Regression Model Selection

Summary

The **Regression Model Selection** procedure is designed help select the independent variables to use in building a multiple regression model to predict a single quantitative dependent variable Y. The procedure considers all possible regressions involving different combinations of the independent variables. It compares models based on the adjusted R-Squared, Mallows' Cp statistic, and the mean squared error.

Sample StatFolio: *select reg.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 several columns from that file:

Make	*Model*	*MPG Highway*	*Weight*	*Horsepower*	*Wheelbase*	*Passengers*
Acura	Integra	31	2705	140	102	5
Acura	Legend	25	3560	200	115	5
Audi	90	26	3375	172	102	5
Audi	100	26	3405	172	106	6
BMW	535i	30	3640	208	109	4
Buick	Century	31	2880	110	105	6
Buick	LeSabre	28	3470	170	111	6
Buick	Roadmaster	25	4105	180	116	6
Buick	Riviera	27	3495	170	108	5
Cadillac	DeVille	25	3620	200	114	6
Cadillac	Seville	25	3935	295	111	5
Chevrolet	Cavalier	36	2490	110	101	5

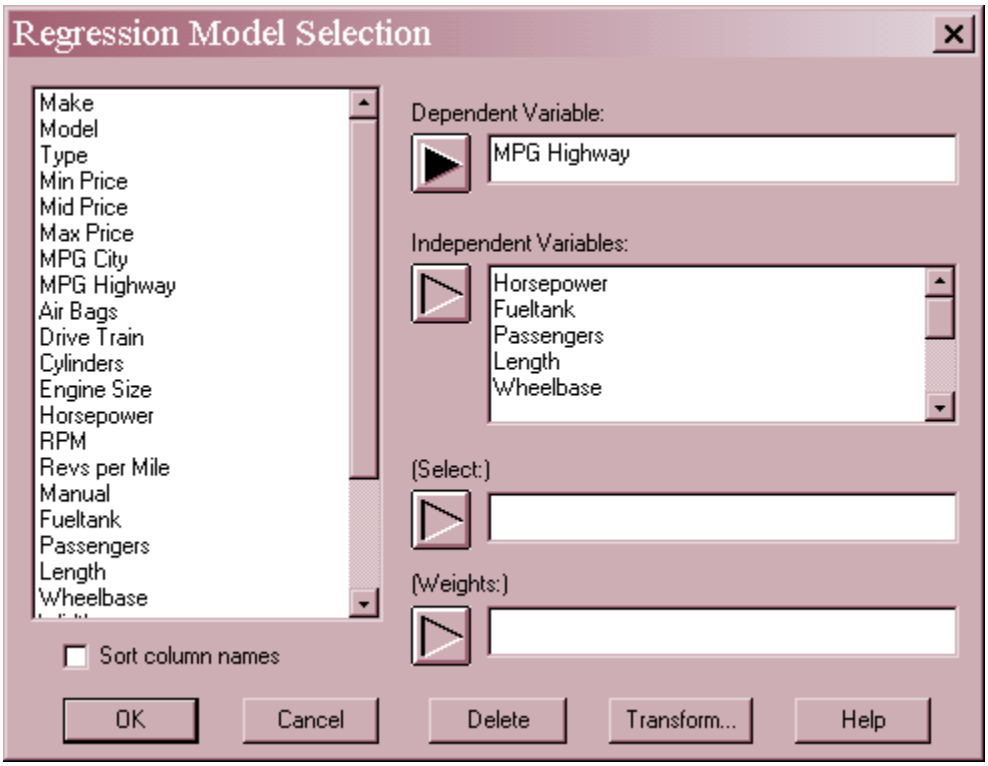
It is desired to construct a model to predict *MPG Highway*. 10 possible predictor variables will be considered:

1. *Horsepower*
2. *Fueltank*
3. *Passengers*
4. *Length*
5. *Wheelbase*
6. *Width*
7. *U-Turn Space*
8. *Rear seat*
9. *Luggage*
10. *Weight*

A parsimonious model is desired, i.e., a model involving as few variables as possible, provided none of the omitted variables would add significantly to the predictive capability of the model.

