

## Factor Analysis

### Summary

The **Factor Analysis** procedure is designed to extract  $m$  common factors from a set of  $p$  quantitative variables  $X$ . In many situations, a small number of common factors may be able to represent a large percentage of the variability in the original variables. The ability to express the covariances amongst the variables in terms of a small number of meaningful factors often leads to important insights about the data being analyzed.

This procedure supports both principal components and classical factor analysis. Factor loadings may be extracted from either the sample covariance or sample correlation matrix. The initial loadings may be rotated using either varimax, equimax, or quartimax rotation.

**Sample StatFolio:** *factor analysis.sgp*

### Sample Data:

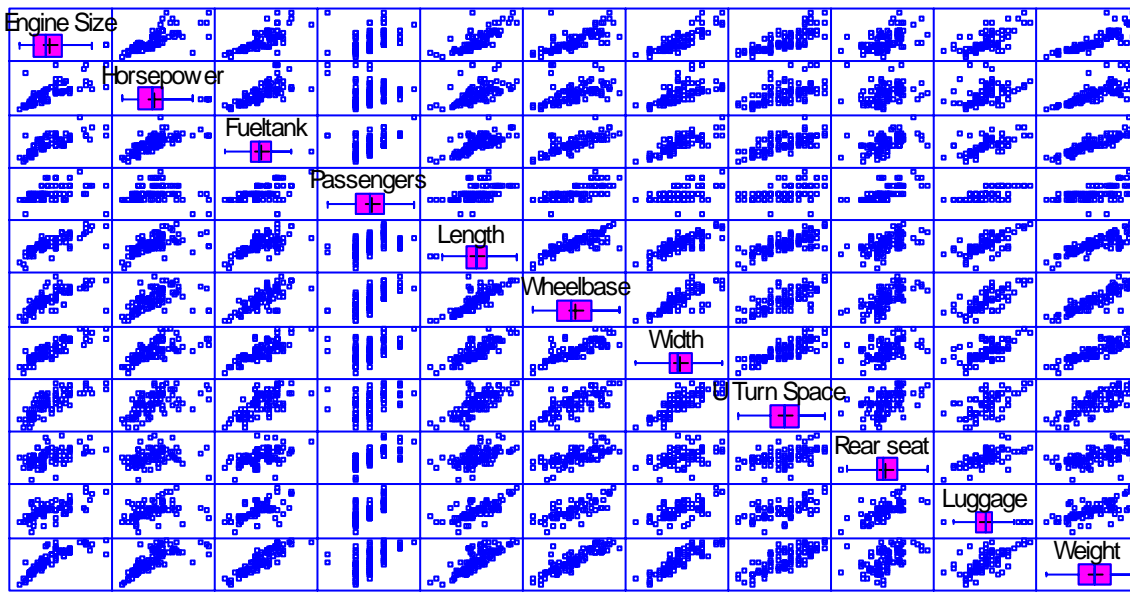
The file *93cars.sf6* contains information on 26 variables for  $n = 93$  makes and models of automobiles, taken from Lock (1993). The table below shows a partial list of the data in that file:

<i>Make</i>	<i>Model</i>	<i>Engine Size</i>	<i>Horsepower</i>	<i>Fuel Tank</i>	<i>Passengers</i>	<i>Length</i>
Acura	Integra	1.8	140	13.2	5	177
Acura	Legend	3.2	200	18	5	195
Audi	90	2.8	172	16.9	5	180
Audi	100	2.8	172	21.1	6	193
BMW	535i	3.5	208	21.1	4	186
Buick	Century	2.2	110	16.4	6	189
Buick	LeSabre	3.8	170	18	6	200
Buick	Roadmaster	5.7	180	23	6	216
Buick	Riviera	3.8	170	18.8	5	198
Cadillac	DeVille	4.9	200	18	6	206
Cadillac	Seville	4.6	295	20	5	204
Chevrolet	Cavalier	2.2	110	15.2	5	182

It is desired to perform a factor analysis on the following variables:

*Engine Size*  
*Horsepower*  
*Fuel tank*  
*Passengers*  
*Length*  
*Wheelbase*  
*Width*  
*U Turn Space*  
*Rear seat*  
*Luggage*  
*Weight*

