Chapter 382

Fractional Polynomial Regression – Y vs One X

Introduction

This program fits fractional polynomial models in situations in which there is one dependent (Y) variable and one independent (X) variable. It creates a model of the variance of Y as a function of X. Using these two models, it calculates reference intervals for Y and stipulated X values.

Fractional Polynomial Model

A generalization of the polynomial function, called fractional polynomials (FP for short), was proposed by Royston and Altman (1994) and Royston and Sauerbrei (2008). FPs are of the form

$$Y = B0 + B1\,X^{P1} + B2\,X^{P2} + \cdots$$

where P1, P2, ... are exponents selected from {-2, -1, -0.5, 0, 0.5, 1, 2, 3}. The convention is that X^0 equals LN(X). Hence the model FP(1, 0, -2) is

$$Y = B0 + B1\,X + B2\,\mathrm{LN}(X) + B3\,1/X^2$$

An additional extension is with models that involve repeated powers such as (1, 1). Here, the second term is multiplied by LN(X). For example, the model FP(2, 2) is

$$Y = B0 + B1\,X^2 + B2\,(X^2)\,\mathrm{LN}(X)$$

It turns out the models that involve only two terms are usually adequate.

Reference Interval

Consider a measurement made on a well-defined population of individuals. A **reference interval** (RI) of this measurement gives the boundaries between which a typical measurement is expected to fall. When a measurement occurs that is outside these reference interval boundaries, the individual is said to be abnormal. That is, the measurement is unusually high or low.

The reference interval is often presented as percentiles of a reference population, such as the 2.5^{th} percentile and the 97.5^{th} percentile. Of course, the choice of a reference population is crucial, and you would expect that the interval varies according to age, region, gender, and so on.

This procedure estimates an **X-specific reference interval** for cross-sectional studies using the methodology of Altman (1993), Royston and Wright (1998), and Royston and Sauerbrei (2008). It provides formulas that may be used to produce percentiles as well as z-scores for new measurements not included in the original analysis.

This methodology gives results that are similar to those obtained by quantile regression.

Technical Details

Data Collection

The data should only include one measurement pair per subject. It is desirable to have approximately equal numbers of individuals at each value of X.

Models of the Mean and Standard Deviation (SD)

The fundamental assumption of this method is that at each X-value, the measurement of interest is normally distributed with a given mean and standard deviation. Furthermore, the means and standard deviations are smooth functions across X. Various types of models are available to model the mean and SD functions, including polynomial, fractional polynomial, and ratios of fractional polynomials.

The reference interval equation takes the form

$$Y = M(X) + z_\alpha \, SD(X), \quad 0 < X < \infty$$

where X is the independent variable, M(X) is an estimate of the mean of Y at X, SD(X) is an estimate of the standard deviation of Y at X, and z_α is the appropriate percentile of the standard normal distribution. M(X) is estimated using nonlinear least squares.

SD(X) is estimated using a separate (possibly nonlinear) least squares regression in which Y is replaced by the scaled absolute residuals. The residuals are scaled so that they directly estimate the SD of Y at each X. The scaling of the residuals (Y - M(X)) is accomplished by multiplying them by √(π/2) (which is approximately equal to 1.2533). This scale factor is based on the normal distribution.

The Six Step Estimation Process

The following six step procedure was suggested by Altman and Chitty (1994).

Step 1 – Fit the Mean Function

The first step is to fit the mean function with a reasonable, well-fitting model. This is usually accomplished by fitting a polynomial, a fractional polynomial, or the ratio of two fractional polynomials. Also, the possibility of transforming Y using the logarithm, square root, or some other power transformation function is considered.

During this step, various models are investigated by considering the goodness-of-fit (R^2), the Y-X scatter plot, and the residual versus X plot.

Step 2 – Study the Residuals from the Mean Fit

During this step, the residuals between the data and the fitted line are examined more closely. Often, the vertical spread of the residuals changes with X. This heteroscedasticity will be treated in the next step. But another feature that should be considered is whether the residuals are symmetric or skewed about zero across X. Skewing is not modelled during the next step, so it must be fixed before proceeding to step 3. Skewing is usually corrected by using the logarithm of Y instead of Y itself.

