Chapter 350

Introduction to Curve Fitting

Introduction

Historians attribute the phrase *regression analysis* to Sir Francis Galton (1822-1911), a British anthropologist and meteorologist, who used the term *regression* in an address that was published in *Nature* in 1885. Galton used the term while talking of his discovery that offspring of seeds "did not tend to resemble their parent seeds in size, but to be always more mediocre [i.e., more average] than they.... The experiments showed further that the mean filial regression towards mediocrity was directly proportional to the parental deviation from it."

The content of Galton's paper would probably be called *correlation analysis* today, a term which he also coined. However, the term *regression* soon was applied to situations other than Galton's and it has been used ever since.

Regression Analysis refers to the study of the relationship between a response (dependent) variable, Y, and one or more independent variables, the X's. When this relationship is reasonably approximated by a straight line, it is said to be *linear*, and we talk of linear regression. When the relationship follows a curve, we call it curvilinear regression.

Usually, you assume that the independent variables are measured exactly (without random error) while the dependent variable is measured with random error. Frequently, this assumption is not completely true, but when it cannot be justified, a much more complicated fitting procedure is required. However, if the size of the measurement error in an independent variable is small relative to the range of values of that variable, least squares regression analysis may be used with legitimacy.

Linear Regression Models

Perhaps the simplest example of a regression model is the familiar straight-line regression between two variables, X and Y, expressed by the formula:

(1) $\qquad Y = B_0 + B_1 X$

where B_0 and B_1 are called parameters, which are known constants linking Y and X. B_0 is the y-intercept, B_1 is the slope.

The relationship in (1) is exact. If you know X, you can determine Y exactly. Exact relationships are hard to find in applied science. Usually, you have to deal with empirical approximations determined from observed data. These relationships are represented as follows:

(2) $\qquad Y_i = B_0 + B_1 X_i + e_i$

where Y_i and X_i are the ith observed values of the dependent variable and the explanatory (regressor, predictor, or independent) variable, respectively. B_0 and B_1 are unknown parameter constants which must be estimated. The error term, e_i, represents the error at the ith data point. It is customary to assume that $E(e_i)=0$ (unbiased) and $V(e_i)=s^2$ (constant variance).

Actually, linear models include a broader range of models than those represented by equation (2). The main requirement is that the model is linear in the parameters (the B-coefficients). Other linear models are:

(3) $\ln(Y_i) = B_0 + B_1 \ln(X_i) + e_i$

and

(4) $Y_i = e^{B_0} + \sqrt{B_1}\, e^{X_i} + e_i$

At first, (4) appears nonlinear in the parameters. However, if you set $C_0 = e^{B_0}$, $\;C_1 = \sqrt{B_1}$, and $Z_i = e^{X_i}$ you will notice that it reduces to the form of (2). Models which may be reduced to linear models with suitable transformations are called intrinsically linear models. Model (5) is a second example of an intrinsically linear model.

(5) $Y_i = B_0 [\, e^{B_1 X_i} \,]\, e_i$

Notice that applying a logarithmic transformation to both sides of (5) results in the following:

(6) $\ln(Y_i) = \ln(B_0) + B_1 X_i + \ln(e_i)$

This is now easily recognized as an intrinsically linear model.

You should note that if the errors are normally distributed in (5), their logarithms in model (6) will not be so distributed. Likewise, if the errors, $\log(e_i)$, in (6) are normally distributed, the detransformed errors, e_i, in (5) will not be. Hence, when you are applying transformations to simplify models, you should check to see that the resulting error term has the desired properties. We will come back to this point later.

Nonlinear Regression Models

Nonlinear regression models are those which are not linear in the parameters to begin with nor can they be made so by transformation. A general representation for the nonlinear regression model is:

(7) $Y_i = f(X_i, e_i;\; B_1, B_2, ..., B_p)$

where B_1, B_2, ..., B_p are the p parameters to be estimated from your data, and e_i is the error term.

Note that e_i is not necessarily additive as in (2), although this is a common form. An example of an additive model is:

(8) $Y_i = B_0\, e^{B_1(X_i)} + e_i$

Linear models, such as those in (5), are preferred over nonlinear models, such as (8), for two main reasons. First, the linear model is mathematically easier to work with. Parameters may be estimated with explicit expressions. Nonlinear models must use iterative schemes, which may converge to several solutions. Second, often the investigator does not know the actual form of the relationship and is looking for an approximation. The linear model is an obvious place to start.

