Multiple Variable Analysis

statgraphics 18® centurion

Revised: 10/11/2017

Summary

The **Multiple-Variable Analysis** procedure is designed to summarize two or more columns of numeric data. It calculates summary statistics for each variable, as well as correlations and covariances between the variables. The graphs include a scatterplot matrix, star plots, and sunray plots. This procedure is often used prior to constructing a multiple regression model.

Sample StatFolio: *multvar.sgp*

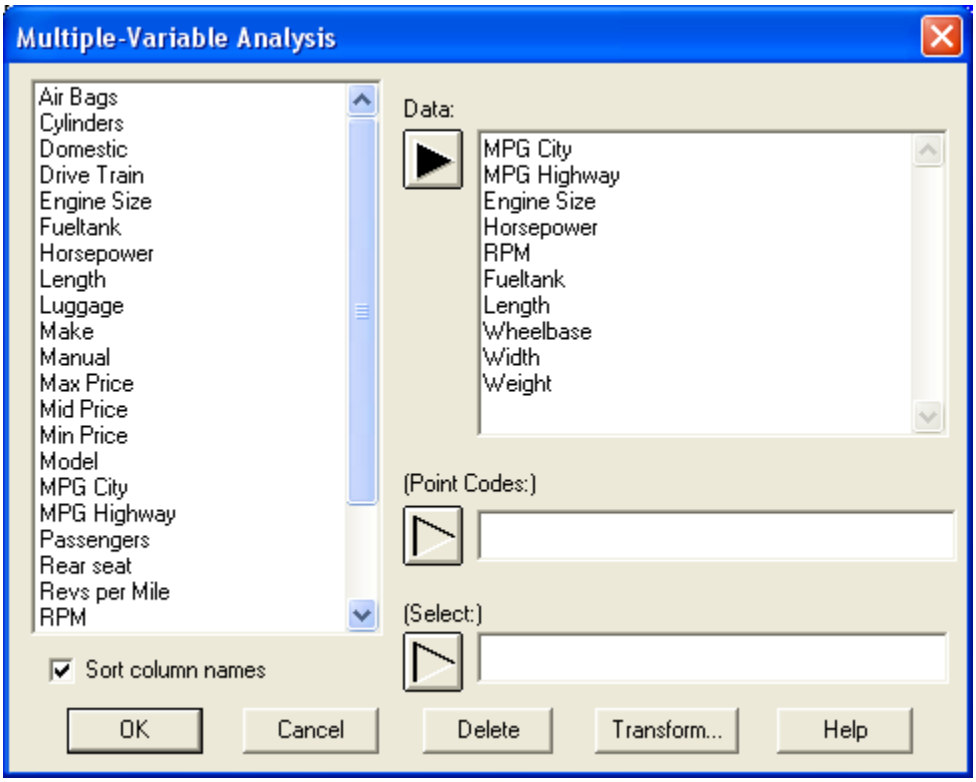
Sample Data

The file *93cars.sgd* contains information on 26 variables for $n = 93$ makes and models of automobiles, taken from Lock (1993). The table below shows a partial list of the data in that file:

Make	*Model*	*MPG City*	*MPG Highway*	*Engine Size*	*Horsepower*	*RPM*	*Fueltank*
Acura	Integra	25	31	1.8	140	6300	13.2
Acura	Legend	18	25	3.2	200	5500	18
Audi	90	20	26	2.8	172	5500	16.9
Audi	100	19	26	2.8	172	5500	21.1
BMW	535i	22	30	3.5	208	5700	21.1
Buick	Century	22	31	2.2	110	5200	16.4
Buick	LeSabre	19	28	3.8	170	4800	18
Buick	Roadmaster	16	25	5.7	180	4000	23
Buick	Riviera	19	27	3.8	170	4800	18.8
Cadillac	DeVille	16	25	4.9	200	4100	18
Cadillac	Seville	16	25	4.6	295	6000	20
Chevrolet	Cavalier	25	36	2.2	110	5200	15.2
Chevrolet	Corsica	25	34	2.2	110	5200	15.6
Chevrolet	Camaro	19	28	3.4	160	4600	15.5

Data Input

The data to be analyzed consists of two or more numeric columns.



- **Data:** numeric columns containing the data to be summarized.
- **Point codes:** optional column with codes to use when creating a scatterplot matrix.
- **Select:** subset selection.

Analysis Summary

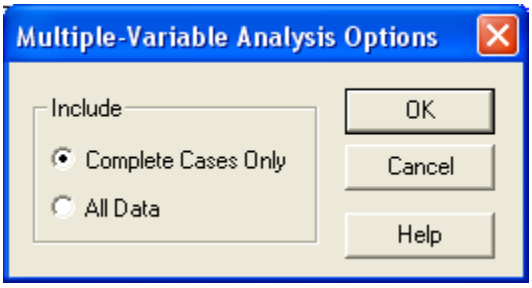
The *Analysis Summary* lists the names of the data columns.

Multiple-Variable Analysis
Data variables:
- MPG City
- MPG Highway
- Engine Size
- Horsepower
- RPM
- Fueltank
- Length
- Wheelbase
- Width
- Weight

There are 93 complete cases for use in the calculations.

Unless changed using *Analysis Options*, only rows containing complete information on all of the variables will be included in the analysis. In the sample data, there are $n = 93$ automobiles with complete information on the $k = 10$ variables listed.

