Canonical Correlations

Summary

The **Canonical Correlations** procedure is designed to help identify associations between two sets of variables. It does so by finding linear combinations of the variables in the two sets that exhibit strong correlations. The pair of linear combinations with the strongest correlation forms the first set of *canonical variables*. The second set of canonical variables is the pair of linear combinations that show the next strongest correlation amongst all combinations that are uncorrelated with the first set. Often, a small number of pairs can be used to quantify the relationships that exist between the two sets.

Sample StatFolio: *canonical.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*	*Mid Price*	*MPG City*	*Engine Size*	*Horsepower*	*Length*	*Weight*
Acura	Integra	15.9	25	1.8	140	177	2705
Acura	Legend	33.9	18	3.2	200	195	3560
Audi	90	29.1	20	2.8	172	180	3375
Audi	100	37.7	19	2.8	172	193	3405
BMW	535i	30	22	3.5	208	186	3640
Buick	Century	15.7	22	2.2	110	189	2880
Buick	LeSabre	20.8	19	3.8	170	200	3470
Buick	Roadmaster	23.7	16	5.7	180	216	4105
Buick	Riviera	26.3	19	3.8	170	198	3495
Cadillac	DeVille	34.7	16	4.9	200	206	3620
Cadillac	Seville	40.1	16	4.6	295	204	3935
Chevrolet	Cavalier	13.4	25	2.2	110	182	2490

The variables will be divided into two sets. The first set of $p = 7$ variables characterizes the physical characteristics of the vehicles:

Engine Size
Horsepower
Length

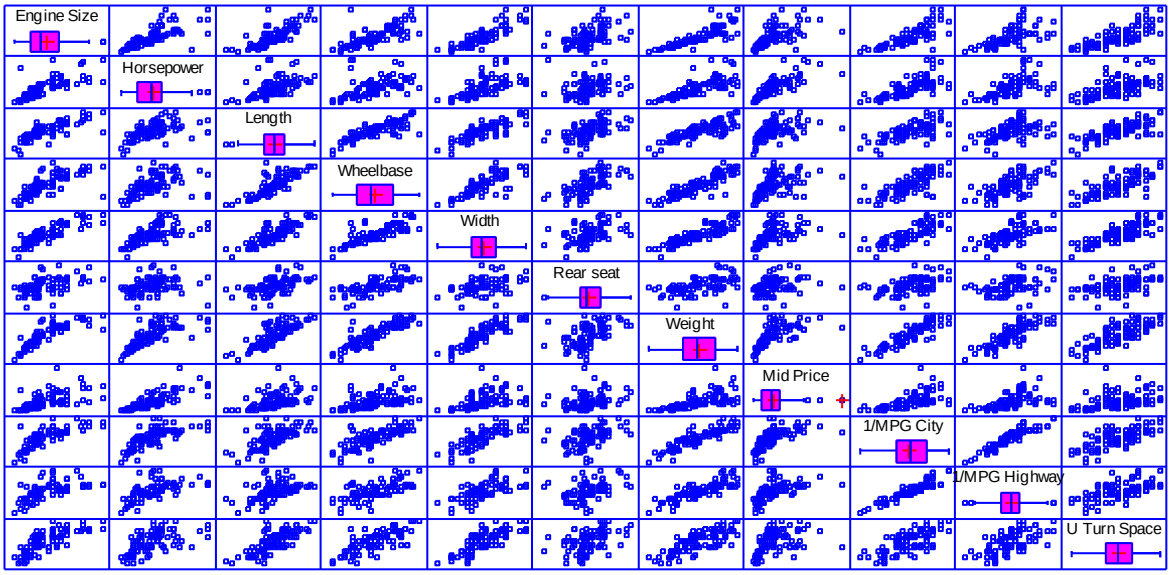
Wheelbase
Width
Rear seat
Weight

The second set of $q = 4$ variables characterizes the price and performance of the automobiles:

Mid Price
1 / MPG Highway
1 / MPG City
U Turn Space

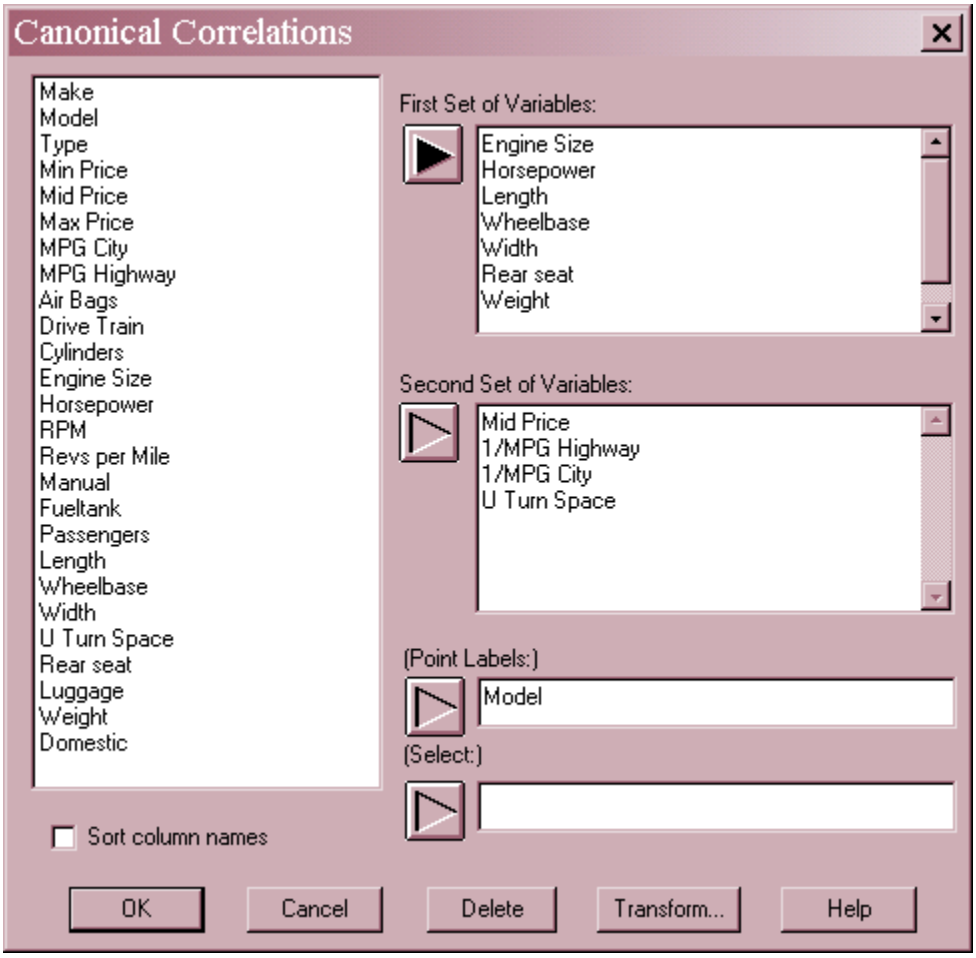
Note that the observed miles per gallon has been reexpressed as gallons per mile, so that all four variables can be expected to increase with the size of the vehicle.

A matrix plot of the 11 variables shows consistent positive correlations amongst all of the variables:



Data Input

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



- **First Set of Variables:** the names of the *p* variables in the larger set.
- **Second Set of Variables:** the names of the *q* variables in the smaller set.
- **Point Labels**: optional labels for each observation.
- **Select:** subset selection.

Note that the sets must be chosen so that $p \geq q$.

Statistical Model

The goal of the canonical correlation procedure is to construct linear combinations of the variables in the two sets that have the strongest correlations. The first set of canonical variables takes the form

$$U_1 = a_{11}X_1 + a_{12}X_2 + ... + a_{1p}X_p \quad (1)$$

$$V_1 = b_{11}Y_1 + b_{12}Y_2 + ... + b_{1q}Y_q \quad (2)$$

where X and Y represent the standardized values of the variables in the first and second set, respectively. The correlation between the first set of linear combinations is called the *first canonical correlation* and will be denoted by ρ_1^*.

An additional $q - 1$ canonical variables can be constructed in a similar manner. The q canonical correlations are found by determining the eigenvalues of

$$\rho_{11}^{-1/2}\rho_{12}\rho_{22}^{-1}\rho_{21}\rho_{11}^{-1/2} \quad (3)$$

where the ρ's represent the correlation matrices amongst variables in each set and between sets. The coefficients of the canonical variables are derived from the eigenvectors.

