Chapter 376

Polynomial Model Fit – Y vs Multiple X's

Introduction

This program fits a model that is the either a polynomial or a ratio of two polynomials. These polynomials may involve up to four independent variables (U, V, W, and X).

An example of a polynomial model is:

$$Y = A0 + A1X + \mathrm{A2U} + \mathrm{A3XU} + A4X^2 + A5U^2$$

An example of a ratio of polynomial models is:

$$Y = \frac{A0 + A1X + A2X^2 + A3U + A4U^2}{1 + B1X + B2X^2 + B3U + B4U^2}$$

These models approximate many different curves. They offer a wide variety of curves from which to choose. Since these are approximating curves and have no physical interpretation, care must be taken outside the range of the data. You must study the resulting model graphically to determine that the model behaves properly between data points.

Usually, you would use the *Polynomial Model Search – Y vs Multiple X's* procedure first to find an appropriate model and then fit that model with this program.

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.

Starting Values

Starting values are determined by the program. You do not have to supply starting values.

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 – Fitting a Multivariate Polynomial Model

This section presents an example of how to fit a multivariate polynomial model. In this example, we will fit a custom model to the variables Y, U, and X of the FnReg4 dataset. The particular form of this model was determined by the corresponding search procedure in shorthand notation as

E1,E2,S3.

When expanded, this model is

$$Y = A0 + A1U + A2U^2 + A3U^3 + A4X + A5UX + A6U^2X + A7X^2 + A8UX^2 + A9U^2X^2 + A10X^3$$

where

$Y = \ln(Y)$

$U = \ln(U)$

$X = \ln(X)$.

Setup

To run this example, complete the following steps:

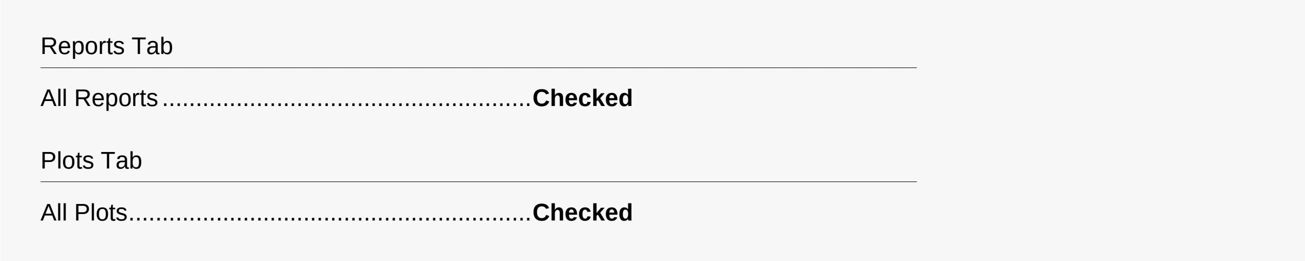
1 Open the FnReg4 example dataset

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

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

- Find and open the **Polynomial Model Fit – 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
Transform Y values before estimation .............Checked
  Y Transformation ...........................................ln(Y)
U Variable ........................................................U
Transform U values before estimation .............Checked
  U Transformation ...........................................ln(U)
X Variable ........................................................X
Transform X values before estimation .............Checked
  X Transformation ...........................................ln(X)
Type of Model ..................................................Polynomial
Polynomial Terms ............................................E1,E2,S3
```



Reports Tab

All Reports ....**Checked**

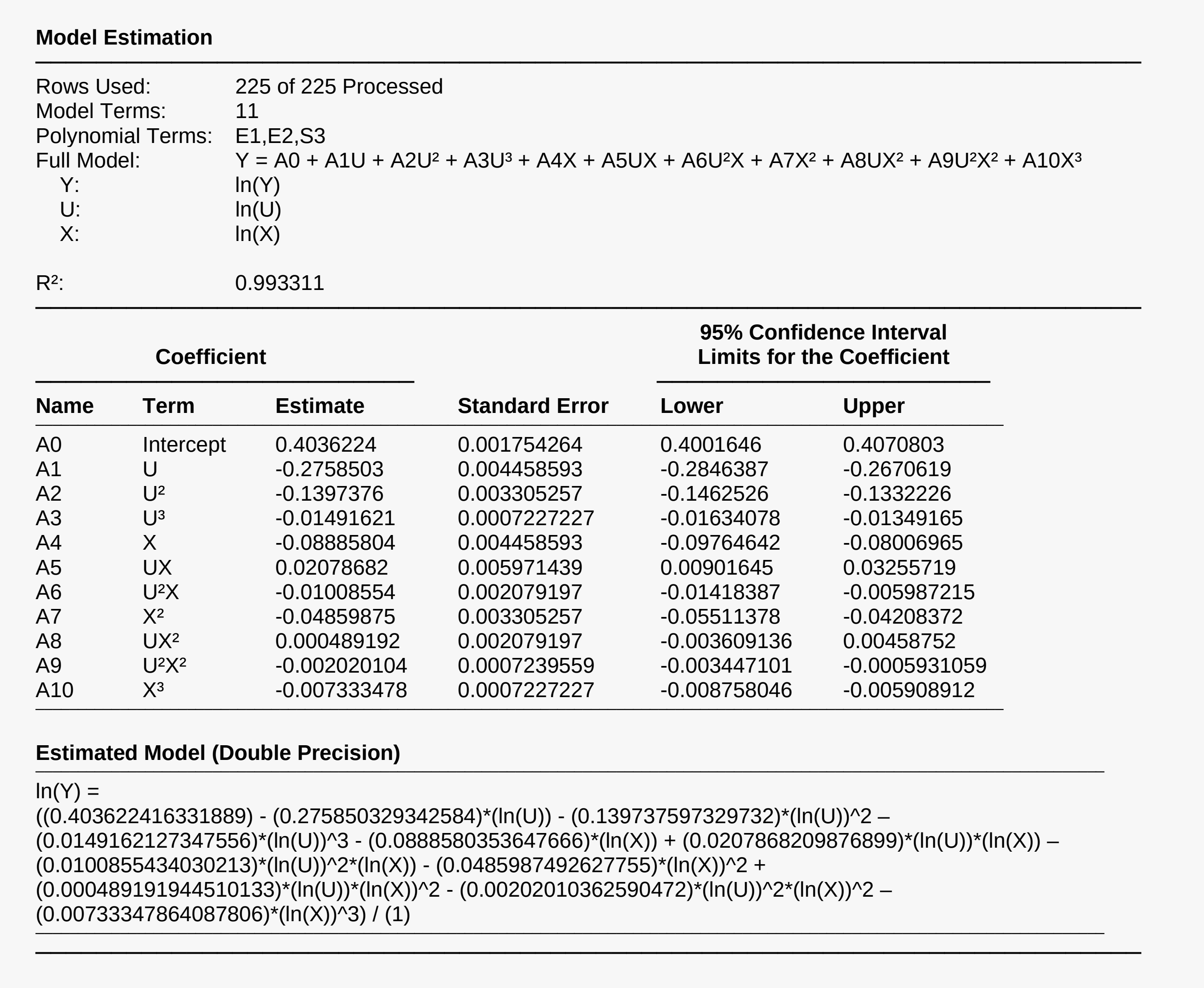
Plots Tab

All Plots...**Checked**

3 Run the procedure

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

Model Estimation



```
Model Estimation
─────────────────────────────────────────────────────────────────────────────────────────
Rows Used:          225 of 225 Processed
Model Terms:        11
Polynomial Terms:   E1,E2,S3
Full Model:         Y = A0 + A1U + A2U² + A3U³ + A4X + A5UX + A6U²X + A7X² + A8UX² + A9U²X² + A10X³
   Y:               ln(Y)
   U:               ln(U)
   X:               ln(X)

R²:                 0.993311
─────────────────────────────────────────────────────────────────────────────────────────

                                                          95% Confidence Interval
          Coefficient                                     Limits for the Coefficient
─────────────────────────────                          ──────────────────────────
Name      Term        Estimate       Standard Error   Lower           Upper
──────────────────────────────────────────────────────────────────────────────
A0        Intercept   0.4036224      0.001754264      0.4001646       0.4070803
A1        U           -0.2758503     0.004458593      -0.2846387      -0.2670619
A2        U²          -0.1397376     0.003305257      -0.1462526      -0.1332226
A3        U³          -0.01491621    0.0007227227     -0.01634078     -0.01349165
A4        X           -0.08885804    0.004458593      -0.09764642     -0.08006965
A5        UX          0.02078682     0.005971439      0.00901645      0.03255719
A6        U²X         -0.01008554    0.002079197      -0.01418387     -0.005987215
A7        X²          -0.04859875    0.003305257      -0.05511378     -0.04208372
A8        UX²         0.000489192    0.002079197      -0.003609136    0.00458752
A9        U²X²        -0.002020104   0.0007239559     -0.003447101    -0.0005931059
A10       X³          -0.007333478   0.0007227227     -0.008758046    -0.005908912
──────────────────────────────────────────────────────────────────────────────

