Chapter 307

Subset Selection in Multiple Regression

Introduction

Multiple regression analysis is documented in *Chapter 305 – Multiple Regression*, so that information will not be repeated here. Refer to that chapter for in depth coverage of multiple regression analysis. This chapter will deal solely with the topic of subset selection.

Subset selection refers to the task of finding a small subset of the available independent variables that does a good job of predicting the dependent variable. Exhaustive searches are possible for regressions with up to 15 IV's. However, when more than 15 IV's are available, algorithms that add or remove a variable at each step must be used. Two such searching algorithms are available in this module: forward selection and forward selection with switching.

An issue that comes up because of categorical IV's is what to do with the internal variables that are generated for a categorical independent variable. If such a variable has six categories, five internal variables are generated. You can see that with two or three categorical variables, a large number of internal variables will result, which greatly increases the total number of variables that must be searched. To avoid this problem, the algorithms search on model terms rather than on the individual internal variables. Thus, the whole set of internal variables associated with a given term are considered together for inclusion in, or deletion from, the model. It's all or none. Because of the time consuming nature of the algorithm, this is the only feasible way to deal with categorical variables. If you want the subset algorithm to deal with them individually, you can save the internal variables in a first run and designate them as Numeric Variables.

Hierarchical Models

Another issue is what to do with interactions. Usually, an interaction is not entered in the model unless the individual terms that make up that interaction are also in the model. For example, the interaction term A*B*C is not included unless the terms A, B, C, A*B, A*C, and B*C are already in the model. Such models are said to be *hierarchical*. You have the option during the search to force the algorithm to consider only hierarchical models during its search. Thus, if C is not in the model, interactions involving C are not even considered. Even though the option for non-hierarchical models is available, we recommend that you only consider hierarchical models.

Selection Methods

Forward Selection

The method of forward selection proceeds as follows.

1. Begin with no terms in the model.
2. Find the term that, when added to the model, achieves the largest value of R^2. Enter this term into the model.

3. Continue adding terms until a target value for R^2 is achieved or until a preset limit on the maximum number of terms in the model is reached. Note that these terms can be limited to those keeping the model hierarchical.

This method is comparatively fast, but it does not guarantee that the best model is found except for the first step when it finds the best single term. You might use it when you have a large number of observations and terms so that other, more time consuming, methods are not feasible.

Forward Selection with Switching

This method is similar to the method of Forward Selection discussed above. However, at each step when a term is added, all terms in the model are switched one at a time with all candidate terms not in the model to determine if they increase the value of R^2. If a switch can be found, it is made and the pool of terms is again searched to determine if another switch can be made. Note that this switching can be limited to those keeping the model hierarchical.

When the search for possible switches does not yield a candidate, the subset size is increased by one and a new search is begun. The algorithm is terminated when a target subset size is reached or all terms are included in the model.

Discussion

These algorithms usually require two runs. In the first run, you set the maximum subset size to a large value such as 10. By studying the Subset Selection reports from this run, you can quickly determine the optimum number of terms. You reset the maximum subset size to this number and make the second run. This two-step procedure works better than relying on some *F*-to-enter and *F*-to-remove tests.

Data Structure

The data are entered in two or more columns. An example of data appropriate for this procedure is shown below. These data are from a study of the relationship of several variables with a person's I.Q. 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 variables being analyzed are ignored. If data are present on a row for all but the dependent variable, a predicted value and confidence limits are generated for that row.

Procedure Options

This section describes the options available in this procedure.

Variables, Model Tab

This panel specifies the variables used in the analysis.

Dependent Variable

Y

This option specifies one or more dependent (*Y*) variables. If more than one variable is specified, a separate analysis is run for each.

Independent Variables

Numeric X's

Specify numeric independent (also called regressor, explanatory, or predictor) variables here. Numeric variables are those whose values are numeric and are at least ordinal. Nominal variables, even when coded with numbers, should be specified as Categorical Independent Variables. Although you may specify binary (0-1) variables here, they are more appropriately analyzed when you specify them as Categorical X's.

If you want to create powers and cross-products of these variables, specify an appropriate model below in the Regression Model section.

If you want to create predicted values of *Y* for values of *X* not in your database, add the *X* new values as rows to the bottom of the database, with the value of Y blank. These rows will not be used during estimation phase, but predicted values will be generated for them on the reports.

Categorical X's

Specify categorical (nominal or group) independent variables in this box. By categorical we mean that the variable has only a few unique, numeric or text, values like 1, 2, 3 or Yes, No, Maybe. The values are used to identify categories.

Regression analysis is only defined for numeric variables. Since categorical variables are nominal, they cannot be used directly in regression. Instead, an internal set of numeric variables must be substituted for each categorical variable.

Suppose a categorical variable has *G* categories. ***NCSS*** automatically generates the *G*-1 internal, numeric variables for the analysis. The way these internal variables are created is determined by the Recoding Scheme and, if needed, the Reference Value. These options can be entered separately with each categorical variable, or they can specified using a default value (see Default Recoding Scheme and Default Reference Value below).

The syntax for specifying a categorical variable is *VarName*(*CType; RefValue*) where *VarName* is the name of the variable, CType is the recoding scheme, and RefValue is the reference value, if needed.

CType

The recoding scheme is entered as a letter. Possible choices are B, P, R, N, S, L, F, A, 1, 2, 3, 4, 5, or E. The meaning of each of these letters is as follows.

- **B** for **binary** (the group with the reference value is skipped).
  Example: Categorical variable Z with 4 categories. Category D is the reference value.

