Fitting curves to data using nonlinear regression: a practical and nonmathematical review

HARVEY J. MOTULSKY AND LENNART A. RANSNAS
Department of Pharmacology, University of California, San Diego, La Jolla, California 92093, USA

ABSTRACT

Many types of data are best analyzed by fitting a curve using nonlinear regression, and computer programs that perform these calculations are readily available. Like every scientific technique, however, a nonlinear regression program can produce misleading results when used inappropriately. This article reviews the use of nonlinear regression in a practical and nonmathematical manner to answer the following questions: Why is nonlinear regression superior to linear regression of transformed data? How does nonlinear regression differ from polynomial regression and cubic spline? How do nonlinear regression programs work? What choices must an investigator make before performing nonlinear regression? What do the final results mean? How can two sets of data or two fits to one set of data be compared? What problems can cause the results to be wrong? This review is designed to demystify nonlinear regression so that both its power and its limitations will be appreciated. — MOTULSKY, H. J.; RANSNAS, L. A. Fitting curves to data using nonlinear regression: a practical and nonmathematical review. *FASEB J.* 1: 365-374; 1987.

Key Words: nonlinear regression • computer programs • curve fitting • data analysis

MANY TYPES OF DATA ARE BEST ANALYZED by fitting a curve using nonlinear regression, and computer programs that can perform these calculations are readily available. However, it is difficult to learn about nonlinear regression because it is a topic virtually ignored by most statistical textbooks, and because many articles on the subject assume an advanced level of mathematical background. We therefore prepared this review to explain simply the theory, use, and pitfalls of nonlinear regression. Our goal is to present a practical approach to the problem, and we largely cite review articles. More mathematical details, as well as citations to the primary literature, can be found in several reviews (see refs 1-5).

TYPES OF CURVE FITTING

Nonlinear regression is a powerful tool for fitting data to an equation to determine the values of one or more parameters. Before discussing nonlinear regression, however, we will first review the other methods used for fitting curves to data.

Linear regression of transformed data

Linear regression is familiar to all scientists. Data are graphed so that the x axis represents the independent variable and the y axis represents the dependent variable. The line drawn by the linear regression procedure is chosen to minimize the sum of the squares of the vertical distances of the points from that line.

An easy method for dealing with curved relationships is to transform the data into a straight line and then perform linear regression. An example is shown in Fig. 1. The data follow an exponential decay ($y = Ae^{-Bx}$). By taking the logarithm of each y value, the data points form a straight line [$\ln(y) = \ln(A) - BX$]. Other common transformations are used to convert saturation radioligand-binding isotherms into Scatchard plots, and enzyme kinetic data into Lineweaver-Burke plots. Linear regression of the transformed data yields a slope and intercept that can be used to determine the parameters of interest.

Performing linear regression on transformed data has several advantages: It is intuitively straightforward; it does not require a computer; and the result may seem easy to evaluate statistically. However, the results are not statistically optimal. Linear regression calculations are valid only when the experimental uncertainty of replicate y values is not related to the values of x or y. This assumption is usually not valid after data have been transformed. For example, Fig. 1 shows an exponential decay curve, and a linear plot constructed by taking the logarithm of all y values. However, this transformation enhanced the errors associated with the points with a small y value. Thus those points will be emphasized by linear regression, and the points with a low x value will be relatively ignored. Accordingly, the resulting slope will not be the best estimate of the rate constant. Linearizing transformations are not optimal because they distort the experimental errors. This distortion is especially severe in transformations that combine x and y values, for example, Scatchard plots used to analyze radioligand-binding data.

Polynomial regression

Polynomial regression is used to fit data points to the following equation:

0892-6638/87/0001-0365/$01.50.

$$Y = A + Bx + Cx^2 + Dx^3 + Ex^4 \cdots$$

The goal of polynomial regression is to determine values for the parameters (A, B, C,. . .) that make the curve best fit the data points. When using a polynomial regression program, you must specify the order of the equation—the number of parameters to be fit. When only A and B are to be fit, the equation describes a line, and polynomial regression is identical to linear regression. With three or more parameters, the equation describes a curve; more parameters create a more flexible curve. In this equation, y is linear with (proportional to) each of the parameters (when x and the other parameters are held constant). Accordingly, a unique solution can be obtained without the use of the iterative procedures described below. The methods used for fitting polynomial equations are extensions of linear regression.

Polynomial regression is often used to create a generic curve through the data points; in such cases the mathematical form of the equation is irrelevant. See, for example, Fig. 2. Polynomial regression is not frequently used for data analysis in biology.

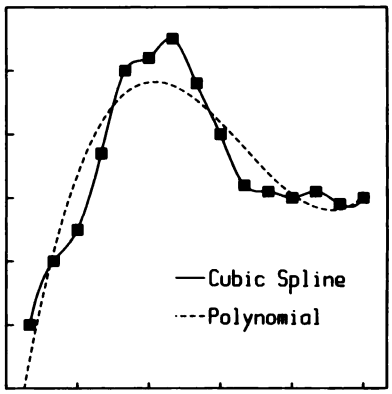

Figure 2. Cubic spline vs. cubic polynomial. The solid curve fit via a cubic spline procedure goes through every data point. The dashed curve was determined by fitting the data to the cubic polynomial equation $y = A + Bx^2 + Cy^3$. Depending on the circumstances, both of these procedures are useful for drawing generic curves through data points. These procedures should be considered the electronic equivalent of a French curve or a flexible ruler; they are not often useful for data analysis.

Cubic spline

A cubic spline is a curve that goes through every point. Cubic spline is useful for plotting a smooth curve through a set of data points (Fig. 2). This technique may also be used for interpolating between data points on a standard curve. Unlike nonlinear regression, cubic spline is not a tool useful for analyzing data. Different cubic spline algorithms may find slightly different curves, especially near the first and last data points.

