Simple Regression

Summary

The **Simple Regression** procedure is designed to construct a statistical model describing the impact of a single quantitative factor X on a dependent variable Y. Any of 27 linear and nonlinear models may be fit, using either least squares or a resistant estimation procedure. Tests are run to determine the statistical significance of the model. The fitted model may be plotted with confidence limits and/or prediction limits. Residuals may also be plotted and influential observations identified.

Sample StatFolio: *simple reg.sgp*

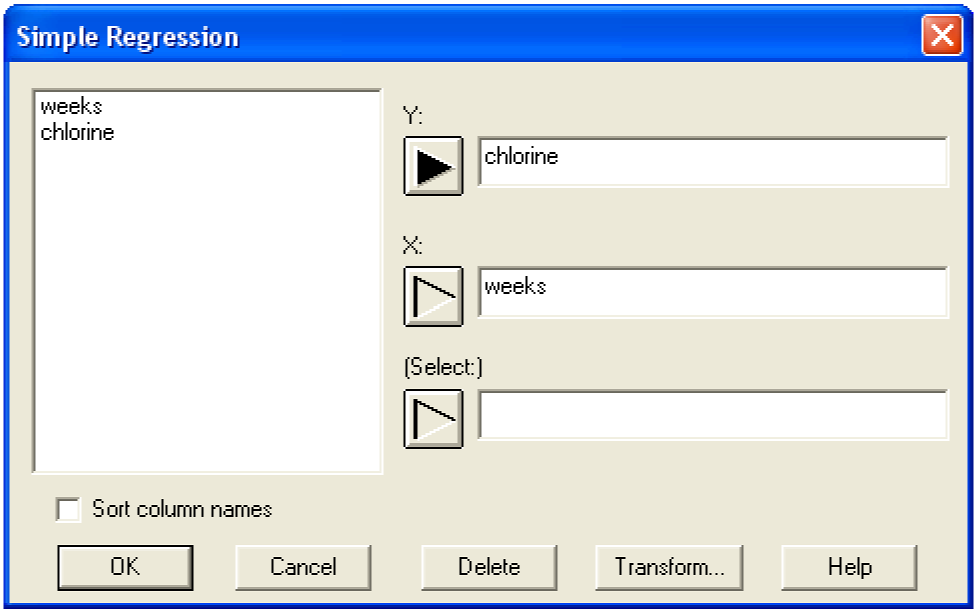
Sample Data:

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), consists 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
14	0.45
14	0.43
14	0.43
…	…

Data Input

The data input dialog box requests the names of the columns containing the dependent variable Y and the independent variable X:



Y: chlorine

X: weeks

- **Y:** numeric column containing the *n* observations for the dependent variable Y.
- **X:** numeric column containing the *n* values for the independent variable X.
- **Select:** subset selection.

Analysis Summary

The *Analysis Summary* shows information about the fitted model.

Simple Regression - chlorine vs. weeks
Dependent variable: chlorine
Independent variable: weeks
Linear model: Y = a + b*X

Coefficients

	Least Squares	*Standard*	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
Intercept	0.48551	0.00589066	82.4204	0.0000
Slope	-0.00271679	0.000243115	-11.1749	0.0000

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	0.0295587	1	0.0295587	124.88	0.0000
Residual	0.00994133	42	0.000236698		
Total (Corr.)	0.0395	43			

Correlation Coefficient = -0.865055
R-squared = 74.8321 percent
R-squared (adjusted for d.f.) = 74.2328 percent
Standard Error of Est. = 0.015385
Mean absolute error = 0.012834
Durbin-Watson statistic = 0.992081 (P=0.0001)
Lag 1 residual autocorrelation = 0.451981

Included in the output are:

- **Variables and model:** identification of the input variables and the model that was fit. By default, a linear model of the form

  $$Y = a + b X \tag{1}$$

  is fit, although a different model may be selected using *Analysis Options*.

- **Coefficients:** the estimated coefficients, standard errors, t-statistics, and P values. The estimates of the model coefficients can be used to write the fitted equation, which in the example is

  $$\textit{chlorine} = 0.48551 - 0.00271679\ \textit{weeks} \tag{2}$$

  The t-statistic tests the null hypothesis that the corresponding model parameter equals 0, versus the alternative hypothesis that it does not equal 0. Small P-Values (less than 0.05 if operating at the 5% significance level) indicate that a model coefficient is significantly different from 0. In the sample data, both the intercept and slope are statistically significant.

- **Analysis of Variance:** decomposition of the variability of the dependent variable Y into a model sums of squares and a residual or error sum of squares. Of particular interest is the F-test and its associated P-value, which tests the statistical significance of the fitted model. A small P-Value (less than 0.05 if operating at the 5% significance level) indicates that a significant relationship of the form specified exists between Y and X. In the sample data, the model is highly significant.

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

  Correlation coefficient - measures the strength of the linear relationship between Y and X on a scale ranging from -1 (perfect negative linear correlation) to +1 (perfect positive linear correlation). In the sample data, the correlation between *chlorine* and *weeks* is relatively strong, with the negative sign indicating that *chlorine* goes down as *weeks* goes up.

  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 75% of the variability in the chlorine measurements. The remaining 25% is attributable to deviations around the line, which may be due to other factors, to measurement error, or to a failure of the linear model to fit the data adequately.

  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.