Analysis Options

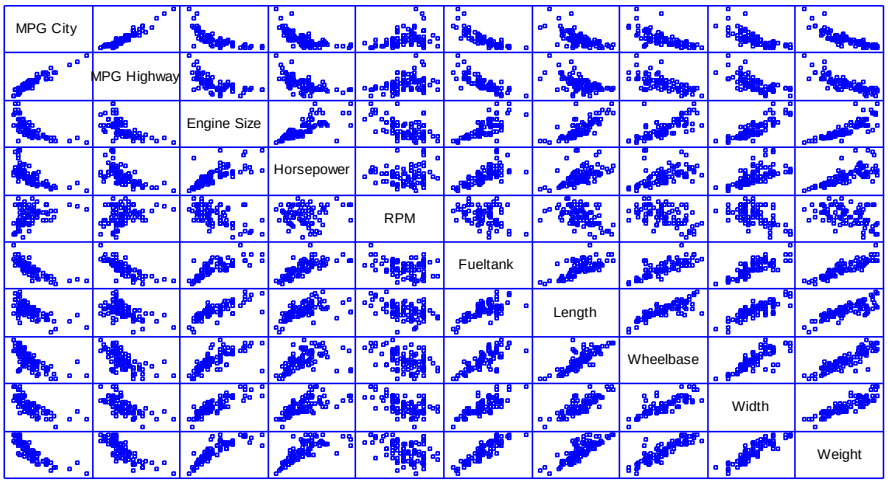


- **Complete Cases Only**: exclude from all plots and statistics any row in which one or more of the input data columns contains a missing value.

- **All Data**: use all data wherever possible.

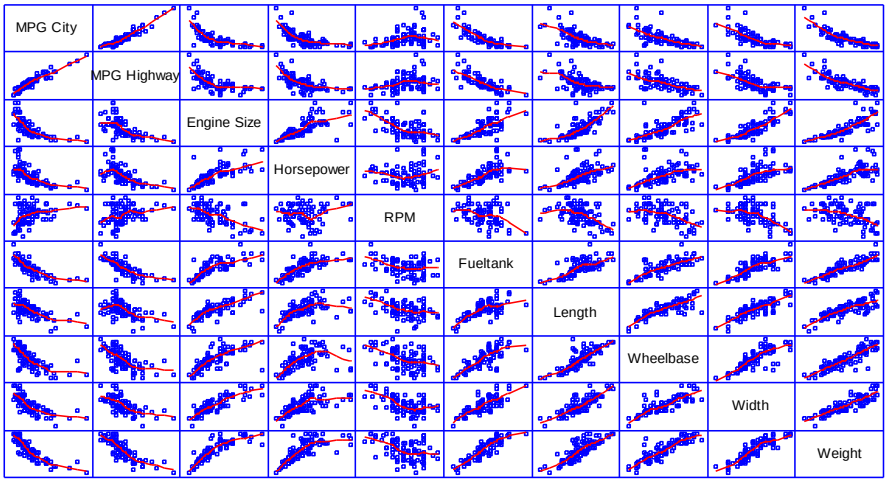
Scatterplot Matrix

The *Scatterplot Matrix* creates a matrix of two variable scatterplots for all pairs of variables.



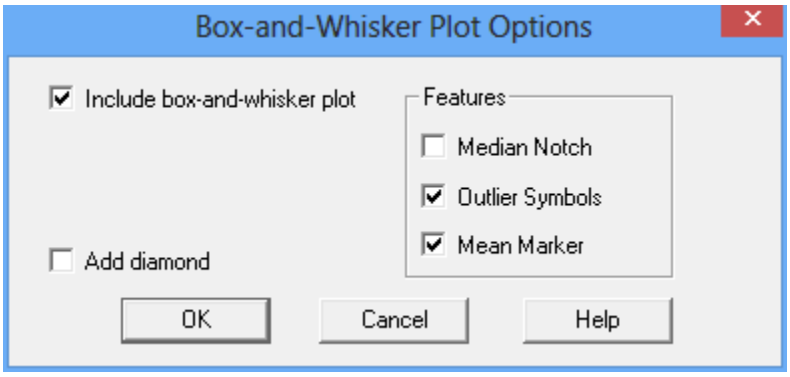
The scatterplot in row *i*, column *j* displays variable *i* on the vertical axis and variable *j* on the horizontal axis. In the matrix, every pair of variables is plotted twice, once with the first variable on the X axis and once with that variable on the Y axis. The plot can often be used to identify those variables which are highly correlated, as well as occasional outliers.

It is sometimes helpful to smooth the scatterplots by pushing the *Smooth/Rotate* button on the analysis toolbar. The plot below uses the default *Robust LOWESS* smoother:



It is now easier to judge the relationships that exist amongst the variables.

Pane Options
If desired, box-and-whisker plots may be added to the diagonal locations on the plot. Properties of the plots are controlled by the following dialog box:



- **Include box-and-whisker plot**: whether to include box-and-whisker plots in the display.

- **Features**: the box-and-whisker plots may include a notch to indicate a 95% confidence interval for the median, outlier symbols to indicate the presence of outside points, and/or a plus sign to indicate the location of the sample mean.

- **Add diamond:** if selected, a diamond will be added to the plot showing a 100(1-α)% confidence interval for the mean at the default system confidence level.

Summary Statistics

The *Summary Statistics* pane calculates a number of different statistics that are commonly used to summarize a sample of *n* observations:

