Chapter 371

Polynomial Model Search – Y vs Multiple X's

Introduction

This procedure searches through hundreds of potential polynomial models looking for the one that fits your data the best. The procedure is heuristic in nature but seems to do well with the data we have tried. This procedure models the relationship between a dependent variable and up to four independent variables.

Besides the common polynomial models, the procedure also considers a general class of models called the ratio of polynomials (see Polynomial Model Fit – Y vs Multiple X's procedure). These models provide a wider variety of curves to search from then do the usual polynomial models. Normally, fitting these models is a slow, iterative process. However, using a shortcut, an approximate solution may be found very quickly so that a large number of models may be searched in a short period of time. After the best fitting model is found, the *Polynomial Model Fit – Y vs Multiple X's* procedure is used to obtain a detailed analysis.

Parsimony

One of the main principles in model building is that you never use three parameters when two parameters will do. Hence, one of our tasks will be to find a model with the fewest number of parameters. A second principle in dealing with the ratio-of-polynomials model is that you should not fit a model with a numerator of higher polynomial order than that of the denominator. The models tried by default by this program follow these rules. A third rule is that all terms in a polynomial up to the desired order must be included. Hence, you would not use $Y=A+CX^2$. Instead, you would fit $Y=A+BX+CX^2$.

Goodness-of-Fit

Measuring how well a given model fits the data so that the various models can be compared is an important part of the search. This is tough since the goodness-of-fit statistics you are familiar with (like R^2) do not have the same meaning in nonlinear regression models. However, because of the lack of other general, goodness-of-fit indices, we have chosen to base our selection on the value of an R^2 like statistic called pseudo-R^2.

Problems with Ratio of Polynomials Models

As stated above, polynomials are used to approximate a function in a specific range close to a fixed point (such as zero). The approximation is only accurate within a narrow range. Outside this range, the polynomial approximation is less accurate.

For example, consider the polynomial ratio model

$$Y = \frac{10 + 11X + X^2}{4 - 5X + X^2}$$

Note that these two polynomials can be factored as follows

$$Y = \frac{(X + 1)(X + 10)}{(X - 1)(X - 4)}$$

Suppose the range of X is from 0 to 10. We note that when X is equal to 1 or 4, a division by zero will occur and the predicted value of Y goes toward infinity, so the model may not be useful. However, if the range of the data was 5 to 10, the roots of the denominator polynomial are missed, and no division by zero occurs.

As this example points out, when the roots of the denominator polynomial are within the range of the data, serious errors in the approximation will often be seen.

Shorthand Version of the Model

These polynomial models can be long, so **NCSS** has developed a shorthand notation that allows you to enter a long, complicated model with only a few terms.

Syntax

The syntax of the lists of terms in the models follow these rules:

Individual Terms

Individual terms may be listed as UiVj. If i or j is one, it may be omitted. For example, "UV2X3" means UV^2X^3 and "U2" means U^2. A list of individual terms in the polynomial is formed by separating terms with commas.

For example, if you had two variables selected, the entry

"U,V,UV,U2,V2,UV2,U2V,U2V2"

would result in the polynomial

$$Y = A0 + A1U + A2V + A3UV + A4U^2 + A5V^2 + A6UV^2 + A7U^2V + A8U^2V^2$$

Oi

The Oi notation includes all terms (not variables) of a particular **order**. The order is the sum of the exponents of the variables in a term. For example, the order of the term UVW^3 (entered as "UVW3") is five.

The maximum value for i is 5.

If you had selected three variables and included "O3" in the list of terms, you would include the terms U^3, V^3, W^3, U^2V, U^2W, V^2W, UV^2, VW^2, and UVW in your model.

No other terms of a different order are added to the model by this option. However, you can enter several of these together to form more complete models. For example, if you had three variables, U, V, and X, the entry "O1,O2" adds the following polynomial to the model

$$Y = A0 + A1U + A2V + A3X + A4U^2 + A5V^2 + A6X^2 + A7UV + A8UX + A9VX$$

Si

The Si notation includes all terms of **single** variables to the power i. The maximum value for i is 5.

For example, if you had selected three variables and included "S2" in the list of terms, your polynomial would include the terms U^2, V^2, and W^2.