Step 3 – Fit a Standard Deviation Function

The next step is to estimate an SD function. This is usually accomplished by fitting a linear polynomial to the scaled absolute residuals (SAR). The scaling factor is √(π/2). Occasionally, a quadratic polynomial is required, but usually nothing more complicated than a linear polynomial is needed.

Step 4 – Calculate Z-Scores

The next step is to calculate a z-score for each observation. The z-score for the k^{th} observation is calculated using

$$Z_k = \frac{Y_k - M(X_k)}{SD(X_k)}$$

Step 5 – Check the Goodness-of-Fit of the Models

The first item to consider is the value of R^2. This value should be as high as possible, although a high R^2 is not the only consideration. But it is a starting point. The plot of the fit of the mean overlaid on the X-Y plot allows you visually determine whether the model is appropriate.

The z-scores should be checked to determine that they are approximately normal. This can be done by looking at a normal probability plot of the z-scores and by considering the results of a normality test such as the Shapiro-Wilk test.

Step 6 – Calculate the Reference Interval

The final step is to calculate the reference interval at various values of X. The reference interval is defined by two percentile boundaries that depend on X and the percentile. Often, a 95% reference interval is desired. This is based on the 2.5^{th} and the 97.5^{th} percentiles. The formula for these values is

$$Y_{(X,\alpha)} = M(X) + z_\alpha \, SD(X)$$

Fractional Polynomials

A polynomial function is of the form

$$Y = B0 + B1\,X + B2\,X^2 + B3\,X^3 + \cdots$$

where the exponents of X are non-negative integers. Although popular, low order polynomials suffer from many deficiencies. They offer only a few model shapes which often do not fit the data well, especially near the ends of the data range. Also, polynomial functions do not have asymptotes, so they can't model this type of behavior.

A generalization of the polynomial function, called *fractional polynomials* (FP for short), was proposed by Royston and Altman (1994) and Royston and Sauerbrei (2008). FPs are of the form

$$Y = B0 + B1\,X^{P1} + B2\,X^{P2} + \cdots$$

where P1, P2, ... are selected from {-2, -1, -0.5, 0, 0.5, 1, 2, 3}. The convention is that X^0 equals LN(X). Hence the model FP(1, 0, -2) is

$$Y = B0 + B1\,X + B2\,\mathrm{LN}(X) + B3\,1/X^2$$

An additional extension is with models that involve repeated powers such as (1, 1). In this case, the second term is multiplied by LN(X). For example, the model FP(2, 2) is

$$Y = B0 + B1\,X^2 + B2\,X^2\,\mathrm{LN}(X)$$

It turns out the models that involve only two terms are usually adequate for creating reference intervals.

Ratio of Two Fractional Polynomials

Another useful extension that **NCSS** provides is the availability of ratios of fractional polynomials. These models are of the form

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

These models approximate many different curve shapes. They offer a wide variety of curves and often provide better fitting models than polynomials and fractional polynomials. Unfortunately, the presence of the terms in the denominator can cause severe problems since the denominator can become zero. When this happens, the model must be discarded.

Data Structure

The data are entered in two variables: one for Y and one for X.

Missing Values

Rows with missing values in the variables being analyzed are ignored in the calculations. If transformations are used which limit the range of X and Y (such as the logarithm), observations that cannot be transformed are treated as missing values.

Example 1 – Creating a Reference Interval Equation

This section presents an example of how to create a reference interval equation from a set of gestation data. In this dataset, the length of gestation (Gestation) and an ultrasonic measurement (Response) of 100 individuals is recorded. The program will conduct a search of 44 possible models and select the model that fits the data the best. A straight-line linear regression model appeared to fit the scaled absolute residuals. These models will be used to create the reference interval equation.

