Chapter 370

# Polynomial Model Search – Y vs One X

## Introduction

This procedure searches through hundreds of polynomial models looking for a model that fits your data the best. The procedure is heuristic in nature and seems to do well with the data we have tried.

In addition to regular polynomial models, a more general class of models called the *ratio of polynomials* 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 *Polynomials Model Fit – Y vs One X* 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

$$g(Y) = \frac{A0 + A1f(X) + A2f^2(X) + A3f^3(X) + A4f^4(X) + A5f^5(X)}{1 + B1f(X) + B2f^2(X) + B3f^3(X) + B4f^4(X) + B5f^5(X)} + e.$$

Here *g(Y)* and *f(X)* represent power transformations of Y and X such as ln(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 + A1 X + e,* is obtained.

### Shortcut

Consider the simple model

$$Y = \frac{A0 + A1X}{1 + B1X} + e$$

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

$$Y = A0 + A1X - B1Z$$

$$= A + BX + CZ$$

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 numerator with no denominator. Next, 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 a pseudo-$R^2$ statistic. 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.

## Problems with Ratio of Polynomials Models

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.

## 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 Polynomial Model Search - Y vs One X procedure options**

- Find and open the **Polynomial Model Search - Y vs One X** 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** |
| X Variable | **X** |
| Include transformed X's in the model search | **Checked** |
| X | **Checked** |
| ln(X) | **Checked** |

**3 Run the procedure**

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

# Model Search Summary

**Model Search Summary**

| Model Rank | Transformation Y | Transformation X | Polynomial Order* | Pseudo-R² (Search Criterion) Value | Best | % of Best | Rows Used |
|---|---|---|---|---|---|---|---|
| 1 | Y | X | 3 / 4 | 0.990115 | 0.990115 | 100.00% | 51 of 51 |
| 2 | Y | X | 1 / 4 | 0.989091 | 0.990115 | 99.90% | 51 of 51 |
| 3 | Y | X | 2 / 4 | 0.989078 | 0.990115 | 99.90% | 51 of 51 |
| 4 | ln(Y) | X | 3 / 4 | 0.988342 | 0.990115 | 99.82% | 51 of 51 |
| 5 | Y | X | 0 / 5 | 0.987972 | 0.990115 | 99.78% | 51 of 51 |
| 6 | Y | X | 4 / 4 | 0.984444 | 0.990115 | 99.43% | 51 of 51 |
| 7 | ln(Y) | X | 0 / 5 | 0.984241 | 0.990115 | 99.41% | 51 of 51 |
| 8 | ln(Y) | X | 2 / 4 | 0.984015 | 0.990115 | 99.38% | 51 of 51 |
| 9 | ln(Y) | X | 1 / 4 | 0.983513 | 0.990115 | 99.33% | 51 of 51 |
| 10 | Y | X | 0 / 4 | 0.983109 | 0.990115 | 99.29% | 51 of 51 |
| 11 | ln(Y) | X | 0 / 4 | 0.977900 | 0.990115 | 98.77% | 51 of 51 |
| 12 | ln(Y) | ln(X) | 4 / 5 | 0.975903 | 0.990115 | 98.56% | 51 of 51 |
| 13 | Y | X | 1 / 5 | 0.975564 | 0.990115 | 98.53% | 51 of 51 |
| 14 | ln(Y) | X | 5 | 0.974421 | 0.990115 | 98.41% | 51 of 51 |
| 15 | Y | X | 2 / 5 | 0.972910 | 0.990115 | 98.26% | 51 of 51 |
| 16 | ln(Y) | X | 1 / 5 | 0.970396 | 0.990115 | 98.01% | 51 of 51 |
| 17 | ln(Y) | X | 4 | 0.967638 | 0.990115 | 97.73% | 51 of 51 |
| 18 | Y | ln(X) | 4 / 5 | 0.956060 | 0.990115 | 96.56% | 51 of 51 |
| 19 | Y | X | 5 | 0.948378 | 0.990115 | 95.78% | 51 of 51 |
| 20 | ln(Y) | ln(X) | 3 / 3 | 0.945165 | 0.990115 | 95.46% | 51 of 51 |

* For polynomial models, Polynomial Order simply displays the order of the polynomial.
For polynomial ratio models, Polynomial Order is displayed as (Numerator Polynomial Order) / (Denominator Polynomial Order).

| Order | (Numerator) Polynomial | Denominator Polynomial |
|---|---|---|
| 0 | A0 | |
| 1 | A0 + A1X | 1 + B1X |
| 2 | A0 + A1X + A2X² | 1 + B1X + B2X² |
| 3 | A0 + A1X + A2X² + A3X³ | 1 + B1X + B2X² + B3X³ |
| 4 | A0 + A1X + A2X² + A3X³ + A4X⁴ | 1 + B1X + B2X² + B3X³ + B4X⁴ |
| 5 | A0 + A1X + A2X² + A3X³ + A4X⁴ + A5X⁵ | 1 + B1X + B2X² + B3X³ + B4X⁴ + B5X⁵ |

This report displays a separate line for each model tried. Note that the results have been sorted by Pseudo-$R^2$ 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 Rank

The ranking of the model displayed on this line.

