Comparison of Regression Lines

Summary

The **Comparison of Regression Lines** procedure is designed to compare the regression lines relating Y and X at two or more levels of a categorical factor. Tests are performed to determine whether there are significant differences between the intercepts and the slopes at the different levels of that factor. The regression lines are plotted, unusual residuals are identified, and predictions are made using the fitting model.

Sample StatFolio: *compare reg.sgp*

Sample Data:

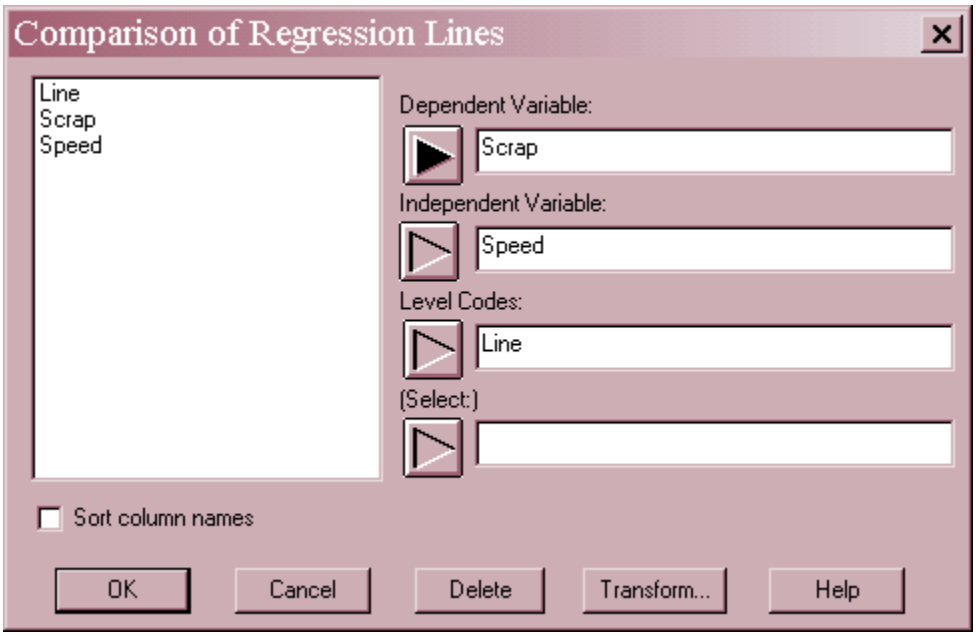
The file *soap.sgd* contains data on the amount of scrap produced by two production lines as a function of the speed at which those lines are run. The data, from Neter et al. (1996), is shown below:

Line	*Scrap*	*Speed*
1	218	100
1	248	125
1	360	220
1	351	205
1	470	300
1	394	255
1	332	225
1	321	175
1	410	270
1	260	170
1	241	155
1	331	190
1	275	140
1	425	290
1	367	265
2	140	105
2	277	215
2	384	270
2	341	255
2	215	175
2	180	135
2	260	200
2	361	275
2	252	155
2	422	320
2	273	190
2	410	295

It is desired to determine whether the relationship between *Scrap* (Y) and *Speed* (X) is different for the two *Lines*.

Data Input

The data input dialog box requests the names of the input variables:



- **Dependent Variable:** numeric column containing the *n* values of Y.
- **Independent Variable:** numeric column containing the *n* values of X.
- **Level Codes**: numeric or nonnumeric columns identifying the level of the categorical factor. A separate intercept and slope will be estimated for each unique value in this column.
- **Select:** subset selection.

Analysis Summary

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

Comparison of Regression Lines - Scrap vs. Speed
Dependent variable: Scrap
Independent variable: Speed
Level codes: Line

Number of complete cases: 27
Number of regression lines: 2

Multiple Regression Analysis

		Standard	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
CONSTANT	97.9653	19.1817	5.10724	0.0000
Speed	1.14539	0.0895535	12.79	0.0000
Line=2	-90.3909	28.3457	-3.18887	0.0041
Speed*Line=2	0.176661	0.128838	1.37119	0.1835

