Biochem. J. (1976) **157**, 489–492
Printed in Great Britain

Methods for Fitting Equations with Two or More Non-Linear Parameters

By IAN A. NIMMO and GORDON L. ATKINS
Department of Biochemistry, University of Edinburgh Medical School, Teviot Place, Edinburgh EH8 9AG, Scotland, U.K.

(*Received* 23 *April* 1976)

1. Descriptions are given of two ways for fitting non-linear equations by least-squares criteria to experimental data. One depends on solving a set of non-linear simultaneous equations, and the other on Taylor's theorem. 2. It is shown that better parameter estimates result when an equation with two or more non-linear parameters is fitted to all the sets of data simultaneously than when it is fitted to each set in turn.

It is fairly easy to estimate the parameters of an enzyme-catalysed reaction so long as the rate equation is non-linear in only one of them, typically K_m (e.g. Atkins & Nimmo, 1975). The problem becomes more complex when there are other non-linear parameters as well, such as the K_m of a second substrate, the K_i of an inhibitor or the Hill coefficient.

Cleland (1963, 1967) has devised methods to handle the first two of these more complex problems. A rectangular hyperbola is fitted to a set of initial velocities determined with one of the independent variables (e.g. second substrate, inhibitor) held constant, and then the coefficients of the hyperbola are replotted against the second independent variable to give the desired parameters ('replot method'). Alternatively, the rate equation can be fitted directly to all the sets of data at once ('direct method'). Cleland (1963) suggested that the direct method gives the better answers, though he did not prove it.

An inspection of the literature published during 1974 and 1975 shows quite clearly that most enzymologists favour the replot method, even though it is likely to give less reliable answers. Consequently the majority opinion seems to be that either the additional refinement gained from the direct method does not justify the extra effort required to use it, or perhaps the method is too formidable mathematically.

The present paper therefore has two objectives. The first is to establish, by analysing simulated data, exactly how much better the direct method is than the replot one. The second objective is to explain how a non-linear equation can be fitted (by least-squares criteria) to several sets of data simultaneously. One of the ways in which this may be done entails the solution of a set of non-linear simultaneous equations; another, which has been described in detail by Cleland (1967), is based on Taylor's theroem (Courant, 1937*a*).

Theory and Methods

Two sets of experiment were simulated. In the first, initial velocities (v) were measured for a reaction involving a single substrate and a competitive inhibitor, and the relevant rate equation was fitted to all the sets of data simultaneously by solving a set of non-linear simultaneous equations. In the second sort of experiment the entire progress curve was followed for a reaction in which a single substrate was converted into a product that was also a competitive inhibitor with respect to the substrate. [This is essentially the approach used by Newman (1974) to characterize human erythrocyte acetylcholine hydrolase, EC 3.1.1.7.] The progress curves were fitted by a technique based on Taylor's theorem (Courant, 1937*a*). In both sorts of experiment there are three parameters to be estimated: K_m, K_i (both of which are non-linear) and V (which is linear for initial velocities and non-linear for progress curves).

Initial velocities

The perfect data were identical with those in Fig. 3 of Cleland (1967), except that there were six observations at each of the four concentrations of inhibitor. An experiment was simulated by adding to each of the 24 perfect initial velocities one of a series of normally distributed pseudo-random numbers of mean zero and standard deviation either 0.01 or 0.05 (generated by the Edinburgh Regional Computing Centre library program Random). The variance of v was therefore constant, and at the lower error level the coefficient of variation ranged from 6.0 to 1.2%. At each error level 33 experiments were simulated.

