Orthogonal Regression

Summary

The **Orthogonal Regression** procedure is designed to construct a statistical model describing the impact of a single quantitative factor X on a dependent variable Y, when both X and Y are observed with error. Any of 27 linear and nonlinear models may be fit. 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 unusual observations identified.

Sample StatFolio: *orthogonalreg.sgp*

Sample Data:

The file *hens.sgd* contains measurements of the number of hen pheasants in Iowa at 2 different times during the years 1962-1976. The data were obtained by the Iowa Conservation Committee and are reported in the classic text on Measurement Error Models by Wayne Fuller (1987).

Year	*August hens*	*Spring hens*
1962	7.3	8.2
1963	10	11.8
1964	10.1	11
1965	7.4	7.4
1966	8.7	10.2
1967	8.7	10.4
1968	8.1	10.9
1969	6.9	7.5
1970	9.2	9.6
1971	9.3	12
1972	9.7	11.9
1973	10.8	11.9
1974	9.8	12.3
1975	6	6.6
1976	8	9

The August count will be treated as the dependent or Y variable. The spring count will be treated as the predictor or X variable. Since the indices are based on counts by trained observers traveling a specified set of routes, both X and Y are subject to error.

Statistical Model

The model assumed by this procedure is the classic error-in-variables model. Let y_i and x_i be the true values of the dependent and independent variables, respectively. Let the observed value of the dependent variable be

$$Y_i = y_i + e_i \tag{1}$$

where e_i is assumed to be a random error generated from a normal distribution with mean 0 and variance σ_e^2. Let the observed value of the independent variable be

$$X_i = x_i + u_i \tag{2}$$

where u_i is assumed to be a random error generated from a normal distribution with mean 0 and variance σ_u^2. Assume also that e_i, the errors in Y, are independent of u_i, the errors in X. Let the relationship between the true values be

$$y_i = \beta_0 + \beta_1 x_i \tag{3}$$

We wish to obtain estimates of β_0 and β_1 as well as predicted values for y given X.

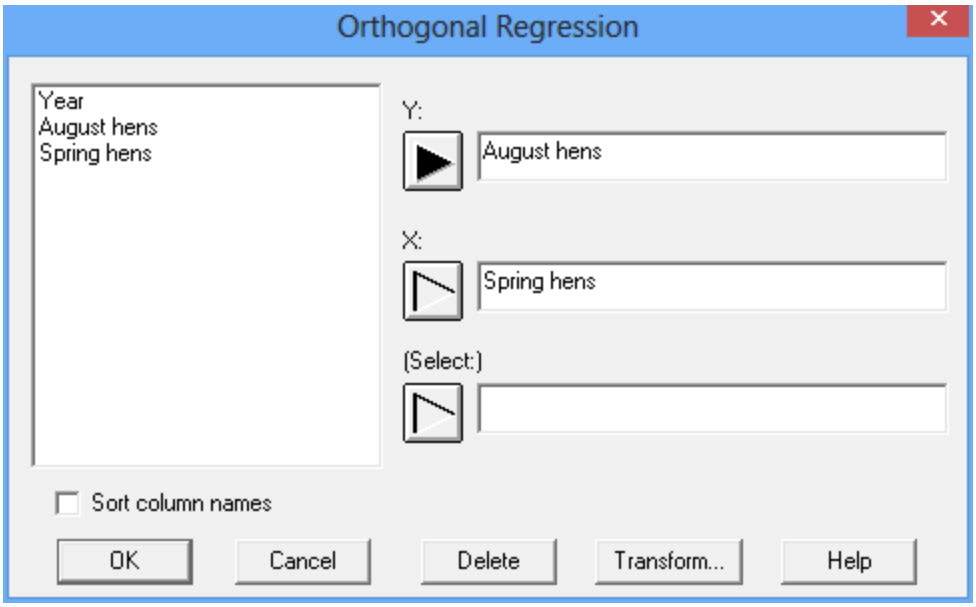
For convenience, we will represent the ratio of the variances between the errors on X and Y by

$$\partial = \frac{\sigma_e^2}{\sigma_u^2} \tag{4}$$

Note: The procedure also fits transformable nonlinear models in which linearity applies after either or both of the variables are transformed. For such models, we assume that equations (1) through (4) hold for the transformed values of Y and X.

Data Input

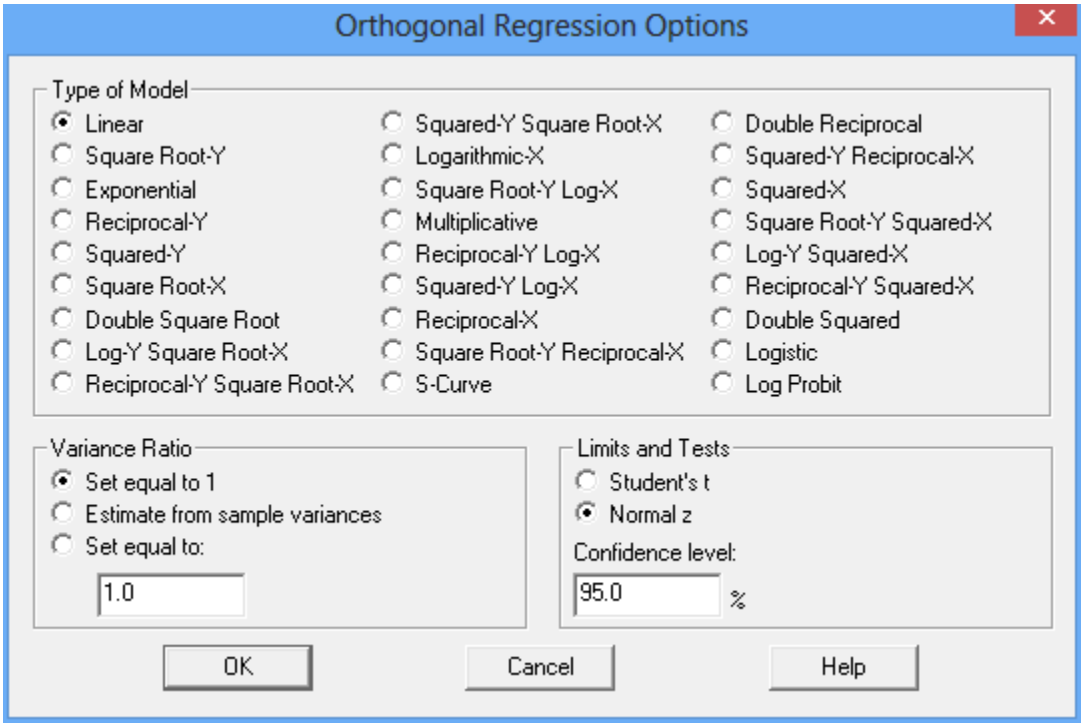
The data input dialog box requests the names of the columns containing the dependent variable Y and the independent variable X:



- **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 Options

The *Analysis Options* dialog box specifies the type of model to be fit, the method for determining the error variance ratio, and the type of statistical tests to be performed:



- **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.