Coefficients

Line	*Intercept*	*Slope*
1	97.9653	1.14539
2	7.57446	1.32205

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	169165.0	d	56388.2	130.95	0.0000
Residual	9904.06	d	430.611		
Total (Corr.)	179069.0	d			

R-Squared = 94.4691 percent
R-Squared (adjusted for d.f.) = 93.7477 percent
Standard Error of Est. = 20.7512
Mean absolute error = 16.0087
Durbin-Watson statistic = 1.79915 (P=0.2120)
Lag 1 residual autocorrelation = 0.0911102

Residual Analysis

	Estimation	*Validation*
n	27	
MSE	430.611	
MAE	16.0087	
MAPE	5.45666	
ME	4.31589E-14	
MPE	-0.437749	

Included in the output are:

- **Data Summary:** a summary of the input data. In the example, there are a total of $n = 27$ observations. The categorical factor contains $m = 2$ levels.

- **Coefficients:** the estimated coefficients, standard errors, t-statistics, and P values. The general equation for the model is

$$Y = \beta_0 + \beta_1 X + \beta_2 I_1 + \beta_3 I_1 X + \beta_4 I_2 + \beta_5 I_2 X + \ldots + \beta_{2m-2} I_{m-1} + \beta_{2m-1} I_{m-1} X \qquad (1)$$

  where $I_1, I_2, \ldots, I_{m-1}$ are indicator variables for the categorical factors, where $I_j = 1$ if the categorical factor is at its (j+1)-st level and 0 otherwise. The terms involving only the indicator variables allow the intercepts to vary amongst levels of the categorical factor, while

the terms containing the cross-products of the indicator variables with X allow the slopes to vary. 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.

- **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. The residual sum of squares is further partitioned into a lack-of-fit component and a pure error component. Of particular interest is the F-test and the associated P-value. The F-test on the *Model* line 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.

- **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 94.5% of the variability amongst the observed amounts of *Scrap*.

  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). This 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.

For the sample data, the fitted model is

$$Scrap = 97.9653 + 1.14539*Speed - 90.3909*(Line=2) + 0.176661*Speed*(Line=2)$$

where *Line=2* takes the value 1 for line #2 and 0 for line #1. To see that this corresponds to 2 separate regression lines, the model can be written separately for each line:

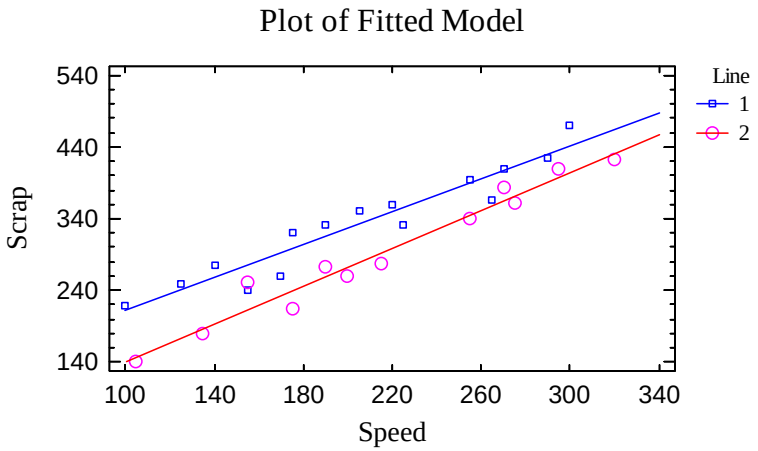
Line #1: Scrap = 97.9653 + 1.14539**Speed*

Line #2: Scrap = (97.9653-90.3909) + (1.14539+0.176661)**Speed*

The third and fourth coefficients in the full model represent the difference in the intercepts and slopes between the two lines.

Plot of Fitted Model

This *Plot of Fitted Model* pane shows the two regression lines:



There is a noticeable offset between the lines, with line #1 producing more scrap at all speeds. In addition, the slope for line #2 is somewhat greater than for line #1.

Conditional Sums of Squares

The *Conditional Sums of Squares* pane displays an analysis of variance table that may be used to determine whether the intercepts and the slopes of the lines are significantly different:

Further ANOVA for Variables in the Order Fitted

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Speed	149661.0	1	149661.0	347.55	0.0000
Intercepts	18694.1	1	18694.1	43.41	0.0000
Slopes	809.623	1	809.623	1.88	0.1835
Model	169165.0	3			

Of primary interest are the F-tests and P-values for the *Intercepts* and *Slopes*.