Setup

To run this example, complete the following steps:

1 Open the ReferenceInterval example dataset

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

2 Specify the Fractional Polynomial Regression – Y vs One X procedure options

- Find and open the **Fractional Polynomial Regression – 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 (Response) Variable	**Response**
X (Covariate) Variable	**Gestation**
Model Type	**Find the Best Fitting Fractional Polynomial**
X	**Checked**

Reports Tab	
Model Search: Candidates Models	**Checked**
Model Summary	**Checked**
Coefficient Estimation	**Checked**
Analysis of Variance Tables	**Checked**
Coefficient Correlation Matrix	**Checked**
Normality Test	**Checked**
Percentiles	**Checked**
Percentiles	**2.5 10 25 50 75 90 97.5**
Xs	**(5)**
Predicted Values and Residuals	**Checked**

3 Run the procedure

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

Best Model from Search

Best Model from Search

Item	Mean Model	Standard Deviation Model
Y Variable	Response	Scaled Absolute Residuals
X Variable	Gestation	Gestation
Rows Used	100 of 100	100 of 100
Residual Scale Factor	√(pi/2)	
Models Tried	44	1
Selected Model	A0 + A1X² + A2(1/X)²	C0 + C1X
R² of Selected Model	0.857448	0.125270
SE = √(MSE)	0.04496217	0.03399233

This report summarizes the fitting of the two models: the first column of the Mean and the second column of the Standard Deviation.

Variable Names

These entrees give the names of the X and Y variables.

Rows Used

The number of rows used in the calculations. This is the number of rows with non-missing values in both X and Y. This number is followed by the number of rows read.

Residual Scale Factor

During the estimation of the standard deviation model, each residual is multiplied by this value.

Models Tried

The number of models considered during the search for the best fitting model.

Selected Model

The selected model in symbolic form.

R² of Selected Model

This value is computed in the usual way for models that do not include a denominator polynomial. When a denominator is included, this value is only approximately correct.

R^2 varies between 0 and 1, with 0 indicating a poor fit and 1 indicating a perfect fit. Note that the R^2 of the standard deviation model will usually be close to zero. That is okay.

The R^2 value allows you to compare various models. This value, combined with the plots, is used to determine the best fitting model.

SE

An estimate of the standard error.

Model Search: Candidate Models Sorted by R^2

Model Search: Candidate Models Sorted by R^2

	Mean Model			Standard Deviation Model		
Rank	Equation	R^2	R^2 - Best R^2	Equation	R^2	Normality Test of Z-Scores P-Value
1	**A0 + A1X² + A2(1/X)²**	**0.857448**	**0.000000**	**C0 + C1X**	**0.125270**	**0.6878**
2	A0 + A1X + A2(1/X)²	0.857414	-0.000034	C0 + C1X	0.124475	0.6539
3	A0 + A1(1/X)² + A2X³	0.857375	-0.000074	C0 + C1X	0.126365	0.7199
4	A0 + A1√(X) + A2(1/X)²	0.857353	-0.000095	C0 + C1X	0.123920	0.6360
5	A0 + A1(1/X) + A2X³	0.857294	-0.000154	C0 + C1X	0.122213	0.5748
6	A0 + A1ln(X) + A2(1/X)²	0.857262	-0.000187	C0 + C1X	0.122963	0.6127
7	A0 + A1(1/X) + A2X²	0.857208	-0.000240	C0 + C1X	0.121813	0.5804
8	A0 + A1(1/√(X)) + A2(1/X)²	0.857139	-0.000309	C0 + C1X	0.121890	0.6070
9	A0 + A1X + A2(1/X)	0.857133	-0.000316	C0 + C1X	0.121470	0.5902
10	A0 + A1√(X) + A2(1/X)	0.857100	-0.000348	C0 + C1X	0.121320	0.5954
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

Rank

The rank number after sorting the models by R^2.

Mean Model: Equation

The generic model of the mean being reported on in this row.

Mean Model: R^2

The R^2 value of this model.

Mean Model: R^2 - Best R^2

The difference between the R^2 value of this model and the R^2 value of the best model encountered.

Standard Deviation Model: Equation

The generic model of the standard deviation reported on in this row.

