Chapter 330

Response Surface Regression

Introduction

This *Response Surface Analysis* (RSA) program fits a polynomial regression model with cross-product terms of variables that may be raised up to the third power. It calculates the minimum or maximum of the surface. The program also has a variable selection feature that helps you find the most parsimonious hierarchical model. **NCSS** automatically scans the data for duplicates so that a lack-of-fit test may be calculated using pure error.

One of the main goals of RSA is to find a polynomial approximation of the true nonlinear model, similar to the Taylor's series expansion used in calculus. Hence, you are searching for an approximation that works well in a specified region. As the region is reduced, the number of terms may also be reduced. In a very small region, a linear (first order) approximation may be adequate. A larger region may require a quadratic (second order) approximation.

Hierarchical Models

In the following discussion, the X's are independent variables with at least three distinct values (up to six X's may be specified). Y is the dependent variable. Z is a covariate (note that covariates do not have to have three or more levels). The β's are the regression coefficients or beta weights.

A polynomial model is one in which the X's occur as multiples of each other. Examples of polynomial models are:

$$Y_j = ß_0 + ß_1 X_{1j} + \varepsilon_j$$

$$Y_j = ß_0 + ß_1 X_{1j} + ß_2 X_{2j} + ß_3 X_{1j} X_{2j} + \varepsilon_j$$

$$Y_j = ß_0 + ß_1 X_{1j} + ß_2 X_{2j}^2 + ß_3 X_{1j}^2 X_{2j}^3 + \varepsilon_j$$

A hierarchical model obeys the following rule: all lower-order terms that can be constructed by reducing the exponents of the variables in a term are also in the model. For example, if the term $X1X_2^2$ is in the model, so are X1, X2, X1X2, and X_2^2.Notice that each of these terms can be created be decreasing the exponents of the two variables that form the original term (noting that X0j = 1). Note that the first two models are hierarchical, but the third is not.

Hierarchical models enjoy several useful properties, including stability, the ability to change the scale (coding) of a variable, and a general relationship with ANOVA modeling. However, they usually require the fitting of more parameters than nonhierarchical models. This **NCSS** procedure fits only hierarchical models. If nonhierarchical models are desired, they can be fit using the Multiple Regression module.

Model Selection

There are several strategies to variable selection and model building in regression analysis: forward selection, backward elimination, stepwise, all possible regressions, and more. However, none of these methods guarantee hierarchical models. We need a method that does. This **NCSS** program adopts a strategy that has been used for quite a while in dealing with hierarchical models. The strategy may be outlined as follows:

1. Begin with the most complicated model desired. **NCSS** allows terms of the form $X_1^i X_2^j$, where i and j are each less than or equal to three.
2. Search through all terms, marking those that are not necessary to maintain the hierarchical constraint on the model. This group of terms is available for removal.
3. Check each of the available terms to determine how much R-Squared is decreased if they are removed.
4. Remove the term that decreases R-Squared the least. Return to step 2. Note that this variable is never reconsidered for inclusion in the model.
5. If no available term can be identified that reduces R-Squared by an amount that is less than the specified cutoff value, the model selection procedure is terminated.