Z	B1	B2	B3
A	1	0	0
B	0	1	0
C	0	0	1
D	0	0	0

- **P** for **Polynomial** of up to 5th order (you cannot use this option with category variables with more than 6 categories.
  Example: Categorical variable Z with 4 categories.

Z	P1	P2	P3
1	-3	1	-1
3	-1	-1	3
5	1	-1	-3
7	3	1	1

- **R** to compare each with the **reference value** (the group with the reference value is skipped).
  Example: Categorical variable Z with 4 categories. Category D is the reference value.

Z	C1	C2	C3
A	1	0	0
B	0	1	0
C	0	0	1
D	-1	-1	-1

- **N** to compare each with the **next** category.
  Example: Categorical variable Z with 4 categories.

Z	S1	S2	S3
1	1	0	0
3	-1	1	0
5	0	-1	1
7	0	0	-1

- **S** to compare each with the **average of all subsequent** values.
  Example: Categorical variable Z with 4 categories.

Z	S1	S2	S3
1	-3	0	0
3	1	-2	0
5	1	1	-1
7	1	1	1

- **L** to compare each with the **prior** category.
  Example: Categorical variable Z with 4 categories.

Z	S1	S2	S3
1	-1	0	0
3	1	-1	0
5	0	1	-1
7	0	0	1

- **F** to compare each with the **average of all prior** categories.
  Example: Categorical variable Z with 4 categories.

Z	S1	S2	S3
1	1	1	1
3	1	1	-1
5	1	-2	0
7	-3	0	0

- **A** to compare each with the **average of all** categories (the Reference Value is skipped).
  Example: Categorical variable Z with 4 categories. Suppose the reference value is 3.

Z	S1	S2	S3
1	-3	1	1
3	1	1	1
5	1	-3	1
7	1	1	-3

- **1** to compare each with the **first** category after sorting.
  Example: Categorical variable Z with 4 categories.

Z	C1	C2	C3
A	-1	-1	-1
B	1	0	0
C	0	-1	0
D	0	0	-1

- **2** to compare each with the **second** category after sorting.
  Example: Categorical variable Z with 4 categories.

Z	C1	C2	C3
A	1	0	0
B	-1	-1	-1
C	0	1	0
D	0	0	1

- **3** to compare each with the **third** category after sorting.
  Example: Categorical variable Z with 4 categories.

Z	C1	C2	C3
A	1	0	0
B	0	1	0
C	-1	-1	-1
D	0	0	1

- **4** to compare each with the **fourth** category after sorting.
  Example: Categorical variable Z with 4 categories.

Z	C1	C2	C3
A	1	0	0
B	0	1	0
C	0	0	1
D	-1	-1	-1

- **5** to compare each with the **fifth** category after sorting.
  Example: Categorical variable Z with 5 categories.

Z	C1	C2	C3	C4
A	1	0	0	0
B	0	1	0	0
C	0	0	1	0
D	0	0	0	1
E	-1	-1	-1	-1

- **E** to compare each with the **last** category after sorting.
  Example: Categorical variable Z with 4 categories.

Z	C1	C2	C3
A	1	0	0
B	0	1	0
C	0	0	1
D	-1	-1	-1

RefValue
A second, optional argument is the reference value. The reference value is one of the categories. The other categories are compared to it, so it is usually a baseline or control value. If neither a baseline or control value is evident, the reference value is the most frequent value.

For example, suppose you want to include a categorical independent variable, State, which has four values: Texas, California, Florida, and NewYork. Suppose the recoding scheme is specified as *Compare Each with Reference Value* with the reference value of *California*. You would enter

State(R;California)

Default Recoding Scheme

Select the default type of numeric variable that will be generated when processing categorical independent variables. The values in a categorical variable are not used directly in regression analysis. Instead, a set of numeric variables is automatically created and substituted for them. This option allows you to specify what type of numeric variable will be created. The options are outlined in the sections below.

The contrast type may also be designated within parentheses after the name of each categorical independent variable, in which case the default contrast type is ignored.

If your model includes interactions of categorical variables, this option should be set to ‘Contrast with Reference’or Compare with All Subsequent' in order to match GLM results for factor effects.

- **Binary** (the group with the reference value is skipped).
  Example: Categorical variable Z with 4 categories. Category D is the reference value.

Z	B1	B2	B3
A	1	0	0
B	0	1	0
C	0	0	1
D	0	0	0

- **Polynomial** of up to 5th order (you cannot use this option with category variables with more than 6 categories.
  Example: Categorical variable Z with 4 categories.

Z	P1	P2	P3
1	-3	1	-1
3	-1	-1	3
5	1	-1	-3
7	3	1	1

- **Compare Each with Reference Value** (the group with the reference value is skipped).
  Example: Categorical variable Z with 4 categories. Category D is the reference value.

Z	C1	C2	C3
A	1	0	0
B	0	1	0
C	0	0	1
D	-1	-1	-1

- **Compare Each with Next**.
  Example: Categorical variable Z with 4 categories.

Z	S1	S2	S3
1	1	0	0
3	-1	1	0
5	0	-1	1
7	0	0	-1

- **Compare Each with All Subsequent**.
  Example: Categorical variable Z with 4 categories.

Z	S1	S2	S3
1	-3	0	0
3	1	-2	0
5	1	1	-1
7	1	1	1

- **Compare Each with Prior**
  Example: Categorical variable Z with 4 categories.