These options can be combined with other options. For example, "O1,O2,H1,S2,E1" is a value choice. Duplicate terms will be removed.

Ei

The Ei notation includes all terms with at least one variable to the power (**exponent**) i and none of the other variables to a power greater than i. The maximum value for i is 5.

For example, if you had selected two variables and included "E2" in the list of terms, you would include the terms U^2, V^2, U^2V, UV^2, and U^2V^2.

Hi

The Hi notation includes all terms in a **hierarchical** model of order i. The maximum value for i is 5.

For example, if you had selected three variables and entered "H2" as the sole entry, the resulting polynomial would be

$$Y = A0 + A1U + A2V + A3W + A4U^2 + A5V^2 + A6W^2 + A7UV + A8UW + A9VW$$

P

The P option includes all simple paired terms. For example, if you had selected three variables and included "P" in the list of terms, the following terms would be added to the polynomial: UV, UW, and VW.

T

The T notation includes all triplet terms. For example, if you had selected four variables and included "T" in the list of terms, you would include the terms UVW, UVX, UWX, and UWX in your model.

Combining Options

You can combine these notations however you like. If a term is specified twice, it will be included in the model only once. The order in which you specify terms is arbitrary.

Examples of valid entries are

E2

U,V,E2,O1

O1,U2V2

Hint

If you want to know what polynomial results from a particular entry, enter it in this option, run the program, and view it in the Model Estimation report.

Assumptions and Limitations

Usually, nonlinear regression is used to estimate the parameters in a nonlinear model without performing hypothesis tests. In this case, the usual assumption about the normality of the residuals is not needed. Instead, the main assumption needed is that the data may be well represented by the model.

Data Structure

The data are entered in two or more variables: one dependent variable and up to four independent variables.

Missing Values

Rows with missing values in the variables being analyzed are ignored in the calculations. When only the value of the dependent variable is missing, predicted values are generated.

Example 1 – Finding a Multivariate Polynomial Model

This section presents an example of how to fit a multivariate ratio of polynomials model. In this example, we will search for a model relating the dependent variable Y to the independent variables U and X of the FnReg4 dataset.

Setup

To run this example, complete the following steps:

1 Open the FnReg4 example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **FnReg4** and click **OK**.

2 Specify the Polynomial Model Search – Y vs Multiple X's procedure options

- Find and open the **Polynomial Model Search – Y vs Multiple X's** 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 Variable	**Y**
Include transformed Y's in the model search	**Checked**
Y	**Checked**
ln(Y)	**Checked**
U Variable	**U**
Include transformed U's in the model search	**Checked**
U	**Checked**
ln(U)	**Checked**
X Variable	**X**
Include transformed X's in the model search	**Checked**
X	**Checked**
ln(X)	**Checked**
Types of Models Searched	**Polynomial and Ratio of Polynomials Models**
Maximum Exponent of Variables	**5**

3 Run the procedure

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

Model Search Summary

Model Search Summary

Variables:	Y = Y, U = U, X = X
Types of Models Searched:	Polynomials and Ratio of Polynomials Models
Maximum Exponent of Variables:	5
Models Searched:	760