Data Input

The data input dialog box requests the name of the dependent variable Y and all candidate independent variables:



- **Dependent Variable:** numeric column containing the *n* observations for the dependent variable Y.

- **Independent Variables:** numeric columns containing the *n* values for all independent variables X to be considered for use in the model. This field scrolls if necessary.

- **Select:** subset selection.

- **Weight:** an optional numeric column containing weights to be applied to the squared residuals when performing a weighted least squares fit.

Analysis Summary

The *Analysis Summary* displays information about the input data and the fitted models. The top section of the output indicates the number of observations *n* and the number of fitted models.

Regression Model Selection - MPG Highway
Dependent variable: MPG Highway
Independent variables:
A=Horsepower
B=Fueltank
C=Passengers
D=Length
E=Wheelbase
F=Width
G=U Turn Space
H=Rear seat
I=Luggage
J=Weight

Number of complete cases: 82
Number of models fit: 1024

Models are fit involving all combinations of variables up to the number specified on the *Analysis Options* dialog box.

The second section of the output shows a summary of all fitted models. A portion of that table is shown below.

Model Results

		Adjusted		*Included*
MSE	*R-Squared*	*R-Squared*	*Cp*	*Variables*
25.1105	0.0	0.0	177.237	
13.4123	47.2463	46.5868	57.7023	A
11.5449	54.5913	54.0237	38.8081	B
21.0657	17.1435	16.1078	135.138	C
15.738	38.0989	37.3252	81.2326	D
17.1595	32.5077	31.664	95.6153	E
16.6421	34.5426	33.7244	90.3808	F
17.4867	31.2207	30.3609	98.9261	G
23.4678	7.69584	6.54204	159.441	H
21.9131	13.8108	12.7335	143.711	I
10.0846	60.3349	59.8391	24.0333	J
10.9961	57.2905	56.2092	33.8648	AB
11.8965	53.793	52.6232	42.8616	AC
12.2353	52.4773	51.2742	46.246	AD
12.9804	49.5831	48.3067	53.6911	AE
12.941	49.7363	48.4638	53.297	AF
12.7091	50.6371	49.3874	50.9799	AG
13.3785	48.037	46.7215	57.6682	AH
13.1223	49.0322	47.7418	55.1083	AI
10.174	60.4834	59.483	25.6514	AJ
11.4782	55.4178	54.2891	38.6821	BC

Included in the table are:

- **MSE**: the mean squared error. This is an estimate of the variance of the deviations from the fitted model, given by:

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n - p - 1} \quad (1)$$

where y_i is the observed value of Y, $\hat{y}_i$ is the predicted value from the fitted model, and p equals the number of independent variables included in the model.

- **R-Squared:** the adjusted coefficient of determination, calculated from

$$R^2 = 100\left(1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}\right)\% \quad (2)$$

R-Squared measures the percentage of the variability in Y that has been explained by the fitted model.

- **Adjusted R-Squared:** the adjusted coefficient of determination, calculated from

$$R^2_{adj} = 100\left(1 - \left(\frac{n-1}{n-p-1}\right)\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}\right)\% \quad (3)$$

The adjusted R-Squared compensates for the number of independent variables in the model. It is more useful than the ordinary R-squared in comparing models with different numbers of independent variables, since the latter statistic will never go down even if unrelated variables are added to the model.

- **Cp:** Mallows' Cp statistic, calculated from

$$C_p = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{MSE(full)} - (n - 2p) \quad (4)$$

where *MSE(full)* is the mean squared error of the model when all independent variables are included in the fit. If a fitted model has little bias, *Cp* should be close to *p*. It is desirable to have a small *Cp*, as long as the value is not much greater than *p*.

