Chapter 312

All Possible Regressions

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. The phrase "throw in the kitchen sink" takes on new meaning here. 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).

After a pool of candidate variables has been formed, the next task is to establish a basis for comparing two models. How do we decide if model A is better than model B? Three statistics have been found useful for selecting among various regression models. These are R-Squared, mean square error, and Cp. Other criteria have been suggested, but these three are the most popular.

Once we have a pool of variables and a selection criterion, the final task in variable selection is to plan a strategy to see how each of the possible models does on the criterion. The problem that now arises is that there are too many possible models to choose from. The number of possible models that can be formed from p regressors is 2 to the power p. If we have p = 4 regressors, there are 16 possible models to choose from. With 15 regressors, there are 32,768 possible models. With 20 regressors, there are 1,048,576 models. Obviously, the number of possible models grows exponentially with the number of regressors. However, with up to 15 regressors, the problem does seem manageable.

This procedure was programmed so that it will efficiently look at up to 32,768 models for up to 15 regressors. That is why it is called *all possible regressions*. It guarantees that you will find the "best" model, since it looks at all of them. Unfortunately, no automatic procedure will find the "best" model in every sense. It will, however, find the model that is best according to your selection criterion. It is still left up to you to determine if the model makes theoretical and practical sense.

All Possible Regressions

This algorithm fits all regressions involving one regressor, two regressors, three regressors, and so on. The selection criterion is recorded for each regression. Once the procedure finishes, the champion for each subset size is determined. You then determine which subset size is optimum for your case.

The All Possible Regressions solution is the target of the popular step-regression procedures. Although it takes longer to run, it guarantees the right answer. Hence, when you have 15 or fewer independent variables to choose from, this is the variable selection procedure you should use.

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 (Chapter 15) for a discussion of these assumptions. We will here mention restrictions necessary for this algorithm to work.

Number of Regressor Variables

This procedure will work with up to fifteen regressor variables, not including the intercept. The intercept is always included in the regression model.

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

Since one of the models that must be solved involves all of the candidate variables (the full model), no linear dependencies can exist among these variables. 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.

Data Structure

An example of data appropriate for this procedure is shown in the table below. These data are 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 dataset 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 variable pool are ignored. The program does not run a separate analysis for each pattern of missing values. It is possible to get slightly different results when you analyze a subset because variables in the subset may not contain missing values. Without the missing values, the rows that were deleted in the original analysis are included in the subset analysis.

Example 1 – All Possible Regressions Analysis

This section presents an example of how to run an all possible regressions 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 All Possible Regressions procedure options

- Find and open the **All Possible Regressions** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 1** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Variables Tab	
Y: Dependent Variable	**IQ**
X's: Independent Variables	**Test1-Test5**
Alternative Models	**5**
Reports Tab	
Descriptive Statistics	**Checked**

3 Run the procedure

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

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 of nonmissing values, the arithmetic mean of the nonmissing values, and the standard deviation of the nonmissing values are computed. This report is particularly useful for checking that the correct variables were selected.

All Possible Results Section

All Possible Results Section

Model Size	R-Squared	Root MSE	Cp	Model
1	0.137941	10.61539	1.910838	D (Test4)
1	0.057913	11.09719	3.109405	B (Test2)
1	0.050917	11.13832	3.214175	A (Test1)
1	0.005486	11.40179	3.894581	C (Test3)
1	0.003371	11.41391	3.926255	E (Test5)
Variables in Best Model: Test4				
2	0.154246	10.94386	3.666643	CD
2	0.144790	11.00487	3.808266	AD
2	0.139411	11.03943	3.888825	BD
2	0.137980	11.0486	3.910256	DE
2	0.098957	11.29591	4.494690	AB
Variables in Best Model: Test3, Test4				
3	0.383854	9.756291	2.227864	ABD
3	0.159103	11.39763	5.593906	BCD
3	0.157158	11.4108	5.623033	ACD
3	0.155707	11.42062	5.644768	CDE
3	0.145431	11.48991	5.798660	ADE
Variables in Best Model: Test1, Test2, Test4				
4	0.396353	10.12816	4.040666	ABCD
4	0.383859	10.23245	4.227794	ABDE
4	0.163351	11.92369	7.530276	BCDE
4	0.157627	11.96441	7.616005	ACDE
4	0.115826	12.25768	8.242057	ABCE
Variables in Best Model: Test1, Test2, Test3, Test4				
5	0.399068	10.65198	6.000000	ABCDE
Variables in Best Model: Test1, Test2, Test3, Test4, Test5				

This report presents the results of the all possible regressions search procedure. The models for each subset (model) size are sorted from best to worst. To use this report, you scan down a criterion column, say R-Squared, for the subset size where this value 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.

Model Size

This is the number of independent variables in the model. Model size will range from 1 to p. The option, Alternative Models, controls the number of models reported from each model subset size.

