Chapter 18
Examining Correlations

Chapter Table of Contents

Chapter 18
Examining Correlations

In this chapter you examine relationships between pairs of variables by looking at correlations.

You can use correlation coefficients to measure the strength of the linear association between two variables. You can also use confidence ellipses in scatter plots as a visual test for bivariate normality and an indication of the strength of the correlation.

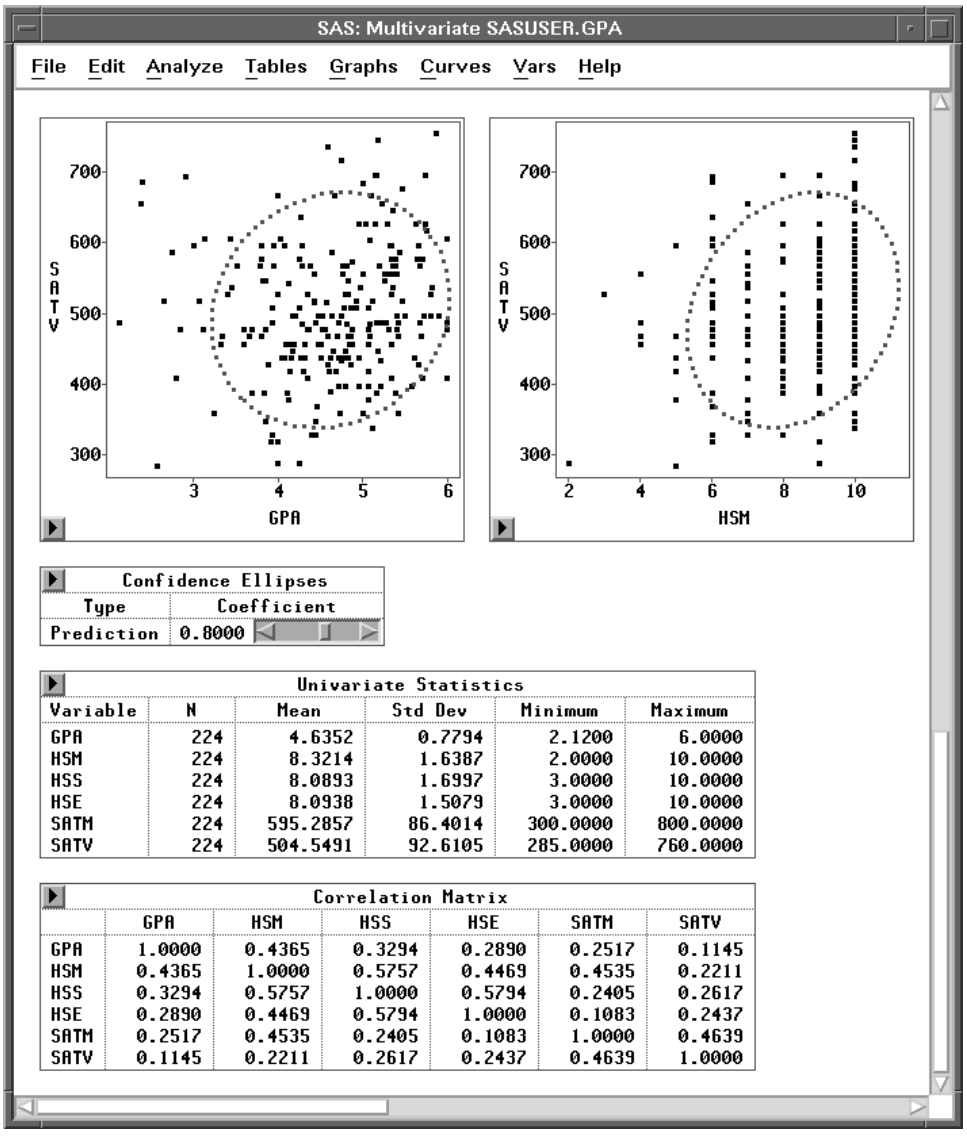


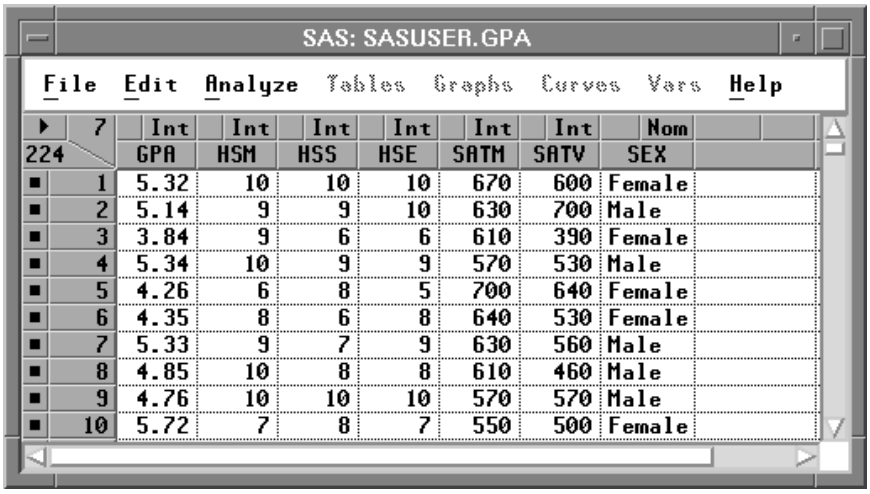
Figure 18.1. Multivariate Window with Correlation Analysis