The idea behind a cubic spline curve is straightforward. Although a pair of data points can be fit only to a single straight line, they can be fit to many cubic (third-degree polynomial) curves. The cubic spline technique finds a different cubic equation for every pair of adjacent points, and selects these equations so that the overall curve is smooth. It does this by ensuring that the first and second derivatives of each curve segment match those of the adjacent segments.

Note the distinction between fitting data to a cubic polynomial equation and using the cubic spline method. *Equation 1* is a cubic equation when E and all variables beyond are set to zero. A polynomial fit with a cubic equation fits one equation to all the data. A cubic spline, in contrast, uses a different cubic equation for each adjacent pair of data points. The cubic spline curve goes through each data point; a polynomial fit usually does not.

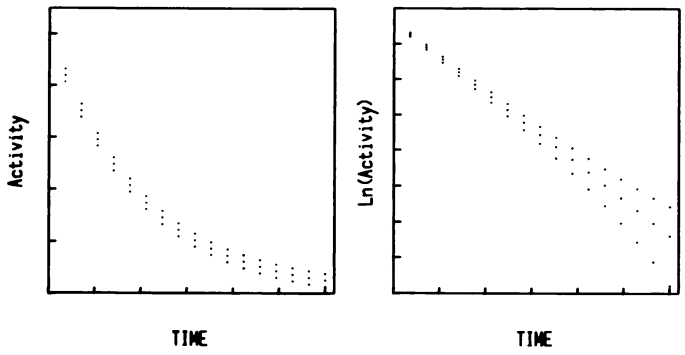

Figure 1. Curve fitting by linear transformation. The left panel shows computer generated data points after an exponential dissociation curve. The replicate values are equally spaced, representing experimental uncertainty that is not related to the values of x or y. The right panel shows the same data after each y value is converted to its logarithm. After this transformation, the relationship is linear and, accordingly, linear regression can be used to analyze the data. Note, however, that the experimental uncertainty of the points now is related to the values of x and y. Linear regression on the transformed data would therefore overemphasize the points with large x values.

THEORY OF NONLINEAR REGRESSION

Nomenclature

We confine our discussion to the common situation in which there is a single *independent* variable x, and a single *dependent* variable y. The independent variable is controlled by the experimenter; the dependent variable is measured. The relationship between the variables x and y can be described by an equation that includes one or more parameters, which we call A, B, C, etc.

Goal

Nonlinear regression is a procedure for fitting data to any selected equation. As with linear regression, nonlinear regression procedures determine values of the parameters that minimize the sum of the squares of the distances of the data points to the curve. If the y value of each data point is called y_{data} and the y value of the curve is called y_{curve}, the goal is to minimize the residual sum of squares (SS):

$$SS = \text{sum}[(y_{data} - y_{curve})^2]$$

Because this criterion minimizes the sum of the square of the distances, it is called a *least-squares method;* such methods are appropriate when the experimental uncertainty is Gaussian (normally distributed), and not related to the values of x or y.

Unlike linear or polynomial regression, a nonlinear regression problem cannot be solved in one step. Instead the problem must be solved iteratively. An initial estimate (first guess) of the value of each parameter must be provided. The nonlinear regression procedure then adjusts these values to improve the fit of the curve to the data. It then adjusts those new values to improve the fit again. These iterations continue until negligible, if any, improvement occurs. Although iterative procedures are not required, iterative nonlinear regression programs may also be used with equations (such as polynomial equations) in which y is linear with each parameter.

In the discussion that follows, we first consider only equations that contain two parameters to be fit, A and B. Accordingly, nonlinear regression can be viewed in a three-dimensional topographical analogy (Fig. 3). In this model, the horizontal plane represents A and B, and the vertical axis represents the sum of squares. Thus every pair of possible values for A and B is associated with a single sum of squares value z. Depending on the data and the equation chosen, this surface may be simple and symmetrical, or it may contain numerous peaks and valleys. The goal of nonlinear regression—to find the pair of values for A and B that minimize the sum of squares—is to find the deepest valley. Note that the data points do not individually appear in this graph; instead goodness of fit (incorporating all the data) is graphed as a function of each pair of possible values for the parameters A and B.

When the equation has more than two parameters, the topographical analogy contains more than three dimensions and cannot be visualized. Nonetheless, the mathematical principles are the same.

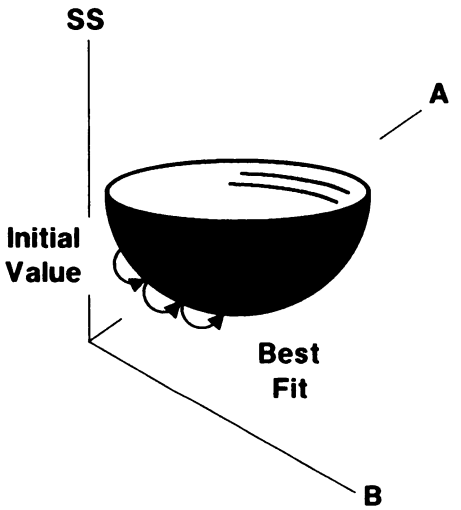

Figure 3. A topographical analogy of nonlinear regression. This three-dimensional graph plots the sum of squares SS (a measure of goodness of fit) as a function of possible values for the two parameters A and B. The purpose of nonlinear regression is to find the values for A and B that minimize the sum of squares. This occurs at the bottom of the valley. In this example the topographical surface is simple; in other examples the surface may contain several peaks and valleys. Note that the individual data points do not appear on this type of plot.

