Piecewise Linear Regression

Revised: 6/28/2020

Summary

The **Piecewise Linear Regression** procedure is designed to fit a regression model where the relationship between the dependent variable Y and the independent variable X is a continuous function consisting of 2 or more linear segments. The function is estimated using nonlinear least squares. The user specifies the number of segments and initial estimates of the locations where the segments join.

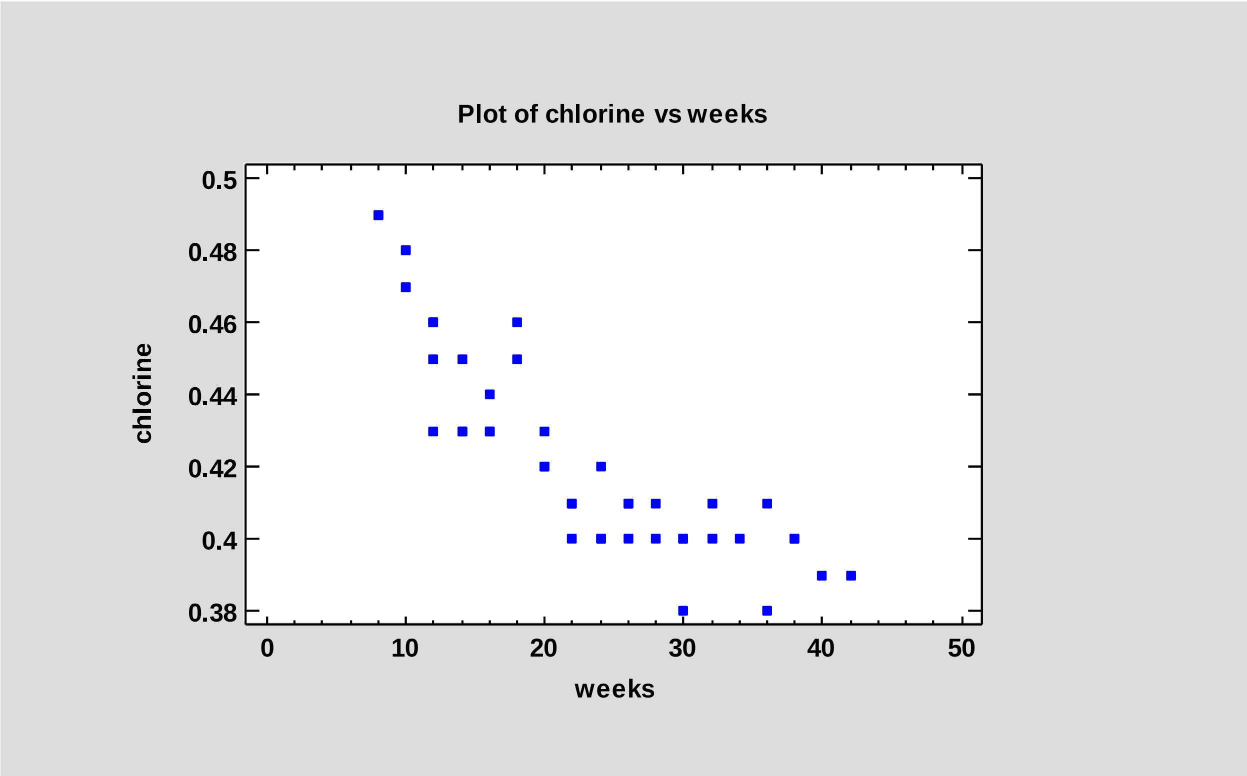
Sample StatFolio: *piecewise.sgp*

Sample Data:

The file *nonlin.sgd* contains data on the amount of available chlorine in samples of a product as a function of the number of weeks since it was produced. The data, from Draper and Smith (1998), consist of n = 44 samples, a portion of which are shown below:

Weeks	*Chlorine*
8	0.49
8	0.49
10	0.48
10	0.47
10	0.48
10	0.47
12	0.46
12	0.46
12	0.45
12	0.43
…	…

As may be seen in the plot below, the relationship between *chlorine* and *weeks* is not linear:



While various types of curvilinear models might be fit to the data, a model consisting of 3 linear segments might also be considered with slope changes in the vicinity and 15 and 25 weeks.

Statistical Model

The statistical model fit by this procedure consists of k linear segments. It may be represented by:

$$Y = \beta_0 + \beta_1 X + \sum_{j=1}^{k-1} \beta_{j+1}(X - \Delta_j)\, I(X - \Delta_j) \tag{1}$$

where

β_0 = Y-intercept

β_j = slope of segment j, j=1, 2, …, k

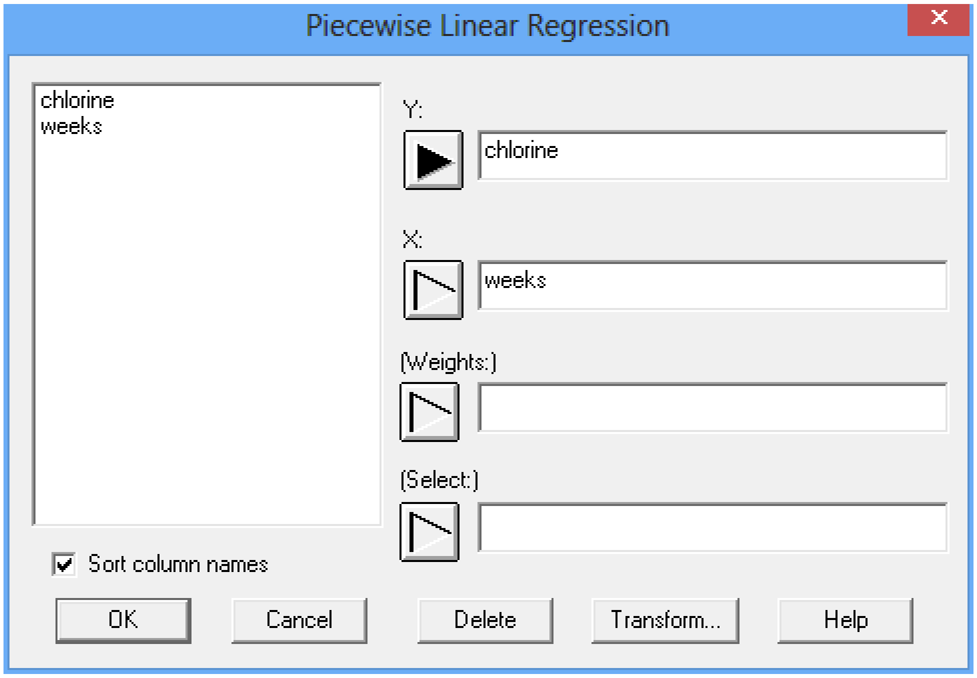
Δ_j = location of slope changes between segment j and segment j+1, j=1, 2, …, k-1

$I(X - \Delta_j)$ = 1 if $X \geq \Delta_j$ and 0 otherwise

The locations Δ_j at which the slope changes may be specified by the user or estimated by the model fitting procedure.

