Critical Reviews in Biochemistry and Molecular Biology, 35(5):359–391 (2000)

Beyond Eyeballing: Fitting Models to Experimental Data

Arthur Christopoulos and Michael J. Lew

Table of Contents

1040-9238/00/$.50

I. INTRODUCTION

A. "Eyeballing"

The oldest and most commonly used tool for examining the relationship between experimental variables is the graphical display. People are very good at recognizing patterns, and can intuitively detect various modes of behavior far more easily from a graph than from a table of numbers. The process of "eyeballing the data" thus represents the experimenter's first attempt at understanding their results and, in the past, has even formed the basis of formal quantitative conclusions. Eyeballing can sometimes be assisted by judicious application of a ruler, and often the utility of the ruler has been enhanced by linearizing data transformations. Nowadays it is more common to use a computer-based curve-fitting routine to obtain an "unbiased" analysis. In some common circumstances there is no important difference in the conclusions that would be obtained by the eye and by the computer, but there are important advantages of the more modern methods in many other circumstances. This chapter will discuss some of those methods, their advantages, and how to choose between them.

B. Models

The modern methods of data analysis frequently involve the fitting of mathematical models to the data. There are many reasons why a scientist might choose to model and many different conceptual types of models. Modeling experiments can be entirely constructed within a computer and used to test "what if" types of questions regarding the underlying mathematical aspects of the system of interest. In one sense, scientists are constructing and dealing with models all the time inasmuch as they form "worldview" models; experiments are designed and conducted and then used in an intuitive fashion to build a mental picture of what the data may be revealing about the experimental system (see Kenakin, this volume). The experimental results are then frequently analyzed by applying either empirical or mechanistic mathematical models to the data. It is these models that are the subject of this article.

II. EMPIRICAL OR MECHANISTIC?

Empirical models are simple descriptors of a phenomenon that serve to approximate the general shape of the relationship being investigated without any theoretical meaning being attached to the actual parameters of the model. In contrast, mechanistic models are primarily concerned with the quantitative properties of the relationship between the model parameters and its variables, that is, the processes that govern (or are thought to govern) the phenomenon of interest. Common examples of mechanistic models are those related to mass action that are applied to binding data to obtain estimates of chemical dissociation constants whereas nonmechanistic, empirical models might be any model applied to drug concentration–response curves in order to obtain estimates of drug potency. In general, mechanistic models are often the most useful, as they consist of a quantitative formulation of a hypothesis.[1] However, the consequences of using an inappropriate mechanistic model are worse than for empirical models because the parameters in mechanistic models provide information about the quantities and properties of real system components. Thus, the appropriate-

ness of mechanistic models needs close scrutiny.

The designation of a mathematical model as either empirical or mechanistic is based predominantly on the purpose behind fitting the model to experimental data. As such, the *same* model can be both empirical and mechanistic depending on its context of use. As an example, consider the following form of the Hill equation:

$$Y = \frac{\alpha[A]^S}{[A]^S + K^S} \quad (1)$$

This equation is often used to analyze concentration–occupancy curves for the interaction of radioligands with receptors or concentration–response curves for the functional interaction of agonist drugs with receptors in cells or tissues. The Hill equation describes the observed experimental curve in terms of the concentration of drug (A), a maximal asymptote (α), a midpoint location (K), and a midpoint slope (S). In practice, these types of curves are most conveniently visualized on a semi-logarithmic scale, as shown in Figure 1.

When Hill first derived this equation,[2,3] he based it on a mechanistic model for the binding of oxygen to the enzyme, hemoglobin. In that context, the parameters that Hill was interested in, K and S, were meant to reveal specific biological properties about the interaction he was studying; K was a measure of the affinity of oxygen for the enzyme and S was the number of molecules of oxygen bound per enzyme. Subsequent experiments over the years have revealed that this model was inadequate in accounting for the true underlying molecular mechanism of oxygen-hemoglobin binding, but the equation remains popular both as a mechanistic model when its validity is accepted, and as an empirical model where its shape approximates that of experimental data. For instance, if the experimental curve is a result of the direct binding of a radioligand to a receptor, then application of Equation (1) to the dataset can be used to detect whether the interaction conforms to the simplest case of one-site mass-action binding and, if S = 1, the parameters K and α can be used as quantitative estimates of the ligand-receptor dissociation constant (K_D) and total density of receptors (B_{max}), respectively. This is an example where the Hill equation is a mechanistic equation, because the resulting parameters provide actual information about the underlying properties of the interaction. In contrast, concentration–response curves represent the final element in a series of sequential biochemical cascades that yield the observed response subsequent to the initial mass-action binding of a drug to its receptor. Thus, although the curve often retains a sigmoidal shape that is similar to the binding curve, the Hill equation is no longer valid as a mechanistic equation. Hence, the Hill equa-

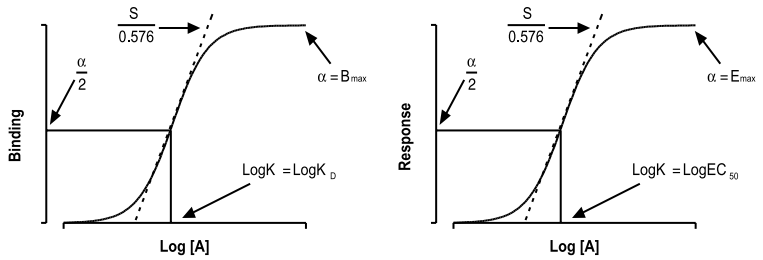

FIGURE 1. Concentration–binding (left) and concentration–response (right) curves showing the parameters of the Hill equation (α, K, and S) as mechanistic (left) or empirical (right) model descriptors.

tion is useful in providing a good fit to sigmoidal concentration–response curves, but the resulting parameters are considered empirical estimates of maximal response, midpoint slope, and midpoint location, and no mechanistic interpretation should be made.

III. TYPES OF FITTING

The variables whose relationships that can be plotted on Cartesian axes do not necessarily have the same properties. Often one variable is controlled by the experimenter and the other variable is a measurement. Thus one variable has substantially more uncertainty or variability than the other, and traditionally that variable would be plotted on the vertical axis. In that circumstance the Y variable can be called the "dependent" variable because of its dependence on the underlying relationship and on the other variable, which is called "independent" to denote its higher reliability. It is important to note that not all datasets have a clearly independent variable. Historically, the statistical determination of the relationship between two or more dependent variables has been referred to as a *correlation analysis*, whereas the determination of the relationship between dependent and independent variables has come to be known as a *regression analysis*. Both types of analyses, however, can share a number of common features, and some are discussed below.

A. Correlation

Correlation is not strictly a regression procedure, but in practice it is often confused with linear regression. Correlation quantifies the degree by which two variables vary together. It is meaningful only when both variables are outcomes of measurement such that there is no independent variable.

1. The Difference between Correlation and Linear Regression

Correlation quantifies how well two dependent variables vary together; linear regression finds the line that best predicts a dependent variable given one or more independent variables, that is, the "line of best-fit."[4] Correlation calculations do not find a best-fit straight line.[5]

2. The Meaning of r^2

The *direction* and *magnitude* of the correlation between two variables can be quantified by the correlation coefficient, r, whose values can range from –1 for a perfect negative correlation to 1 for a perfect positive correlation. A value of 0, of course, indicates a lack of correlation. In interpreting the meaning of r, a difficulty can arise with values that are somewhere between 0 and –1 or 0 and 1. Either the variables do influence each other to some extent, or they are under the influence of an additional factor or variable that was not accounted for in the experiment and analysis. A better "feel" for the covariation between two variables may be derived by squaring the value of the correlation coefficient to yield the *coefficient of determination*, or r^2 value. This number may be defined as the fraction of the variance in the two variables that is shared, or the fraction of the variance in one variable that is explained by the other (provided the following assumptions are valid). The value of r^2, of course, will always be between 0 and 1.

3. Assumptions of Correlation Analysis

1. The subjects are randomly selected from a larger population. This is often not true in biomedical research, where randomization is more common than sampling, but may be sufficient to assume that the subjects are at least representative of a larger population.
2. The samples are paired, i.e., each experimental unit has both X and Y values.
3. The observations are independent of each other. Sampling one member of the population should not affect the probability of sampling another member (e.g., making measurements in the same subject twice and treating them as separate datapoints; making measurements in siblings).
4. The measurements are independent. If X is somehow involved or connected to the determination of Y, or vice versa, then correlation is not valid. This assumption is very important because artifactual correlations can result from its violation. A common cause of such a problem is where the Y value is expressed as either a change from the X value, or as a fraction of the corresponding X value (Figure 2).
5. The X values were measurements, not controlled (e.g., concentration, etc.). The confidence interval for r^2 is otherwise meaningless, and we must then use linear regression.
6. The X and Y values follow a Gaussian distribution.
7. The covariation is linear.

4. Misuses of Correlation Analysis

Often, biomedical investigators are interested in comparing one method for measuring a biological response with another. This usually involves graphing the results as an X, Y plot, but what to do next? It is quite common to see a correlation analysis applied to the two methods of measurement and the correlation coefficient, r, and the resulting *P* value utilized in hypothesis testing. However, Ludbrook[6] has outlined some serious criticisms of this approach, the major one being that although correlation analysis will identify the strength of the linear association between X and Y, as it is intended to do, it will give no indication of any bias between the two methods of measurement. When the purpose of the exercise is to identify and quantify fixed and proportional biases be-

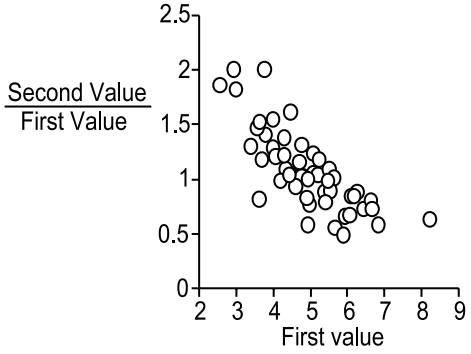

FIGURE 2. An apparent correlation between two sets of unrelated random numbers (pseudo-random numbers generated with mean = 5 and standard deviation = 1) comes about where the Y value is expressed as a function of the X value (here each Y value is expressed as a fraction of the corresponding X value).

tween two methods of measurement, then correlation analysis is inappropriate, and a technique such as ordinary or weighted least products regression[6] should be used.

