Chapter 370

Ratio of Polynomials Search – One Variable

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.

A general class of models called the ratio of polynomials (see the previous 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 Ratio of Polynomials Fit chapter to provide a detailed analysis of it.

For each model, various transformations of X and Y can be tried. This expands the number of models that may be tried to several hundred.

The general ratio of polynomials model fit is

$$g(Y) = \frac{A0 + A1\,f(X) + A2\,f^2(X) + A3\,f^3(X) + A4\,f^4(X) + A5\,f^5(X)}{1 + B1\,f(X) + B2\,f^2(X) + B3\,f^3(X) + B4\,f^4(X) + B5\,f^5(X)} + e\,.$$

Here *g(Y)* and *f(X)* represent power transformations of Y and X such as LOG(X), SQRT(X), etc. The parameters *A0*, *A1*, *A2*, ..., *B5* are constants that are estimated from the data. The value *e* represents the error or residual of that observation. By setting some constants to zero, various simplified models are obtained. For example, if only *A0* and *A1* are nonzero, the familiar linear model, *Y=A0+A1X+e*, is obtained.

A Shortcut

Consider the simple model

$$Y = \frac{A0 + A1\,X}{1 + B1\,X} + e\,.$$

If you ignore *e* (set it to zero for a moment) and multiply both sides of this equation by *(1+B1X)* you will get

$$Y+B1XY=A0+A1X.$$

Now if you subtract *B1XY* from both sides you will get

$$Y=A0+A1X-B1XY.$$

Finally, if you relabel *XY* as *Z* you get

$$Y=A+BX+CZ.$$

Note that the variable *Z* is a direct transformation of *X* and *Y*. This last equation is in standard linear form. The parameters *A*, *B*, and *C* may be estimated using standard multiple regression! Note that the parameter *B1* in our original equation is equal to *-C* in the final equation.

One catch in using this procedure is that you have to assume the *e* to be zero. When the model fits well, the *e* will be near zero. When the model does not fit well, these *e* will be relatively large and our method breaks down. However, the large *e* will warn us that the model has not fit well.

Parsimony

One of the main principles in model building is that you should 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 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 $Y=A+BX+CX^2$.

The program tries the five models having a fifth-order polynomial in the denominator. The numerator polynomials are $A0+A1X$, $A0+A1X+A2X^2$, ..., $A0+A1X+A2X^2+A3X^3+A4X^4+A5X^5$. Next the four models having a fourth-order polynomial denominator are tried. This continues on down to the simple equation $Y=(A0+A1X)/(1+B1X)$. This process is repeated for each combination of transformations that are specified for *Y* and *X*.

Goodness-of-Fit

The final issue measuring of how well a given model fits the data so that the various models can be compared. This is tough since the goodness-of-fit statistics you are familiar with (like R^2) 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^2. We justify this because this procedure is only an intermediate step in the modeling process. You must take several steps before making your final model selection.

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 variables: one dependent variable and one independent variable.

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 – Searching for the Best Ratio of Polynomials Model

This section presents an example of how to search for the best fitting ratio of polynomials model. In this example, we will search for the best fitting model using the variables Y and X of the FnReg3 dataset. We will also consider the log transformation of each variable in our search.

Setup

To run this example, complete the following steps:

1 Open the FnReg3 example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **FnReg3** and click **OK**.

2 Specify the Ratio of Polynomials Search – One Variable procedure options

- Find and open the **Ratio of Polynomials Search – One Variable** 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 (Dependent) Variable	**Y**
X (Independent) Variable	**X**

3 Run the procedure

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

Search Summary Section

Search Summary Section

Model No.	F(Y)	F(X)	Model	Current R-Squared	Best R-Squared	Percent of Best
1	y	x	3 / 4	0.990115	0.990115	100.00
2	y	x	1 / 4	0.989091	0.990115	99.90
3	y	x	2 / 4	0.989078	0.990115	99.90
4	LN(y)	x	3 / 4	0.988342	0.990115	99.82
5	y	x	0 / 5	0.987972	0.990115	99.78
6	y	x	4 / 4	0.984444	0.990115	99.43
7	LN(y)	x	0 / 5	0.984241	0.990115	99.41
8	LN(y)	x	2 / 4	0.984015	0.990115	99.38
9	LN(y)	x	1 / 4	0.983513	0.990115	99.33
10	y	x	0 / 4	0.983109	0.990115	99.29
11	LN(y)	x	0 / 4	0.977900	0.990115	98.77
12	LN(y)	LN(x)	4 / 5	0.975903	0.990115	98.56
13	y	x	1 / 5	0.975564	0.990115	98.53
14	LN(y)	x	5 / 0	0.974421	0.990115	98.41
15	y	x	2 / 5	0.972910	0.990115	98.26
16	LN(y)	x	1 / 5	0.970396	0.990115	98.01
17	LN(y)	x	4 / 0	0.967638	0.990115	97.73
18	y	LN(x)	4 / 5	0.956060	0.990115	96.56
19	y	x	5 / 0	0.948378	0.990115	95.78
20	LN(y)	LN(x)	3 / 3	0.945165	0.990115	95.46