### Transformation: Y

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

### Transformation: X

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

### Polynomial Order

This gives the maximum orders of the numerator and denominator polynomials used in the ratio 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.

### Pseudo-$R^2$ (Search Criterion): Value

The value of pseudo-$R^2$ for this model and transformations.

There is no direct $R^2$ defined for nonlinear regression. This is a pseudo R-Squared 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). 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.

### Pseudo-$R^2$ (Search Criterion): Best

The pseudo-$R^2$ of the first (best) model.

### Pseudo-$R^2$ (Search Criterion): % of Best

The percent that the pseudo $R^2$ 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 fewer parameters.

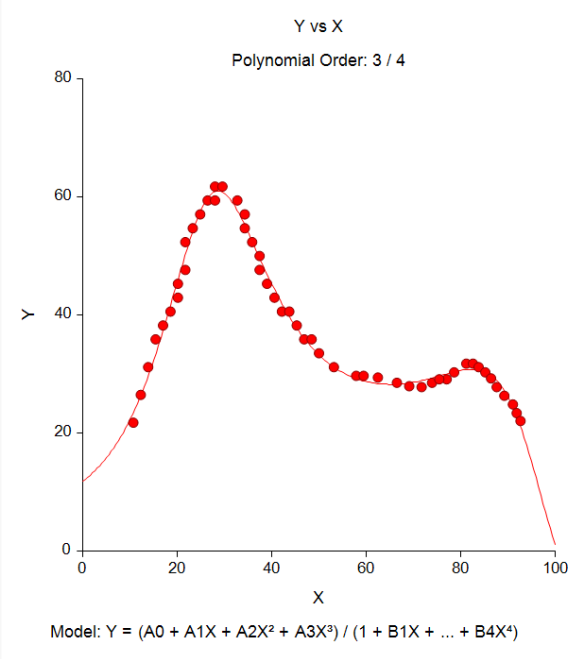
### Rows Used

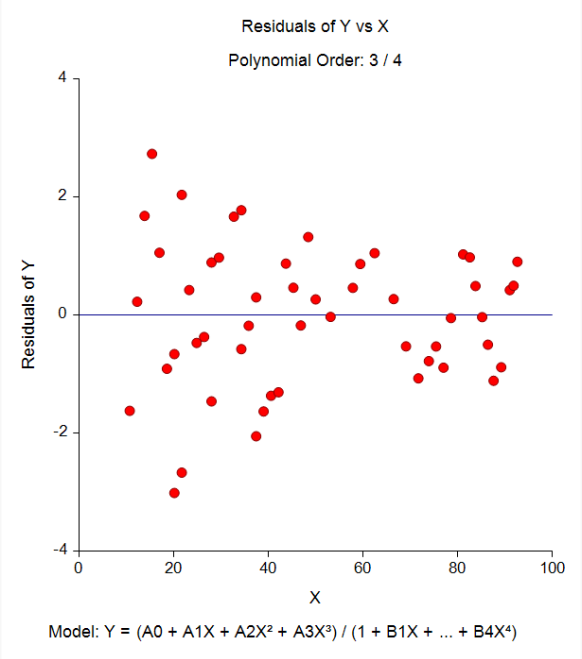
The number of rows used followed by the number of rows available. This is especially useful when a transformation causes some rows to be ignored. For example, if the data contain negative and zero values and you have selected a logarithmic transformation, the negative and zero rows will be ignored.

# Model 1 Plots and Reports

Individual plots and reports are displayed for each model so you can pick the best model.

**Model 1 Plots**




These plots show a plot of the data with the model fit overlayed on the left and the residuals on the right. These plots will allow you to detect any unwanted characteristics in this model. Note that none can be seen in these plots.