B. Regression

The actual term "regression" is derived from the latin word "*regredi*," and means "to go back to" or "to retreat." Thus, the term has come to be associated with those instances where one "retreats" or "resorts" to approximating a response variable with an estimated variable based on a functional relationship between the estimated variable and one or more input variables. In regression analysis, the input (independent) variables can also be referred to as "regressor" or "predictor" variables.

1. Linear Regression

The most straightforward methods for fitting a model to experimental data are those of linear regression. Linear regression involves specification of a linear relationship between the dependent variable(s) and certain properties of the system under investigation. Surprisingly though, linear regression deals with some curves (i.e., nonstraight lines) as well as straight lines, with regression of straight lines being in the category of "ordinary linear regression" and curves in the category of "multiple linear regressions" or "polynomial regressions."

2. Ordinary Linear Regression

The simplest general model for a straight line includes a parameter that allows for inexact fits: an "error parameter" which we will denote as ε. Thus we have the formula:

$$Y = \alpha + \beta X + \varepsilon \qquad (2)$$

The parameter, α, is a *constant*, often called the "intercept" while b is referred to as a *regression coefficient* that corresponds to the "slope" of the line. The additional parameter, ε, accounts for the type of error that is due to random variation caused by experimental imprecision, or simple fluctuations in the state of the system from one time point to another. This error term is sometimes referred to as the *stochastic* component of the model, to differentiate it from the other, *deterministic*, component of the model (Figure 3).[7] When data are fitted to the actual straight-line model, the error term denoted by ε is usually not included in the fitting procedure so that the output of the regression forms a perfect straight line based solely on the deterministic component of the model. Nevertheless, the regression procedure assumes that the scatter of the datapoints about the best-fit straight line reflects the effects of the error term, and it is also implicitly assumed that ε follows a Gaussian distribution with a mean of 0. This assumption is often violated, however, and the implications are discussed elsewhere in this article. For now, however, we will assume that the error is Gaussian; Figure 4 illustrates the output of the linear model with the inclusion of the error term. Note that the Y values of the resulting "line" are randomly distributed above and below the ideal (dashed) population line defined by the deterministic component of the model.

3. Multiple Linear Regression

The straight line equation [Equation (2)] is the simplest form of the linear regression

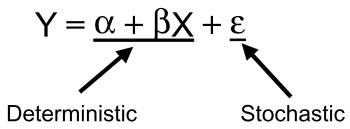

FIGURE 3. The simple linear population model equation indicating the deterministic component of the model that is precisely determined by the parameters α and β, and the stochastic component of the model, ε, that represents the contribution of random error to each determined value of Y.

model, because it only includes one independent variable. When the relationship of interest can be described in terms of more than one independent variable, the regression is then defined as "multiple linear regression." The general form of the linear regression model may thus be written as:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i + \varepsilon \quad (3)$$

where Y is the dependent variable, and X1, X2 … Xi are the (multiple) independent variables. The output of this model can deviate from a straight line, and one may thus question the meaning of the word "linear" in "linear regression." Linear regression implies a linear relationship between the dependent variable and the *parameters*, not the independent variables of the model. Thus Equation (3) is a linear model because the parameters α, β_1, β_2 … β_i have the (implied) exponent of unity. Multiple linear regression models also encompass polynomial functions:

$$Y = \alpha + \beta_1 X + \beta_2 X^2 + \ldots + \beta_i X^i + \varepsilon \quad (4)$$

The equation for a straight line [Equation (2)] is a first-order polynomial. The quadratic equation, $Y = \alpha + \beta_1 X + \beta_2 X^2$, is a second-order polynomial whereas the cubic equation, $Y = \alpha + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$ is a third-order polynomial. Each of these higher order polynomial equations defines curves, not straight lines. Mathematically, a linear model can be identified by taking the first derivative of its deterministic component with respect to the parameters of the model. The resulting derivatives should not include any of the parameters; otherwise, the model is said to be "nonlinear." Consider the following second-order polynomial model:

$$Y = \alpha + \beta_1 X + \beta_2 X^2 \quad (5)$$

Taking first derivatives with respect to each of the parameters yields:

$$\frac{\partial Y}{\partial \alpha} = 1 \quad (6)$$

$$\frac{\partial Y}{\partial \beta_1} = X \quad (7)$$

$$\frac{\partial Y}{\partial \beta_2} = X^2 \quad (8)$$

The model is linear because the first derivatives do not include the parameters. As a consequence, taking the second (or higher) order derivative of a linear function with respect to its parameters will always yield a value of zero.[8] Thus, if the independent variables and all but one parameter are held constant, the relationship between the dependent variable and the remaining parameter will always be linear.

It is important to note that linear regression does not actually test whether the data sampled from the population follow a linear relationship. It assumes linearity and attempts to find the best-fit straight line relationship based on the data sample.

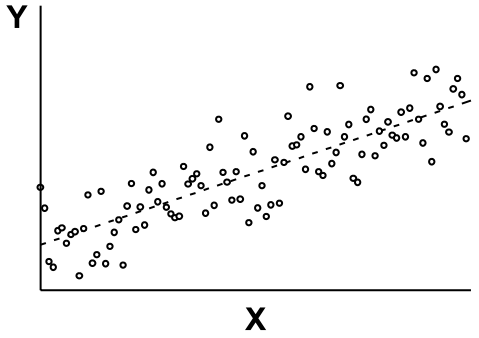

FIGURE 4. A linear model that incorporates a stochastic (random error) component. The dashed line is the deterministic component, whereas the points represent the effect of random error [denoted by the symbol ε in Equation (2)].

4. Nonlinear Regression

Because there are so many types of nonlinear relationships, a general model that encompasses all their behaviors cannot be defined in the sense used above for linear models, so we will define an explicit nonlinear function for illustrative purposes. In this case, we will use the Hill equation [Equation (1); Figure 1] which contains one independent variable [A], and 3 parameters, α, K, and S. Differentiating Y with respect to each model parameter yields the following:

$$\frac{\partial Y}{\partial \alpha} = \frac{[A]^S}{[A]^S + K^S} \tag{9}$$

$$\frac{\partial Y}{\partial K} = \frac{-\alpha S(K[A])^S}{K([A]^S + K^S)^2} \tag{10}$$

$$\frac{\partial Y}{\partial K} = \frac{-\alpha S(K[A])^S}{K([A]^S + K^S)^2} \tag{11}$$

All derivatives involve at least two of the parameters, so the model is nonlinear. However, it can be seen that the partial derivative in Equation (9) does not contain the parameter, α. A linear regression of Y on $[A]^S/(K^S + [A]^S)$ will thus allow the estimation of α. Because this last (linear) regression is conditional on knowing the values of K and S, α is referred to as a "conditionally linear" parameter. Nonlinear models that contain conditionally linear parameters have some advantages when it comes to actual curve fitting.[7]

5. Assumptions of Standard Regression Analyses[4,7]

1. The subjects are randomly selected from a larger population. The same caveats apply here as with correlation analyses.
2. The observations are independent.
3. X and Y are not interchangeable. Regression models used in the vast majority of cases attempt to predict the dependent variable, Y, from the independent variable, X and assume that the error in X is negligible. In special cases where this is not the case, extensions of the standard regression techniques have been developed to account for nonnegligible error in X.
4. The relationship between X and Y is of the correct form, i.e., the expectation function (linear or nonlinear model) is appropriate to the data being fitted.
5. The variability of values around the line is Gaussian.

6. The values of Y have constant variance. Assumptions 5 and 6 are often violated (most particularly when the data has variance where the standard deviation increases with the mean) and have to be specifically accounted for in modifications of the standard regression procedures.
7. There are enough datapoints to provide a good sampling of the random error associated with the experimental observations. In general, the minimum number of independent points can be no less than the number of parameters being estimated, and should ideally be significantly higher.

IV. HOW IT WORKS

A. Minimizing an Error Function (Merit Function)

The goal of both linear and nonlinear regression procedures is to derive the "best fit" of a particular model to a set of experimental observations. To obtain the best-fit curve we have to find parameter values that minimize the difference between the observed experimental observations and the chosen model. This difference is assumed to be due to the error in the experimental determination of the datapoints, and thus it is common to see the entire model-fitting process described in terms of "minimization of an error function" or minimization of a "merit function." [9]

The most common representation ("norm") of the merit function for regression models is based on the chi-square distribution. This distribution and its associated statistic, χ^2, have long been used in the statistical arena to assess "goodness-of-fit" with respect to identity between observed and expected frequencies of measures. Because regression analyses also involve the determination of the best model estimates of the dependent variables based on the experimentally observed dependent variables, it is quite common to see the function used to determine the best-fit of the model parameters to the experimental data referred to as the "χ^2 function," and the procedure referred to as "chi-square fitting." [9]

B. Least Squares

The most widely used method of parameter estimation from curve fitting is the method of least squares. To explain the principle behind least squares methods, we will use an example, in this case the simple linear model. Theoretically, finding the slope, β, and intercept, α, parameters for a perfect straight line is easy: any two X,Y pairs of points can be utilized in the familiar "rise-over-run" formulation to obtain the slope parameter, which can then be inserted into the equation for the straight line to derive the intercept parameter. In reality, however, experimental observations that follow linear relationships almost never fall exactly on a straight line due to random error. The task of finding the parameters describing the line is thus no longer simple; in fact, it is unlikely that values for α and β defined by any pair of experimental points will describe the best line through all the points. This is illustrated in Figure 5; although the dataset appears to follow a linear relationship, it can be seen that different straight lines, each characterized by different slopes and intercepts, are derived depending on which two X,Y pairs are used.