Maximum Orders of Two-Way Terms

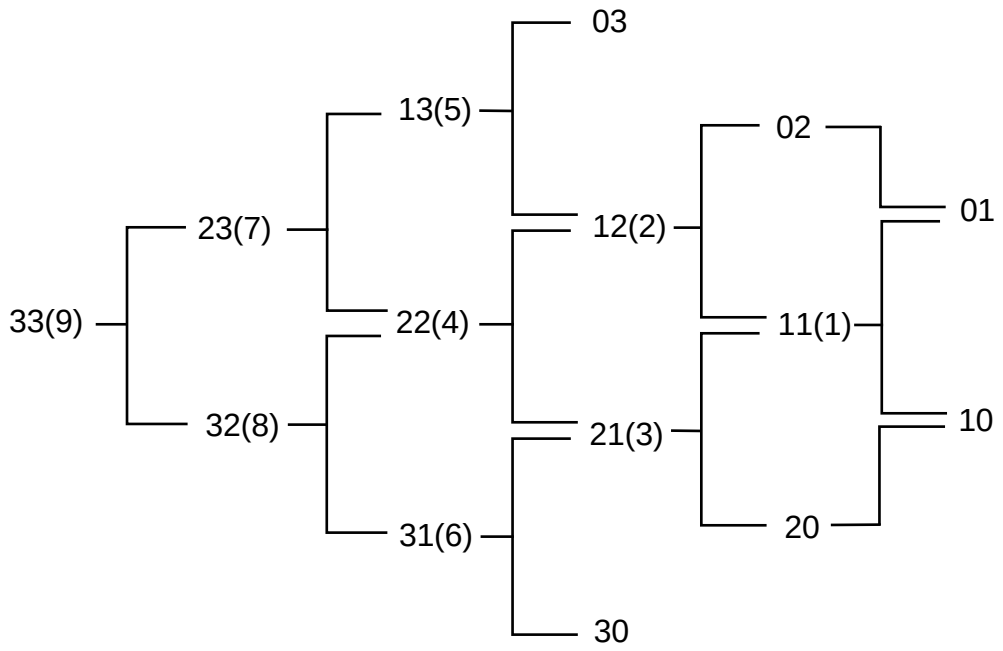
These boxes define the maximum exponent for each factor in the cross-product term of the corresponding variables. For example, “AC” represents the product of factors A and C.

All subset terms (children) are also included in the model, so that the hierarchical nature of the model is maintained.

A code is used to specify the maximum exponents of each term. Up to three of these may be needed to specify the desired hierarchical model. The following table relates each coded value to the terms that it generates. This table will be generated for the *AB* term. The pattern extends in an obvious manner to the other cross-products. Note that a “10” is used to represent A, a “01” represents B, a “20” represents A^2, and so on.

Code	Term(s)	Cross Product Terms Actually Included
1	11	AB
2	12	AB, AB2
3	21	AB, A2B
23	12,21	AB, AB2, A2B
4	22	AB, A2B, AB2, A2B2
5	13	AB, AB2, AB3
35	21,13	AB, A2B, AB2, AB3
45	22,13	AB, A2B, AB2, A2B2, AB3
6	31	AB, A2B, A3B
26	12,31	AB, AB2, A2B, A3B
46	22,31	AB, AB2, A2B, A2B2, A3B
56	13,31	AB, AB2, A2B, AB3, A3B
456	22,13,31	AB, AB2, A2B, AB3, A3B, A2B2
7	23	AB, A2B, AB2, A2B2, AB3, A2B3
67	31,23	AB, A2B, AB2, A2B2, AB3, A3B, A2B3
8	32	AB, AB2, A2B, A2B2, A3B, A3B2
58	13,32	AB, AB2, A2B, A2B2, AB3, A3B, A3B2
78	23,32	AB, AB2, A2B, A2B2, AB3, A3B, A2B3, A3B2
9	33	AB, AB2, A2B, A2B2, A3B, A3B2, AB3, A2B3, A3A3

The following tree diagram shows the hierarchical structure of this system. Each term generates all terms to the right of it. These terms are called *children.* A term that is a child of one term is not specified with that term. For example, terms 13, 22, and 31 could be selected together. However, the terms 23 and 21 could not be entered together since 21 is a child of 23 and will automatically be included when 23 is specified. Note the cross-product terms include their codes in parentheses. Also note that terms like 03 and 20 are specified in the One-Way Terms section.



The actual specification of the term is accomplished by selecting one or more codes from a list of possible models. For example, you might select “58.” This model represents the 13 and the 32 terms plus all their children.

The usual quadratic model is specified by selecting 2’s for the Order terms and 1’s for the Two-Way terms.

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 terms (including all cross products). A popular rule-of-thumb when using any variable selection procedure is that you have at least five observations for each term.

Unique Data Values

Since various powers of the variables are included, the structure of your data must allow for these powers to be fit. This means that if the maximum exponent on a variable is k, the number of unique values in that variable must be at least k+1. For example, suppose a variable consisted of two values: -1 and 1. You could not fit a model that included more than a linear (k=1) term in this variable. Again, suppose your data consisted of three values: -1,0,1. The maximum exponent that could be used with this variable is 2.