The parameters were first estimated by using replots, exactly as described by Cleland (1963). Then

they were estimated directly, i.e. by using least-squares criteria to fit to all the data the equation:

$$v = \frac{V \cdot s}{\left(1+\frac{i}{K_i}\right)K_m + s} \quad (1)$$

where s is the concentration of substrate, and i that of inhibitor. This involves writing down the function for the weighted sum of squares of deviations (SS):

$$SS = \sum w\left(v - \frac{V \cdot s}{\left(1+\frac{i}{K_i}\right)K_m + s}\right)^2$$

partially differentiating it with respect to K_m, K_i and V, and then setting the three partial differentials to zero. (The weighting factor for each point, w, is the reciprocal of its variance.) Here:

$$\frac{\partial SS}{\partial K_m} = V\sum w \cdot \frac{s^2\left(1+\frac{i}{K_i}\right)}{D^3} - \sum w \cdot \frac{v \cdot s\left(1+\frac{i}{K_i}\right)}{D^2} = 0$$

$$\frac{\partial SS}{\partial K_i} = V\sum w \cdot \frac{s^2 \cdot i}{D^3} - \sum w \cdot \frac{v \cdot s \cdot i}{D^2} = 0$$

$$\frac{\partial SS}{\partial V} = V\sum w \cdot \frac{s^2}{D^2} - \sum w \cdot \frac{v \cdot s}{D} = 0$$

where $D = (1+i/K_i)K_m + s$ and $w = 1$ (because the variance of v is assumed to be constant). Any one of the equations can be rearranged to give an expression for V in terms of K_m, K_i and the data, which can be substituted into the other two equations to turn them into expressions in K_m and K_i only. These equations cannot be solved algebraically for K_m and K_i, but they can be solved numerically by any one of a number of iterative techniques. In general such techniques require initial estimates of the unknowns (i.e. of K_m and K_i), which they then refine by successive approximations; several of them are available from computer-program libraries as routines which can be incorporated into larger programs (e.g. in the United Kingdom there are two suitable routines in the Numerical Algorithms Group library, and another two in the Harwell one). We chose the Edinburgh Regional Computing Centre library's routine Imp Davden, which uses the method of Davidenko as described by Broyden (1969). As initial estimates, we used the answers given by the replot method.

Progress curves

When the product of a reaction is a competitive inhibitor with respect to the substrate, the rate equation is:

$$\frac{dp}{dt} = \frac{V \cdot (s_0 - p)}{\left(1+\frac{p}{K_i}\right)K_m + (s_0 - p)} \quad (2)$$

where p is the concentration of product at time t, and s_0 is the initial concentration of substrate (i.e. when $t = p = 0$). The equation of the progress curve, derived by integrating eqn. (2), is:

$$V \cdot t = \left(1-\frac{K_m}{K_i}\right)p - K_m\left(1+\frac{s_0}{K_i}\right) \cdot \ln\left(1-\frac{p}{s_0}\right) \quad (3)$$

The perfect data were formed by setting $K_m = V = 1$, $K_i = 0.01$, 0.1, 1.0, 10 or 100, and $s_0 = 0.5$, 2.2, 10, 44.7 or 200, and then by using eqn. (3) to calculate t at 31 values of p in the range $0.05\,s_0 - 0.95\,s_0$, such that the values of t were roughly equispaced. An experimental progress curve was simulated by adding to each of the 31 perfect values of p one of a series of normally distributed pseudo-random numbers of mean zero and S.D. = 0.01; the variance of p was therefore constant. In all, 25 curves were simulated for each pair of values of K_i and s_0.

Eqn. (3) could easily have been fitted to these curves by solving a set of non-linear simultaneous equations, as described above. However, for the purposes of illustration, Taylor's theorem was used instead.

The parameters were first estimated by the equivalent of the replot method. This entails using the iterative method of Fernley (1974) to fit eqn. (4) to a set of five progress curves, each determined at a different substrate concentration:

$$V' \cdot t = p - K_m' \cdot \ln\left(1-\frac{p}{s_0}\right) \quad (4)$$

The result is five apparent values of K_m and V (K_m' and V' respectively), together with their variances. Comparison of eqn. (4) with eqn. (3) shows that K_m' depends on s_0, and that K_m and K_i can be derived from the coefficients of the linear regression of K_m on s_0 (the values of K_m' should be weighted in inverse proportion to their variances). Similarly V can be derived from the (weighted) mean of V' and these regression coefficients.

The three parameters were also estimated directly by fitting eqn. (3) to all five progress curves at once. This was achieved by an extension of Fernley's (1974) method, and is based on Taylor's theorem [see Cleland (1967) and Wilkinson (1961) for lucid descriptions of the principles involved].