- **Variance Ratio:** the method for setting the value of δ. The choices are to assume equal variances, estimate the ratio from the sample variances, and set the ratio equal to the specified value. Note: do not select *Estimate from sample variances* unless the data are a random sample from a bivariate normal distribution.

- **Limits and tests**: whether to approximate the distribution of the estimated parameters using Student's t distribution or using a normal distribution. Also, the confidence level for the parameter interval estimates may be specified.

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	$y = e^{(\beta_0 + \beta_1 x)}$	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	$y = e^{(\beta_0+\beta_1\sqrt{x})}$	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	$y = e^{(\beta_0+\beta_1/x)}$	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	$y = e^{(\beta_0+\beta_1 x^2)}$	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	$y = \dfrac{e^{(\beta_0+\beta_1 x)}}{\left[1+e^{(\beta_0+\beta_1 x)}\right]}$	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 the correlation coefficient.

Analysis Summary

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

Orthogonal Regression - August hens
Dependent variable: August hens
Independent variable: Spring hens
Linear model: Y = a + b*X
Number of observations: 15
Assumed ratio of error variances: 1.0

Coefficients

	Orthogonal	*Standard*	*Z*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
Intercept	1.71535	0.899585	1.90682	0.0565
Slope	0.691903	0.0882247	7.84251	0.0000

Correlation Coefficient = 0.908693

Estimated error variances

	Variance	*Sigma*
Y	0.222399	0.471592
X	0.222399	0.471592
Residual	0.328868	0.57347

95.0% Confidence Intervals

	Lower limit	*Upper limit*
Intercept	-0.0478083	3.47851
Slope	0.518985	0.864821

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

$$Y = \beta_o + \beta_1 X \quad (5)$$

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

- **Assumed ratio of error variances:** the value of δ used in the calculations.

- **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

$$\textit{August hens} = 1.71535 + 0.691903\ \textit{Spring hens} \quad (6)$$

  The t or Z-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, the slope is statistically significant at the 5% significance level but the intercept is not.

- **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).

- **Estimated error variances:** the estimated values of σ_e^2, σ_u^2, and σ_v^2, which are the variances of the errors in Y, the errors in X, and the residual errors, respectively.

- **Confidence Intervals:** confidence intervals for the intercept and slope using the method selected on the *Analysis Options* dialog box.

Comparison of Estimates

This table shows the differences in the estimated intercept and slope when the model is fit using orthogonal regression, least squares regression of Y on X, and reverse least squares regression of X on Y:

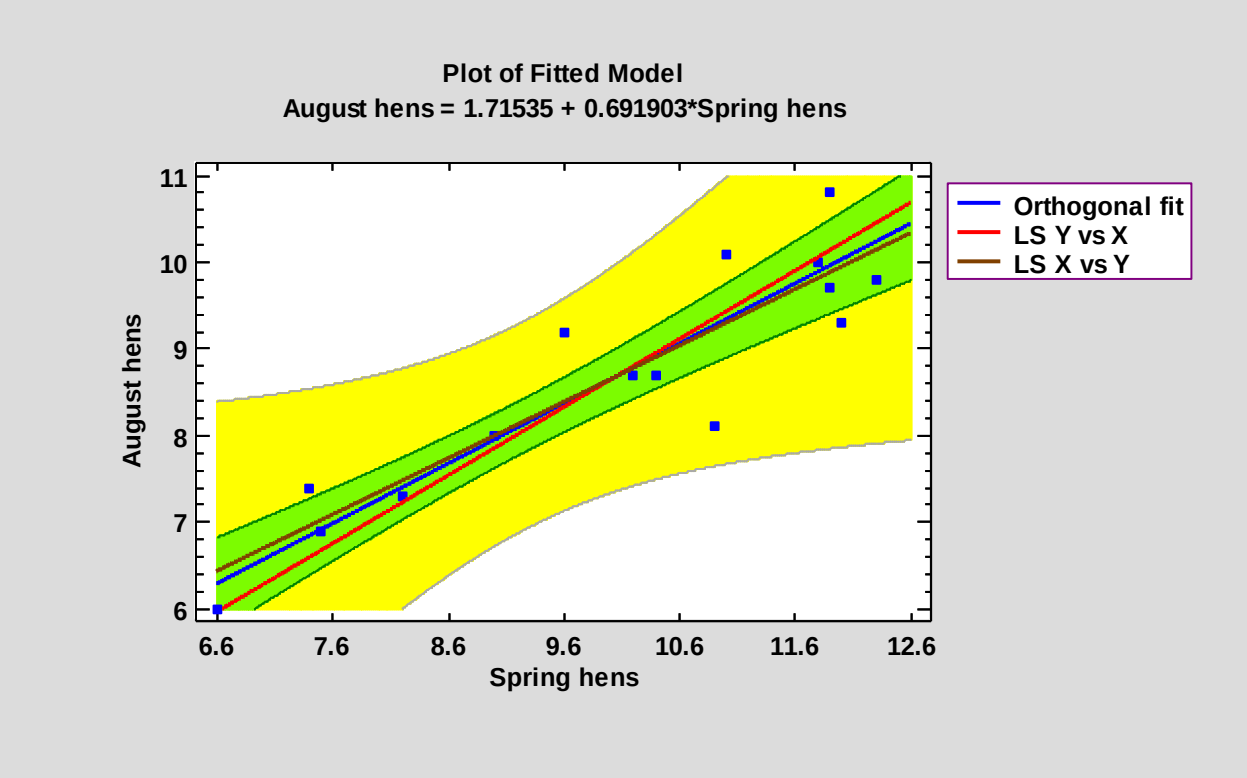
Comparison of Estimates

	Orthogonal (ratio=1.0)	*Least squares*	*Reverse least squares*
Intercept	1.71535	2.14227	0.765228
Slope	0.691903	0.64941	0.78647

Note that the estimated slope for the orthogonal regression is located between the estimated slope using the other 2 fitting methods.