Least Squares Estimation of Nonlinear Models

The method of least squares minimizes the error sum of squares, Q, which is given by

(9) $Q = \displaystyle\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$

where $\hat{Y}_i = f(X_i; \hat{B}_1, \hat{B}_2, ...)$ is the value predicted for a specific X_i using the parameters estimated by least squares. If the errors are normally distributed, the least squares estimates are also the maximum likelihood estimates. This is one of the reasons we strive for normally distributed errors.

The values of the B's that minimize Q in (9) may be found either of two ways. First, if $f()$ is a simple function, such as in (2), you may find an analytic solution by differentiating Q with respect to B_1, B_2, ..., B_p, setting the resulting partial derivatives equal to zero, and solving the resulting p normal equations. Unfortunately, very few nonlinear models may be estimated this way.

The second method is to try different values for the parameters, calculating Q each time, and work towards the smallest Q possible. Three general procedures work toward a solution in this manner.

The *Gauss-Newton*, or *linearization,* method uses a Taylor series expansion to approximate the nonlinear model with linear terms. These may be used in a linear regression to come up with trial parameter estimates which may then be used to form new linear terms. The process iterates until a solution is reached.

The *steepest descent* method searches for the minimum Q value by iteratively determining the direction in which the parameter estimates should be changed. It is particularly useful when poor starting values are used.

The *Marquardt* algorithm uses the best features of both the Gauss-Newton and the steepest descent methods. This is the procedure that is implemented in this program. Note that *numerical derivatives* are used whenever derivatives are called for.

Starting Values

All iterative procedures require starting values for the parameters. This program finds the starting values for you. However, the values so found may fail to converge or you may be using a user-defined function which does not have preprogrammed starting values. Hence, you will have to supply your own starting values.

Unfortunately, there is no easy method for generating starting values for the B's in every case. However, we can provide you with some guidelines and a general method of attack that will work in many cases.

1. Try entering a 1 or 0 for each parameter and letting the program crank through a few iterations for you. You must be careful not to give impossible values (like taking the square root of a negative number), or the procedure will halt immediately. Even though the procedure may take longer to converge, the elapsed time will often be shorter than when using steps 2 and 3 below, since they require much more time and effort on your part.

2. Pick p observations that spread across the range of the independent variable and solve the model ignoring the error term. The resulting solution will often provide reasonable starting values. This includes transforming the model to a simpler form.

3. Consider the behavior of $f()$ as X approaches zero or infinity and substitute in appropriate observations that most closely approximate these conditions. This might be accomplished from a plot of your data or from an examination of the data directly. Once some of the parameters have been estimated in this manner, others may be found by applying step 2 above.

Inferences about Nonlinear Regression Parameters

The following results are from Seber (1989), chapter 5. They require the assumption that the errors are normally distributed with equal variance.

Confidence Intervals for Parameters

Let

(10) $Y_i = f(X_i; B_1, B_2,...) + e_i$ $(i = 1,2,...,n)$

represent the nonlinear model that we are interested in fitting. Let B represent the parameters B_1, B_2, ..., B_p. The asymptotic distribution of the estimates of B, which we call $\hat{B}$, is given by

(11) $\hat{B} \sim N_p(B, \sigma^2 C^{-1})$, $C = F.' F.$, $F. = [(\frac{\partial f}{\partial Bj})]$

For large n we have, approximately,

(12) $\hat{B}_r \pm t_{n-p}^{\alpha/2} s\sqrt{\hat{c}^{rr}}$

which gives approximate, large-sample 100(1-a)% confidence limits for the individual parameters. Note the *s* is an estimate of σ in (11), based on the residuals from the fit of (10).

These intervals are often referred to as the asymptotic-linearization confidence intervals because they are based on a local linearization of the function (10). If the curvature of (10) is sharp near $\hat{B}_i$, then the approximation will have considerable error and (12) will be unreliable.