Algorithms

All iterative techniques must be given a starting point on the surface—an initial estimate of the parameters obtained by calculation or intelligent guessing. The nonlinear regression procedure then iteratively moves along the surface by altering the values of the parameters to improve the fit. Several methods can be used to calculate these iterations. In pharmacological and biochemical research, the Marquardt method is most commonly used. This method is a hybrid between two older algorithms, the method of steepest descent and the method of Gauss-Newton.

The method of steepest descent moves along the direction of steepest descent with an arbitrary step length. Then the slope is calculated in the new spot and the procedure is repeated. Essentially this method finds the minimum by repeatedly moving downhill. Although the initial iterations rapidly advance toward the goal, later iterations often zigzag, especially if the step length is large. Many iterations are often required before the method converges on a solution. In some implementations of this method, the step length is varied to hasten the process.

The Gauss-Newton algorithm utilizes another principle. With the topological analogy introduced above, equations in which y is linearly dependent on A and B always generate a single smooth ellipsoid crater. Other equations generate an asymmetrical surface. The Gauss-Newton method approximates the equation so that it does generate a symmetrical ellipsoid surface. From the initial position on that surface, the algorithm can project the entire ellipsoid. It then alters the values of A and B to jump straight to the minimum. With functions linear with the parameters, therefore, the Gauss-Newton method immediately finds the solution without need for further iterations. Nonlinear functions do not generate an ellipsoid surface, but that simplifying assumption usually (but not always) leads to an improvement of the fit. With further iterations, the fit improves. The Gauss-Newton method sometimes works poorly in the initial iterations, and it may move in the wrong direction altogether, making the fit worse. However, this method works well when close to the minimum, where the surface usually can be well approximated by an ellipsoid.

Marquardt designed a method that combines the advantages of both the steepest descent and Gauss-Newton methods, while avoiding their limitations (6). The method of steepest descent works best in initial iterations, and the Gauss-Newton method works best in later iterations. Marquardt's algorithm uses a blend of the two methods. The steepest descent method is emphasized in initial steps, and this method is used repeatedly

as long as the residual sum of squares decreases considerably. When the sum of squares is no longer decreasing, Marquardt's method gradually switches over to the Gauss-Newton principle. This approach has been found to be useful for fitting many types of data to various types of equations.

The simplex algorithm (reviewed in refs 7 and 8) is an alternative method for performing nonlinear regression analysis. One must provide both an initial value and an initial increment for each parameter. From these values, the algorithm generates $N + 1$ starting points, where N is the number of parameters to be fit. Each of these starting points, called a vertex, consists of a possible value for each parameter. The goodness of fit of each vertex is evaluated. The worst vertex is rejected, and a new vertex is generated by blending the best of the others. This algorithm repeatedly rejects the worst vertex and generates a new one until the vertices coalesce to values within a specified tolerance. On our topological analogy, the simplex method starts with three starting values that form a triangle. With each iteration the algorithm rejects the worst of the points, and creates a new one. This can be visualized as a triangular amoeba oozing down the surface. Eventually all three points merge together at the minimum. With more parameters, the vertices form an N-dimensional shape called a simplex. The simplex algorithm for nonlinear regression bears little relationship to the simplex algorithm used for linear programming.

The simplex method has three advantages over the more commonly used methods discussed above: *1*) It is fast. It does not require calculating derivatives. *2*) It rarely converges at a local minimum (see below). *3*) It can be used with noncontinuous functions. The disadvantage of the simplex method is that it does not estimate the standard error of each parameter. It is also somewhat more difficult to use, inasmuch as you must provide starting increments for each parameter.

CHOICES TO BE MADE WHEN USING A NONLINEAR REGRESSION PROGRAM

Scaling the data

In theoretical mathematics, the units used to express data are irrelevant. When using a computer to solve a numerical problem, however, it is important to choose appropriate units. Computers keep track of only a certain number of significant digits (which varies between programs), and all computer calculations involving floating point (noninteger) values are therefore subject to round-off error. It is important, therefore, to express data in reasonable units to avoid very large or very small numbers. For example, round-off errors may be severe if concentrations were entered as $0\text{–}10^{-14}$ M, but would be trivial if those data were expressed as 0–10 fM.

Equation

A nonlinear regression program cannot find the best curve through a set of data points; it can only optimize the parameters in a specified equation. The equation must calculate y as a function of x and one or more parameters. The function should be selected because it describes a hypothetical physical or molecular model. Instead of entering a function, some programs allow one to enter a system of differential equations.

In providing the equation, one must make the distinction between constants (fixed by the experimenter), and parameters that are to be fit by the program. This distinction may greatly affect the final results.

All nonlinear regression procedures (except simplex) repeatedly calculate the derivative (slope) of y with respect to each parameter. Some programs require one to determine the derivatives using calculus and to enter the resulting equations. Other programs calculate the derivatives numerically by evaluating the equation before and after altering the value of a parameter by a small amount Δ, and then dividing the change in y by Δ. Although the calculations are slower, this latter method is of more general utility.

When writing an equation to express a physical model, parameters may be entered in various ways. Thus acidity may enter an equation as $[H^+]$ or as pH. Similarly, a binding affinity may enter an equation as either an equilibrium dissociation constant (K_D, expressed in concentration units), the logarithm of the K_D, or the reciprocal of the K_D (an equilibrium association constant, expressed in reciprocal concentration units). Switching between alternative forms is called reparameterization. Note that this is quite different from the linear transformations mentioned earlier; in those cases the variables x and y were transformed. In reparameterization, the parameters (A, B, . . .) are transformed without altering the x and y values. Reparameterization is useful because many of the statistical inferences that can be made from nonlinear regression are strictly valid only for equations in which y is linear with respect to the parameters. Reparameterizing a nonlinear equation can make it behave more nearly linear, and thus can improve the validity of statistical values. Unfortunately it is quite difficult to measure the extent of nonlinearity of an equation (9). This makes it difficult to ascertain the effectiveness of reparameterization.