Data Input

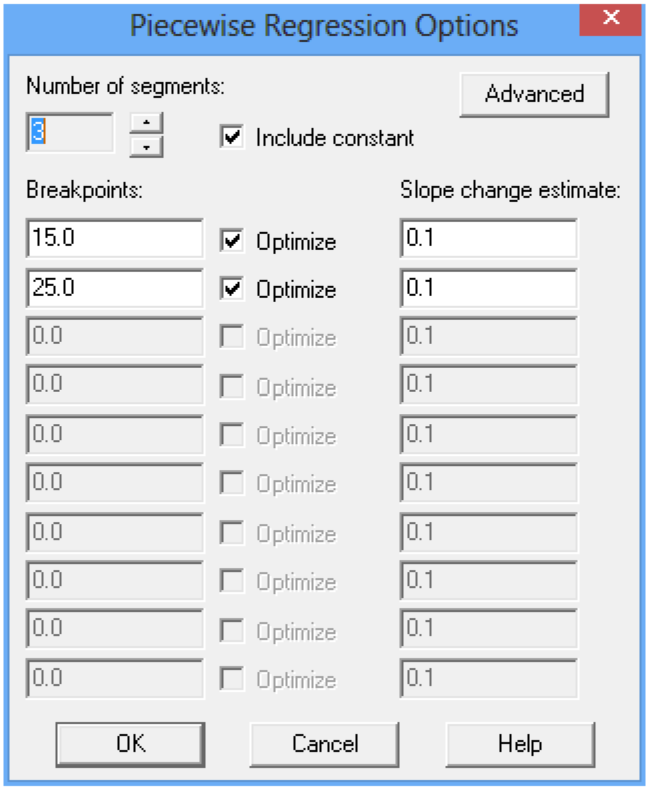
The data input dialog box requests information about the input variables:



- **Y**: a numeric variable containing the *n* values of the dependent variable.
- **X**: a numeric variable containing the *n* values of the independent variable.
- **Weights:** optional weights to be applied to each of the *n* observations.
- **Select**: optional subset selection.

Analysis Options

The *Analysis Options* dialog box is used to specify the number of linear segments and initial estimates for the breakpoints and slope changes:



- **Number of segments:** the number of piecewise linear segments in the model.

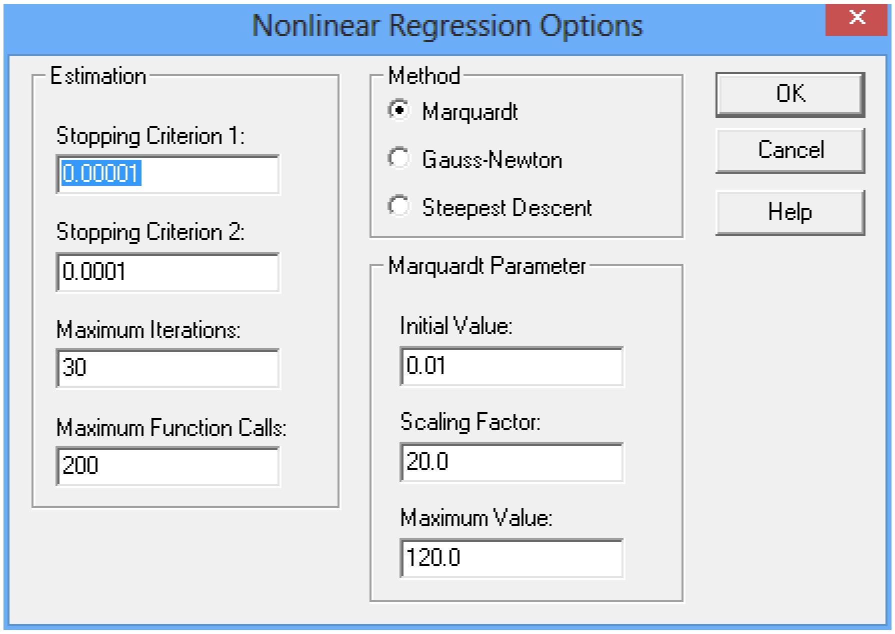
- **Include constant:** whether the constant term β_0 should be included in the model. If not checked, the model will be forced to go through the origin.

- **Breakpoints:** initial estimates of the locations Δ_j where the slope changes.

- **Optimize:** whether the model fitting procedure should optimize the location of the breakpoints. If not, the breakpoints will be fixed at the specified values.

- **Slope change estimates:** initial estimates of the amount at which the slope changes at each breakpoint. Since a numerical procedure is being used to minimize the residual sum of squares, care should be given to providing good estimates of the amount of change in the slope at each breakpoint. It is particularly important that each slope change have the correct sign (positive or negative).

You may control the options used by the nonlinear regression estimation procedure by pressing the *Advanced* button, which displays the following dialog box:



- **Method:** method used to estimate the model parameters. The *Gauss-Newton* method uses a linearization technique that fits a sequence of linear regression models to locate the minimum residual sum of squares. The *Steepest-Descent* method follows the gradient of the residual sum of squares surface. *Marquardt's* method, the default, is a fast and reliable compromise between the other two.

- **Stopping Criterion 1:** The algorithm is assumed to have converged when the relative change in the residual sums of squares from one iteration to the next is less than this value.

- **Stopping Criterion 2:** The algorithm is assumed to have converged when the relative change in all parameter estimates from one iteration to the next is less than this value.

- **Maximum Iterations:** Estimation stops if convergence is not achieved within this many iterations.

- **Maximum Function Calls:** Estimation stops if convergence is not achieved when the function being fit has been evaluated this many times. Multiple function evaluations are done during each iteration.

- **Marquardt Parameter:** The magnitude of the Marquardt parameter controls the extent to which the other two methods are traded off against each other. For details on the Marquardt algorithm, see Box, Jenkins and Reinsel (1994).

Analysis Summary

The *Analysis Summary* summarizes the fitted model:

Piecewise Linear Regression - chlorine
Dependent variable: chlorine (percent available)
Independent variable: weeks (weeks since production)
Number of observations: 44

Coefficients

		Asymptotic	*Asymptotic*	*Asymptotic*
Parameter	*Estimate*	*Standard Error*	*Lower 95.0% C.I.*	*Upper 95.0% C.I.*
Intercept	0.575714	0.0240969	0.526933	0.624496
Initial slope	-0.0103571	0.00229338	-0.0149999	-0.00571442
Slope change #1	0.00674228	0.00241319	0.00185702	0.0116275
Change point #1	12.0915	1.15966	9.74392	14.4392
Slope change #2	0.00281699	0.000918817	0.000956942	0.00467704
Change point #2	24.448	2.46663	19.4545	29.4414

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*
Model	0.035023	5	0.0070046
Residual	0.00447698	38	0.000117815
Total (Corr.)	0.0395	43	

R-squared = 88.6659 percent
R-Squared (adjusted for d.f.) = 87.1745 percent
Standard Error of Est. = 0.0108543
Mean absolute error = 0.00767586
Durbin-Watson statistic = 2.15687
Lag 1 residual autocorrelation = -0.07971

Residual Analysis

	Estimation	*Validation*
n	44	
MSE	0.000117815	
MAE	0.00767586	
MAPE	1.81203	
ME	-5.69753E-7	
MPE	-0.0571027	

The StatAdvisor
The output shows the results of fitting a piecewise linear regression model to describe the relationship between chlorine and weeks. The equation of the fitted model is

chlorine = 0.575714 - 0.0103571*weeks + 0.00674228*(weeks-12.0915)*(weeks>=12.0915) + 0.00281699*(weeks-24.448)*(weeks>=24.448)

In performing the fit, the estimation process terminated successfully after 7 iterations, at which point the residual sum of squares appeared to approach a minimum.