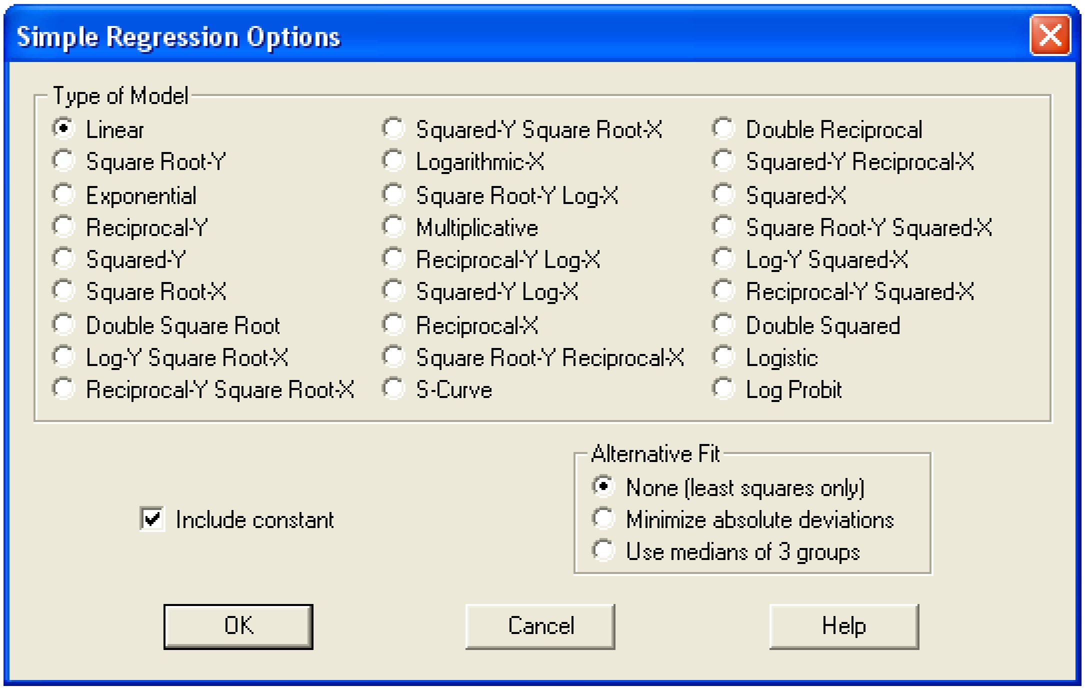
  Standard Error of Est. – the estimated standard deviation of the residuals (the deviations around the model). This value is used to create prediction limits for new observations.

  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. In the current example, a small P-value is indicative of the fact that the linear model has not picked up all of the structure in the data, as will be seen when the residuals are plotted.

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.

Analysis Options



Simple Regression Options

Type of Model

● Linear
○ Square Root-Y
○ Exponential
○ Reciprocal-Y
○ Squared-Y
○ Square Root-X
○ Double Square Root
○ Log-Y Square Root-X
○ Reciprocal-Y Square Root-X
○ Squared-Y Square Root-X
○ Logarithmic-X
○ Square Root-Y Log-X
○ Multiplicative
○ Reciprocal-Y Log-X
○ Squared-Y Log-X
○ Reciprocal-X
○ Square Root-Y Reciprocal-X
○ S-Curve
○ Double Reciprocal
○ Squared-Y Reciprocal-X
○ Squared-X
○ Square Root-Y Squared-X
○ Log-Y Squared-X
○ Reciprocal-Y Squared-X
○ Double Squared
○ Logistic
○ Log Probit

☑ Include constant

Alternative Fit

● None (least squares only)
○ Minimize absolute deviations
○ Use medians of 3 groups

- **Type of Model:** the model to be estimated. All of the models displayed can be linearized by transforming either X, Y, or both. When fitting a nonlinear model, STATGRAPHICS first transforms the data, then fits the model, and then inverts the transformation to display the results.

- **Include constant:** whether to include a constant term in the model. A linear model without a constant term will go through the origin.

- **Alternative Fit**: an alternative estimation procedure. If selected, an additional set of estimates will be added to the output. Two methods of estimation are available, both of which are resistant to outliers:

  Minimize absolute deviations – minimizes the sum of the absolute values of the deviations around the fitted model.

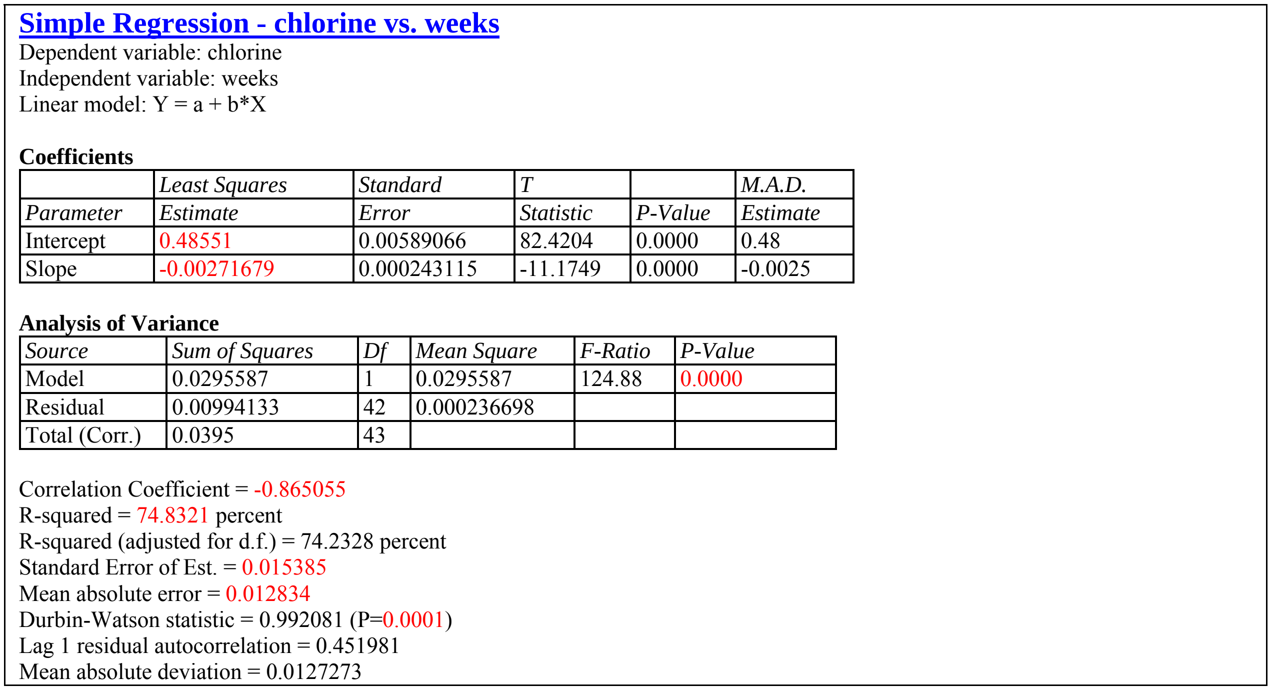
  Use medians of 3 groups – using Tukey's method of fitting a straight line, in which the data are divided into 3 groups according to the value of X, medians computed within each group, and a line determined from the 3 medians.