Summary Statistics

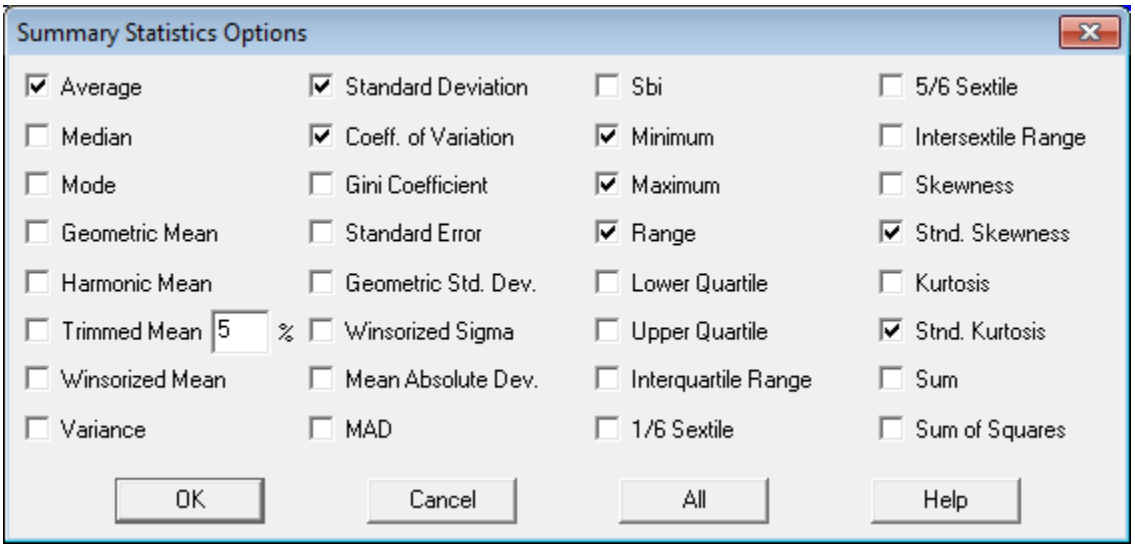
	MPG City	*MPG Highway*	*Engine Size*	*Horsepower*	*RPM*	*Fueltank*
Count	93	93	93	93	93	93
Average	22.3656	29.086	2.66774	143.828	5280.65	16.6645
Standard deviation	5.61981	5.33173	1.03736	52.3744	596.732	3.27937
Coeff. of variation	25.127%	18.3309%	38.8854%	36.4146%	11.3004%	19.6788%
Minimum	15.0	20.0	1.0	55.0	3800.0	9.2
Maximum	46.0	50.0	5.7	300.0	6500.0	27.0
Range	31.0	30.0	4.7	245.0	2700.0	17.8
Interquartile range	7.0	5.0	1.5	67.0	950.0	4.3
Stnd. skewness	6.71035	4.84211	3.38353	3.74696	-1.01784	0.425772
Stnd. kurtosis	7.88248	5.14606	0.750048	2.18677	-0.80606	0.250406

Most of the statistics fall into one of three categories:

1. measures of *central tendency* – statistics that characterize the "center" of the data.
2. measure of *dispersion* – statistics that characterize the spread of the data.
3. measures of *shape* – statistics that characterize the shape of the data relative to a normal distribution.

The statistics included in the table by default are controlled by the settings on the *Stats* pane of the *Preferences* dialog box. Within the procedure, the selection may be changed using *Pane Options*. The meaning of each statistic is described in the *One Variable Analysis* documentation.

Pane Options



Select the desired statistics.

Confidence Intervals

The *Confidence Intervals* pane displays confidence intervals for the mean and standard deviation of each variable.

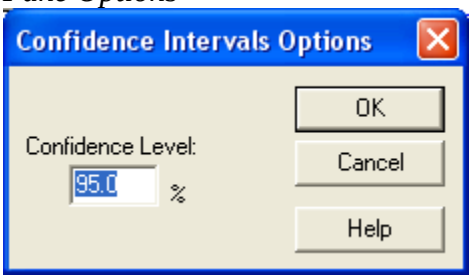
95.0 percent confidence intervals

	Mean	*Stnd. error*	*Lower limit*	*Upper limit*
MPG City	22.3656	0.582747	21.2082	23.523
MPG Highway	29.086	0.552874	27.988	30.1841
Engine Size	2.66774	0.10757	2.4541	2.88138
Horsepower	143.828	5.43097	133.042	154.614
RPM	5280.65	61.8782	5157.75	5403.54
Fueltank	16.6645	0.340055	15.9891	17.3399
Length	183.204	1.5142	180.197	186.212
Wheelbase	103.946	0.707167	102.542	105.351
Width	69.3763	0.391863	68.5981	70.1546
Weight	3072.9	61.1694	2951.42	3194.39

	Sigma	*Lower limit*	*Upper limit*
MPG City	5.61981	4.91195	6.56794
MPG Highway	5.33173	4.66015	6.23125
Engine Size	1.03736	0.906698	1.21238
Horsepower	52.3744	45.7774	61.2106
RPM	596.732	521.568	697.407
Fueltank	3.27937	2.8663	3.83264
Length	14.6024	12.7631	17.066
Wheelbase	6.81967	5.96067	7.97023
Width	3.77899	3.30299	4.41654
Weight	589.897	515.594	689.419

95% confidence intervals are constructed in such a way that, in repeated sampling, 95% of such intervals will contain the true value of the parameter being estimated. You can also view a confidence interval as specifying the "margin of error" in the same manner as stated when taking an opinion poll. For example, the confidence interval for the mean miles per gallon in city driving runs from 21.2 to 23.5.

Pane Options



- **Confidence Level**: level of confidence for the intervals.

Correlations

Correlation coefficients measure the strength of the linear relationship between two columns on a scale of –1 to +1. The larger the absolute value of the correlation, the stronger the linear relationship between the two variables. STATGRAPHICS presents correlation coefficients as a matrix, a section of which is shown below:

Correlations

	MPG City	MPG Highway	Engine Size	Horsepower	RPM
MPG City		0.9439	-0.7100	-0.6726	0.3630
		(93)	(93)	(93)	(93)
		0.0000	0.0000	0.0000	0.0003
MPG Highway	0.9439		-0.6268	-0.6190	0.3135
	(93)		(93)	(93)	(93)
	0.0000		0.0000	0.0000	0.0022
Engine Size	-0.7100	-0.6268		0.7321	-0.5479
	(93)	(93)		(93)	(93)
	0.0000	0.0000		0.0000	0.0000
Horsepower	-0.6726	-0.6190	0.7321		0.0367
	(93)	(93)	(93)		(93)
	0.0000	0.0000	0.0000		0.7270
RPM	0.3630	0.3135	-0.5479	0.0367	
	(93)	(93)	(93)	(93)	
	0.0003	0.0022	0.0000	0.7270	

Correlation
(Sample Size)
P-Value

For each pair of variables, the table shows:

1. r_{ij}, the estimated Pearson product moment correlation coefficient between the row variable *i* and the column variable *j*.

