Chapter 311

Stepwise Regression

Introduction

Often, theory and experience give only general direction as to which of a pool of candidate variables (including transformed variables) should be included in the regression model. The actual set of predictor variables used in the final regression model must be determined by analysis of the data. Determining this subset is called the *variable selection* problem.

Finding this subset of regressor (independent) variables involves two opposing objectives. First, we want the regression model to be as complete and realistic as possible. We want every regressor that is even remotely related to the dependent variable to be included. Second, we want to include as few variables as possible because each irrelevant regressor decreases the precision of the estimated coefficients and predicted values. Also, the presence of extra variables increases the complexity of data collection and model maintenance. The goal of variable selection becomes one of parsimony: achieve a balance between simplicity (as few regressors as possible) and fit (as many regressors as needed).

There are many different strategies for selecting variables for a regression model. If there are no more than fifteen candidate variables, the *All Possible Regressions* procedure (discussed in the next chapter) should be used since it will always give as good or better models than the stepping procedures available in this procedure. On the other hand, when there are more than fifteen candidate variables, the four search procedures contained in this procedure may be of use.

These search procedures will often find very different models. Outliers and collinearity can cause this. If there is very little correlation among the candidate variables and no outlier problems, the four procedures should find the same model.

We will now briefly discuss each of these procedures.

Variable Selection Procedures

Forward (Step-Up) Selection

This method is often used to provide an initial screening of the candidate variables when a large group of variables exists. For example, suppose you have fifty to one hundred variables to choose from, way outside the realm of the all-possible regressions procedure. A reasonable approach would be to use this forward selection procedure to obtain the best ten to fifteen variables and then apply the all-possible algorithm to the variables in this subset. This procedure is also a good choice when multicollinearity is a problem.

The forward selection method is simple to define. You begin with no candidate variables in the model. Select the variable that has the highest R-Squared. At each step, select the candidate variable that increases R-Squared the most. Stop adding variables when none of the remaining variables are significant. Note that once a variable enters the model, it cannot be deleted.

Backward (Step-Down) Selection

This method is less popular because it begins with a model in which all candidate variables have been included. However, because it works its way down instead of up, you are always retaining a large value of R-Squared. The problem is that the models selected by this procedure may include variables that are not really necessary. The user sets the significance level at which variables can enter the model.

The backward selection model starts with all candidate variables in the model. At each step, the variable that is the least significant is removed. This process continues until no nonsignificant variables remain. The user sets the significance level at which variables can be removed from the model.

Stepwise Selection

Stepwise regression is a combination of the forward and backward selection techniques. It was very popular at one time, but the Multivariate Variable Selection procedure described in a later chapter will always do at least as well and usually better.

Stepwise regression is a modification of the forward selection so that after each step in which a variable was added, all candidate variables in the model are checked to see if their significance has been reduced below the specified tolerance level. If a nonsignificant variable is found, it is removed from the model.

Stepwise regression requires two significance levels: one for adding variables and one for removing variables. The cutoff probability for adding variables should be less than the cutoff probability for removing variables so that the procedure does not get into an infinite loop.

Min MSE

This procedure is similar to the Stepwise Selection search procedure. However, instead of using probabilities to add and remove, you specify a minimum change in the root mean square error. At each step, the variable whose status change (in or out of the model) will decrease the mean square error the most is selected and its status is reversed. If it is currently in the model, it is removed. If it is not in the model, it is added. This process continues until no variable can be found that will cause a change larger than the user-specified minimum change amount.

Assumptions and Limitations

The same assumptions and qualifications apply here as applied to multiple regression. Note that outliers can have a large impact on these stepping procedures, so you must make some attempt to remove outliers from consideration before applying these methods to your data.

The greatest limitation with these procedures is one of sample size. A good rule of thumb is that you have at least five observations for each variable in the candidate pool. If you have 50 variables, you should have 250 observations. With less data per variable, these search procedures may fit the randomness that is inherent in most datasets and spurious models will be obtained.

This point is critical. To see what can happen when sample sizes are too small, generate a set of random numbers for 20 variables with 30 observations. Run any of these procedures and see what a magnificent value of R-Squared is obtained, even though its theoretical value is zero!