Confidence Intervals for a Predicted Value

Using (10) - (12) it is easy to give approximate, asymptotic 100(1-a)% confidence intervals (or prediction intervals) for predicted values. These are:

(13) $\hat{Y}_0 \pm t_{n-p}^{\alpha/2} s[1 + f_{0'}(F.' F.)^{-1} f_0]^{1/2}$, $f_0 = (\frac{\partial f(X_0)}{\partial B_1}, \frac{\partial f(X_0)}{\partial B_2}, \dots)$

Note that f_0 and $F.$ must be estimated using the $\hat{B}$. Hence, if the fit of (10) is good and there is little curvature, these confidence intervals will be accurate. If the fit is poor or there is sharp curvature near the region of interest, these confidence limits may be unsatisfactory.

Parameterization

One of the first choices you must make is the way parameters are attached to the functional form of a model. For example, consider the following two models:

(14) $Y_i = \frac{B_0 X_i}{X_i + B_1} + e_i$

(15) $Y_i = \frac{X_i}{C_0 X_i + C_1} + e_i$

These are actually the same basic model. Note that if we let $C_0 = 1/B_0$ and $C_1 = B_1/B_0$, model (15) is simply a rearrangement of (14). However, the statistical properties of these two models are very different. Equations (14) and (15) are two parameterizations of the same basic model.

If there is no precedent for a particular model parameterization, then you should use that model with the best statistical properties. If this case, trial-and-error methods will have to be used to find a model. Often this will include comparing a plot of your data to a plot of the functional forms that are available, until a good match is found. If there are several models possible, a careful study of the error terms (residuals) is necessary to help in your selection.

A common misconception is the view that whether a parameter appears linearly or nonlinearly in the nonlinear model relates directly to its estimation behavior. This is just not the case. (See Ratkowsky (1989) section 2.5.2.)

Another common misconception is that a complicated model is superior to a simple model. In general, the simpler the model, the better the behavior of the estimation process. Adding an extra parameter has unpredictable results on the estimation process. In some cases, it has little effect, while in others it has disastrous consequences.

Overparameterization (using too complicated a model) often leads to convergence problems. These models may have multiple solutions. The estimates from these models are usually biased and nonnormally distributed. They

show high correlation among the parameter estimates. This problem may also occur when you use only a portion of a complicated function to fit a set of data. It is always better to find a simpler function that exhibits the functional behavior of your data. (See Ratkowsky (1989) section 2.5.4.)

The Stochastic Term e_i

A regression model such as (10) may be thought of as having a deterministic part $f(X;B1,B2,...)$ and a stochastic (random) part e_i. Often, assumptions about the e_i are necessary. The most common are:

1.   Independently distributed

2.   Identically distributed with constant variance

3.   Normally distributed

Independence

Independence means that the error at one value of i (say i=4) is not related to the error at another value of i (say i=5). Independence is often violated when data are taken over time and some carry-over effects are active.

Identicalness

Identicalness means that the distribution of the errors is the same for all values of i (for all data pairs X_i and Y_i). In practice, this assumption is equated with constant variance in the errors. If the variance of the e_i increases or decreases, then this assumption is violated.

Normality

The question of normality is very difficult to assess with small sample sizes (under 100). With large sample sizes, normal probability plots (discussed later) do a pretty good job. Least-squares methods (those used by this program) tend to create normality in the observed residuals even if the actual e_i's are not normal.

Some normality tests are available in the *Descriptive Statistics* module, so you can try them on your residuals. However, most technicians agree that if your observed residuals have a bell-shaped distribution with no outliers, the normality assumption is okay.

Summary

These assumptions are ideals that are only approximately met in practice. Least squares tends to be robust to minor departures from these assumptions. Only when there are major departures such as outliers, a large shift in the size of the variance, or a large serial correlation between successive residuals will estimates be significantly in error.

Interpretation of R-Squared

R-Squared is computed as

$$(16) \qquad R^2 = 1 - \frac{\sum_{1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{1}^{n}(Y_i - \overline{Y})^2}$$

That is, it measures the variance accounted for by the nonlinear model over and above that which is accounted for by the mean of Y. When the model does not contain an intercept-type term, this representation of R-Squared must be used carefully. You should also note that the predicted values, $\hat{Y}_i$, might be in the original (detransformed) metric or in a transformed metric. The program selects what we feel is the appropriate metric in each situation.

A common misconception is the view that R-Squared, the proportion of explained variation, is useful as a goodness-of-fit index in all nonlinear regression situations. Only when you have a linear model with a constant term does R-Squared represent the proportion of variation explained. (See Ratkowsky (1989) section 2.5.3.)