Analysis Summary

The *Analysis Summary* table is shown below:

Canonical Correlations

Variables in set 1:
- Engine Size
- Horsepower
- Length
- Wheelbase
- Width
- Rear seat
- Weight

Variables in set 2:
- Mid Price
- 1/MPG Highway
- 1/MPG City
- U Turn Space

Number of complete cases: 91

Canonical Correlations

		Canonical	*Wilks*			
Number	*Eigenvalue*	*Correlation*	*Lambda*	*Chi-Squared*	*D.F.*	*P-Value*
1	0.895275	0.94619	0.0275328	301.76	28	0.0000
2	0.495819	0.704144	0.262906	112.22	18	0.0000
3	0.462885	0.680356	0.521453	54.6955	10	0.0000
4	0.0291608	0.170765	0.970839	2.48593	4	0.6472

Coefficients for Canonical Variables of the First Set

Engine Size	0.261726	0.698443	-0.0737052	2.04984
Horsepower	0.127466	0.404309	1.23884	-0.784463
Length	0.0241777	1.06291	0.279635	-0.0542533
Wheelbase	0.0411746	0.344853	0.710682	-1.45037
Width	-0.0676957	0.292913	-1.51189	-1.08908
Rear seat	0.00425793	-0.0929359	-0.0789944	-0.261572
Weight	0.657779	-2.42508	-0.470777	1.19131

Coefficients for Canonical Variables of the Second Set

Mid Price	0.256618	0.15463	1.21063	-0.401701
1/MPG Highway	-0.0971257	-2.20547	0.175652	-1.51504
1/MPG City	0.652062	1.42486	-0.796365	2.80861
U Turn Space	0.32219	0.454982	-0.340661	-1.33714

Displayed in the top section of the table are:

- **Data variables**: the names of the $p+q$ input columns.
- **Number of complete cases**: the number of cases n for which none of the observations were missing.

The section of the output labeled *Canonical Correlations* tabulates:

- **Number**: the index of the canonical correlation j.
- **Eigenvalue**: the eigenvalues of $\rho_{11}^{-1/2}\rho_{12}\rho_{22}^{-1}\rho_{21}\rho_{11}^{-1/2}$.
- **Canonical Correlation**: the canonical correlations ρ_j^*.

- **Wilk's Lambda**: a statistic calculated from the canonical correlations according to

$$\Lambda_j = \prod_{i=j}^{q} \left(1 - \rho_i^{*2}\right) \tag{4}$$

- **Chi-Squared**: a test statistic used to test the hypothesis that all canonical correlations numbered *j* and higher are equal to 0. It is calculated from

$$X^2 = -\left(n - 1 - \frac{1}{2}(p+q+1)\right) \ln \Lambda_j \tag{5}$$

- **D.F.**: the degrees of freedom *(p-j+1)(q-j+1)* associated with the chi-squared statistic.

- **P-Value**: a one-sided P-Value for the observed chi-squared statistic. Small P-values (less than 0.05 if operating at the 5% significance level) correspond to canonical correlations that are significantly different from zero.

The last two tables show the coefficients *a* and *b* in the construction of the canonical variables *U* and *V*.

In the example, the first 3 canonical correlations are statistically significant. The first correlation, with a magnitude of 0.94, is particularly strong. The associated canonical correlations are

U_1 = 0.262 *Engine Size* + 0.127 *Horsepower* + 0.024 *Length* + 0.041 *Wheelbase* - 0.068 *Width* + 0.004 *Rear Seat* + 0.658 *Weight*

V_1 = 0.257 *Mid Price* – 0.097 * *GPM Highway* + 0.652 *GPM City* + 0.322 *U Turn Space*

where the variables are understood to have been standardized by subtracting their sample means and dividing by their sample standard deviations. This appears to be primarily a relationship between vehicle weight and gallons per mile (GPM) used in city driving, with some contribution of engine size, price, and space required to make a U turn.

The second canonical correlation is also fairly strong, at 0.70. The canonical variable for *X* is a contrast of *Engine Size*, *Horsepower* and *Length* against vehicle *Weight*. This contrast is correlated with something resembling the difference between the miles per gallon observed in city driving versus highway driving.

The third canonical correlation equals 0.68. The canonical variable for *X* is a contrast of *Horsepower* and *Wheelbase* against vehicle *Width*. This is correlated with something resembling a contrast of vehicle *Price* versus *GPM* in city driving and *U Turn Space*.