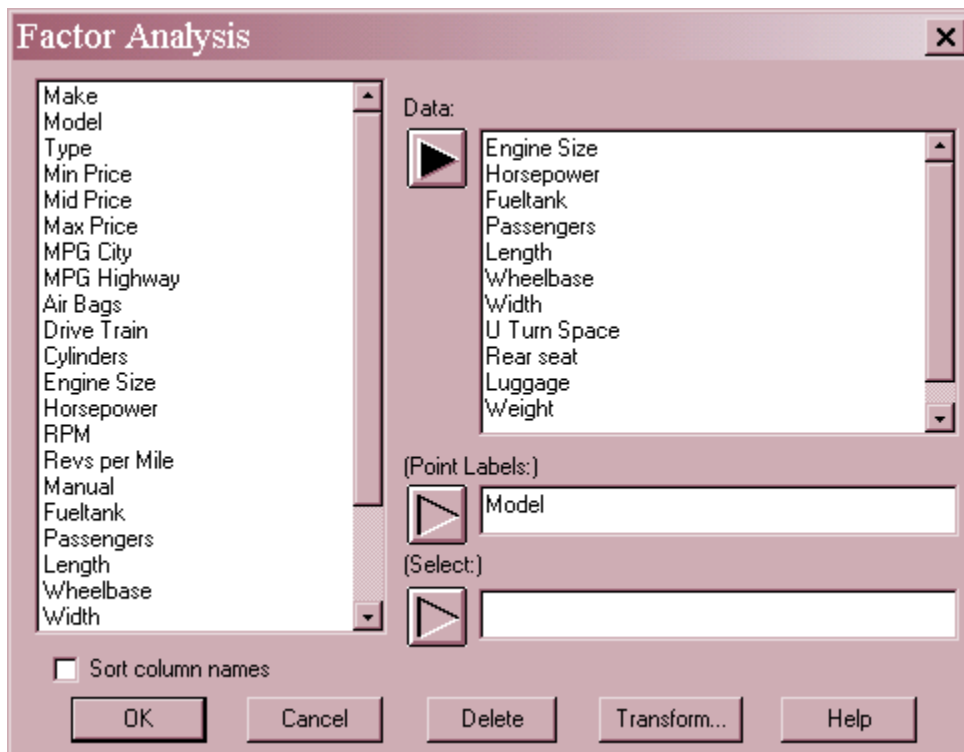
A matrix plot of the data is shown below:



As might be expected, the variables are highly correlated, since most are related to vehicle size.

## Data Input

The data input dialog box requests the names of the columns containing the data:



- **Data:** either the original observations or the sample covariance matrix  $\hat{\Sigma}$ . If entering the original observations, enter  $p$  numeric columns containing the  $n$  values for each column of  $X$ .

If entering the sample covariance matrix, enter  $p$  numeric columns containing the  $p$  values for each column of  $\hat{\Sigma}$ . If the covariance matrix is entered, some of the tables and plots will not be available.

- **Point Labels:** optional labels for each observation.
- **Select:** subset selection.

## Statistical Model

The goal of a factor analysis is to characterize the  $p$  variables in  $X$  in terms of a small number  $m$  of *common factors*  $F$ , which impact all of the variables, and a set of *errors* or *specific factors*  $\varepsilon$ , which affect only a single  $X$  variable. Following Johnson and Wichern (2002), the orthogonal common factor model expresses the observed variables as

$$\begin{aligned} X_1 - \mu_1 &= l_{11}F_1 + l_{12}F_2 + \dots + l_{1m}F_m + \varepsilon_1 \\ X_2 - \mu_2 &= l_{21}F_1 + l_{22}F_2 + \dots + l_{2m}F_m + \varepsilon_2 \\ &\dots \\ X_p - \mu_p &= l_{p1}F_1 + l_{p2}F_2 + \dots + l_{pm}F_m + \varepsilon_p \end{aligned} \quad (1)$$

In matrix notation,

$$X - \mu = LF + \varepsilon \quad (2)$$

where  $\mu$  is a vector of means and  $L$  is called the *factor loading* matrix. It is assumed that the common factors and specific factors are all independent of one another. To avoid ambiguity in scaling, the variances of the common factors are assumed to equal 1, while the covariance matrix of the specific factors  $\Psi$  is a diagonal matrix with diagonal elements  $\Psi_j$ . The covariance matrix  $\Sigma$  of the original observations  $X$  is related to the factor loading matrix by

$$\Sigma = LL' + \Psi \quad (3)$$

An important result of the above model is the relationship between the variances of the original  $X$  variables and the variances of the derived factors. In particular,

$$\text{Var}(X_j) = l_{j1}^2 + l_{j2}^2 + \dots + l_{jm}^2 + \Psi_j \quad (4)$$

This variance is expressed as the sum of two quantities:

1. the *communality*:  $l_{j1}^2 + l_{j2}^2 + \dots + l_{jm}^2$
2. the *specific variance*:  $\Psi_j$

The communality is the variance attributable to factors that all the  $X$  variables have in common, while the specific variance is specific to a single factor.