Data Structure

The data are entered in two or more columns. An example of data appropriate for this procedure is shown in the following table and is found in the Odor dataset. This dataset relates a measurement of odor to three variables in a chemical process. Fifteen rows of data were obtained. The values of the three independent variables have been recoded so that they are -1, 0, and 1. A sixteenth row has been added. Notice that it does not contain a value in the *Odor* column. A predicted value will be generated for this row, but its values will not be used in the estimation process.

We suggest that you open this dataset now so that you can follow along with the example.

Odor dataset

Odor	Temp	Ratio	Height
66	-1	-1	0
58	-1	0	-1
65	0	-1	-1
-31	0	0	0
39	1	-1	0
17	1	0	-1
7	0	1	-1
-35	0	0	0
43	-1	1	0
-5	-1	0	1
43	0	-1	1
-26	0	0	0
49	1	1	0
-40	1	0	1
-22	0	1	1
	1	1	0

Missing Values

Rows with missing values in the variables being analyzed are ignored. If data are present for all but the dependent variable, a predicted value is generated for this row.

Example 1 – Response Surface Analysis

This section presents an example of how to run a response surface analysis of the data contained in the Odor dataset.

Setup

To run this example, complete the following steps:

1 Open the Odor example dataset

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

2 Specify the Response Surface Regression procedure options

- Find and open the **Response Surface 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	**Odor**
Factor A Variable	**Temp**
Factor B Variable	**Ratio**
Factor C Variable	**Height**
Plots Tab	
Number of X Slices	**50**
Number of Y Slices	**50**

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	Minimum	Maximum
Temp	15	0	-1	1
Ratio	15	0	-1	1
Height	15	0	-1	1
Odor	15	15.2	-40	66

This report provides the count, mean, minimum, and maximum of each of the variables in the analysis. It allows you to determine if the data fall within reasonable limits.

Hierarchical Model Summary Section

Hierarchical Model Summary Section

Number of Terms Removed	0
Number of Terms Remaining	9
R-Squared Cutoff Value	0.010000
R-Squared of Final Model	0.881989

Coded Hierarchical Model

	A	B	C
A Temp	2	1(11)	1(11)
B Ratio		2	1(11)
C Height			2

Notes:
For off-diagonal entries:
1=u1w1, 2=u1w2, 3=u2w1, 4=u2w2, 5=u1w3,6=u3w1, 7=u2w3, 8=u3w2, 9=u3w3.
For diagonal entries:
1=u1, 2=u2, 3=u3.
Where u1=u, u2=u^2=u*u, and u3=u^3=u*u*u.

This report shows the hierarchical model that was specified. It also shows the final R-Squared value as well as the R-Squared cutoff that was used. It is mainly used to document the model used.

The specified model is determined by considering all nonzero entries. The *Notes* section at the bottom shows how the model is determined. For example, the first line is nonzero for the *terms AA, AB,* and *AC*. These codes represent the terms: $Temp^2$, (Temp)(Ratio), and (Temp)(Height). All child terms necessary to make this a hierarchical model are also generated.

Sequential ANOVA Section

Sequential ANOVA Section

Source	df	Sequential Sum-Squares	Mean Square	F-Ratio	Prob Level	Incremental R-Squared
Regression	9	18881.98	2097.998	4.15	0.065691	0.881989
Linear	3	7143.25	2381.083	4.71	0.064071	0.333666
Quadratic	3	11445.23	3815.078	7.55	0.026426	0.534614
Lin x Lin	3	293.5	97.83334	0.19	0.896470	0.013710
Total Error	5	2526.417	505.2833			0.118011
Lack of Fit	3	2485.75	828.5833	40.75	0.024047	0.116111
Pure Error	2	40.66667	20.33333			0.001900

Sequential ANOVA Section Using Pure Error

Source	df	Sequential Sum-Squares	Mean Square	F-Ratio	Prob Level	Incremental R-Squared
Regression	9	18881.98	2097.998	103.18	0.009635	0.881989
Linear	3	7143.25	2381.083	117.10	0.008479	0.333666
Quadratic	3	11445.23	3815.078	187.63	0.005306	0.534614
Lin x Lin	3	293.5	97.83334	4.81	0.176872	0.013710
Total Error	5	2526.417	505.2833			0.118011
Lack of Fit	3	2485.75	828.5833	40.75	0.024047	0.116111
Pure Error	2	40.66667	20.33333			0.001900

This display actually shows two reports. The top is the regular Sequential ANOVA Section defined below. Note that the denominator of the F-Ratios is the Total Error Mean Square. The bottom report is identical to the top, except that the denominator of the F-Ratios is now the Pure Error Mean Square.