2. n_{ij}, the number of cases used to estimate that correlation. Depending on *Analysis Options*, the correlation may be calculated using rows with complete information on all variables or using all rows with non-missing values for the selected pair of variables.

3. P_{ij}, a P-value that can be used to test the hypothesis that the correlation between the two variables equals 0.

Small P-Values (less than 0.05 if operating at the 5% significance level) correspond to statistically significant correlations. In the above table, all pairs of variables show significant correlations except *RPM* and *Horsepower*.

Rank Correlations

If the presence of outliers is suspected, then the correlation between each pair of variables can be calculated using a rank correlation coefficient instead of a product-moment correlation. The *Rank Correlations* pane displays a table of correlations based on either of the following:

1. **Spearman rank correlations** – These correlations are calculated by first replacing the data values in each variable by their ranks (on a scale of 1 to *n*) and then calculating the degree of disagreement between the ranks.

2. **Kendall rank correlations** – These correlations are based on the number of concordant and discordant pairs of observations, where a concordant pair is one in which the variables in the first row are either both larger than the variables in the second row or both smaller than the variables in the second row.

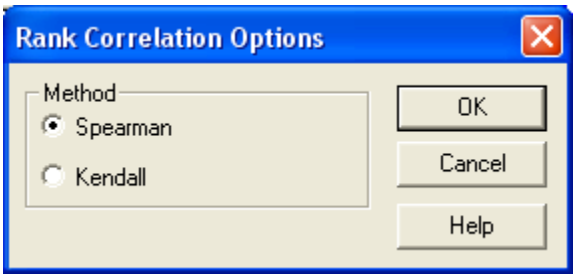
The output is similar to that for the product-moment correlations:

Spearman Rank Correlations

	MPG City	MPG Highway	Engine Size	Horsepower	RPM
MPG City		0.9359	-0.8212	-0.7893	0.3896
		(93)	(93)	(93)	(93)
		0.0000	0.0000	0.0000	0.0002
MPG Highway	0.9359		-0.7257	-0.7100	0.3156
	(93)		(93)	(93)	(93)
	0.0000		0.0000	0.0000	0.0025
Engine Size	-0.8212	-0.7257		0.8142	-0.5295
	(93)	(93)		(93)	(93)
	0.0000	0.0000		0.0000	0.0000
Horsepower	-0.7893	-0.7100	0.8142		-0.0587
	(93)	(93)	(93)		(93)
	0.0000	0.0000	0.0000		0.5731
RPM	0.3896	0.3156	-0.5295	-0.0587	
	(93)	(93)	(93)	(93)	
	0.0002	0.0025	0.0000	0.5731	

Correlation
(Sample Size)
P-Value

Pane Options



- **Method** – the method used to calculate the rank correlation coefficients.

Covariances

Covariances provide a measure of the extent to which two variables vary together.

Covariances

	MPG City	MPG Highway	Engine Size	Horsepower	RPM
MPG City	31.5823	28.2834	-4.13917	-197.98	1217.48
	(93)	(93)	(93)	(93)	(93)
MPG Highway	28.2834	28.4273	-3.46676	-172.865	997.335
	(93)	(93)	(93)	(93)	(93)
Engine Size	-4.13917	-3.46676	1.07612	39.777	-339.164
	(93)	(93)	(93)	(93)	(93)
Horsepower	-197.98	-172.865	39.777	2743.08	1146.63
	(93)	(93)	(93)	(93)	(93)
RPM	1217.48	997.335	-339.164	1146.63	356089.0
	(93)	(93)	(93)	(93)	(93)

Covariance
(Sample Size)

The covariance between variable x and variable y is calculated from

$$\text{cov}(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} \qquad (1)$$

The covariances can be saved to the datasheet for use in other calculations if desired.

Partial Correlations

The *Partial Correlations* pane displays coefficients that measure the strength of the relationship between each pair of variables having already accounted for the relationships with other variables:

Partial Correlations
Variables partialed out: all others

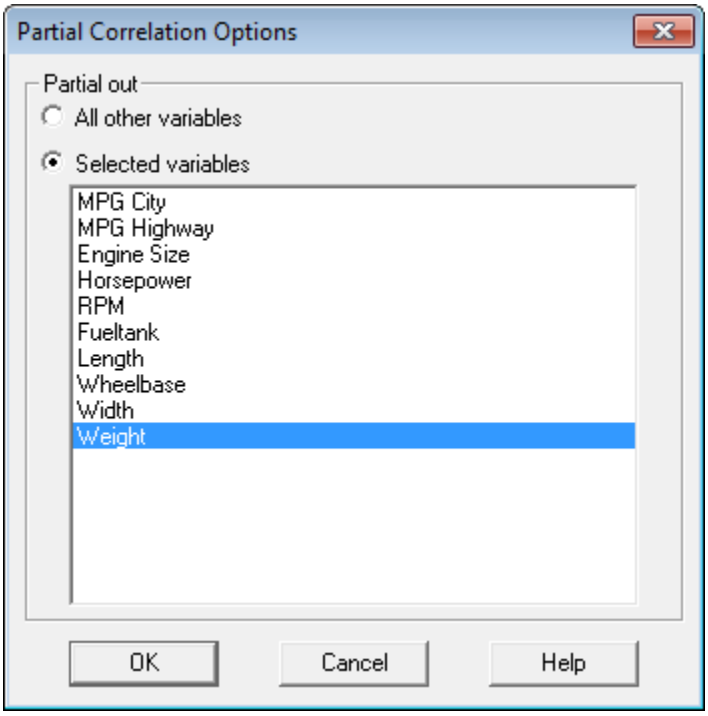
	MPG City	MPG Highway	Engine Size	Horsepower	RPM
MPG City		0.8507	0.0891	-0.2369	0.1960
		(93)	(93)	(93)	(93)
		0.0000	0.4174	0.0290	0.0723
MPG Highway	0.8507		-0.0192	0.1925	-0.1126
	(93)		(93)	(93)	(93)
	0.0000		0.8613	0.0776	0.3051
Engine Size	0.0891	-0.0192		0.6729	-0.6704
	(93)	(93)		(93)	(93)
	0.4174	0.8613		0.0000	0.0000
Horsepower	-0.2369	0.1925	0.6729		0.7994
	(93)	(93)	(93)		(93)
	0.0290	0.0776	0.0000		0.0000
RPM	0.1960	-0.1126	-0.6704	0.7994	
	(93)	(93)	(93)	(93)	
	0.0723	0.3051	0.0000	0.0000	

