Polynomial Regression

Summary

The **Polynomial Regression** procedure is designed to construct a statistical model describing the impact of a single quantitative factor X on a dependent variable Y. A polynomial model involving X and powers of X is fit to the data. Tests are run to determine the proper order of the polynomial. The fitted model may be plotted with confidence limits and/or prediction limits. Residuals may also be plotted and influential observations identified.

Sample StatFolio: *polynomial reg.sgp*

Sample Data:

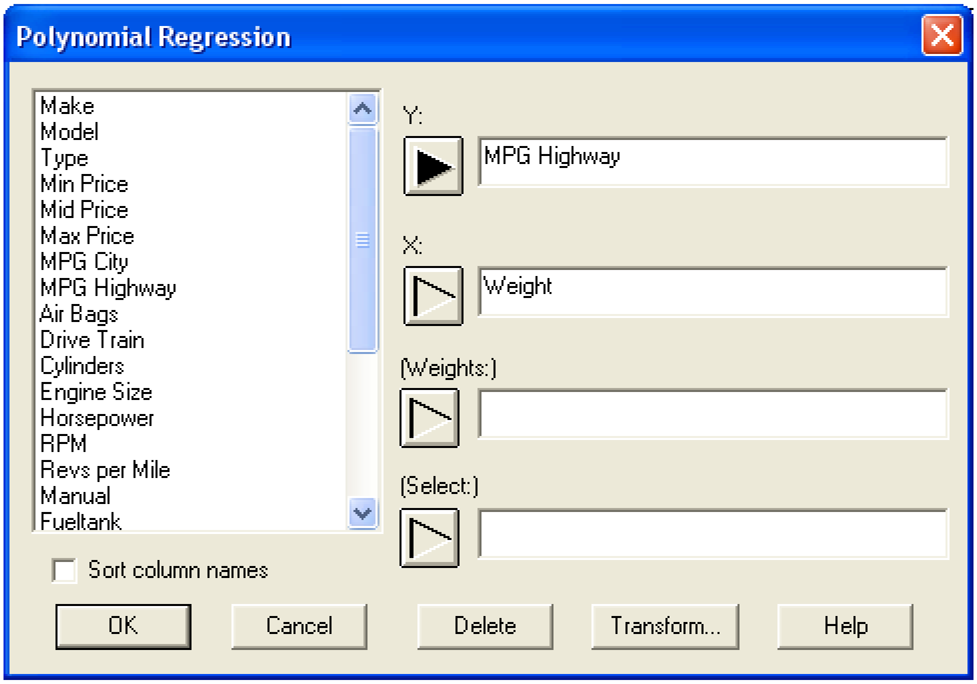
The file *93cars.sgd* contains information on 26 variables for *n* = 93 makes and models of automobiles, taken from Lock (1993). The table below shows a partial list of 4 columns from that file:

Make	*Model*	*MPG Highway*	*Weight*
Acura	Integra	31	2705
Acura	Legend	25	3560
Audi	90	26	3375
Audi	100	26	3405
BMW	535i	30	3640
Buick	Century	31	2880
Buick	LeSabre	28	3470
Buick	Roadmaster	25	4105
Buick	Riviera	27	3495
Cadillac	DeVille	25	3620
Cadillac	Seville	25	3935
Chevrolet	Cavalier	36	2490

A model is desired relating *MPG Highway* to the *Weight* of the vehicles.

Data Input

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



Y: MPG Highway

X: Weight

(Weights:)

(Select:)

□ Sort column names

- **Y:** numeric column containing the *n* observations for the dependent variable Y.