What is needed, therefore, is a "compromise" method for obtaining an objective best-fit. We begin with our population model [Equation (2)]:

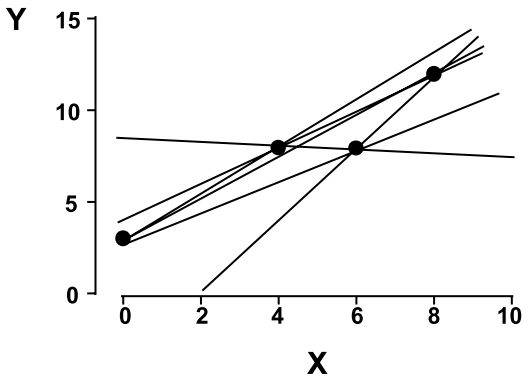

FIGURE 5. All possible straight lines that can be drawn through a four-point dataset when only two points are used to define each line.

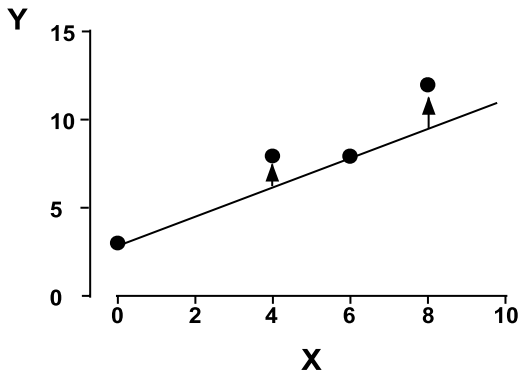

FIGURE 6. A combination of zero and nonzero residuals. The dataset is the same as in Figure 5, with only one of the lines now drawn through the points. The vertical distance of each point from the line (indicated by the arrows) is defined as the "residual."

$$Y = \alpha + \beta X + \varepsilon$$

and derive an equation that is of the same form:

$$\hat{Y} = \hat{\alpha} + \hat{\beta} X \qquad (12)$$

where $\hat{Y}$ is the *predicted response* and $\hat{\alpha}$ and $\hat{\beta}$ are the *estimates* of the population intercept and slope parameters, respectively. The difference between the response variable, Y, and its predictor, $\hat{Y}$, is called the "*residual*" and its magnitude is therefore a measure of how well $\hat{Y}$ predicts Y. The closer the residual is to a value of zero for each experimental point, the closer the predicted line will be to that point. However, because of the error in the data (the ε term in the population model), no prediction equation will fit all the datapoints exactly and, hence, no equation can make the residuals all equal zero. In the example above, each straight line will yield a residual of zero for two points, but a nonzero residual for the other two points; Figure 6 illustrates this for one of the lines.

A best-fit compromise is found by *minimizing the sum of the squares of the residuals*, hence the name "least squares." Mathematically, the appropriate merit function can be written as:

$$\chi^2 = \sum_{i=1}^{N}\left(\frac{Y_i - f(X_i,\theta)}{w_i}\right)^2 = \sum_{i=1}^{N}\left(\frac{r_i}{w_i}\right)^2 \quad (13)$$

where χ^2 is the weighted sum of the squares of the residuals (ri) and is a function of the parameters (the vector, θ), and the N datapoint, X_iY_i. The term, w_i, is the statistical weight (see below) of a particular datapoint, and when used, most often relates to the standard error of that point. For standard (unweighted) least squares procedures such as the current example, wi equals 1. The least squares fit of the dataset outlined above is shown in Figure 7. Note that the best-fit straight line yields nonzero residuals for three of the four datapoints. Nevertheless, the resulting line is based on parameter estimates that give the smallest sum-of-squares of those residuals.

Why do we use the sum of the square of the residuals and not another norm of the deviation, such as the average of the absolute values of the residuals? Arguably, simply because of convention! Different norms of deviation have different relative sensitivities to small and large deviations and conventional usage suggests that sums of the square residuals represent a sensible compromise.[4,10] The popularity of least squares estimators may also be based on the fact that they are relatively easy to determine and that they are accurate estimators if certain assumptions are met regarding the independence of errors and a Gaussian distribution of errors in the data.[8,9,11] Nonetheless, for extremely large deviations due to outlier points, least squares procedures can fail in providing a sensible fit of the model to the data.

Although the example used above was based on a linear model, nonlinear least squares follow the same principles as linear least squares and are based on the same assumptions. The main difference is that the sum-of-squares merit function for linear models is well-behaved and can be solved analytically in one step, whereas for nonlinear models, iterative or numerical procedures must be used instead.

In most common applications of the least squares method to linear and nonlinear models, it is assumed that the majority of the error lies in the dependent variable. However, there can be circumstances when both X and Y values are attended by random error, and different fitting approaches are warranted. One such approach has been described by Johnson,[12] and is particularly useful for fitting data to nonlinear models. In essence, Johnson's method utilizes a form of the standard χ^2 merit function, given above, that has been expanded to include the "best-fit" X value and its associated variance. The resulting merit function is then minimized using an appropriate least squares curve fitting algorithm.

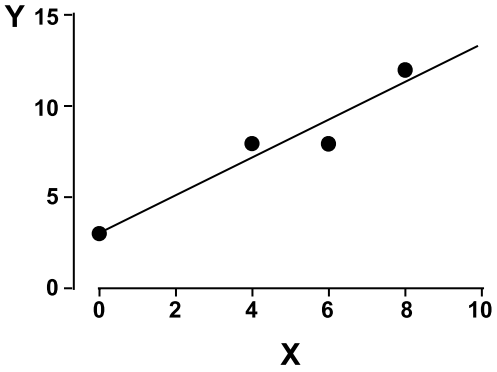

FIGURE 7. The minimized least squares fit of the straight line model [Equation (2)] to the dataset shown in Figures 5 and 6.

C. Nonleast Squares

Cornish-Bowden[11] has listed the minimal requirements for optimal behavior of the least squares method:

a. Correct choice of model.
b. Correct data weighting is known.
c. Errors in the observations are independent of one another.
d. Errors in the observations are normally distributed.
e. Errors in the observations are unbiased (have zero mean).

And we can add:

f. None or the datapoints are erroneous (outliers).

Often, however, the requirements for optimal behavior cannot be met. Other techniques are available for deriving parameter estimates under these circumstances, and they are generally referred to as "robust estimation" or "robust regression" techniques. Because the word "robustness" has a particular connotation, it is perhaps unfair to class all of the diverse nonleast squares procedures under the same umbrella. Overall, however, the idea behind robust estimators is that they are more insensitive to deviations from the assumptions that underlie the fitting procedure than least squares estimators.

"Maximum likelihood" calculations are one class of robust regression techniques that are not based on a Gaussian distribution of errors. In essence, regression procedures attempt to find a set of model parameters that generate a curve that best matches the observed data. However, there is no way of knowing which parameter set is the correct one based on the (sampled) data, and thus there is no way of calculating a probability for any set of fitted parameters being the "correct set." Maximum likelihood calculations work in the opposite direction, that is, given a particular model with a particular set of parameters, maximum likelihood calculations derive a probability for the data being obtained. This (calculated) probability of the *data*, given the parameters, can also be considered to be the likelihood of the parameters, given the data.[9] The goal is then to fit for a set of parameters that maximize this likelihood, hence the term "maximum likelihood," and the calculations attempt to find the regression that has the maximum likelihood of producing the observed dataset. It has been pointed out that there is no formal mathematical basis for the maximum likelihood procedure and because maximum likelihood calculations are quite involved, they are not routinely utilized explicitly.[9] Fortunately the simpler least squares methods described above are equivalent to maximum likelihood calculations where the assumptions of linear and nonlinear regression (particularly the independence and Gaussian distribution of the errors in the data) are valid.[8,9,11]

Certain robust regression techniques focus on using measures of central tendency other than the mean as the preferred statistical parameter estimator. For instance, Cornish-Bowden[11] has described how the median is more insensitive to outlier points in linear regression and certain cases of nonlinear regression than the mean. A drawback of this approach, however, is that it quickly becomes cumbersome when extended to more complex linear problems.

D. Weighting

The simplest minimization functions make no distinction between different experimental points, and assume that each observation contributes equally to the esti-

mation of model parameters. This is appropriate when the variance of all the observations is uniform, and the error is referred to as *homoscedastic*. However, in reality it is common that different points have different variances associated with them with the result that the points with the most variance may have an undue influence on the parameters obtained from an unweighted curve fit. For example, results from many biological experiments are often expressed as a change from a baseline value, with the consequence that the points near the baseline become small numbers (near zero) with a low variance. Points representing larger responses will naturally have a larger variance, a situation that can be described as *heteroscedasticity*. An unweighted curve fit through heteroscedastic data will allow the resulting curve to deviate from the well-defined (tight) near-zero values to improve the fit of the larger, less well-defined values. Clearly it would be better to have the fit place more credence in the more reliably estimated points, something that can be achieved in a weighted curve fit.

Equation (13) was used previously to define the general, least squares, minimization function. There are a number of variations available for this function that employ differential data weighting.[13] These functions explicitly define a value for the w_i term in Equation (13). For instance, if $w_i = 1$ or a constant, then the weighting is said to be "uniform"; if $w_i = Y_i$, then

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{r_i}{Y_i} \right)^2 = \sum_{i=1}^{N} \frac{1}{Y_i^2} (r_i)^2$$

and the weighting is said to be "relative." Relative weighting is also referred to as "weighting by $1/Y^2$" and is useful where the experimental uncertainty is a constant fraction of Y. For example, counts of radioactive decay will have variances described by the Poisson distribution where the variance scales with the mean, and thus the likely error in each estimate is a constant percentage of counts rather than a constant value for any number of counts. Thus, a curve fit allowing for relative weighting can adjust for the resulting heteroscedastic variance. Another useful weighting value is $w_i = \sqrt{Y_i}$. This yields "weighting by 1/Y" and is appropriate, for example, when most of the experimental uncertainty in the dependent variable is due to some sort of counting error.[5] Other weighting schemes utilize the number of replicates that are measured for each value of Y to determine the appropriate weight for the datapoints.[13]

E. Regression Algorithms

What are the actual "mechanics" that underlie the χ^2 minimization process behind least squares regression techniques? The χ^2 merit function for linear models (including polynomials) is quadratic in nature, and is thus amenable to an exact analytical solution. In contrast, nonlinear problems must be solved iteratively, and this procedure can be summarized as follows:

a. Define the merit function.
b. Start with a set of initial estimates (guesses) of the regression parameters and determine the value of the merit function for this set of estimates.
c. Adjust the parameter estimates and recalculate the merit function. If the merit function is improved, then keep the parameter values as new estimates.
d. Repeat step c (each repeat is an "iteration"). When further iterations yield a negligible improvement in the fit, stop adjusting the parameter estimates and generate the curve based on the last set of estimates.