Using This Procedure

This procedure performs one portion of a regression analysis: it obtains a set of independent variables from a pool of candidate variables. Once the subset of variables is obtained, you should proceed to the Multiple Regression procedure to estimate the regression coefficients, study the residuals, and so on.

Data Structure

An example of data appropriate for this procedure is shown in the table below. This data is from a study of the relationships of several variables with a person's IQ. Fifteen people were studied. Each person's IQ was recorded along with scores on five different personality tests. The data are contained in the IQ dataset. We suggest that you open this database now so that you can follow along with the example.

IQ dataset

Test1	Test2	Test3	Test4	Test5	IQ
83	34	65	63	64	106
73	19	73	48	82	92
54	81	82	65	73	102
96	72	91	88	94	121
84	53	72	68	82	102
86	72	63	79	57	105
76	62	64	69	64	97
54	49	43	52	84	92
37	43	92	39	72	94
42	54	96	48	83	112
71	63	52	69	42	130
63	74	74	71	91	115
69	81	82	75	54	98
81	89	64	85	62	96
50	75	72	64	45	103

Missing Values

Rows with missing values in the active variables are ignored.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

Specify the variables on which to run the analysis.

Dependent Variable

Y: Dependent Variable

Specifies a dependent (Y) variable. If more than one variable is specified, a separate analysis is run for each.

Weight Variable

Weight Variable

Specifies a variable containing observation (row) weights for generating weighted regression analysis. These weights might be those saved during a robust regression analysis.

Independent Variables

X's: Independent Variables

Specify the independent (X or candidate) variables.

Model Selection

Selection Method

This option specifies which of the four search procedures should be used: Forward, Backward, Stepwise, or Min MSE.

Prob to Enter

Sometimes call PIN, this is the probability required to enter the equation. This value is used by the Forward and the Stepwise procedures. A variable, not currently in the model, must have a t-test probability value less than or equal to this in order to be considered for entry into the regression equation. You must set PIN < POUT.

Prob to Remove

Sometimes call POUT, this is the probability required to be removed from the equation. This value is used by the Backward and the Stepwise procedures. A variable, currently in the model, must have a t-test probability value greater than this in order to be considered for removal from the regression equation. You must set PIN < POUT.

Min RMSE Change

This value is used by the Minimum MSE procedure to determine when to stop. The procedure stops when the maximum relative decrease in the square root of the mean square error brought about by changing the status of a variable is less than this amount.

Maximum Iterations

This is the maximum number of iterations that will be allowed. This option is useful to prevent the unlimited looping that may occur. You should set this to a high value, say 50 or 100.

Remove Intercept

Unchecked indicates that the intercept term is to be included in the regression. Checked indicates that the intercept should be omitted from the regression model. Note that deleting the intercept distorts most of the diagnostic statistics (R-Squared, etc.).

Reports Tab

These options control the reports that are displayed.

Select Reports

Descriptive Statistics and Selected Variables Reports

This option specifies whether the indicated report is displayed.

Report Options

Report Format

Two output formats are available: brief and verbose. The *Brief* output format consists of a single line for each step (scan through the variables). The *Verbose* output format gives a complete table of each variable's statistics at each step. If you have many variables, the Verbose option can produce a lot of output.

Precision

Specifies the precision of numbers in the report. Single precision will display seven-place accuracy, while double precision will display thirteen-place accuracy.

Variable Names

This option lets you select whether to display only variable names, variable labels, or both.

Example 1 – Stepwise Regression Analysis

This section presents an example of how to run a stepwise regression analysis of the data presented in the IQ dataset.

You may follow along here by making the appropriate entries or load the completed template **Example 1** by clicking on Open Example Template from the File menu of the Stepwise Regression window.

1 Open the IQ dataset.
- From the File menu of the NCSS Data window, select **Open Example Data**.
- Click on the file **IQ.NCSS**.
- Click **Open**.

