Biochem. J. (1978) **171**, 513–517
Printed in Great Britain

What Happens when Data are Fitted to the Wrong Equation?

By KEITH J. ELLIS and RONALD G. DUGGLEBY
Department of Biochemistry, John Curtin School of Medical Research, Australian National University, Canberra, A.C.T. 2601, *Australia*

(*Received* 16 *August* 1977)

In many problems of data analysis it is necessary to fit the data to a mathematical equation. Random errors of measurement will be responsible for deviations between the data and the equation, but superimposed on this there may be deviations that result from the equation being an inadequate description of the system from which the data were obtained. Plots of the residual (i.e. the difference between the experimental and calculated values of the dependent variable) against each of the experimental variables have been previously used to detect a misfit between the data and the equation. In the present paper, we show that the shape of the residual plots may be used as a guide in choosing a more appropriate equation. In addition, residual plots give useful information on the error structure of the data, and hence the weighting factors that should be used in the analysis.

In biochemistry, as in other quantitative sciences, it is often necessary to fit experimental data to a mathematical equation or model. This process usually involves minimizing a suitably weighted sum of squares of the deviations between the data and the model. For this purpose, methods of data treatment have been steadily refined, particularly since computers have become widely available. For example, in kinetic studies of enzymes, the simple method of determining values for the kinetic parameters by drawing a straight line through the data in a Lineweaver–Burk plot has largely been supplanted by the more sophisticated methods of non-linear regression, in which the data are directly fitted to an assumed equation. These advances in the treatment of data have largely been aimed at refining the values of the parameters of a particular equation, although the subjectivity associated with the choice of the equation still remains. It is axiomatic that fitting data to the wrong equation will give meaningless values for the parameters, no matter how sophisticated the method of fitting.

The purpose of the present paper is to show that, when data are fitted to the wrong equation, the shape of the residual plot contains valuable information that can be utilized to determine the way in which the equation should be modified to achieve a better description of the data.

Several workers (Anscombe & Tukey, 1963; Haarhoff, 1969; Bártfai & Mannervik, 1972; Reich *et al.*, 1972, 1974; Storer *et al.*, 1975; Atkins, 1976) have examined the properties of the normalized residual (r) defined by eqn. (1):

$$r = (\hat{y} - y)/\sigma \qquad (1)$$

where y is the experimentally determined value of the dependent variable, $\hat{y}$ is the expected value, calculated from the equation to which the data have been fitted, and σ is the expected standard deviation of y. Provided that the correct equation has been chosen and the fitting procedure is appropriate, the values of the residuals should be unrelated to the values of the dependent or independent variables, or to extraneous factors. Violation of this expectation indicates that either the equation or the fitting procedure is inappropriate, and we have sought to use this fact in a semi-quantitive fashion. It will be shown that plots of the residual against each of the experimental variables can have characteristic shapes that, when compared with the shapes of standard sets of residual plots for a given system, may be used to identify and correct for deficiencies in the original model. In addition, residual plots can also be used to give information about the error structure of the data, and hence the appropriate weighting factors to be used in the analysis. Since it is only the shape of the residual plot that is important, it is not essential to know the absolute values of σ in eqn. (1), but merely a factor proportional to it. Thus, if all data are thought to have equal standard deviations, σ may be given a value of 1.0.

The proposed procedure is best illustrated by a simple example. Consider the data of Fig. 1(*a*), to which a straight line has been fitted. Visual examination of the fit reveals no obvious deficiency, but if a residual plot is constructed (Fig. 1*b*) it is clear that the residual depends to some extent on the value of the independent variable. The residual plot shows that the fitted line passes below the majority of points at the extremes of x, and above them at intermediate values of x. This suggests that a curved line such as a parabola would describe the data more accurately (Fig. 1*c*), and this is confirmed by the residual plot (Fig. 1*d*). In this particular example,