The available models are shown in the following table:

Model	*Equation*	*Transformation on Y*	*Transformation on X*
Linear	$y = \beta_0 + \beta_1 x$	none	none
Square root-Y	$y = \left(\beta_0 + \beta_1 x\right)^2$	square root	none
Exponential	$y = e^{\left(\beta_0 + \beta_1 x\right)}$	log	none
Reciprocal-Y	$y = \left(\beta_0 + \beta_1 x\right)^{-1}$	reciprocal	none
Squared-Y	$y = \sqrt{\beta_0 + \beta_1 x}$	square	none
Square root-X	$y = \beta_0 + \beta_1 \sqrt{x}$	none	square root
Double square root	$y = \left(\beta_0 + \beta_1 \sqrt{x}\right)^2$	square root	square root
Log-Y square root-X	$y = e^{\left(\beta_0 + \beta_1 \sqrt{x}\right)}$	log	square root
Reciprocal-Y square root-X	$y = \left(\beta_0 + \beta_1 \sqrt{x}\right)^{-1}$	reciprocal	square root
Squared-Y square root-X	$y = \sqrt{\beta_0 + \beta_1 \sqrt{x}}$	square	square root
Logarithmic-X	$y = \beta_0 + \beta_1 \ln(x)$	none	log
Square root-Y log-X	$y = \left(\beta_0 + \beta_1 \ln(x)\right)^2$	square root	log
Multiplicative	$y = \beta_0 x^{\beta_1}$	log	log
Reciprocal-Y log-X	$y = \dfrac{1}{\beta_0 + \beta_1 \ln(x)}$	reciprocal	log
Squared-Y log-X	$y = \sqrt{\beta_0 + \beta_1 \ln(x)}$	square	log
Reciprocal-X	$y = \beta_0 + \beta_1 / x$	none	reciprocal
Square root-Y reciprocal- X	$y = \left(\beta_0 + \beta_1 / x\right)^2$	square root	reciprocal
S-curve	$y = e^{\left(\beta_0 + \beta_1 / x\right)}$	log	reciprocal
Double reciprocal	$y = \left[\beta_0 + \beta / x\right]^{-1}$	reciprocal	reciprocal
Squared-Y reciprocal-X	$y = \sqrt{\beta_0 + \beta_1 / x}$	square	reciprocal
Squared-X	$y = \beta_0 + \beta_1 x^2$	none	square
Square root-Y squared-X	$y = \left(\beta_0 + \beta_1 x^2\right)^2$	square root	square
Log-Y squared-X	$y = e^{\left(\beta_0 + \beta_1 x^2\right)}$	log	square
Reciprocal-Y squared-X	$y = \left(\beta_0 + \beta_1 x^2\right)^{-1}$	reciprocal	square
Double squared	$y = \sqrt{\beta_0 + \beta_1 x^2}$	square	square
Logistic	$y = \dfrac{e^{\left(\beta_0 + \beta_1 x\right)}}{\left[1 + e^{\left(\beta_0 + \beta_1 x\right)}\right]}$	y/(1-y)	none
Log probit	$y = \varphi(\beta_0 + \beta_1 \ln(x))$	$\varphi^{-1}(y)$ (inv. normal)	log

To determine which model to fit to the data, the output in the *Comparison of Alternative Models* pane described below can be helpful, since it fits all of the models and lists them in decreasing order of R-squared.

Example – Resistant Fit
Selecting *Minimum absolute deviations* on the *Analysis Options* dialog box shows an alternative estimate of the line relating chlorine and weeks:



Simple Regression - chlorine vs. weeks
Dependent variable: chlorine
Independent variable: weeks
Linear model: Y = a + b*X

Coefficients

	Least Squares	*Standard*	*T*		*M.A.D.*
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*	*Estimate*
Intercept	0.48551	0.00589066	82.4204	0.0000	0.48
Slope	-0.00271679	0.000243115	-11.1749	0.0000	-0.0025

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	0.0295587	1	0.0295587	124.88	0.0000
Residual	0.00994133	42	0.000236698		
Total (Corr.)	0.0395	43			

Correlation Coefficient = -0.865055
R-squared = 74.8321 percent
R-squared (adjusted for d.f.) = 74.2328 percent
Standard Error of Est. = 0.015385
Mean absolute error = 0.012834
Durbin-Watson statistic = 0.992081 (P=0.0001)
Lag 1 residual autocorrelation = 0.451981
Mean absolute deviation = 0.0127273

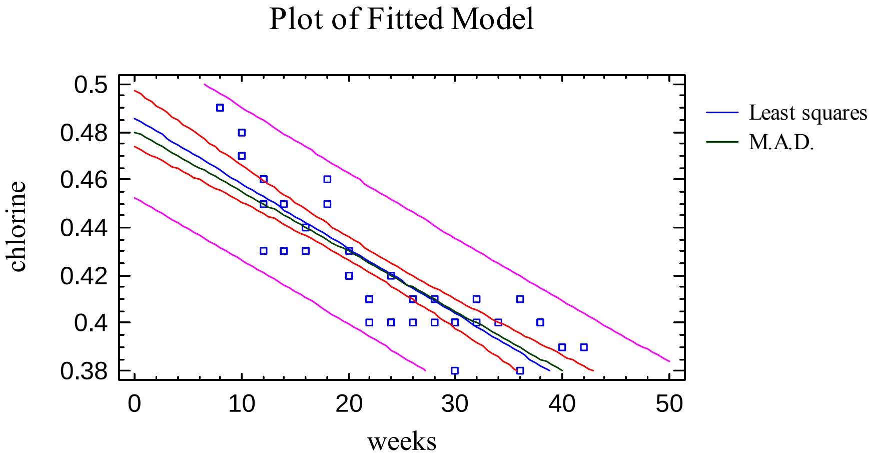
The column labeled *M.A.D. estimate* shows the alternative fit:

$$\textit{chlorine} = 0.48 - 0.0025\ \textit{weeks} \qquad (3)$$

The difference between the 2 fitted models is relatively minor.