Standard Deviation Model: R^2

The R^2 value of this model.

Normality Test of Z-Scores P-Value

The p-value of the Shapiro-Wilk normality test of the z-scores. If this value is greater than 0.05, there is not enough evidence to conclude that the data are not normally distributed.

Coefficient Estimation Reports

Mean Equation - Coefficient Estimation

Model: $Y = A0 + A1X^2 + A2(1/X)^2$

		Coefficient		95% Confidence Interval Limits for the Coefficient			
Name	**Term**	**Estimate**	**Standard Error**	**Lower**	**Upper**	**T-Value**	**P-Value**
A0	Intercept	10.31614	0.03384301	10.24897	10.38331	304.82	0.0000
A1	X^2	-8.090378E-05	2.342309E-05	-0.0001273921	-3.441544E-05	-3.45	0.0008
A2	$1/X^2$	75.65155	9.254083	57.28476	94.01834	8.17	0.0000

Estimated Model of Response (Double Precision)

```
Response =
(10.3161380531194 - (8.09037797269359E-05)*Gestation^2 + (75.6515482561129)*1/(Gestation*Gestation))
```

Standard Deviation Equation - Coefficient Estimation

Model: Scaled Absolute Residuals = C0 + C1X

		Coefficient		95% Confidence Interval Limits for the Coefficient			
Name	**Term**	**Estimate**	**Standard Error**	**Lower**	**Upper**	**T-Value**	**P-Value**
C0	Intercept	-0.00397401	0.01280035	-0.02937588	0.02142786	-0.31	0.7569
C1	X	0.001716751	0.0004582565	0.0008073563	0.002626146	3.75	0.0003

Estimated Model of Scaled Absolute Residuals of Response (Double Precision)

```
Scaled Absolute Residuals of Response =
(-0.00397401029375437 + (0.0017167513612774 3)*Gestation)
```

Estimated Z-Score Model (Double Precision)

```
Z =
(Response - (10.3161380531194 - (8.09037797269359E-05)*Gestation^2 + (75.6515482561129)*1/(Gestation*Gestation))) /
(-0.00397401029375437 + (0.0017167513612774 3)*Gestation)
```

Coefficient and Term

The name of the coefficient and term whose results are shown on this line.

Coefficient Estimate

The estimated value of this coefficient.

Coefficient Standard Error

An estimate of the standard error of the coefficient.

Lower and Upper 95% Confidence Interval Limits for the Coefficient

The lower and upper limits of a 95% confidence interval for this coefficient.

T-Value

The value of the t-statistic used to test whether this term is statistically significant.

P-Value

The significance level or p-value of the test statistic. If this value is 0.05 or less, the t-test is statistically significant.

Estimated Model

This is the estimated model written out so that it can be copied and pasted into another program such as Excel.

Estimated Z-Score Model

This is the estimated z-score model written out so that it can be copied and pasted into another program such as Excel.

Shapiro-Wilk Normality Test of Z-Scores

Shapiro-Wilk Normality Test of Z-Scores

Test Name	Test Statistic	P-Value	Reject Normality at 5% Level?
Shapiro-Wilk	0.99	0.6878	No

This report shows the result of a test of the normality of the z-scores. If normality is rejected, a different model should be used, possibly one that uses LN(y).

Percentile Report

Percentile Report

Model: $Y = A0 + A1X^2 + A2(1/X)^2 + Z\alpha(C0 + C1X)$

	Percentiles of Response						
Gestation	2.5	10.0	25.0	50.0	75.0	90.0	97.5
8	11.47389	11.48051	11.48643	11.49302	11.4996	11.50552	11.51215
16	10.54489	10.56083	10.57509	10.59094	10.60679	10.62105	10.63699
24	10.32791	10.35317	10.37577	10.40088	10.42599	10.44859	10.47384
32	10.20729	10.24186	10.2728	10.30717	10.34154	10.37248	10.40705
40	10.10717	10.15106	10.19034	10.23397	10.27761	10.31689	10.36078

This report shows the estimated percentiles at the Gestation values and Percentile values that were selected. Note that 'Z' stands for standard normal deviate corresponding to the indicated percentile.