Creating the Analysis

The **GPA** data set contains information collected to determine which applicants at a university were likely to succeed in its computer science program. The variable **GPA** is the grade point average; **HSM**, **HSS**, and **HSE** are average high school grades in mathematics, science, and English; and **SATM** and **SATV** are scores on the mathematics and verbal portion of the SAT exam (Moore and McCabe 1989).

Follow these steps to create a correlation analysis of the **GPA** data.

⟹ **Open the GPA data set.**



224	GPA (Int)	HSM (Int)	HSS (Int)	HSE (Int)	SATM (Int)	SATV (Int)	SEX (Nom)
1	5.32	10	10	10	670	600	Female
2	5.14	9	9	10	630	700	Male
3	3.84	9	6	6	610	390	Female
4	5.34	10	9	9	570	530	Male
5	4.26	6	8	5	700	640	Female
6	4.35	8	6	8	640	530	Female
7	5.33	9	7	9	630	560	Male
8	4.85	10	8	8	610	460	Male
9	4.76	10	10	10	570	570	Male
10	5.72	7	8	7	550	500	Female

Figure 18.2. **GPA** Data

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

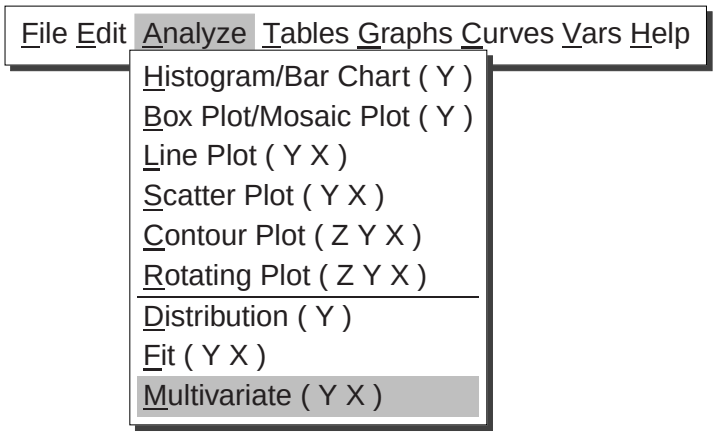


Figure 18.3. Analyze Menu

⟹ **Select GPA, HSM, HSS, HSE, SATM, and SATV. Then click the Y button to assign these variables the Y role.**
Your variables dialog should now appear, as shown in Figure 18.4.

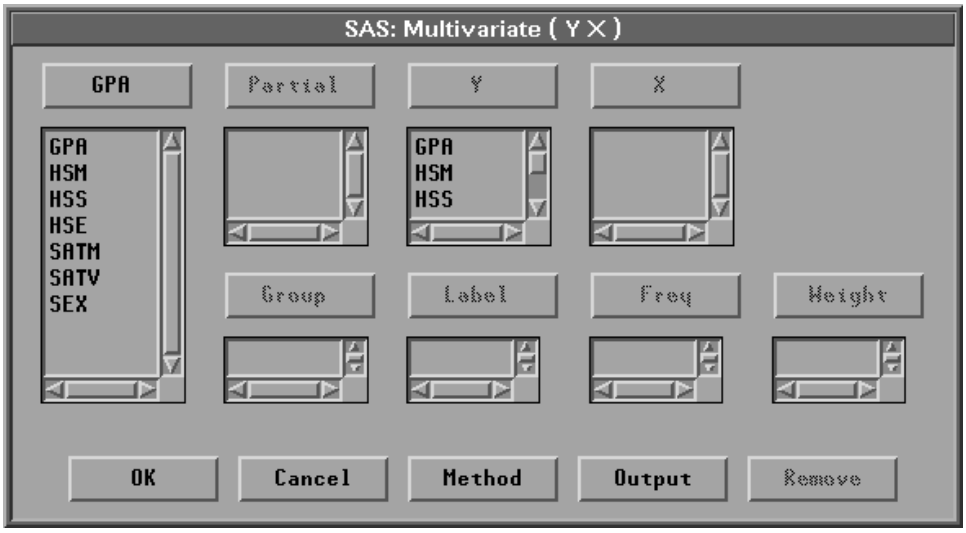


Figure 18.4. Multivariate Variables Dialog

⟹ **Click OK to create the multivariate window.**
By default, the multivariate window contains tables of **Univariate Statistics** and the **Correlation Matrix**.