Z	S1	S2	S3
1	-1	0	0
3	1	-1	0
5	0	1	-1
7	0	0	1

- **Compare Each with All Prior**
  Example: Categorical variable Z with 4 categories.

Z	S1	S2	S3
1	1	1	1
3	1	1	-1
5	1	-2	0
7	-3	0	0

- **Compare Each with Average**
  Example: Categorical variable Z with 4 categories. Suppose the reference value is 3.

Z	S1	S2	S3
1	-3	1	1
3	1	1	1
5	1	-3	1
7	1	1	-3

- **Compare Each with First**
  Example: Categorical variable Z with 4 categories.

Z	C1	C2	C3
A	-1	-1	-1
B	1	0	0
C	0	-1	0
D	0	0	-1

- **Compare Each with Second**
  Example: Categorical variable Z with 4 categories.

Z	C1	C2	C3
A	1	0	0
B	-1	-1	-1
C	0	1	0
D	0	0	1

- **Compare Each with Third**
  Example: Categorical variable Z with 4 categories.

Z	C1	C2	C3
A	1	0	0
B	0	1	0
C	-1	-1	-1
D	0	0	1

- **Compare Each with Fourth**
  Example: Categorical variable Z with 4 categories.

Z	C1	C2	C3
A	1	0	0
B	0	1	0
C	0	0	1
D	-1	-1	-1

- **Compare Each with Fifth**
  Example: Categorical variable Z with 5 categories.

Z	C1	C2	C3	C4
A	1	0	0	0
B	0	1	0	0
C	0	0	1	0
D	0	0	0	1
E	-1	-1	-1	-1

- **Compare Each with Last**
  Example: Categorical variable Z with 4 categories.

Z	C1	C2	C3
A	1	0	0
B	0	1	0
C	0	0	1
D	-1	-1	-1

Default Reference Value

This option specifies the default reference value to be used when automatically generating indicator variables during the processing of selected categorical independent variables. The reference value is often the baseline, and the other values are compared to it. The choices are

- **First Value after Sorting – Fifth Value after Sorting**
  Use the first (through fifth) value in alpha-numeric sorted order as the reference value.

- **Last Value after Sorting**
  Use the last value in alpha-numeric sorted order as the reference value.

Weight Variable

Weights

When used, this is the name of a variable containing observation weights for generating a weighted-regression analysis. These weight values should be non-negative.

Regression Model

These options control which terms are included in the pool of candidate variables from which the subset is selected.

Terms

This option specifies which terms (terms, powers, cross-products, and interactions) are included in the candidate pool.

The options are

- **Up to 1-Way**
  This option generates a candidate pool in which each variable is represented by a single model term. No cross-products, interactions, or powers are added. Use this option when you want to use the variables you have specified, but you do not want to generate other terms.

  This is the option to select when you want to analyze the independent variables specified without adding any other terms.

  For example, if you have three independent variables A, B, and C, this would generate the candidate pool:

  A, B, C

- **Up to 2-Way**
  This option specifies that all individual variables, two-way interactions, and squares of numeric variables are included in the candidate pool. For example, if you have three numeric variables A, B, and C, this would generate the candidate pool:

  A, B, C, A*B, A*C, B*C, A*A, B*B, C*C

On the other hand, if you have three categorical variables A, B, and C, this would generate the candidate pool:

A, B, C, A*B, A*C, B*C

- **Up to 3-Way**
  All individual variables, two-way interactions, three-way interactions, squares of numeric variables, and cubes of numeric variables are included in the candidate pool. For example, if you have three numeric, independent variables A, B, and C, this would generate the candidate pool:

  A, B, C, A*B, A*C, B*C, A*B*C, A*A, B*B, C*C, A*A*B, A*A*C, B*B*C,A*C*C, B*C*C

  On the other hand, if you have three categorical variables A, B, and C, this would generate the candidate pool:

  A, B, C, A*B, A*C, B*C, A*B*C

- **Up to 4-Way**
  All individual variables, two-way interactions, three-way interactions, and four-way interactions are included in the candidate pool. Also included would be squares, cubes, and quartics of numeric variables and their cross-products.

  For example, if you have four categorical variables A, B, C, and D, this would generate the candidate pool:

  A + B + C + D + A*B + A*C + A*D + B*C + B*D + C*D + A*B*C + A*B*D + A*C*D + B*C*D + A*B*C*D

- **Interaction**
  Mainly used for categorical variables. A saturated model (all terms and their interactions) is generated. This requires a dataset with no missing categorical-variable combinations (you can have unequal numbers of observations for each combination of the categorical variables). No squares, cubes, etc. are generated.

  For example, if you have three independent variables A, B, and C, this would generate the candidate pool:

  A, B, C, A*B, A*C, B*C, A*B*C

  Note that the discussion of the Custom option discusses the interpretation of this candidate pool.

- **Custom**
  The candidate pool specified in the *Custom* box is used.

Remove Intercept

Unchecked indicates that the intercept term, β_0, 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^2, etc.). In most situations, you should include the intercept in the model.

Replace Custom Model with Preview Model (button)

When this button is pressed, the Custom Model is cleared and a copy of the Preview model is stored in the Custom Model. You can then edit this Custom Model as desired.

Maximum Order of Custom Terms

This option specifies that maximum number of variables that can occur in an interaction (or cross-product) term in a custom candidate pool. For example, A*B*C is a third order interaction term and if this option were set to 2, the A*B*C term would not be included in the candidate pool.

This option is particularly useful when used with the bar notation of a custom candidate pool to allow a simple way to remove unwanted high-order interactions.

