Chapter 371

Ratio of Polynomials Search – Many Variables

Introduction

This procedure searches through hundreds of potential curves looking for the model 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 is more general that the Ratio of Polynomials Search procedure because it models up to four independent variables.

A general class of models called the ratio of polynomials (see Multivariate Ratio of Polynomials Fit chapter) provides a wide variety of curves to search from. 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, use the procedure discussed in the Multivariate Ratio of Polynomials Fit chapter 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-Squared) do not have the same meaning in this setting. However, because of the lack of other general, goodness-of-fit indices, we have chosen to base our selection on the value of R-Squared.

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.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

Specify the variables on which to run the analysis.

Variables

Y (Dependent) Variable

Specifies a single dependent (Y) variable. This is the variable being predicted.

Variables – Transformations (Y)

1/Y^2, 1/Y, 1/SQRT(Y), LN(Y), SQRT(Y), Y, Y^2

Specifies whether this transformation of Y should be searched.

Variables – Independent Variables

U, V, W, X

Each option specifies a single independent variable. When a variable is selected, designate which of the transformations of that variable are searched.

Variables – Transformations (U,V,W,X)

1/U^2, 1/U, 1/SQRT(U), LN(U), SQRT(U), U, U^2

Specifies whether this transformation of U, V, W, and/or X should be searched.

Model Specification

These options specify which models are searched. Care must be taken when selecting models because it is very easy to overwhelm the algorithm by selecting too many candidate models.

For each model type, select one of the following options.

- **Omit**

  Do not add this type of model to the pool of models searched.

- **Numerator**

  Add models involving the numerator polynomial only to the pool of models that is searched.

- **Denominator**

  Add models involving the denominator polynomial only to the pool of models that is searched.

- **Numer. + Denom.**

  Add both numerator only and denominator only models to the pool of models that is searched.

- **Ratio**

  Add ratio of polynomial models to the pool of models that is searched.

- **Numer. + Ratio**

  Add numerator only and ratio models to the pool of models that is searched.

- **All Three**

  Add numerator only, denominator only, and ratio models to the pool of models that is searched.

(1-5) Single Variable

This option selects candidate models consisting of one variable only, up to the order specified to the left of the list box (1-5). For example, selecting Numerator in the second row (order 2) would include quadratic, single variable models such as

$Y=B0+B1X+B2X^2, Y=B0+B1U+B2U^2$, etc.

(1-5) Hierarchical

This option selects hierarchical-polynomial models (see Hi below) to the order specified by the number on the left (1 to 5).

1 Max Power

This option selects models using the "E"-notation (see below). These models would be of the form

$Y = E1.$

2 Max Power

This option selects models using the "E"-notation (see below). These models would be of the form

$Y = E1,E2.$

3 Max Power

This option selects models using the "E"-notation (see below). These models would be of the form

$Y = E1,E2,E3.$

4 Max Power

This option selects models using the "E"-notation (see below). These models would be of the form

$Y = E1,E2,E3,E4.$

5 Max Power

This option selects models using the "E"-notation (see Ei below). These models would be of the form

$Y = E1,E2,E3,E4,E5.$

Pairs - (1 and 2)

This option selects models consisting of two independent variables with no cross products, up to the order first (Pairs -1) or second (Pairs -2) order. For example, selecting Numerator in the Pairs- 1 box would include models like

Y=B0+B1X+B2U.

Triplets - 1

This option selects models consisting of three independent variables with no cross products, up to a first order. For example, selecting Numerator in the Triplets -1 box would include models like

Y=B0+B1X+B2U+B3V.

Models Reported

This option limits the number of models that are reported on. For example, if you select 20 here, then the report shows the 20 best models.

Zero

This is the value used as zero by the nonlinear algorithm. Because of rounding error, values lower than this value are reset to zero. If unexpected results are obtained, you might try using a smaller value, such as 1E-16. Note that 1E-5 is an abbreviation for the number 0.00001.

Custom Models Tab

Up to ten custom models may be specified using the basic syntax for model input described below by specifying the numerator and denominator polynomials.

The syntax of these lists of terms follows these rules:

1. Individual terms may be listed as UiVj. If i or j is one, it may be omitted. For example, UV2X3 means $(U)(V*V)(X*X*X)$ and U2 means U^2, which means $U*U$. A list of individual terms is formed by separating such terms with commas.

2. The **Oi** notation includes all terms 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 U2VW3 is six. If you had selected three variables and included "O2" in the list of terms, you would include the terms U3, V3, W3, U2V, U2W, V2W, UV2, VW2, and UVW in your model.