Correlation
(Sample Size)
P-Value

They are useful in measuring the unique correlation between 2 variables not explainable by the others. For example, *MPG City* is moderately correlated with both *Engine Size* (-0.71) and *Horsepower* (-0.67), but the partial correlations are much smaller since *Engine Size* and *Horsepower* tend to explain the same characteristic of the automobiles.

Pane Options

The *Pane Options* dialog box specifies which variables' effects are removed before calculating the partial correlations:



If "All other variables" is chosen, the partial correlations quantify the relationship between each pair of variables having controlled for the effect of all the others. If "Selected variables" is chosen, then only the effect of the selected variables is controlled for.

For example, the dialog box shown above requests partial correlations between variables controlling for the effect of *weight*. Compare the following tables:

Pearson Product-Moment Correlations

	MPG Highway	Engine Size	Horsepower	RPM	Fueltank	Length	Wheelbase
MPG City	0.9439	-0.7100	-0.6726	0.3630	-0.8131	-0.6662	-0.6671
	(93)	(93)	(93)	(93)	(93)	(93)	(93)
	0.0000	0.0000	0.0000	0.0003	0.0000	0.0000	0.0000

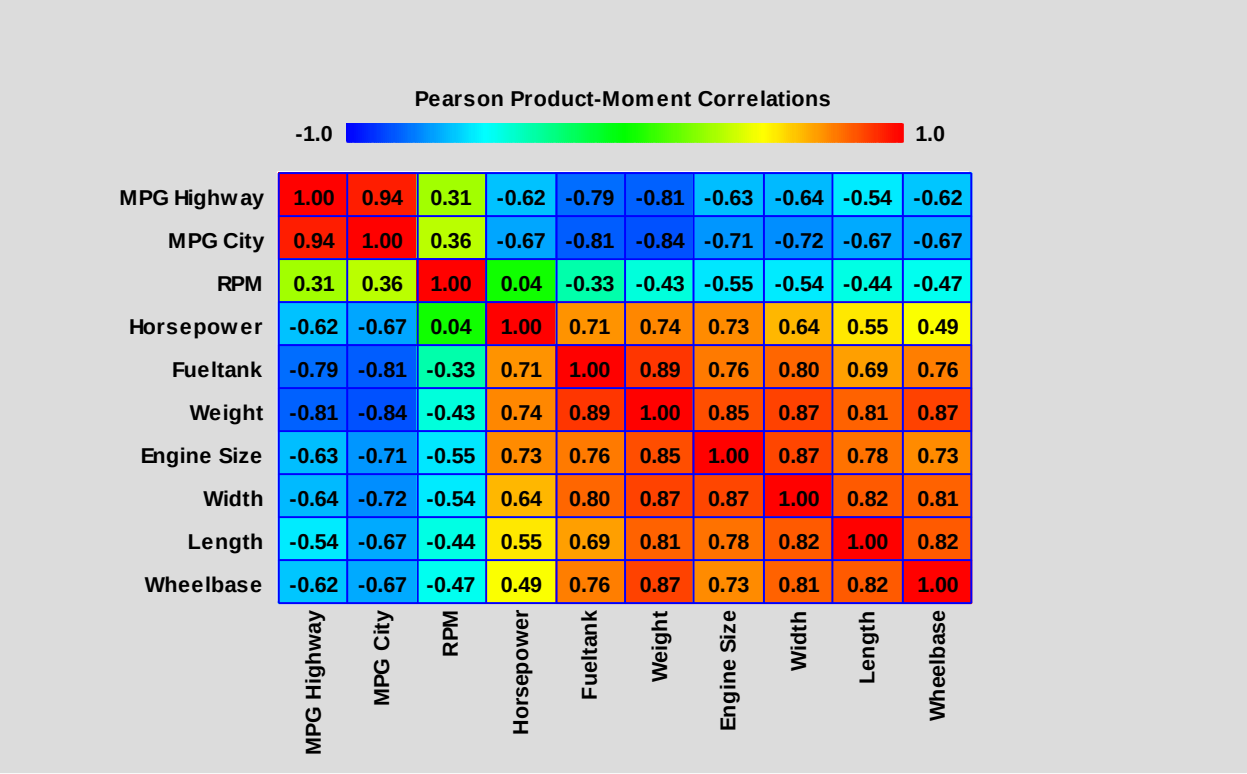
Partial Correlations After Partialing Out Weight

	MPG Highway	Engine Size	Horsepower	RPM	Fueltank	Length	Wheelbase
MPG City	0.8272	0.0087	-0.1372	0.0046	-0.2464	0.0426	0.2583
	(93)	(93)	(93)	(93)	(93)	(93)	(93)
	0.0000	0.9341	0.1921	0.9652	0.0179	0.6865	0.0129