It should also be noted that the factor loadings  $L$  are not unique. Multiplication by any orthogonal matrix yields another acceptable set of factor loadings. Subsequent to the initial factor

extraction, it is therefore common to rotate the factor loadings until they can be most easily interpreted.

## Analysis Summary

The *Analysis Summary* table is shown below:

### Factor Analysis

Data variables:

Engine Size  
Horsepower  
Fuel tank  
Passengers  
Length  
Wheelbase  
Width  
U Turn Space  
Rear seat  
Luggage  
Weight

Data input: observations

Number of complete cases: 82

Missing value treatment: listwise

Standardized: yes

Type of factoring: principal components

Number of factors extracted: 2

### **Factor Analysis**

<i>Factor</i>		<i>Percent of</i>	<i>Cumulative</i>
<i>Number</i>	<i>Eigenvalue</i>	<i>Variance</i>	<i>Percentage</i>
1	7.92395	72.036	72.036
2	1.32354	12.032	84.068
3	0.47071	4.279	88.347
4	0.353248	3.211	91.559
5	0.269048	2.446	94.004
6	0.190242	1.729	95.734
7	0.172892	1.572	97.306
8	0.107148	0.974	98.280
9	0.0824071	0.749	99.029
10	0.0694689	0.632	99.660
11	0.0373497	0.340	100.000

	<i>Initial</i>
<i>Variable</i>	<i>Communality</i>
Engine Size	1.0
Horsepower	1.0
Fuel tank	1.0
Passengers	1.0
Length	1.0
Wheelbase	1.0
Width	1.0
U Turn Space	1.0
Rear seat	1.0
Luggage	1.0
Weight	1.0

Displayed in the table are:

- **Data variables:** the names of the  $p$  input columns.

- **Data input:** either *observations* or *matrix*, depending upon whether the input columns contain the original observations or the sample covariance matrix.
- **Number of complete cases:** the number of cases  $n$  for which none of the observations were missing.
- **Missing value treatment:** how missing values were treated in estimating the covariance or correlation matrix. If *listwise*, the estimates were based on complete cases only. If *pairwise*, all pairs of non-missing data values were used to obtain the estimates.
- **Standardized:** *yes* if the analysis was based on the correlation matrix. *No* if it was based on the covariance matrix.
- **Type of factoring:** either *principal components* if the factor extraction was done directly on the sample covariance or correlation matrix, or *classical* if the diagonal elements were adjusted using estimates of the communalities.
- **Number of components extracted:** the number of components  $m$  extracted from the data. This number is based on the settings on the *Analysis Options* dialog box.

A table is also displayed showing information for each of the  $p$  possible factors:

- **Factor number:** the factor number  $j$ , from 1 to  $p$ .
- **Eigenvalue:** the eigenvalue of the estimated covariance or correlation matrix,  $\hat{\lambda}_j$ , after adjusting for the estimated communalities if using the *classical* method.
- **Percentage of variance:** the percent of the total estimated population variance represented by this factor, equal to

$$100 \left( \frac{\hat{\lambda}_j}{\hat{\lambda}_1 + \hat{\lambda}_2 + \dots + \hat{\lambda}_m} \right) \% \quad (5)$$

- **Cumulative percentage:** the cumulative percentage of the total estimated population variance accounted for by the first  $j$  factors.
- **Initial communality:** the initial communality used in the calculations, either input by the user or estimated from the sample covariances or correlations.

In the example, the first  $m = 2$  factors account for over 84% of the overall variance amongst the 11 variables.

## Analysis Options

**Factor Analysis Options**

Missing Value Treatment

☒ Listwise  
☐ Pairwise

☒ Standardize

Type of Factoring

☒ Principal Components  
☐ Classical

Rotation

☐ None  
☒ Varimax  
☐ Equimax  
☐ Quartimax

Extract by

☒ Minimum Eigenvalue  
☐ Number of Factors

Minimum Eigenvalue: 1.