- **Included Variables:** an indication of which independent variables are included in the model.

In the sample data, every possible combinations of variables was fit, from the simplest model, which includes only a constant term:

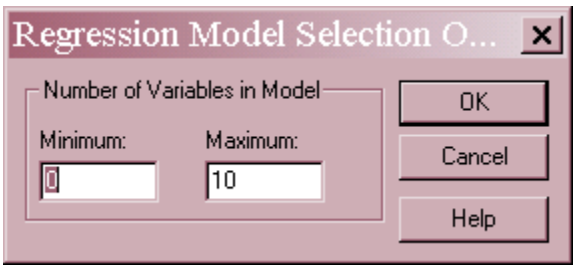
$$\text{MPG Highway} = \beta_0 \quad (5)$$

to the most complicated, which includes all 10 candidate variables:

$$MPG\ Highway = \beta_0 + \beta_1 Horsepower + \beta_2 Fueltank + \beta_3 Passengers + \beta_4 Length + \beta_5 Wheelbase + \beta_6 Width + \beta_7 U\ Turn\ Space + \beta_8 Rear\ Seat + \beta_9 Luggage + \beta_{10} Weight \quad (6)$$

This represents a total of 1,024 models.

Analysis Options



All possible regression models will be fit containing at least the *Minimum* number of independent variables but no more than the *Maximum*.

Best Adjusted R-Squared

This table summarizes the fitted models, sorted in decreasing order of the adjusted R-squared statistic:

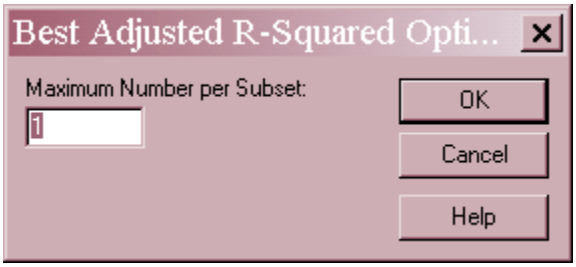
Models with Largest Adjusted R-Squared

		Adjusted		*Included*
MSE	*R-Squared*	*R-Squared*	*Cp*	*Variables*
7.67032	72.0935	69.4537	5.78582	BCEFHIJ
7.69476	72.3829	69.3564	7.04138	ABCEFHIJ
7.7374	71.4691	69.1866	5.39211	BCEFIJ
7.75017	71.041	69.1358	4.49343	BCEIJ
7.79712	72.3989	68.9488	9.00031	ABCDEFHIJ
7.90691	72.399	68.5116	11.0	ABCDEFGHIJ
8.01411	69.6607	68.0846	6.04399	CEIJ
8.26017	68.3231	67.1047	7.4849	CEJ
8.7836	65.8839	65.0202	11.7592	EJ
10.0846	60.3349	59.8391	24.0333	J
25.1105	0.0	0.0	177.237	

By default, the table shows the best model for each number of independent variables. For example, the best model involving only 3 independent variables includes variables C, E, and J, and gives an adjusted R-squared of 67.1%.

The model with the best adjusted R-squared includes 7 variables, BCEFHIJ.

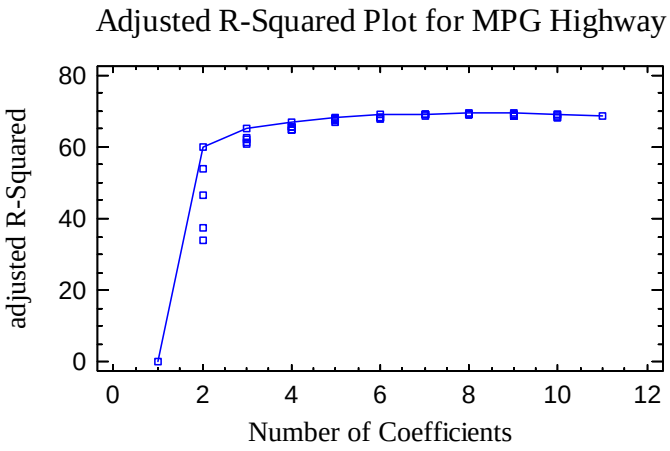
Pane Options



- **Maximum Number per Subset:** the maximum number of models included in the table that contain the indicated number of independent variables.