The output includes:
Included in the output are:

- **Data Summary:** a summary of the input data.

- **Coefficients:** the estimated model coefficients with approximate confidence intervals. Confidence intervals that do not contain 0 indicate that the corresponding model parameter is statistically significant at the stated confidence level.

- **Analysis of Variance:** decomposition of the variability of the dependent variable Y into a model sum of squares and a residual or error sum of squares.

- **Statistics:** summary statistics for the fitted model, including:

  R-squared - represents the percentage of the variability in Y which has been explained by the fitted regression model, ranging from 0% to 100%. For the sample data, the regression has accounted for about 88.7% of the variability amongst the observed *chlorine* concentrations.

  Adjusted R-Squared – the R-squared statistic, adjusted for the number of coefficients in the model. This value is often used to compare models with different numbers of coefficients.

  Standard Error of Est. – the estimated standard deviation of the residuals (the deviations around the model). This value is used to create prediction limits for new observations.

  Mean absolute error – the average absolute value of the residuals.

  Durbin-Watson statistic – a measure of serial correlation in the residuals. If the residuals vary randomly, this value should be close to 2. A small P-value indicates a non-random pattern in the residuals. For data recorded over time, a small P-value could indicate that some trend over time has not been accounted for.

  Lag 1 residual autocorrelation – the estimated correlation between consecutive residuals, on a scale of –1 to 1. Values far from 0 indicate that significant structure remains unaccounted for by the model.

- **Residual Analysis:** if a subset of the rows in the datasheet have been excluded from the analysis using the *Select* field on the data input dialog box, the fitted model is used to make predictions of the Y values for those rows. This table shows statistics on the prediction errors, defined by

$$e_i = y_i - \hat{y}_i \qquad (2)$$

  Included are the mean squared error (MSE), the mean absolute error (MAE), the mean absolute percentage error (MAPE), the mean error (ME), and the mean percentage error (MPE). These validation statistics can be compared to the statistics for the fitted model to determine how well that model predicts observations outside of the data used to fit it.

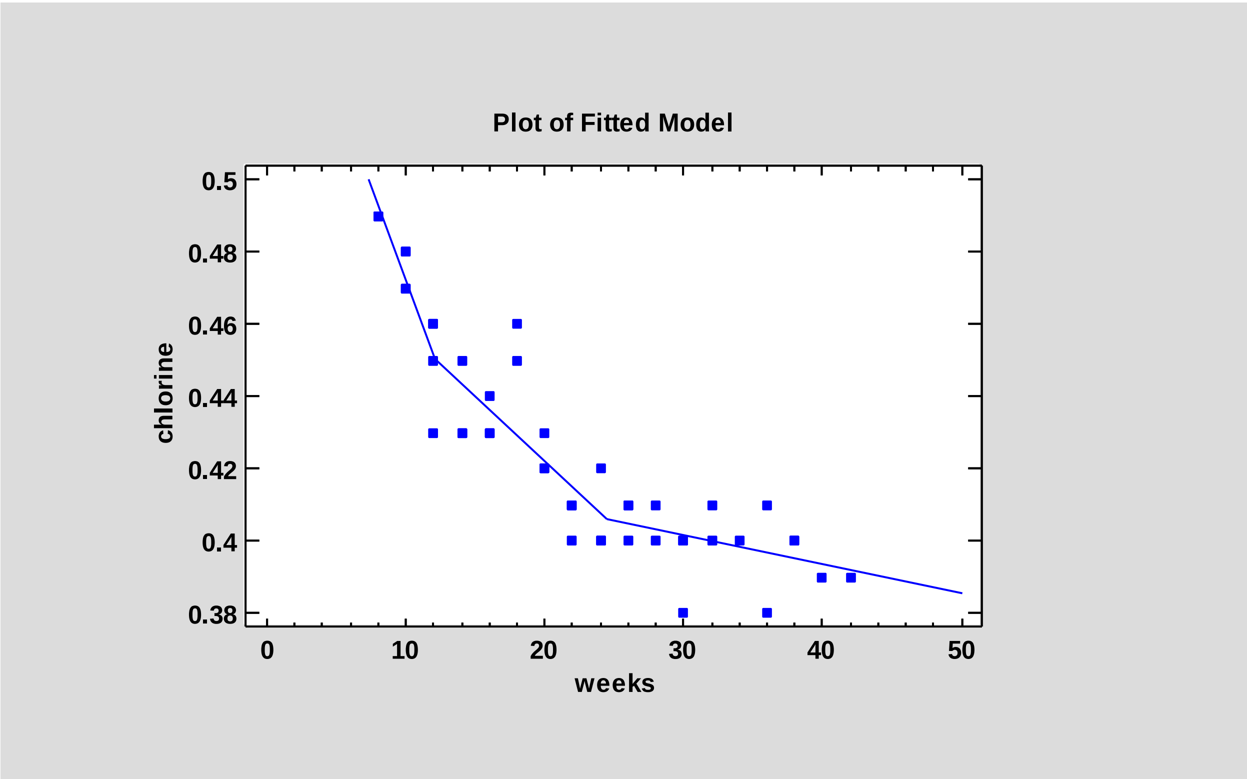
- **The StatAdvisor:** displays the equation of the fitted model. For the sample data, the fitted model is

*chlorine = 0.575714 - 0.0103571*weeks + 0.00674228*(weeks-12.0915)*(weeks>=12.0915) + 0.00281699*(weeks-24.448)*(weeks>=24.448)*

The model changes slope at 12.0915 weeks and 24.448 weeks.

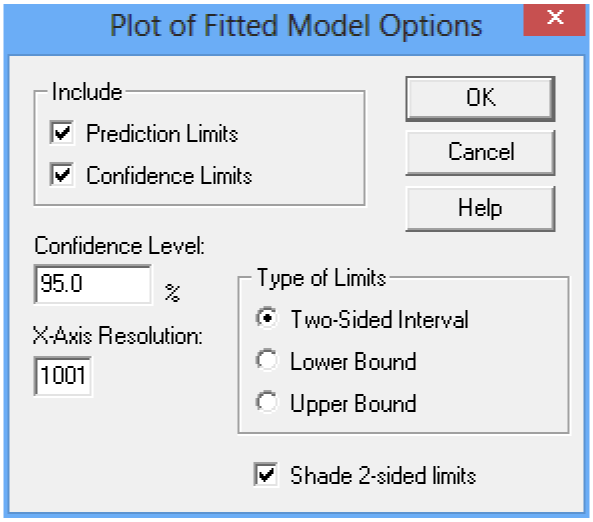
Plot of Fitted Model

This plot shows the fitted piecewise linear model:



To add approximate prediction limits for new observations and confidence limits for the mean response, use *Pane Options*.