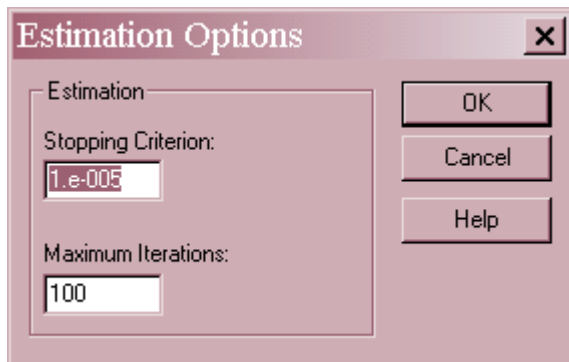
Number of Factors: 11

OK  
Cancel  
Estimation...  
Communalities...  
Help

- **Missing Values Treatment:** method of handling missing values when estimating the sample covariances or correlations. Specify *Listwise* to use only cases with no missing values for any of the input variables. Specify *Pairwise* to use all pairs of observations in which neither value is missing.
- **Standardize:** Check this box to base the analysis on the sample correlation matrix rather than the sample covariance matrix. This corresponds to standardizing each input variable before calculating the covariances, by subtracting its mean and dividing by its standard deviation.
- **Type of Factoring:** Select *principal components* to extract the factors directly from the covariance or correlation matrix. Select *classical* to replace the diagonal elements with estimated communalities. If using the *classical* method, you can specify the communalities by pressing the *Communalities* button or let the program use an iterative procedure to estimate them.
- **Rotation:** the method used to rotate the factor loading matrix after it has been extracted. *Varimax* rotation maximizes the variance of the squared loadings in each column. *Quartimax* maximizes the variance of the squared loadings in each row. *Equimax* attempts to achieve a balance between rows and columns.
- **Extract By:** the criterion used to determine the number of factors to extract.
- **Minimum Eigenvalue:** if extracting by magnitude of the eigenvalues, the smallest eigenvalue for which a factor will be extracted.
- **Number of Factors:** if extracting by number of factors, the number  $k$ .

There are also two buttons that access additional dialog boxes:

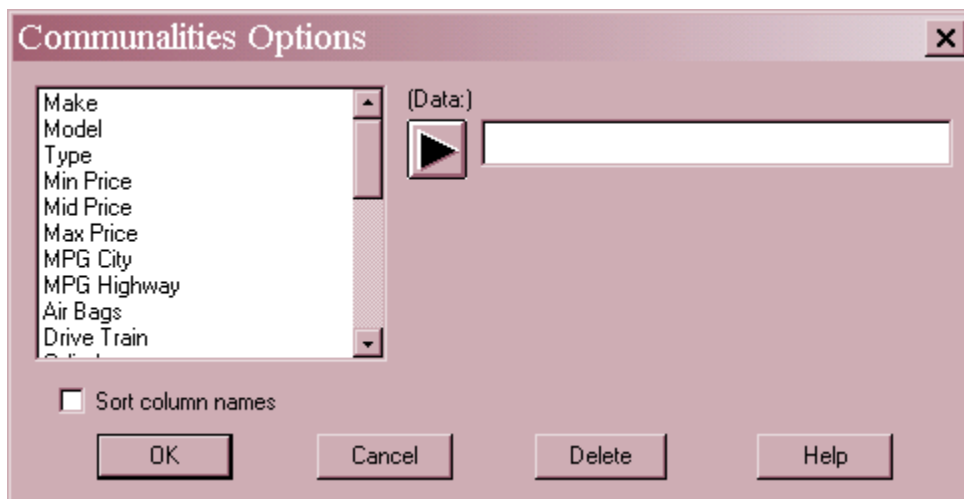
### Estimation Button



These fields control the iterations used in:

1. The *classical* method of factor extraction. Estimated communalities are revised until the proportional change in their sum is less than the *stopping criterion*, or the *maximum iterations* is reached.
2. *Rotation* of the factor loadings. The stopping criteria apply to the variance of the squared elements on the diagonal of the factor loading matrix.