R-Squared

R-Squared is the ratio of the variation explained by the model to the total variation in the dependent variable. R-Squared ranges from zero to one. The larger the R-Squared, the better the model. A comprehensive definition of R-Squared is given in the Multiple Regression chapter.

Root MSE

This is the square root of the mean square error. The smaller this value is, the better the model.

Cp

Another criterion for variable selection and importance is Mallow's Cp statistic. The optimum model will have a Cp value close to p+1, where p is the number of independent variables. A Cp greater than (p+1) indicates that the regression model is over specified (contains too many variables and stands a chance of having collinearity problems). On the other hand, a model with a Cp less than (p+1) indicates that the regression model is underspecified (at least one important independent variable has been omitted). The formula for the Cp statistic is as follows, where k is the maximum number of independent variables available:

$$C_p = \left[\frac{MSE_p}{MSE_k}\right][n - p - 1] - [n - 2(p + 1)]$$

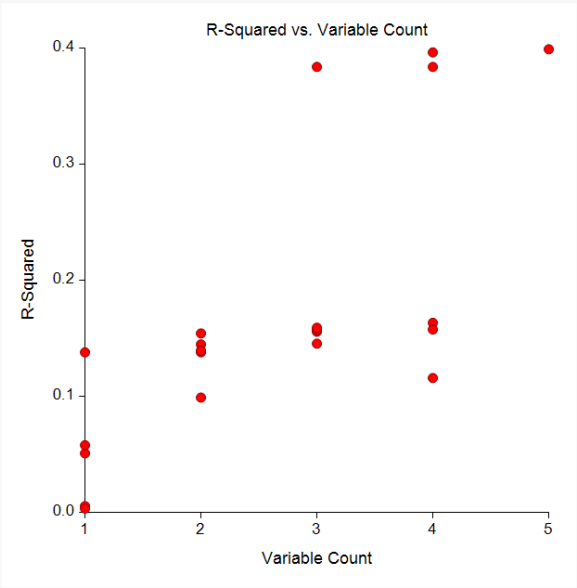
Model

This column labels the model whose statistics are being reported. Letters are given in a shorthand notation to represent the independent variables. The letter A is associated with the first variable, the letter B with the second, and so on. The letters and corresponding variable names are displayed for all variables in the first section of the report. Two-variable models are represented by two letters. Hence, in this example, the model CD represents the two-variable model consisting of variables Test3 and Test4.

R-Squared vs. Variable Count Plot

This plot displays the values of R-Squared on the vertical axis and the number of independent variables on the horizontal axis for the data displayed in the All Possible Results Section above. Note the large disparity in the three-variable models. There is one model that is way above the rest. We can also quickly see that the four- and five-variable models do not do much better.

R-Squared vs Variable Count Plot

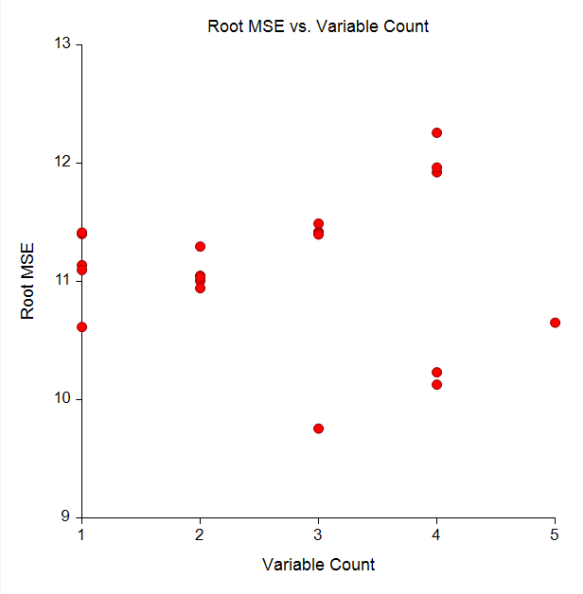


Root MSE vs. Variable Count Plot

This plot displays the values of the square root of the mean square error on the vertical axis and the number of independent variables on the horizontal axis for the data displayed in the All Possible Results Section above. Note the large disparity in the three-variable models. There is one model that is way below the rest.

Root MSE is often considered a better criterion for choosing a best model than R-Squared. The root MSE decreases initially as p increases, stabilizes at some subset size, and eventually begins to increase with further increments of p. You should choose a best model based on the minimum MSE or a value of p near the point where the smallest MSE turns upward. The model subset that minimizes MSE will usually maximize R-Squared.

Root MSE vs Variable Count Plot



Cp vs. Variable Count Plot

This plot displays the values of Cp on the vertical axis and the number of independent variables on the horizontal axis for the data displayed in the All Possible Results Section above. The Cp plot is more difficult to interpret because we are looking for the model where Cp is closest to p+1. You will most likely want to stick with the numeric report when considering Cp.

Cp vs Variable Count Plot