Custom

This option specifies a custom candidate pool. It is only used when the *Terms* option is set to *Custom*. A custom model specifies the terms (single variables, cross-products, and interactions) that are to be kept in the model.

Interactions
An interaction expresses the combined relationship between two or more variables and the dependent variable by creating a new variable that is the product of the variables. The interaction (cross-product) between two numeric variables is generated by multiplying them. The interaction between to categorical variables is generated by multiplying each pair of internal variables. The interaction between a numeric variable and a categorical variable is created by generating all products between the numeric variable and the generated, numeric variables.

Syntax
A candidate pool is written by listing one or more terms. The terms are separated by a blank or plus sign. Terms include variables and interactions. Specify regular variables (main effects) by entering the variable names. Specify interactions by listing each variable in the interaction separated by an asterisk (*), such as Fruit*Nuts or A*B*C.

You can use the bar (|) symbol as a shorthand technique for specifying many interactions quickly. When several variables are separated by bars, all of their interactions are generated. For example, A|B|C is interpreted as A, B, C, A*B, A*C, B*C, A*B*C.

You can use parentheses. For example, A*(B+C) is interpreted as A*B, A*C.

Some examples will help to indicate how the model syntax works:

A|B = A, B, A*B

A|B A*A B*B = A, B, A*B, A*A, B*B

Note that you should only repeat numeric variables. That is, A*A is valid for a numeric variable, but not for a categorical variable.

A|A|B|B (Max Term Order=2) = A, B, A*A, A*B, B*B

A|B|C = A, B, C, A*B, A*C, B*C, A*B*C

(A + B)*(C + D) = A*C, A*D, B*C, B*D

(A + B)|C = (A + B) + C + (A + B)*C = A, B, C, A*C, B*C

Subset Search Options

Search Method

This option specifies the subset selection algorithm used to reduce the number of independent variables used in the regression model. The *Forward* algorithm is much quicker than the *Forward with Switching* algorithm, but the *Forward* algorithm does not usually find as good of a model.

Also note that in the case of categorical independent variables, the algorithm searches among the original categorical variables, not among the individual generated variables. That is, either all numeric variables associated with a particular categorical variable are included or not—they are not considered individually.

Hierarchical models are such that if an interaction is in the model, so are the terms that can be derived from it. For example, if A*B*C is in the model, so are A, B, C, A*B, A*C, and B*C. Statisticians usually adopt hierarchical models rather than non-hierarchical models. The subset selection procedure can be made to consider only hierarchical models during its search.

The subset selection options are:

- **(Hierarchical) Forward Selection**

  With this algorithm, the term that adds the most to R^2 is entered into the model. Next, the term that increases the R^2 the most is added. This selection is continued until all the terms have been entered or until the maximum subset size has been reach.

  If hierarchical models are selected, only those terms that will keep the model hierarchical are candidates for selection. For example, the interaction term A*B will not be considered unless both A and B are already in the model.

  When using this algorithm, you must make one run that allows a large number of terms to find the appropriate number of terms. Next, a second run is made in which you decrease the maximum terms in the subset to the number after which the R^2 does not change significantly.

- **(Hierarchical) Forward Selection with Switching**

  This algorithm is similar to the Forward algorithm described above. The term with the largest R^2 is entered into the regression model. The term which increases largest R^2 the most when combined with the first term is entered next. Now, each term in the current model is removed and the rest of the terms are checked to determine if, when they are used instead, R^2 is increased. If a term can be found by this switching process, the switch is made and the whole switching operation is begun again. The algorithm continues until no term can be found that improves the likelihood. This model then becomes the best two-term model.

  Next, the subset size is increased by one, the best third term is entered into the model, and the switching process is repeated. This process is repeated until the maximum subset size is reached. Hence, this model finds the optimum subset for each subset size. You must make one run to find an appropriate subset size by looking at the change in R^2. You then reset the maximum subset size to this value and rerun the analysis.

  If hierarchical models are selected, only those terms that will keep the model hierarchical are candidates for addition or deletion. For example, the interaction term A*B will not be considered unless both A and B are already in the model. Likewise, the term A cannot be removed from a model that contains A*B.

Stop Search when Number of Terms Reaches

Once this number of terms has been entered into the model, the subset selection algorithm is terminated. Often you will have to run the procedure twice to find an appropriate value. You would set this value high for the first run and then reset it appropriately for the second run, depending upon the values of R^2.

Note that the intercept is not counted in this number.

Reports Tab

These options control which reports are displayed. Note that many of these reports are only needed is special situations. You will only need a few reports for a typical regression analysis.

Alphas and Confidence Levels

Test Alpha

Alpha is the significance level used in conducting the hypothesis tests. The value of 0.05 is usually used. This corresponds to a chance of 1 out of 20. However, you should not be afraid to use other values since 0.05 became popular in pre-computer days when it was the only value available.

Typical values range from 0.01 to 0.20.

Assumptions Alpha

This value specifies the significance level that must be achieved to reject a preliminary test of an assumption. In regular hypothesis tests, common values of alpha are 0.05 and 0.01. However, most statisticians recommend that preliminary tests of assumptions use a larger alpha such as 0.10, 0.15, or 0.20.

We recommend 0.20.

Confidence Level

Enter the confidence level (or confidence coefficient) as a percentage for the confidence intervals reported. The interpretation of confidence level is that if confidence intervals are constructed across many experiments at the same confidence level, the percentage of such intervals that surround the true value of the parameter is equal to the confidence level.

Typical values range from 80 to 99.99. Usually, 95 is used.