Plot of Fitted Model

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



The plot may include:

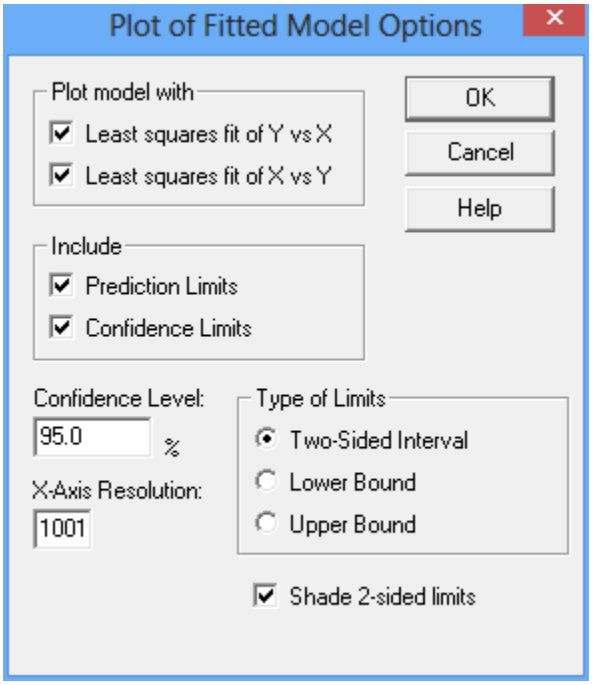
- The line of **orthogonal fit**. This is the equation that would be used to predict values of the dependent variable Y given values of the independent variable X.

- **Alternative models** estimated by a least squares regression of Y on X and a least squares regression of X on Y. These fits assume that X is known without error and that Y is known without error, respectively.

- **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.

- **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.

Pane Options



- **Plot model with:** which alternative regression lines to include on the plot.
- **Include:** the type of limits to include.
- **Type of Limits:** whether to include two-sided limits or one-sided bounds.
- **Shade 2-sided limits:** if two-sided intervals are plotted, whether to shade the area within the limits.
- **Confidence Level:** used to calculate the prediction and confidence limits.
- **X-Axis Resolution:** the number of locations along the X axis at which the line and limits are calculated.

Fitted Values and Residuals

This table shows the fitted values and residuals for each observation used to fit the model:

Fitted Values and Residuals

	Spring hens	*August hens*				
Row	Observed X	Observed Y	Fitted X	Fitted Y	Residual	Std. Residual
1	8.2	7.3	8.15838	7.36015	-0.0889527	-0.155113
2	11.8	10.0	11.8562	9.91872	0.120197	0.209596
3	11.0	10.1	11.362	9.57677	0.773719	1.34919
4	7.4	7.4	7.66416	7.01821	0.56457	0.98448
5	10.2	8.7	10.166	8.7492	-0.0727584	-0.126874
6	10.4	8.7	10.3012	8.84278	-0.211139	-0.368178
7	10.9	8.1	10.3586	8.88249	-1.15709	-2.0177
8	7.5	6.9	7.49784	6.90312	-0.00462066	-0.00805736

9	9.6	9.2	9.99415	8.63033	0.842383	1.46892	
10	12.0	9.3	11.664	9.78568	-0.718184	-1.25235	
11	11.9	9.7	11.7835	9.86838	-0.248993	-0.434187	
12	11.9	10.8	12.2982	10.2245	0.851007	1.48396	
13	12.3	9.8	12.1008	10.0879	-0.425754	-0.742418	
14	6.6	6.0	6.46809	6.19064	-0.281908	-0.491583	
15	9.0	8.0	9.02692	7.9611	0.057525	0.10031	

Included in the table are:

1. **Row:** row number *i* in the datasheet containing Y and X.
2. **Observed X:** the observed value of the independent variable X_i.
3. **Observed Y:** the observed value of the dependent variable Y_i.
4. **Fitted X:** the fitted value of the independent variable $\hat{x}_i$.
5. **Fitted Y:** the fitted value of the dependent variable $\hat{y}_i$.
6. **Residual:** the residual $\hat{v}_i$.
7. **Std. Residual:** the standardized value of the ith residual.

The fitted values for X are calculated from

$$\hat{x}_i = X_i - \frac{\hat{\sigma}_{uv}\hat{v}_i}{\hat{\sigma}_{vv}} \tag{7}$$

where

$$\hat{\sigma}_{vv} = \hat{\sigma}_e^2 + \hat{\beta}_1^2\hat{\sigma}_u^2 \tag{8}$$

$$\hat{\sigma}_{uv} = -\hat{\beta}_1\ \hat{\sigma}_u^2 \tag{9}$$

and

$$\hat{v}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \tag{10}$$

The fitted values for Y are calculated from

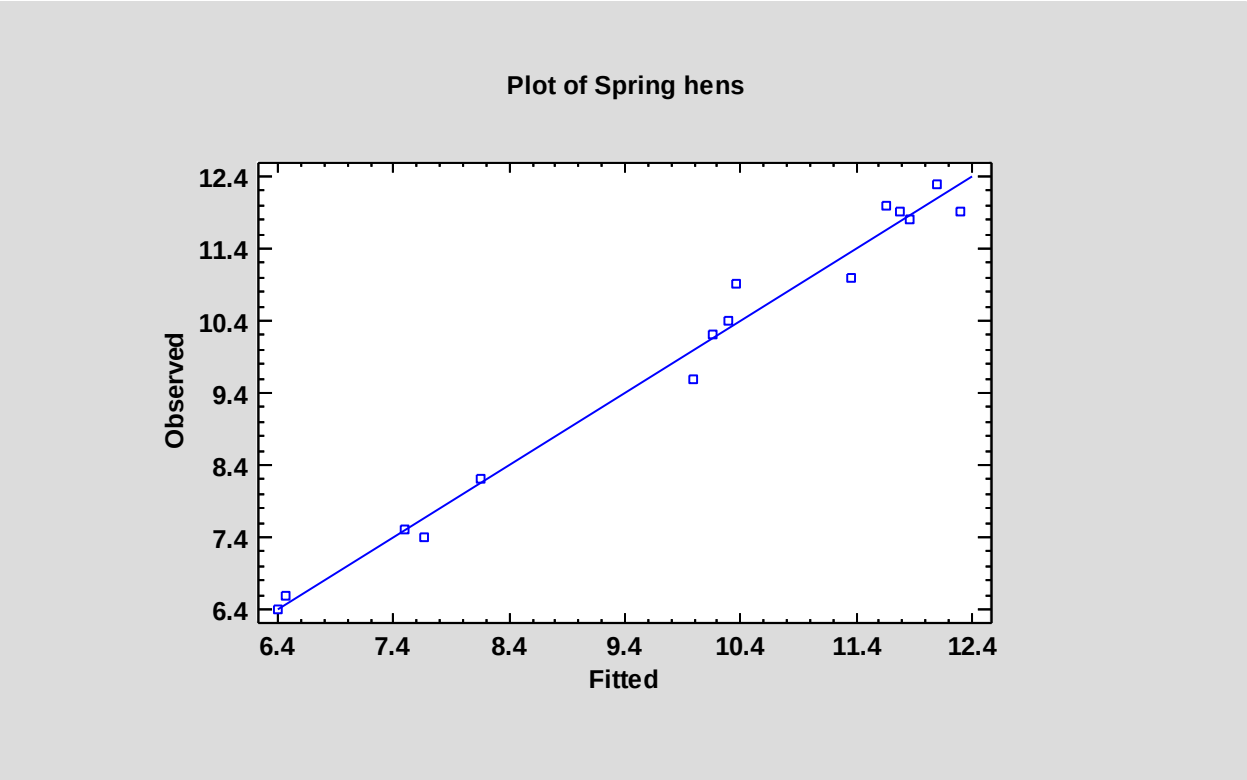
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1\hat{x}_i \tag{11}$$