Data Table

The *Data Table* pane displays the values of the canonical variables for each of the *n* observations. A portion of that table is shown below:

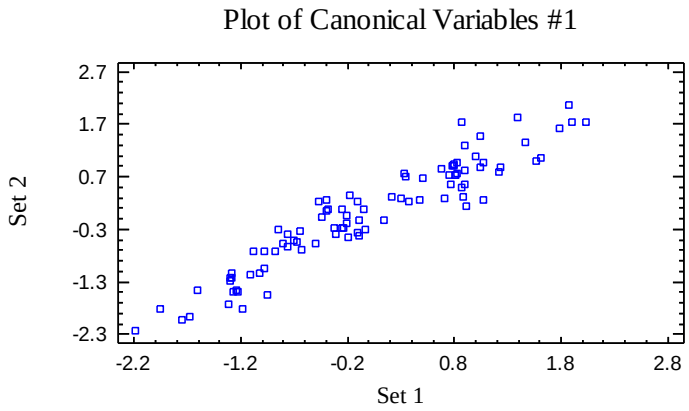
Table of Canonical Variables

		Set-Variable	*Set-Variable*	*Set-Variable*	*Set-Variable*	*Set-Variable*
Row	*Label*	*1-1*	*2-1*	*1-2*	*2-2*	*1-3*
1	Integra	-0.633815	-0.68972	0.252376	-0.191381	0.552317
2	Legend	0.89817	0.829715	0.341564	-0.320696	1.70703
3	90	0.479834	0.246904	-1.41249	-0.794373	1.17577
4	100	0.509327	0.659087	-0.251631	-0.2627	0.541995
5	535i	1.06715	0.245875	-0.723874	0.732446	1.90213
6	Century	-0.390542	0.0756147	0.625743	1.1806	-0.272698
7	…	…	…	…	…	…

The order of the columns is U_1, V_1, U_2, V_2, …, U_q, V_q.

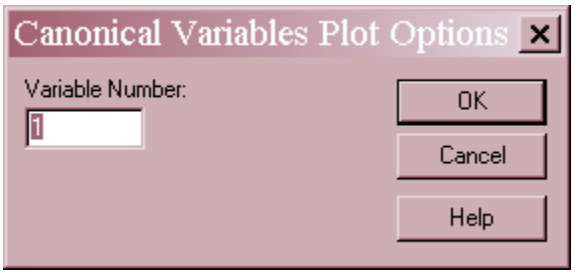
Canonical Variables Plot

The *Canonical Variables* plot displays the n values of a selected set of canonical variables



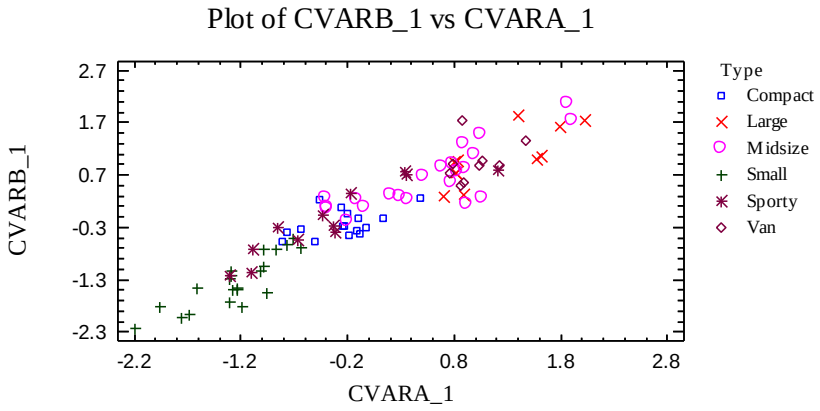
U is plotted on the horizontal axis, while *V* is plotted on the vertical axis. Note the very strong correlation for the first canonical variables.

Pane Options



- **Variable Number**: specify *1* to plot the first set of canonical variables, *2* to plot the second set, up to *q* for the last set.

An interesting variation of this plot is one in which the canonical 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 *Canonical Variables* to new columns of the datasheet.
2. Select the *X-Y Plot* procedure from the top menu and input the first canonical variable from each set.
3. Select *Analysis Options* and specify *Type* in the *Point Codes* field.

Note the grouping of the automobiles by type.

Save Results

The following results may be saved to the datasheet:

1. *Coefficients – First Set – q* columns containing the *p* coefficients *a* of the canonical variables corresponding to *X*.

2. *Coefficients – Second Set – q* columns containing the *q* coefficients *b* of the canonical variables corresponding to *Y*.

3. *Canonical Variables – First Set – q* columns containing the values of the canonical variables *U* corresponding to each of the *n* observations in *X*.

4. *Canonical Variables – Second Set – q* columns containing the values of the canonical variables *V* corresponding to each of the *n* observations in *Y*.