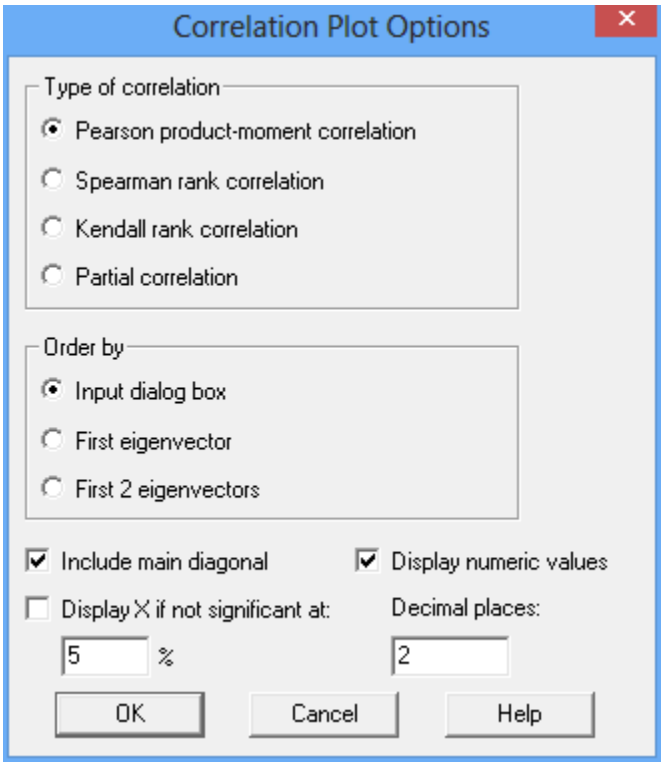
Note how the correlation of variables such as *Engine Size* and *RPM* with *MPG City* nearly vanish after controlling for *Weight*.

Correlation Plot

The *Correlation Plot* pane displays the estimated correlations or partial correlations in the form of a matrix with colored cells. A typical example is shown below:



Pane Options



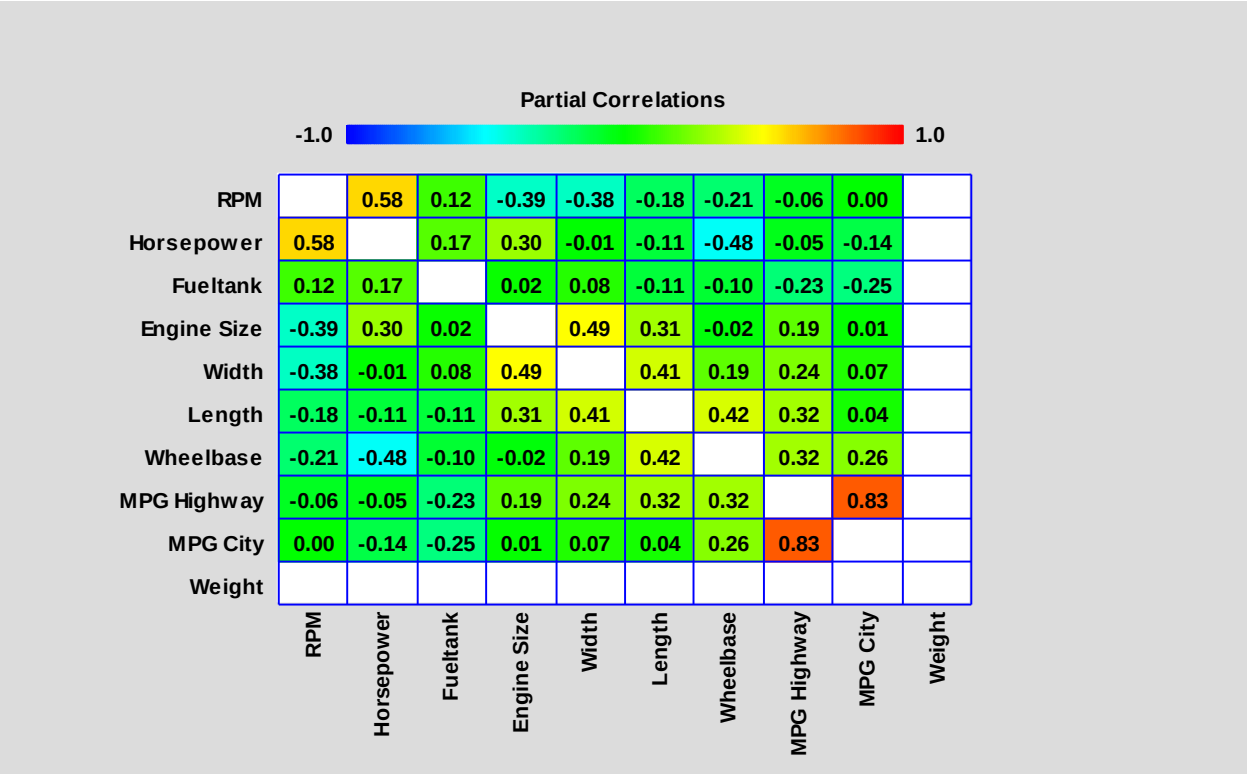
- *Type of correlation* – the type of correlation to be displayed.
- *Order by* – controls the order of the rows and columns. Ordering by the first eigenvector or the first 2 eigenvectors tends to group similar variables together.
- *Include main diagonal* – whether to color the cells on the main diagonal of the matrix.
- *Display X if not significant at* – replaces any numeric values that are not statistically significant at the indicated significance level by an X.
- *Display numeric values* – whether to display the numeric values in each cell.
- *Decimal places* – the number of decimal places displayed if the numeric values are shown.