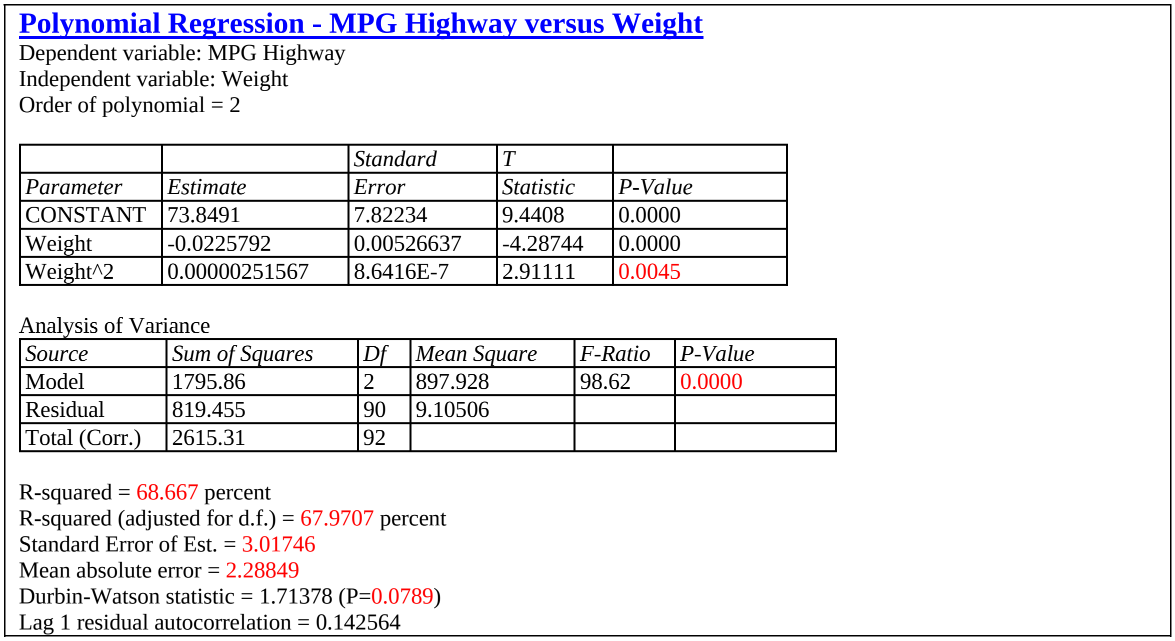
- **X:** numeric column containing the *n* values for the independent variable X.

- **Weight:** an optional numeric column containing weights to be applied to the squared residuals when performing a weighted least squares fit.

- **Select:** subset selection.

Analysis Summary

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



Polynomial Regression - MPG Highway versus Weight

Dependent variable: MPG Highway
Independent variable: Weight
Order of polynomial = 2

		Standard	T	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
CONSTANT	73.8491	7.82234	9.4408	0.0000
Weight	-0.0225792	0.00526637	-4.28744	0.0000
Weight^2	0.00000251567	8.6416E-7	2.91111	0.0045

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	1795.86	2	897.928	98.62	0.0000
Residual	819.455	90	9.10506		
Total (Corr.)	2615.31	92			

R-squared = 68.667 percent
R-squared (adjusted for d.f.) = 67.9707 percent
Standard Error of Est. = 3.01746
Mean absolute error = 2.28849
Durbin-Watson statistic = 1.71378 (P=0.0789)
Lag 1 residual autocorrelation = 0.142564

Included in the output are:

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

  $$Y = \beta_0 + \beta_1 X + \beta_2 X^2 \qquad (1)$$

  is fit, although a different order polynomial 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

  $$\textit{MPG Highway} = 73.8491 - 0.0225792*\textit{Weight} + 0.00000251567*\textit{Weight}^2 \qquad (2)$$

  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. Of particular interest when fitting a polynomial is the P-value for the highest order term. If this term is not significant, then the model might reasonably be simplified by lowering the order of the polynomial. In the sample data, the P-value for *Weight*^{2} is small, so that a model of at least order 2 is needed to adequately describe the relationship between Y and X.

- **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. 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:

  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 68.5% of the variability in the miles per gallon. The remaining 31.5% is attributable to deviations around the line, which may be due to other factors, to measurement error, or to a failure of the current polynomial 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.