Initial estimates

Initial values for each parameter in the equation must be specified, based on previous experience, on preliminary analyses based on linear transformations, or on a hunch. Although it is impossible to give any general guidelines, it is usually not difficult to estimate the parameters if one understands the physical model that the equation represents, and understands the meaning of each parameter. A poor selection of initial values may have several consequences: *1*) In a well-behaved system, the amount of computer time required to reach a solution will be increased, but the solution will be correct. *2*) In other situations, poorly selected initial values can lead the nonlinear regression program to go in the wrong direction and never converge on a solution. *3*) It

is also possible that poorly selected initial values can cause the program to converge on the wrong solution, a local minimum (see below). The choice of initial values is less important with equations containing only one or a few parameters than with equations containing many parameters.

Weighting scheme

Most common nonlinear regression schemes, like linear regression, minimize the sum of the square of the vertical distances of the points from the curve. This approach is statistically valid when the experimental uncertainties do not relate in a systematic way to the values of x or y. Often this is not true. For example, in many experimental situations the experimental uncertainty is (on the average) a constant fraction of the value of y. In these situations, the usual regression methods are not optimal, inasmuch as the points with large y values tend to be further from the curve than are points with small y values. In minimizing the sum of squares, the program would therefore tend to relatively emphasize those points with large y values, and ignore points with small y values. To circumvent this problem, the procedure for quantitating goodness of fit can be altered, so that the deviation of each point from the curve is divided by the y value of that point, and then squared. With this modification, the algorithm will not minimize the square of the distances (expressed in the units of the y variable) of the points from the curve, but instead will sum the square of the relative distances (expressed as fractions). This option, commonly available in regression programs, is termed weighting the data points. In fact this procedure reduces the inappropriate weight given the points with large y values, and should be thought of as unweighting.

The concept of differential weighting of data points can be somewhat confusing. It may help to consider instead the criteria used to measure how far a point is from the curve. Normally this is measured as a distance, expressed in units of the y axis. With the weighting scheme described above, the relative distance, expressed as a fraction, is used instead.

Other choices for differentially weighting the data points are commonly used. One possibility is for the distance of the point from the curve to be divided by the square root of y, and then squared; this is a blend of the two methods listed above. Alternatively, it is common to weight data points according to the number of replicate determinations. Thus points that are the mean of five replicates will be weighted more than points that are the mean of only two replicate determinations. Another possibility is to inversely weight each point according to the standard deviation of the replicate values that went into determining that point. Thus data points with tight replicates will be weighted more heavily than points with a great deal of variability.

When in doubt, it is best to weight each point equally. Use differential weighting only when the relationship between the experimental uncertainties and the value of y is clear. It is not appropriate to weight the data points by $1/y$ or $1/y^2$ when some of the y values are equal to, or nearly equal to, zero, as commonly occurs when one defines y as an experimentally determined value minus a baseline (or nonspecific) value.

Convergence criterion

As nonlinear regression calculations occur iteratively, the computer must be told when to stop—when the calculations have converged. Some programs have built-in criteria; other programs are more flexible. For example, some nonlinear regression programs terminate when an iteration reduces the sum of squares by less than one part in a thousand. Others terminate when an iteration alters the value of each variable by less than one part in a thousand. Too loose a criterion can cause a program to stop before it has reached the best fit. Too tight a criterion can consume a great deal of computer time.

INTERPRETING THE RESULTS

Nonlinear regression programs typically spew out several pages of information about the final equation. This information is designed to answer several questions: How well does the model fit the data? Does this model fit the data better than an alternative model? How much uncertainty is there in the values of the parameters? Does the equation fit this set of data differently from another set of data?

Assessing goodness of fit

In assessing goodness of fit, it is essential to first examine a graph of a curve superimposed on the data points. Many potential problems are easiest to spot graphically. It is inappropriate to use the results of a nonlinear regression program without first examining a graph of the data together with the fit curve. In addition to viewing the graph, several statistical methods can be used for quantitating goodness of fit.

The average deviation of the curve from the points is the square root of SS/df, where SS is the sum of squares, and df is the degrees of freedom. This is also called the root mean square (RMS). If goodness of fit was quantitated using the actual deviation of the points from the curve (no weighting), then the RMS will be expressed in the units used by the y values.

In a residual plot, the y value of each point is replaced by the distance of that point from the curve. Two examples are shown in Fig. 4. If the equation is appropriate for the data, the residuals represent only experimental error. Accordingly, the residuals should not be systematically related to the x values, and the residual plot will have a random arrangement of positive and negative residuals. If the equation is inappropriate, the positive residuals may tend to cluster together at some parts of the graph, whereas negative residuals cluster together at other parts. Such clustering indicates that the data points differ systematically (not just randomly) from the predictions of the curve.

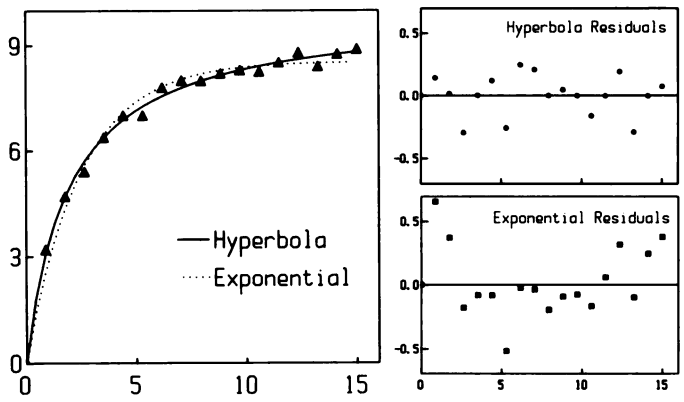

Figure 4. Residual plots. The solid line shows the results of nonlinear regression using the equation describing a rectangular hyperbola, $y = Ax/(B + x)$. The dotted line is the result of nonlinear regression using an equation describing an exponential association, $y = A[1 - \exp(-Bx)]$. Which equation better describes the data? It is difficult to answer that question by examining the curves on the left. The residual plots on the right, however, make the answer clear. The residual plots show how far each point is from the curve. The residuals from the rectangular hyperbola are randomly above and below zero. The residuals from the exponential equation, however, show a systematic pattern, with positive residuals for the first and last few points and negative residuals for the middle points. Such systematic deviations indicate that the data are not well-described by that equation. If these were real data, the choice between the two equations would be based not only on goodness of fit, but also on the physical meaning of the two equations.