Estimated Model (Double Precision)
──────────────────────────────────────────────────────────────────────────────────────
ln(Y) =
((0.403622416331889) - (0.275850329342584)*(ln(U)) - (0.139737597329732)*(ln(U))^2 –
(0.0149162127347556)*(ln(U))^3 - (0.0888580353647666)*(ln(X)) + (0.0207868209876899)*(ln(U))*(ln(X)) –
(0.0100855434030213)*(ln(U))^2*(ln(X)) - (0.0485987492627755)*(ln(X))^2 +
(0.000489191944510133)*(ln(U))*(ln(X))^2 - (0.00202010362590472)*(ln(U))^2*(ln(X))^2 –
(0.00733347864087806)*(ln(X))^3) / (1)
──────────────────────────────────────────────────────────────────────────────────────
─────────────────────────────────────────────────────────────────────────────────────────
```

This section reports the estimated coefficients.

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).

Model Terms

This is the number of terms in the model.

Polynomial Terms

This is the model that was input.

Full Model

The model that was fit in the expanded form with all terms listed.

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.

Coefficient Name

The name of the parameter whose results are shown on this line.

Coefficient Term

The name of the term in the model.

Coefficient Estimate

The estimated value of this parameter.

Standard Error

An estimate of the standard error of the parameter based on asymptotic (large sample) results.

Lower and Upper 95% Confidence Interval Limits for the Parameter

The lower and upper values of the 95% confidence interval limits for this parameter. This is a large sample (at least 25 observations for each parameter) confidence limit.

Estimated Model

This is a copy of the full model in which the coefficient names have been replaced by their double-precision estimates.

Analysis of Variance

Analysis of Variance

Source	DF	Sum of Squares	Mean Square
Mean	1	67.28918	67.28918
Model	11	68.47485	6.224987
Model (Adjusted)	10	1.185668	0.1185668
Error	214	0.007984756	3.731194E-05
Total (Adjusted)	224	1.193653	
Total	225	68.48284	

Source

The labels of the various sources of variation.

DF

The degrees of freedom.

Sum of Squares

The sum of squares associated with this term. Note that these sums of squares are based on Y, the dependent variable. Individual terms are defined as follows:

Mean The sum of squares associated with the mean of Y. This may or may not be a part of the model. It is presented since it is the amount used to adjust the other sums of squares.

Model The sum of squares associated with the model.

Model (Adjusted) The model sum of squares minus the mean sum of squares.

Error The sum of the squared residuals. This is often called the sum of squares error or just "SSE."

Total The sum of the squared Y values.

Total (Adjusted) The sum of the squared Y values minus the mean sum of squares.

Mean Square

The sum of squares divided by the degrees of freedom. The Mean Square for Error is an estimate of the underlying variation in the data.

Asymptotic Correlation Matrix of Parameters

```
Asymptotic Correlation Matrix of Parameters
───────────────────────────────────────────────────────────────────────────────

Section 1
             A0          A1          A2          A3          A4          A5          A6

A0     1.000000    0.722308    0.497426    0.328961    0.722308    0.562704    0.461919
A1     0.722308    1.000000    0.899575    0.730436    0.296523    0.544539    0.516329
A2     0.497426    0.899575    1.000000    0.947151    0.114328    0.242514    0.255763
A3     0.328961    0.730436    0.947151    1.000000    0.000000    0.000000    0.000000
A4     0.722308    0.296523    0.114328    0.000000    1.000000    0.544539    0.447007
A5     0.562704    0.544539    0.242514    0.000000    0.544539    1.000000    0.948195
A6     0.461919    0.516329    0.255763    0.000000    0.447007    0.948195    1.000000
A7     0.497426    0.114328    0.044081    0.000000    0.899575    0.242514    0.199077
A8     0.461919    0.447007    0.199077    0.000000    0.516329    0.948195    0.899073
A9     0.379184    0.423850    0.209954    0.000000    0.423850    0.899073    0.948195
A10    0.328961    0.000000    0.000000    0.000000    0.730436    0.000000    0.000000