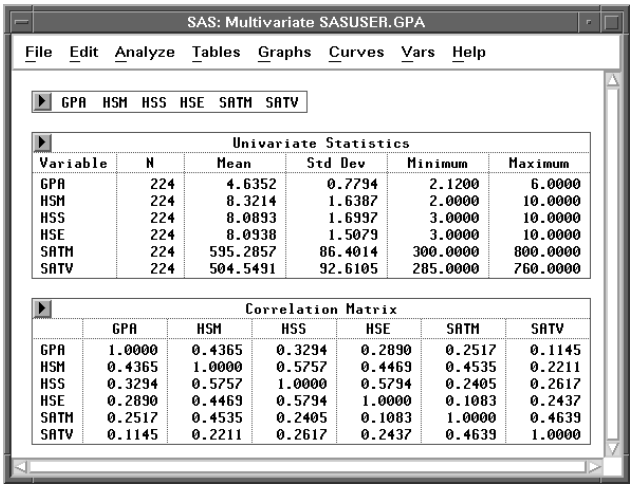


Figure 18.5. Multivariate Window

Correlation Matrix

Examine the **Correlation Matrix** table. The *correlation coefficient* is a numerical measure that quantifies the strength of linear relationships. **GPA**, the grade point average, shows a correlation of 0.4365 with **HSM**, the high school math average. This is not surprising since you would expect the more successful computer science majors to have stronger quantitative skills.

GPA is not as strongly correlated with the other variables and shows a correlation of only 0.1145 with **SATV**. The verbal portion of the SAT exam does not measure the quantitative skills needed by computer science majors.

Confidence Ellipses

To learn more about correlations in the data, add a scatter plot matrix with confidence ellipses for all of the variables under consideration.

⟹ **Choose Curves:Confidence Ellipse:Prediction: 80%.**

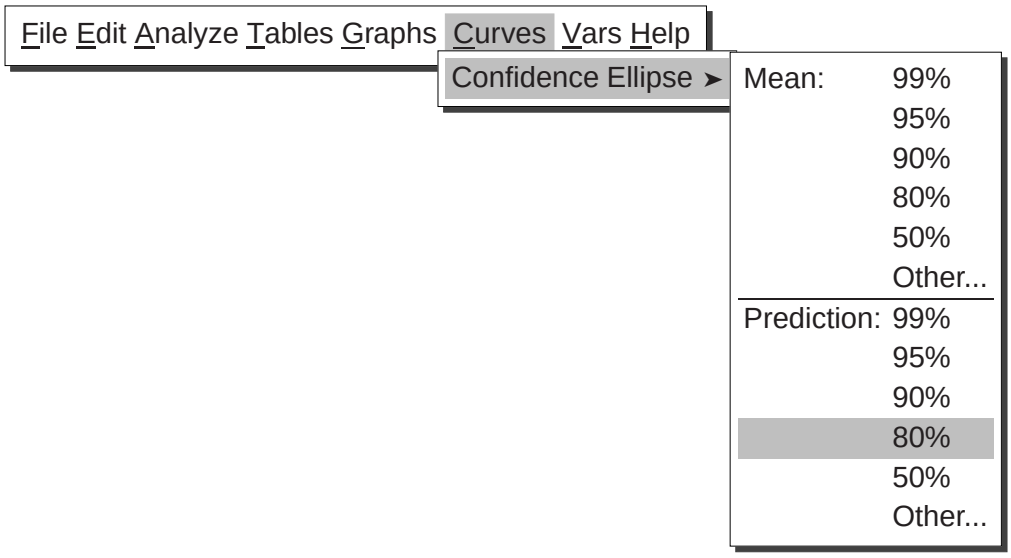


Figure 18.6. Curves Menu

The lower half of the scatter plot matrix for the six variables appears on your display with the 80% prediction confidence ellipses drawn, as shown in Figure 18.7.