Transformed Y and Unequal Variance

One of the challenges of nonlinear modeling is selecting the appropriate form of the error term. For example, consider the two models (17) and (18) which differ only in the way the error term is represented.

(17) $Y = AX^B + e_1$

(18) $Y = CX^D e_2$

If you take the logs of both sides in (18) you will get

(19) $\ln(T) = \ln(C) + D\ln(X) + \ln(e_2)$

which is linear in the parameters and can be estimated using simple linear regression. Most of us would rather fit (19) with linear regression than fit (17) with a nonlinear least-squares algorithm. Does it matter? Of course it does.

The difference lies in the pattern of the residuals, the e's. If the true relationship is (17) and you fit (19), you will see a strange pattern in the plot of the residuals. They will exhibit nonconstant variance. Again, if the true relationship is (18), then using (17) will result in an improper model.

The point is, the pattern of the residuals, not convenience, dictates the form of the error term. Hence, you should not use (18) and (19) on a curve with constant variance. Instead, you should use (17). Similarly, if the variance is increasing, a variance-stabilizing transformation of Y (like the log) will be useful in making the variance constant across all values of X.

In summary, there are three reasons for transforming Y: firstly, to obtain linearity; secondly, to obtain errors that are normally distributed; and thirdly to obtain a constant error variance. An examination of the residuals both from the fits before and after transformation is the only way to assess which model is appropriate

Reducing Transformation Bias in Curve Fitting

This material is taken from Miller (1984). If possible, read his article.

One of the most common ways of modeling a nonlinear relationship between two variables is to find a transformation for either the dependent or independent variables (or both) that results in a linear relationship. This relationship may then be estimated with standard linear regression methods. The residuals, calculated from the fit after the transformations, are studied to see that the various assumptions hold. This is taught in many courses and textbooks.

In many situations, the real interest lies in the relationship between the variables in the original metric. When the dependent variable has been transformed, a reverse transformation is used to return a transformed predicted value to the original metric. Prediction intervals in the fitted (transformed) metric are detransformed to arrive at corresponding prediction intervals in the original metric. These detransformed predicted values and prediction intervals give values for the median response, not the mean response as is often supposed.

When the estimated mean response is sought and the above methods are used, the resulting estimates are severely biased. This program provides bias correction factors that may be applied when an estimate of the mean response is desired.

Without going into the details of how and why this biasing occurs, we present the following correction procedures that may be used to correct for this bias. Remember, if the median response is okay then these correction factors do not have to be applied. Note that $\hat{\sigma}^2$ is the mean square error from the transformed model. Also $\hat{Y}$ refers to the detransformed predicted value of Y.

The following table shows the dependent variable transformation and the bias correction factor used.

Transformation	**Bias Correction Formula**
Ln(Y)	$\hat{Y}\left(exp\left(\dfrac{\hat{\sigma}^2}{2}\right)\right)$
Sqrt(Y)	$\hat{Y} + \hat{\sigma}^2$
1/Y	$\hat{Y}(1 + \hat{Y}^2\hat{\sigma}^2)$

Further Reading

This has been a brief introduction to curve fitting. If you want to get into the issues of variable transformations more deeply, we suggest that you begin with Box and Draper (1987), chapters 7 and 8.

If you want to see examples of fitting curves to data, we suggest Draper and Smith (1981), Hastings (1957), Davis (1962), and Ezekiel and Fox (1967). The first of these is a modern account of nonlinear regression, which goes through several examples. The last three books were written before the computer revolution when the emphasis was on hand calculation. Even though the calculation methods are out of date in these books, they work many examples and provide a great deal of insight into the art of curve fitting.

Curve Fit Plot Format Window Options

This section describes the specific options available on the Curve Fit Plot Format window, which is displayed when the Curve Fit Plot Format button is clicked. Common options, such as axes, labels, legends, and titles are documented in the Graphics Components chapter.