Select Reports

Check those reports that you want to see. If you want to see a common set of reports, click the *Uncheck All* button followed by the *Common Set* button.

Show All Rows

This option makes it possible to display fewer observations in the row-by-row lists. This is especially useful when you have a lot of observations.

Bootstrap Calculation Options – Sampling

If you check the *Bookstrap CI's* report, additional options will appear at the bottom of the page.

Confidence Levels

These are the confidence coefficients of the bootstrap confidence intervals. They are entered as percentages. All values must be between 50 and 99.99. For example, you might enter *90 95 99*.

You may enter several values, separated by blanks or commas. A separate confidence interval is generated for each value entered.

Samples

This is the number of bootstrap samples used. A general rule of thumb is that you use at least 100 when standard errors are your focus or at least 1000 when confidence intervals are your focus. If computing time is available, it does not hurt to do 4000 or 5000.

We recommend setting this value to at least 3000.

Retries

If the results from a bootstrap sample cannot be calculated, the sample is discarded and a new sample is drawn in its place. This parameter is the number of times that a new sample is drawn before the algorithm is terminated. We recommend setting the parameter to at least 50.

Percentile Type

The method used to create the percentiles when forming bootstrap confidence limits. You can read more about the various types of percentiles in the Descriptive Statistics chapter. We suggest you use the *Ave X(p[n+1])* option.

C.I. Method

This option specifies the method used to calculate the bootstrap confidence intervals. The reflection method is recommended.

- **Percentile**
  The confidence limits are the corresponding percentiles of the bootstrap values.

- **Reflection**
  The confidence limits are formed by reflecting the percentile limits. If *X0* is the original value of the parameter estimate and *XL* and *XU* are the percentile confidence limits, the reflection interval is (2 *X0* - *XU*, 2 *X0* - *XL*).

Format Tab

Variable Labels

Precision

This option is used when the number of decimal places is set to ***All***. It specifies whether numbers are displayed as single (7-digit) or double (13-digit) precision numbers in the output. All calculations are performed in double precision regardless of the Precision selected here.

Single

Unformatted numbers are displayed with 7-digits

Double

Unformatted numbers are displayed with 13-digits. This option is most often used when the extremely accurate results are needed for further calculation. For example, double precision might be used when you are going to use the Multiple Regression model in a transformation.

Double Precision Format Misalignment

Double precision numbers may require more space than is available in the output columns, causing column alignment problems. The double precision option is for those instances when accuracy is more important than format alignment.

Variable Names

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

Stagger label and output if label length is ≥

When writing a row of information to a report, some variable names/labels may be too long to fit in the space allocated. If the name (or label) contains more characters than specified here, the rest of the output for that line is moved down to the next line. Most reports are designed to hold a label of up to 15 characters.

Enter ***1*** when you always want each row's output to be printed on two lines.

Enter ***100*** when you want each row printed on only one line. Note that this may cause some columns to be miss-aligned.

Decimal Places

Probability ... Mean Square Decimals

Specify the number of digits after the decimal point to display on the output of values of this type. This option in no way influences the accuracy with which the calculations are done.

All

Select ***All*** to display all digits available. The number of digits displayed by this option is controlled by whether the *Precision* option is *Single* (7) or *Double* (13).

Plots Tab

These options control the inclusion and the settings of each of the plots.

Select Plots

Histogram ... Partial Resid vs X Plot

Indicate whether to display these plots. Click the plot format button to change the plot settings.

Edit During Run

This is the small check-box in the upper right-hand corner of the format button. If checked, the graphics format window for this plot will be displayed while the procedure is running so that you can format it with the actual data.

Storage Tab

These options let you specify if, and where on the dataset, various statistics are stored.

Warning: Any data already in these variables are replaced by the new data. Be careful not to specify columns that contain important data.

Data Storage Options

Storage Option

This option controls whether the values indicated below are stored on the dataset when the procedure is run.

- **Do not store data**
  No data are stored even if they are checked.
- **Store in empty columns only**
  The values are stored in empty columns only. Columns containing data are not used for data storage, so no data can be lost.
- **Store in designated columns**
  Beginning at the *First Storage Variable*, the values are stored in this column and those to the right. If a column contains data, the data are replaced by the storage values. Care must be used with this option because it cannot be undone.

Store First Item In

The first item is stored in this column. Each additional item that is checked is stored in the columns immediately to the right of this variable.

Leave this value blank if you want the data storage to begin in the first blank column on the right-hand side of the data.

Warning: any existing data in these columns is automatically replaced, so be careful.

Data Storage Options – Select Items to Store

Predicted Y ... VC(Betas) Matrix

Indicate whether to store these row-by-row values, beginning at the column indicated by the *Store First Variable In* option.

Example 1 – Subset Selection in Multiple Regression (All Reports)

This section presents an example of how to run a subset selection from the data presented earlier in this chapter. The data are in the IQ dataset. This example will find a subset of three IV's from the candidate pool of *Test1* through *Test5*. The dependent variable is IQ. This program outputs over thirty different reports and plots, many of which contain duplicate information. Only those reports that are specifically needed for a robust regression will be present here.

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 Subset Selection in Multiple 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 procedure window.

- Using the Analysis menu or the Procedure Navigator, find and select the **Subset Selection in Multiple Regression** procedure.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- Select the **Variables, Model tab**.
- Set the **Y** box to **IQ**.
- Set the **Numeric X's** box to **Test1-Test5**.
- Set the **Terms** box to **1-Way**.
- Set the **Stop search when number of terms reaches** box to **3**.