Section 2
             A7          A8          A9         A10

A0     0.497426    0.461919    0.379184    0.328961
A1     0.114328    0.447007    0.423850    0.000000
A2     0.044081    0.199077    0.209954    0.000000
A3     0.000000    0.000000    0.000000    0.000000
A4     0.899575    0.516329    0.423850    0.730436
A5     0.242514    0.948195    0.899073    0.000000
A6     0.199077    0.899073    0.948195    0.000000
A7     1.000000    0.255763    0.209954    0.947151
A8     0.255763    1.000000    0.948195    0.000000
A9     0.209954    0.948195    1.000000    0.000000
A10    0.947151    0.000000    0.000000    1.000000
───────────────────────────────────────────────────────────────────────────────
```

This report displays the asymptotic correlations of the parameter estimates. When these correlations are high (absolute value greater than 0.95), the precision of the parameter estimates is suspect.

Predicted Values and Residuals

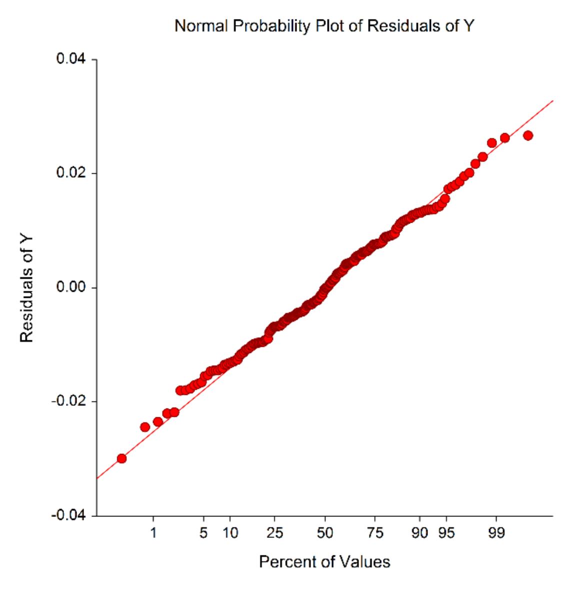
Predicted Values and Residuals

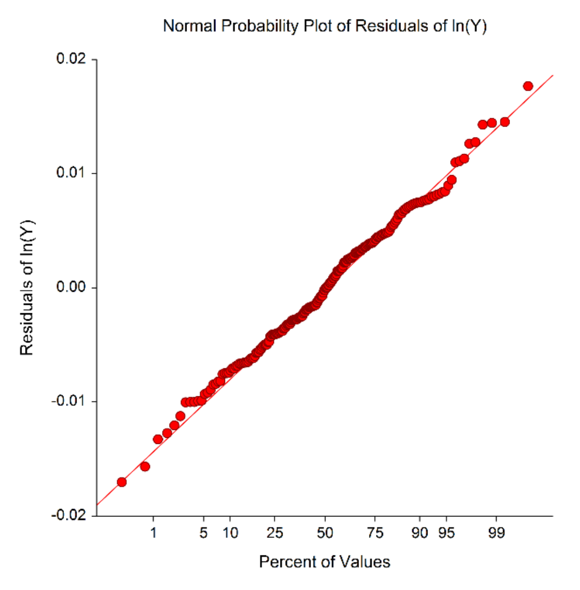
			95% Prediction Interval		
Row Number	Y	Predicted Value	Lower	Upper	Residual
1	1.981996	1.985781	1.954117	2.017957	-0.00378463
2	2.028455	2.027117	2.000246	2.054349	0.001337741
3	2.027451	2.017885	1.991534	2.044585	0.009565716
4	1.982509	1.987645	1.961936	2.01369	-0.005135759
5	1.955915	1.948343	1.923401	1.973608	0.007572046
6	1.903501	1.905211	1.881063	1.929668	-0.001709758
7	1.859508	1.860794	1.837407	1.884478	-0.0012858
8	1.829529	1.816414	1.793718	1.839397	0.01311511
9	1.742849	1.77278	1.750686	1.795152	-0.02993087
10	1.734567	1.730272	1.708683	1.752134	0.00429506

The section shows the values of the residuals and predicted values. If you have observations in which the independent variables are given, but the dependent (Y) variable was left blank, a predicted value and prediction limits will be generated and displayed in this report.