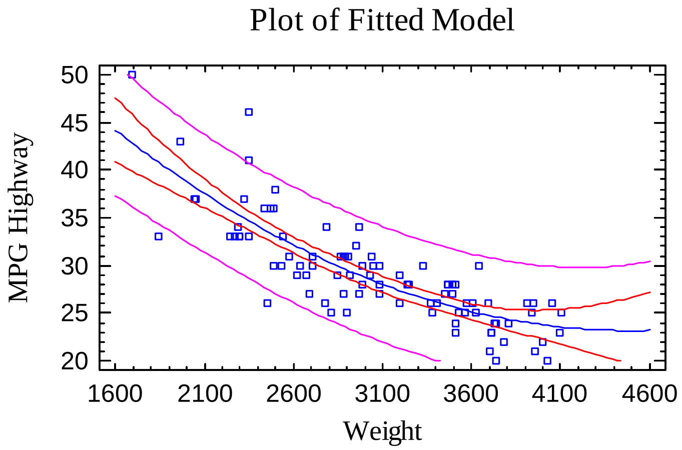
  Mean Absolute Error – the average absolute value of the residuals.

  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, the P-value is greater than 0.05, so there is not a significant correlation at the 5% significance level.

  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.

Plot of Fitted Model

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



The plot includes:

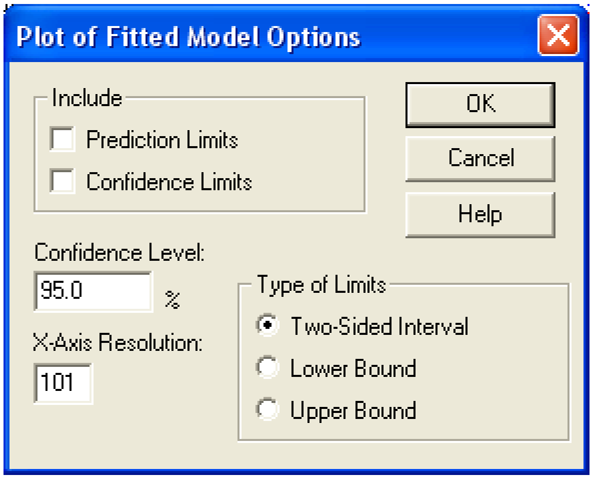
- The line of best fit or **prediction equation**. This is the equation that would be used to predict values of the dependent variable Y given values of the independent variable X. Note that it does a relatively good job of picking up much of the relationship between *MPG Highway* and *weight*.

- **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

Include
□ Prediction Limits
□ Confidence Limits

Confidence Level: 95.0 %

X-Axis Resolution: 101

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

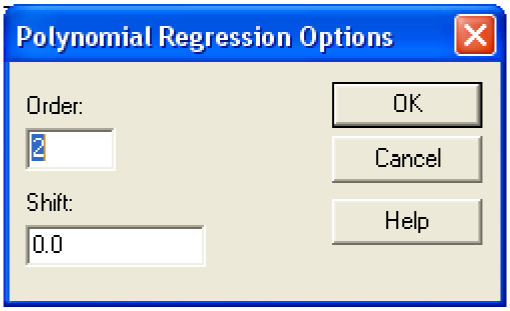
- **Include**: the limits to include on the plot.

- **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.

Analysis Options



Order: 2

Shift: 0.0

- **Order:** the order of the polynomial to be fit to the data.

- **Shift:** value to be subtracted from X before estimating the coefficients. When fitting high order polynomials, it may be necessary to specify an offset near the middle of the observed X data values to avoid numerical problems when fitting the model.

Example – Fitting a Third Order Polynomial

If a third order polynomial is fit to the data, the results are shown below:

Polynomial Regression - MPG Highway versus Weight

Dependent variable: MPG Highway
Independent variable: Weight
Order of polynomial = 3

		Standard	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
CONSTANT	114.476	31.385	3.64748	0.0004
Weight	-0.0660918	0.0329821	-2.00387	0.0481
Weight^2	0.0000175809	0.0000113068	1.55489	0.1235
Weight^3	-1.69022E-9	1.26487E-9	-1.33628	0.1849

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	1811.97	3	603.991	66.91	0.0000
Residual	803.337	89	9.02626		
Total (Corr.)	2615.31	92			