Analysis of Variance Tables

Mean Equation - Analysis of Variance

Model Term(s)	DF	Sum of Squares	Mean Square
Mean	1	10787.3	10787.3
Model	3	10788.48	10788.61
Model (Adjusted)	2	1.179512	0.5897558
Error	97	0.1960949	0.002021597
Total (Adjusted)	99	1.375606	
Total	100	10788.67	

Standard Deviation Equation - Analysis of Variance

Model Term(s)	DF	Sum of Squares	Mean Square
Mean	1	0.1785716	0.1785716
Model	2	0.1947882	0.2514067
Model (Adjusted)	1	0.01621659	0.01621659
Error	98	0.1132369	0.001155479
Total (Adjusted)	99	0.1294535	
Total	100	0.3080251	

Model Term(s)

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

Total The sum of the squared Y values.

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.

Correlation Matrix of Parameters

Mean Equation - Coefficient Correlation Matrix

	A0	A1	A2
A0	1.000000	-0.965022	-0.958157
A1	-0.965022	1.000000	0.882923
A2	-0.958157	0.882923	1.000000

Standard Deviation Equation - Coefficient Correlation Matrix

	C0	C1
C0	1.000000	-0.964095
C1	-0.964095	1.000000

This report displays the correlations of the coefficient estimates.

Predicted Values and Residuals Section

Predicted Values, Residuals, and Z-Scores

Model: $Y = A0 + A1X^2 + A2(1/X)^2$

Row	Gestation (X)	Response (Y)	Predicted Response (Yhat|X)	Residual	Scaled Residual	Standard Deviation of Y	Z-Score of Y	
							Value	P-Value
1	38.26939	10.24195	10.24931	-0.007357729	-0.009221545	0.06172502	-0.12	0.4526
2	30.56216	10.29393	10.32156	-0.02763455	-0.03463478	0.04849362	-0.57	0.2844
3	21.19614	10.42426	10.44818	-0.02391429	-0.02997211	0.03241448	-0.74	0.2303
4	22.50685	10.50074	10.4245	0.07624204	0.09555522	0.03466465	2.20	0.9861
5	33.06033	10.33878	10.29693	0.0418499	0.05245107	0.05278236	0.79	0.7861
6	22.33047	10.4488	10.42751	0.02129137	0.02668477	0.03436186	0.62	0.7322
7	35.60615	10.34229	10.27324	0.06905431	0.08654675	0.05715289	1.21	0.8865
8	34.34106	10.27968	10.28488	-0.005198971	-0.006515944	0.05498106	-0.09	0.4623
9	30.76532	10.32139	10.31949	0.001897267	0.002377871	0.0488424	0.04	0.5155
10	15.66574	10.5926	10.60454	-0.01193815	-0.01496225	0.02292017	-0.52	0.3012

This report shows the predicted values, residuals, and z-scores.

Row

The row number from the dataset.

X

The value of the covariate.

Y

The value of the response.

Predicted Y (Yhat|X)

The predicted value of the response using only the mean model.

Residual

The value of the residual, the difference between Y and the predicted Y.

Scaled Residual

The value of the residual times the scale factor.

Standard Deviation of Y

The value of the standard deviation using the standard deviation model.

Z-Score of Y: Value

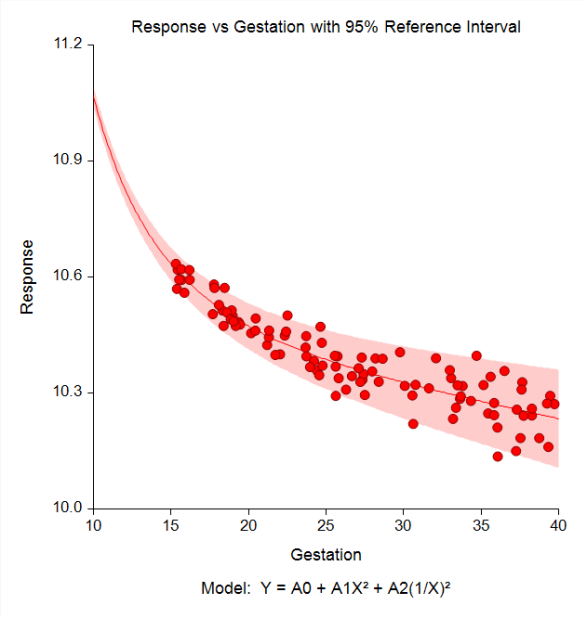
The z-score of this row. Most z-scores should be between ±2 if the data are normally distributed.