The standardized residuals are calculated from

$$\frac{\hat{v}_i}{\sqrt{\hat{\sigma}_{vv}}} \tag{12}$$

Observed versus Fitted X

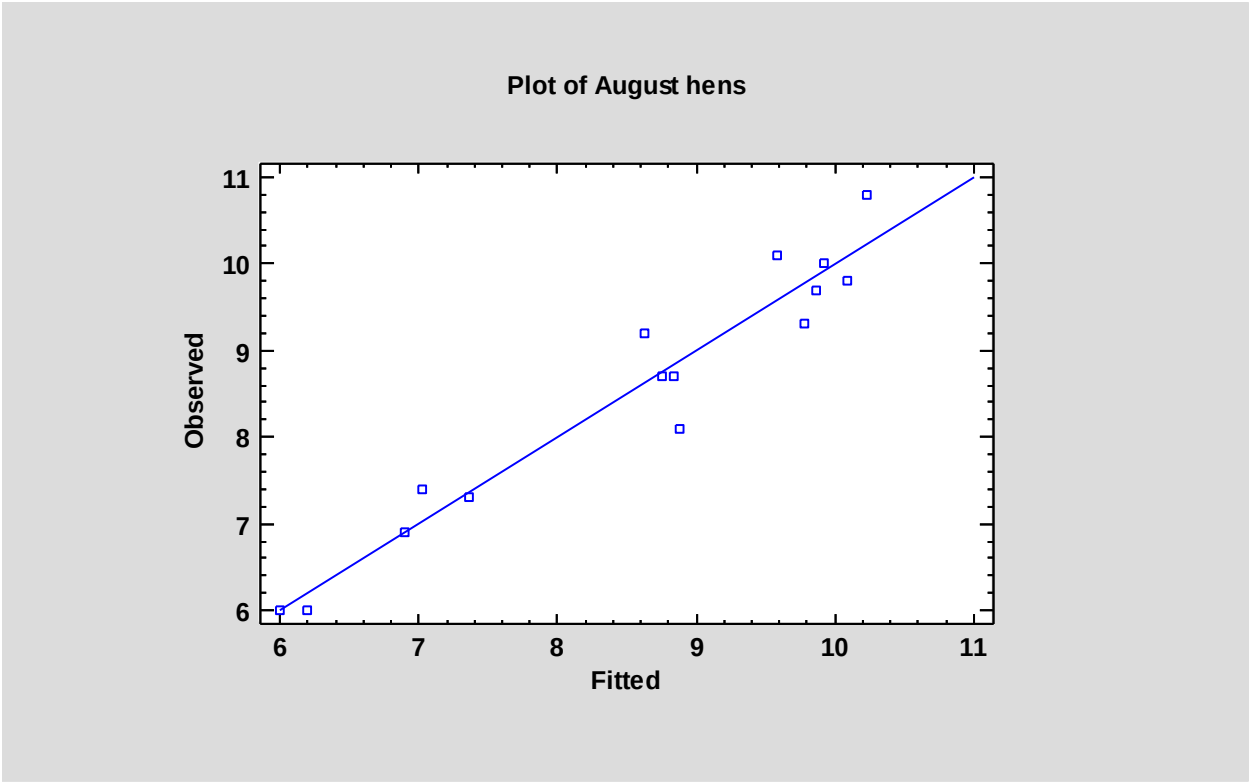
The *Observed versus Fitted X* plot shows the observed values of *X* on the vertical axis and the fitted values $\hat{x}$ on the horizontal axis.



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.

Observed versus Fitted Y

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



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

Predictions

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

Predicted Values

	Predicted	*Standard*	*Lower 95%*	*Upper 95%*	*Standard*	*Lower 95%*	*Upper 95%*
X	*Y*	*Deviation*	*Pred. Limit*	*Pred. Limit*	*Error*	*Conf. Limit*	*Conf. Limit*
6.0	5.86677	*1.12912*	*1.65637*	*10.0772*	*0.291539*	*3.65372*	*8.07981*
8.0	7.25057	0.674321	4.59489	9.90625	0.174109	5.92893	8.57222
10.0	8.63438	0.600493	6.79444	10.4743	0.155047	7.45743	9.81132
12.0	10.0182	0.996278	7.4248	12.6116	0.257238	8.06551	11.9709

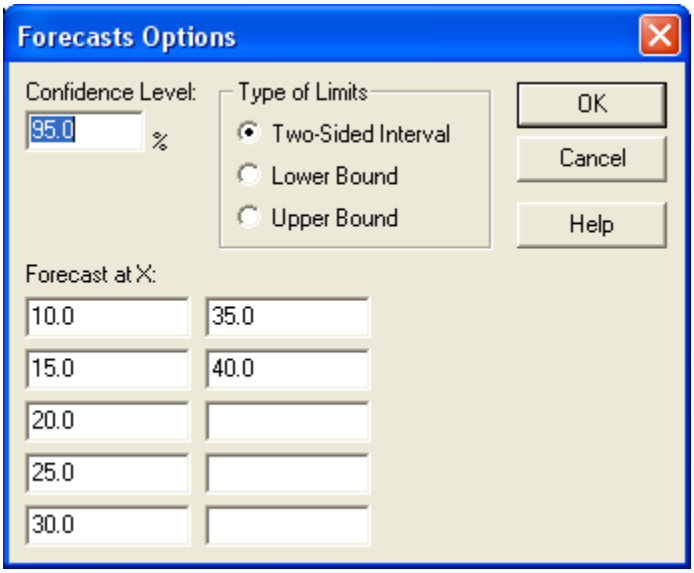
Included in the table are:

- **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.
- **Standard Deviation –** the estimated standard deviation of the predictions, determined using jackknifing.
- **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).
- **Standard Error –** the estimated standard error of the predictions, determined using jackknifing.

