a measure of the combined performance of home runs and runs batted in versus other career performance. The observations with large values of **PCR2** correspond to Mike Schmidt and Reggie Jackson. As one might expect, both players have high career-long home runs and runs batted in.

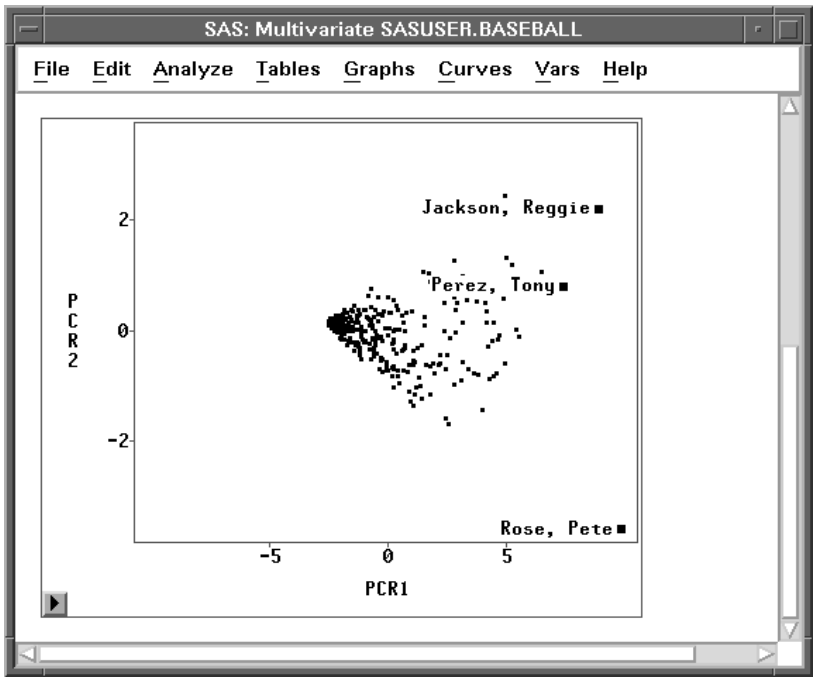


Figure 19.8. Scatter Plot of First Two Principal Components

Plotting Against Original Variables

Now that you have reduced the dimensionality of the career performance variables to two dimensions, you can easily examine scatter plots of these principal components versus the **SALARY** variable. The two principal component scores are automatically stored in the data window.

⟹ **Choose Analyze:Scatter Plot (Y X).**
This displays the scatter plot variables dialog.

⟹ **Select SALARY in the list at the left, then click the Y button.**
SALARY appears in the **Y** variables list.

⟹ **Select PCR1 and PCR2, then click the X button.**
PCR1 and **PCR2** appear in the **X** variables list.

⟹ **Select NAME in the list at the left, then click the Label button.**
NAME appears in the **LABEL** variables list.

A scatter plot variables dialog should now appear as in Figure 19.9.

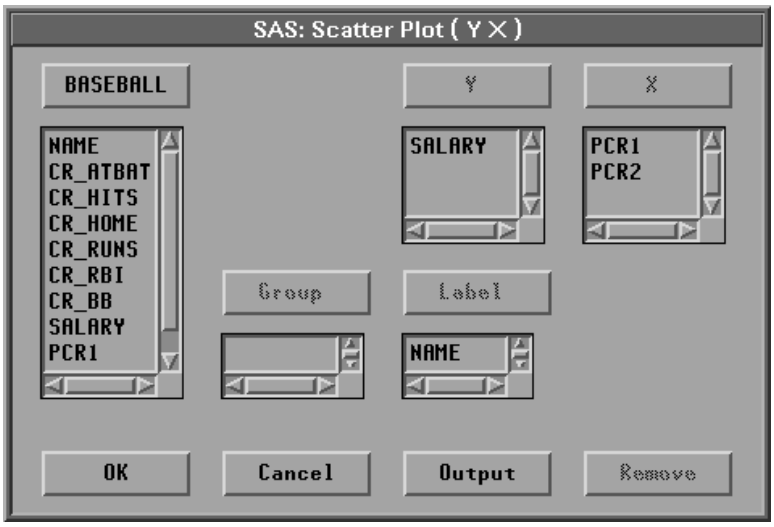


Figure 19.9. Variable Roles Assigned

⟹ **Click the OK button.**
A scatter plot window appears, as shown in Figure 19.10.