R-squared = 69.2833 percent
R-squared (adjusted for d.f.) = 68.2479 percent
Standard Error of Est. = 3.00437
Mean absolute error = 2.25416
Durbin-Watson statistic = 1.68521 (P=0.0615)
Lag 1 residual autocorrelation = 0.157148

The fitted model now includes X, X^2, and X^3. Note that the P-Value for *Weight*^{3} is well above 0.05, indicating that the third-order term is *not* statistically significant. This indicates that the second-order model was probably adequate for this data. Note: although the P-value for the second order term is not significant, it should not be assumed that a second-order model is unnecessary, since the P-value for *Weight*^{2} will change if *Weight*^{3} is removed from the model. To select a reasonable order for the polynomial, see the *Conditional Sums of Squares* pane described below.

Conditional Sums of Squares

The *Conditional Sums of Squares* pane displays a table showing the statistical significance of each coefficient in the model as it added to the fit:

Further ANOVA for Variables in the Order Fitted

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Weight	1718.7	1	1718.7	188.22	0.0000
Weight^2	77.1615	1	77.1615	8.45	0.0046
Weight^3	16.1176	1	16.1176	1.77	0.1875
Weight^4	7.94288	1	7.94288	0.87	0.3536
Weight^5	0.969712	1	0.969712	0.11	0.7453
Model	1820.89	5			

The table decomposes the model sum of squares SSR into contributions due to each coefficient by showing the increase in SSR as each term is added to the model. These sums of squares are often called *Type I sums of squares*. The F-Ratios compare the mean square for each term to the MSE of the highest order model, in this case a fifth order polynomial. In the above table, all terms beyond the second have P-values well in excess of 0.05, suggesting that a second-order model is sufficient for this data.

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	1795.86	2	897.928	98.62	0.0000
Residual	819.455	90	9.10506		
Lack-of-Fit	739.455	78	9.48019	1.42	0.2563
Pure Error	80.0	12	6.66667		
Total (Corr.)	2615.31	92			

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

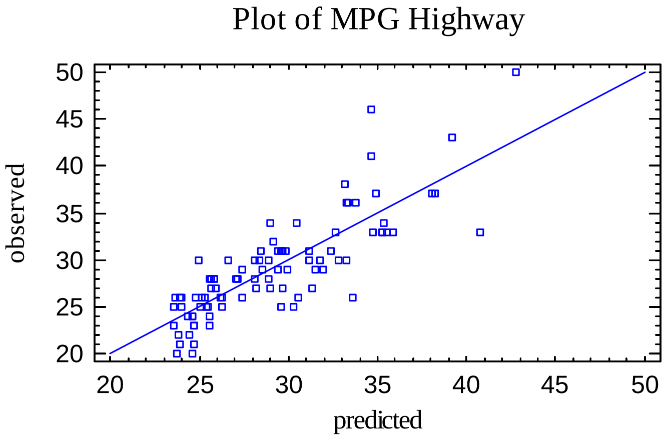
1. *Pure error:* variability of the Y values at the same value of X.
2. *Lack-of-fit:* variability of the average Y values around the fitted model.

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 lack-of-fit P-value is well above 0.05, indicates that the second-order polynomial adequately explains the relationship between *MPG Highway* and *weight*.

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 higher order polynomial. 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 increase somewhat as the predicted values get large.

Residual Plots

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

$$e_i = y_i - \hat{y}_i \qquad (3)$$

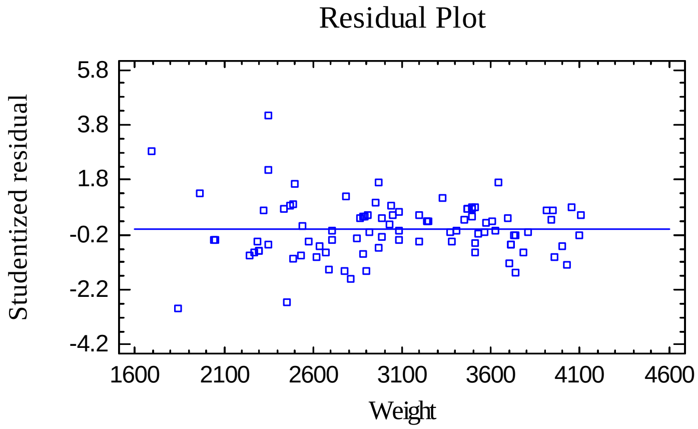
i.e., the residuals are the differences between the observed data values and the fitted model.