Curve Fit Plot Tab

Symbols Section

You can modify the shape, color, and size of the plot symbols. To change them, click the Symbol Format button to display the Symbol Format window. Here are some of the graphics effects that can easily be achieved.
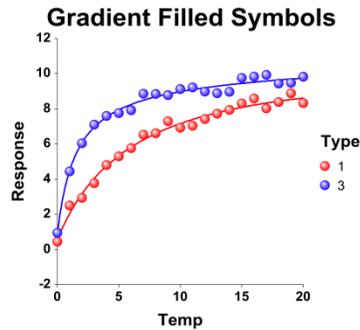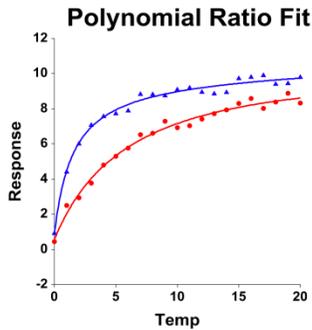
Regression Section

You can add the regression curve, residuals, and prediction limits to the plot using the options available in this section.
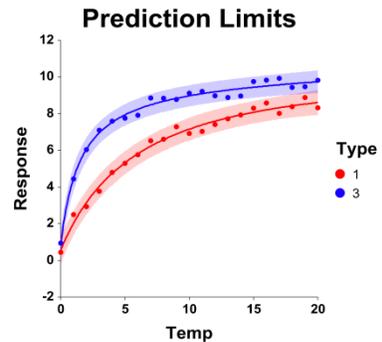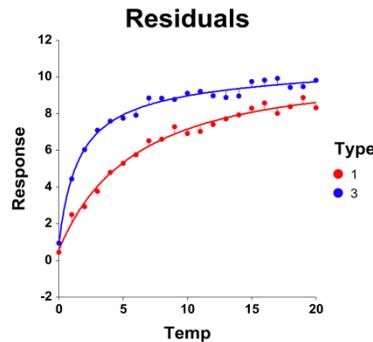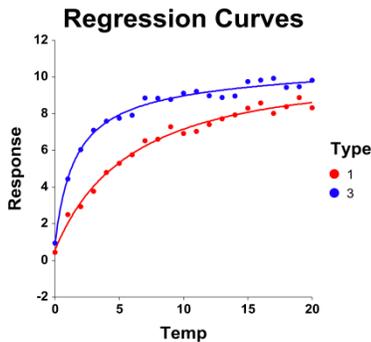

Estimation Section

With these options, you can set the resolution of any curved lines (such as prediction limits and regression fits).

More Lines Tab

Loess Smooth Section

You can add a loess smooth line as a comparison line using the options available in this section. The loess smooth line is documented in the Linear Regression and Correlation chapter.
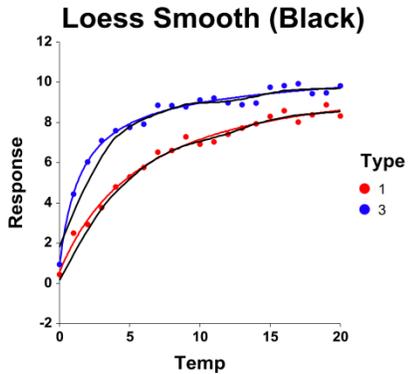
Connect Data Points Section

You can connect the symbols with a line. Along with setting the format of the connecting line, this section also lets you specify the order in which the points are connected. The connection can be done with a straight line or a curved spline.
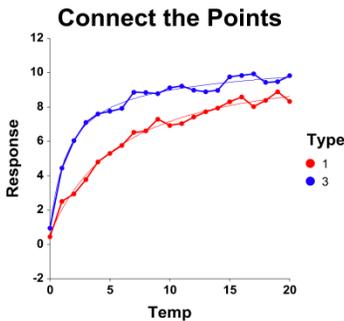

Points-to-Axis Tab

Tick Marks for each Data Point Section

You can display a tick mark along each axis for each point.

Lines from each Data Point Section

You can display a line between the data point and an axis.

Data Lines
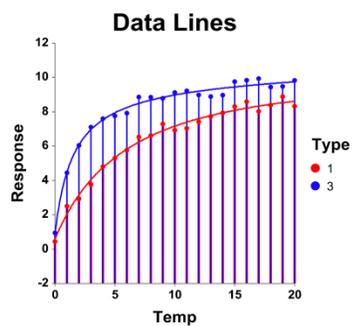
Bars from each Data Point Section

You can display bars between data points and an axis.

Titles, Legend, X Axis, Y Axis, Grid Lines, and Background Tabs

Details on setting the options in these tabs are given in the Graphics Components chapter.