Pseudo-R²		Transformations			Number of				
Rank	Value	Y	U	X	Independent Variables	Model Terms	Coefficients Equal Zero	Rows Used	Abbreviated Model
1	0.99997	ln(Y)	ln(U)	ln(X)	2	15	0	225 of 225	E1,E2,S3,S4,S5
2	0.99997	Y	ln(U)	ln(X)	2	15	0	225 of 225	E1,E2,S3,S4,S5
3	0.99997	Y	ln(U)	ln(X)	2	15	0	225 of 225	(1) / (E1,E2,S3,S4,S5)
4	0.99997	ln(Y)	ln(U)	ln(X)	2	13	0	225 of 225	E1,E2,S3,S4
5	0.99997	Y	ln(U)	ln(X)	2	13	0	225 of 225	E1,E2,S3,S4
6	0.99997	Y	ln(U)	ln(X)	2	13	0	225 of 225	(1) / (E1,E2,S3,S4)
7	0.99997	ln(Y)	ln(U)	ln(X)	2	11	0	225 of 225	E1,E2,S3
8	0.99996	Y	ln(U)	ln(X)	2	11	0	225 of 225	(1) / (E1,E2,S3)
9	0.99996	ln(Y)	ln(U)	X	2	15	0	225 of 225	E1,E2,S3,S4,S5
10	0.99996	Y	ln(U)	X	2	15	0	225 of 225	E1,E2,S3,S4,S5
11	0.99996	Y	ln(U)	X	2	15	0	225 of 225	(1) / (E1,E2,S3,S4,S5)
12	0.99996	ln(Y)	ln(U)	X	2	13	0	225 of 225	E1,E2,S3,S4
13	0.99996	Y	ln(U)	X	2	13	0	225 of 225	(1) / (E1,E2,S3,S4)
14	0.99996	Y	ln(U)	ln(X)	2	11	0	225 of 225	E1,E2,S3
15	0.99996	Y	ln(U)	X	2	11	0	225 of 225	(1) / (E1,E2,S3)
16	0.99996	Y	ln(U)	X	2	13	0	225 of 225	E1,E2,S3,S4
17	0.99996	ln(Y)	ln(U)	X	2	11	0	225 of 225	E1,E2,S3
18	0.99996	ln(Y)	ln(U)	X	2	15	0	225 of 225	(1) / (E1,E2,S3,S4,S5)
19	0.99996	ln(Y)	ln(U)	ln(X)	2	15	0	225 of 225	(1) / (E1,E2,S3,S4,S5)
20	0.99996	ln(Y)	ln(U)	ln(X)	2	13	0	225 of 225	(1) / (E1,E2,S3,S4)

Pseudo-R²	Compares the R² of the estimated model with the R² of the null model. The null model contains no parameters besides the intercept. Pseudo-R² is a measure of improvement a particular model over the null model and thus is a measure of the goodness of fit. Unfortunately, in nonlinear regression analysis, the pseudo-R² can be, and often is, a large negative number. We interpret this to mean that the model does not fit well.
Model Terms	This is the number of terms in the model. The selected model should have as few independent variables as possible while maintaining a relatively large value of pseudo-R².
Equal Zero	This is the number of estimated coefficients that are exactly zero. A coefficient of zero is a signal that the corresponding term is not used, indicating that this model has more terms than needed and should not be used.
Rows Used	This is the number of rows used followed by the number of rows processed. This allows you to note how many rows were omitted because of a missing value in the data or a missing value caused by a transformation (e.g. trying to take the log of a negative number).
Si	Includes all single variables to the power i. For example, if you had selected three variables and included 'S2' in the list of terms, you would include the terms U², V², and W² in your model.
Ei	Includes all combinations of variables with at least one variable to the power i and none of the other variables to a power greater than i. For example, if you had selected two variables and included 'E2' in the list of terms, you would include the terms U², V², U²V, UV², and U²V² in your model.
Hi	Includes all terms in the hierarchical model of order i. All terms for which the sum of the exponents is less than or equal to i are included. For example, if you had selected two variables and included 'H2' in the list of terms, you would include the terms U, V, U², V², and UV in your model. However, terms like U²V, UV², and U²V² are not include since the sum of the exponents in the term is > 2.

This report displays the best models (in terms of R^2) found. Each row describes the results for a single model.

Note that the first seven models have the same value of pseudo-R^2 to five decimal places. Note that models ranked 3, 6, 8, 11, 13, 15, 18, 19, and 20 are ratio of polynomial models. These models are more complicated and should be avoided unless they result in a much higher value of pseudo-R^2 (which they do not in this example).

Models Searched

This value is the total number of models that were evaluated.

Pseudo-R^2

This is the usual value of R^2 for regular polynomial models. However, there is no direct R^2 defined for ratio of polynomial models. In this case, a pseudo R^2 is constructed to approximate the usual R^2 value used in multiple regression. We use the following generalization of the usual R^2 formula:

$$R^2 = (ModelSS - MeanSS)/(TotalSS-MeanSS)$$

where *MeanSS* is the sum of squares due to the mean, *ModelSS* is the sum of squares due to the model, and *TotalSS* is the total (uncorrected) sum of squares of Y (the dependent variable).

This version of R^2 tells you how well the model performs after removing the influence of the mean of Y. Since many nonlinear models do not explicitly include a parameter for the mean of Y, this R^2 may be negative (in which case we set it to zero) or difficult to interpret. However, if you think of it as a direct extension of the R^2 that you use in multiple regression, it will serve well for comparative purposes.