The *Polynomial Regression* procedure creates 3 residual plots:

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

Residuals versus X

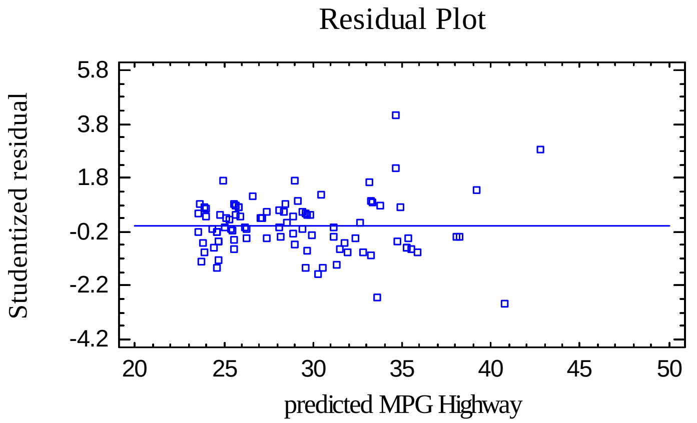
This plot is helpful in visualizing any need for a higher order polynomial.



No obvious curvature is detectable.

Residuals versus Predicted

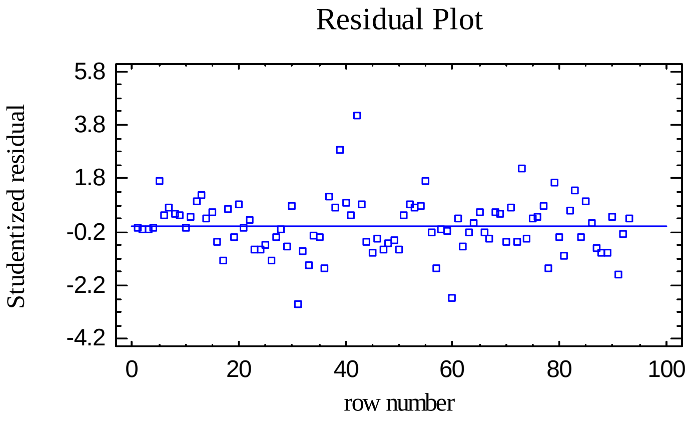
This plot is helpful in detecting any heteroscedasticity in the data.



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. In the plot above, some increased variability in miles per gallon can be seen at high predicted values, which corresponds to the smaller cars. For the smaller cars, the miles per gallon appears to vary more than for the larger cars.

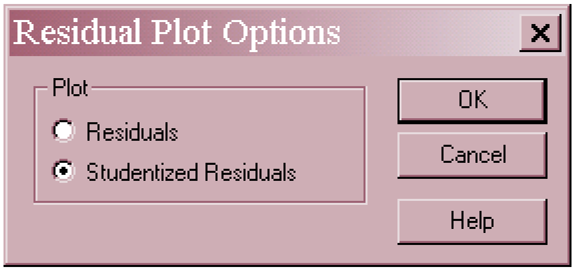
Residuals versus Observation

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, no obvious trend is present, although there is a standardized residual in excess of 4, indicating that it is more than 4 standard deviations from the fitted curve!

Pane Options


Residual Plot Options

Plot

○ Residuals

● Studentized Residuals

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.