1. The F-test for *Slopes* tests the hypotheses:

Null Hypothesis: slopes of the lines are all equal

Alt. Hypothesis: slopes of the lines are not all equal

If the P-Value is small (less than 0.05 if operating at the 5% significance level), then the slopes of the lines vary significant amongst the levels of the categorical factor.

2. The F-test for *Intercepts* tests the hypotheses:

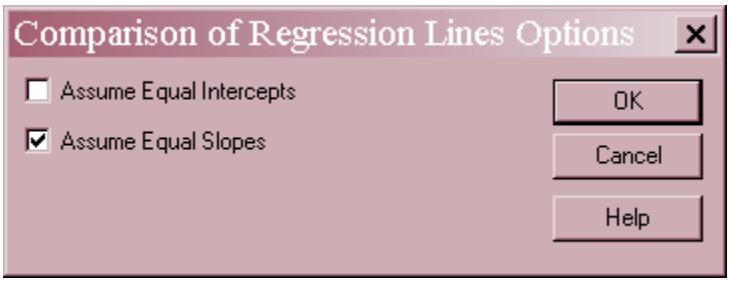
   Null Hypothesis: intercepts of the lines are all equal

   Alt. Hypothesis: intercepts of the lines are not all equal

   If the P-Value is small (less than 0.05 if operating at the 5% significance level), then the intercepts of the lines vary significant amongst the levels of the categorical factor.

In the sample data, the large P-value for *Slopes* indicates that the slope of line #1 and line #2 are *not* significantly different. However, the intercepts of the two lines *are* significantly different.

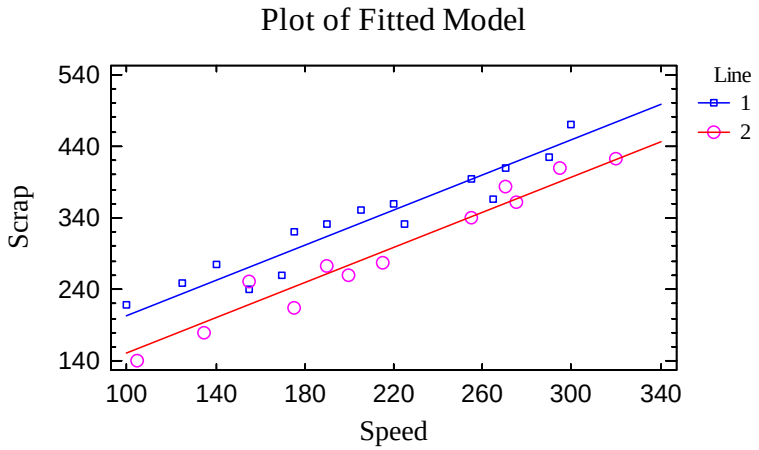
Analysis Options



- **Assume Equal Intercepts:** select this option to fit a model with equal intercepts at all levels of the categorical factor.

- **Assume Equal Slopes:** select this option to fit a model with equal slopes at all levels of the categorical factor.

Example – Fitting Parallel Regression Lines
Selecting *Assume Equal Slopes* forces the regression lines to be parallel:



This is accomplished by removing the terms in the model that contain cross-products between the indicator variables and X. In the current example, the new model is:

Scrap = 80.411 + 1.23074**Speed* - 53.1292*(*Line=2*)

The last coefficient in the model indicates that line #2 produces 53.1292 units less scrap on average than line #1.

Forecasts

The model can be used to predict Y at selected values of X.