4 Specify the reports.

- Select the **Variables, Model tab**.
- In addition to the reports already checked, check **Correlations**, **Subset Summary**, **Subset Detail**, and **Residuals**.
- Uncheck **ANOVA Summary** and **Predicted for Individuals**.

5 Run the procedure.

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

Run Summary Section

Item	Value	Rows	Value
Dependent Variable	IQ	Rows Processed	17
Number Ind. Variables	3	Rows Filtered Out	0
Weight Variable	None	Rows with X's Missing	0
R^2	0.1591	Rows with Weight Missing	0
Adj R^2	0.0000	Rows with Y Missing	2
Coefficient of Variation	0.1092	Rows Used in Estimation	15
Mean Square Error	129.91	Sum of Weights	15.000
Square Root of MSE	11.40		
Ave Abs Pct Error	6.933		
Completion Status	Normal Completion		

This report summarizes the results. It presents the variables used, the number of rows used, and some basic results.

R^2

R^2, officially known as the coefficient of determination, is defined as

$$R^2 = \frac{SS_{Model}}{SS_{Total(Adjusted)}}$$

R^2 is probably the most popular statistical measure of how well the regression model fits the data. R^2 may be defined either as a ratio or a percentage. Since we use the ratio form, its values range from zero to one. A value of R^2 near zero indicates no linear relationship between the Y and the X's, while a value near one indicates a perfect linear fit. Although popular, R^2 should not be used indiscriminately or interpreted without scatter plot support. Following are some qualifications on its interpretation:

1. *Additional independent variables*. It is possible to increase R^2 by adding more independent variables, but the additional independent variables may actually cause an increase in the mean square error, an unfavorable situation. This case happens when your sample size is small.
2. *Range of the independent variables*. R^2 is influenced by the range of each independent variable. R^2 increases as the range of the X's increases and decreases as the range of the X's decreases.
3. *Slope magnitudes*. R^2 does not measure the magnitude of the slopes.
4. *Linearity*. R^2 does not measure the appropriateness of a linear model. It measures the strength of the linear component of the model. Suppose the relationship between x and Y was a perfect circle. The R^2 value of this relationship would be zero.
5. *Predictability*. A large R^2 does not necessarily mean high predictability, nor does a low R^2 necessarily mean poor predictability.
6. *No-intercept model*. The definition of R^2 assumes that there is an intercept in the regression model. When the intercept is left out of the model, the definition of R^2 changes dramatically. The fact that your R^2 value increases when you remove the intercept from the regression model does not reflect an increase in the goodness of fit. Rather, it reflects a change in the underlying meaning of R^2.
7. *Sample size*. R^2 is highly sensitive to the number of observations. The smaller the sample size, the larger its value.

Adjusted R^2

This is an adjusted version of R^2. The adjustment seeks to remove the distortion due to a small sample size.

Coefficient of Variation

The coefficient of variation is a relative measure of dispersion, computed by dividing root mean square error by the mean of the dependent variable. By itself, it has little value, but it can be useful in comparative studies.

$$CV = \frac{\sqrt{MSE}}{\bar{y}}$$

Ave Abs Pct Error

This is the average of the absolute percent errors. It is another measure of the goodness of fit of the regression model to the data. It is calculated using the formula

$$AAPE = \frac{100\sum_{j=1}^{N}\left|\frac{y_j - \hat{y}_j}{y_j}\right|}{N}$$

Note that when the dependent variable is zero, its predicted value is used in the denominator.

Descriptive Statistics Section

Variable	Count	Mean	Standard Deviation	Minimum	Maximum
Test1	15	67.933	17.392	37	96
Test2	15	61.400	19.397	19	89
Test3	15	72.333	14.734	43	96
Test4	15	65.533	13.953	39	88
Test5	15	69.933	16.153	42	94
IQ	15	104.333	11.017	92	130

For each variable, the count, arithmetic mean, standard deviation, minimum, and maximum are computed. This report is particularly useful for checking that the correct variables were selected.

Correlation Matrix Section

	Test1	Test2	Test3	Test4
Test1	1.0000	0.1000	-0.2608	0.7539
Test2	0.1000	1.0000	0.0572	0.7196
Test3	-0.2608	0.0572	1.0000	-0.1409
Test4	0.7539	0.7196	-0.1409	1.0000
Test5	0.0140	-0.2814	0.3473	-0.1729
IQ	0.2256	0.2407	0.0741	0.3714

	Test5	IQ
Test1	0.0140	0.2256
Test2	-0.2814	0.2407
Test3	0.3473	0.0741
Test4	-0.1729	0.3714
Test5	1.0000	-0.0581
IQ	-0.0581	1.0000

Pearson correlations are given for all variables. Outliers, nonnormality, nonconstant variance, and nonlinearities can all impact these correlations. Note that these correlations may differ from pair-wise correlations generated by the correlation matrix program because of the different ways the two programs treat rows with missing values. The method used here is row-wise deletion.

These correlation coefficients show which independent variables are highly correlated with the dependent variable and with each other. Independent variables that are highly correlated with one another may cause collinearity problems.

Subset Selection Summary

No. Terms	No. X's	R^2 Value	R^2 Change
1	1	0.1379	0.1379
2	2	0.1542	0.0163
3	3	0.1591	0.0049

This report shows the number of terms, number of IV's, and R^2 values for each subset size. This report is used to determine an appropriate subset size for a second run. You search the table for a subset size after which the R^2 increases only slightly as more variables are added.