This report is designed with two main goals:

1. Determine the sequential influence of the various power and cross-product terms.
2. Test for model lack of fit if repeated observations are available.

Source

The group of independent variables being tested.

Regression	Total of all terms in the model.
Linear	The total for X_i terms.
Quadratic	The total for X_i^2 terms.
Cubic	The total for X_i^3 terms.
Lin x Lin	The total for X_iX_j terms.
Lin x Quad	The total for $X_iX_j^2$ terms.
Quad x Quad	The total for $X_i^2X_j^2$ terms.
Lin x Cubic	The total for $X_iX_j^3$ terms.
Quad x Cubic	The total for $X_i^2X_j^3$ terms.
Cubic x Cubic	The total for $X_i^3X_j^3$ terms.

DF

The degrees of freedom associated with the group of terms.

Sequential Sum-Squares

The regression sum of squares added sequentially by each group of terms. Each group of terms adds this amount of sum of squares after accounting for the terms above it in the report.

Mean Square

The sum of squares divided by the degrees of freedom.

F-Ratio

The F-value formed by dividing the Mean Square by the Total Error Mean Square. Note that these tests are sequential in nature and should be considered from the bottom up. Note that in the second report, the Total Error Mean Square is replaced by the Pure Error Mean Square as the denominator of the F-ratio.

In the above example, the Lin x Lin F-ratio tests whether the linear-by-linear terms are significant in the regression model after considering the linear and quadratic terms. The Quadratic F-ratio tests whether the quadratic terms add significantly to a model consisting of the linear terms (ignoring the linear-by-linear terms).

In terms of the ODOR data, the tests are interpreted as follows:

Group	Terms	Hypothesis Tested
Lin x Lin	Temp x Ratio Temp x Height Ratio x Height	All coefficients of these variables are zero.
Quadratic	Temp x Temp Ratio x Ratio Height x Height	All coefficients of these variables are zero, ignoring the influence of the cross-product terms.
Linear	Temp Ratio Height	All coefficients of these variables are zero, ignoring the influence of the cross-product and quadratic terms

Prob Level

This is the right-tail probability or significance level of this test. Reject the hypothesis that the influence of the terms is zero when this value is less than a predetermined value of alpha, say 0.05.

Incremental R-Squared

The first line displays the total R-Squared for the complete model. The other lines display the amount of R-Squared that is added by each group of terms. Hence, the total of the rest of the lines equals the first.

Lack of Fit and Pure Error

These lines are only displayed if you have repeated observations from which the variability between identical observations may be estimated. The lack of fit tests the adequacy of the specified model. A significant F-test implies that a higher-order polynomial (such as cubic) or a different functional form would fit the data better.

If pure error is available, the F-tests are recalculated using the Pure Error Mean Square as the denominator rather than the Total Error Mean Square.

ANOVA Section

ANOVA Section

Factor	df	Last Sum-Squares	Mean Square	F-Ratio	Prob Level	Term R-Squared
Temp	4	5258.016	1314.504	2.60	0.161334	0.245605
Ratio	4	11044.6	2761.151	5.46	0.045377	0.515900
Height	4	3813.016	953.254	1.89	0.251025	0.178108
Total Error	5	2526.417	505.2833			0.118011
Lack of Fit	3	2485.75	828.5833	40.75	0.024047	0.116111
Pure Error	2	40.66667	20.33333			0.001900