Unusual Residuals

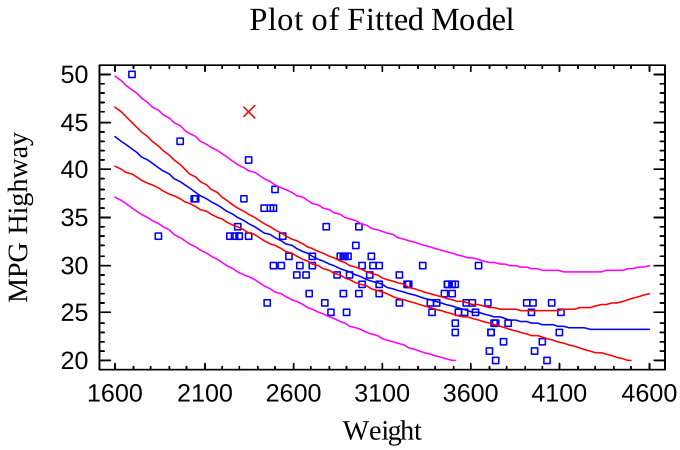
Once the model has been fit, it is useful to study the residuals to determine whether any outliers exist that should be removed from the data. The *Unusual Residuals* pane lists all observations that have Studentized residuals of 2.0 or greater in absolute value.

Unusual Residuals

		Predicted		*Studentized*
Row	*Y*	*Y*	*Residual*	*Residual*
31	33.0	40.7538	-7.75378	-2.90
39	50.0	42.8049	7.19515	2.84
42	46.0	34.6806	11.3194	4.13
60	26.0	33.6302	-7.63024	-2.64
73	41.0	34.6806	6.31936	2.17

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 #42 is more 4 standard deviations out. Row #42 is a Honda Civic, which was listed in the dataset as achieving 46 miles per gallon, while the model predicted less than 35.

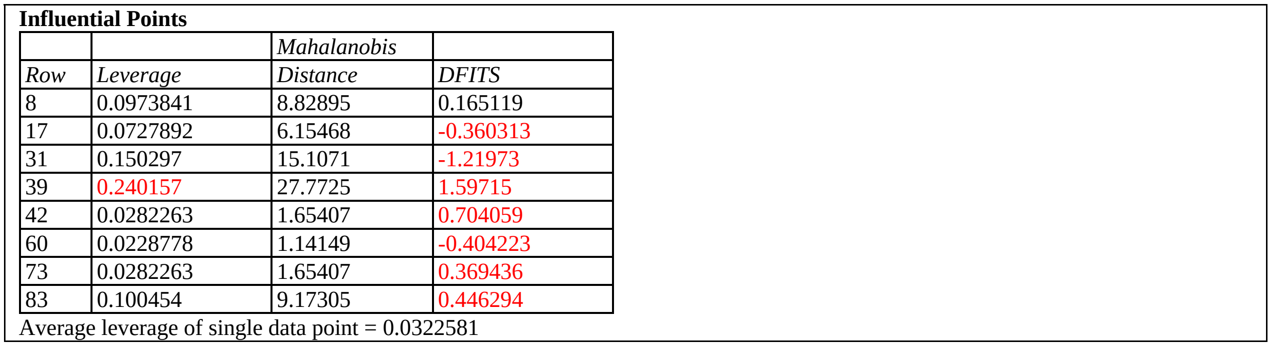
Points can be removed from the fit while examining the *Plot of the Fitted Model* by clicking on a point and then pressing the *Exclude/Include* button on the analysis toolbar:



Excluded points are marked with an X. For the sample data, removing row #42 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

		Mahalanobis	
Row	*Leverage*	*Distance*	*DFITS*
8	0.0973841	8.82895	0.165119
17	0.0727892	6.15468	-0.360313
31	0.150297	15.1071	-1.21973
39	0.240157	27.7725	1.59715
42	0.0282263	1.65407	0.704059
60	0.0228778	1.14149	-0.404223
73	0.0282263	1.65407	0.369436
83	0.100454	9.17305	0.446294

Average leverage of single data point = 0.0322581

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.