The method is as follows. Eqn. (3) can be thought of as a function relating p to t, s_0 and the unknowns K_m, K_i and V:

$$p = f(t, s_0, K_m, K_i, V)$$

If K_m^0, K_i^0 and V^0 are provisional estimates of the parameters (e.g. those found from replots), then Taylor's theorem states that:

$$p \approx f(t, s_0, K_m^0, K_i^0, V^0) + \Delta K_m \cdot \frac{\partial f}{\partial K_m} + \Delta K_i \cdot \frac{\partial f}{\partial K_i} + \Delta V \cdot \frac{\partial f}{\partial V} \quad (5)$$

where ΔK_m, ΔK_i and ΔV are the corrections that have to be added to K_m^0, K_i^0 and V^0 to give the improved estimates. Putting $\hat{p} = f(t, s_0, K_m^0, K_i^0, V^0)$, eqn. (5) becomes:

$$p - \hat{p} \approx \Delta K_m \cdot \frac{\partial f}{\partial K_m} + \Delta K_i \cdot \frac{\partial f}{\partial K_i} + \Delta V \cdot \frac{\partial f}{\partial V}$$

where

$$\frac{\partial f}{\partial K_m} = \frac{\frac{p}{K_i^0} + \left(1 + \frac{s_0}{K_i^0}\right) \cdot \ln\left(1 - \frac{p}{s_0}\right)}{D}$$

$$\frac{\partial f}{\partial K_i} = \frac{-\frac{K_m^0}{(K_i^0)^2} \cdot \left[p + s_0 \cdot \ln\left(1 - \frac{p}{s_0}\right)\right]}{D}$$

$$\frac{\partial f}{\partial V} = \frac{t}{D}$$

and

$$D = \left(1 - \frac{K_m^0}{K_i^0}\right) + \left(\frac{K_m}{1 - \frac{p}{s_0}}\right) \cdot \left(\frac{1}{s_0} + \frac{1}{K_i^0}\right)$$

The quantity $\hat{p}$ can easily be found from eqn. (3) by using the Newton–Raphson process (Courant, 1937*b*) with p as the starting value. The multiple linear regression (without a constant term) of $(p - \hat{p})$ on the three partial derivatives then gives the corrections to be added to the parameters. The whole process is repeated until they become vanishingly small, which implies that the best-fit estimates have been reached.

Results

Initial velocities

Table 1 gives the means and standard deviations of the three parameters (K_m, K_i and V) and of the SS that were calculated from the data sets with the greater error by the replot and the direct methods. The results for the data sets with the lower error were qualitatively the same and have not been tabulated. The direct method gave on average the lower SS (in fact in every instance its SS was the lower), and its estimates of K_m, V and (especially) K_i were the more precise. The relative values of the standard deviations are about 0.70, 0.72 and 0.42 for K_m, V and K_i respectively.

Progress curves

Eqn. (4) could not be fitted to any of the error-containing progress curves for which $K_i \leqslant 1.0$, $K_i = 10$ and $s_0 \geqslant 10$, or $K_i = 100$ and $s_0 = 200$ (this was because the SS did not converge to a minimum). Consequently the only curves examined had $K_i = 100$ and $s_0 = 0.5$–44.7. Even with these curves realistic estimates were not obtained in four instances with the replot method and three with the direct one. The results (Table 1) are qualitatively similar to those for initial velocities. The relative standard deviations are about 0.50, 0.18 and 0.52 for K_m, V and K_i respectively. Both the direct and the replot methods gave large coefficients of variation for the mean value of K_i, which suggests that the individual estimates of this parameter at least were not normally distributed.

Table 1. *Estimates of K_m, K_i and V determined by the replot and direct-fit methods*

Values are means±S.D. n = number of data sets successfully analysed. For the initial-velocity data the perfect values are $K_m = K_i = V = 1.00$. For the progress curves they are $K_m = V = 1.00$, $K_i = 100$. None of the means are significantly different from their perfect values (P < about 0.05, calculated from interval estimate of median; Campbell, 1967).