ANOVA Section Using Pure Error

Factor	df	Last Sum-Squares	Mean Square	F-Ratio	Prob Level	Term R-Squared
Temp	4	5258.016	1314.504	64.65	0.015291	0.245605
Ratio	4	11044.6	2761.151	135.79	0.007324	0.515900
Height	4	3813.016	953.254	46.88	0.020994	0.178108
Total Error	5	2526.417	505.2833			0.118011
Lack of Fit	3	2485.75	828.5833	40.75	0.024047	0.116111
Pure Error	2	40.66667	20.33333			0.001900

This report tests the significance of each factor. This display actually shows two reports. The top is the regular ANOVA Section defined below. Note that the denominator of the F-Ratios is the Total Error Mean Square. The second report is identical to the top, except that the denominator of the F-Ratios is now the Pure Error Mean Square.

Factor

This line lists the factor being tested for deletion. All terms that include this factor are included in the test. In our example, the terms being tested are as follows:

Factor	Individual Terms Referred To
Temp	Temp, Temp x Ratio, Temp x Height, Temp x Temp.
Ratio	Ratio, Temp x Ratio, Ratio x Height, Ratio x Ratio.
Height	Height, Height x Ratio, Height x Temp, Height x Height.

Note that there is overlap in these terms (some cross-products occur twice).

DF

The degrees of freedom associated with the term(s).

Last Sum-Squares

The regression sum of squares that would be lost if this factor were omitted.

Mean Square

The sum of squares divided by the degrees of freedom.

F-Ratio

In the top report, the F-value is formed by dividing the Mean Square by the Total Error Mean Square. In the second report, the F-value is formed by dividing the Mean Square by the Pure Error Mean Square. Note that these tests are not sequential, but each tests the importance of the factor after considering all other factors.

Prob Level

This is the right-tail probability or significance level of this test. Reject the hypothesis that the influence of the terms is zero when this value is less than a predetermined value of alpha, say 0.05.

Term R-Squared

The amount that the R-Squared would decrease if this factor were removed from the model.

Lack of Fit / Pure Error

These lines are only displayed if you have repeated observations from which the variability between like observations may be estimated. The lack of fit tests the adequacy of the specified model. If this test is significant, conclude that a higher order polynomial (such as cubic), or a different functional form, would fit the data better.

Estimation Section

Estimation Section

Parameter	df	Regression Coefficient	Standard Error	T-Ratio	Prob Level	Last R-Squared
Intercept	1	-30.66667				
Temp	1	-12.125	7.947353	-1.53	0.187613	0.054938
Ratio	1	-17	7.947353	-2.14	0.085417	0.107995
Height	1	-21.375	7.947353	-2.69	0.043321	0.170733
Temp^2	1	32.08333	11.69819	2.74	0.040667	0.177530
Ratio^2	1	47.83333	11.69819	4.09	0.009457	0.394616
Height^2	1	6.083333	11.69819	0.52	0.625242	0.006383
Temp*Ratio	1	8.25	11.23925	0.73	0.495884	0.012717
Temp*Height	1	1.5	11.23925	0.13	0.899034	0.000420
Ratio*Height	1	-1.75	11.23925	-0.16	0.882357	0.000572

Estimation Section Using Pure Error

Parameter	df	Regression Coefficient	Standard Error	T-Ratio	Prob Level	Last R-Squared
Intercept	1	-30.66667				
Temp	1	-12.125	1.594261	-7.61	0.016853	0.054938
Ratio	1	-17	1.594261	-10.66	0.008680	0.107995
Height	1	-21.375	1.594261	-13.41	0.005517	0.170733
Temp^2	1	32.08333	2.346688	13.67	0.005307	0.177530
Ratio^2	1	47.83333	2.346688	20.38	0.002398	0.394616
Height^2	1	6.083333	2.346688	2.59	0.122137	0.006383
Temp*Ratio	1	8.25	2.254625	3.66	0.067241	0.012717
Temp*Height	1	1.5	2.254625	0.67	0.574315	0.000420
Ratio*Height	1	-1.75	2.254625	-0.78	0.518860	0.000572

This report shows the regression coefficient estimates of each term and their test of significance. This display actually shows two reports. The top is the regular Estimation Section defined below. Note that the Standard Errors are based on the Total Error Mean Square. The second report is identical to the top, except that the Standard Errors are now based on the Pure Error Mean Square.