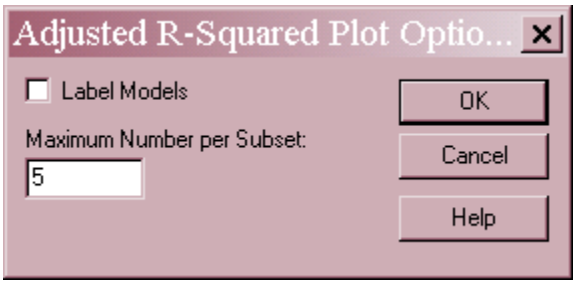
Adjusted R-Squared Plot

This plot shows the models with the highest adjusted R-Squared values.



The line connects the models with the best adjusted R-squared values for each number of coefficients. Notice that the best adjusted R-squared increases noticeably until the number of coefficients equals 6 (corresponding to 5 independent variables). Referring back to the table above, the best model with 5 variables is BCEIJ, which has an adjusted R-squared = 69.1%.

Pane Options



- **Label Models:** if selected, model labels will be added to the plot.

- **Maximum Number per Subset:** the maximum number of models included in the plot that contain the same number of independent variables.

Best Cp

This table shows the models, sorted in increasing order of the Cp statistic:

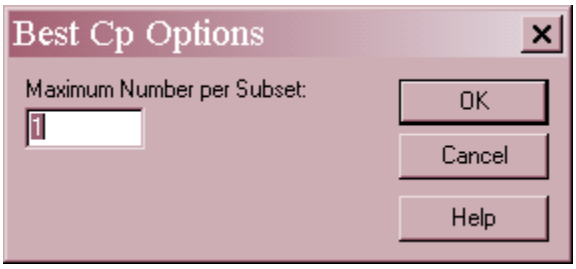
Models with Smallest Cp

		Adjusted		*Included*
MSE	*R-Squared*	*R-Squared*	*Cp*	*Variables*
7.75017	71.041	69.1358	4.49343	BCEIJ
7.7374	71.4691	69.1866	5.39211	BCEFIJ
7.67032	72.0935	69.4537	5.78582	BCEFHIJ
8.01411	69.6607	68.0846	6.04399	CEIJ
7.69476	72.3829	69.3564	7.04138	ABCEFHIJ
8.26017	68.3231	67.1047	7.4849	CEJ
7.79712	72.3989	68.9488	9.00031	ABCDEFHIJ
7.90691	72.399	68.5116	11.0	ABCDEFGHIJ
8.7836	65.8839	65.0202	11.7592	EJ
10.0846	60.3349	59.8391	24.0333	J
25.1105	0.0	0.0	177.237	

By default, the table shows the best model for each number of independent variables. For example, the best model involving only 3 independent variables includes variables C, E, and J, and has a Cp equal to 7.48. Small Cp is desirable as long as it is less than the number of independent variables in the model.

The model with the smallest Cp is BCEIJ. Since its Cp value is less than 5, that model would appear to be the best.

Pane Options



- **Maximum Number per Subset:** the maximum number of models included in the table that contain the same number of independent variables.