	Initial velocities		Progress curves	
	Replot	Direct	Replot	Direct
K_m	1.03±0.20	1.05±0.14	0.96±0.07	1.00±0.03
K_i	1.14±0.44	1.06±0.18	67±103	103±54
V	0.98±0.07	1.01±0.05	1.00±0.03	1.00±0.01
$10^2 \times$SS	8.54±4.55	5.71±1.65	—	—
n	33	33	21	22

Discussion

The main conclusion to be drawn from these results might have been expected intuitively, namely that the direct method (in which all the data sets are fitted simultaneously) gives a closer fit, as judged by the magnitude of the SS, and more precise parameter estimates than does the replot method (in which they are fitted one at a time). Atkins (1973) came to the same conclusion about methods for fitting the Hill equation. It seems reasonable to generalize it to other situations as well: for instance, when the reaction involves two substrates or a different sort of inhibitor. Our results also suggest by how much the use of the direct method is likely to improve the precision of the parameter estimates, and are intended to help enzymologists decide whether the increase in precision is worth the extra computational effort involved.

To show that it is actually quite easy to fit a non-linear equation to data by the method of least squares we deliberately used two of the possible approaches. [Although one of them has been expounded by Cleland (1967), it does not seem to have

found general acceptance and consequently has been recapitulated here.] Neither approach is particularly difficult mathematically, requiring only familiarity with partial differentiation and either access to a routine for solving non-linear simultaneous equations (as illustrated in the initial-velocity problem) or a knowledge of multiple linear regression (as in fitting the progress curves). However, as both approaches depend on iterative techniques the calculations are tedious and are best handled by a digital computer. We consider that it is often more efficient in terms of computer time and space to derive specific solutions to specific problems in this way, rather than to use more general programmes such as those of Cleland (1967) or Nelder & Mead (1965). We also think that, although it is possible to calculate standard errors for the parameters fitted by all the above methods, they are of little value, as they cannot be interpreted rigorously. Naturally all these least-squares direct-fit methods should converge to the same answers. In our experience the rapidity with which they do so depends on factors such as the particular problem under consideration, the number of data points and the amount of error, so that it is difficult to say at the outset which will be the fastest.

A by-product of the simulations is the demonstration that progress curves analysed as described above are unlikely to be of much use in determining the parameters of an enzymic reaction whose product is a competitive inhibitor, unless K_i is much larger than K_m. [One such reaction which seems to fulfil this criterion is the hydrolysis of acetylcholine by human erythrocyte acetylcholine hydrolase (Newman, 1974).] However, if K_i were less than K_m, the analysis of Philo & Selwyn (1973) could be used instead.

We thank Miss Caroline Thompson for her inestimable assistance.

References

Atkins, G. L. (1973) *Eur. J. Biochem.* **33**, 175–180
Atkins, G. L. & Nimmo, I. A. (1975) *Biochem. J.* **149**, 775–777
Broyden, C. G. (1969) *Comput. J.* **12**, 94–99
Campbell, R. C. (1967) *Statistics for Biologists*, p. 34, Cambridge University Press, Cambridge
Cleland, W. W. (1963) *Nature (London)* **198**, 463–465
Cleland, W. W. (1967) *Adv. Enzymol. Relat. Areas Mol. Biol.* **29**, 1–32
Courant, R. (1937*a*) *Differential and Integral Calculus*, vol. 1, 2nd edn., pp. 320–325, Blackie and Son, London and Glasgow
Courant, R. (1937*b*) *Differential and Integral Calculus*, vol. 1, 2nd edn., pp. 335–357, Blackie and Son, London and Glasgow
Fernley, H. N. (1974) *Eur. J. Biochem.* **43**, 377–378
Nelder, J. A. & Mead, R. (1965) *Comput. J.* **7**, 308–313
Newman, P. F. J. (1974) M.Phil. Thesis, University of Edinburgh
Philo, R. D. & Selwyn, M. J. (1973) *Biochem. J.* **135**, 525–530
Wilkinson, G. N. (1961) *Biochem. J.* **80**, 324–332