Pane Options



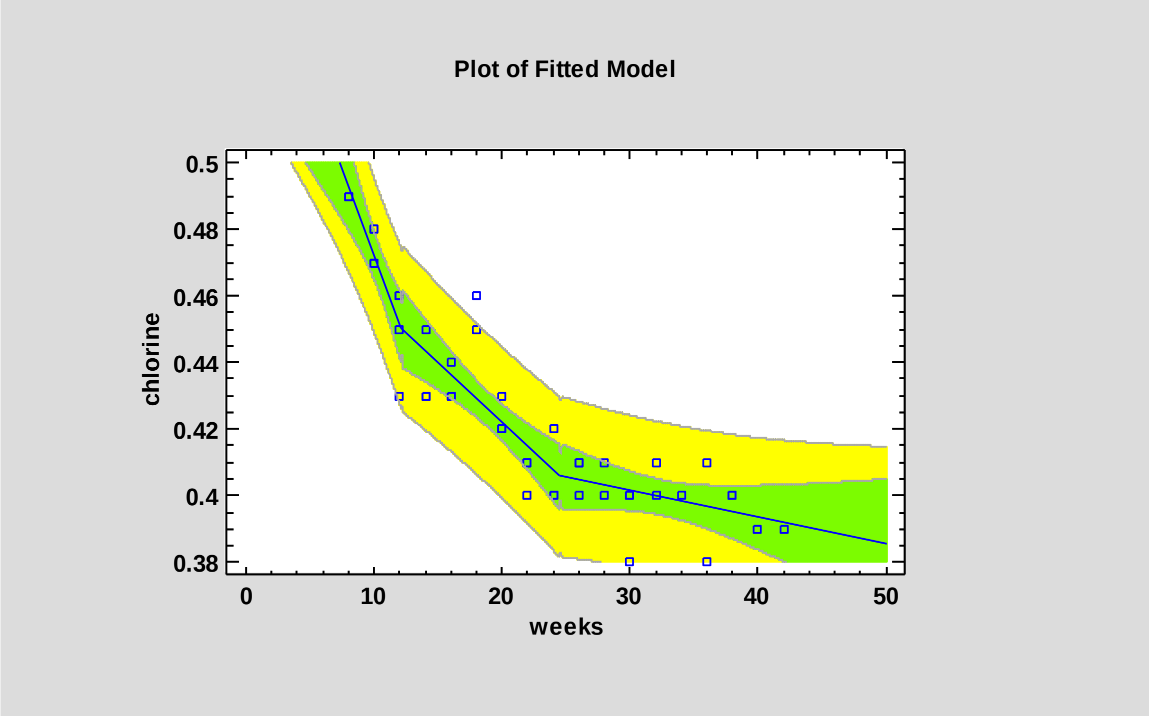
- **Include**: the limits to include on the plot.

- **Confidence Level:** the confidence percentage for the limits.

- **X-Axis Resolution**: the number of values of X at which the model is determined when plotting. Higher resolutions result in smoother plots.

- **Type of Limits**: whether to plot two-sided confidence intervals or one-sided confidence bounds.

The plot below displays both prediction limits (outer bands) and confidence limits for the mean (inner bands). Note that the limits are approximate and obtained from numerically estimated second derivatives at the nonlinear least squares solution.



Lack-of-Fit Test

When more than one observation has been recorded at the same value of X, a lack-of-fit test can be performed to determine whether the fitted model adequately describes the relationship between Y and X. The *Lack-of-Fit* pane displays the following table:

Analysis of Variance with Lack-of-Fit

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Residual	0.00447698	38	0.000117815		
Lack-of-Fit	0.00211032	12	0.00017586	1.93	0.0778
Pure Error	0.00236667	26	0.0000910256		

The lack-of-fit test decomposes the residual sum of squares into 2 components:

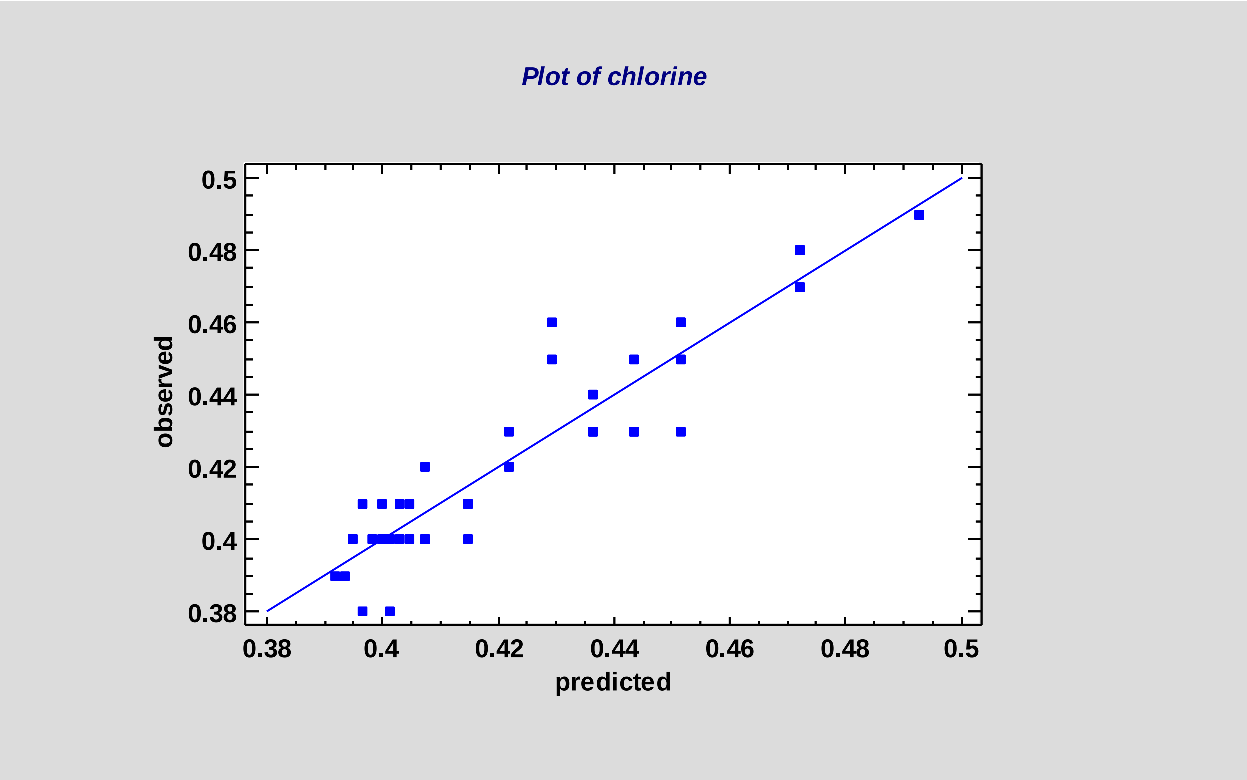
1. *Pure error:* variability of the Y values at the same value of X.
2. *Lack-of-fit:* variability of the average Y values around the fitted model.

Of primary interest is the P-Value for lack-of-fit. A small P-value (below 0.05 if operating at the 5% significance level) indicates that the selected model does *not* adequately describe the observed relationship.

For the example data, the P-value is above 5% and indicates that the selected model may adequately explain the relationship between *chlorine* and *weeks*.

Observed Versus Predicted

The *Observed versus Predicted* plot shows the observed values of the dependent variable on the vertical axis and the predicted values on the horizontal axis.



If the model fits well, the points should be randomly scattered around the diagonal line.