The rules for adjusting the parameters of the nonlinear model are based on matrix algebra and are formulated as computer algorithms. The merit function can be viewed as a multidimensional surface that has all possible sum-of-squares values as one plane and all possible values of each of the model parameters as the other planes. This surface may thus vary from a smooth, symmetrical shape to one characterized by many crests and troughs. The role of the nonlinear regression algorithm is to work its way down this surface to the deepest trough that should then correspond to the set of model parameters that yield the minimum sum-of-squares value.

There are a number of different algorithms that have been developed over the years, and they all have their pros and cons. One of the earliest algorithms is the *method of steepest descent* (or the *gradient search method*[8]). This method proceeds down the steepest part of the multidimensional merit function surface in fixed step lengths that tend to be rather small.[9] At the end of each iteration, a new slope is calculated and the procedure repeated. Many iterations are required before the algorithm converges on a stable set of parameter values. This method works well in the initial iterations, but tends to drag as it approaches a minimum value.[13]

The *Gauss-Newton method* is another algorithm that relies on a linear approximation of the merit function. By making this approximation, the merit function approaches a quadratic, its surface becomes a symmetrical ellipsoid, and the iterations of the Gauss-Newton algorithm allow it to converge toward a minimum much more rapidly than the method of steepest descent. The Gauss-Newton method works best when it is employed close to the surface minimum, because at this point most merit functions are well approximated by linear (e.g., quadratic) functions.[9] In contrast, the Gauss-Newton method can work poorly in initial iterations, where the likelihood of finding a linear approximation to the merit function is decreased.

A method exploiting the best features of the methods of steepest descent and Gauss-Newton was described by Marquardt, based on an earlier suggestion by Levenberg,[9] and the resulting algorithm is thus often referred to as the *Levenberg-Marquardt method*. Marquardt realized that the size of the increments in an interative procedure poses a significant scaling problem for any algorithm, and proceeded to refine the scaling issue and derive a series of equations that can approximate the steepest descent method at early iterations and the Gauss-Newton method at later stages closer to the minimum. The Levenberg-Marquard method (sometimes simply referred to as the Marquardt method) has become one of the most widespread algorithms used for computerized nonlinear regression.

Another type of algorithm that is geometric rather than numeric in nature is the *Nelder-Mead Variable Size Simplex method*.[8,14] Unlike the methods outlined above, this method does not require the calculation of any derivatives. Instead, this algorithm depends on the generation of a number of starting points, called "vertices," based on initial estimates for each parameter of the model, as well as an initial increment step. The vertices form a multidimensional shape called a "simplex." The goodness of fit is evaluated at each vertex in the simplex, the worst vertex is rejected and a new one is generated by combining desirable features of the remaining vertices. This is repeated in an iterative fashion until the simplex converges to a minimum. The big advantage of the Nelder-Mead method is that it is very successful in converging to a minimum; its main disadvantage is that it does not provide any information regarding the errors associated with the final parameter estimates.[8]

V. WHEN TO DO IT (APPLICATION OF CURVE FITTING PROCEDURES)

A. Calibration Curves (Standard Curves)

Calibration curves are most convenient when they are linear, but even for assays where a linear relationship is expected on theoretical grounds, nonlinear curves can result from instrumentation nonlinearities and other factors. The equation of a curve fitted through the calibration data will allow convenient conversion between the raw measurement and the required value. In cases where there is no theoretical basis for choosing one model over another, calibration curves can be considered to be a smoothing rather than a real fitting problem and one might decide to apply a polynomial model to the data because of the availability of an analytical solution. In such a case the order of the chosen polynomial would need to be low so that noise in the calibration measurements is not converted into wobbles on the calibration curve.

B. Parameterization of Data (Distillation)

It is often desirable to describe data in an abbreviated way. An example of this is the need to summarize a concentration–response curve into a potency estimate and maximum response value. These parameters are easily obtained by eyeballing the data, but an unbiased estimate from an empirical curve fit is preferable and probably more acceptable to referees!

VI. HOW TO DO IT

A. Choosing the Right Model

1. Number of Parameters

The expectation function should include the *minimum* number of parameters that adequately define the model and that allow for a successful convergence of the fit.

If a model is overparameterized, it is considered to possess "redundant" parameters (often used interchangeably with the term "redundant variables"), and the regression procedure will either fail or yield meaningless parameter estimates. Consider the "operational model" of Black and Leff.[15] This is a model that is often used in pharmacological analyses to describe the concentration–response relationship of an agonist (A) in terms of its affinity (dissociation constant) for its receptor (K_A), its "operational" efficacy (τ), and the maximum response (E_m) that the tissue can elicit. One common form of the model is:

$$E = \frac{E_m \cdot \tau \cdot [A]}{([A] + K_A) + \tau \cdot [A]} \qquad (14)$$

where E denotes the observed effect. Figure 8 shows a theoretical concentration–response curve, plotted in semilogarithmic space, that illustrates the relationship between the operational model parameters and the maximal asymptote (α) and midpoint location (EC_{50}) of the resulting sigmoidal curve. A concentration–response curve like the one in Figure 8 can be successfully fitted using the two-parameter version of the Hill equation, which describes the curve in terms of only the EC_{50} and α (the slope being equal to 1):

$$E = \frac{\alpha \cdot [A]}{[A] + EC_{50}} \qquad (15)$$

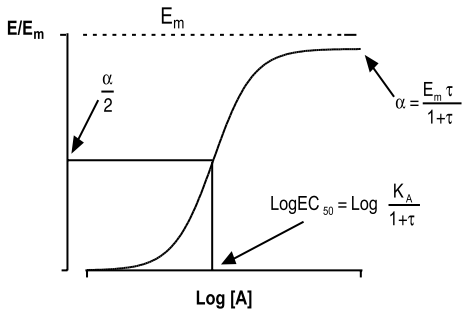

FIGURE 8. The relationship between the Hill equation [Equation (15)] parameters, a and EC_{50}, and the operational model [Equation (14)] parameters K_A, t, and E_m, in the description of a concentration–response curve of an agonist drug. It can be seen that each parameter of the Hill equation is composed of two operational model parameters.

However, it can be seen in Figure 8 that the midpoint and maximal asymptote of the curve are related to the operational model in a more complicated manner; each parameter of the sigmoidal Hill equation is comprised of two operational model parameters. If someone were to try directly fitting Equation (14) to this curve in order to derive individual estimates of E_m, K_A, and τ, they would be unsuccessful. As it stands, the operational model is overparameterized for fitting to a single curve; the regression algorithm simply will not be able to apportion meaningful estimates between the individual operational model parameters as it tries to define the midpoint and maximal asymptote of the concentration–response curve. In practice, the successful application of the operational model to real datasets requires additional experiments to be incorporated in the curve fitting process that allow for a better definition of the individual model parameters.[16,17]

2. Shape

When fitting empirical models to data the most important feature of the model must be that its shape should be similar to the data. This seems extraordinarily obvious, but very little exploration of the literature is needed to find examples where the curve and the data have disparate shapes! Empiricism allows one a great deal of freedom in choosing models, and experimenters should not be overly shy of moving away from the most common models (e.g., the Hill equation) when their data ask for it. Even for mechanistic models it is important to look for a clear shape match between the model and data: a marked difference can only mean that the model is inappropriate or the data of poor quality.

Perhaps the only feature that practically all biological responses have in common is that they can be approximated by nonlinear, saturating functions. When plotted on a logarithmic concentration scale, responses usually lie on a sigmoid curve, as shown in Figures 71 and 8, and a number of functions have been used in the past to approximate the general shape of such responses. Parker and Waud,[18] for instance, have highlighted that the rectangular hyperbola, the integral of the Gaussian distribution curve, the arc-tangent, and the logistic function have all been used by various researchers to empirically fit concentration–response data. Some of these functions are more flexible than others; for instance, the rectangular hyperbola has a fixed slope of 1. In contrast, the logistic equation has proven very popular in the fitting of concentration–response data:

$$E = \frac{1}{1 + e^{-(\alpha+\beta X)}} \qquad (16)$$

Part of the popularity of this equation is its flexibility and its ability to match the parameters of the Hill equation [Equation (1)] for empirical fitting purposes.

In general, the correct choice of expectation function is most crucial when fitting mechanistic models. The difficulty in ascertaining the validity of the underlying model in these cases arises because the curve fitting process is undertaken with the automatic assumption that the model is a plausible one prior to actually fitting the model and applying some sort of diagnostics to the fit (see Assessing the Quality of the Fit, below). We must always remain aware, therefore, that we will never really know the "true" model, but can at least employ a reasonable one that accommodates the experimental findings and, importantly, allows for the prediction of testable hypotheses. From a practical standpoint, this may be seen as having chosen the "right" mechanistic model.