Plot of Fitted Model

This pane shows the fitted model or models, together with confidence limits and prediction limits if desired.



The plot includes:

- The line of best fit or **prediction equation**:

$$\hat{y} = \hat{a} + \hat{b}x \tag{4}$$

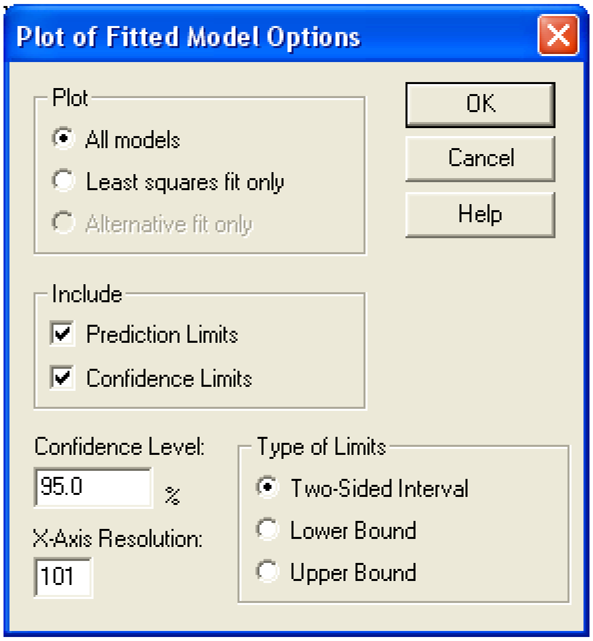
  This is the equation that would be used to predict values of the dependent variable Y given values of the independent variable X. Note that it does a relatively good job of picking up much of the negative correlation between *chlorine* and *weeks*.

- **Confidence intervals** for the mean response at X. These are the inner bounds in the above plot and describe how well the location of the line has been estimated given the available data sample. As the size of the sample *n* increases, these bounds will become tighter. You should also note that the width of the bounds varies as a function of X, with the line estimated most precisely near the average value $\bar{x}$.

- **Prediction limits** for new observations. These are the outer bounds in the above plot and describe how precisely one could predict where a single new observation would lie. Regardless of the size of the sample, new observations will vary around the true line with a standard deviation equal to σ.

The inclusion of confidence limits and prediction limits and their default confidence level is determined by settings on the *ANOVA/Regression* tab of the *Preferences* dialog box, accessible from the *Edit* menu.

Pane Options



Plot of Fitted Model Options

Plot
● All models
○ Least squares fit only
○ Alternative fit only

Include
☑ Prediction Limits
☑ Confidence Limits

Confidence Level: 95.0 %

X-Axis Resolution: 101

Type of Limits
● Two-Sided Interval
○ Lower Bound
○ Upper Bound

- **Plot**: the model or models to plot.

- **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 line 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.

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 selected 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*
Model	0.0295587	1	0.0295587	124.88	0.0000
Residual	0.00994133	42	0.000236698		
Lack-of-Fit	0.00757467	16	0.000473417	5.20	0.0001
Pure Error	0.00236667	26	0.0000910256		
Total (Corr.)	0.0395	43			

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

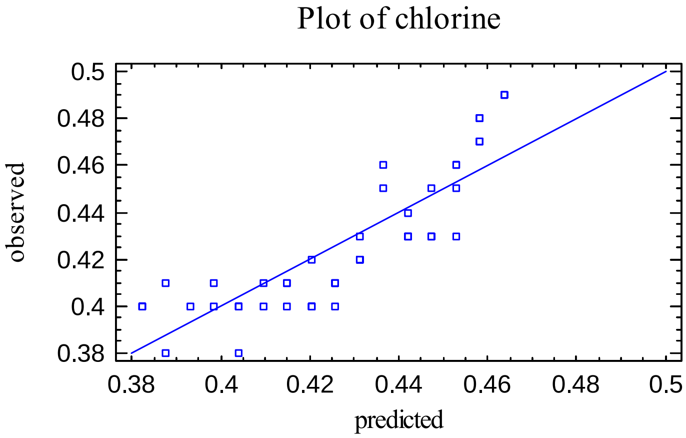
1. *Pure error:* variability of the Y values at the same value of X.
2. *Lack-of-fit:* variability of the average Y values around the fitted model.

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 very small P-value indicates that the linear model does not adequately explain the relationship between *chlorine* and *weeks*.

Observed versus Predicted

The *Observed versus Predicted* plot shows the observed values of Y on the vertical axis and the predicted values $\hat{Y}$ on the horizontal axis.



If the model fits well, the points should be randomly scattered around the diagonal line. It is sometimes possible to see curvature in this plot, which would indicate the need for a curvilinear model rather than a linear model. Any change in variability from low values of X to high values of X might also indicate the need to transform the dependent variable before fitting a model to the data. In the above plot, the variability appears to be fairly constant. However, some evidence of curvature is present.