This report displays a separate line for each model tried. Note that the results have been sorted by R-Squared so that the best model is displayed at the top.

For this example, the best model is the ratio of a third order numerator polynomial and a fourth order denominator polynomial, with no transformations of Y or X needed. We would now fit this model using the Ratio of Polynomial Fit procedure.

Model No.

The ranking of the model displayed on this line.

F(Y)

The transformation (if any) applied to the Y (dependent) variable.

F(X)

The transformation (if any) applied to the X (independent) variable.

Model

The ratio of polynomial model whose results are displayed on this row. The syntax of the model statement is N/D where N represents the order of the numerator polynomial and D represents the order of the denominator polynomial. If N or D is set to zero, that polynomial is ignored.

For example, the model 1/2 means A0+A1X in the numerator and 1+B1X+B2X^2 in the denominator.

Current R-Squared

The value of pseudo R-Squared for this model and transformations.

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:

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

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.

Best R-Squared

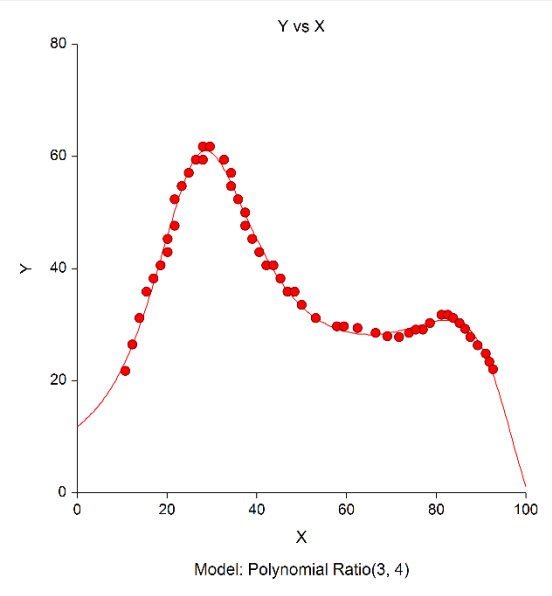
The pseudo R-Squared of the first (best) model.