- **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 = 6, the best prediction of the mean value of Y is 5.867, although it could easily be anywhere between 3.654 and 8.080. In addition, one could predict with 95% confidence that any sample for which X = 6 would fall between 1.166 and 10.077. Obviously, the mean can be estimated much more closely that the observed value of any single random sample.

Pane Options



- **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.

Comparison of Alternative Models

The *Comparison of Alternative Models* pane shows the correlation coefficients associated with each of the 27 available models:

Comparison of Alternative Models

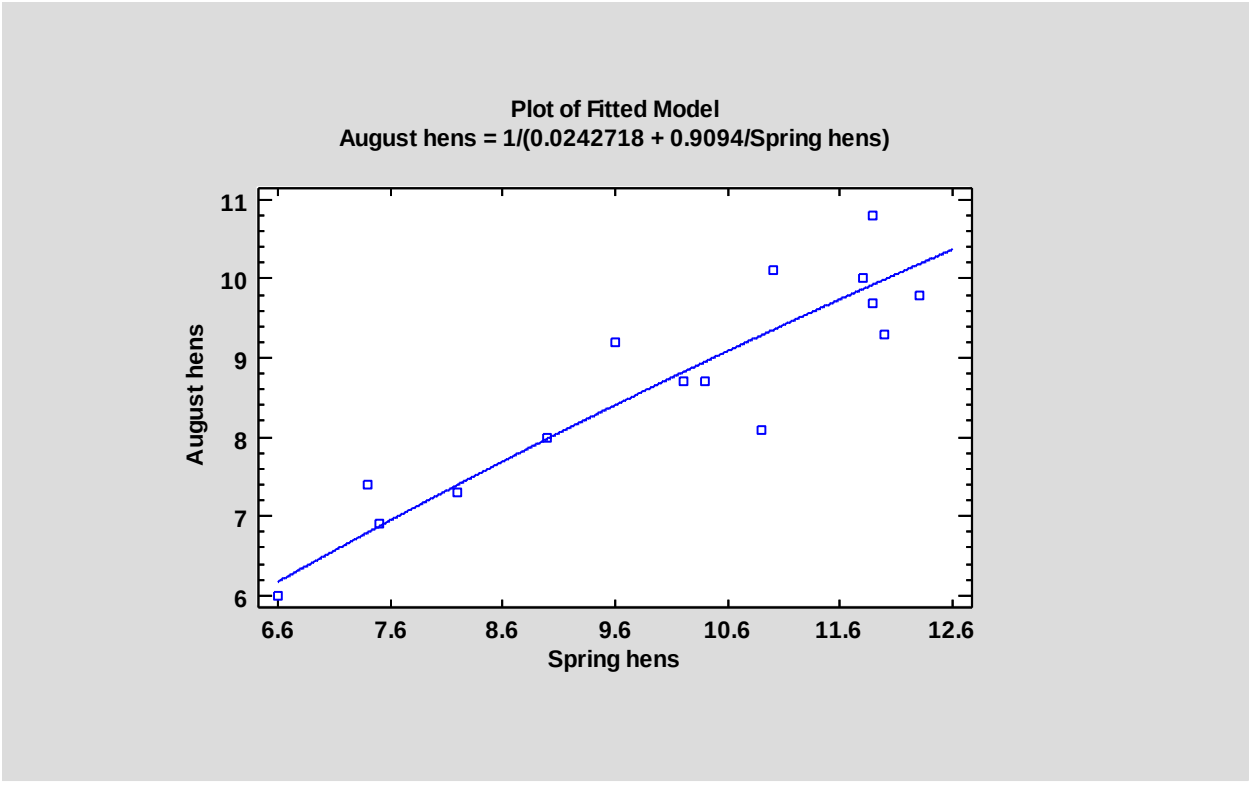
Model	*Correlation*
Double reciprocal	0.9361
Reciprocal-Y logarithmic-X	-0.9275
S-curve model	-0.9270
Multiplicative	0.9238
Reciprocal-Y square root-X	-0.9213
Logarithmic-Y square root-X	0.9202
Square root-Y reciprocal-X	-0.9195
Square root-Y logarithmic-X	0.9189
Double square root	0.9166

Exponential	0.9153
Reciprocal-Y	-0.9139
Square root-Y	0.9130
Logarithmic-X	0.9122
Square root-X	0.9111
Reciprocal-X	-0.9103
Linear	0.9087
Logarithmic-Y squared-X	0.9024
Square root-Y squared-X	0.9023
Squared-X	0.9003
Reciprocal-Y squared-X	-0.8964
Squared-Y square root-X	0.8952
Squared-Y	0.8950
Squared-Y logarithmic-X	0.8940
Double squared	0.8908
Squared-Y reciprocal-X	-0.8875
Logistic	<no fit>
Log probit	<no fit>

The models are listed in decreasing order of the absolute value of the correlation between X and Y after transforming one or both variables to linearize the model. When selecting an alternative model, consideration should be given to those models near the top of the list. However, since the correlations are calculated after transforming X and/or Y, the model with the highest correlation 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 *Double Reciprocal* model has the highest absolute correlation, it is a reasonable candidate for the sample data. Selecting that model using *Analysis Options* shows the following result:



While nonlinear, the amount of curvature in the model is very small.