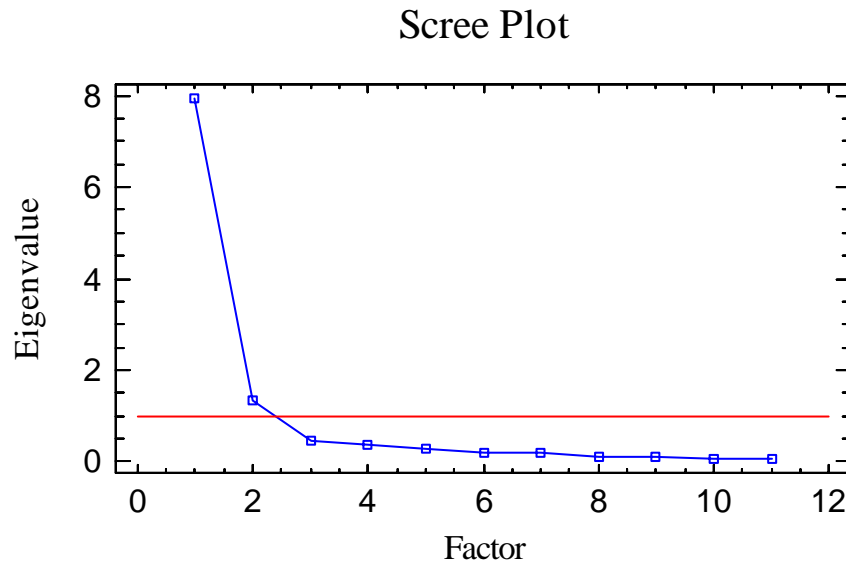
### Communalities Button



When using the *classical* estimation method, you may specify a column containing the communalities instead of having the program estimate them iteratively.

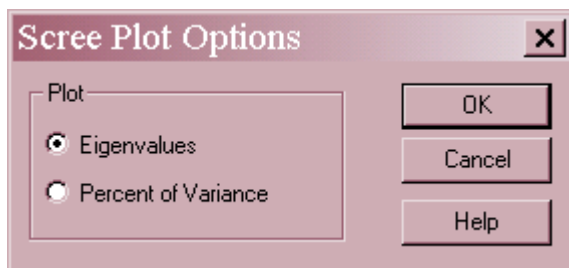
## Scree Plot

The *Scree Plot* can be very helpful in determining the number of factors to extract. By default, it plots the size of the eigenvalues corresponding to each of the  $p$  possible factors:



A horizontal line is added at the minimum eigenvalue specified on the *Analysis Options* dialog box. In the plot above, note that only the first 2 factors have large eigenvalues.

### Pane Options



- **Plot:** value plotted on the vertical axis.



## Extraction Statistics

The *Extraction Statistics* pane shows the estimated value of the coefficients  $l$  for each factor extracted, before any rotation is applied:

Factor Loading Matrix Before Rotation		
	<i>Factor</i>	<i>Factor</i>
	<i>1</i>	<i>2</i>
Engine Size	0.936606	-0.154035
Horsepower	0.754754	-0.50948
Fuel tank	0.876138	-0.241737
Passengers	0.671882	0.610074
Length	0.944075	0.0244126
Wheelbase	0.944096	0.0702147
Width	0.914567	-0.154446
U Turn Space	0.842284	-0.0955416
Rear seat	0.650975	0.613778
Luggage	0.778316	0.371338
Weight	0.948687	-0.237682

	<i>Estimated</i>	<i>Specific</i>
<i>Variable</i>	<i>Communality</i>	<i>Variance</i>
Engine Size	0.900958	0.0990419
Horsepower	0.829223	0.170777
Fuel tank	0.826054	0.173946
Passengers	0.823616	0.176384
Length	0.891874	0.108126
Wheelbase	0.896247	0.103753
Width	0.860287	0.139713
U Turn Space	0.71857	0.28143
Rear seat	0.800491	0.199509
Luggage	0.743667	0.256333
Weight	0.9565	0.0435005

Also displayed are the communalities and the specific variances. The weights within each column often have interesting interpretations. In the example, note that the weights in the first column are all approximately the same. This implies that the first component is basically an average of all of the input variables. The second component is weighted most heavily in a positive direction on the number of *Passengers*, the *Rear Seat* room, and the amount of *Luggage* space, and in a negative direction on *Horsepower*. It likely differentiates amongst the different types of vehicles. Note also that *U Turn Space* and *Luggage* have a larger specific variance than the others, implying that they are not as well accounted for by the two extracted factors.