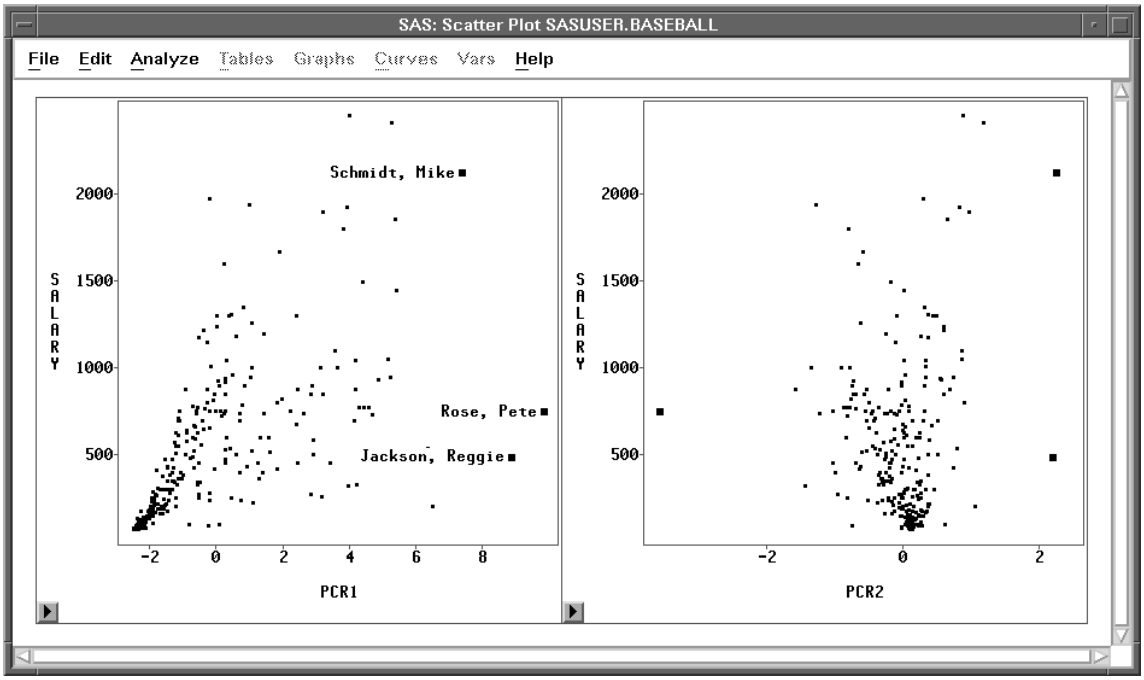


Figure 19.10. SALARY versus First Two Principal Components

Examine the scatter plot of **SALARY** versus **PCR1**, recalling that **PCR1** is highly associated with overall career performance. The linear trend evident in the plot indicates a strong linear relationship between a player's salary and his overall performance. On the other hand, if you examine the scatter plot of **SALARY** versus **PCR2** (which is the contrast between the combined performance of career home runs and runs batted in versus the other performance), you can see that there is no evident relationship.

You can also examine these scatter plots for potential outliers. Click on the observations with large values of **PCR1** in the scatter plot of **SALARY** versus **PCR1**. These observations correspond to players who have had outstanding careers.

Saving Principal Components

This completes the principal component analysis. You began with a high dimensional set of data (six variables) and reduced it to two dimensions (two variables representing principal component scores) that accounted for over 95% of the variation. You were then able to plot the principal component scores against the variable of interest, **SALARY**.

At this point, you may want to save the principal component scores for use in subsequent analyses.

⟹ **Choose Vars:Principal Components:2.**

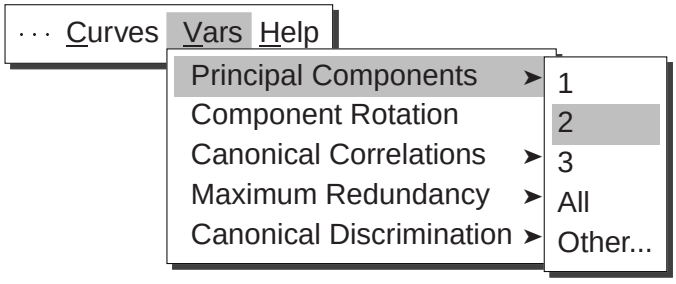


Figure 19.11. Vars Menu

This causes the two variables, **PCR1** and **PCR2**, to be retained in the data window even after you delete the multivariate window. You can then include these variables in later analyses.

⊕ **Related Reading:** Principal Components, Chapter 40.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc., *SAS/INSIGHT User's Guide, Version 8*, Cary, NC: SAS Institute Inc., 1999. 752 pp.

SAS/INSIGHT User's Guide, Version 8

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

ISBN 1–58025–490–X

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, in 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.

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.