Best Information Criteria

This table shows the models, sorted in increasing order of a selected information criterion:

Models with Best Information Criteria

MSE	*Coefficients*	*AIC*	*HQC*	*SBIC*	*Included Variables*
7.75017	6	2.19406	2.26476	2.37016	BCEIJ
8.01411	5	2.20316	2.26207	2.34991	CEIJ
8.26017	4	2.20901	2.25614	2.32641	CEJ
8.10874	5	2.21489	2.27381	2.36164	BCEJ
7.7374	7	2.2168	2.29928	2.42225	BCEFIJ
8.12583	5	2.217	2.27592	2.36375	CEFJ
7.94334	6	2.21867	2.28938	2.39478	CEFHJ
7.76198	7	2.21997	2.30246	2.42542	ABCEIJ
7.96728	6	2.22169	2.29239	2.39779	CEFIJ
7.77775	7	2.222	2.30449	2.42745	BCEHIJ
7.79739	7	2.22452	2.30701	2.42997	BCEFHJ
7.99798	6	2.22553	2.29623	2.40163	BCEFJ
7.82658	7	2.22826	2.31074	2.43371	BCEGIJ
8.22842	5	2.22955	2.28846	2.3763	CEHJ
8.04899	6	2.23189	2.30259	2.40799	CEHIJ
7.67032	8	2.23248	2.32675	2.46728	BCEFHIJ
8.28297	5	2.23615	2.29507	2.3829	CEGJ
7.73722	8	2.24116	2.33543	2.47597	ABCEFIJ
8.7836	3	2.24606	2.28141	2.33411	EJ
7.80764	8	2.25022	2.34449	2.48503	ABCEHIJ
7.8196	8	2.25176	2.34603	2.48656	ABCDEIJ
7.82246	8	2.25212	2.34639	2.48692	ABCEGIJ
8.65087	4	2.25522	2.30236	2.37262	EFJ
8.66666	4	2.25704	2.30418	2.37445	BEJ
7.69476	9	2.26005	2.36611	2.5242	ABCEFHIJ
7.77425	9	2.27033	2.37638	2.53448	BCEFGHIJ
7.77538	9	2.27047	2.37653	2.53463	BCDEFHIJ
8.81713	4	2.27426	2.32139	2.39166	EIJ
8.81806	4	2.27436	2.3215	2.39176	AEJ
7.83658	9	2.27831	2.38437	2.54247	ABCDEFIJ
7.83798	9	2.27849	2.38455	2.54265	ABCEFGIJ
7.79712	10	2.29766	2.41549	2.59116	ABCDEFHIJ
7.80155	10	2.29823	2.41606	2.59173	ABCEFGHIJ
7.8822	10	2.30851	2.42635	2.60201	BCDEFGHIJ
9.41106	3	2.31506	2.35041	2.40311	FJ
7.94228	10	2.3161	2.43394	2.6096	ABCDEFGIJ
7.94628	10	2.31661	2.43444	2.61011	ABCDEGHIJ
9.55281	3	2.33001	2.36536	2.41806	IJ
7.90691	11	2.33603	2.46565	2.65888	ABCDEFGHIJ
9.71534	3	2.34688	2.38223	2.43493	DJ
10.0846	2	2.35979	2.38336	2.41849	J
9.87462	3	2.36314	2.39849	2.45119	HJ
11.5449	2	2.49502	2.51859	2.55372	B
13.4123	2	2.64495	2.66852	2.70365	A
15.738	2	2.80486	2.82842	2.86356	D
16.6421	2	2.86072	2.88429	2.91942	F
25.1105	1	3.24768	3.25946	3.27703	

Information criteria are based on the residual mean squared error of the fitted model, together with a penalty for using a large number of coefficients in the model. The smaller the value of the criterion, the better the model.

There are three criteria to choose from:

Akaike Information Criterion
The Akaike Information Criterion (AIC) is calculated from

$$AIC = 2\ln(RMSE) + \frac{2p}{n} \tag{7}$$

where *RMSE* is the root mean squared error during the estimation period, *p* is the number of estimated coefficients in the fitted model, and *n* is the sample size used to fit the model. Notice that the AIC is a function of the variance of the model residuals, penalized by the number of estimated parameters. In general, the model will be selected that minimizes the mean squared error without using too many coefficients (relative to the amount of data available).