**Model 1 Details**

| | |
|---|---|
| Polynomial Order: | 3 / 4 |
| Polynomial Model: | Y = (A0 + A1X + A2X² + A3X³) / (1 + B1X + B2X² + B3X³ + B4X⁴) |
| Y: | Y |
| X: | X |
| Number of Rows Used: | 51 of 51 |

Pseudo-R² (Search Criterion)

| | | |
|---|---|---|
| This Model: | 0.990115 | (100% of the Best Model R²) |
| Best Model: | 0.990115 | |

| Numerator | | Denominator | |
|---|---|---|---|
| **Coefficient** | **Estimate** | **Coefficient** | **Estimate** |
| A0 | 11.78766 | B1 | 0.07834508 |
| A1 | -0.2798519 | B2 | -0.002391857 |
| A2 | 0.005809555 | B3 | 2.86472E-05 |
| A3 | -4.172031E-05 | B4 | -1.171313E-07 |

**Estimated Model (Double Precision)**

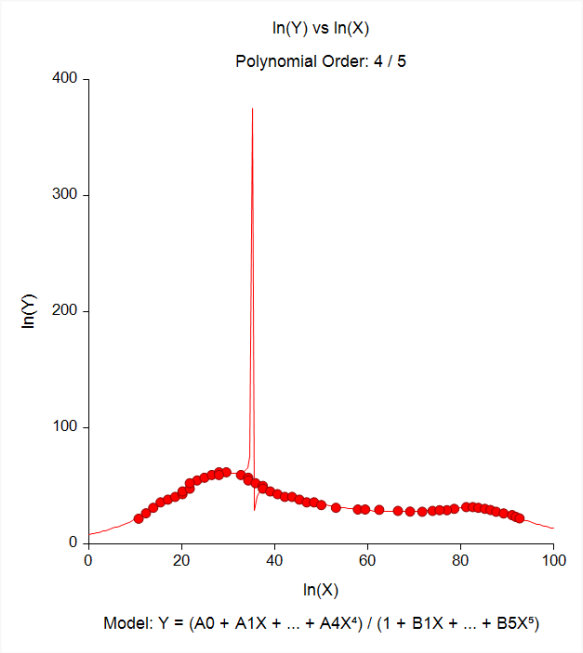
Y =
(11.7876572084676 - (0.279851898180577)*(X) + (0.00580955472564643)*(X)^2 –
(4.17203122468724E-05)*(X)^3) / (1 + (0.078345075095875 1)*(X) - (0.00239185672828884)*(X)^2 +
(2.86471961423406E-05)*(X)^3 - (1.17131250064605E-07)*(X)^4)

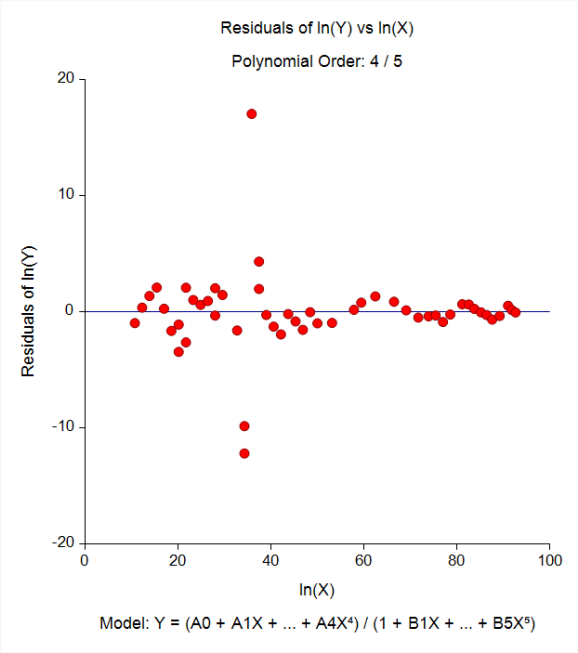
This report presents the numeric details of the fit. Note that this fit comes from the special approximation that was used. The actual fit provided by the nonlinear regression in the 'Fit' routine will usually give slightly different results.

## Model 12 Plots (To Illustrate Problems)

To point out the type of problem that may appear with these models, we next display the details for Model 12.

**Model 12 Plots**




These plots show a problem that can occur with ratio of polynomial models. The strange plot occurs because the denominator polynomial includes a value near zero when X is just below 40. Since the model divides by this close to zero amount, the function value increases towards infinity. This problem may also be seen in the residual plot on the right. Note that the model seems to perform very well with a pseudo-$R^2$ of 0.978. The only remedy is to discard this model.

This points out why investigating the plots is so important to the analysis.

Once the best model has been found, you would move the *Polynomial Model Fit - Y vs One X* procedure to analyze this model more closely.