

How To: Establish a Reasonable Shelf Life

Using STATGRAPHICS Centurion

by

Dr. Neil W. Polhemus

July 19, 2005

Introduction

Many products have a limited shelf life. As soon as they are produced, changes begin to occur. After some period of time, the product loses its effectiveness and must be pulled from sale or use. Establishing the length of time during which there is reasonable belief that the product will still be effective is an interesting statistical question.

Although there are different approaches to this problem, one common method is to take samples at different lengths of time after production and construct a statistical model for one or more critical variables. The model can then be used to predict that point in time after which the probability that the product will still be effective falls below some specified threshold. For example, we might want to pull the product when the probability of its being effective falls below 90%. This document will examine a data-based approach to the construction of a statistical model that could be used to determine a reasonable shelf life.

Sample Data

As an example, we will consider the following data, presented by Draper and Smith in Applied Regression Analysis, third edition (Wiley, 1998). It shows the measured amount of available chlorine in 44 samples of a product examined anywhere between 8 and 42 weeks after production. The data, which include multiple samples at certain weeks, are shown below:

Weeks since Produced	Available Chlorine
8	0.49, 0.49
10	0.48, 0.47, 0.48, 0.47
12	0.46, 0.46, 0.45, 0.43
14	0.45, 0.43, 0.43
16	0.44, 0.43, 0.43
18	0.46, 0.45
20	0.42, 0.42, 0.43
22	0.41, 0.41, 0.40
24	0.42, 0.40, 0.40
26	0.41, 0.40, 0.41
28	0.41, 0.40
30	0.40, 0.40, 0.38
32	0.41, 0.40
34	0.40
36	0.41, 0.38
38	0.40, 0.40
40	0.39
42	0.39


Figure 1: Sample Chlorine Data

It is quite clear that the amount of available chlorine drops as the product ages. Determining when to pull it off the shelf, however, requires a detailed analysis.

Step 1: Plot the Data

When beginning to analyze a new set of data, it is always a good idea to plot it. Much can be seen by eye that might not be caught by numerical statistics. In addition, visual inspection of the data brings the analyst into the data analysis process, which often avoids constructing models that make no sense.

Procedure: X-Y Scatterplot

The data in this case involve two quantitative variables, one which clearly depends on the other. To display the relationship between them, a simple *X-Y Scatterplot* will suffice. This plot is used so heavily in data analysis that it may be invoked by pressing the *X-Y Scatterplot* button  on the main toolbar. On the data input dialog box, indicate the variables to be plotted on each axis:

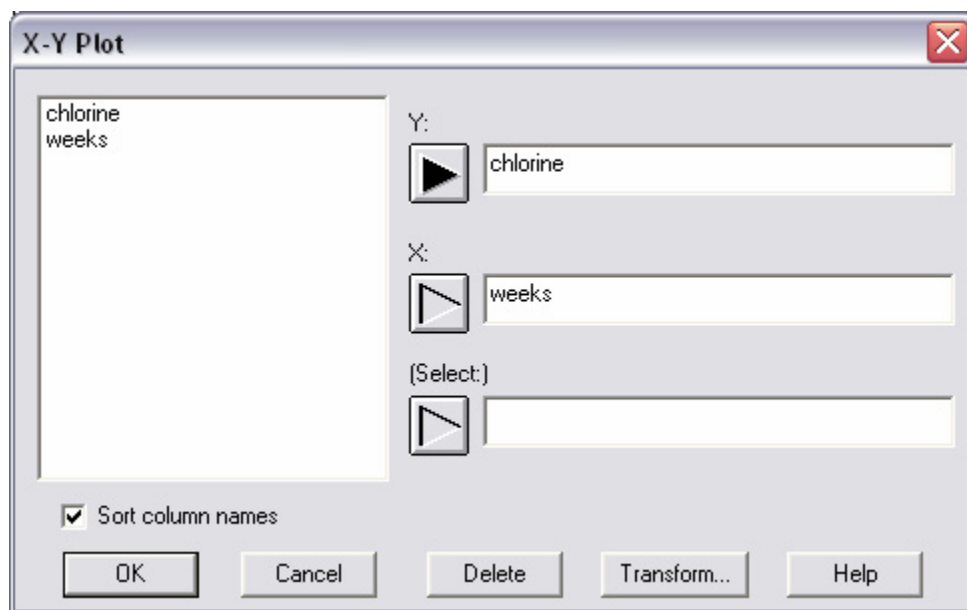


Figure 2: Data Input Dialog Box for X-Y Scatterplot

The resulting plot shows a strong negative correlation, as expected:

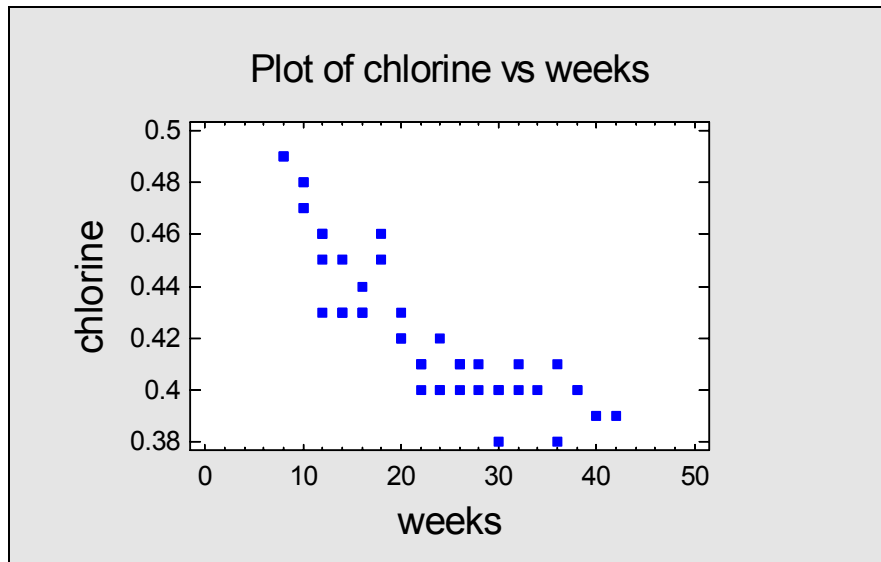



Figure 3: X-Y Scatterplot for Chlorine Data

To help visualize the form of relationship between the variables, double-click on the plot to enlarge it and then press the *Smooth/Rotate* button . On the subsequent dialog box, ask for a robust LOWESS smoother to be added to the chart:

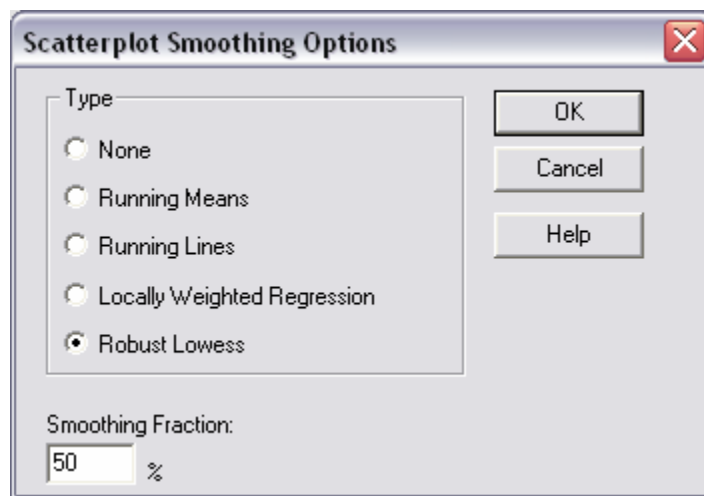


Figure 4: Smooth/Rotate Dialog Box

LOWESS stands for *Locally Weighted Scatterplot Smoothing* and is a technique that can be applied to any X-Y scatterplot to help visualize the relationship between the variables plotted on each axis. It is made “robust” or resistant to outliers by smoothing the data twice, down-weighting points far from the first smooth when the second smooth is performed. In this case, the smooth shows some obvious non-linearity, with the slope of the curve decreasing as the value of weeks increases:

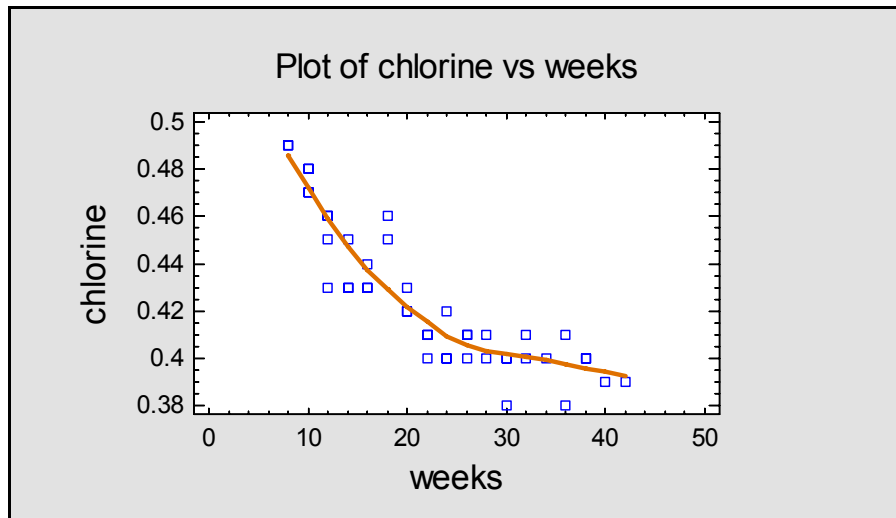


Figure 5: X-Y Scatterplot with LOWESS Smoother

A change in slope is not surprising. It simply means that the rate of loss of chlorine decreases as the amount of available chlorine decreases, a type of behavior often well-modeled by an exponential function.

Step 2: Fit a Curve to the Data

Having looked at the data and observed a curvilinear trend, we can now proceed to fit a model. Two types of models may be useful:

1. Nonlinear models such as exponentials, growth curves, and other types of functions.
2. Polynomial models involving powers of X.

Procedure: Simple Regression

The *Simple Regression* procedure is designed to fit many types of functions involving a single dependent variable Y and a single independent variable X. It is found in the following location on the main STATGRAPHICS Centurion menu:

- If using the Classic menu: *Relate – One Factor – Simple Regression*.
- If using the Six Sigma menu: *Improve – Regression Analysis – One Factor – Simple Regression*.

The data input dialog box is shown below:

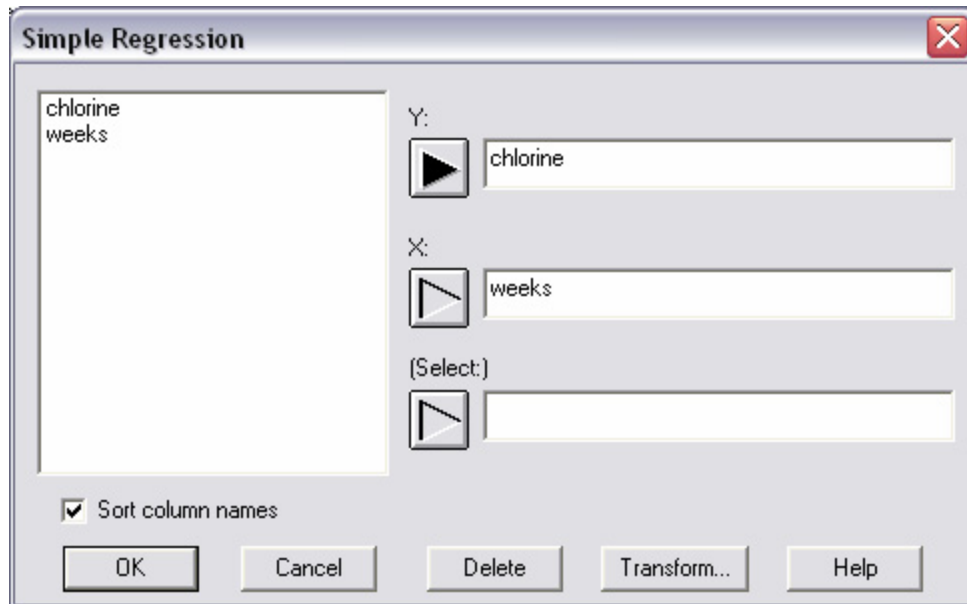


Figure 6: Data Input Dialog Box for Simple Regression

Initially, the procedure fits a straight line to the data, as displayed on the *Plot of Fitted Model*:

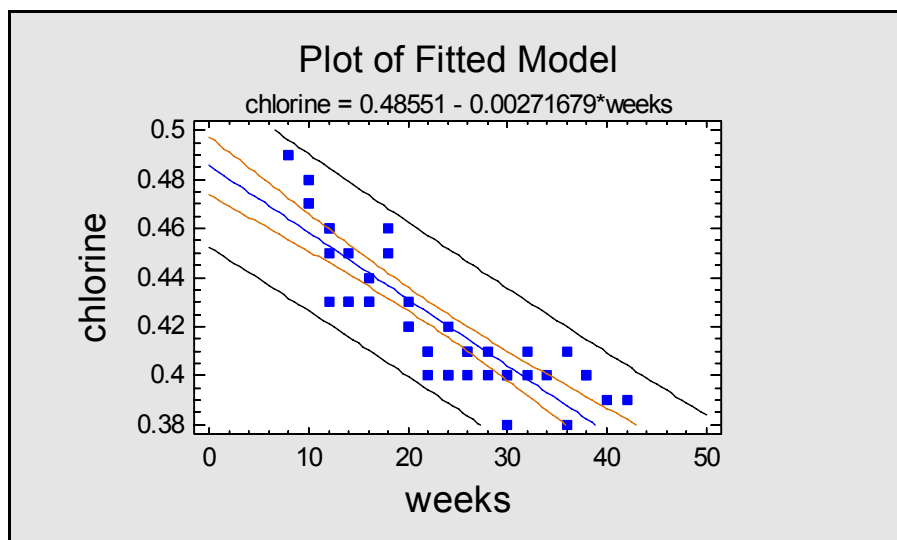



Figure 7: Plot of Fitted Linear Model

Included on the plot by default are:

- (1) The least squares regression line, which is the line for which the sum of the squared vertical distances from each point to the line is as small as possible.
- (2) 95% confidence intervals for the mean value of Y as a function of X. These are the inner bounds in the above plot.
- (3) 95% prediction intervals for new observations Y as a function of X. These are the outer bounds in the above plot.

Of particular interest are the outer prediction limits, which bound the distribution of the samples around the line. 95% of all samples are expected to fall within those bounds.


While the linear model describes some of the relationship between the variables, a nonlinear model is likely to do a much better job. If you press the Tabular Options button  on the analysis toolbar, you can select an option called *Comparison of Alternative Models*. This option fits a large number of nonlinear functions and displays them in decreasing order of R-squared:

Comparison of Alternative Models		
<i>Model</i>	<i>Correlation</i>	<i>R-Squared</i>
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 StatAdvisor
 This table shows the results of fitting several curvilinear models to the data. Of the models fitted, the squared-Y reciprocal-X model yields the highest R-Squared value with 87.7494%. This is 12.9174% higher than the currently selected linear model. To change models, select the Analysis Options dialog box.

Figure 8: Table of Alternative Models in Decreasing Order of R-Squared

R-squared is usually interpreted as the percentage of the variability in the response variable Y that has been explained by the model. In general, the higher R-squared is, the better the model. There is no guarantee that the model at the top of the list will be the best, however, especially when several other models give a similar statistic.

That said, it makes sense to try the model with the highest R-squared. This model can be fit by pressing the *Analysis Options* button  and selecting the desired model:

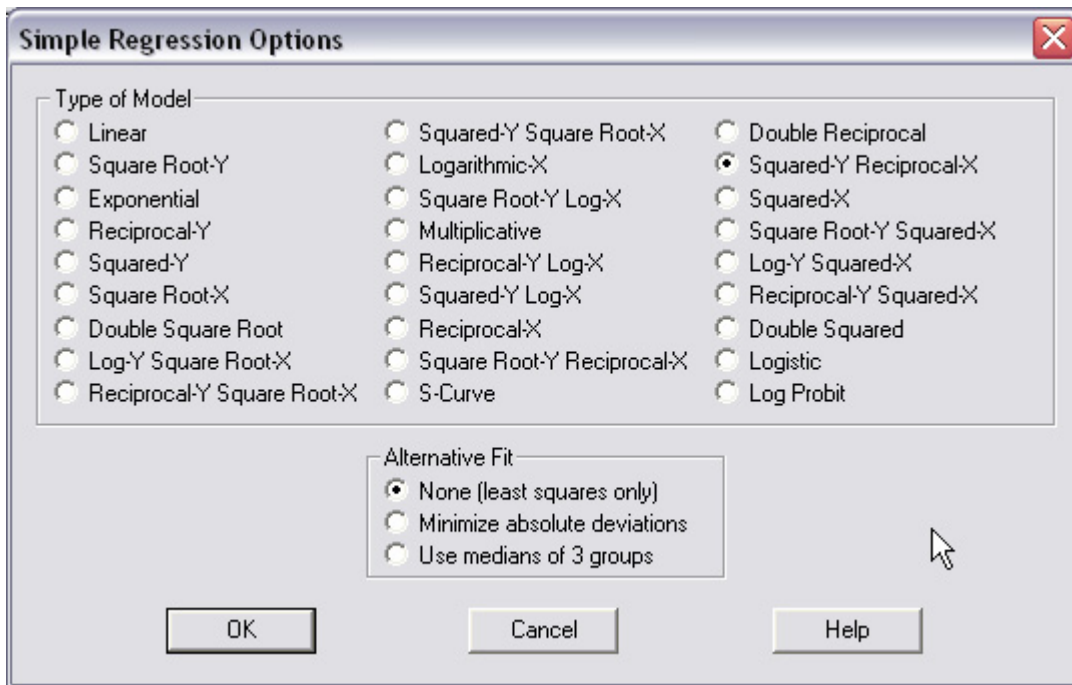


Figure 9: Simple Regression Analysis Options Dialog Box

When the OK button is pressed, all of the tables and graphs will automatically change to reflect the new function:

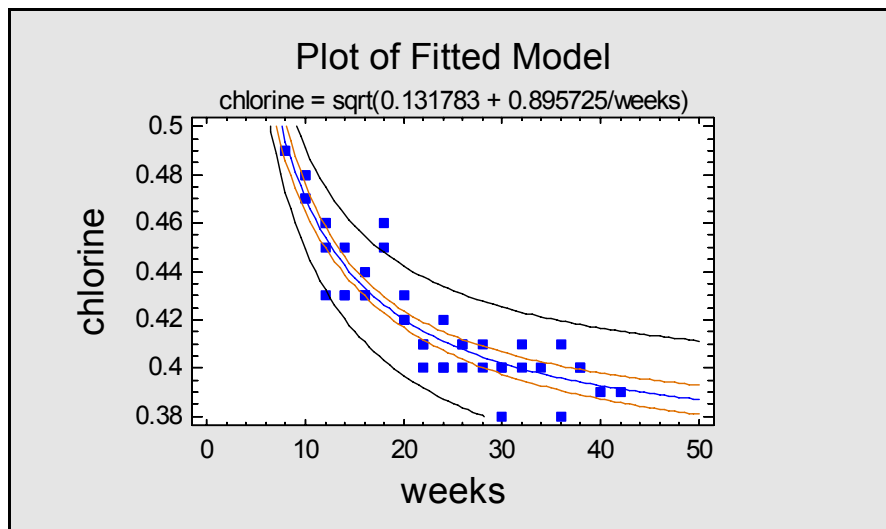


Figure 10: Fitted Nonlinear Model

The resulting fit matches the data much more closely than the linear model.