Forecasts

		95.00%		*95.00%*	
Speed	*Predicted*	*Prediction*	*Limits*	*Confidence*	*Limits*
Line	*Scrap*	*Lower*	*Upper*	*Lower*	*Upper*
100.0					
1	203.485	156.234	250.736	185.288	221.682
2	150.356	102.339	198.373	130.255	170.457
200.0					
1	326.559	281.516	371.602	315.274	337.844
2	273.43	227.992	318.868	260.661	286.199
300.0					
1	449.633	402.823	496.443	432.614	466.652
2	396.504	349.71	443.298	379.53	413.478

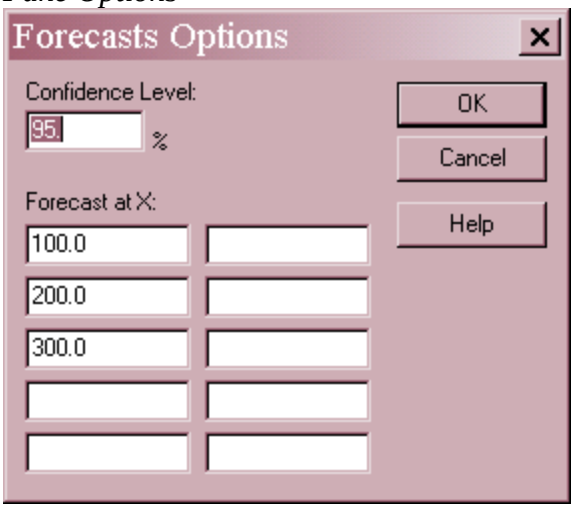
Included in the table are:

- **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…
- **Confidence limits** - confidence limits for the mean value of Y.

For example, the predicted *Scrap* for *Line* #1 at *Speed* = 200 is approximately 327. New observations from that line can be expected to be between 282 and 372 with 95% confidence. On

average, the scrap from that line at that speed is estimated to be somewhere between 315 and 338.

Pane Options



- **Confidence Level:** confidence percentage for the intervals.
- **Forecast at X**: up to 10 values of X at which to make predictions.

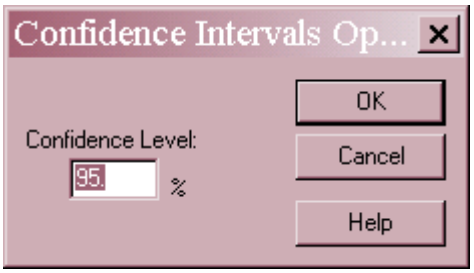
Confidence Intervals

The *Confidence Intervals* pane shows the potential estimation error associated with each coefficient in the model.

95.0% confidence intervals for coefficient estimates

		Standard		
Parameter	*Estimate*	*Error*	*Lower Limit*	*Upper Limit*
CONSTANT	80.411	14.5438	50.394	110.428
Speed	1.23074	0.0655522	1.09545	1.36603
Line=2	-53.1292	8.21003	-70.0739	-36.1845

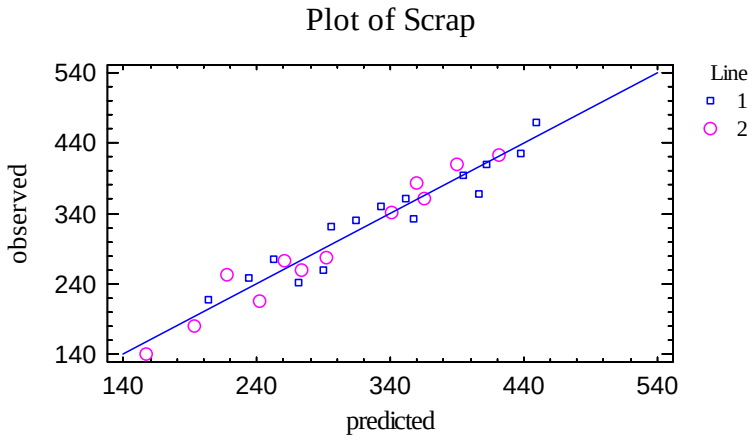
Pane Options



- **Confidence Level:** percentage level for the interval or bound.

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 Y to high values of Y might also indicate the need to transform the dependent variable before fitting a model to the data.

