Chapter 310

Subset Selection in Multivariate Y Multiple Regression

Introduction

Often theory and experience give only general direction as to which of a pool of candidate 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 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 search procedure contained in this procedure is an excellent choice.

While studying at Texas A&M University, Dr. Claude McHenry (1978) developed a heuristic algorithm that often yields the same subset as the all possible regressions routine, but with a lot less work. The algorithm is of a more general nature than the other variable selection procedures in **NCSS** because it allows more than one dependent variable to be studied. Hence, it is useful for variable selection in multivariate multiple regression and in discriminant analysis.

McHenry's Select Algorithm

The algorithm seeks a subset that provides a maximum value of R-Squared (or a minimum Wilks' lambda in the multivariate case). The algorithm first finds the best single variable. To find the best pair of variables, it tries each of the remaining variables and selects the one that adds the most. It then omits the first variable and determines if any other variable would add more. If a better variable is found, it is kept and the worst variable is removed. Another search is now made through the remaining variables. This switching process continues until no switching will result in a better subset.

Once the optimal pair of variables is found, the best three variables is searched for in much the same manner. First, the best third variable is found to add to the optimal pair of variables from the last step. Next, each of the first two variables is omitted and another, even better, variable is searched for. The algorithm continues until no switching improves R-Squared.

This algorithm is extremely fast. It seems to find the best (or very near best) subset in most situations. An interesting feature is the ability to specify more than one dependent variable. This is useful in discriminant analysis where each group may be considered as a binary (0, 1) variable. It is also useful when you want to predict several dependent variables using a minimum number of independent variables.

Assumptions and Limitations

The same assumptions and qualifications apply here as applied to multiple regression. We refer you to the Assumptions section in the Multiple Regression chapter for a discussion of these assumptions. We will here mention a couple of restrictions necessary for this algorithm to work.

Number of Observations

The number of observations must be at least one greater than the number of candidate regressors. A popular rule-of-thumb when using any variable selection procedure is that you have at least five observations for each candidate variable.

No Linear Dependencies

This algorithm begins by fitting the full model with all candidate variables. In order to solve this full model, no linear dependencies may exist in the data. A linear dependency occurs when one variable is a weighted average of the rest. For example, if one variable is the total of several others, it cannot be included in the candidate pool.

This same restriction applies to the set of dependent variables.

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 set 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

The data are entered in two or more variables. An example of data appropriate for this procedure is shown below. These data are contained in the IQ dataset.

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

Missing Values

Rows with missing values in the variable pool are ignored. This may cause differences in the results between this procedure and regression analysis. Suppose that through the selection process, none of the variables with missing values end up in the final subset. When a regression analysis is run on the subset, the rows with missing values will not be deleted (since those variables are no longer active). This will obviously change the estimated values.

Example 1 – Variable Selection Analysis

This section presents an example of how to run a variable selection analysis of the data contained in the IQ dataset.

Setup

To run this example, complete the following steps:

1 Open the IQ example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **IQ** and click **OK**.

2 Specify the Subset Selection in Multivariate Y Multiple Regression procedure options

- Find and open the **Subset Selection in Multivariate Y Multiple Regression** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 1** settings template. To load this template, click **Open Example Template** in the Help Center or File menu.

Option	Value
Variables Tab	
Y's: Dependent Variables	**IQ**
X's: Independent Variables	**Test1-Test5**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Code Cross-Reference Section

Code Cross-Reference Section

Code	Variable	Count	Mean
	IQ	15	104.3333
A	Test1	15	67.93333
B	Test2	15	61.4
C	Test3	15	72.33334
D	Test4	15	65.53333
E	Test5	15	69.93333

The code, count, and mean are displayed for each variable. This report is particularly useful for checking that the correct variables were selected. The letters A to Z are assigned to each of the independent variables involved in the regression model. These are used to specify which variables are in each subset.

Selection Results Section

Selection Results Section

Model Size	R-Squared	R-Squared Change	Coded Variables
1	0.137941	0.137941	D
2	0.154246	0.016305	CD
3	0.383854	0.229608	ABD
4	0.396353	0.012499	ABCD
5	0.399068	0.002715	ABCDE

This report presents the results of the search procedure. The model for each subset (model) size is presented. To use this report, you scan down the R-Squared values, looking for the subset size where R-Squared stabilizes. In this example, the R-Squared value for the best three-variable model is 0.383854, and the R-Squared for the best four-variable model is 0.396353. This is a minor increase. We would select the three-variable model as our final model.

If more the one dependent variable is specified, the R-Squared column will be replaced by a Wilks' Lambda column.

Model Size

This is the number of independent variables in the model.

R-Squared

This is the value of R-Squared achieved for this subset. Note that if multiple dependent variables are specified, this column will be labeled Wilks' Lambda. Wilks' Lambda is the multivariate extension of R-Squared. It behaves like 1-(R-Squared). Hence, when you have multiple dependent variables, you look for a value close to zero, rather than close to one as you do with R-Squared.

R-Squared Change

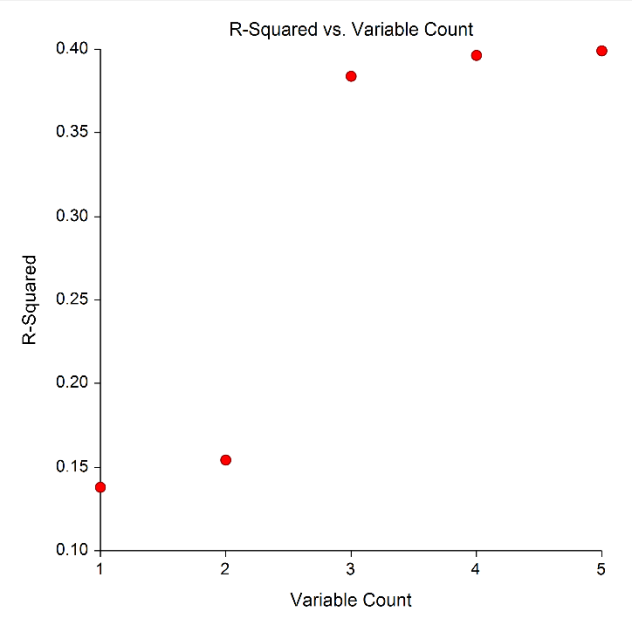
This is the amount that is added to R-Squared (or Wilks' Lambda) when an additional variable is added to the model.

Coded Variables

This is a list of the variables in the model. The variable which each letter represents is listed in the Code Cross-Reference Section.

R-Squared vs Variable Count Plot

Plots Section



This plot displays the values of R-Squared on the vertical axis and the subset size on the horizontal axis for the data displayed in the Selection Results Section, above. Note the large jump between the two-variable model and the three-variable model. We quickly see that the four and five variable models do not do much better. Hence, our conclusion is to use the three-variable model. The three-variable model is ABD, which translates to variables Test1, Test2, and Test4.