3. The **Si** notation 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 U, V, W, U2, V2, and W2 in your model.

4. The **Ei** notation 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 three variables and included "E2" in the list of terms, you would include the terms U2, V2, W2, U2V, U2W, U2V2, U2W2, UV2, UW2, VW2, V2W, and V2W2 in your model.

5. The **Hi** notation includes all terms in the hierarchical model of order i. 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.

6. The **P** notation includes all simple paired terms. For example, if you had selected three variables and included "P" in the list of terms, you would include the terms UV, UW, and VW in your model.

7. 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.

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 are:

E2

U,V,E2,O1

O1,U2V2

Example 1 – Fitting a Multivariate Ratio of Polynomials 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.

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 Ratio of Polynomials Search – Many Variables window.

1   Open the FnReg4 dataset.
- From the File menu of the NCSS Data window, select **Open Example Data**.
- Click on the file **FnReg4.NCSS**.
- Click **Open**.

2   Open the Ratio of Polynomials Search – Many Variables window.
- Using the Analysis menu or the Procedure Navigator, find and select the **Ratio of Polynomials Search – Many Variables** procedure.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3   Specify the variables.
- On the Ratio of Polynomials Search – Many Variables window, select the **Variables tab**.
- Double-click in the **Y (Dependent)** box. This will bring up the variable selection window.
- Select **Y** from the list of variables and then click **Ok**.
- Check the box next to **Ln(Y)**.
- Double-click in the **U Variable** box. This will bring up the variable selection window.
- Select **X** from the list of variables and then click **Ok**.
- Check the box next to **Ln(U)**.
- Double-click in the **V Variable** box. This will bring up the variable selection window.
- Select **U** from the list of variables and then click **Ok**.
- Check the box next to **Ln(V)**.
- Leave all other options at their default settings.

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

Multivariate Search Report

Results of Multivariate Search

Variables   Y=Y, U=X, V=U
Models Fit   192

R-Squared	Percent of Best	Y	U	V	W	X	No. of X's	Model Numerator / Denominator
0.991356	100.00	N	L	L			10	(H2) / (H2)
0.991349	100.00	L	L	L			10	(H2) / (H2)
0.982629	99.12	L	N	L			10	(H2) / (H2)
0.982347	99.09	N	N	L			10	(H2) / (H2)
0.980970	98.95	N	N	L			8	(E1,E2) / 1
0.979765	98.83	N	L	L			8	(E1,E2) / 1
0.978064	98.66	L	N	L			8	(E1,E2) / 1
0.977188	98.57	N	L	L			5	(H2) / 1
0.976036	98.45	L	L	L			8	(E1,E2) / 1
0.975872	98.44	N	N	L			5	(H2) / 1
0.972253	98.07	L	N	L			5	(H2) / 1
0.971677	98.01	L	L	L			5	(H2) / 1
0.899874	90.77	L	N	N			8	(E1,E2) / 1
0.898429	90.63	N	L	N			10	(H2) / (H2)
0.897845	90.57	L	L	N			8	(E1,E2) / 1
0.893617	90.14	N	N	N			8	(E1,E2) / 1
0.892422	90.02	N	L	N			8	(E1,E2) / 1
0.887341	89.51	L	N	L			4	(U1,V1,U2,V2) / 1
0.884633	89.23	L	L	L			4	(U1,V1,U2,V2) / 1
0.884020	89.17	N	N	L			4	(U1,V1,U2,V2) / 1

Transformation key:
2=1/X^2, I=1/X, Q=1/SQRT(X), L=LN(X), R=SQRT(X), N=X, and S=X^2.

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

Models Fit

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

R-Squared

There is no direct R-Squared defined for nonlinear regression. This is a pseudo R-Squared constructed to approximate the usual R-Squared value used in multiple regression. We use the following generalization of the usual R-Squared formula:

$$R\text{-}Squared = (ModelSS - MeanSS)/(TotalSS\text{-}MeanSS)$$

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-Squared 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-Squared 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-Squared that you use in multiple regression, it will serve well for comparative purposes.

Percent of Best

This is the percent that the R-Squared value of this model is of the best (top) model.

Transformations Y U V W X

The letters correspond to the transformations that were used for each variable. The transformation key is listed at the bottom of the report. For example, the entry for the first row is N L L. This means that Y was regressed on ln(X) and ln(U).

No. of X's

This is the number of parameters fit in this model. Of course, we want models with only a few parameters and a large value of R-Squared.

Model Numerator/Denominator

This gives the model using the shorthand notation described in Models - Custom section above. You can apply this shorthand notation directly in Multivariate Ratio of Polynomial Fit procedure to obtain detailed results for a particular model.

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