The runs test is a simple and robust method for determining whether data differ systematically from a theoretical curve. A run is a series of consecutive points with a residual of the same sign (positive or negative). The runs test statistic is calculated from the number of runs associated with a particular fit of the data. For example the residuals of fit of the exponential equation in Fig. 4 have the following signs: + + − − − − − − − − − − + + − + + . Thus these 17 points have only 5 runs. This is associated with a P value of < 0.05 as determined from an appropriate table (9). A low P value indicates that the curve deviates systematically from the points. In contrast the residuals from the fit of those data to the hyperbolic equation have 11 runs.

The r^2 value (square of the correlation coefficient) often accompanies the results of linear regression, and most scientists have developed a good intuitive grasp of its meaning. This value represents the fraction of the overall variance of the y values that is reduced (or explained) by the line. The r^2 value is traditionally defined only for linear, and not for curved, relationships. Nonetheless it can easily be calculated after nonlinear regression as the fraction of the variance of the y values (from their overall mean) that is reduced (or explained) by the curve. In a perfect fit, $r^2 = 1.00$; in a very poor fit, $r^2 = 0.00$.

Uncertainty in the value of the parameters

After obtaining the converged values of the parameters, one wants to know the reliability of those values. Nonlinear regression programs generally print out estimates of the standard error of the parameters, but these values should not be taken too seriously. In nonlinear functions, errors are neither additive nor symmetrical, and exact confidence limits cannot be calculated. The reported standard error values are based on linearizing assumptions, and will always underestimate the true uncertainty of any nonlinear equation. The extent to which a particular error value is underestimated depends on the particular equation and data being analyzed. As we will discuss below, reparameterizing the equation may improve the accuracy of the error estimates. In a nonlinear equation, the uncertainty of all the parameters will be underestimated, even parameters that are linear with y [for example, A in the equation $y = A\exp(Bx)$].

For the reasons given above, it is not appropriate to use the standard error values printed by a nonlinear regression program in further formal statistical calculations. They are quite useful, however, as an informal measure of the goodness of a fit. Because the estimated standard errors are always underestimates, large error values point to a real problem. Large standard errors will occur in several situations: when the data points have a lot of scatter, when the parameters are highly correlated or redundant (see below), or when data have not been collected over a large enough range of x values.

We have found another approach useful for qualitatively demonstrating the certainty of the values determined by a nonlinear regression program. Along with the data points and the best-fit curve, one may superimpose various other curves calculated by changing the value of one of the parameters altered by a specified amount (see Fig. 5).

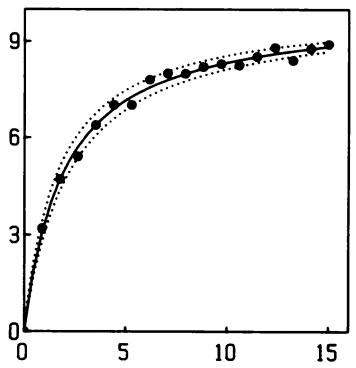

Figure 5. A qualitative method for demonstrating the uncertainty in the values of a variable determined by nonlinear regression. The data points are the same as those in Fig. 4. The solid curve was determined by nonlinear regression using an equation describing a rectangular hyperbola, $y = Ax/(B + x)$. The best-fit values determined by the nonlinear regression program (without weighting) were $A = 10.00 \pm 0.12$ and $B = 2.00 \pm 0.10$. Inasmuch as those uncertainties are estimated standard errors, the 95% confidence limits for the value of B is between 1.8 and 2.2. However, the error values determined by nonlinear regression are not precise. To understand intuitively the uncertainty of the value of B, the dotted curves were computer generated leaving $A = 10.0$, but setting B to either 1.7 (top) or 2.3 (bottom). Clearly these curves do not fit the data well, and B must therefore be between 1.7 and 2.3.

This approach can be extended systematically in the Monte Carlo method. First the data are fitted using any of the algorithms. Then a set of ideal data are created using the same x values as the actual data, but replacing the y values with values predicted by the best-fit curve. Next pseudoexperimental data are generated by adding random error to the ideal data points. This is done by adding to each point a random number calculated from a Gaussian distribution with a standard deviation equal to the mean standard deviation of replicate values in the actual data. Multiple sets of pseudo data should be generated in this manner, and each should be analyzed by nonlinear regression. The standard deviation of estimates of a parameter from multiple sets of pseudo data is a reasonable estimate of the error of the value of that parameter determined from analysis of the actual data. This method is especially useful when used in conjunction with the simplex method, which does not otherwise provide any estimate of the errors.