3. Correlation of Parameters

When a model, either mechanistic or empirical, is applied to a dataset we generally consider each of the parameters to be responsible for a single property of the curve. Thus, in the Hill equation, there is a slope parameter, S, a parameter for the maximum asymptote (α), and a parameter for the location (K or EC_{50}). Ideally, each of these parameters would be entirely independent so that error or variance in one does not affect the values of the others. Such a situation would mean that the parameters are entirely uncorrelated. In practice it is not possible to have uncorrelated parameters (see Figure 9), but the parameters of some functions are less correlated than others. Strong correlations between parameters reduce the reliability of their estimation as well as making any estimates from the fit of their variances overly optimistic.[19]

4. Distribution of Parameters

The operational model example can also be used to illustrate another practical consideration when entering equations for curve fitting, namely the concept of "reparameterization."[13] When fitting the operational model or the Hill equation to concentration–response curves, the parameters may be entered in the equation in a number of ways; for instance, the EC_{50} is

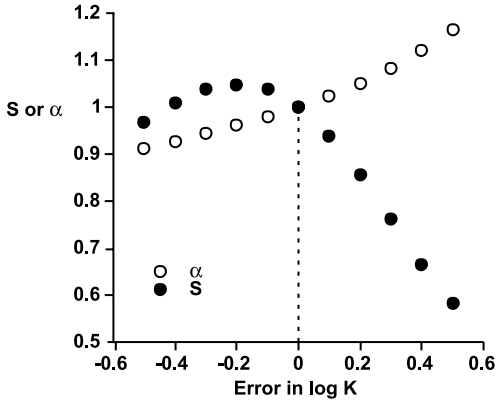

FIGURE 9. Altered estimates of the maximal asymptote, a, and the slope, S, obtained by fitting the Hill equation to logistic data where the parameter K (log K) was constrained to differ from the correct value. The systematic relationship between the error in K and the values of the parameters S and α indicates that each is able to partially correct for error in K and thus are correlated with K.

commonly entered as $10^{LogEC_{50}}$. This reparameterization means that the regression algorithm will actually provide the best-fit estimate of the logarithm of the EC_{50}. Why reparameterize? As mentioned earlier, many of the assumptions of nonlinear regression rely on a Gaussian distribution of experimental uncertainties. Many model parameters, including the EC_{50} of the Hill equation, the dissociation constant of a hyperbolic radioligand binding equation, and the τ parameter of the operational model, follow an approximately Gaussian distribution only when transformed into logarithms.[17] Thus, although not particularly important for the estimation of the parametric value, reparameterization can improve the validity of statistical inferences made from nonlinear regression algorithms.[13] Other examples of reparameterizations that can increase the statistical reliability of the estimation procedure include recasting time parameters as reciprocals and counts of radioactive decay as square roots.[5]

B. Assessing the Quality of the Fit

The final determination of how "appropriate" the fit of a dataset is to a model will always depend on a number of factors, including the degree of rigor the researcher actually requires. Curve fitting for the determination of standard curves, for instance, will not warrant the same diagnostic criteria one may apply to a curve fit of an experimental dataset that was designed to investigate a specific biological mechanism. In the case of standard curves, an eyeball inspection of the curve superimposed on the data is usually sufficient to indicate the reliability of the fit for that specific purpose. However, when the fitting of models to experimental data is used to provide insight into underlying biological mechanisms, the ability to ascribe a high degree of appropriateness to the resulting curve fit becomes paramount.

1. Inspection

Although usually sufficient for empirical models, an initial test for conformity of the data to any selected model is a simple inspection of the curve fit superimposed on the data. Although rudimentary, this procedure is quite useful in highlighting *really bad* curve fits, i.e., those that are almost invariably the consequence of having inadvertently entered the wrong equation or setting certain parameter values to a constant value when they should have been allowed to vary as part of the fitting process. Assuming that visual inspection does not indicate a glaring inconsistency of the model with the data, there are a number of statistical procedures that can be used to quantify the goodness of the fit.

2. Root Mean Square

Figure 10 shows a schematic of an experimental dataset consisting of 6 observations (open circles labeled obs1 – obs6) and the superimposed best-fit of a sigmoidal concentration–response model [Equation (15)] to the data. The solid circles (exp1 – exp6) represent the expected response corresponding to each X-value used for the determination of obs1 – obs6, derived from the model fit. The sum of the squared residuals, i.e., the sum of the squared differences between the observed and expected responses has also been defined as the *Error Sum of Squares* (SSE), and it is this quantity that most researchers think of when discussing the sum-of-squares derived from their curve fitting exercises [see Equation (13)]:

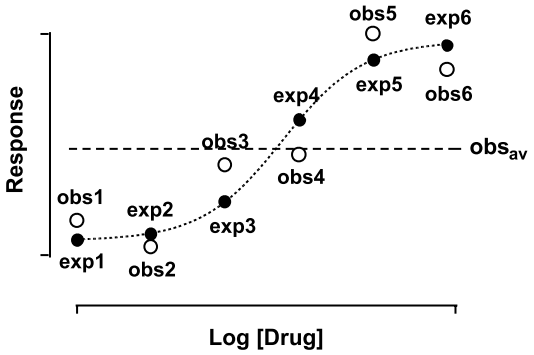

FIGURE 10. Relationship between a set of experimental observations (open circles; obs1 – obs6) and their corresponding least squares estimates (solid circles; exp1 – exp6). The horizontal dashed line represents the average of all the experimental observations (obs_{av}).

$$SSE = (obs1 - exp1)^2 + (obs2 - exp2)^2 + \ldots + (obs6 - exp6)^2 \quad (17)$$

The SSE is sometimes used as an index of goodness-of-fit; the smaller the value, the better the fit. However, in order to use this quantity more effectively, an allowance must also be made for the "degrees of freedom" of the curve fit. For regression procedures, the degrees of freedom equal the total number of datapoints minus the number of model parameters that are estimated. In general, the more parameters that are added to a model, the greater the likelihood of observing a very close fit of the regression curve to the data, and thus a smaller SSE. However, this comes at the cost of degrees of freedom. The "mean square error" (MSE) is defined as the SSE divided by the degrees of freedom (df):

$$MSE = \frac{SSE}{df} \quad (18)$$

Finally, the square root of MSE is equal to the root mean square, RMS:

$$RMS = \sqrt{\frac{SSE}{df}} \quad (19)$$

The RMS (sometimes referred to as $S_{y.x}$) is a measure of the standard deviation of the residuals. It should be noted, however, that although RMS is referred to as the "standard deviation" or "standard error" of the model, this should not be confused with the standard deviation or error associated with the individual parameter estimates. The degree of uncertainty associated with any model parameter is derived by other methods (see below).

3. R^2 (Coefficient of Determination)

Perhaps more common than the RMS, the R^2 value is often used as a measure of goodness of fit. Like the r^2 value from linear regression or correlation analyses, the value of R^2 can range from 0 to 1; the closer to 1 this value is, the closer the model fits the dataset. To understand the derivation of R2, it is important to first appreciate the other "flavors" of sums-of-squares that crop up in the mathematics of regression procedures in addition to the well-known SSE.

Using Figure 10 again as an example, the sum of the squared differences between each observed response and the average of all responses (obsav) is defined as the *Total Sum of Squares* (SST; sometimes denoted as Syy):

$$SST = (obs1 - obs_{av})^2 + (obs2 - obs_{av})^2 + \ldots + (obs6 - obs_{av})^2 \quad (20)$$

where

obs_{av} = (obs1 + obs2 + obs3 + obs4 + obs5 + obs6)/6 (21)

The sum of the squared differences between each estimated (expected) response, based on the model, and the average of all observed responses is defined as the Regression Sum of Squares (SSR):

SSR = (exp1 – obs_{av})2 + (exp2 – obs_{av})2 + … + (exp6 – obs_{av})2 (22)

The total sum of squares, SST, is equal to the sum of SSR and SSE, and the goal of regression procedures is to minimize SSE (and, as a consequence, SST).

Using the definitions outlined above, the value of R^2 can be calculated as follows:[5,10]

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \quad (23)$$

R^2 is the proportion of the adjusted variance in the dependent variables that is attributed to (or explained by) the estimated regression model. Although useful, the R^2 value is often overinterpreted or overutilized as the main factor in the determination of goodness of fit. In general, the more parameters that are added to the model, the closer R^2 will approach a value of 1. It is simply an index of how close the datapoints come to the regression curve, not necessarily an index of the correctness of the model, so while R^2 may be used as a starting point in the assessment of goodness of fit, it should be used in conjunction with other criteria.

4. Analysis of Residuals

Because the goal of least squares regression procedures is to minimize the sum of the squares of the residuals, it is not surprising that methods are available for analyzing the final residuals in order to assess the conformity of the chosen model to the dataset. The most common analysis of residuals relies on the construction of a scatter diagram of the residuals.[13,20] Residuals are usually plotted as a function of the values of the independent variable. If the model is adequate in describing the behavior of the data, then the residuals plot should show a random scatter of positive and negative residuals about the regression line. If, however, there is a systematic deviation of the data from the model, then the residuals plot will show nonrandom clustering of positive and negative residuals. Figure 11 illustrates this with an example of a radioligand competition binding experiment. When the data are fitted to a model of binding to a single site, a systematic deviation of the points from the regression curve is manifested as clustering in the residuals plot. In contrast, when the same dataset is fitted to a model of binding to two sites, a random scatter of the residuals about the regression line indicates a better fit of the second model. This type of residual analysis is made more quantitative when used in conjunction with the "runs test" (see below).