Transformations (Y U V W X)

The symbols refer to the transformations that were used for that variable.

Number of Independent Variables

This is the number of variables fit in this model.

Number of Model Terms

This is the number of terms in the model. The selected model should have as few independent variables as possible while maintaining a relatively large value of pseudo-R^2.

Number of Coefficients Equal Zero

This is the number of estimated coefficients that are exactly zero. A coefficient of zero is a signal that the corresponding term is not used, indicating that this model has more terms than needed and should not be used.

Number of Rows Used

This is the number of rows used followed by the number of rows processed. This allows you to note how many rows were omitted because of a missing value in the data or a missing value caused by a transformation (e.g. trying to take the log of a negative number).

Abbreviated Model

This gives the model using the shorthand notation described next. You can apply this shorthand notation directly in *Polynomial Model Search – Y vs Multiple X's* procedure to obtain detailed results for a particular model.

Note that the numeral one (1) is used when no polynomial is specified.

Si Include all single variables to the power i. For example, if you had selected three variables and included 'S2' in the list of terms, you would include the terms U^2, V^2, and W^2 in your model.

Ei Includes all combinations of variables with at least one variable to the power i and none of the other variables to a power greater than i. For example, if you had selected two variables and included 'E2' in the list of terms, you would include the terms U^2, V^2, U^2V, UV^2, and U^2V^2 in your model.

Hi Includes all terms in the hierarchical model of order i. All terms for which the sum of the exponents is less than or equal to i are included. For example, if you had selected two variables and included 'H2' in the list of terms, you would include the terms U, V, U^2, V^2, and UV in your model. However, terms like U^2V, UV^2, and U^2V^2 are not include since the sum of the exponents in the term is > 2.

Model 1 Details

Model 1 Details

Independent Variables:	2
Model Exponent:	5
Model Terms:	15 (Number of Zero Coefficients = 0)
Abbreviated Model:	E1,E2,S3,S4,S5
Full Model:	$(A0 + A1U + A2U^2 + A3U^3 + A4U^4 + A5U^5 + A6X + A7UX + A8U^2X + A9X^2 + A10UX^2 + A11U^2X^2 + A12X^3 + A13X^4 + A14X^5)$
Y:	ln(Y)
U:	ln(U)
X:	ln(X)
Rows Used:	225 of 225

Pseudo-R^2 (Search Criterion)

This Model:	0.99997 (100.00% of the Best Model R^2)
Best Model:	0.99997

Coefficient	Term	Estimate
A0	Intercept	0.40628
A1	U	-0.23594
A2	U^2	-0.03416
A3	U^3	0.08602
A4	U^4	0.03963
A5	U^5	0.00542
A6	X	-0.09168
A7	UX	0.02079
A8	U^2X	-0.01009
A9	X^2	-0.05408
A10	UX^2	0.00049
A11	U^2X^2	-0.00202
A12	X^3	-0.01090
A13	X^4	-0.00084
A14	X^5	-0.00005

Estimated Model (Double Precision)

ln(Y) =
((0.406280349175502) - (0.235944235358469)*(ln(U)) - (0.0341576298714983)*(ln(U))^2 +
(0.086016729918113)*(ln(U))^3 + (0.039625505353205)*(ln(U))^4 + (0.00542316969985191)*(ln(U))^5 -
(0.0916761537798921)*(ln(X)) + (0.020786820987610 7)*(ln(U))*(ln(X)) - (0.010085543403049)*(ln(U))^2*(ln(X)) -
(0.0540822197146366)*(ln(X))^2 + (0.000489191944476615)*(ln(U))*(ln(X))^2 -
(0.00202010362591643)*(ln(U))^2*(ln(X))^2 - (0.0108974989738289)*(ln(X))^3 -
(0.000835317789244781)*(ln(X))^4 - (5.14566538217058E-05)*(ln(X))^5)

This report displays the details of each candidate model sorted in order of pseudo-R^2. Thus, the first model shown is the model with the best fit.

Abbreviated Model

This gives the model using the shorthand notation. The next line gives the full model so that you can determine which terms where used in the model.

Full Model

This gives the model expanded to show each term in the model.