If you elect to order the rows and columns using eigenvectors, the program computes the eigenvectors of the correlation matrix. It then sorts the variables using either the values of the first eigenvector or the first 2 eigenvectors. In the latter case, it first calculates the angles of the variables in the space of the first 2 eigenvectors according to

$$\alpha_i = \begin{cases} \tan^{-1}(e_{i2}/e_{i1}) & e_{i1} > 0 \\ \tan^{-1}(e_{i2}/e_{i1}) + \pi & otherwise \end{cases} \quad (2)$$

and then plots the variables in order of those angles, starting at the largest gap. For more information, see "Corrgrams: Exploratory Displays for Correlation Matrices" by Michael Friendly, *The American Statistician*, August 19, 2002.

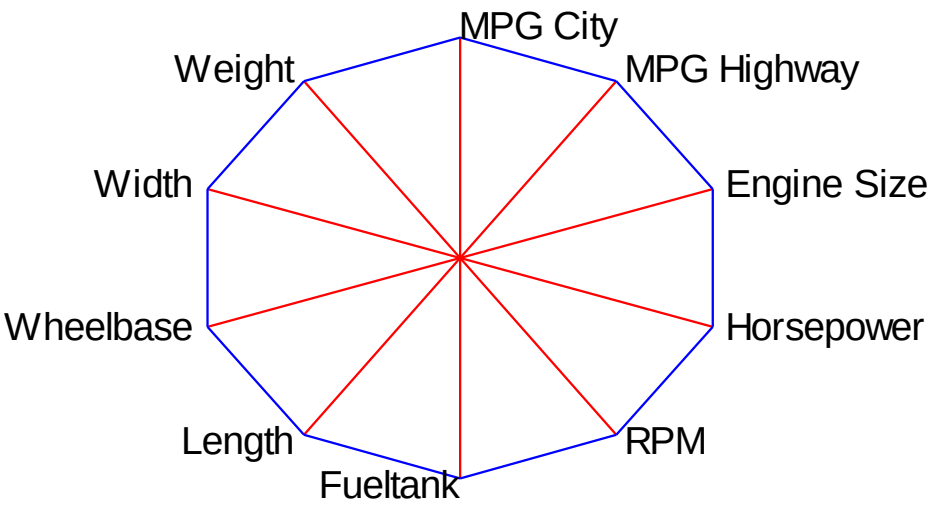
Note: If partial correlations are plotted, the program uses the options for the *Partial Correlations* table to determine which variables to partial out. For example, partialing out *Weight* results in the following display:



Note how similar pairs of variables tend to be grouped together, such as *RPM* with *Horsepower*, *Engine Size* with *Width*, and *MPG Highway* with *MPG City*.

Key Glyph

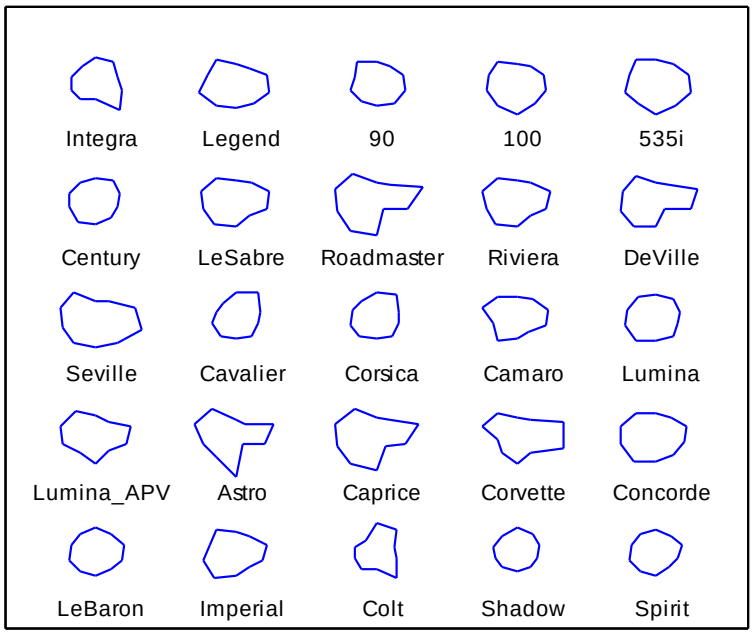
Many methods have been developed to display multivariate data. One useful method is that of a *glyph*. A glyph is a symbolic figure constructed to display the value of multiple quantitative variables. The *Multiple Variable Analysis* procedure generates glyphs in the form of polygons:



The distance from the center of the figure to each vertex is used to represent the relative value of a selected variable. For example, the vertex at the 6 o-clock position represents the size of fuel tank. A car with a large capacity fuel tank will have a vertex located far away from the center in that direction, while the vertex for a car with a small fuel tank will be much closer to the center.

Star Plots

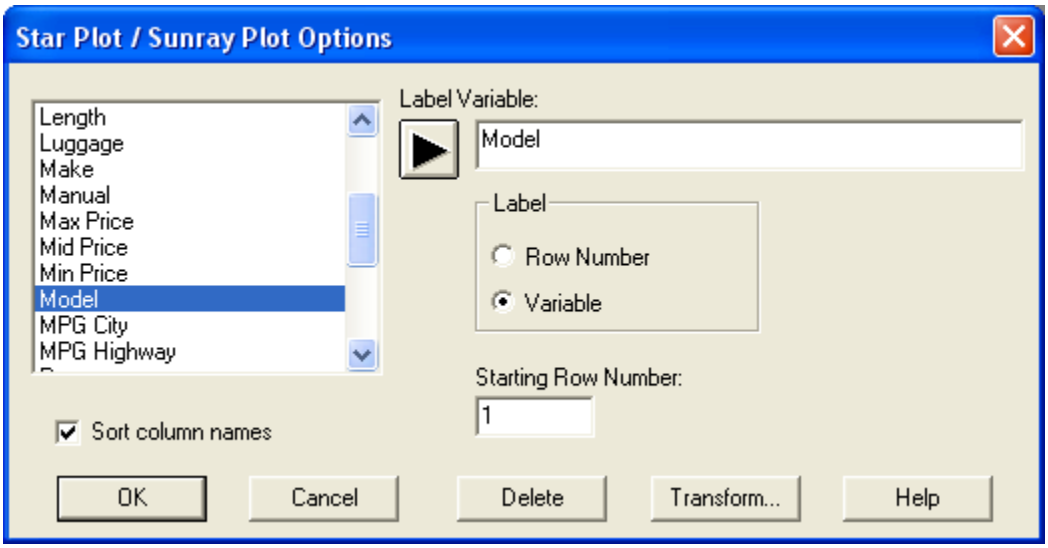
The *Star Plots* pane creates glyphs with the following format:



Glyphs for up to 25 rows may be displayed at one time. The polygons are structured such that the distance of a vertex from the center is very small for the row with a minimum value of the relevant variable and of maximum length for the row with the largest value.