Unusual Residuals

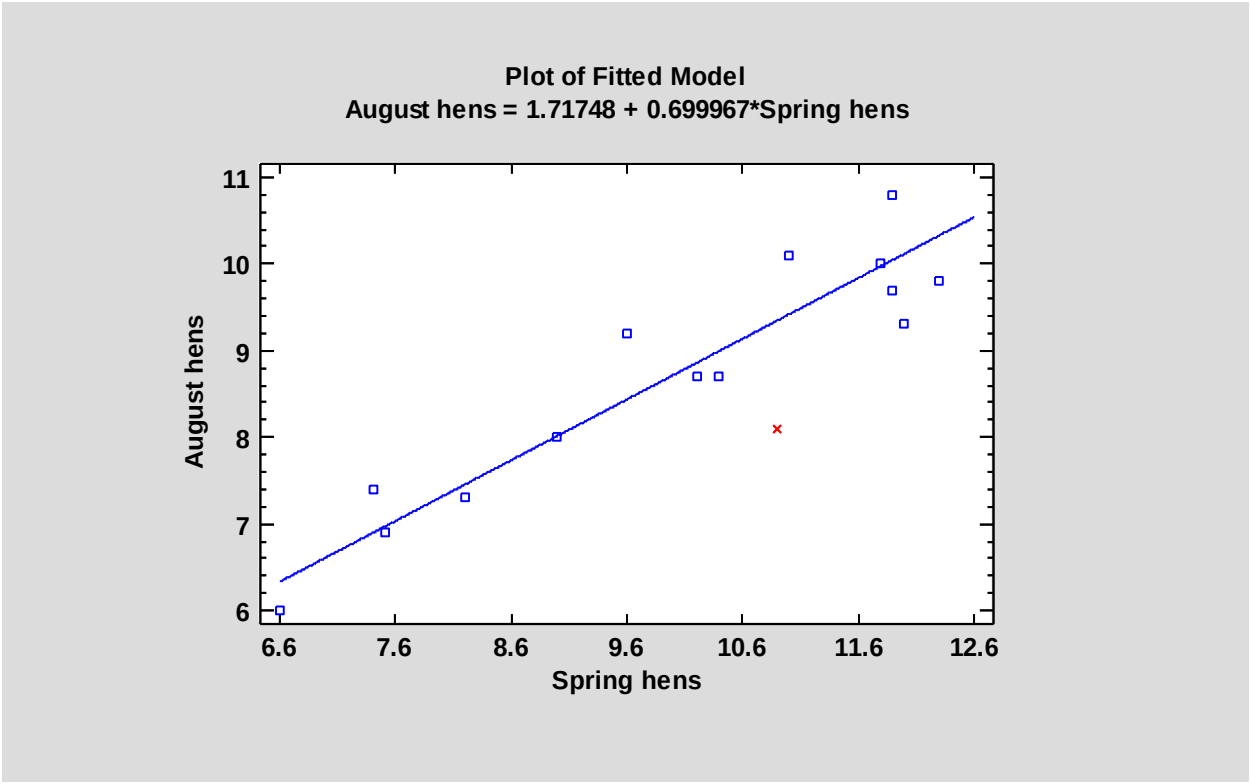
Once the model has been fit, it is useful to study the residuals to determine whether any outliers exist that should be removed from the data. The *Unusual Residuals* pane lists all observations that have standardized residuals of 2.0 or greater in absolute value.

Unusual Residuals

	Spring hens	*August hens*				
Row	Observed X	Observed Y	Fitted X	Fitted Y	Residual	Std. Residual
7	10.9	8.1	10.3586	8.88249	-1.15709	-2.0177

Standardized 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 #7 is slightly more than 2 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 #7 has little effect on the fitted model.

Residual Plots

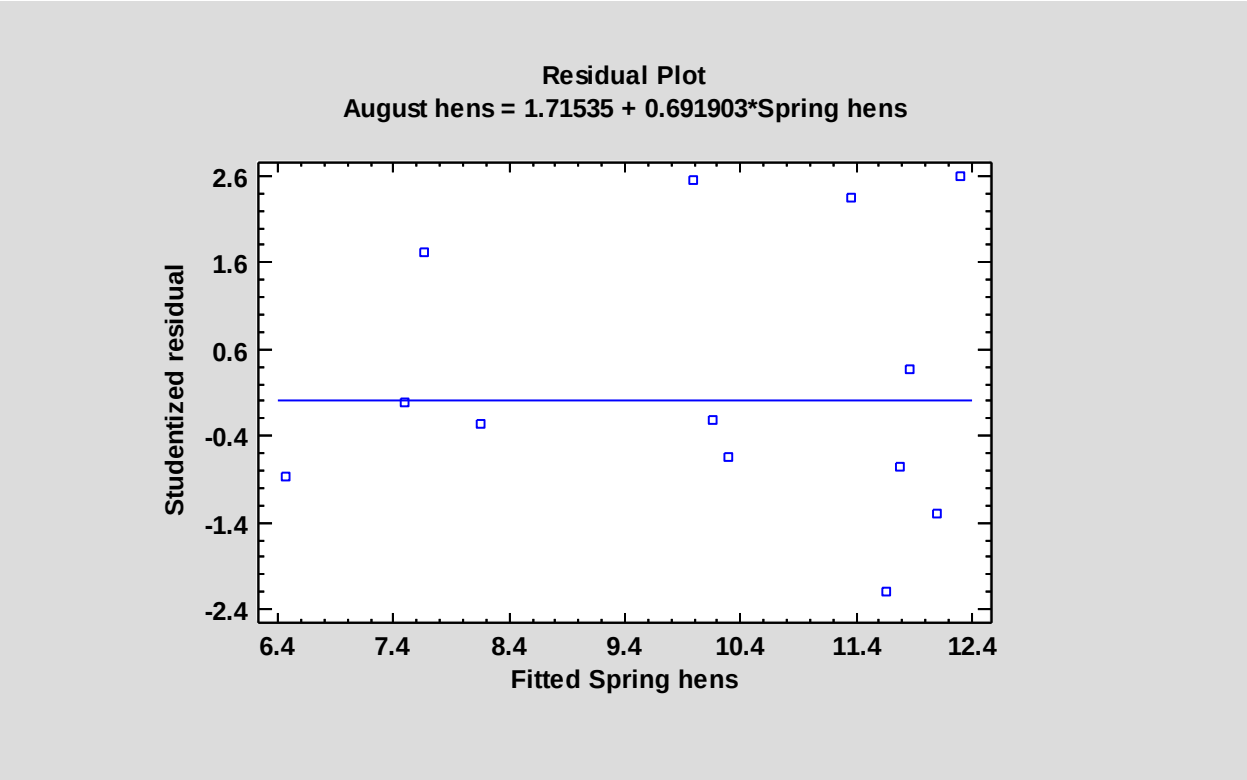
As with all statistical models, it is good practice to examine the residuals. The *Orthogonal Regression* procedure creates 4 residual plots:

1. Residuals versus fitted X.
2. Residuals versus fitted Y.
3. Residuals versus row number.
4. Normal probability plot of residuals.