2 Open the Stepwise Regression window.
- Using the Analysis menu or the Procedure Navigator, find and select the **Stepwise Regression** procedure.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.
- On the Stepwise Regression window, select the **Variables tab**.
- Double-click in the **Y: Dependent Variable** text box. This will bring up the variable selection window.
- Select **IQ** from the list of variables and then click **Ok**. "IQ" will appear in the Y: Dependent Variable box.
- Double-click in the **X's: Independent Variables** text box. This will bring up the variable selection window.
- Select **Test1** through **Test5** from the list of variables and then click **Ok**. "Test1-Test5" will appear in the X's: Independent Variables.
- In the **Selection Method** list box, select **Backward**.

4 Specify the reports.
- On the Stepwise Regression window, select the **Reports tab**.
- In the **Report Format** list box, select **Verbose**.
- Check the **Descriptive Statistics** checkbox.

5 Run the procedure.
- From the Run menu, select **Run Procedure**. Alternatively, just click the green Run button.

Descriptive Statistics Section

Descriptive Statistics Section

Variable	Count	Mean	Standard Deviation
Test1	15	67.93333	17.39239
Test2	15	61.4	19.39735
Test3	15	72.33334	14.73415
Test4	15	65.53333	13.95332
Test5	15	69.93333	16.15314
IQ	15	104.3333	11.0173

For each variable, the Count, Mean, and Standard Deviation are calculated. This report is especially useful for making certain that you have selected the right variables and that the appropriate number of rows was used.

Iteration Detail Section (Verbose Version)

Iteration Detail Section

Iteration 0: Unchanged

In	Variable	Standard. Coefficient	R-Squared Increment	R-Squared Other X's	T-Value	Prob Level	Pct Change Sqrt(MSE)
Yes	Test1	-3.0524	.2357	.9747	-1.8789	.0930	11.9387
Yes	Test2	-2.9224	.2414	.9717	-1.9016	.0897	12.3210
Yes	Test3	.1404	.0152	.2280	.4773	.6445	-3.9386
Yes	Test4	4.7853	.2832	.9876	2.0596	.0695	15.0741
Yes	Test5	-.0595	.0027	.2329	-.2017	.8447	-4.9176
R-Squared=0.3991		Sqrt(MSE)=10.65198					

Iteration 1: Removed Test5 from equation

In	Variable	Standard. Coefficient	R-Squared Increment	R-Squared Other X's	T-Value	Prob Level	Pct Change Sqrt(MSE)
Yes	Test1	-3.0612	.2373	.9747	-1.9825	.0756	12.5340
Yes	Test2	-2.9032	.2392	.9716	-1.9906	.0745	12.6640
Yes	Test3	.1163	.0125	.0752	.4550	.6588	-3.6717
Yes	Test4	4.7850	.2832	.9876	2.1660	.0555	15.5681
No	Test5		.0027	.2329	.2017	.8447	5.1719
R-Squared=0.3964		Sqrt(MSE)=10.12816					

Iteration 2: Removed Test3 from equation

In	Variable	Standard. Coefficient	R-Squared Increment	R-Squared Other X's	T-Value	Prob Level	Pct Change Sqrt(MSE)
Yes	Test1	-3.1020	.2444	.9746	-2.0890	.0607	13.1519
Yes	Test2	-2.9024	.2391	.9716	-2.0659	.0632	12.7977
Yes	Test4	4.7988	.2849	.9876	2.2553	.0455	15.7808
No	Test3		.0125	.0752	.4550	.6588	3.8116
No	Test5		.0000	.0810	.0087	.9932	4.8805
R-Squared=.3839		Sqrt(MSE)=9.756291					

Iteration 3: Unchanged

In	Variable	Standard. Coefficient	R-Squared Increment	R-Squared Other X's	T-Value	Prob Level	Pct Change Sqrt(MSE)
Yes	Test1	-3.1020	.2444	.9746	-2.0890	.0607	13.1519
Yes	Test2	-2.9024	.2391	.9716	-2.0659	.0632	12.7977
Yes	Test4	4.7988	.2849	.9876	2.2553	.0455	15.7808
No	Test3		.0125	.0752	.4550	.6588	3.8116
No	Test5		.0000	.0810	.0087	.9932	4.8805
R-Squared=.3839		Sqrt(MSE)=9.756291					

This report presents information about each step of the search procedures. You can scan this report to see if you would have made the same choice. Each report shows the statistics after the specified action (entry or removal) was taken.