In both linear and nonlinear regression, the parameters are usually not entirely independent. Thus, altering one parameter will make the fit worse, but this can be partially offset by adjusting the value of another variable. The degree to which two parameters depend on one another is shown in the covariance matrix, which is often reported as correlation coefficients for every pair of parameters. A correlation of zero means that two parameters are completely independent; a correlation of 1.0 or −1.0 means that the two parameters are redundant (see below). Parameters in most models are correlated, and correlation coefficients as large as 0.8 are commonly seen. Very high correlations can indicate that the equation includes redundant variables, that the data points do not span a sufficiently large range of x values to define a curve well, or that too few data points are collected. The joint distribution of the confidence limits of two parameters is sometimes presented as a confidence ellipse, as shown in Fig. 6.

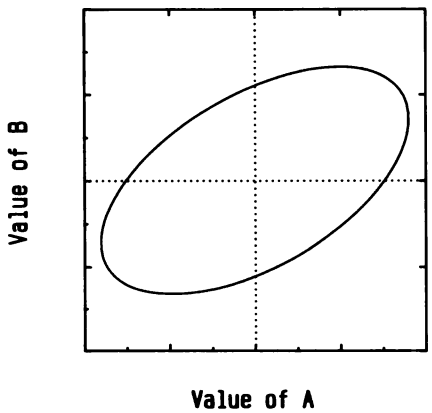

Figure 6. A confidence ellipse. The values of the parameters reported by a nonlinear regression program are not independent of one another. Thus increasing the value of one parameter may worsen the fit, but altering the value of another parameter may partially offset this decrease. The joint distribution of two parameters can be plotted as a confidence ellipse. In the example shown the 95% confidence limits of A and B are plotted. Any pair of values of A and B can be plotted as a single point, and we are 95% certain that the true values of A and B lie within the ellipse.

Comparing two fits to one set of data

The sum of squares value allows one to compare two fits to the same data. Such a comparison is meaningful only when the data have not been changed or transformed between fits, and when the same weighting scheme was used for each fit. This allows one to compare two different equations (models) fit to the same data. Clearly, the selection of a model to explain a particular set of data should not be based entirely on statistical measures. More important are the physical plausibility of the model, and its consistency with other types of data.

Comparing two models with the same number of parameters is easy: the fit with the lower sum of squares is superior, for its curve lies closer to the points. The statistical significance is obtained by $F = SS_1/SS_2$, where both numerator and denominator have $N - V$ degrees of freedom (SS = residual sum of squares of each fit, N = number of data points; V = number of parameters fit by the program).

Comparing two models with a different number of parameters is less straightforward because increasing the number of parameters gives more flexibility to the curve-fitting procedure, and almost always leads to a curve that is closer to the points. The question is whether the improved fit is worth the cost (in lost degrees of freedom) of the additional parameter or parameters. This question is usually answered statistically by performing an F test with the following equation:

$$F = \frac{(SS_1 - SS_2)/(df_1 - df_2)}{SS_2/df_2}$$

Here SS refers to the sum of squares, and df refers to the number of degrees of freedom (number of data points minus number of parameters). The subscript 1 refers to the fit with fewer parameters, the more simple model. A P value is obtained from the F value by consulting a standard table using $(df_1 - df_2)$ and df_2 degrees of freedom. A small P value indicates that the more complex model (with more parameters) fits the data significantly better than the simpler model.

Comparing data sets

Usually one wishes to fit two sets of data to the same general model, and use the results to determine whether the two sets of data differ significantly. For example, one may perform dissociation rate experiments under two different conditions and ask whether the resulting exponential dissociation rate constants differ significantly. Despite the fact that this is a common problem, there is no clear consensus for its solution. One intuitively straightforward method is not appropriate. As discussed above, it is not proper to use the standard error values reported by the nonlinear regression program to compare two models, as those standard error values are underestimates of the actual uncertainty.

One approach is to repeat the experiment several times, and to compare the resulting parameters using paired t tests. We prefer this method and use it in our work inasmuch as it is straightforward to calculate and easy to explain to others. However, this method does not use all available information for each experiment; it uses only the estimated value of each parameter without taking into account its standard error. Accordingly, the method is statistically conservative, and the resulting P value may be too high. Thus small differences may be missed by this method, especially if the experiment has been performed only two or three times.

A more powerful approach can also be used (10). First the two sets of data are analyzed separately. The overall values for the sum of squares for the two sets of data analyzed separately are the sums of the individual values from each fit. Similarly, the number of degrees freedom is the sum of the values from each fit:

$$SS_{sep} = SS_1 + SS_2 \qquad df_{sep} = df_1 + df_2$$

Next the two sets of data are pooled (combined) and analyzed simultaneously. This pooled fit yields values for SS_{pool} and df_{pool}. The question is whether the separate fit is significantly better than the pooled fit. The significance of the improvement is determined from the F ratio calculated as

$$F = \frac{(SS_{pool} - SS_{sep})/(df_{pool} - df_{sep})}{SS_{sep}/df_{sep}}$$

To interpret the meaning of this F value, a statistical table is used to convert to a P value. In using such a table, the numerator has $(df_{pool} - df_{sep})$ degrees of freedom; the denominator has df_{sep} degrees of freedom. A large F value (with corresponding low P value) indicates that the separate fit is much better than the pooled fit—that the two sets of data are not well fit by one curve.

The t test method mentioned above is used to determine one P value from a series of paired experiments. With this method, N refers to the number of experiments. With the F test method, a P value is calculated from one paired experiment. With this method, N refers to the number of data points in each half of that experiment. Although more powerful, this method is potentially misleading inasmuch as an extremely small P value can come from a single experiment. Despite a low P value, such results cannot be considered to be significant until the experiment is repeated.

Linear assumptions applied to nonlinear models

The methods used for calculating standard errors of the parameters and for comparing two fits are strictly valid only when applied to equations in which y is linear with each of the parameters. With other equations, the uncertainties do not follow a Gaussian distribution. Thus the calculated confidence limits are only approximate. Moreover, the sums of squares also do not follow a Gaussian distribution, and statistical inferences based on comparisons of sums of squares are also only approximate. In a practical sense, these problems rarely have a substantial impact on conclusions based on nonlinear regression.