In each case, *Pane Options* gives you the choice of plotting either the ordinary residuals $\hat{v}_i$ or the standardized residuals.

Residuals versus Fitted X

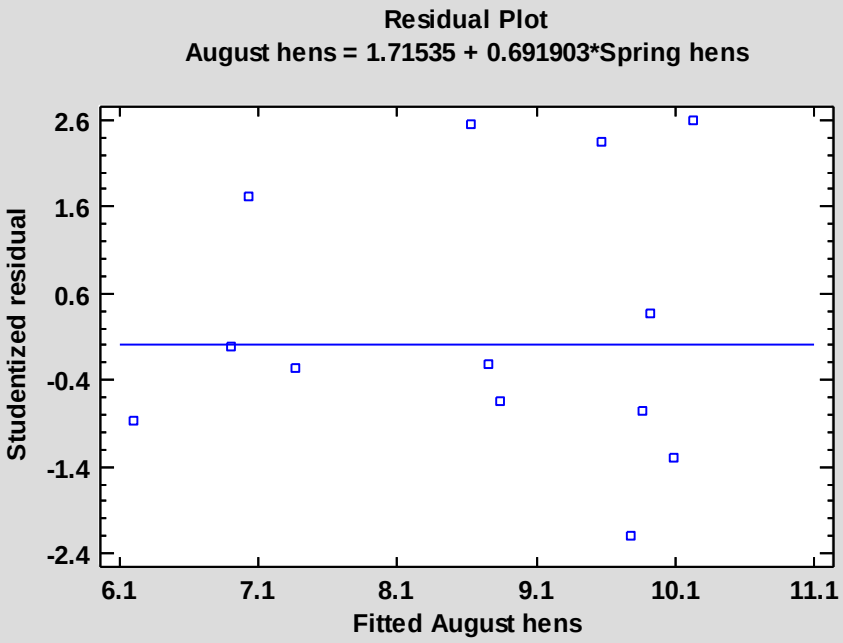
This plot displays $\hat{v}_i$ versus $\hat{x}_i$.



It is helpful in detecting the need for a curvilinear model and visualizing any change in variability as a function of X.

Residuals versus Fitted Y

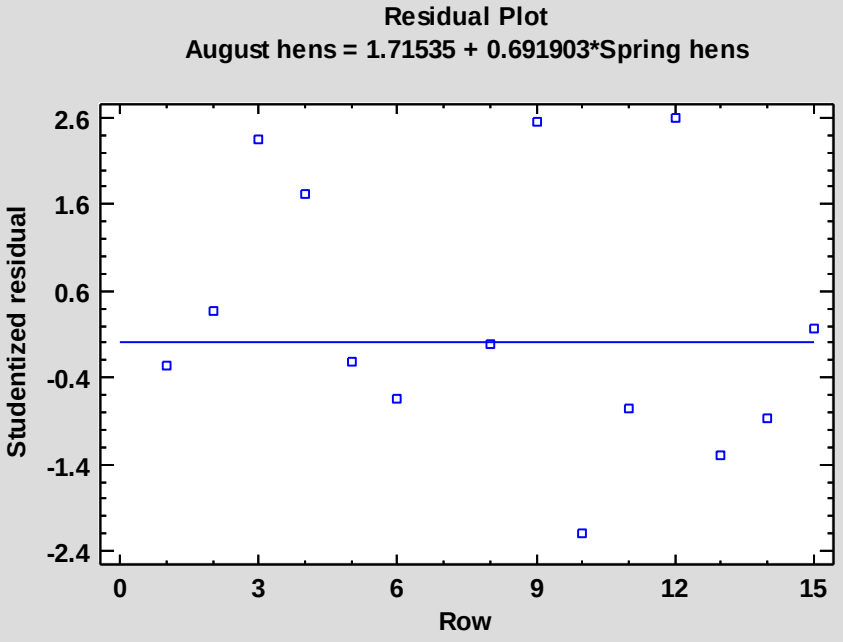
This plot displays $\hat{v}_i$ versus $\hat{y}_i$.



It is helpful in detecting the need for a curvilinear model and visualizing any change in variability as a function of Y.

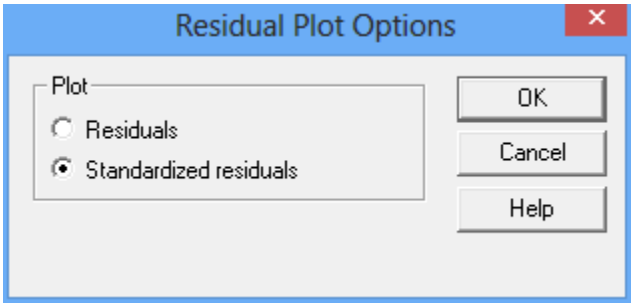
Residuals versus Row Order

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.