Residual Plots

As with all statistical models, it is good practice to examine the residuals. In a regression, the residuals are defined by

$$e_i = y_i - \hat{y}_i \tag{5}$$

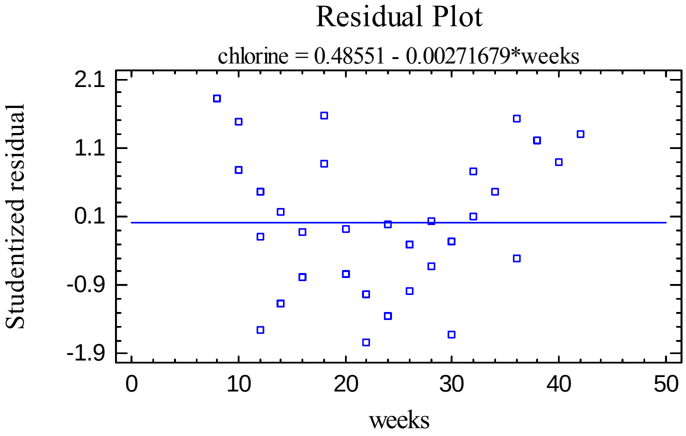
i.e., the residuals are the differences between the observed data values and the fitted model.

The *Simple Regression* procedure creates 3 residual plots:

1. versus X.
2. versus predicted value $\hat{Y}$.
3. versus row number.

Residuals versus X

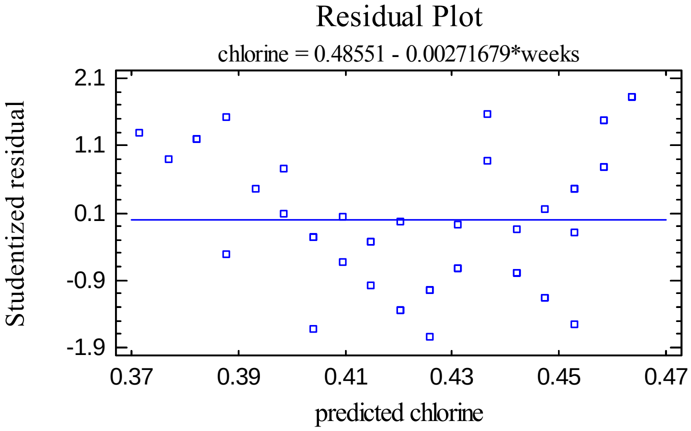
This plot is helpful in visualizing any need for a curvilinear model.



Note that between 20 and 30 weeks, all of the residuals lie below 0 (shown by the horizontal line). Within this range, the straight line over-predicts the amount of available chlorine. It also tends to under-predict the amount beyond 30 weeks.

Residuals versus Predicted

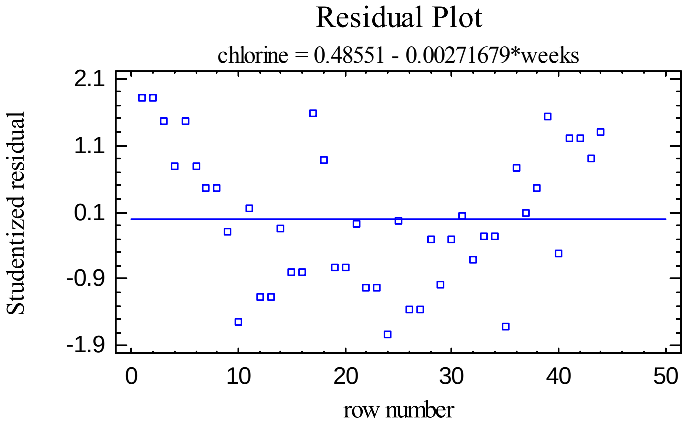
This plot is helpful in detecting any heteroscedasticity in the data.



Heteroscedasticity occurs when the variability of the data changes as the mean changes, and might necessitate transforming the data before fitting the regression model. It is usually evidenced by a funnel-shaped pattern in the residual plot.

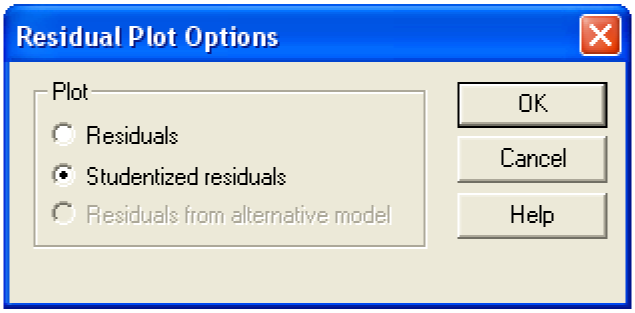
Residuals versus Observation

This plot shows the residuals versus row number in the datasheet:



If the data are arranged in chronological order, any pattern in the data might indicate an outside influence. In the above plot, curvature can be seen since the example data file is sorted according to the values of X.