Hannan-Quinn Criterion
The Hannan Quinn Criterion (HQC) is calculated from

$$HQC = 2\ln(RMSE) + \frac{2p\ln(\ln(n))}{n} \tag{8}$$

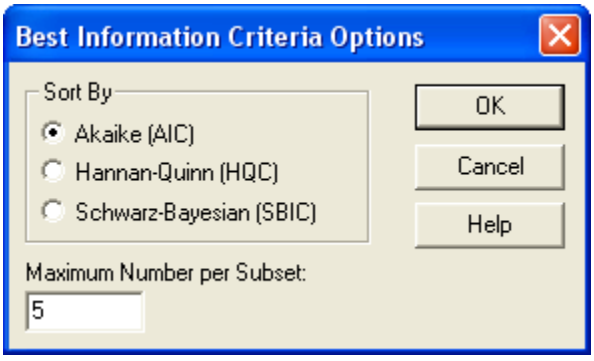
This criterion uses a different penalty for the number of estimated parameters.

Schwarz-Bayesian Information Criterion
The Schwarz-Bayesian Information Criterion (SBIC) is calculated from

$$SBIC = 2\ln(RMSE) + \frac{p\ln(n)}{n} \tag{9}$$

Again, the penalty for the number of estimated parameters is different than for the other criteria.

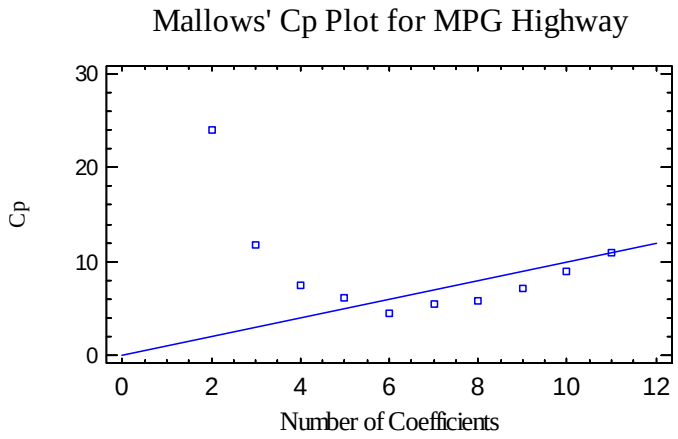
Pane Options



- **Sort By:** Select the criterion by which to sort the models for display in the table.
- **Maximum Number per Subset:** the maximum number of models included in the table that contain the same number of independent variables.

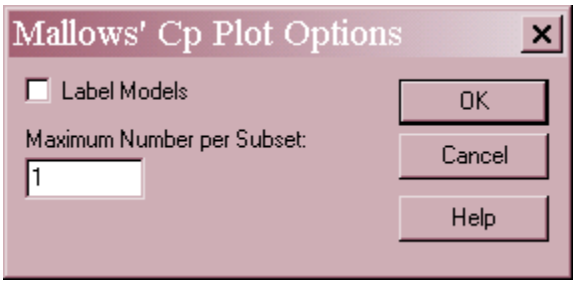
Mallows' Cp Plot

This plot shows the models with the smallest Cp values. After rescaling the vertical axis, the plot shows:



Small values are desired, provided they lie below the diagonal line, defined by Cp = p. Increasing the number of independent variables up to 5 (plus a constant) improves the statistic. Beyond p = 5, Cp increases.