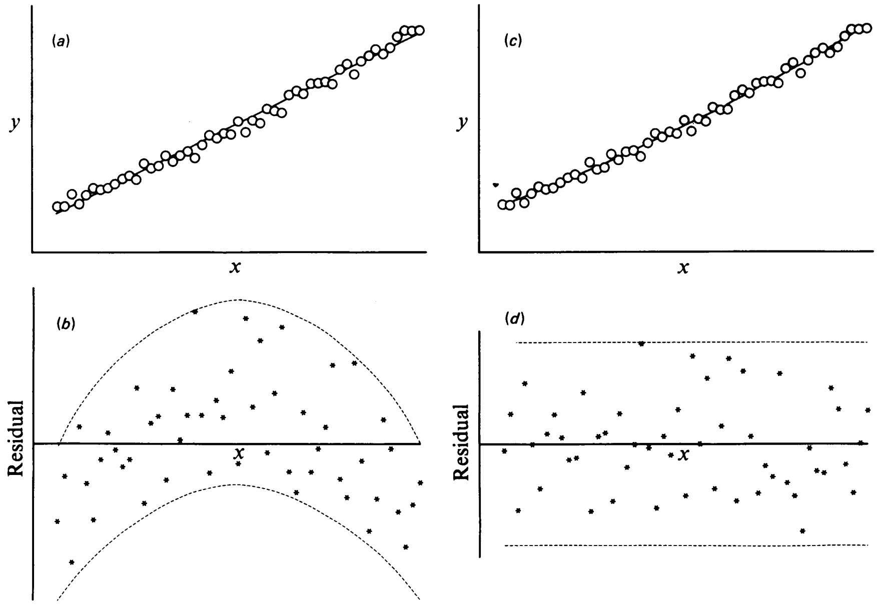

Fig. 1. *Data and residual plots in a simple system*
Identical sets of data are shown in (*a*) and (*c*). The data were fitted to a straight line (*a*), and a plot was constructed of residual against the variable x (*b*). The same data were fitted to a parabola (*c*) and residuals plotted (*d*). The broken lines in the residual plots are to emphasize the distribution of residuals.

virtually any curvilinear model would have eliminated the correlation between the residual and the independent variable. In practice, the type of equation would often be dictated by the underlying theory. For instance, a quadratic function might indicate a particular mechanism, whereas an exponential function might be meaningless.

When a residual plot is found to indicate a misfit between the equation and the data, it is apparent that the equation must be modified if it is to be an accurate description of the experimental results. However, it is perhaps not so apparent that the shape of the residual plot can indicate the manner in which such a modification might be made. Frequently, of course, the aim is not merely to describe the data, but also to deduce a mechanism from the form of the equation. Consider, for example, an enzyme kinetic experiment, designed to study the effect of an inhibitor. Visual inspection of the data suggested that the inhibition might be competitive, but, when the data were fitted to the appropriate equation, residual plots revealed a misfit. In principle, the data could then be fitted to each of several other possible equations, each of which would have different mechanistic implications. However, the characteristic shape of the original residual plots itself reveals which of several feasible alternative equations is most likely to be consistent with the data.

This process is illustrated below, by using a set of simulated data computed from the equation for hyperbolic competitive inhibition. Fig. 2(*a*) shows the data, and the fit to the equation for competitive inhibition, and Fig. 2(*b*) shows one of the residual plots (residual against inhibitor concentration). The curved envelope in which the residuals lie clearly shows that the data are in conflict with the model. Fig. 3 shows the general shape of the residual plots when various types of data are fitted to the equation for competitive inhibition. Comparison of Fig. 2(*b*) with four plots of residual against inhibitor concentration (Fig. 3, bottom row) suggests that the data are likely to fit the equation for hyperbolic competitive inhibition. Fig. 4(*a*) shows one of the residual plots obtained when a set of actual data was fitted to the equation for competitive inhibition. The data are for the inhibition of *Aerobacter aerogenes* prephenate dehydrogenase (EC 1.3.1.12) by bicarbonate with respect to NAD$^+$ (P. K. Dudziński, unpublished work). Clearly, the plot resembles that expected for parabolic inhibition (Fig. 3). When the data were fitted to this latter model, the residual plot shown in Fig. 4(*b*) was obtained. No obvious trends are apparent, and this model may therefore be considered to be an adequate description of the data.

In these particular examples, the plot of residual against inhibitor concentration has the greatest

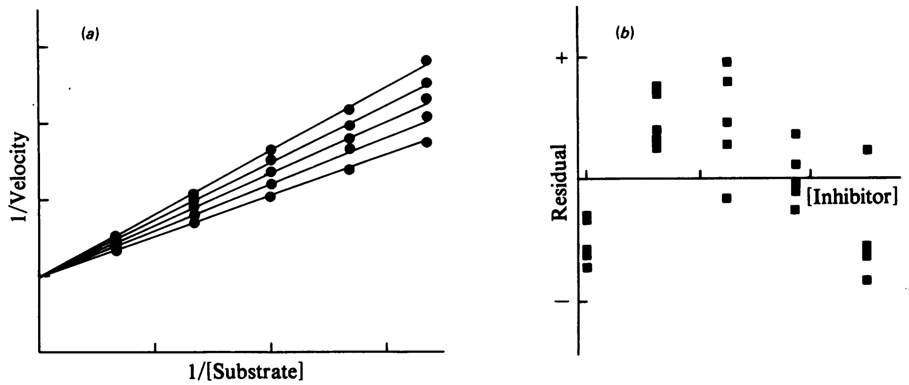

Fig. 2. *Data and residual plots for simulated enzyme-inhibition data*
Theoretical velocities were calculated for five different substrate concentrations at each of five inhibitor concentrations, by using the equation for hyperbolic competitive inhibition. A randomly selected normally distributed error was applied to each data point, and these 25 simulated data points were fitted to the equation for competitive inhibition. (*a*) Double-reciprocal plot of the data and the fitted lines. Each line corresponds to a different inhibitor concentration. (*b*) Plot of residual against inhibitor concentration.