Z-Score of Y: P-Value

The probability level of the above z-score assuming the normal distribution.

Y vs X with Reference Interval

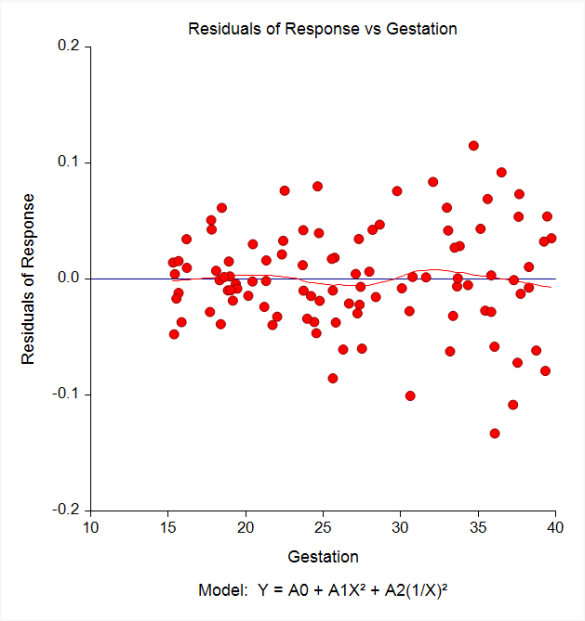
Plot of Y vs X with Reference Interval



This plot displays the data along with the estimated function and reference interval. It is useful in deciding if the fit is adequate and the reference interval is appropriate.

Residuals of Y vs X

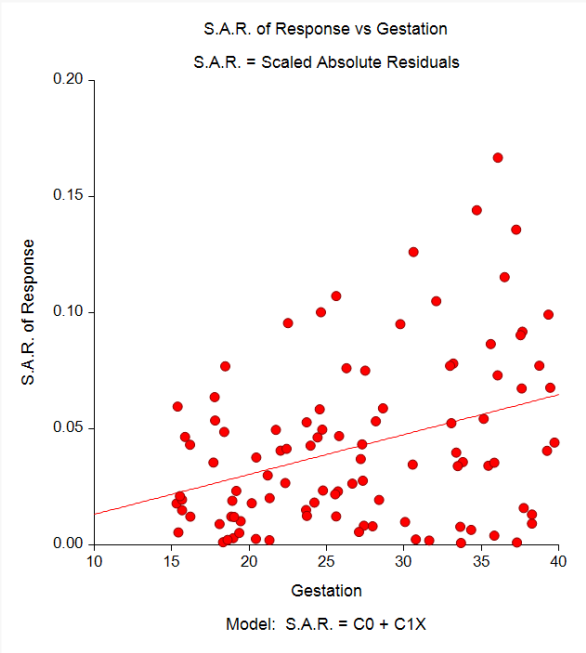
Plot of Residuals of Y vs X



This is a scatter plot of the residuals versus the independent variable, X. The preferred pattern is a rectangular shape or point cloud. Any nonrandom pattern may require a redefining of the model. A loess curve is overlaid to give you a better understanding of the trends in the data.