## Rotation Statistics

The *Rotation Statistics* pane shows the estimated value of the coefficients  $l$  after the requested rotation is applied:

Factor Loading Matrix After Varimax Rotation		
	<i>Factor</i>	<i>Factor</i>
	<i>1</i>	<i>2</i>
Engine Size	0.859769	0.402188
Horsepower	0.910596	0.00617243
Fuel tank	0.859441	0.295661
Passengers	0.209571	0.883004
Length	0.765091	0.553632
Wheelbase	0.739226	0.591432
Width	0.841818	0.389395
U Turn Space	0.748896	0.397145
Rear seat	0.190229	0.874245
Luggage	0.43229	0.746186
Weight	0.917004	0.340004

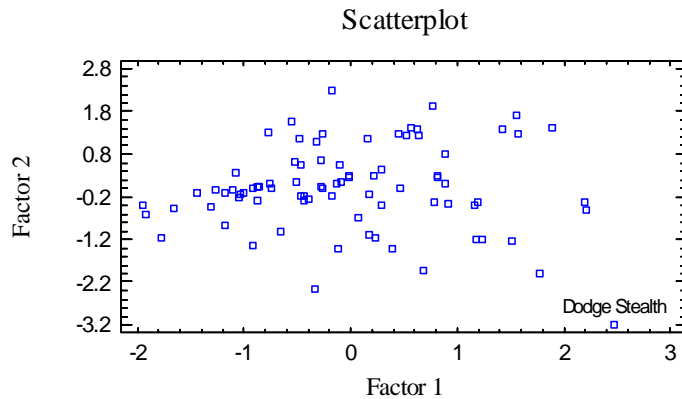
  

	<i>Estimated</i>	<i>Specific</i>
<i>Variable</i>	<i>Communality</i>	<i>Variance</i>
Engine Size	0.900958	0.0990419
Horsepower	0.829223	0.170777
Fuel tank	0.826054	0.173946
Passengers	0.823616	0.176384
Length	0.891874	0.108126
Wheelbase	0.896247	0.103753
Width	0.860287	0.139713
U Turn Space	0.71857	0.28143
Rear seat	0.800491	0.199509
Luggage	0.743667	0.256333
Weight	0.9565	0.0435005

Note that the rotation has substantially decreased the loading of *Passengers*, *Rear seat*, and *Luggage* on the first factor and made them the dominant variables in the second factor. The second factor seems to distinguish large family vehicles such as minivans and SUV's from the other automobiles.

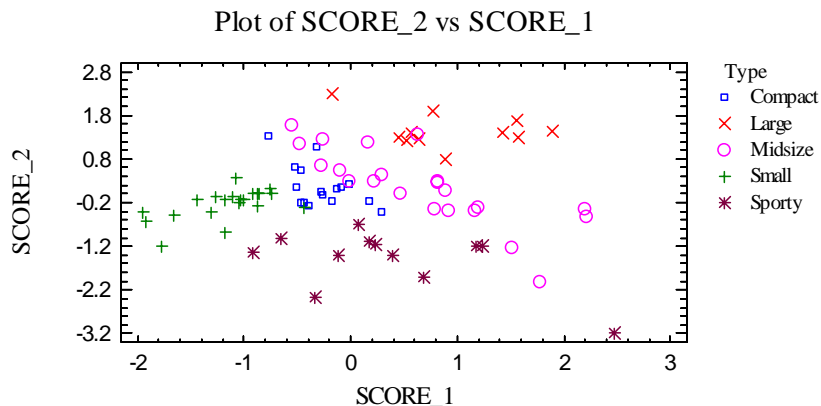
## 2D and 3D Scatterplots

These plots display the values of 2 or 3 selected factors for each of the  $n$  cases, after rotation.



It is useful to examine any points far away from the others, such as the highlighted *Dodge Stealth*, which has a very low value for the second factor.

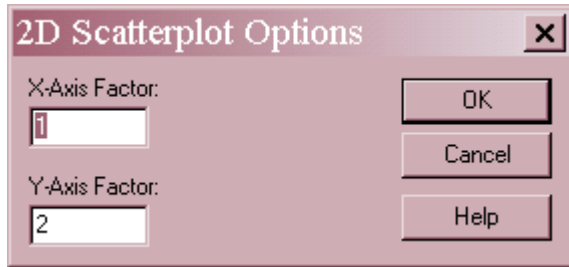
An interesting variation of this plot is one in which the variables are coded according to another column, such as the type of vehicle:



To produce the above plot:

1. Press the *Save Results* button and save the *Factor Scores* to new columns of the datasheet.
2. Select the *X-Y Plot* procedure from the top menu and input the new columns.
3. Select *Analysis Options* and specify *Type* in the *Point Codes* field.