Pane Options


Residual Plot Options

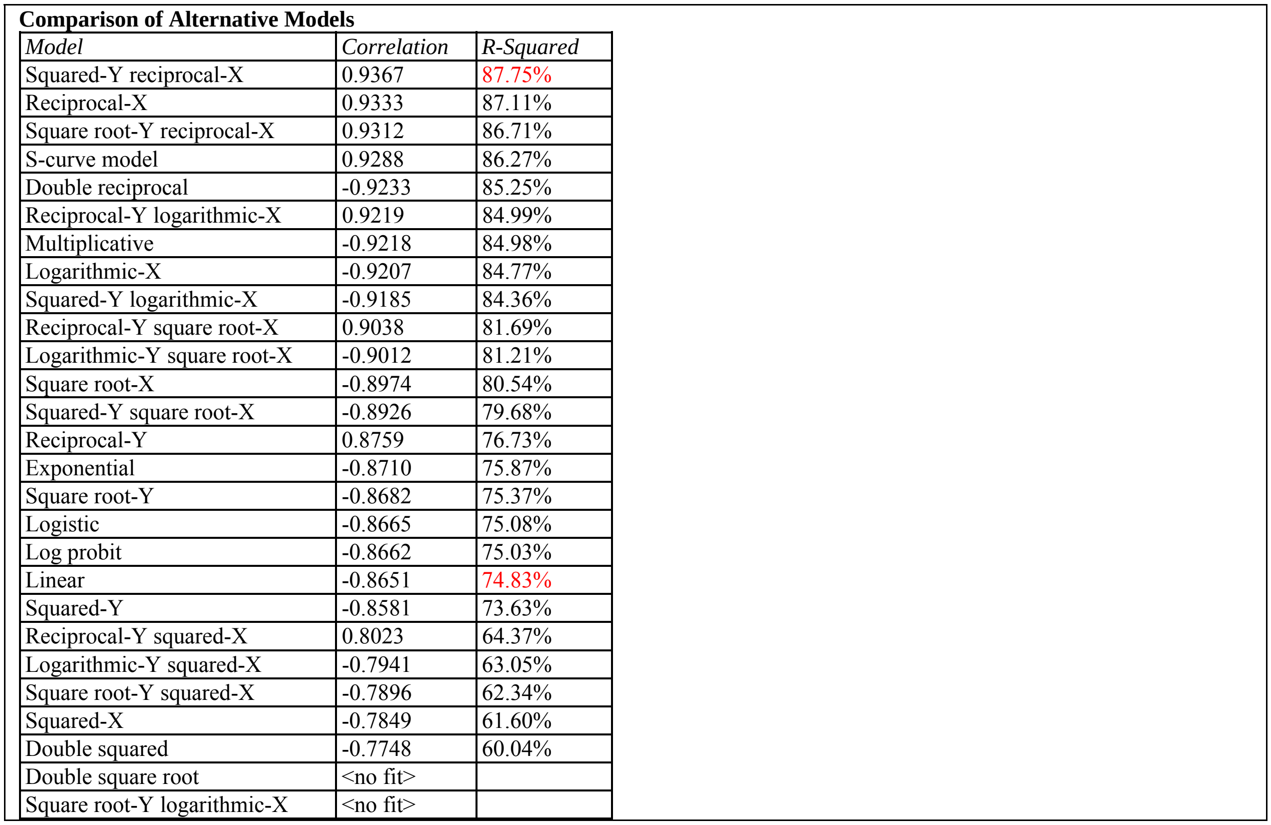
Plot
○ Residuals
● Studentized residuals
○ Residuals from alternative model

The following residuals may be plotted on each residual plot:

1. *Residuals* – the residuals from the least squares fit.
2. *Studentized residuals* – the difference between the observed values y_i and the predicted values $\hat{y}_i$ when the model is fit using all observations except the *i-th*, divided by the estimated standard error. These residuals are sometimes called *externally deleted residuals*, since they measure how far each value is from the fitted model when that model is fit using all of the data except the point being considered. This is important, since a large outlier might otherwise affect the model so much that it would not appear to be unusually far away from the line.
3. *Residuals from alternative model* – the residuals from the model when estimated using the selected resistant method.

Comparison of Alternative Models

The *Comparison of Alternative Models* pane shows the R-squared values obtained when fitting each of the 27 available models:



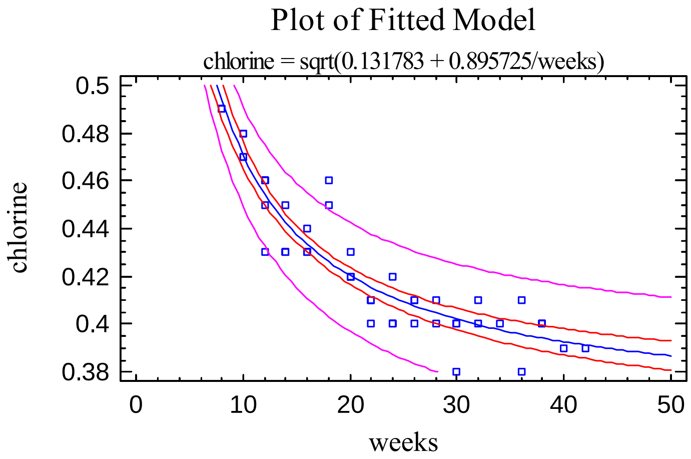
Comparison of Alternative Models

Model	*Correlation*	*R-Squared*
Squared-Y reciprocal-X	0.9367	87.75%
Reciprocal-X	0.9333	87.11%
Square root-Y reciprocal-X	0.9312	86.71%
S-curve model	0.9288	86.27%
Double reciprocal	-0.9233	85.25%
Reciprocal-Y logarithmic-X	0.9219	84.99%
Multiplicative	-0.9218	84.98%
Logarithmic-X	-0.9207	84.77%
Squared-Y logarithmic-X	-0.9185	84.36%
Reciprocal-Y square root-X	0.9038	81.69%
Logarithmic-Y square root-X	-0.9012	81.21%
Square root-X	-0.8974	80.54%
Squared-Y square root-X	-0.8926	79.68%
Reciprocal-Y	0.8759	76.73%
Exponential	-0.8710	75.87%
Square root-Y	-0.8682	75.37%
Logistic	-0.8665	75.08%
Log probit	-0.8662	75.03%
Linear	-0.8651	74.83%
Squared-Y	-0.8581	73.63%
Reciprocal-Y squared-X	0.8023	64.37%
Logarithmic-Y squared-X	-0.7941	63.05%
Square root-Y squared-X	-0.7896	62.34%
Squared-X	-0.7849	61.60%
Double squared	-0.7748	60.04%
Double square root	<no fit>	
Square root-Y logarithmic-X	<no fit>	

The models are listed in decreasing order of R-squared. When selecting an alternative model, consideration should be given to those models near the top of the list. However, since the R-Squared statistics are calculated after transforming X and/or Y, the model with the highest R-squared may not be the best. You should always plot the fitted model to see whether it does a good job for your data.