There are many other methods of performing detailed analyses of residuals in addition to the common method described above. These methods include cumulative probability distributions of residuals, χ^2 tests, and a variety of tests for serial correlation.[7,10,11,20]

5. The Runs Test

The runs test is used for quantifying trends in residuals, and thus is an additional measure of systematic deviations of the model from the data. A "run" is a consecu-

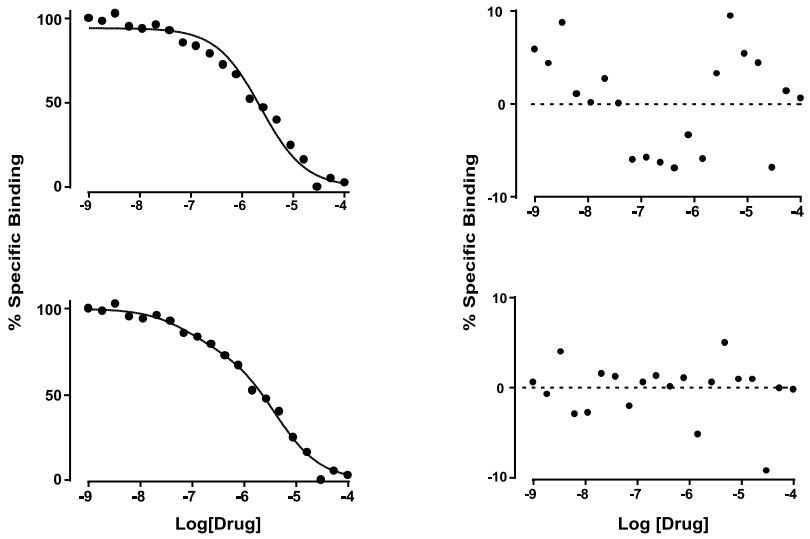

FIGURE 11. An example of residuals plots. The top panel represents a curve fit based on a one binding site model to a data set obtained from a radioligand competition binding assay (left) and its corresponding residuals plot (right). Note the clustering of positive and negative residuals. The bottom panel represents a curve fit based on a two binding site model to the same dataset (left) and its corresponding residuals plot (right). Note the random scatter of positive and negative residuals in this case.

tive series of residuals of the same sign (positive or negative). The runs test involves a calculation of the expected number of runs, given the total number of residuals and expected variance.[20] The test uses the following two formulae:

$$\text{Expected Runs} = \frac{2N_pN_n}{N_p + N_n} + 1 \qquad (24)$$

$$\text{Expected Variance} = \frac{2N_pN_n(2N_pN_n - N_p - N_n)}{(N_p + N_n)^2(N_p + N_n - 1)} \qquad (25)$$

where N_p and N_n denote the total number of positive and negative residuals, respectively. The results are used in the determination of a *P* value.[5,13] A low *P* value indicates a systematic deviation of the model from the data. In the example shown in Figure 11, the one-site model fit was associated with a P value of less than 0.01 (11 runs expected, 4 observed), whereas the two-site model gave a *P* value of 0.4 (10 runs expected, 9 observed).

C. Optimizing the Fit

With the ubiquitous availability of powerful computers on most desktops, the impressive convergence speed of modern curve fitting programs can often lead to a false sense of security regarding the reliability of the resulting fit. Assuming that the appropriate model has been chosen, there are still a number of matters the biomedical investigator must take into account in order to ensure that the curve fitting procedure will be optimal for their dataset.

1. Data Transformations

Most standard regression techniques assume a Gaussian distribution of experimental uncertainties and also assume that any errors in Y and X are independent. As mentioned earlier, however, these assumptions are not always valid. In particular, the variance in the experimental dataset can be

heteroscedastic, that is, it changes in a systematic fashion with the variables. One method for optimizing the curve fitting process to adjust for heteroscedastic errors is to weight the data, as discussed earlier, while another approach is to transform the data to a form where the errors become more homoscedastic prior to the application of the regression technique. Transformations such as the square root or logarithm of the dependent or independent variables do not necessarily cause any problems of their own, provided they reduce rather than increase any heteroscedasticity in the data. In contrast, classical "linearising" transformations, where a new variable is derived from both the original dependent and independent variables, are quite dangerous and it is unfortunate that they are still common practice in some laboratories. Indiscriminate data transforms of the latter kind are troublesome because they have the potential of distorting homoscedastic errors in experimental uncertainties and thus violating the assumptions of any subsequent regression procedure. Transforms are appropriate if they have a normalizing effect on heteroscedastic errors; they are *not* valid otherwise. In addition, some data transforms, embodied in reciprocal plots (e.g., Lineweaver-Burk) or the Scatchard transformation, violate the assumption of independence between X and Y variables and are equally inappropriate. In contrast, transformation of model parameters (as described earlier) may often have an optimising effect on the fitting procedure.

2. Initial Estimates

All curve fitting algorithms require the specification of initial estimates of the parameters that are then optimized to yield the best fit. No regression algorithm is perfect, and failure to specify reasonable parameter estimates may result in a failure of the algorithm to converge or, more insidiously, a convergence of the curve fit on a "local minimum." If we recall our earlier discussion of the surface of the merit function that the various algorithms travel down, it is possible to envisage a multiparameter model that results in a series of troughs such that the algorithm may settle in one as if it has converged on the best fit when, in fact, a deeper trough is available elsewhere on the merit function surface. This is an example of the program converging on a local minimum (Figure 12), where the curve fit is not optimal although the user may think that the best fit has been obtained. The best safeguard against this problem is to perform the regression analysis a number of times using different initial estimates. A well-behaved model should converge on essentially the same final estimates each time.

Some commercial programs make the process of finding initial parameter estimates relatively painless by incorporating approximate rules that find initial estimates for the user. Although this is expedient, there is no substitute for the researcher personally addressing the issue of initial parameter estimates. This forces one to focus on the underlying model and the meaning of the model parameters, and it is then not too difficult to come up with a best guess. If further assistance is required, or if there are some parameters that the user does not have a particular "feel" for, then a simplex algorithm or a Monte Carlo-based algorithm (see below) may be utilized to derive estimates that can subsequently be improved upon by the more standard derivative-based algorithms.

D. Reliability of Parameter Estimates

The determination of the reliability of the estimated parameters derived from a curve fit is as important as the actual esti-

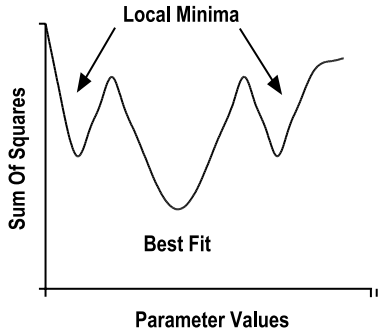

FIGURE 12. Multiple minima in parameter space. The best fit is obtained at that set of parameter values yielding the smallest possible sum of squares. Depending on the initial estimates, however, the fitting algorithm may converge on parameter sets which, although yielding a reduced sum of squares, do not correspond to the minimum possible sum of squares. The regression is then said to have converged on a "local minimum."

mation of the parametric values themselves. All known methods for the calculation of standard errors and confidence intervals from regression algorithms are based on the mathematics of linear models. Since nonlinear models are more common in biology than linear models, it is perhaps disheartening to have to accept that there are no exact theories for the evaluation of parametric errors in nonlinear regression. However, there are a number of procedures available for approximating these errors such that, in most practical applications, a reasonable measure of parameter error is obtained.

1. Number of Datapoints

The number of experimental datapoints collected and analyzed will play a crucial role in the curve fitting process in one (or both) of two ways:

a. Determination of the appropriateness of the model.
b. Determination of the accuracy of the parameter estimates.

Different measures for goodness-of-fit have already been covered, but some discussion on the influence of datapoint number is also warranted at this point, since it can form an important component of choosing the right model, that adequately accounts for the data. Figure 13 illustrates the effect of datapoint number on one of the most common statistical procedures utilized in discriminating between variants of the same model, i.e., the "*F*-test" (or "extra-sum-of-squares" test). The actual test is described in greater detail in the next section. For now, it is sufficient to point out that the *F*-test relies heavily on the degrees of freedom associated with the fit to any model, which are in turn dependent on the number of datapoints minus the number of parameters estimated. Although all the points in each of the panels in Figure 13 are taken from the same simulated dataset, the "correct" model (a two binding site model) can only be statistically resolved when the datapoints were increased from 6 (panel A) or 10 (panel B), to 20 (panel C).

Assuming that the researcher has *a priori* reasons for deciding that a particular model is most appropriate under their circumstances, the number of datapoints will still be crucial in determining the accuracy of the parameters based on that model. Table 1 lists the parameter estimates and corresponding 95% confidence intervals of a two binding site model (i.e., the correct model) applied to the datasets

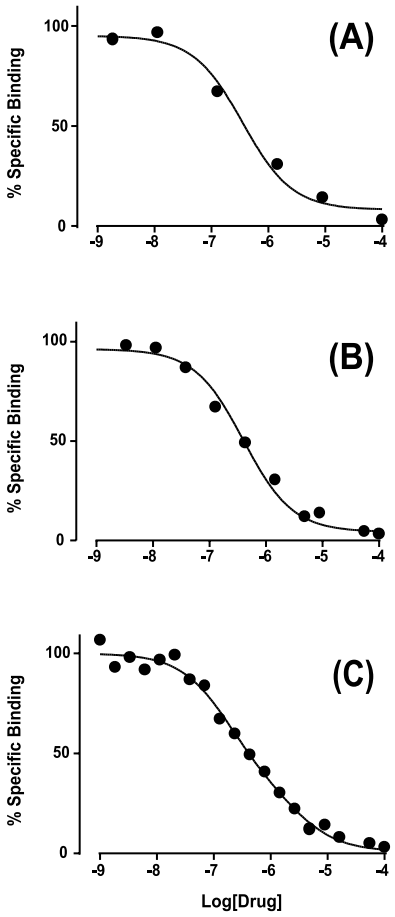

FIGURE 13. Influence of data point number on choice of model. The radioligand competition binding curves above were simulated (with random error) according to a model for binding to two-sites. The sampling of points in each of the panels is from exactly the same simulated dataset. The curves in each panel are the least squares fit of the data to either a one- or two-site binding model, as determined by an F-test (see Section VI. E). Panels A (6 points) and B (10 points) were not statistically significant from a one-site model. Only in panel C (20 points) were the data able to be statistically resolved into the (correct) two-site model fit.

of Panel A and Panel C, respectively, of Figure 13. Although the final parameter estimates appear comparable in each instance, the fit based on the small number of datapoints is associated with unacceptably large confidence intervals. There are simply not enough points to accurately define all the parameters of the model. In contrast, increasing the number of datapoints to 20 allowed for reasonable estimates of the error associated with each parameter estimate. The confidence intervals reported in the table were calculated from the asymptotic standard errors derived by the computer program from the fitting algorithm and are most likely *underestimates* of the true error (see below), thus rendering our (already shaken) confidence in the accuracy

of minimal-data-point parameter estimates virtually nonexistent. There have been some methods presented in the literature for maximizing the reliability of parameter estimates under conditions of minimal datapoint number (e.g., References 21 and 22), but there really is no substitute for a good sampling of experimental datapoints.