It is now clear that the first factor is related to the size of the vehicle, while the second factor separates the sporty cars from the others.

*Pane Options*


The dialog box titled "2D Scatterplot Options" has a close button (X) in the top right corner. It contains two input fields: "X-Axis Factor:" with a dropdown menu showing "1" and "Y-Axis Factor:" with a dropdown menu showing "2". To the right of these fields are three buttons: "OK", "Cancel", and "Help".

Specify the factors to plot on each axis.

**Factor Scores**

The *Factor Scores* pane displays the values of the rotated factor scores for each of the  $n$  cases.

Table of Factor Scores			
		<i>Factor</i>	<i>Factor</i>
<i>Row</i>	<i>Label</i>	<i>1</i>	<i>2</i>
1	Integra	-0.440603	-0.294691
2	Legend	0.817275	0.299261
3	90	0.177176	-0.154546
4	100	0.155524	1.17616
5	535i	1.5048	-1.23631
6	Century	-0.474803	1.14786
7	LeSabre	0.63412	1.25438
8	Roadmaster	1.88652	1.43271
9	Riviera	1.18707	-0.321997
...	...	...	...

The factor scores show where each observation falls with respect to the extracted factors.

**Factor Score Coefficients**

The table of *Factor Score Coefficients* shows the coefficients used to create the factor scores from the original variables.

Factor Score Coefficients		
	<i>Factor</i>	<i>Factor</i>
	<i>1</i>	<i>2</i>
Engine Size	0.163284	0.29611
Horsepower	-0.0292234	-0.263759
Fuel tank	19.7073	14.343
Passengers	4.48584	10.7923
Length	39.5473	46.3067
Wheelbase	8.18626	26.1997
Width	-59.5975	-13.7139
U Turn Space	3.83938	15.7456
Rear seat	-17.5316	-16.1991
Luggage	-6.00197	19.3445
Weight	-114.779	-181.049

If the sample covariance matrix  $S$  has been factored, then the coefficients are the leading terms multiplying the deviation of each variable from its mean in

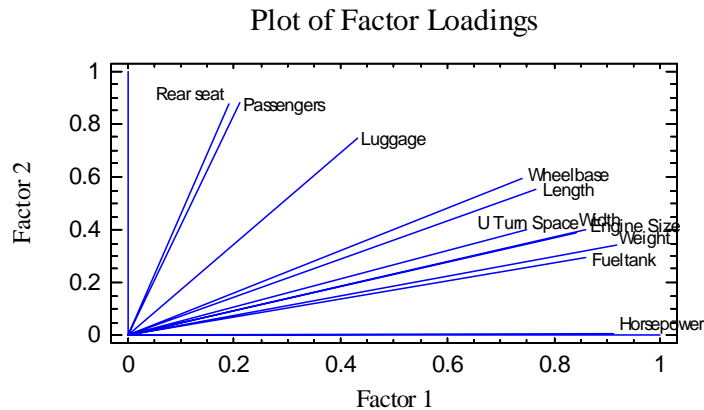
$$\hat{f}_j = \hat{L}'S^{-1}(x_j - \bar{x}) \quad (6)$$

If the correlation matrix  $R$  has been factored, then the coefficients are the leading terms multiplying the standardized values of each variable according in

$$\hat{f}_j = \hat{L}'R^{-1}z_j \quad (7)$$

## 2D and 3D Factor Plots

The *Factor Plots* show the location of each variable in the space of 2 or 3 selected factors:



Variables furthest from the reference lines at 0 make the largest contribution to the factors.

## Save Results

The following results may be saved to the datasheet:

1. *Eigenvalues* – the  $m$  eigenvalues.
2. *Factor Matrix* –  $m$  columns, each containing  $p$  estimates of the coefficients  $l$  before rotation.
3. *Rotated Factor Matrix*–  $m$  columns, each containing  $p$  estimates of the coefficients  $l$  after rotation.
4. *Transition Matrix* – the  $m$  by  $m$  matrix that multiplies the original factor loadings to yield the rotated factor loadings.
5. *Communalities* – the  $p$  estimated communalities after rotation.
6. *Specific variances* – the  $p$  specific variances after rotation.

7. *Factor Scores* –  $m$  columns, each containing  $n$  values corresponding to the extracted factors.
8. *Factor Score Coefficients* –  $m$  columns, each containing the  $p$  values of the factor score coefficients.