Example: Fitting a Nonlinear Model

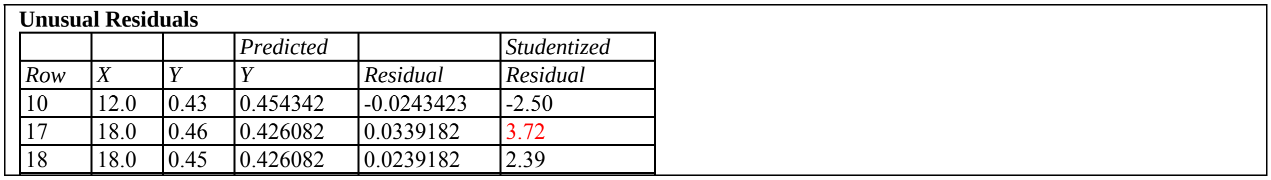
Since the *Squared-Y Reciprocal-X* model has the highest R-squared value, it is a reasonable candidate for the sample data. Selecting that model using *Analysis Options* shows the following result:



Visually, it appears to capture well the observed curvature in the data. Several of the other models give very similar results.

Unusual Residuals

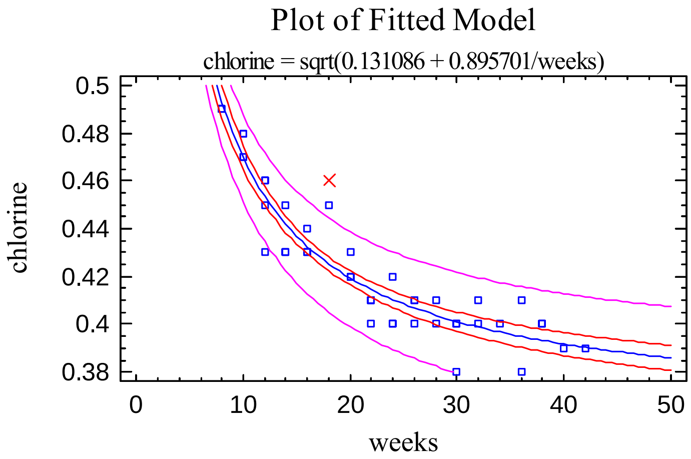
Once the model has been fit, it is useful to study the residuals to determine whether any outliers exist that should be removed from the data. The *Unusual Residuals* pane lists all observations that have Studentized residuals of 2.0 or greater in absolute value.


Unusual Residuals

			Predicted		*Studentized*
Row	*X*	*Y*	*Y*	*Residual*	*Residual*
10	12.0	0.43	0.454342	-0.0243423	-2.50
17	18.0	0.46	0.426082	0.0339182	3.72
18	18.0	0.45	0.426082	0.0239182	2.39

Studentized residuals greater than 3 in absolute value correspond to points more than 3 standard deviations from the fitted model, which is an extremely rare event for a normal distribution. In the sample data, row #17 is almost 4 standard deviations out.

Points can be removed from the fit while examining the *Plot of the Fitted Model* by clicking on a point and then pressing the *Exclude/Include* button on the analysis toolbar:



Excluded points are marked with an X. For the sample data, removing row #17 has little effect on the fitted model.

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*
1	8.0	0.49	0.493709	-0.42	0.170244
2	8.0	0.49	0.493709	-0.42	0.170244

Average leverage of single data point = 0.0454545

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, the two values at X = 8 have a moderately large influence on the fitted model, since these values correspond to the minimum value of X. Compared to the average leverage $\bar{h}$ = 0.045, these points have close to 4 times the influence of an average point. Ideally, one would prefer a data set in which all values had approximately the same leverage, since no point would then have excessive impact on the fitted model. In many cases, this cannot be achieved, but the high leverage points should at least be checked to insure their validity.

Forecasts

The *Forecasts* pane creates predictions using the fitted least squares model.

Predicted Values

		95.00%		*95.00%*	
	Predicted	*Prediction*	*Limits*	*Confidence*	*Limits*
X	*Y*	*Lower*	*Upper*	*Lower*	*Upper*
10.0	0.470485	0.449151	0.490892	0.464671	0.476227
15.0	0.437605	0.41521	0.458909	0.434084	0.441099
20.0	0.420202	0.396859	0.442314	0.416737	0.423638
25.0	0.409405	0.385331	0.432139	0.405391	0.41338
30.0	0.402046	0.377409	0.425258	0.397462	0.406577
35.0	0.396706	0.371626	0.420291	0.391636	0.401711
40.0	0.392653	0.367218	0.416538	0.387182	0.398048

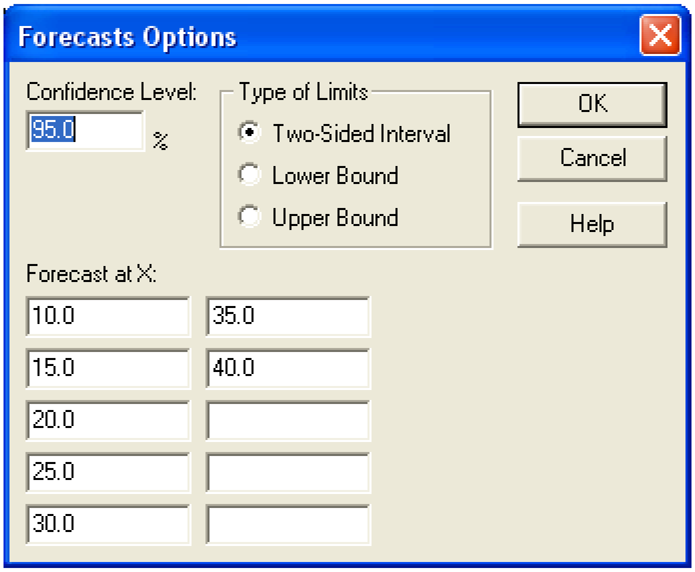
Included in the table are:

- **X** - the value of the independent variable at which the prediction is to be made.
- **Predicted Y** - the predicted value of the dependent variable using the fitted model.
- **Prediction limits** - prediction limits for new observations at the selected level of confidence (corresponds to the outer bounds on the plot of the fitted model).

- **Confidence limits** - confidence limits for the mean value of Y at the selected level of confidence (corresponds to the inner bounds on the plot of the fitted model).