Table 1
Parameter Estimates and Associated Confidence Intervals from Fitting a Two-Site Model of Radioligand Competition Binding to Different Data Point Numbers Taken from the Same Dataset (Panels A and C; Figure 13)

Parameter	Estimate	95% Confidence Interval
	Datapoints = 6	
Maximum Asymptote[a]	–96.7	45.6 to 148.3
Minimum Asymptote[b]	1.3	–84.8 to 87.56
Log IC50High[c]	–6.7	–9.93 to –3.53
Log IC50Low[d]	–5.1	–16.1 to 5.9
Fraction High[e]	0.74	–1.4 to 2.9
	Datapoints = 20	
Maximum Asymptote	99.9	95.4 to 104.4
Minimum Asymptote	0.9	–5.5 to 7.4
Log IC_{50High}	–6.8	–7.3 to –6.5
Log IC_{50Low}	–5.6	–6.4 to –4.6
Fraction High	0.64	0.4 to 0.8

[a] Y-axis value in the absence of competing drug.
[b] Y-axis value in the presence of saturating concentrations of competing drug.
[c] Potency estimate for competition at the high affinity binding site.
[d] Potency estimate for competition at the low affinity binding site.
[e] Fraction of high affinity binding sites.

2. Parameter Variance Estimates from Repeated Experiments

The most straightforward and conservative approach to building up an error profile of a given parameter is to simply repeat the same experiment many times, obtain single parameter estimates from each individual curve fit, and then derive the mean and standard deviation (and error) of the parameters using standard textbook methods. Assuming that each curve fit is performed under optimal conditions, e.g., appropriate number of datapoints, appropriate transformation and weighting, etc., biomedical research is still fraught with small overall sample sizes; it is not uncommon to see n = 3 – 6 given in many publications as the number of times an experiment is repeated. As such, the conservative, albeit straightforward approach to parameter error estimation just described may not have the power to resolve small differences between experimental treatments, as it is based on small sample sizes and, furthermore, does not utilize all the available datapoints. The remaining methods for parameter error estimation utilize all the datapoints in some form or other.

3. Parameter Variance Estimates from Asymptotic Standard Errors

The standard errors reported by practically all commercially available least squares regression programs fall under this category. Asymptotic standard errors are computationally the easiest to determine and, perhaps not surprisingly, the least accurate. In most instances, these standard errors will underestimate the true error that is likely to be associated with the parameter of interest.

The calculation of the asymptotic standard error and associated confidence inter-

vals involves matrix algebra, but may be summarized as follows:[23]

1. Determine the *Hessian (or "information") matrix.* This is the matrix containing the second derivatives of the parameters with respect to the minimized χ^2 merit function.
2. Evaluate the *variance-covariance matrix* by multiplying the inverse of the Hessian matrix by the variance of the residuals of the curve fit.
3. The diagonal elements of the resulting variance-covariance matrix are the squares of the asymptotic standard errors; the off-diagonal elements of the matrix are the *covariances* of the parameters, and are a measure of the extent to which the parameters in the model are correlated with one another.

The computer program then reports the resulting standard errors. For these errors to actually be a good measure of the accuracy of the parameter estimates, the following assumptions must hold:[23]

a. The fitting equation is linear.
b. The number of datapoints is very large.
c. The covariance terms in the variance-covariance matrix are negligible.

For nonlinear models, the first assumption is invalid, however, the impact of failure to conform to this assumption may be lessened for models that are well behaved, e.g., contain conditionally linear parameters or can be approximated by linear functions. The second assumption can also be reasonable provided the experimenter is able to ensure an adequate sampling of datapoints. Unfortunately, the third assumption is almost never realized. As described earlier, most parameters in nonlinear models show some degree of correlation with one another; indeed, high correlations are indicative of parameter redundancies in the model. As such, ignoring the covariances from the variance-covariance matrix in the reporting of parameter errors will underestimate the true error.

Nevertheless, asymptotic standard errors may serve a useful diagnostic role. Since they will invariably be underestimates of the true error, very large standard errors or confidence intervals reported after a curve fit are indicative of a very poor fit of the associated parameter (see Table 1). This may occur, for instance, because the parameter is ill defined by the available data.

4. Monte Carlo Methods

The most reliable method for the determination and validation of model parameter confidence intervals is also the most computer-intensive. Monte Carlo simulations involve the generation of multiple (hundreds to thousands) of pseudodatasets, based on a chosen model, and the subsequent analysis of the simulated datasets with the same model used to generate them followed by construction of a frequency histogram showing the distribution of parameter estimates.[17,24] Figure 14 shows a flowchart summarizing the general approach to Monte Carlo simulation.

The crucial factor in the implementation of the Monte Carlo approach is the ability to add random "error" to the pseudodataset points that accurately reflects the distribution of experimental uncertainties associated with the determination of "real" datasets. The best determinant of this error is the variance of the fit of the chosen model to real experimental data, provided that the standard assumptions underlying least squares regression analyses are valid. In addition to the appropriate choice of variance for the simulations, other key features in this approach are the choice and the number of independent variables, which again should match those determined in a typical experiment.

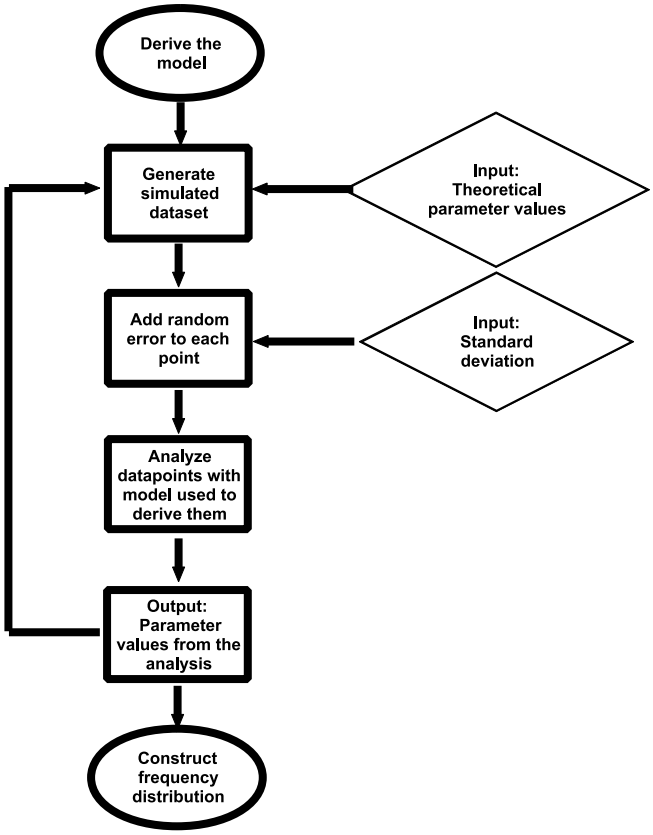

FIGURE 14. A general approach to Monte Carlo simulation.

The beauty of the Monte Carlo approach is that the level of accuracy with regard to the confidence interval profiles is very much in the hands of the researcher; the greater the number of simulated datasets, the greater the resolution of the confidence intervals. However, this comes at the expense of computer time; a Monte Carlo simulation of 1000 datasets may take 1000 times longer than a least squares fit of the actual experimental dataset used to pattern the simulations. Coupled with the fact that many commercially available curve fitting packages do not contain Monte Carlo-compatible programming features, the time factor involved in generating parameter confidence intervals from the Monte Carlo approach dissuades many researchers from routinely using this method. Nonetheless, great insight can be gained from Monte Carlo approaches. For instance, in addition to providing the greatest degree of accuracy in parameter error estimation, Monte Carlo methods can also guide the experimenter toward the most appropriate model reparameterizations in order to optimize the actual curve fitting procedure.[17]

One potential problem with the standard Monte Carlo approach is that it is necessary to define the population distributions for the errors applied to the datapoints. A normal distribution is most commonly used, but it is not always clear that it is appropriate. The bootstrap, described below, explicitly overcomes that problem.

5. The Bootstrap

"Bootstrapping" is an oddly-named process that allows an approximate reconstruction of the parameters of the population from which the data have been (at least conceptually)

sampled.[25] Bootstrapping differs from standard Monte Carlo methods in that it makes no assumption about the form of the population, and instead assumes that the best estimate of the properties of the population is the experimentally determined dataset. The population is reconstructed by repeated resampling of the datapoints to give a large number (hundreds or even thousands) of new pseudodatasets. The resampling is done "with replacement," which is to say that any particular real datapoint can appear in each pseudodataset more than one time. The result is a population of pseudodatasets that represents a pseudo-population that has approximately the same properties as the original population.

Bootstrapping can be used in several ways relevant to model fitting. First, it can provide a pseudopopulation of any parameter calculable from each pseudodataset. Thus it can be used to give confidence intervals for fitted parameters obtained from methods that do not directly provide estimates of parameter variance, such as the simplex method. Similarly, it has been used to estimate the reliability of the variance estimates obtained from other methods that rely on the covariance matrix.[19]

Bootstrapping is not without potential problems. One arises from the fact that the real dataset is unlikely to include any samples from the extreme tails of the overall population of possible datapoints. This means that bootstrapped populations generally have less area under the extreme tails than the real population from which the data were sampled. There are corrections that can be applied,[25] but bootstrapping is not universally accepted by statisticians.

6. Grid Search Methods

Another computer-intensive approach to error determination involves the construction of multidimensional grids based on model parameter values and then "searching" for those parameter value combinations where the variance of the overall fit increases significantly. The confidence intervals are then defined as those regions of the grid (which resemble a multidimensional ellipsoid) that surround the minimum over which the variance does not change significantly.[8,23]

7. Evaluation of Joint Confidence Intervals

As discussed earlier, the parameters in most models tend to show some correlation with one another. The evaluation of joint confidence intervals is a procedure that is designed to include the covariance of the parameters in the determination of parameter error estimates.[8,13] The equations underlying this approach, however, assume that the fitting equation is linear in order to derive a symmetrical elliptically-shaped confidence interval profile of parameters. Unfortunately, this method yields asymmetric confidence regions for those nonlinear models that cannot approximate to a linear model, and is thus not as reliable as Monte Carlo or Grid search methods.

E. Hypothesis Testing

Often, the desire to ascertain the standard error or confidence interval associated with model parameters is a prelude to the statistical testing of the parameters according to a particular hypothesis. Therefore, some objective statistical test is required in order to allow for comparisons between parameters or comparisons between models.