Fitted Values and Residuals

This table shows the predicted value of Y and the residual for each row in the datasheet used to estimate the model:

Fitted Values and Residuals

			Predicted		*Studentized*
Row	*X*	*Y*	*Y*	*Residual*	*Residual*
1	8.0	0.49	0.492857	-0.00285714	-0.324414
2	8.0	0.49	0.492857	-0.00285714	-0.324414
3	10.0	0.48	0.472143	0.00785714	0.761836
4	10.0	0.47	0.472143	-0.00214286	-0.206281
5	10.0	0.48	0.472143	0.00785714	0.761836
6	10.0	0.47	0.472143	-0.00214286	-0.206281
7	12.0	0.46	0.451429	0.00857143	0.888408
8	12.0	0.46	0.451429	0.00857143	0.888408
9	12.0	0.45	0.451429	-0.00142857	-0.146556
10	12.0	0.43	0.451429	-0.0214286	-2.35691
11	14.0	0.45	0.443582	0.00641833	0.64852
12	14.0	0.43	0.443582	-0.0135817	-1.40028
13	14.0	0.43	0.443582	-0.0135817	-1.40028
14	16.0	0.44	0.436352	0.00364806	0.350866
15	16.0	0.43	0.436352	-0.00635194	-0.612998
16	16.0	0.43	0.436352	-0.00635194	-0.612998
17	18.0	0.46	0.429122	0.0308778	3.30165
18	18.0	0.45	0.429122	0.0208778	2.07278
19	20.0	0.42	0.421892	-0.00189248	-0.177816
20	20.0	0.42	0.421892	-0.00189248	-0.177816
21	20.0	0.43	0.421892	0.00810752	0.767486
22	22.0	0.41	0.414663	-0.00466275	-0.448084
23	22.0	0.41	0.414663	-0.00466275	-0.448084
24	22.0	0.4	0.414663	-0.0146628	-1.44433
25	24.0	0.42	0.407433	0.012567	1.28614
26	24.0	0.4	0.407433	-0.00743302	-0.749895
27	24.0	0.4	0.407433	-0.00743302	-0.749895
28	26.0	0.41	0.404575	0.00542465	0.539689
29	26.0	0.4	0.404575	-0.00457535	-0.454678
30	26.0	0.41	0.404575	0.00542465	0.539689
31	28.0	0.41	0.40298	0.00702039	0.679338
32	28.0	0.4	0.40298	-0.00297961	-0.286862
33	30.0	0.4	0.401384	-0.00138386	-0.13072
34	30.0	0.4	0.401384	-0.00138386	-0.13072
35	30.0	0.38	0.401384	-0.0213839	-2.14089
36	32.0	0.41	0.399788	0.0102119	0.969271
37	32.0	0.4	0.399788	0.000211884	0.0198607
38	34.0	0.4	0.398192	0.00180763	0.169957
39	36.0	0.41	0.396597	0.0134034	1.3053
40	36.0	0.38	0.396597	-0.0165966	-1.6365
41	38.0	0.4	0.395001	0.00499912	0.489257
42	38.0	0.4	0.395001	0.00499912	0.489257
43	40.0	0.39	0.393405	-0.00340514	-0.345303
44	42.0	0.39	0.391809	-0.00180939	-0.193271

The Studentized residuals equal the ordinary residuals divided by their approximate standard errors.

Predictions

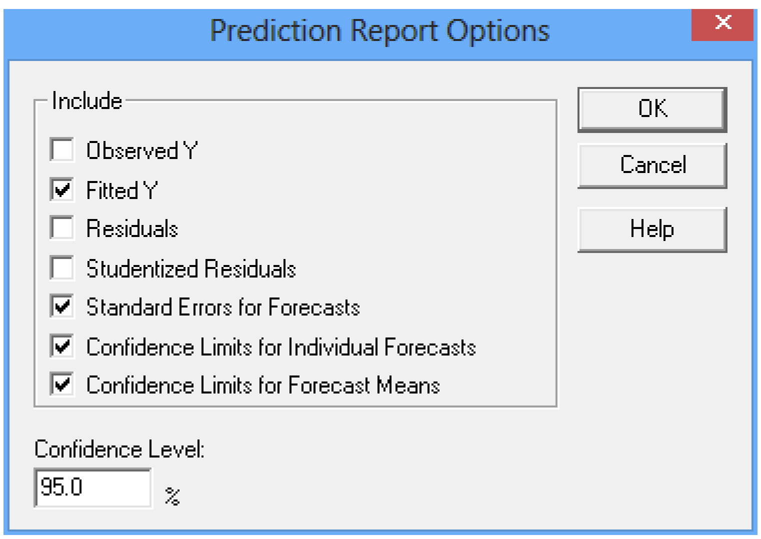
The fitted regression model may be used to predict the outcome of new samples by adding additional rows to the datasheet with specified values of X but missing values (blank cells) for Y.

Predictions for chlorine

		Fitted	*Stnd. Error*	*Lower 95.0% CL*	*Upper 95.0% CL*	*Lower 95.0% CL*	*Upper 95.0% CL*
Row	*weeks*	*Value*	*for Forecast*	*for Forecast*	*for Forecast*	*for Mean*	*for Mean*
45	44.0	0.390214	0.0127284	0.364446	0.415981	0.376755	0.403672
46	46.0	0.388618	0.0132689	0.361756	0.415479	0.373168	0.404068
47	48.0	0.387022	0.0138693	0.358945	0.415099	0.369544	0.4045

The *Fitted Value* shows the predicted mean amount of chlorine for samples aged 44, 46 and 48 weeks. Note that this is an extrapolation of the fitted model and is correct only if the model is valid through 48 weeks.

Pane Options



The table may include:

- *Observed Y* – as entered in the datasheet.
- *Fitted Y* – predicted values from the fitted model.
- *Residuals* – observed minus fitted values.
- *Studentized Residuals* – residuals divided by their approximate standard errors.
- *Standard Errors for Forecasts* – the estimated standard errors for new observations at the indicated values of X.

- *Confidence Limits for Individual Forecasts* – approximate confidence limits for new observations at the indicated values of X.
- *Confidence Limits for Forecast Means* – approximate confidence limits for the mean value of Y at the indicated values of X.

If *Observed Y*, *Residuals* or *Studentized Residuals* is selected, the table will include all rows in the datasheet. Otherwise, only rows with missing values for Y will be included.

Unusual Residuals

Once the model has been fit, it is useful to study the residuals to determine whether any outliers exist that should be removed from the data. The *Unusual Residuals* pane lists all observations that have unusually large residuals.