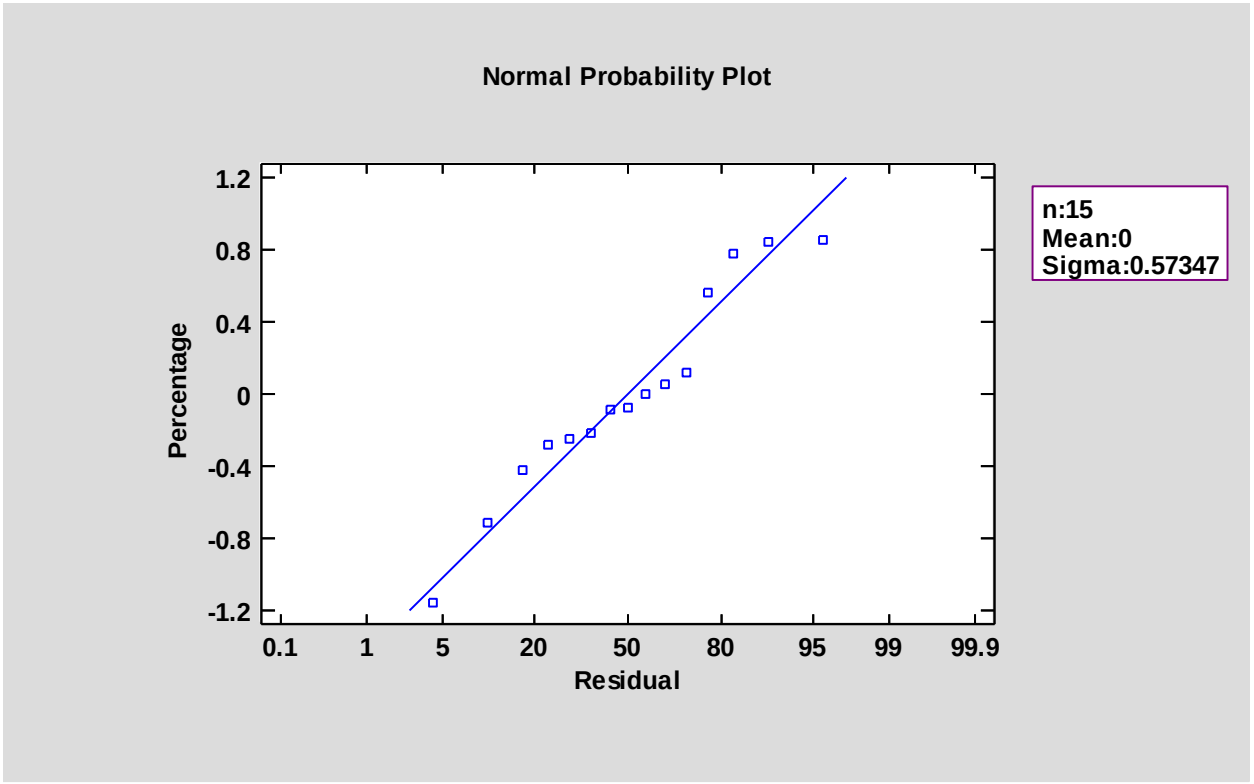
Pane Options



Select either the residuals or standardized residuals.

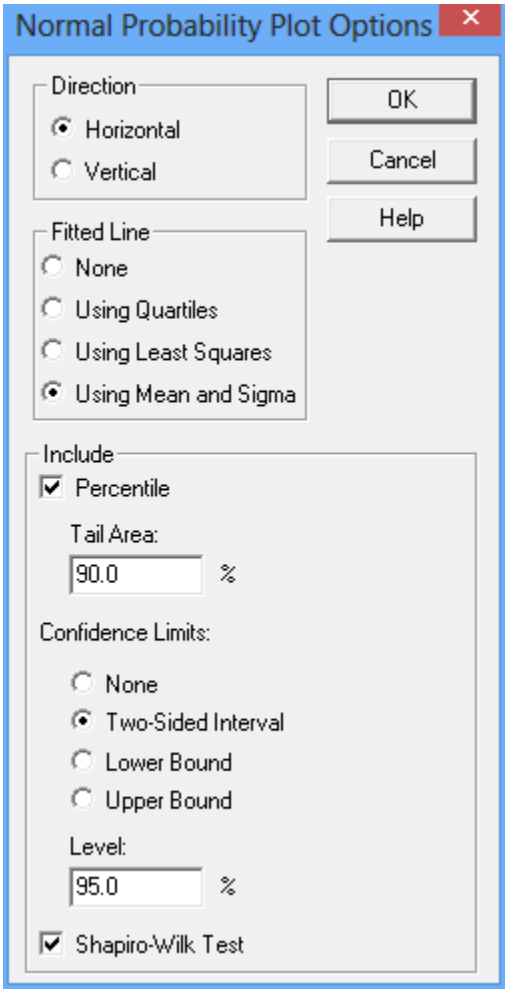
Residuals Probability Plot

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 as in the plot below:



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

Pane Options



- **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.

- **Include - Percentile:** displays a reference line at the specified percentile. The percentile is estimated using the fitted line.

- **Include – Confidence Limits:** displays confidence limits around the fitted line (only available when estimating the line *Using mean and sigma*). The confidence level applies to each percentile separately. The limits are the same as those displayed in the *Percentiles* table.

- **Include – Shapiro-Wilk Test**: displays the calculated value of the Shapiro-Wilk W test and its associated P value. This test is described in the document titled *Distribution Fitting – Uncensored Data*. A small P-value indicates that the data are not well modeled by a normal distribution.

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

Save Results

The following results may be saved to the datasheet:

1. *Fitted X Values* – the fitted values $\hat{x}_i$.
2. *Fitted Y Values* – the fitted values $\hat{y}_i$.
3. *Residuals* – the residuals $\hat{v}_i$.
4. *Standardized Residuals* – the standardized residuals.
5. *Coefficients* – the estimated model coefficients.

Calculations

Coefficient Estimates

$$\hat{\beta}_1 = \frac{m_{YY} - \partial m_{XX} + \left[(m_{YY} - \partial m_{XX})^2 + 4\partial m_{XY}^2 \right]^{1/2}}{2m_{XY}} \tag{13}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \tag{14}$$

where

$$m_{XX} = \sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2 / (n-1) \tag{15}$$

$$m_{YY} = \sum_{i=1}^{n} \left(Y_i - \bar{Y} \right)^2 / (n-1) \tag{16}$$

$$m_{XY} = \sum_{i=1}^{n} \left(X_i - \bar{X} \right)\left(Y_i - \bar{Y} \right) / (n-1) \tag{17}$$

Standard Errors

The standard errors are calculating using Equation (1.3.12) in Fuller (1987).

Correlation Coefficient

$$r = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X} \right)\left(Y_i - \bar{Y} \right)}{\sqrt{\sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2 \sum_{i=1}^{n} \left(Y_i - \bar{Y} \right)^2}} \tag{18}$$

Prediction Intervals

Prediction intervals for y given x are obtained using a method known as jackknifing. Jackknifing estimates the standard deviation and standard error of the predictions by refitting the model *n* times, each time removing one observation. A prediction is made each time the model is fit. The variability amongst the *n* predictions is then used to estimate the prediction standard error.

References

Carroll, R.J. and Ruppert, D. (1996). "The Use and Misuse of Orthogonal Regression in Linear Errors-in-Variables Models". The American Statistician. Vol. 50 (1) 1-6.

Fuller, Wayne A. (1987) Measurement Error Models. John Wiley and Sons, New York.

Linnet, K. (1990) "Estimation of the Linear Relationship Between the Measurements of Two Methods with Proportional Errors". Statistics in Medicine. John Wiley and Sons. 1463-1471.