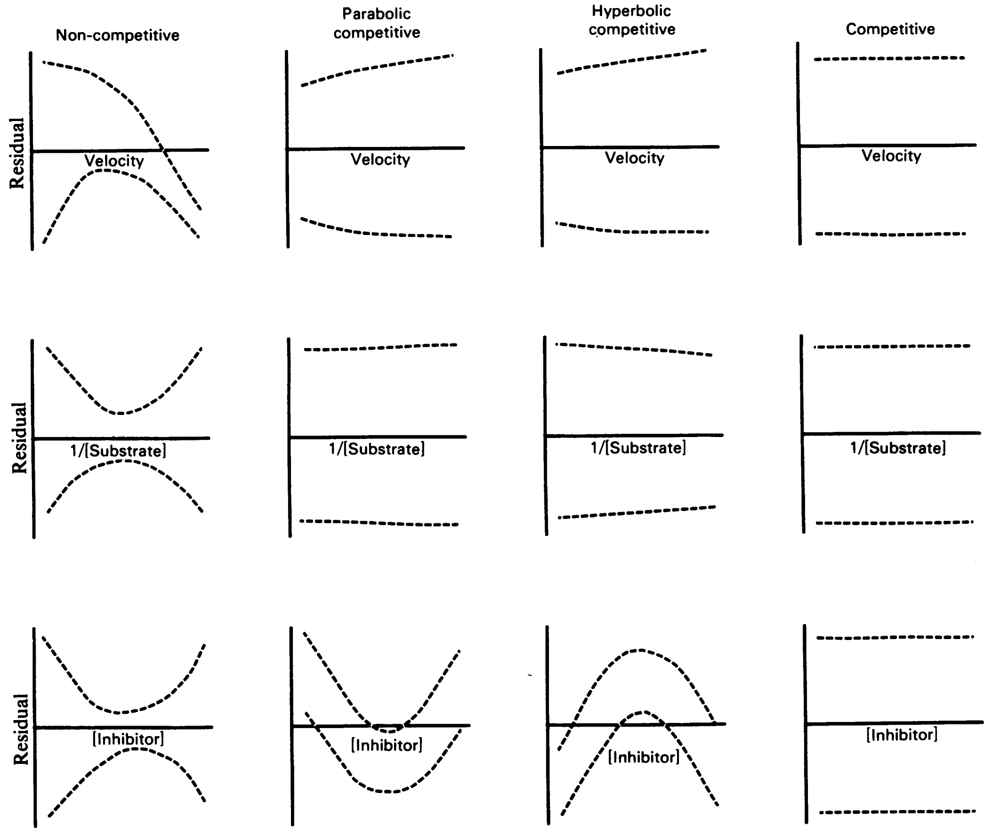

Fig. 3. *Diagnostic residual plots for data fitted to the equation for competitive inhibition*
Theoretical sets of non-competitive, parabolic competitive, hyperbolic competitive and linear competitive inhibition data were calculated, and fitted to the equation for competitive inhibition. Plots of residual against velocity, $[\text{substrate}]^{-1}$ and [inhibitor] were then constructed for each set. The equation from which the data was derived is listed at the top. In each case, the broken lines indicate the general shape of the residual plots.