Probability Plot(s)

Probability Plot(s)



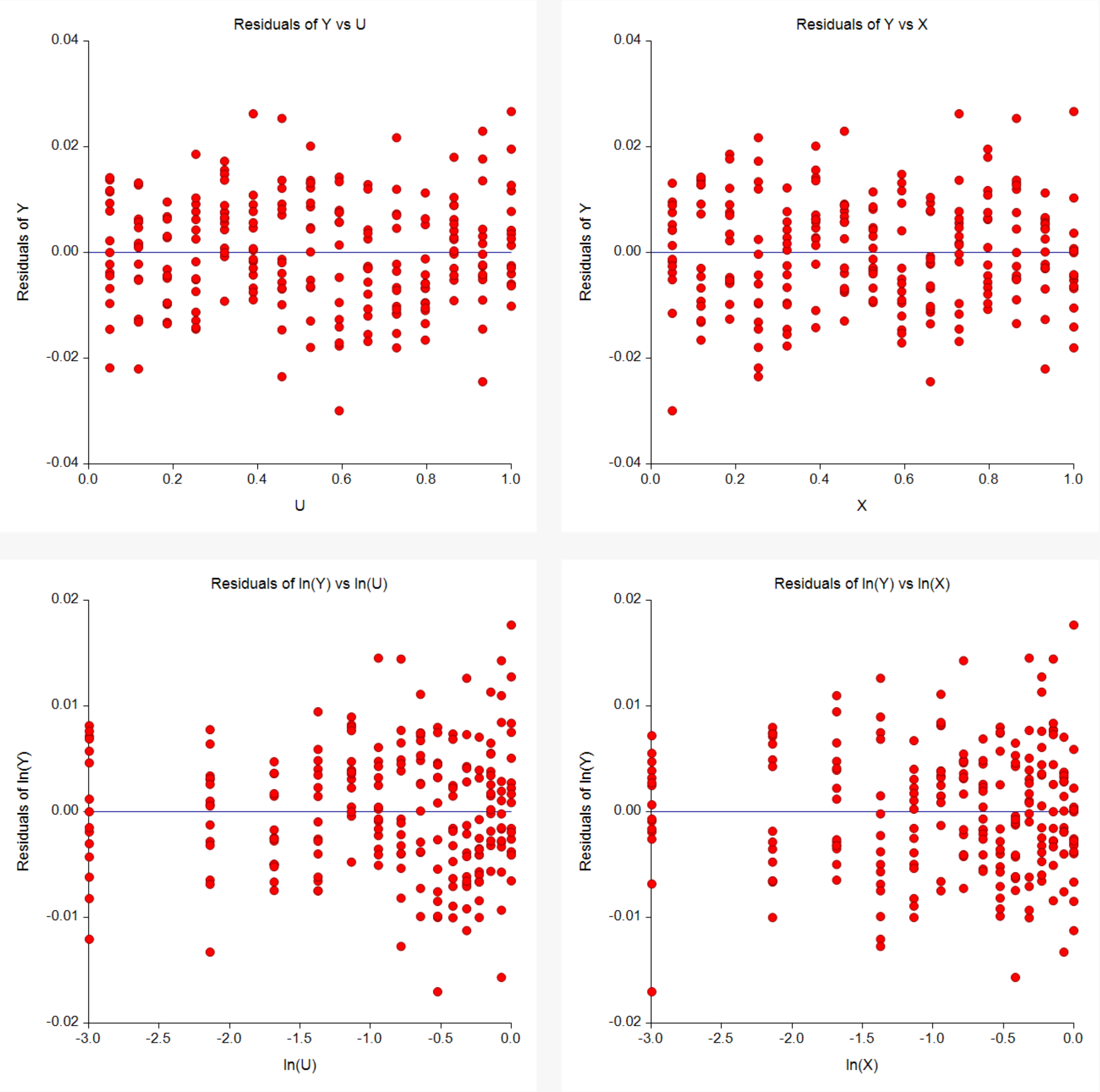


If the residuals are normally distributed, the data points of the normal probability plot will fall along a straight line. Major deviations from this ideal picture reflect departures from normality. Stragglers at either end of the normal probability plot indicate outliers, curvature at both ends of the plot indicates long or short distributional tails, convex or concave curvature indicates a lack of symmetry, and gaps or plateaus or segmentation in the normal probability plot may require a closer examination of the data or model. We do not recommend that you use this diagnostic with small sample sizes.

In this example, the Y variable was transformed to ln(Y). So the most appropriate plot is the second which shows the residuals of ln(Y).

Residual Plot(s)

Residual Plot(s)



These are scatter plots of the residuals versus each of the independent variables. Note that some of these plots show the data in the transformed (ln(Y), ln(U), and ln(X)) scale.

The preferred pattern is a rectangular shape or point cloud. Any nonrandom pattern may require a redefining of the model.

Predicting for New Values

You can use your model to predict Y for new values of the independent variables. Here's how. Add new rows to the bottom of your database containing the values of the independent variables that you want to create predictions for. Leave the dependent variable blank. When the program analyzes your data, it will skip these rows during the estimation phase, but it will generate predicted values for all rows, regardless of whether the Y variable is missing or not.