PROBLEMS

Nonconvergence

Sometimes the calculations terminate without converging because of overflow or underflow (a number becomes too large or too small for the computer to handle), or because the system becomes ill-conditioned or a matrix becomes singular. This occurs in several situations: *1*) The data contain numbers that are too large or too small. *2*) The selected equation does not reasonably fit the data. *3*) The initial values are far from correct. *4*) The data points are quite scattered. *5*) The data were not collected over a sufficiently wide range of x values. *6*) The computer calculations were not sufficiently precise (not enough significant digits).

Slow convergence

Nonlinear regression analyses usually converge in fewer than 5–10 iterations, even when the initial estimates are far from the final values. Slow convergence indicates that the program is having difficulty finding a solution. This may occur for several reasons: *1*) The convergence criterion is too stringent. *2*) The data are not adequate to define the parameters; more data points or more widely spaced points are required. *3*) The equation contains too many parameters; it may help to fix one or more parameters as a constant. *4*) The equation needs to be reparameterized, as discussed above.

Redundant variables

Nonlinear regression fails when one tries to fit data to an equation with too many parameters. For example, it would be a mistake to try to fit exponential growth data to the equation $y = A\exp(Bx + C)$, as the parameters A and C are essentially interchangeable, both representing the y intercept. {This can be seen by rewriting the equation as $y = [A\exp(C)]\exp(Bx)$; only a single parameter is needed to represent the y intercept, and a nonlinear regression program has no basis for deciding how to apportion that value between A and C.} With a complicated equation, it can be difficult to detect such redundancy. Consider this possibility if the reported standard error values of the parameters are extremely large, or if the nonlinear regression program consistently fails to converge on a solution.

Local minima

In nonlinear regression, it is possible that the final (converged) fit is not the best possible fit. Recall that nonlinear regression makes relatively small changes during each iteration, and stops when making such small

changes no longer improves the fit of the equation to the data. It is possible, however, that making much larger changes in the values of the variables would lead to a much better fit. This is usually not a problem when only one or two variables are being fit; it is more likely to occur when many variables are being fit. This problem is known as finding a local minimum. In our topographical analogy, the goal is to find the spot with the lowest altitude. The nonlinear regression procedures nearly always move downhill, so it is possible that it converges at the bottom of one valley without finding a much deeper valley just over a ridge (Fig. 7). To guard against this problem, one can repeat nonlinear regression several times, using different starting values. More important, it helps to have a good intuitive grasp of the equation used.

MODIFICATIONS TO NONLINEAR REGRESSION

Robust nonlinear regression

Most statistical analyses, including nonlinear regression, are based on the assumption that the experimental variation of each data point is the sum of many small random inaccuracies. From this assumption comes the notion that experimental data ought to be distributed in a Gaussian (normal) fashion. Accordingly, about 5% of the data points should lie more than two standard deviations from the mean, and fewer than 1 in 10,000 data points should be more than four standard deviations from the mean. In real life, outliers occur more commonly than that! This is because experimental variation results not only from the sum of small inaccuracies, but also from larger mistakes. As used by statisticians, the term experimental error does not include mistakes. For example, the Gaussian distribution accounts for the imprecision of pipettes; it does not account for the fact that the experimenter may accidentally include a large bubble, thus reducing the volume of reagent.

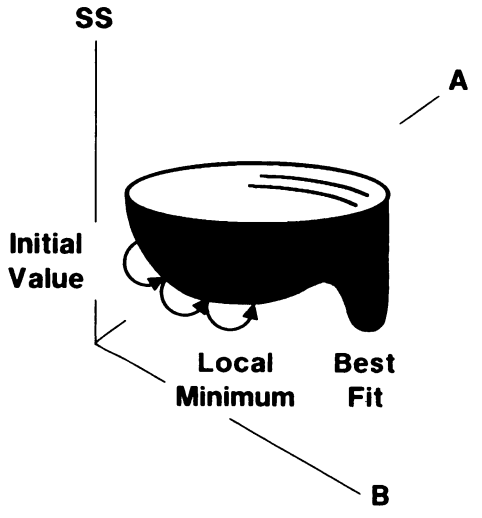

Figure 7. Local minimum. This figure is similar to Fig. 3. The goal of nonlinear regression is to find the values of the parameters that minimize the sum of squares. With reasonable starting values, nonlinear regression algorithms nearly always move stepwise downhill until the minimum (best fit) is found. With an inappropriate starting place, however, a local minimum might be found instead.

Several approaches can be used to deal with outlier. One approach is to let them be—to include all points in the analysis. Although this seems like a compulsive and conservative approach, it may lead to incorrect results, as outlier can have a big influence on the resulting curve. The square of the distance of an outlying point from the curve can be huge, and the nonlinear regression will twist the curve considerably to reduce that distance (see Fig. 8). A more common approach is to delete outlying points, either on an ad hoc or a systematic basis.

Another approach for dealing with outlier is to modify the criteria used to assess goodness of fit to reduce the weight given outlying points and thus make the nonlinear regression process more robust. Several such methods have been described. Each method is designed to be identical to the sum of squares method for most of the points, but to reduce the importance of outlying points. Each such method requires that you choose a robustness constant c, a deviation (z, the absolute value of the distance of a point from the curve) beyond which a point is likely to be invalid. Wahrendorf proposed that the sum of squares (z^2) be replaced by $2Cz - C^2$, when $z > C$ (11). Thus points that are close to the curve ($z < C$) are treated as usual, but as the deviations increase their contribution will be proportional to the distance, not the square, of the distance. Mosteller and Tukey proposed an alternate scheme where points with $z > C$ are ignored, and the other points are used to minimize the sum of $[1 - (z/C)^2]^2$ (12).