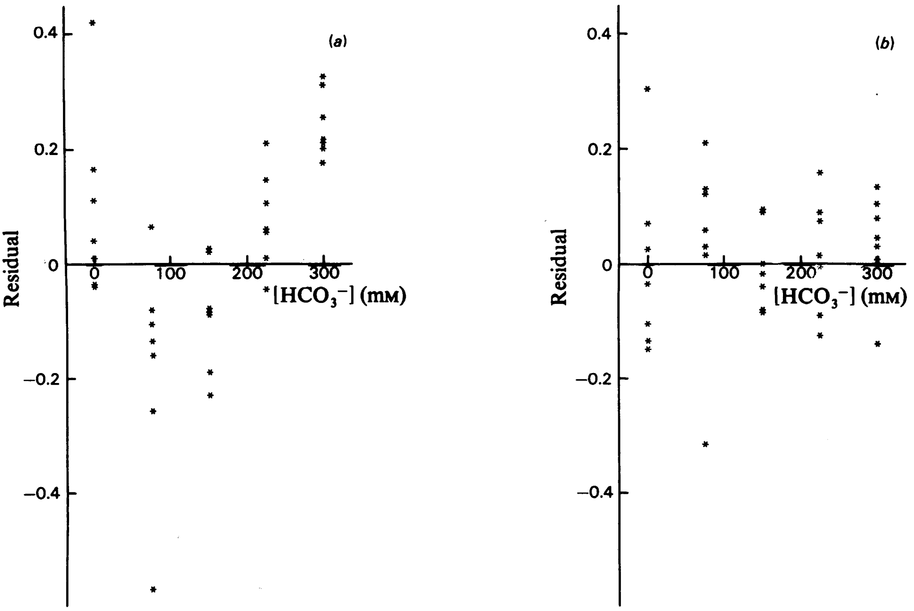

Fig. 4. *Residual plots for the inhibition of Aerobacter aerogenes prephenate dehydrogenase by bicarbonate with respect to NAD*$^+$

Residuals are plotted against bicarbonate concentration, for situations in which the data have been fitted to either competitive inhibition (*a*) or parabolic inhibition (*b*).

diagnostic power. Nevertheless, we recommend that plots of residual against each of the experimental variables should be examined, to ensure that all patterns are consistent with the proposed model. It is a relatively simple matter to generate the residual plots during computer fitting of the data.

In any data-fitting problem, it is essential that individual data points should be weighted according to their expected variance (Wilkinson, 1961; Reich, 1970; Storer *et al.*, 1975), and residual plots can be used to estimate the appropriate weighting factors. For example, a wedge-shaped plot of residual against the dependent variable (Storer *et al.*, 1975) will result from incorrectly assuming the data points to have equal variances and using equal weights in the analysis, when, in fact, the variance is proportional to the dependent variable. It should be borne in mind that both a misfit and incorrect weighting may be present simultaneously, which could lead to some confusion. It is recommended that a few replicate measurements should be made so that approximate weighting factors may be used until a reasonably good model is found, and then the weighting factors refined. This process can be repeated until residual plots reveal no trends.

It will occasionally happen, particularly with small data sets, that a residual plot may appear to indicate a trend that is, in fact, illusory. For this reason, whenever a complex model is chosen over a simpler one, it is essential that the fit to the more complex model should be shown to be significantly better by an appropriate statistical test. We have found the *F* test (Haarhoff, 1969; Duggleby & Dennis, 1974; McMinn & Ottaway, 1977) useful in this connection.

F is calculated from eqn. (2):

$$F=\frac{(R_2-R_1)(n-p_1)}{R_1(p_1-p_2)} \qquad (2)$$

Where R_1 and p_1 are the sum of squares of the residuals and the number of parameters associated with the more complex model, R_2 and p_2 are the corresponding parameters of the simpler model, and *n* is the number of data points. The degrees of freedom of *F* are (p_1-p_2) and $(n-p_1)$. For example, the comparison between competitive and parabolic inhibition for the data of Fig. 4 gives an *F* value that would be obtained by chance in fewer than 1 in 1000 experiments if the system really exhibits competitive inhibition. This is sufficiently unlikely to provide strong support for the view that the system does exhibit parabolic inhibition. We would endorse the view of Atkins (1976) that several statistical tests should be used wherever practical.

Although most of the results presented in the present paper were obtained with simulated data, the procedure has been used by members of this Department with experimental data over the past 2 years, and has proved to be valuable in practice. The examples presented are mainly concerned with enzyme kinetic data, but we expect that residual plots would be useful in a variety of data-fitting problems, in both biochemistry and other sciences.

We gratefully acknowledge the interest and encouragement shown by Dr. John F. Morrison during the development of this work, and thank Ms. Paulina Dudziński for the data on the inhibition of prephenate dehydrogenase.

References

Anscombe, F. J. & Tukey, J. W. (1963) *Technometrics* **5**, 141–160

Atkins, G. L. (1976) *Biochem. Soc. Trans.* **4**, 357–361

Bártfai, T. & Mannervik, B. (1972) *FEBS Lett.* **26**, 252–256

Duggleby, R. G.: & Dennis, D. T. (1974) *J. Biol. Chem.* **249**, 167–174, but see correction, p. 6027

Haarhoff, K. N. (1969) *J. Theor. Biol.* **22**, 117–150

McMinn, C. L. & Ottaway, J. H. (1977) *Biochem. J.* **161**, 569–581

Reich, J. G. (1970) *FEBS Lett.* **9**, 245–251

Reich, J. G., Wangermann, G., Falck, M. & Rohde, K. (1972) *Eur. J. Biochem.* **26**, 368–379

Reich, J. G., Winkler, J. & Zinke, I. (1974) *Stud. Biophys.* **43**, 77–90

Storer, A. C., Darlison, M. G. & Cornish-Bowden, A. (1975) *Biochem. J.* **151**, 361–367

Wilkinson, G. N. (1961) *Biochem. J.* **80**, 324–332