Parameter

The particular term being displayed.

DF

The degrees of freedom associated with the term.

Regression Coefficient

The estimated value of the regression coefficient.

Standard Error

The standard error of the above regression coefficient. Note that the Total Error Mean Square is used for the top report, and the Pure Error Mean Square is used for the bottom report.

T-Ratio

The t-value for testing that this regression coefficient is zero after considering all other terms in the model. Note that the Total Error Mean Square is used for the top report, and the Pure Error Mean Square is used for the bottom report.

Prob Level

The probability or significance level of this test. If you were testing at the alpha equals 0.05 level of significance, this value would have to be less than 0.05 in order for the test to be deemed significant and the regression coefficient different from zero.

Last R-Squared

The amount that the R-Squared would decrease if this term were removed from the model.

Optimum Solution Section

Optimum Solution Section

Parameter	Maximum Exponent	Optimum Value
Temp	2	0.1219125
Ratio	2	0.1995746
Height	2	1.770525

Function at optimum	-52.02463
Number of Function Evaluations	378
Maximum Functions Evaluations	500

This report gives the results of the function minimization (or maximization) calculation.

Optimum Value

The value for each of the factors at the computed critical point. Covariates were evaluated at their means. Note that this solution is not constrained to fall within the design space. Note also that the values of some variables may be very large or small. This indicates that the function did not have a minimum (maximum) and that the search procedure was terminated by the maximum number of function evaluations. In this case, you might switch from finding a minimum to finding a maximum in the Optimization Goal box.

Function at Optimum

The value of the estimated function evaluated at the optimal values of each of the factors.

Residual Section

Residual Section

Row	Odor	Predicted	Residual
1	66	86.625	-20.625
2	58	42.5	15.5
3	65	59.875	5.125
4	-31	-30.66667	-0.3333333
5	39	45.875	-6.875
6	17	15.25	1.75
7	7	29.375	-22.375
8	-35	-30.66667	-4.333333
9	43	36.125	6.875
10	-5	-3.25	-1.75
11	43	20.625	22.375
12	-26	-30.66667	4.666667
13	49	28.375	20.625
14	-40	-24.5	-15.5
15	-22	-16.875	-5.125
16		28.375	

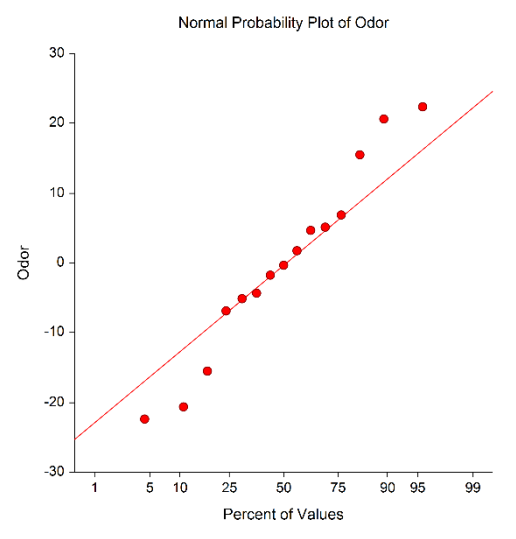
This report shows the response variable, the predicted value based on the response surface equation, and the residual (the difference between the two).

Notice that a predicted value is given for row sixteen, but no residual or Odor value is given. If you look at row sixteen on the database, you will note that it has a missing value for Odor and thus was not used in estimating the regression equation. However, since there are values for the three independent variables, a predicted value can be generated. This shows how to automatically generate predicted values for a set of X's when the observed Y is not on your database.

Normal Probability Plot

This plot displays a normal probability plot for assessing the validity of the assumption of normality. Note that you should ignore this plot when you have less than about five observations per term in the model, since the assumption of independence of residuals cannot be demonstrated and thus the probability plot may give inaccurate results.

Probability Plot



Contour Plots

These contour (or grid) plots show the value of the estimated equation at the center of each grid of rectangles. All factors not on either axis are evaluated at their mean value (unless a constant value was specified in the Factor Constant box).

Contour Plots