1. Assessing Changes in a Model Fit between Experimental Treatments

There are three broad approaches to performing statistical comparisons between the same model parameters before and after an experimental treatment. The first relies on the use of standard parametric tests, such as the Student's *t*-test. The second approach relies on more computer-intensive, but preferable comparisons between parameters based on permutation tests. The third approach differs from the other two in that it uses all the experimental data generated before and after a particular treatment in a comparison of global changes in goodness of fit. The last procedure may be summarized as follows.[5,26]

1. Analyze each dataset separately.
2. Sum the SSE resulting from each fit to give a new "total" sum-of-squares value (SS_A). Similarly, sum the two degrees of freedom values from each fit to give a "total" degrees of freedom (df_A).
3. Pool the two sets of data into one large set.
4. Analyze this new "global" dataset to obtain a new sum-of-squares value (SS_B) and degrees of freedom (df_B).
5. Calculate the following *F* ratio:

$$F = \frac{(SS_B - SS_A) / (df_B - df_A)}{SS_A / df_A} \qquad (26)$$

The *F* value is used to obtain a P value, with the numerator having ($df_B - df_A$) degrees of freedom and the denominator having df_A degrees of freedom. A small P value (i.e., large *F* value) indicates that the individual fits are better than the global, pooled fit, i.e., the experimental treatment resulted in a significant difference in the model parameters between the two datasets.

2. Choosing between Models

The *F* ratio can also be used to compare the fit of a single dataset to two different versions of the same model:

$$F = \frac{(SS1 - SS2) / (df1 - df2)}{SS2 / df2} \qquad (27)$$

In this instance, SS1 and df1 are defined as the SSE and degrees of freedom, respectively, of the model with fewer parameters, whereas SS2 and df2 are defined as the SSE and degrees of freedom, respectively, of the model with the greater number of parameters. The addition of more parameters to a model will result in an improvement of the goodness of fit and a reduction in SSE, but at the cost of degrees of freedom. The *F* test [Equation (27)] attempts to quantify whether the loss of degrees of freedom on going from a simpler to a more complicated model is worth the gain in goodness of fit. A low *P* value is indicative of the more complicated model being the statistically better model. It should be noted, however, that the *F* test can only be applied to two different versions of the same model, e.g., a one binding-site versus a two binding-site curve fit. In addition, the *F* test is particularly harsh since it relies so heavily on degrees of freedom and, hence, datapoints and number of parameters. As a consequence, the test may be too conservative and reject the more complicated model for the simpler one, even when this is not the case. Thus, results from the test should be regarded with caution if the number of datapoints is limited and other measures of goodness of fit appear to indicate that the simpler model is not a reasonable fit to the data. When in doubt, repeat

the experiment with greater numbers of datapoints.

VII. FITTING VERSUS SMOOTHING

Throughout this article, the process of fitting empirical or mechanistic models to experimental data has generally been encompassed within the umbrella term "curve fitting." However, some distinctions can be made. *Simulation* refers to the process whereby the properties of the model are examined in order to determine the theoretical consequences of imposing specified conditions on the parameters and variables. The term *fitting* refers to the process whereby the model parameters are altered to discover which set of parameter values best approximate a set of experimental observations derived from the actual system of interest. A special case of the fitting process is the procedure known as *smoothing*, whereby a model is chosen to generate a fit that simply passes near or through all the experimental datapoints in order to act as a guide for the eye.

If the purpose of the curve fitting procedure is simply to smooth or to generate a standard curve for extrapolation, then the nature of the underlying model and accompanying regression technique is not crucial. If, however, the purpose of the curve fitting procedure is to obtain insight into the features of the model that describe an aspect of the biological system of interest, then the choice of model is paramount. Although linear models can give curved lines, (e.g., the polynomial equations described earlier), most biological experiments that yield data described by a curve are probably best analyzed using nonlinear regression. This is because it is much more common to find a nonlinear model that can be related in a meaningful and realistic fashion to the system under study than a general linear model.

VIII. CONCLUSION

Computerized curve fitting has become nearly ubiquitous in the analysis of biomedical research. The ease of use and speed of the modern curve fitting programs encourage researchers to use them routinely for obtaining unbiased parameter estimates where in the not very distant past, they might

Table 2
Selected List of Commercially-Available Curve Fitting Programs and Their Associated Least Squares Algorithms. Distributors are listed in parentheses

Program	Algorithm
Enzfitter (Biosoft)	Levenberg-Marquardt; Simplex
Excel (Microsoft)	Simplex
Fig. P (Biosoft)	Levenberg-Marquardt
Kaleidagraph (Synergy)	Levenberg-Marquardt
KELL (Biosoft)	Levenberg-Marquardt
Origin (Microcal)	Levenberg-Marquardt
Prism (GraphPad)	Levenberg-Marquardt
ProFit (QuantumSoft)	Levenberg-Marquardt; Robust; Monte Carlo (Simplex)
Scientist (Micromath)	Levenberg-Marquardt; Simplex
SigmaPlot (SPSS)	Levenberg-Marquardt

have used eyeballing or linearization processes that would have contained substantial subjective elements and systematic distortions. Nevertheless, indiscriminate use of curve fitting without regard to the underlying features of the model and data is a hazardous approach. We hope that the content of this chapter is useful in illustrating both strengths and some pitfalls of computer-based curve fitting, and some ways to optimize the quality and utility of the parameters so obtained.

IX. SOFTWARE

Table 2 contains a limited sampling of commercially available curve fitting programs. Some of them (e.g., EnzFitter and KELL) are more specialized in their applications than others, but all are commonly applied to curve fitting of biological models to data. Also shown in the table are the associated regression algorithms utilized by each program.

REFERENCES

1. Wells, J.W., Analysis and interpretation of binding at equilibrium, in *Receptor-Ligand Interactions: A Practical Approach*, E.C. Hulme, Ed., Oxford University Press, Oxford, 1992, 289–395.
2. Hill, A.V., The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves, *J. Physiol.*, 40, iv–vii, 1910.
3. Hill, A.V., The combinations of haemoglobin with oxygen and with carbon monoxide. I, *Biochem. J.*, 7, 471–80, 1913.
4. Motulsky, H.J., *Intuitive Biostatistics*, Oxford University Press, New York, 1995.
5. Motulsky, H.J., *Analyzing Data with GraphPad Prism*, GraphPad Software Inc., San Diego, CA, 1999.
6. Ludbrook, J., Comparing methods of measurements, *Clin. Exp. Pharmacol. Physiol.*, 24(2), 193–203, 1997.
7. Bates, D.M. and Watts, D.G., *Nonlinear Regression Analysis and Its Applications*, Wiley and Sons, New York, 1988.
8. Johnson, M.L. and Faunt, L.M., Parameter estimation by least-squares methods, *Methods Enzymol.*, 210, 1–37, 1992.
9. Press, W.H., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P., *Numerical Recipes in C. The Art of Scientific Computing*, Cambridge University Press, Cambridge, MA, 1992.
10. Gunst, R.F. and Mason, R.L., *Regression Analysis and Its Applications: A Data Oriented Approach*, Marcel Dekker, New York, 1980.
11. Cornish-Bowden, A., *Analysis of Enzyme Kinetic Data*, Oxford University Press, New York, 1995.
12. Johnson, M.L., Analysis of ligand-binding data with experimental uncertainties in independent variables, *Methods Enzymol.*, 210, 106–17, 1992.
13. Motulsky, H.J. and Ransnas, L.A., Fitting curves to data using nonlinear regression: a practical and nonmathematical review, *FASEB J.*, 1, 365–74, 1987.
14. Jurs, P.C., Curve fitting, in *Computer Software Applications in Chemistry*, Wiley and Sons Inc., New York, 1996, 25–51.
15. Black, J.W. and Leff, P., Operational models of pharmacological agonism, *Proc. Roy. Soc. (Lond.) B.*, 220, 141–62, 1983.
16. Leff, P., Prentice, D.J., Giles, H., Martin, G.R., and Wood, J., Estimation of agonist affinity and efficacy by direct, operational

model-fitting, *J. Pharm. Methods*, 23, 225–37, 1990.

17. Christopoulos, A., Assessing the distribution of parameters in models of ligand-receptor interaction: to log or not to log, *Trends Pharmacol. Sci*., 19, 351–7, 1998.

18. Parker, R.B. and Waud, D.R., Pharmacological estimation of drug-receptor dissociation constants. Statistical evaluation. I. Agonists, *J. Pharmacol. Exp. Ther*., 177, 1–12, 1971.

19. Lew, M.J. and Angus, J.A., Analysis of competitive agonist-antagonist interactions by nonlinear regression, *Trends Pharmacol. Sci.*, 16(10), 328–37, 1995.

20. Straume, M. and Johnson, M.L., Analysis of residuals: criteria for determining goodness-of-fit, *Methods Enzymol.*, 210, 8–105, 1992.

21. Rovati, G.E., Rodbard, D., and Munson, P.J., DESIGN: computerized optimization of experimental design for estimating K_d and B_{max} in ligand binding experiments. I. Homologous and heterologous binding to one or two classes of sites, *Anal. Biochem*., 174(2), 636–49, 1988.

22. Rovati, G.E., Rodbard, D., and Munson, P.J., DESIGN: computerized optimization of experimental design for estimating K_d and B_{max} in ligand binding experiments. II. Simultaneous analysis of homologous and heterologous competition curves and analysis of "multiligand" dose-response surfaces, *Anal. Biochem*., 184(1), 172–83, 1990.

23. Johnson, M.L., Use of least-squares techniques in biochemistry, *Methods Enzymol.*, 240, 1–22, 1994.

24. Straume, M., Veldhuis, J.D., and Johnson, M.L., Model-independent quantification of measurement error: empirical estimation of discrete variance function profiles based on standard curves, *Methods Enzymol.*, 240, 121–50, 1994.

25. Efron, B. and Tibshirani, R.J., *An Introduction to the Bootstrap.*, Vol. 57. Chapman and Hall, New York, 1993.

26. Ratkowsky, D., *Nonlinear Regression Modelling: A Unified and Practical Approach*, Marcel Dekker, New York, 1983.