The glyphs are quite useful in clustering the rows, i.e., identifying rows that are similar to each other. For example, the *LeBaron , Shadow*, and *Spirit* have average values for all of the variables and thus have a similar shape. Unusual cases such as the *Astro* also stand out (it has an unusually large fuel tank).

Pane Options



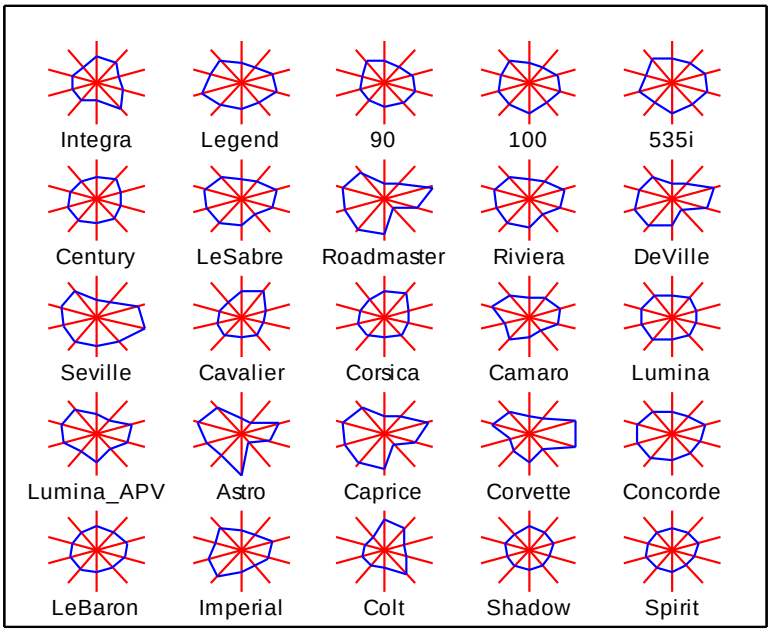
- **Label Variable**: variable (if any) used to label each glyph.

- **Label**: Glyphs may be labeled by their row number of by the value of a selected column in the datasheet.

- **Starting Row Number**: Glyphs for up to 25 rows will be displayed at one time, beginning with the specified row number.

Sun Ray Plots

The *Sun Ray Plots* are similar to the star plots but have a slightly different format:



The major difference is in the placement of the vertices. For each variable, the vertex is located in the middle of the ray if the value of that variable equals the sample mean. It is located at the end of the ray if it is 3 or more standard deviations above the mean and very close to the center of the figure if the value is 3 or more standard deviations below the sample mean.

For the *Astro*, note that the size of its fuel tank is at least 3 standard deviations greater than the mean of the 93 automobiles.

Save Results

The following results may be saved to the datasheet:

1. *Variable Names* – the variables associated with each row of the matrices.
2. *Correlations* – the product moment correlations, one row after the other.
3. *Rank Correlations* – the calculated rank correlations.
4. *Covariances* – the estimated covariances.
5. *Partial Correlations* – the estimated partial correlations.

Calculations

Pearson Product-Moment Correlation Coefficient

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3}$$

$$t = \frac{(n-2)r^2}{1-r^2} \tag{4}$$

The t statistic is compared to a t distribution with *n-2* degrees of freedom.

Spearman Rank Correlation

If U_i equals the rank of x_i and V_i equals the rank of y_i, then Spearman's rank correlation is given by

$$R = \frac{A + B - \sum_{i=1}^{n} D_i^2}{2\sqrt{AB}} \tag{5}$$

where

$$D_i = U_i - V_i \tag{6}$$

$$A = \frac{n^3 - n - \sum_{j=1}^{g_x}\left(t_{j,x}^3 - t_{j,x}\right)}{12} \tag{7}$$

$$B = \frac{n^3 - n - \sum_{j=1}^{g_y}\left(t_{j,y}^3 - t_{j,y}\right)}{12} \tag{8}$$

The quantities *A* and *B* are corrections for tied ranks. They involve summing the number of tied observations $t_{j,x}$ for each of the g_x tied groups. The significance of the correlation is found by comparing

$$z = R\sqrt{n-1} \tag{9}$$

to a standard normal distribution.

Kendall Rank Correlation

If U_i equals the rank of x_i and V_i equals the rank of y_i, then

$$R = \frac{S}{\sqrt{\frac{n(n-1)}{2} - \sum_{j=1}^{g_x}\left(t_{j,x}^3 - t_{j,x}\right)}\sqrt{\frac{n(n-1)}{2} - \sum_{j=1}^{g_y}\left(t_{j,y}^3 - t_{j,y}\right)}} \tag{10}$$

where S is the total number of concordant pairs of observations (pairs in which $(U_i - U_j)(V_i - V_j)$ is positive) minus the number of discordant pairs of observations (pairs in which $(U_i - U_j)(V_i - V_j)$ is negative). The significance of the correlation is found by comparing

$$z = \frac{S}{\sqrt{n(n-1)(2n+5)/18}} \tag{11}$$

to a standard normal distribution.

12q