			Predicted		Studentized
Row	X	Y	Y	Residual	Residual
10	12.0	0.43	0.451429	-0.0214286	-2.36
17	18.0	0.46	0.429122	0.0308778	3.30
18	18.0	0.45	0.429122	0.0208778	2.07
35	30.0	0.38	0.401384	-0.0213839	-2.14

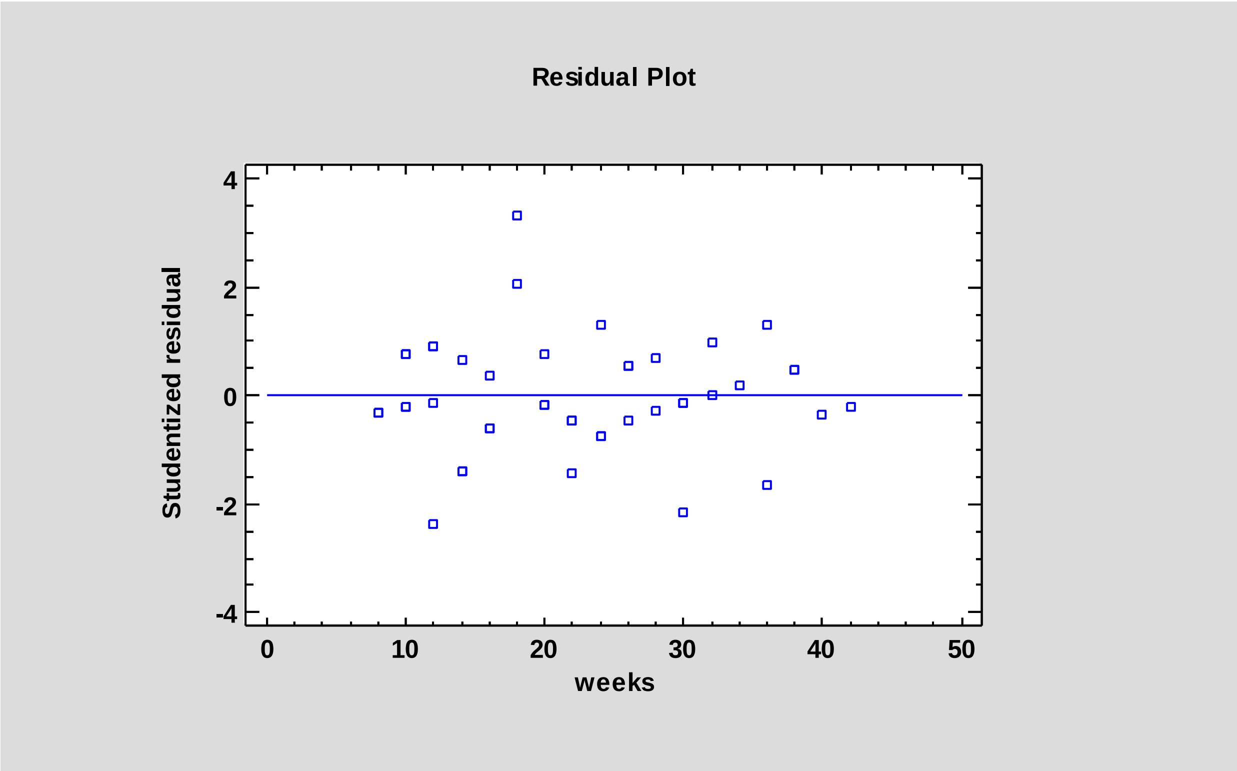
The table displays:

- **Row** – the row number in the datasheet.
- **Observed Value** – the observed value of the dependent variable.
- **Predicted Value** – the value predicted by the fitted model.
- **Residual** – the difference between the observed and predicted values.
- **Studentized Residual** – a standardized residual in which each residual is divided by an estimate of its standard error.

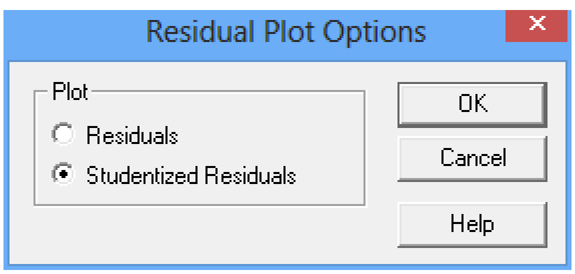
The table includes all rows for which the absolute value of the Studentized residual is greater than 2.0. Absolute values in excess of 3 such as row #17 should be examined carefully to determine whether they correspond to outliers.

Residuals versus X

This graph plots the residuals from the fitted model versus the observed values of X. It is helpful in visualizing whether all of the curvature has been captured by the fitted model. If so, the points should be randomly scattered around the horizontal line.



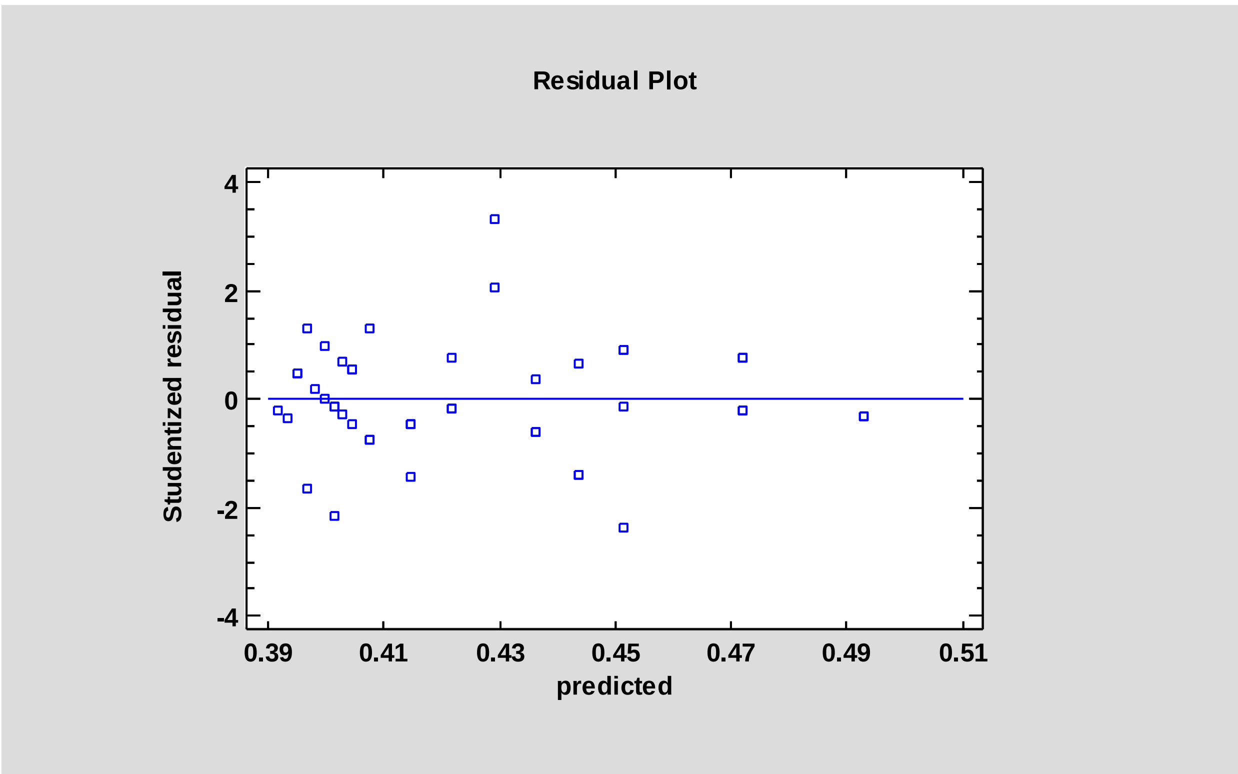
Pane Options



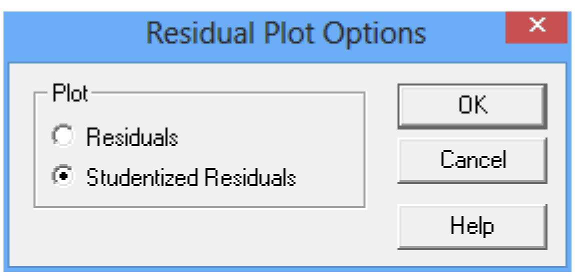
- **Plot:** selects the type of residuals to display.