Residual Plots

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

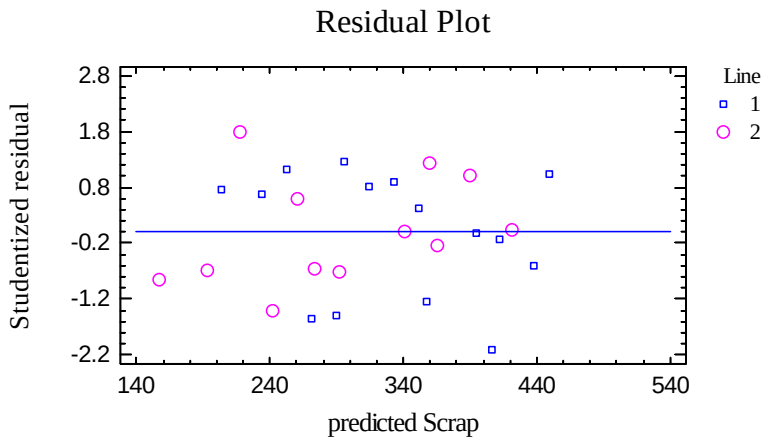
$$e_i = y_i - \hat{y}_i \tag{3}$$

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

The *Comparison of Regression Lines* procedure various type of residual plots, depending on *Pane Options*.

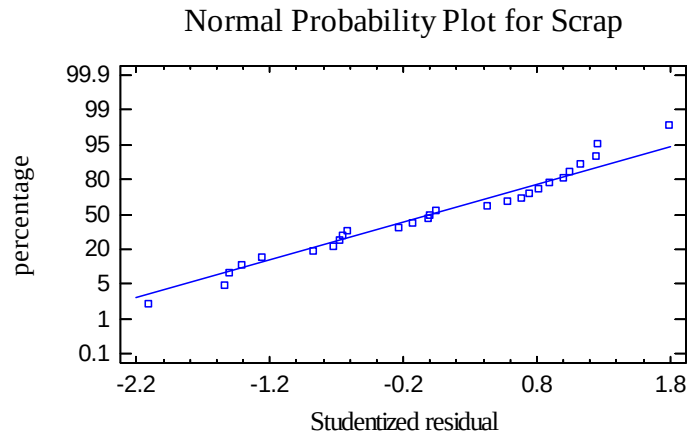
Scatterplot versus X

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



Normal Probability Plot

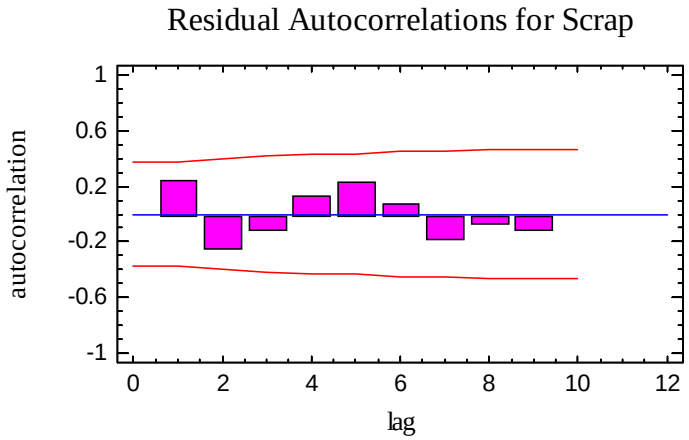
This plot can be used to determine whether or not the deviations around the line follow a normal distribution, which is the assumption used to form the prediction intervals.



If the deviations follow a normal distribution, they should fall approximately along a straight line.