Pane Options

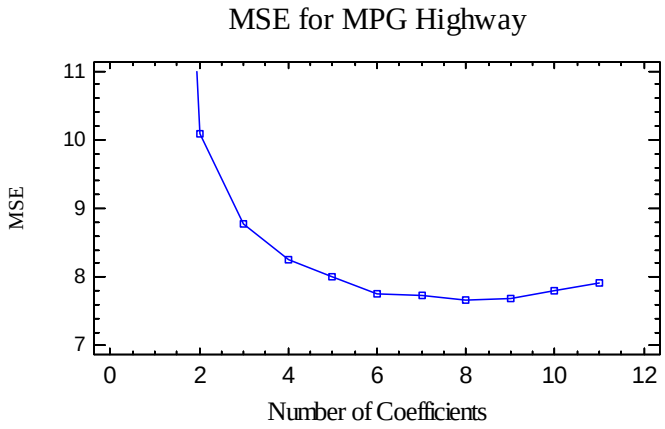


- **Label Models:** if selected, model labels will be added to the plot.

- **Maximum Number per Subset:** the maximum number of models included in the plot that contain the same number of independent variables.

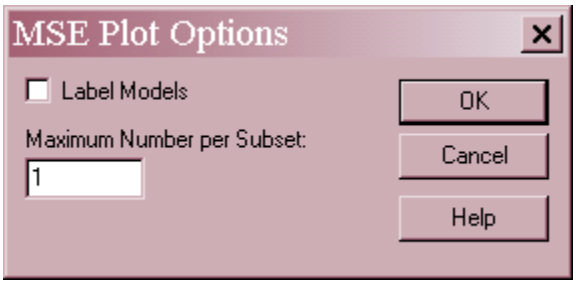
MSE Plot

This plot shows the models with the smallest mean squared error values. After rescaling the vertical axis, the plot shows:



The MSE continues to drop through 8 coefficients, although the drop after 6 is very small.

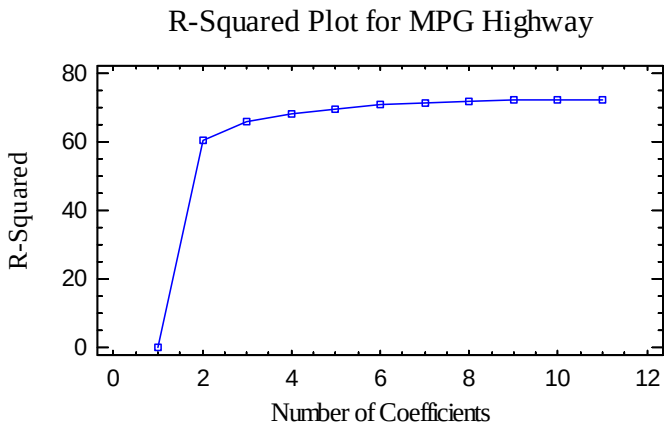
Pane Options



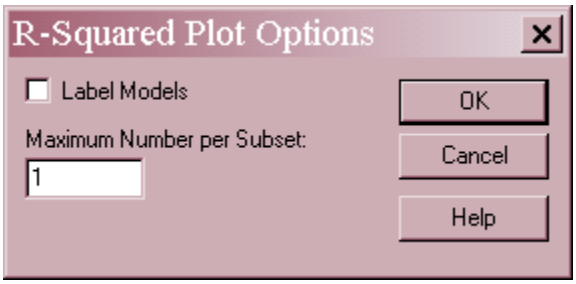
- **Label Models:** if selected, model labels will be added to the plot.
- **Maximum Number per Subset:** the maximum number of models included in the plot that contain the same number of independent variables.

R-Squared Plot

This plot shows the models with the largest R-Squared values:



Pane Options



- **Label Models:** if selected, model labels will be added to the plot.
- **Maximum Number per Subset:** the maximum number of models included in the plot that contain the same number of independent variables.