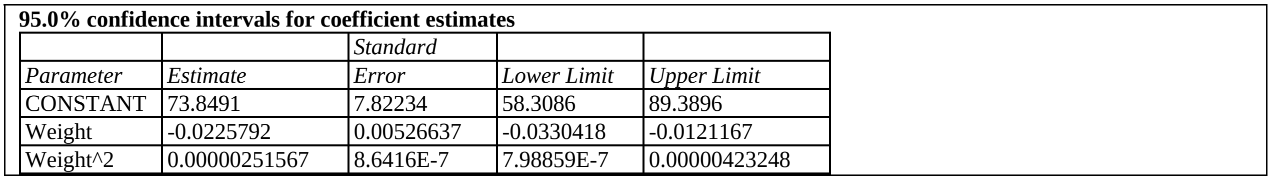
- **Mahalanobis Distance –** measures the distance of a point from the center of the collection of points in the multivariate space of the independent variables. Since this distance is related to *leverage*, it is not used to select points for the table.

- **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.

In the sample data, row #39 shows a leverage value of nearly 8 times that of an average data point. Row #39 is a Geo Metro, the lightest car in the dataset.

Confidence Intervals

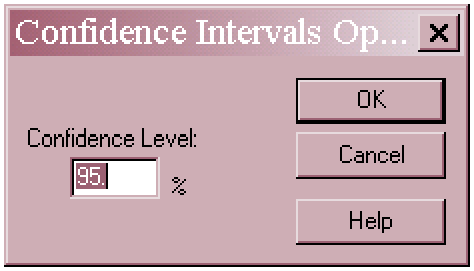
The *Confidence Intervals* pane shows the potential estimation error associated with each coefficient in the model.



95.0% confidence intervals for coefficient estimates

		Standard		
Parameter	*Estimate*	*Error*	*Lower Limit*	*Upper Limit*
CONSTANT	73.8491	7.82234	58.3086	89.3896
Weight	-0.0225792	0.00526637	-0.0330418	-0.0121167
Weight^2	0.00000251567	8.6416E-7	7.98859E-7	0.00000423248

Pane Options



Confidence Level: 95. %

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

Forecasts

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

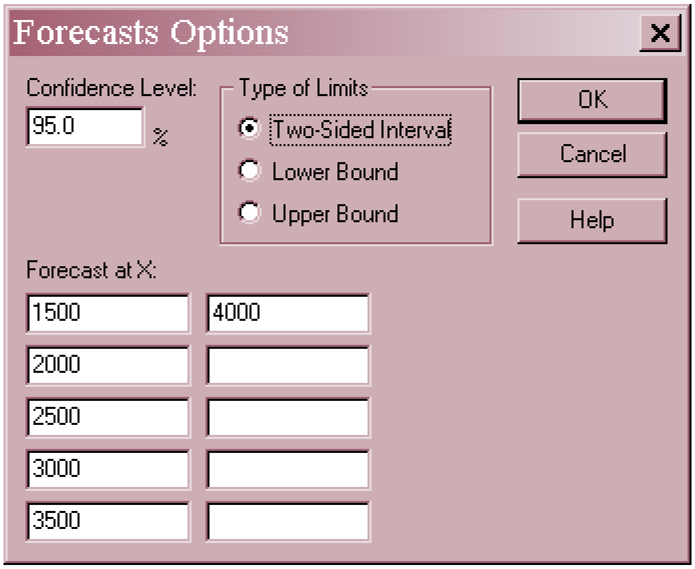
Predicted Values

		95.00%		*95.00%*	
	Predicted	*Prediction*	*Limits*	*Confidence*	*Limits*
X	*Y*	*Lower*	*Upper*	*Lower*	*Upper*
1500.0	45.6405	38.5073	52.7737	41.7745	49.5064
2000.0	38.7533	32.4975	45.009	36.9651	40.5415
2500.0	33.1239	27.0656	39.1822	32.2483	33.9995
3000.0	28.7524	22.696	34.8088	27.8903	29.6144
3500.0	25.6387	19.5909	31.6864	24.8397	26.4376
4000.0	23.7828	17.5924	29.9732	22.2385	25.3271

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).

Pane Options



Forecasts Options

Confidence Level: 95.0 %

Type of Limits

● Two-Sided Interval
○ Lower Bound
○ Upper Bound

Forecast at X:

1500	4000
2000	
2500	
3000	
3500	

- **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. *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.

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

The polynomial regression model is a special case of a multiple variable linear regression model. See the *Multiple Regression* documentation for details regarding the calculations.