Subset Selection Detail Section

Step	Action	No. of Terms	No. of X's	R^2	Term Entered	Term Removed
0	Add	0	0	0.0000	Intercept	
1	Add	1	1	0.1379	Test4	
2	Add	2	2	0.1542	Test3	
3	Add	3	3	0.1591	Test2	

This report shows the details of which variables were added or removed at each step in the search procedure. The final model for three IV's would include Test2, Test3, and Test4.

Because of the restrictions due to our use of hierarchical models, you might run an analysis using the Forward with Switching option as well as a search without 2-way interactions. Because of the small sample size, these options produce models with much larger *R*-squared values. However, it is our feeling that this larger *R*-squared values occur because the extra variables are actually fitting random error rather than a reproducible pattern.

Regression Coefficients T-Tests

Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)	Standard-ized Coefficient	T-Statistic to Test H0: β(i)=0	Prob Level	Reject H0 at 5%?	Power of Test at 5%
Intercept	75.93027	23.076055	0.0000	3.290	0.0072	Yes	0.8492
Test2	-0.05859	0.232438	-0.1032	-0.252	0.8056	No	0.0561
Test3	0.10893	0.214619	0.1457	0.508	0.6218	No	0.0751
Test4	0.36808	0.325849	0.4662	1.130	0.2827	No	0.1784

This report gives the coefficients, standard errors, and significance tests.

Independent Variable

The names of the independent variables are listed here. The intercept is the value of the *Y* intercept.

Regression Coefficient b(i)

The regression coefficients are the least squares estimates of the parameters. The value indicates how much change in *Y* occurs for a one-unit change in that particular X when the remaining *X's* are held constant. These coefficients are often called partial-regression coefficients since the effect of the other *X's* is removed. These coefficients are the values of $b_0, b_1, \cdots, b_p$.

Standard Error Sb(i)

The standard error of the regression coefficient, s_{b_j}, is the standard deviation of the estimate. It is used in hypothesis tests or confidence limits.

Standardized Coefficient

Standardized regression coefficients are the coefficients that would be obtained if you standardized the independent variables and the 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 the independent variables have vastly different scales of measurement, this value provides a way of making comparisons among 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 j^{th} independent variable.

T-Statistic to test H0: β(i)=0

This is the t-test value for testing the hypothesis that $\beta_j = 0$ versus the alternative that $\beta_j \neq 0$ after removing the influence of all other *X's*. This *t*-value has *n*-*p*-1 degrees of freedom.

Prob Level

This is the *p*-value for the significance test of the regression coefficient. The *p*-value is the probability that this *t*-statistic will take on a value at least as extreme as the actually observed value, assuming that the null hypothesis is true (i.e., the regression estimate is equal to zero). If the *p*-value is less than alpha, say 0.05, the null hypothesis of equality is rejected. This *p*-value is for a two-tail test.

Reject H0 at 5%?

This is the conclusion reached about the null hypothesis. It will be either reject *H0* at the 5% level of significance or not.

Note that the level of significance is specified in the Test Alpha box on the Format tab panel.

Power (5%)

Power is the probability of rejecting the null hypothesis that $\beta_j = 0$ when $\beta_j = \beta_j^* \neq 0$. The power is calculated for the case when $\beta_j^* = b_j$, $\sigma^2 = s^2$, and alpha is as specified in the Alpha of C.I.'s and Tests option.

High power is desirable. High power means that there is a high probability of rejecting the null hypothesis that the regression coefficient is zero when this is false. This is a critical measure of sensitivity in hypothesis testing. This estimate of power is based upon the assumption that the residuals are normally distributed.

Regression Coefficients Confidence Intervals

Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)	Lower 95% Conf. Limit of β(i)	Upper 95% Conf. Limit of β(i)
Intercept	75.93027	23.076055	25.14022	126.72033
Test2	-0.05859	0.232438	-0.57018	0.45300
Test3	0.10893	0.214619	-0.36345	0.58130
Test4	0.36808	0.325849	-0.34911	1.08526

Note: The T-Value used to calculate these confidence limits was 2.201.

This report gives the coefficients, standard errors, and confidence interval.

Independent Variable

The names of the independent variables are listed here. The intercept is the value of the *Y* intercept.

Regression Coefficient

The regression coefficients are the least squares estimates of the parameters. The value indicates how much change in *Y* occurs for a one-unit change in *x* when the remaining *X's* are held constant. These coefficients are often called partial-regression coefficients since the effect of the other *X's* is removed. These coefficients are the values of $b_0, b_1, \cdots, b_p$.

Standard Error Sb(i)

The standard error of the regression coefficient, s_{b_j}, is the standard deviation of the estimate. It is used in hypothesis tests and confidence limits.

Lower - Upper 95% Conf. Limit of β(i)

These are the lower and upper values of a $100(1-\alpha)\%$ interval estimate for β_j based on a *t*-distribution with *n-p*-1 degrees of freedom. This interval estimate assumes that the residuals for the regression model are normally distributed.

The formulas for the lower and upper confidence limits are:

$$b_j \pm t_{1-\alpha/2,n-p-1}\ s_{b_j}$$

Note: The T-Value …

This is the value of $t_{1-\alpha/2,n-p-1}$ used to construct the confidence limits.

Estimated Equation

IQ =
75.9302747014515 - 0.0585872131040306 * Test2 + 0.108927169070947 * Test3 + 0.368076016587041 * Test4

This is the least squares regression line presented in double precision. Besides showing the regression model in long form, it may be used as a transformation by copying and pasting it into the Transformation portion of the spreadsheet.