If we want to test the adequacy of the fit, we can do so by selecting *Lack-of-Fit Test* from the list of tabular options. This test is displayed using an analysis of variance table:

Analysis of Variance with Lack-of-Fit					
Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	0.0262486	1	0.0262486	300.84	0.0000
Residual	0.00366453	42	0.0000872506		
Lack-of-Fit	0.00197697	16	0.00012356	1.90	0.0700
Pure Error	0.00168756	26	0.0000649062		
Total (Corr.)	0.0299131	43			

The StatAdvisor
The lack of fit test is designed to determine whether the selected model is adequate to describe the observed data, or whether a more complicated model should be used. The test is performed by comparing the variability of the current model residuals to the variability between observations at replicate values of the independent variable X. Since the P-value for lack-of-fit in the ANOVA table is greater or equal to 0.05, the model appears to be adequate for the observed data at the 95% confidence level.

Figure 11: Lack-of-Fit Test for the Fitted Nonlinear Model

The lack-of-fit test, which can only be performed when there are replicate observations at the same values of X, compares the variability amongst the replicates to the variability around the fitted model. If the P-Value for lack-of-fit is less than 0.05, then the selected model does not adequately describe the observed relationship. In the above table, the model passes the lack-of-fit test, but just barely.

Step 3: Examine the Residuals

Part of the reason that the selected non-linear model barely passes the lack-of-fit test could be the presence of outliers. In the plot, 3 of the 44 observations used to fit the model lie outside the 95% prediction limits. To determine whether individual data values are significant outliers, select *Unusual Residuals* from the list of tabular options. This will display a table showing all residuals that are 2 or more standard deviations away from the fitted model:

Unusual Residuals					
Row	X	Y	Predicted Y	Residual	Studentized 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

The StatAdvisor
The table of unusual residuals lists all observations which have Studentized residuals greater than 2.0 in absolute value. Studentized residuals measure how many standard deviations each observed value of chlorine deviates from a model fitted using all of the data except that observation. In this case, there are 3 Studentized residuals greater than 2.0, one greater than 3.0. You should take a careful look at the observations greater than 3.0 to determine whether they are outliers which should be removed from the model and handled separately.

Figure 12: Table of Unusual Residuals

Two types of residuals are included in the table:

- (1) *ordinary residuals*, defined as the difference between each observed Y and the value of Y predicted by the model at corresponding value of X.
- (2) *Studentized residuals*, defined as the difference between the observed and predicted value of Y when the observation is not used to fit the model, divided by its standard error.

The latter residuals are of particular interest, especially any less than -3 or greater than +3. In this case, observation number 17 is 3.72 standard deviations from the fitted model, which is

extremely unusual if the deviations around the line follow a normal distribution. If an assignable cause could be found to explain why that observation is different than the rest, it would clearly be advantageous to remove it, since it would otherwise inflate the estimated variability of the observations around the fitted model.

Step 4: Fit a Resistant Model

To determine how sensitive the fitted model is to these possible outliers, we will try two approaches:

- (1) Remove the most extreme value and refit the model.
- (2) Fit the model using a procedure that is less sensitive than least squares to the presence of outliers.

To determine how sensitive the current fit is to the outliers, let's do the following:

- (1) Double-click on the plot of the fitted model to maximize it.
- (2) Press the alternate mouse button and select *Copy Pane to StatGallery*.
- (3) When the StatGallery appears, select *Edit – Paste* to place the graph in the upper left quadrant:

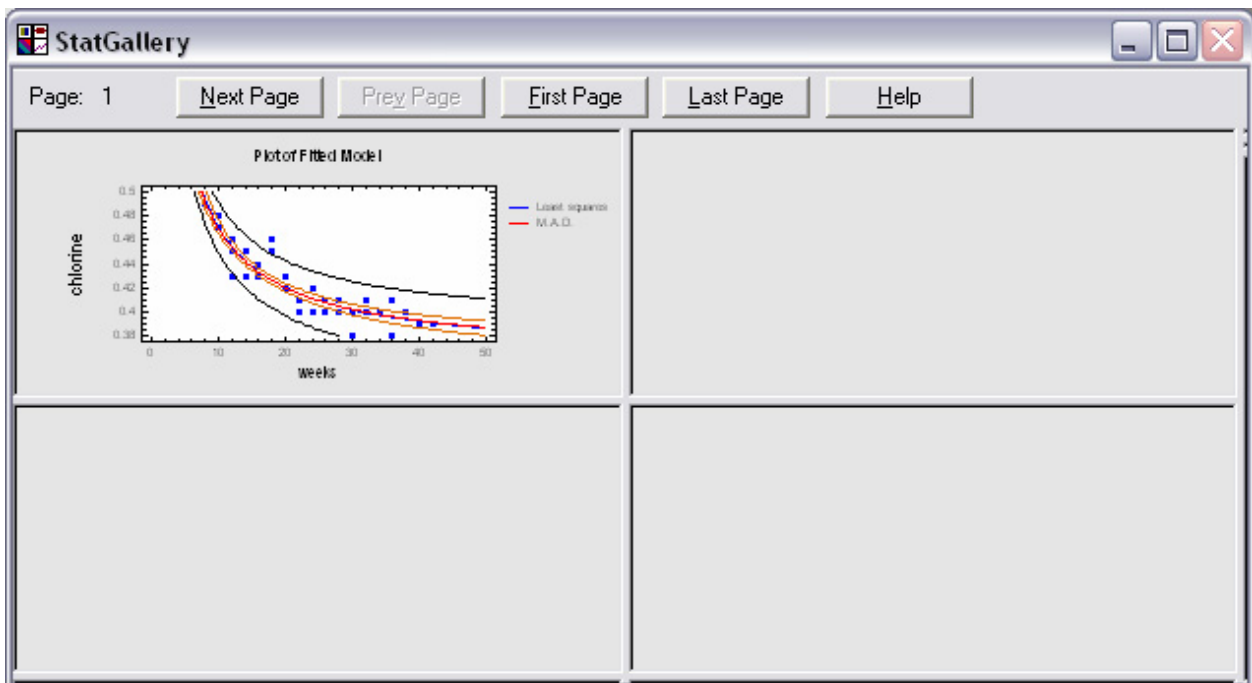



Figure 13: StatGallery after First Graph Has Been Pasted

- (4) Return to the *Simple Regression* plot. Click on the most extreme data value with the mouse and then press the *Exclude* button  on the analysis toolbar. The model will be automatically refit without that data value:

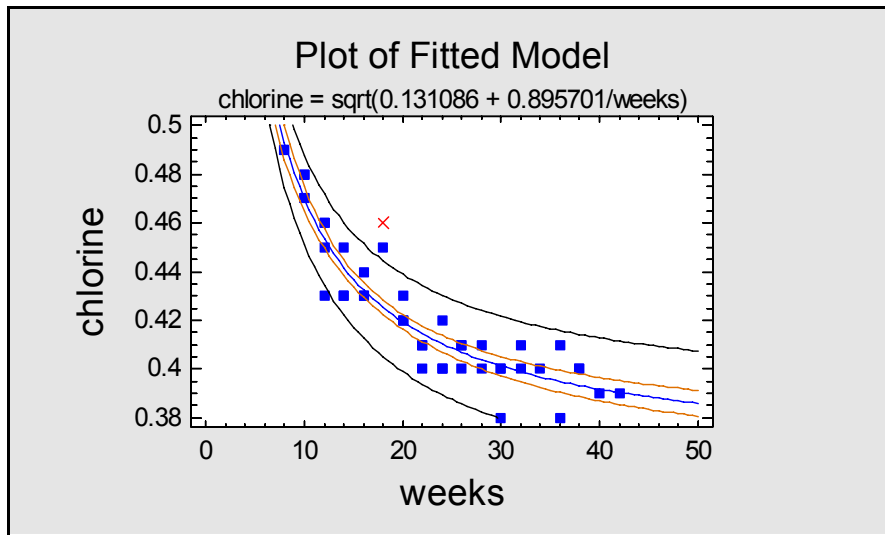


Figure 14: Fitted Nonlinear Model After Excluding Outlier

- (5) Press the alternate mouse button again and select *Copy Pane to StatGallery*.
- (6) When the StatGallery appears, attempt to paste the new graph in the same position as the first. When a dialog box appears, select *Overlay*:

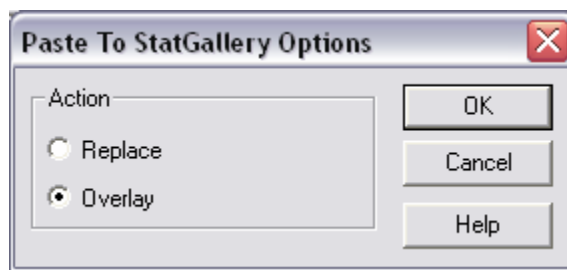


Figure 15: StatGallery Paste Options

The new graph will be pasted on top of the first.

- (7) Finally, double-click on the upper left quadrant of the StatGallery to maximize the plot:

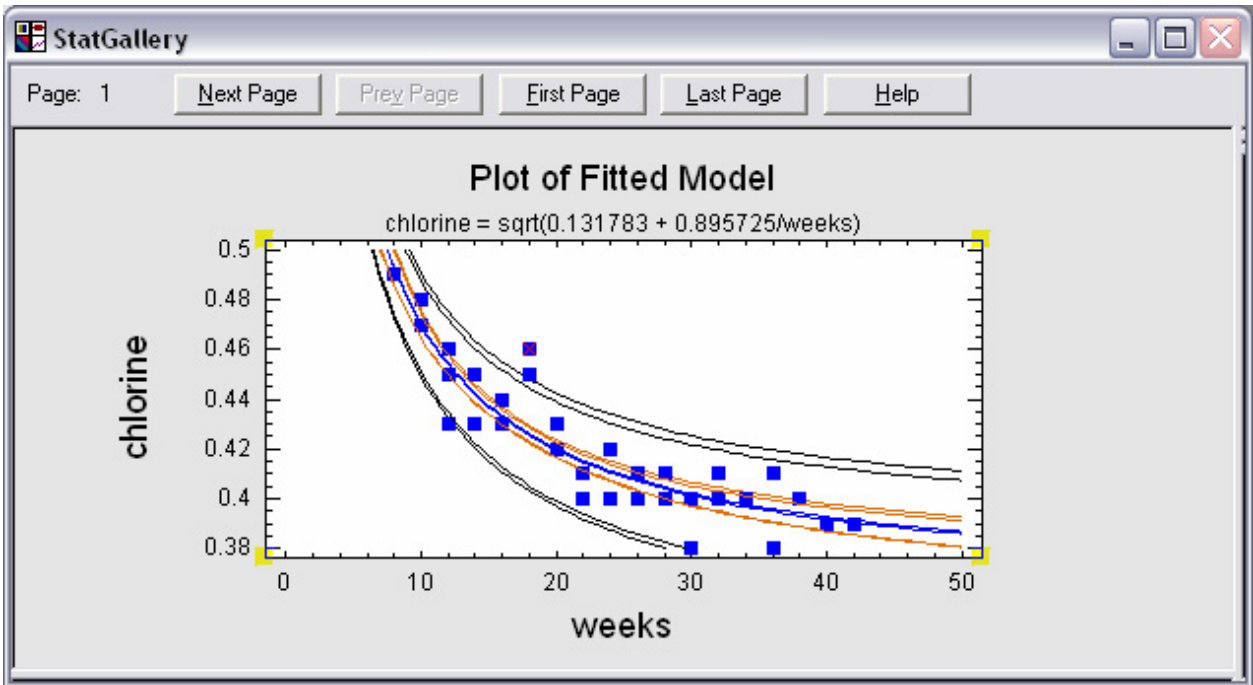


Figure 16: Fitted Models With and Without the Outlier

The best-fitting line has changed very little. There was also a very small change in the confidence and prediction limits, but not enough to make a major difference to this analysis. If you read the discussion of this data in Draper and Smith (1998), you will discover that they attempted to find an assignable cause for the outlier but, failing to do so, left it in, which is the conservative approach.

As mentioned, a second possibility for dealing with the potential outlier is to fit the model using a method that is less sensitive to outliers than is least squares. If you now return to the *Simple Regression* window and select *Analysis Options*, you will see two alternative methods for fitting the curve:

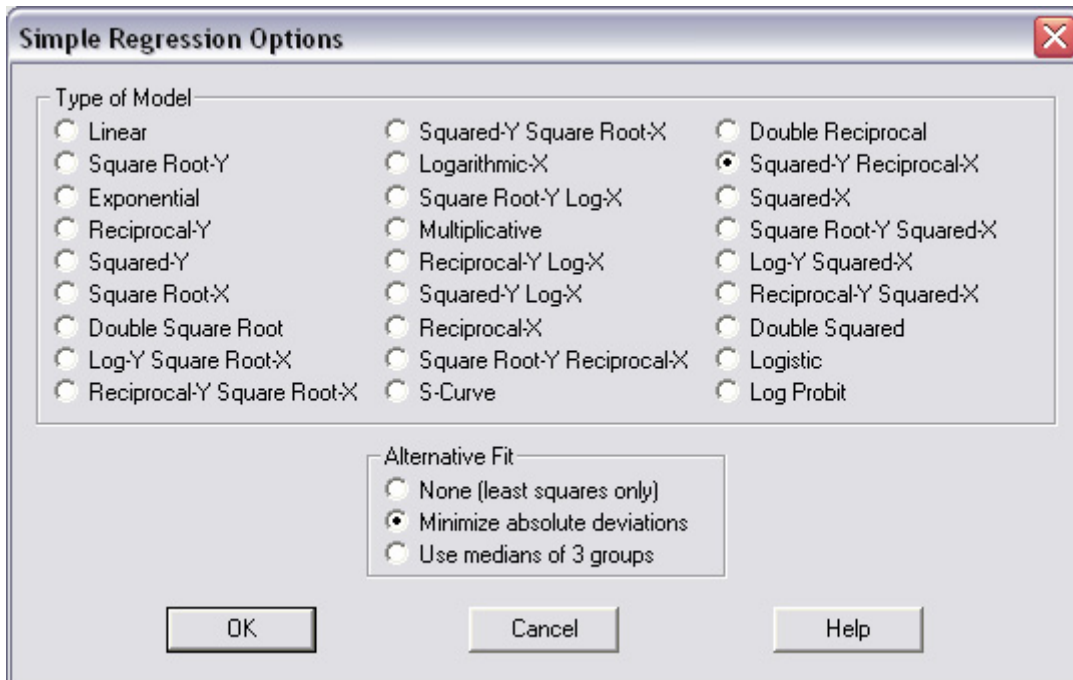


Figure 17: Selecting an Alternative Fit

The methods are:

- (1) *minimize absolute deviations* – fits the model so as to minimize the sum of the absolute values of the residuals rather than the sum of squares. Not squaring the residuals lessens the impact of the data values far away from the line
- (2) *use medians of 3 groups* – a method due to John Tukey that divides the data into 3 groups, locates a median position within each group, and then fits the model based on those 3 median positions.

Taking the first approach, after putting the outlier back into the data set, yields the following plot:

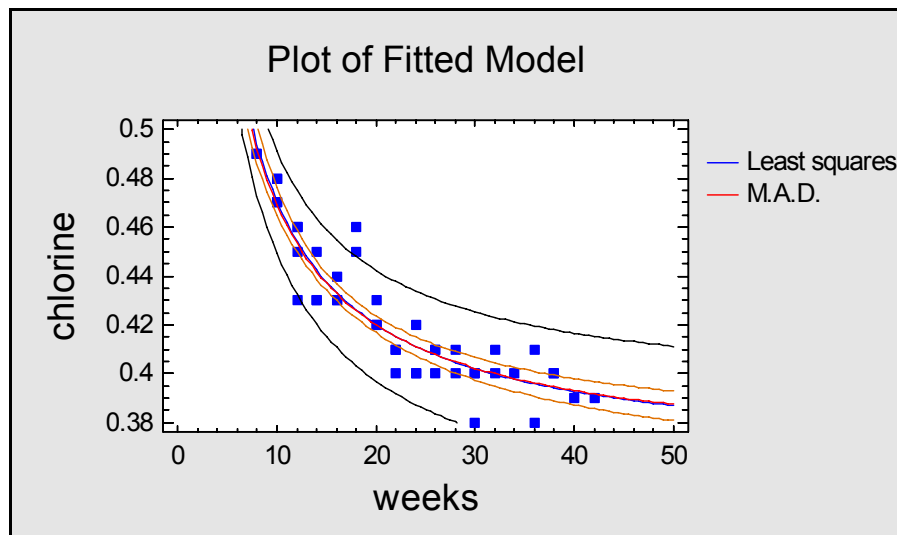


Figure 18: Alternative Minimum Absolute Deviation Fit

You can barely see the difference between the two models. There consequently seems to be no compelling reason not to use the nonlinear model as originally fit.

Step 5: Determine a Reasonable Shelf Life

Having modeled the data, we are now ready to select a shelf life for the product. It will be recalled that we wish to find the largest value of weeks for which 90% of the product is predicted to be at least 0.40. This can be done entirely graphically, by:

- (1) Displaying the *Plot of Fitted Model* in the *Simple Regression Procedure*.
- (2) Pressing the *Pane Options* button and requesting 80% prediction limits:

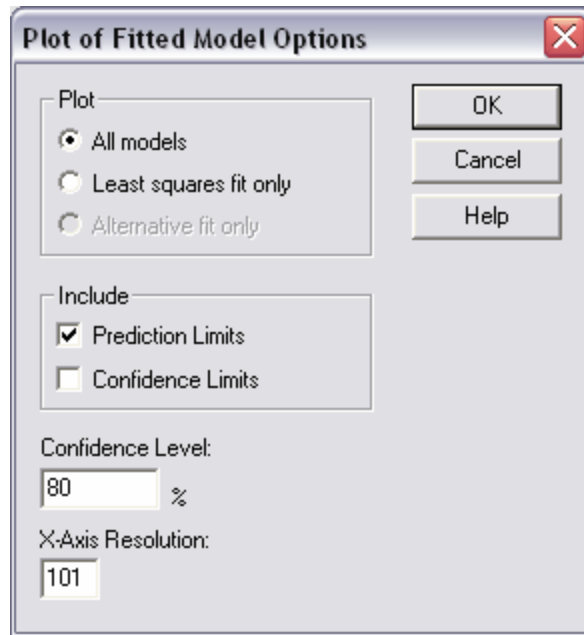


Figure 19: Changing the Pane Options

This creates the following plot:

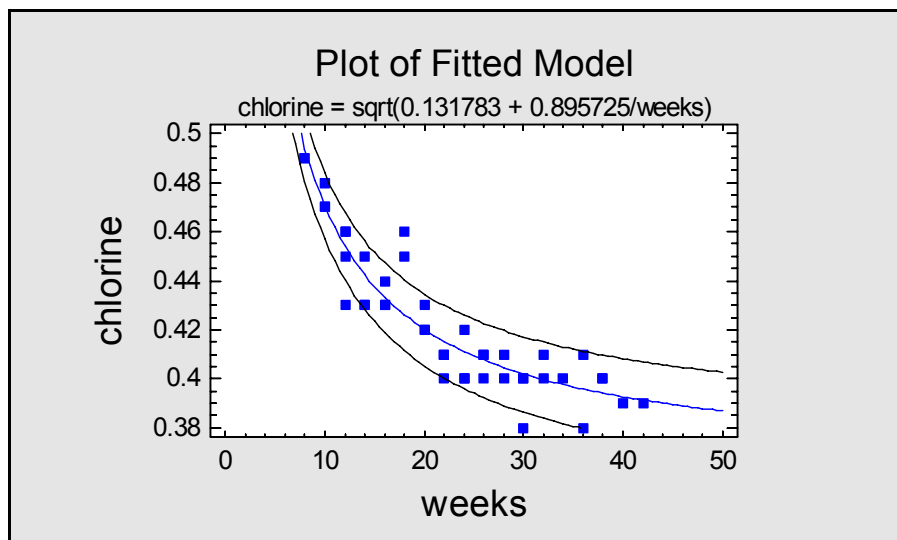


Figure 20: Fitted Model with 80% Prediction Limits

The reason for selecting 80% limits is that the lower bound will then be exceeded 90% of the time: 80% inside the bounds plus 10% above the upper bound.