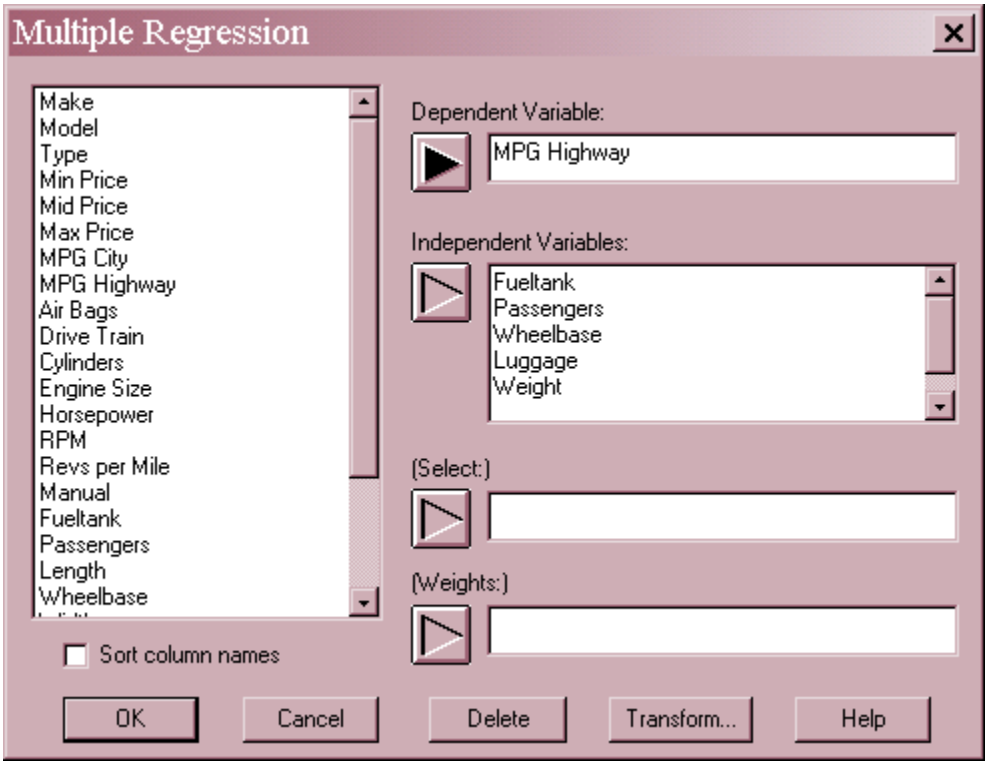
Save Results

A row for each fitted model may be saved to the datasheet, including:

1. *Model Identifiers* – an indication of the independent variables included in the model.
2. *Adjusted R-squared* – the adjusted R-squared statistic.
3. *Cp* – Mallows' Cp statistic.
4. *MSE* – the mean squared error.
5. *R-squared* – the unadjusted R-squared statistic.

Fitting the Best Model

If the BCEIJ model is judged to be best, it can be fit using the *Multiple Regression* procedure.



The output is shown below:

Multiple Regression - MPG Highway

Dependent variable: MPG Highway

		Standard	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
CONSTANT	22.9571	8.36713	2.74372	0.0076
Fueltank	-0.454275	0.238684	-1.90325	0.0608
Passengers	-1.96767	0.621566	-3.16567	0.0022
Wheelbase	0.451142	0.120719	3.73713	0.0004
Luggage	0.34953	0.163638	2.136	0.0359
Weight	-0.0091559	0.00161292	-5.67659	0.0000

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	1444.94	5	288.988	37.29	0.0000
Residual	589.013	76	7.75017		
Total (Corr.)	2033.95	81			

R-squared = 71.041 percent
R-squared (adjusted for d.f.) = 69.1358 percent
Standard Error of Est. = 2.78391
Mean absolute error = 2.00394
Durbin-Watson statistic = 1.55279 (P=0.0179)
Lag 1 residual autocorrelation = 0.221875

The final model is:

$$MPG\ Highway = 22.9571 - 0.454275\ Fueltank - 1.96767\ Passengers + 0.451142\ Wheelbase + 0.34953\ Luggage - 0.0091559\ Weight \quad (7)$$