Residuals versus Predicted

This graph plots the residuals from the fitted model versus the predicted values of Y. It is helpful in detecting any heteroscedasticity in the data. Heteroscedasticity occurs when the variability of the data changes as the mean changes, and might necessitate transforming the data before fitting the regression model or doing a weighted regression. It is usually evidenced by a funnel-shaped pattern in this residual plot.



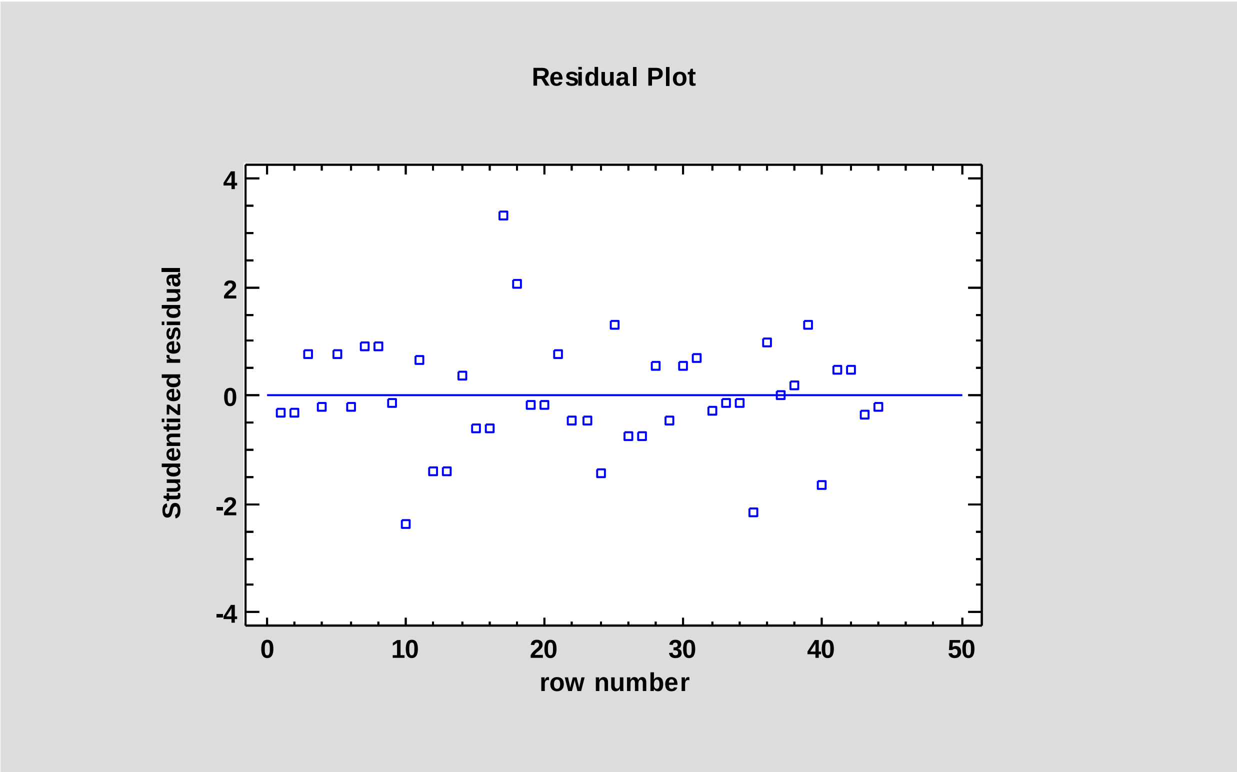
Pane Options



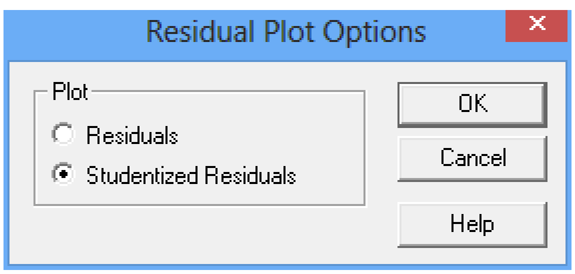
- **Plot:** selects the type of residuals to display.

Residuals versus Row Number

This graph plots the residuals from the fitted model versus the row number in the datasheet. If the data are arranged in chronological order, any pattern in the data might indicate an outside influence.



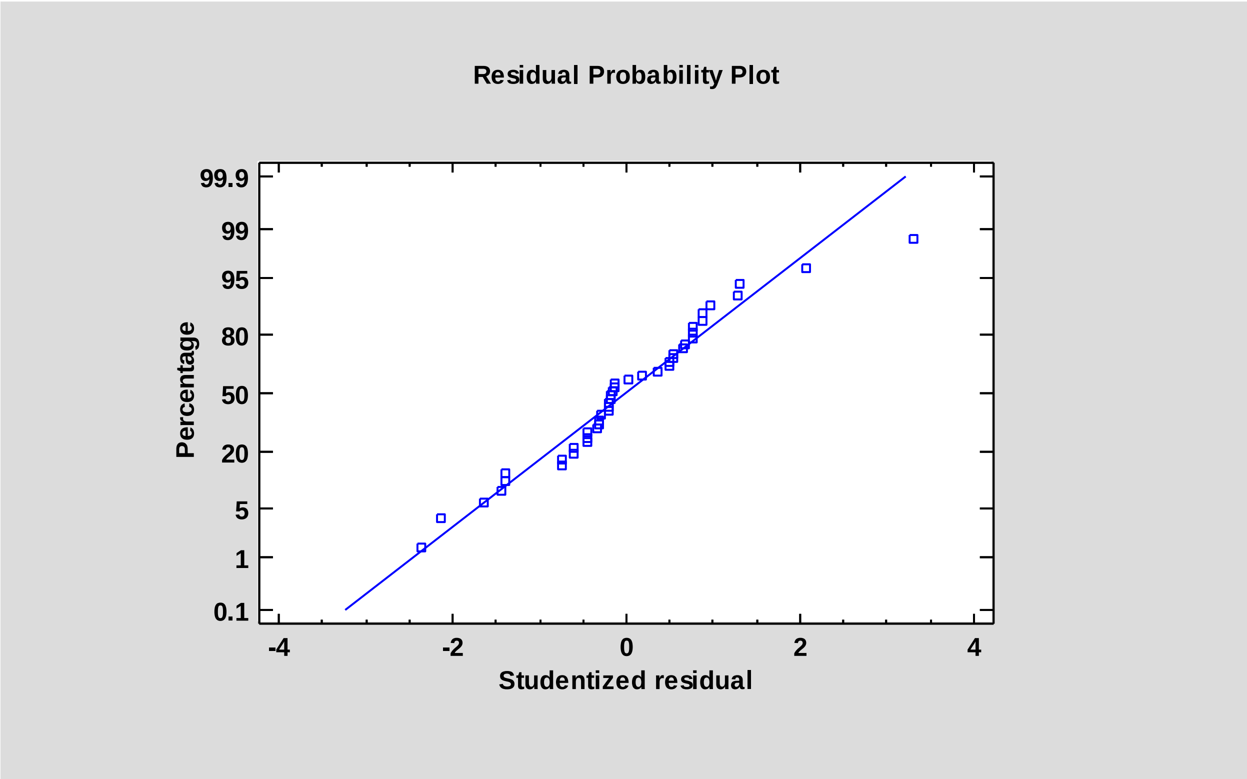
Pane Options



- **Plot:** selects the type of residuals to display.