For example, at X = 30 weeks, the best prediction of the mean amount of available chlorine is 0.402, although it could easily be anywhere between 0.397 and 0.407. In addition, one could predict with 95% confidence that any sample produced 30 weeks after production would fall between 0.377 and 0.425. Obviously, the mean can be estimated much more closely that the observed value of any single random sample.

Pane Options



Forecasts Options

Confidence Level: 95.0 %

Type of Limits

● Two-Sided Interval

○ Lower Bound

○ Upper Bound

Forecast at X:

10.0 | 35.0

15.0 | 40.0

20.0

25.0

30.0

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

- **Type of Limits:** whether to display two-sided limits or one-sided bounds.

- **Forecast at X**: up to 10 values of X at which to make predictions.

Save Results

The following results may be saved to the datasheet:

1. *Predicted Values* – the predicted value of Y corresponding to each of the *n* observations.
2. *Lower Limits for Predictions* – the lower prediction limits for each predicted value.
3. *Upper Limits for Predictions* – the upper prediction limits for each predicted value.
4. *Lower Limits for Forecast Means* – the lower confidence limits for the mean value of Y at each of the *n* values of X.
5. *Upper Limits for Forecast Means*– the upper confidence limits for the mean value of Y at each of the *n* values of X.
6. *Residuals* – the *n* residuals.
7. *Studentized Residuals* – the *n* Studentized residuals.
8. *Leverages* – the leverage values corresponding to the *n* values of X.

9. *Predictions from resistant model* – the predicted values of Y made with the model estimated using the alternative fit resistant method.
10. *Residuals from resistant model* – the residuals calculated from the model estimated using the alternative fit resistant method.
11. *Model statistics* – summary statistics for the regression model.
12. *Statistics labels* – identifiers for each of the model statistics.

Note: If limits are saved, they will correspond to the settings on the *Forecasts* pane. If two-sided limits are displayed in the Forecasts table, then the saved limits will also be two-sided. If one-sided bounds are displayed in the table, then the saved limits will also be one-sided.

Calculations

Least Squares Estimates

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \tag{6}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{7}$$

where

$$S_{xx} = \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2 \tag{8}$$

$$S_{xy} = \sum_{i=1}^{n} \left(x_i - \bar{x} \right)\left(y_i - \bar{y} \right) \tag{9}$$

ANOVA Table

Model sum of squares: $SSR = \hat{\beta}_1^2 S_{XX}$ (10)

Error sum of squares: $SSE = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$ (11)

Mean squares error: $MSE = \dfrac{SSE}{n-2}$ (12)

F-ratio: $F = \dfrac{SSR}{MSE}$ (13)

Lack-of-fit: $SSLOF = \sum_{j=1}^{c} \sum_{i=1}^{n_j} \left(\bar{y}_j - \hat{y}_{ij} \right)^2$ (14)

Pure Error: $SSPE = \sum_{j=1}^{c} \sum_{i=1}^{n_j} \left(y_{ij} - \bar{y}_j \right)^2$ (15)

F-Ratio for lack-of-fit: $F = \dfrac{SSLOF/(c-2)}{SSPE/(n-c)}$ (16)

where c = number of unique values of X.

Standard Errors

$$s(\hat{\beta}_0) = \sqrt{MSE\left[\frac{1}{n} + \frac{\overline{X}^2}{S_{XX}}\right]} \tag{17}$$

$$s(\hat{\beta}_1) = \sqrt{\frac{MSE}{S_{XX}}} \tag{18}$$

Correlation Coefficient

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{19}$$

R-Squared

$$R^2 = \frac{SSR}{SSR + SSE} \tag{20}$$

Adjusted R-Squared

$$R_{adj}^2 = 100\left[1 - \left(\frac{n-1}{n-2}\right)\frac{SSE}{SSR + SSE}\right]\% \tag{21}$$

Standard Error of Est.

$$\hat{\sigma} = \sqrt{MSE} \tag{22}$$

Predictions

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{23}$$

Confidence limits: $\hat{y} \pm t_{\alpha/2,n-2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}$ (24)

Prediction limits: $\hat{y} \pm t_{\alpha/2,n-2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}$ (25)

Leverage

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \tag{26}$$

Durbin-Watson Statistic

$$D = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \tag{27}$$

If $n > 500$, then

$$D^* = \frac{|D - 2|}{\sqrt{4/n}} \tag{28}$$

is compared to a standard normal distribution. For 100 < n ≤ 500, *D/4* is compared to a beta distribution with parameters

$$\alpha = \beta = \frac{n-1}{2} \tag{29}$$

For smaller sample sizes, *D/4* is compared to a beta distribution with parameters which are based on the trace of certain matrices related to the *X* matrix, as described by Durbin and Watson (1951) in section 4 of their classic paper.

Lag 1 Residual Autocorrelation

$$r_1 = \frac{\sum_{i=2}^{n} e_i e_{i-1}}{\sum_{i=1}^{n} e_i^2} \tag{30}$$

Modifications if No Constant

The following formulas are modified as shown if no constant is included in the model:

$$\hat{\beta}_0 = 0 \tag{7A}$$

$$S_{xx} = \sum_{i=1}^{n} x_i^{\ 2} \tag{8A}$$

$$S_{xy} = \sum_{i=1}^{n} x_i y_i \tag{9A}$$

$$MSE = \frac{SSE}{n-1} \tag{12A}$$

$$F = \frac{SSLOF/(c-1)}{SSPE/(n-c)} \tag{16A}$$

$$R^2_{adj} = 100\left[1 - \left(\frac{n-1}{n-1}\right)\frac{SSE}{SSR+SSE}\right]\% \tag{21A}$$

$$\hat{y} \pm t_{\alpha/2,n-1}\hat{\sigma}\sqrt{\frac{x^2}{S_{xx}}} \tag{24A}$$

$$\hat{y} \pm t_{\alpha/2,n-1}\hat{\sigma}\sqrt{1+\frac{x^2}{S_{xx}}} \tag{25A}$$

$$h_i = \frac{{x_i}^2}{S_{xx}} \tag{26A}$$