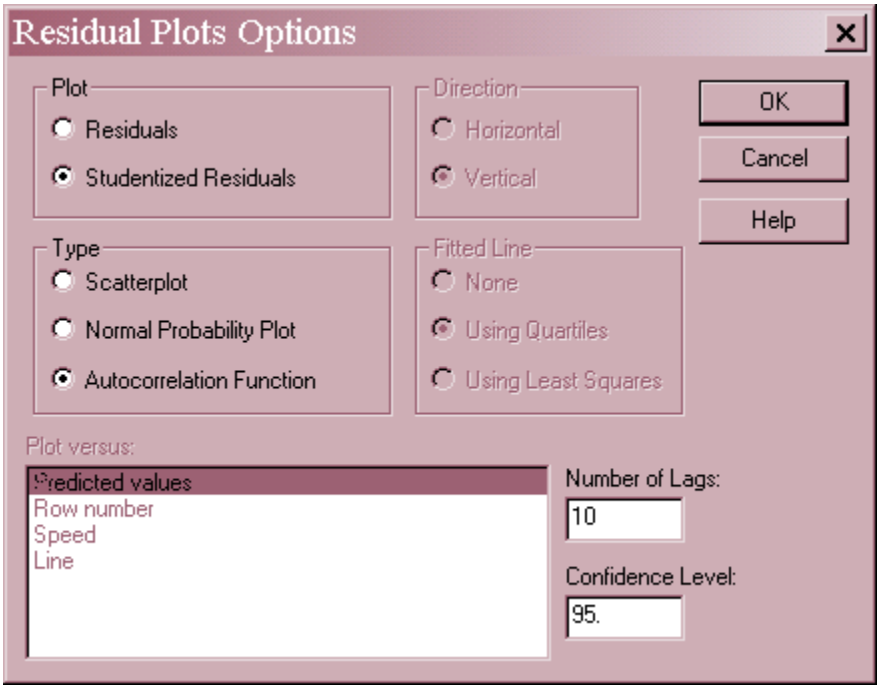
Residual Autocorrelations

This plot calculates the autocorrelation between residuals as a function of the number of rows between them in the datasheet.



It is only relevant if the data have been collected sequentially. Any bars extending beyond the probability limits would indicate significant dependence between residuals separated by the indicated "lag", which would violate the assumption of independence made when fitting the regression model.

Pane Options



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

- **Type:** the type of plot to be created. A *Scatterplot* is used to test for curvature. A *Normal Probability Plot* is used to determine whether the model residuals come from a normal distribution. An *Autocorrelation Function* is used to test for dependence between consecutive residuals.

- **Plot Versus**: for a *Scatterplot*, the quantity to plot on the horizontal axis.

- **Number of Lags**: for an *Autocorrelation Function*, the maximum number of lags. For small data sets, the number of lags plotted may be less than this value.

- **Confidence Level:** for an *Autocorrelation Function*, the level used to create the probability limits.

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	*Y*	*Y*	*Residual*	*Residual*
15	367.0	406.557	-39.5573	-2.11

Studentized residuals greater than 3 in absolute value correspond to points more than 3 standard deviations from the fitted model, which is a rare event for a normal distribution.

Note: 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.

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

		Mahalanobis		*Cook's*
Row	*Leverage*	*Distance*	*DFITS*	*Distance*
15	0.100555	1.83337	-0.706032	0.0131351

Average leverage of single data point = 0.111111

Points are placed on this list for one of the following reasons:

- **Leverage** – measures how distant an observation is from the mean of all *n* observations in the space of the *independent* variables. The higher the leverage, the greater the impact of the point on the fitted values $\hat{y}$. Points are placed on the list if their leverage is more than 3 times that of an average data point.

- **Mahalanobis Distance –** measures the distance of a point from the center of the collection of points in the multivariate space of the independent variables. Since this distance is related to *leverage*, it is not used to select points for the table.

- **DFITS** – measures the difference between the predicted values $\hat{y}_i$ when the model is fit with and without the i-th data point. Points are placed on the list if the absolute value of DFITS exceeds $2p/\sqrt{n}$, where *p* is the number of coefficients in the fitted model.

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. *Standard Errors of Predictions* - the standard errors for the *n* predicted values.
3. *Lower Limits for Predictions* – the lower prediction limits for each predicted value.
4. *Upper Limits for Predictions* – the upper prediction limits for each predicted value.
5. *Standard Errors of Means* - the standard errors for the mean value of Y at each of the *n* values of X.
6. *Lower Limits for Forecast Means* – the lower confidence limits for the mean value of Y at each of the *n* values of X.
7. *Upper Limits for Forecast Means*– the upper confidence limits for the mean value of Y at each of the *n* values of X.
8. *Residuals* – the *n* residuals.
9. *Studentized Residuals* – the *n* Studentized residuals.
10. *Leverages* – the leverage values corresponding to the *n* values of X.
11. *DFITS Statistics* – the value of the DFITS statistic corresponding to the *n* values of X.
12. *Mahalanobis Distances* – the Mahalanobis distance corresponding to the *n* values of X.
13. *Coefficients* – the estimated model coefficients.

Calculations

Details on the calculations in this procedure may be found in the *Multiple Regression* documentation.