Extended nonlinear regression

We discussed earlier the importance of choosing the appropriate weighting scheme, as nonlinear regression will converge on the correct solution only when the points are appropriately weighted according to the expected experimental uncertainties. To account for common patterns of experimental uncertainty, the sum of squares is often calculated as the sum of (D^2/y^N), where D is the distance of the point from the curve, y is the y value of the point, and N is 0, 1, or 2. If n is zero, this method is the same as ordinary sum of squares; it sums the square of the distances of the points from the curve. When $N = 2$, this equation reduces to $(D/y)^2$, which sums the square of the relative distance (as a fraction) of the points from the curve. Setting $N = 1$ is a blend between using absolute and relative distances.

In many experimental contexts it is difficult to know what value to give N, and N does not even have to be given an integer value. The idea of extended nonlinear regression is to let the regression procedure fit the value of N as well as the other parameters in the equation. In two simulations of pharmacokinetics, extended nonlinear regression was found to work as well as ordinary weighted nonlinear regression using the appropriate

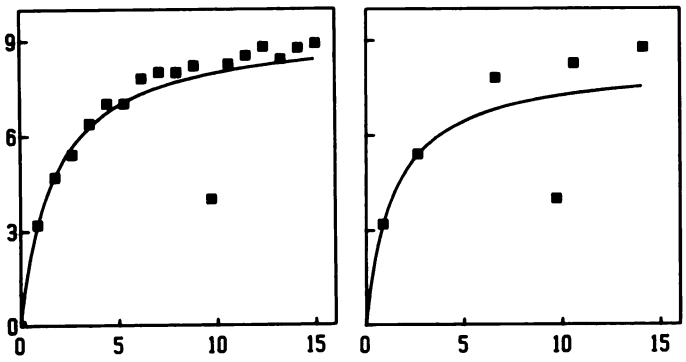

Figure 8. Effects of an outlier. These graphs clearly show how a single outlier point can greatly distort the curve determined by nonlinear regression. Outlier are more influential in data sets with few points (right). Except for the outlier, the data are the same as in Figs. 4 and 5.

value for N, and much better than ordinary regression given the wrong value for N (13, 14). Because this method increases the number of parameters to be fit by regression procedure, it will be less useful when the number of data points is small.

PERSPECTIVE

Nonlinear regression is a powerful technique for data analysis. Although nonlinear regression calculations cannot be reasonably performed by hand, computer programs that perform these calculations are available for microcomputers found in many laboratories. Nonlinear curve-fitting analyses are easy, fast, and practical for routine use. No sophisticated knowledge of computers or mathematics is required to use such programs. To use nonlinear regression properly, however, it is necessary to have an intuitive feel for the selected equation and for the physical model it represents. Without such an understanding, the use of nonlinear regression will be frustrating and potentially misleading. Results from a nonlinear regression program should be carefully scrutinized and viewed graphically. Like every scientific technique, a nonlinear regression program may produce misleading results when used inappropriately.

Computers are associated with a powerful mystique, and incorrect or incomplete computer analyses are sometimes published and accepted uncritically. We hope that this review will demystify nonlinear regression, so that both its power and limitations will be appreciated. FJ

This work was supported by a grant from the National Institutes of Health and the Swedish Medical Research Council.

REFERENCES

1. Landaw, E. M.; DiStefana, A. J. Multiexponential, multicompartmental, and noncompartmental modelling. II. Data analysis and statistical considerations. *Am. J. Physiol.* 246: R665-R677; 1984.
2. Duggleby, R. G. A nonlinear regression program for small computers. *Anal. Biochem.* 110: 9-18; 1981.
3. Jennrich, R. I.; Ralston, M. L. Fitting nonlinear models to data. *Annu. Rev. Biophys. Bioeng.* 8: 195-238; 1979.
4. Draper, N. R.; Smith, H. An introduction to nonlinear estimation. *Applied regression analysis.* New York: Wiley; 1981: 458-517.
5. Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. Modelling of data. *Numerical recipes: the art of scientific computing.* Cambridge: Cambridge University Press; 1986: 498-546.
6. Caceci, M. S.; Cacheris, W. P. Fitting curves to data. The simplex algorithm is the answer. *Byte* 9: 340-362; 1984.
7. Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. Minimization or maximization of functions. *Numerical recipes: the art of scientific computing.* Cambridge: Cambridge University Press; 1986: 289-293.
8. Marquardt, D. W. An algorithm for least-squares estimation of nonlinear parameters. *J. Soc. Ind. Appl. Math.* 11: 431-444; 1963.
9. Beyer, W. H. *Handbook of tables for probability and statistics.* Cleveland: CRC Press; 1968: 397-398.
10. Ratkowsky, D. Comparing parameter estimates from more than one data set. *Nonlinear regression modelling: a unified and practical approach.* New York: Dekker; 1983: 135-152.
11. Wahrendorf, J. The application of robust nonlinear regression methods for fitting hyperbolic scatchard plots. *Int. J. Bio-Med. Comput.* 10: 75-87; 1979.
12. Mosteller, F.; Tukey, J. W. *Data analysis and regression.* New York: Addison-Wesley; 1977.
13. Peck, C. C.; Beal, S. L.; Sheiner, L. B.; Nichols, A. I. Extended least squares nonlinear regression: a possible solution to the "choice of weights" problem in analysis of individual pharmacokinetic data. *J. Pharmacokinet. Biopharm.* 12: 545-558; 1984.
14. Thomson, A. H.; Kelman, A. W.; Whitting, B. Evaluation of nonlinear regression with extended least squares: simulation study. *J. Pharm. Sci.* 74: 1327-1331; 1985.