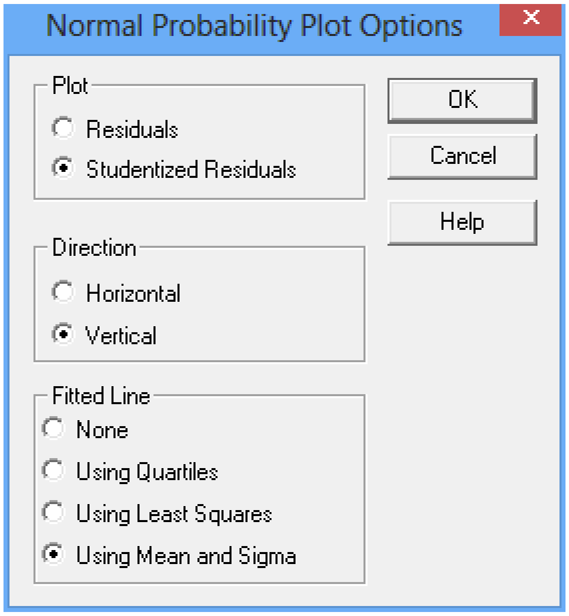
Residual Probability Plot

This plot displays the residuals on a graph scaled to help determine whether the residuals could reasonably have come from a normal distribution. If so, they should fall close to the diagonal line:



For details on how the normal probability plot is constructed, refer to the document titled *One Variable Analysis*.

Pane Options



- **Plot:** the type of residual to plot.

- **Direction**: the orientation of the plot. If vertical, the *Percentage* is displayed on the vertical axis. If *Horizontal*, *Percentage* is displayed on the horizontal axis.

- **Fitted Line**: the method used to fit the reference line to the data. If *Using Quartiles*, the line passes through the median when *Percentage* equals 50 with a slope determined from the interquartile range. If *Using Least Squares*, the line is fit by least squares regression of the normal quantiles on the observed order statistics. If *Using Mean and Sigma*, the line is determined from the mean and standard deviation of the *n* observations. The method based on quartiles puts more weight on the shape of the data near the center and is often able to show deviations from normality in the tails that would not be evident using the other methods.

The *Direction* and *Fitted Line* defaults are determined from the settings on the *EDA* tab of the *Preferences* dialog box on the *Edit* menu.

Influential Points

In fitting a regression model, all observations do not have an equal influence on the parameter estimates in the fitted model. In a simple regression, points located at very low or very high values of X have greater influence than those located nearer to the mean of X. The *Influential Points* pane displays any observations that have high influence on the fitted model:

Influential Points

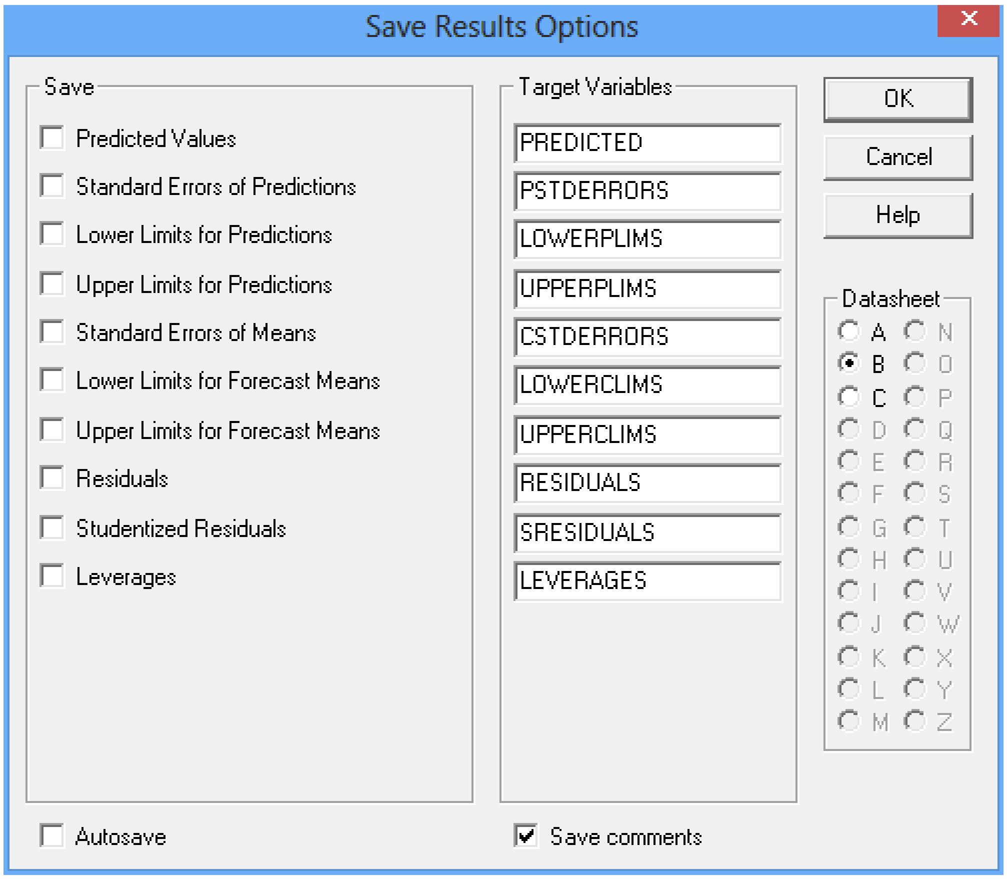
			Predicted	Studentized	
Row	*X*	*Y*	*Y*	*Residual*	*Leverage*

Average leverage of single data point = 0.136364

The above table shows every point with *leverage* equal to 3 or more times that of an average data point, where the leverage of an observation is a measure of its influence on the estimated model coefficients. In general, values with leverage exceeding 5 times that of an average data value should be examined closely, since they have unusually large impact on the fitted model.

In the sample data, there are no high leverage points.

Save Results



The following results may be saved to the datasheet:

1. *Predicted Values* – the fitted values corresponding to each row of the datasheet.
2. *Standard Errors of Predictions* – the estimated standard error for new observations at each value of X.
3. *Lower Limits for Predictions* – lower prediction limits for new observations at each value of X.
4. *Upper Limits for Predictions* – upper prediction limits for new observations at each value of X.
5. *Standard Errors of Means* – the estimated standard error for the mean value of Y at each value of X.
6. *Lower Limits for Forecast Means* – lower confidence limits for the mean value of Y at each value of X.
7. *Upper Limits for Forecast Means* – upper confidence limits for the mean value of Y at each value of X.
8. *Residuals* – the ordinary residuals.
9. *Studentized Residuals* – the standardized residuals.

10. *Leverages* – the leverage associated with each observation.

References

Box, G. E. P., Jenkins, G. M. and Reinsel, G. (1994) Time Series Analysis: Forecasting and Control, 3rd edition. Pearson Education.

Draper, N. R. and Smith, H. (1998). Applied Regression Analysis, 3rd edition. New York: John Wiley and Sons.