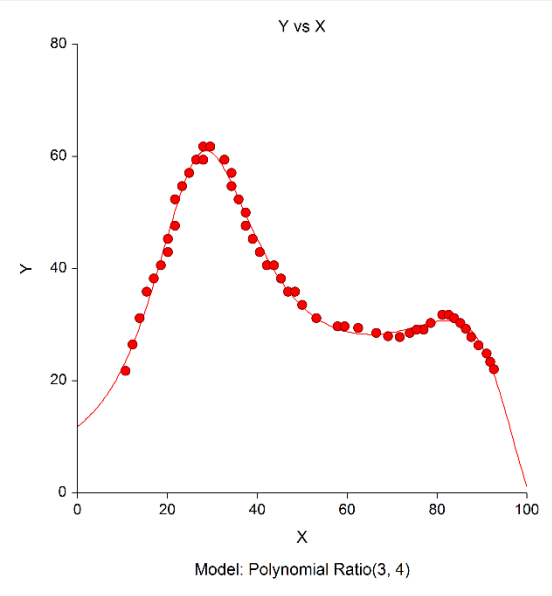
Percent of Best

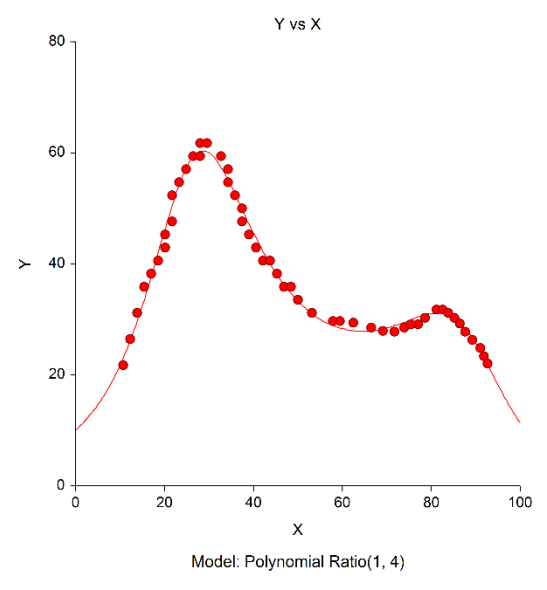
The percent that the pseudo R-Squared of this model is of the overall best model. Often you will be able to find models that are nearly as good as the best model but have many fewer parameters.

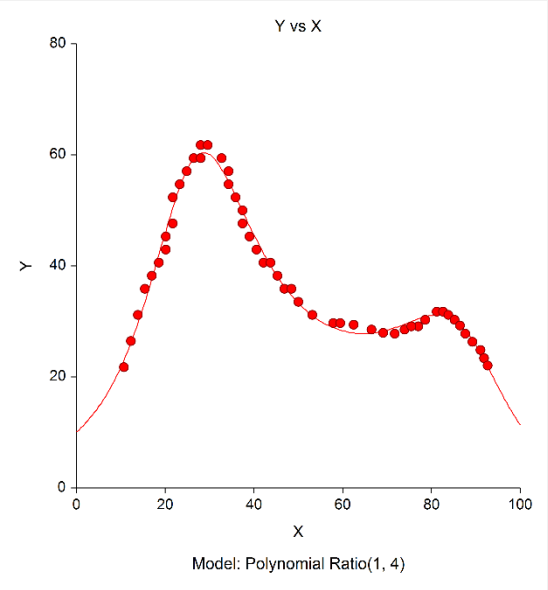
Function Plots

Function Plots

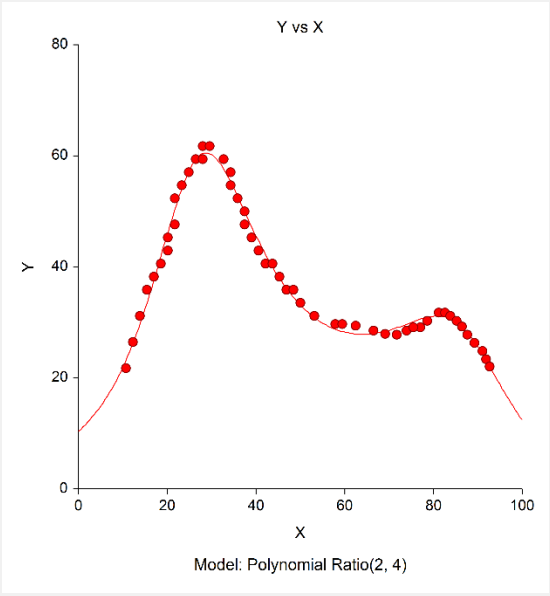


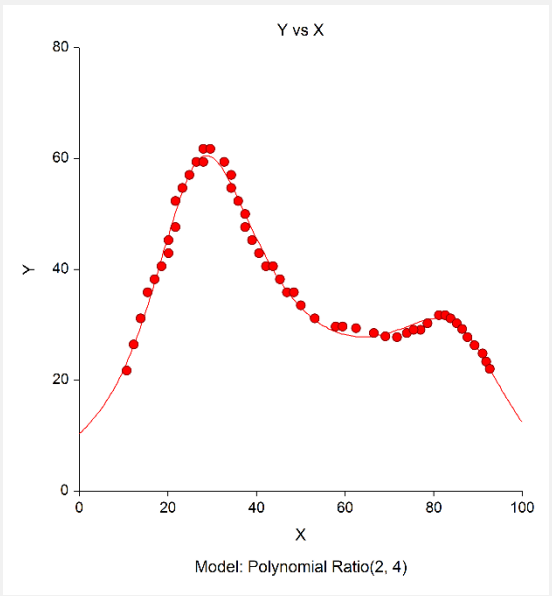






Ratio of Polynomials Search – One Variable





These plots show the best few models plotted in the original (on the left) and transformed (on the right) scales. They will help you determine which model (or models) you want to evaluate further using the Ratio of Polynomial Fit procedure.