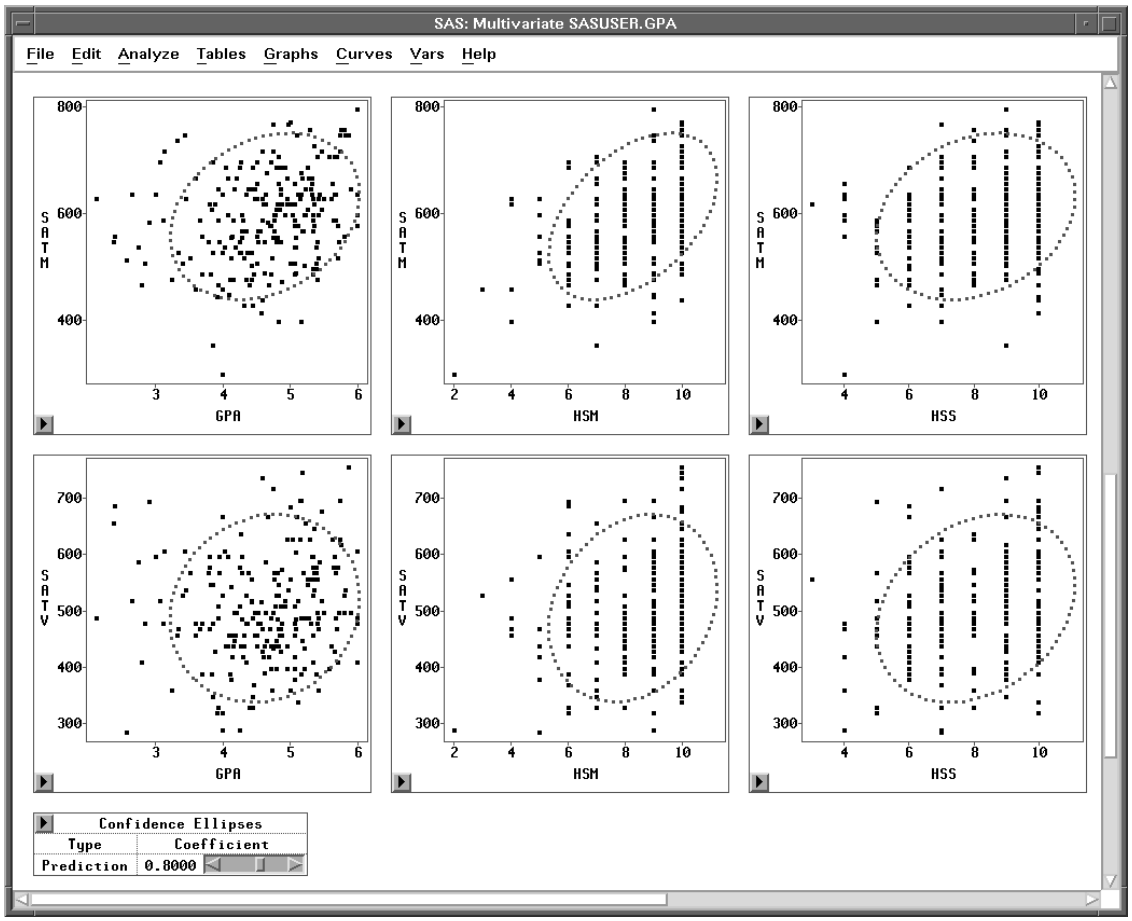


Figure 18.7. Multivariate Window with Confidence Ellipses

There are two ways to interpret the ellipses: as confidence curves for bivariate normal distributions and as indicators of correlation.

As confidence curves, the ellipses show where the specified percentage of the data should lie, assuming a bivariate normal distribution. Under bivariate normality, the percentage of observations falling inside the ellipse should closely agree with the specified confidence level. You can examine the effect of increasing or decreasing the confidence level by adjusting the slider in the **Confidence Ellipses** table below the scatter plot matrix.

Confidence ellipses can also serve as visual indicators of correlations. The confidence ellipse collapses diagonally as the correlation between two variables approaches 1 or -1. The confidence ellipse is more circular when two variables are uncorrelated.

In this case the scatter plots for high school scores (**HSM**, **HSS**, and **HSE**) show a granular appearance that indicates the data are not continuous. These scatter plots clearly do not follow a bivariate normal distribution; therefore, it is not appropriate to interpret confidence ellipses.

The confidence ellipses for **GPA**, **SATM**, and **SATV** can be interpreted. These confidence ellipses contain observations appropriate to the 80% confidence level you specified. The nearly circular appearance of the confidence ellipse in the plot of **GPA** versus **SATV** reflects the small correlation you observed in the **Correlation Matrix** table. The ellipse in the plot of **GPA** versus **SATM** is somewhat more elongated, reflecting a higher correlation.

† **Note:** Visual interpretation of correlations can be subjective because changes in scale affect your perception (Moore and McCabe 1989). When examining correlations, you should use correlation coefficients as well as confidence ellipses.

⊕ **Related Reading:** Correlation Coefficients, Confidence Ellipses,Chapter 40.

References

Moore, D.S. and McCabe, G.P. (1989), *Introduction to the Practice of Statistics*, New York: W.H. Freeman and Company, 179–199.

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

ISBN 1–58025–490–X

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

1st printing, October 1999