For each iteration, there are three possible actions:

1. *Unchanged*. No action was taken because of the scan in this step. Because of the "backward look" in the stepwise search method, this will show up a lot when this method is used. Otherwise, it will usually show up as the first and last steps.

2. *Removal*. A variable was removed from the model.

3. *Entry*. A variable was added to the model.

Individual definitions of the items on the report are as follows:

In

A *Yes* means the variable is in the model. A *No* means it is not.

Variable

This is the name of the candidate variable.

Standard. Coefficient

Standardized regression coefficients are the coefficients that would be obtained if you standardized each independent and dependent variable. Here *standardizing* is defined as subtracting the mean and dividing by the standard deviation of a variable. A regression analysis on these standardized variables would yield these standardized coefficients.

When there are vastly different units involved for the variables, this is a way of making comparisons between variables. The formula for the standardized regression coefficient is:

$$b_{j,std} = b_j \left(\frac{s_{x_j}}{s_y} \right)$$

where s_y and s_{x_j} are the standard deviations for the dependent variable and the corresponding j^{th} independent variable.

R-Squared Increment

This is the amount that R-Squared would be changed if the status of this variable were changed. If the variable is currently in the model, this is the amount the R-Squared value would be decreased if it were removed. If the variable is currently out of the model, this is the amount the overall R-Squared would be increased if it were added. Large values here indicate important independent variables.

You want to add variables that make a large contribution to R-Squared and to delete variables that make a small contribution to R-Squared.

R-Squared Other X's

This is a collinearity measure, which should be as small as possible. This is the R-Squared value that would result if this independent variable were regressed on all of the other independent variables currently in the model.

T-Value

This is the t-value for testing the hypothesis that this variable should be added to, or deleted from, the model. The test is adjusted for the rest of the variables in the model. The larger this t-value is, the more important the variable.

Prob Level

This is the two-tail p-value for the above t-value. The smaller this p-value, the more important the independent variable is. This is the significance value that is compared to the values of PIN and POUT (see *Stepwise Method* above).

Pct Change Sqrt(MSE)

This is the percentage change in the square root of the mean square error that would occur if the specified variable were added to, or deleted from, the model. This is the value that is used by the *Min MSE* search procedure. This percentage change in root mean square error (RMSE) is computed as follows:

$$Percent\ change\ = \left[\frac{RMSE_{previous}\ -\ RMSE_{current}}{RMSE_{current}} \right] 100$$

R-Squared

This is the R-Squared value for the current model.

Sqrt(MSE)

This is the square root of the mean square error for the current model.

Iteration Detail Section (Brief Version)

This report was not printed because the Report Format box was set to Verbose. If this option had been set to Brief, this is the output that would have been displayed.

Iteration Detail Section

Iter. No.	Action	Variable	R-Squared	Sqrt(MSE)	Max R-Squared Other X's
0	Unchanged		0.399068	10.65198	0.987631
1	Removed	Test5	0.396353	10.12816	0.987631
2	Removed	Test3	0.383854	9.756291	0.987628
3	Unchanged		0.383854	9.756291	0.987628

This is an abbreviated report summarizing the statistics at each iteration. Individual definitions of the items on the report are as follows:

Iter. No.

The number of this iteration.

Action

For each iteration, there are three possible actions:

1. *Unchanged*. No action was taken because of the scan in this step. Because of the "backward look" in the stepwise search method, this will show up a lot when this method is used. Otherwise, it will show up at the first and last steps.

2. *Removed*. A variable was removed from the model.

3. *Added*. A variable was added to the model.

Variable

This is the name of the variable whose status is being changed.

R-Squared

The value of R-Squared for the current model.

Sqrt(MSE)

This is the square root of the mean square error for the current model.

Max R-Squared Other X's

This is the maximum value of R-Squared Other X's (see verbose report definitions) for all the variables in the model. This is a collinearity model. You want this value to be as small as possible. If it approaches 0.99, you should be concerned with the possibility that multicollinearity is distorting your results.