Scaled Absolute Residuals vs X

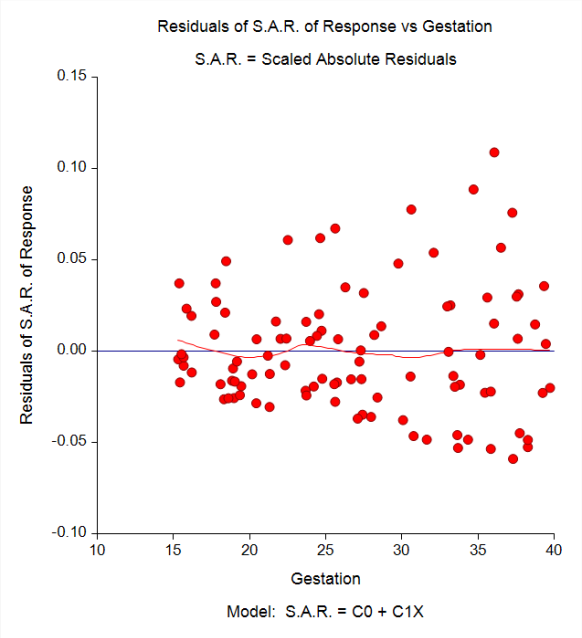
Plot of Scaled Absolute Residuals vs X



This a scatter plot of the scaled absolute residuals versus X. The line is the model of the standard deviation.

Residuals of Scaled Absolute Residuals vs X

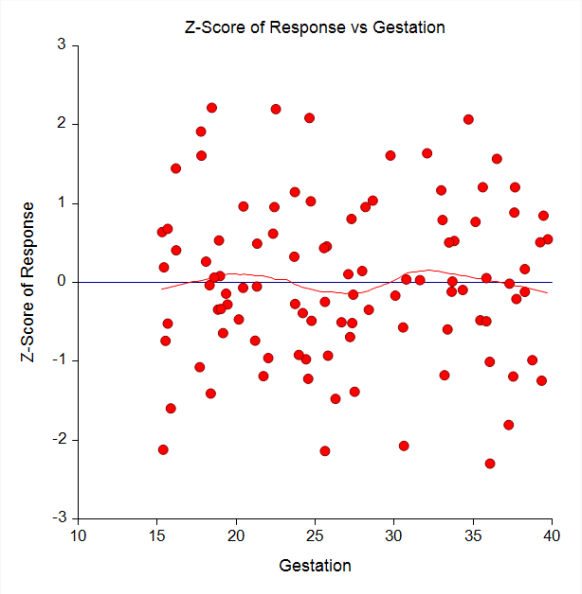
Plot of Residuals of Scaled Absolute Residuals vs X



This is a scatter plot of the residuals from the S.A.R. fit versus the independent variable, X. Often, the plot will exhibit a funnel shape indicating the changing nature of these residuals. This is to be expected. A loess curve is overlaid to give you a better understanding of any patterns that should be modelled.

Z-Scores vs X

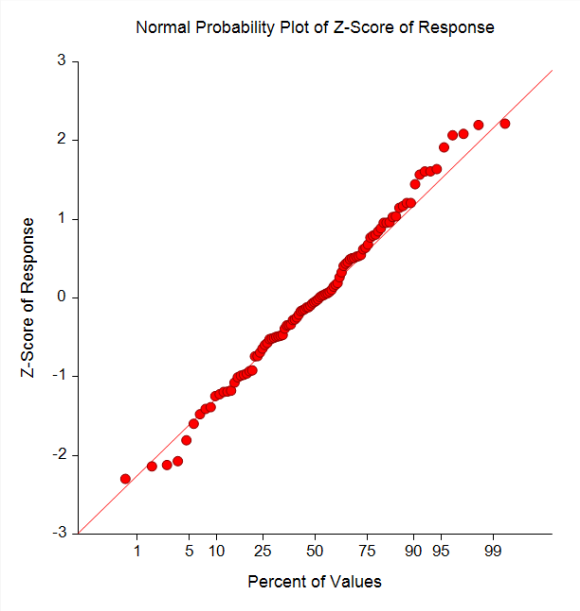
Plot of Z-Scores vs X



This scatter plot displays the z-scores versus the covariate, X. If all has gone well, this plot should show a random pattern.

Normal Probability Plot of Z-Scores

Normal Probability Plot of Z-Scores



If the z-scores 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.