(3) Press the alternate mouse button and select *Locate* from the popup menu. This displays a set of crosshair cursor that you can position at the intersection of 0.4 on the Y-axis and the lower prediction limits:

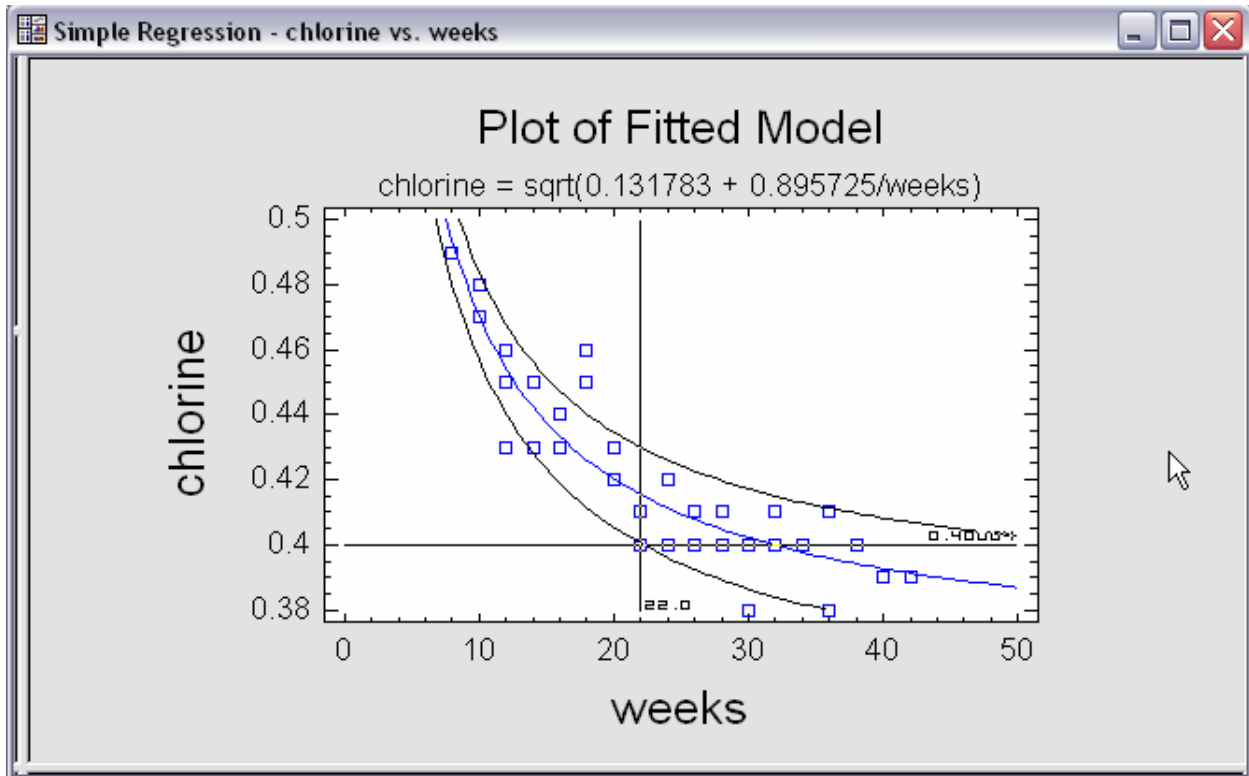


Figure 21: Using the Locate Option to Find an Intersection

The lower prediction limit drops below 0.40 at approximately 22 weeks, which is the value at which we should set the shelf life.

Step 6: Try Fitting a Polynomial Model as an Alternative

The nonlinear models fit by the *Simple Regression* procedure are usually sufficient to find a good model for the data. An alternative approach is to fit a polynomial model involving X , X^2 , X^3 , and other powers. STATGRAPHICS Centurion provides a special procedure for fitting such models:

- If using the Classic menu: *Relate – One Factor – Polynomial Regression*.
- If using the Six Sigma menu: *Improve – Regression Analysis – One Factor – Polynomial Regression*.

The data input dialog box is shown below:

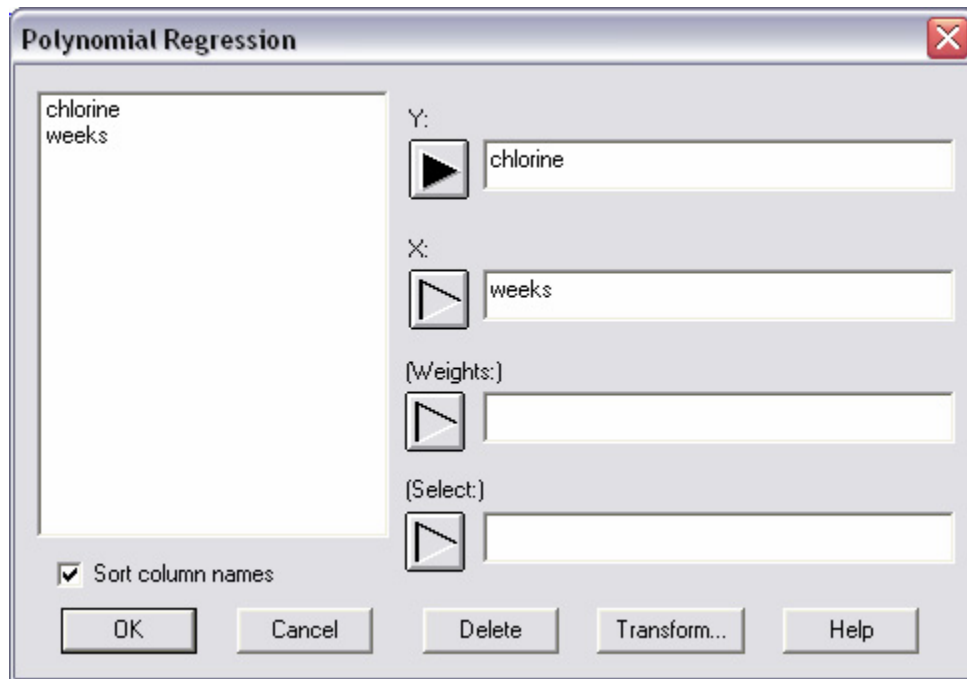


Figure 22: Data Input Dialog Box for Polynomial Regression

By default, the procedure fits a second-order polynomial:

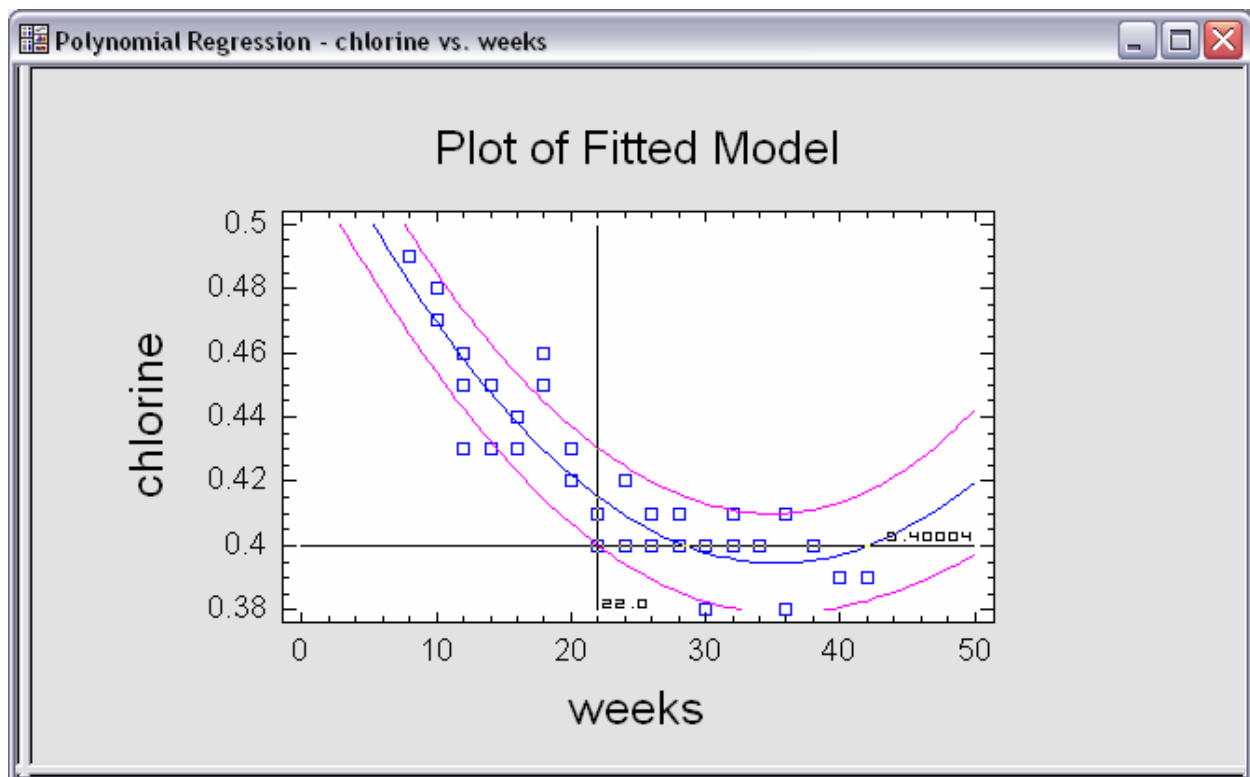


Figure 23: Fitted Second Order Polynomial with 80% Limits

which has the equation

$$\text{chlorine} = 0.540367 - 0.00826602 * \text{weeks} + 0.000117085 * \text{weeks}^2$$

The above model is usually good for a couple laughs, since it seems to imply that chlorine will begin to increase after some point in time, as is the nature of polynomials. In fact, for our purposes, which are interpolative, it gives practically the same answer as the earlier nonlinear model. Extrapolation, of course, would be unthinkable with this model, whereas the fitted nonlinear model

$$\text{chlorine} = \text{sqrt}(0.131783 + 0.895725/\text{weeks})$$

decays smoothly to an asymptotic value of 0.363.

Analysis Options allows us to use a higher order polynomial if desired:

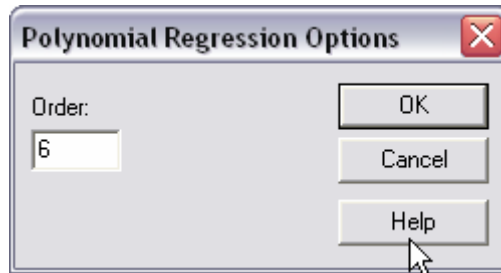


Figure 24: Polynomial Regression Analysis Options

We should, however, be very careful not to overmodel the data. Using a more complicated model than necessary is only likely to lead to less reliable predictions.

To determine the order of polynomial to use, specify a large number (such as 6). Then select *Conditional Sums of Squares* from the list of tabular options. This will display an analysis of variance table that shows the significance of each power as it is added to the model:

Further ANOVA for Variables in the Order Fitted					
Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
weeks	0.0295587	1	0.0295587	234.69	0.0000
weeks^2	0.00458676	1	0.00458676	36.42	0.0000
weeks^3	0.000323963	1	0.000323963	2.57	0.1173
weeks^4	0.0000100531	1	0.0000100531	0.08	0.7791
weeks^5	0.0002908	1	0.0002908	2.31	0.1371
weeks^6	0.0000696804	1	0.0000696804	0.55	0.4617
Model	0.0348399	6			

The StatAdvisor
 This table shows the statistical significance of each power of weeks as it was added to the polynomial regression model. The table can be used to help determine whether a lower-order polynomial than that currently fit to the data might be sufficient to describe the observed relationship between chlorine and weeks. Since the P-value corresponding to the term of order 2 is less than 0.05, a model of order 2 is suggested by this table at the 95% confidence level.

Figure 25: Conditional Sums of Squares Table

The largest order for which the P-Value is less than 0.05 would be the order to select. In this case, the second-order model is the most complicated that can be justified given the data.

Conclusion

Regression models are extremely useful in many applications. This How To guide describes a particular use in setting shelf life for a product. We saw that the *Simple Regression* procedure fits 27 different nonlinear functions and lists them in decreasing order of R-Squared. We also examined the impact of possible outliers on the fit and looked at methods to reduce their influence.

The type of analysis described here can only be done when the analyst is prepared to examine the output at each stage. No black box is likely to give a satisfactory answer to this type of problem. With the right tools, however, developing a good statistical model should not be difficult.

Note: The author welcomes comments about this guide. Please address your responses to neil@statgraphics.com.