Analysis of Variance Detail

Source	DF	R²	Sum of Squares	Mean Square	F-Ratio	Prob Level	Power (5%)
Intercept	1		163281.67	163281.67			
Model	3	0.1591	270.37	90.12	0.694	0.5748	0.1523
Test2	1	0.0049	8.25	8.25	0.064	0.8056	0.0561
Test3	1	0.0197	33.46	33.46	0.258	0.6218	0.0751
Test4	1	0.0975	165.76	165.76	1.276	0.2827	0.1784
Error	11	0.8409	1428.96	129.91			
Total(Adjusted)	14	1.0000	1699.33	121.38			

This analysis of variance table provides a line for each term in the model. It is especially useful when you have categorical independent variables.

Source

This is the term from the design model.

DF

This is the number of degrees of freedom that the model is degrees of freedom is reduced when this term is removed from the model. This is the numerator degrees of freedom of the *F-test*.

R^2

This is the amount that R^2 is reduced when this term is removed from the regression model.

Sum of Squares

This is the amount that the model sum of squares that are reduced when this term is removed from the model.

Mean Square

The mean square is the sum of squares divided by the degrees of freedom.

F-Ratio

This is the *F*-statistic for testing the null hypothesis that all β_j associated with this term are zero. This *F-statistic* has *DF* and *n*-*p*-1 degrees of freedom.

Prob Level

This is the *p*-value for the above *F*-test. The *p*-value is the probability that the test statistic will take on a value at least as extreme as the observed value, assuming that the null hypothesis is true. If the *p*-value is less than α, say 0.05, the null hypothesis is rejected. If the *p*-value is greater than α, then the null hypothesis is accepted.

Power(5%)

Power is the probability of rejecting the null hypothesis that all the regression coefficients associated with this term are zero, assuming that the estimated values of these coefficients are their true values.

Normality Tests Section

Test Name	Test Statistic to Test H0: Normal	Prob Level	Reject H0 at 20%?
Shapiro Wilk	0.896	0.0833	Yes
Anderson Darling	0.593	0.1220	Yes
D'Agostino Skewness	2.274	0.0230	Yes
D'Agostino Kurtosis	1.765	0.0775	Yes
D'Agostino Omnibus	8.287	0.0159	Yes

This report gives the results of applying several normality tests to the residuals. The Shapiro-Wilk test is probably the most popular, so it is given first. These tests are discussed in detail in the Normality Test section of the Descriptive Statistics procedure.

Residual Report

Row	Actual IQ	Predicted IQ	Residual	Absolute Percent Error	Sqrt(MSE) Without This Row
1	106.000	104.207	1.793	1.69	11.93
2	92.000	100.436	-8.436	9.17	11.42
3	102.000	104.042	-2.042	2.00	11.93
4	121.000	114.015	6.985	5.77	11.51
5	102.000	105.697	-3.697	3.62	11.89
6	105.000	107.652	-2.652	2.53	11.92
7	97.000	104.666	-7.666	7.90	11.68
8	92.000	96.883	-4.883	5.31	11.76
9	94.000	97.787	-3.787	4.03	11.85
10	112.000	100.891	11.109	9.92	11.14
11	130.000	103.301	26.699	20.54	7.29
12	115.000	105.789	9.211	8.01	11.55
13	98.000	107.722	-9.722	9.92	11.47
14	96.000	108.974	-12.974	13.51	10.97
15	103.000	102.936	0.064	0.06	11.95
16		96.851			
17		101.035			

This section reports on the sample residuals, or e_i's.

Actual

This is the actual value of *Y*.

Predicted

The predicted value of *Y* using the values of the IV's given on this row.

Residual

This is the error in the predicted value. It is equal to the *Actual* minus the *Predicted.*

Absolute Percent Error

This is percentage that the absolute value of the *Residual* is of the *Actual* value. Scrutinize rows with the large percent errors.

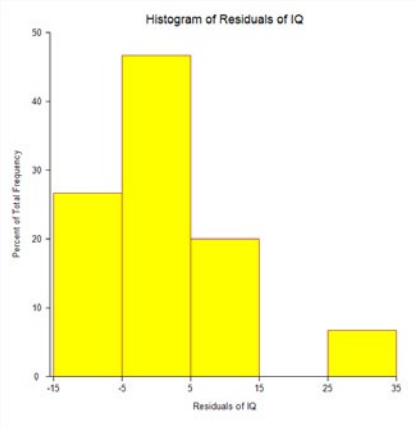
Sqrt(MSE) Without This Row

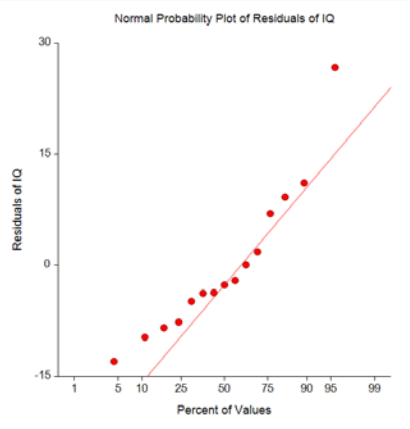
This is the value of the square root of the mean square error that is obtained if this row is deleted. A perusal of this statistic for all observations will highlight observations that have an inflationary impact on mean square error and could be outliers.

Plots of Residuals

These plots let you assess the residuals. Any nonrandom pattern may require a redefining of the regression model.

Distributional Plots of Residuals





Residuals vs X's Plots