*J. Pharm. Pharmac.*, 1974, **26**, 807–813 *Received January* 29, 1974

# Improved parameter estimates in drug-protein binding studies by non-linear regression

B. W. MADSEN AND J. S. ROBERTSON

*Department of Pharmacy, Sydney University, Sydney 2006, Australia*

A quantitative comparison of various methods of fundamental binding parameter estimation is presented. For optimal reliability, choice of a particular approach should be made on a critical assessment of the experimental procedure and various calculated statistical criteria, rather than on the ease of data treatment.

Interest in the relative merit of different procedures for analysing binding data developed in these laboratories through studies on the oral anticoagulants. There is a good deal of variation between different workers on supposedly similar systems, and inspection of the data suggested that choice of treatment could be responsible in part. For example, consider the bishydroxycoumarin—human plasma albumin interaction where Chignell (1970) reported the number of sites in the highest affinity class ($n_1$) as 3 and an association constant ($K_1$) of $7{\cdot}7 \times 10^5$ M$^{-1}$; Garten & Wosilait (1971) found $n_1 = 2{\cdot}08$ and $K_1 = 2{\cdot}08 \times 10^6$ M$^{-1}$; O'Reilly (1971) gave $n_1 = 2{\cdot}0 \pm 0{\cdot}1$ and $K_1 = 2{\cdot}31 \pm 0{\cdot}05 \times 10^5$ M$^{-1}$ and Cho, Mitchell & Pernarowski (1971) gave $n_1 = 3$ and $K_1 = 3{\cdot}5 \times 10^5$ M$^{-1}$.

Recent work by Noval & Champion (1973) on the binding of chlorpromazine to human albumin provides a striking example of the dependence of parameter estimates on the method of treatment. Equation 1 is the general form derived from the law of mass action for the reversible interaction of drugs and proteins.

$$\bar{v} = \sum_{i=1}^{c} \frac{n_1 \quad K_1 \quad D_f}{1 + K_1 \quad D_f} \qquad .. \qquad .. \quad (1)$$

where v = molar ratio of drug bound per mole of protein, c = number of classes of sites, $n_i$ = number of binding sites within the ith class, $K_1$ = association constant of the ith class, $D_f$ = molar free drug. In the simplest case where there is only one class of sites involved, equation 1 is often rearranged to give two linear forms making parameter estimation simpler.

$$\frac{1}{\bar{v}} = \frac{1}{n_1 \quad K_1 \quad D_f} + \frac{1}{n_1} \qquad .. \qquad .. \quad (2)$$

$$\frac{v}{D_f} = n_1 K_1 - K_1 \bar{v} \qquad .. \qquad .. \quad (3)$$

Noval and Champion used both these equations on the one data set and found $n_1 = 16$, $K_1 = 1355$ M$^{-1}$ with equation 2 and $n_1 = 25$, $K_1 = 810$ M$^{-1}$ with equation 3.

Kruger-Thiemer (1967) has previously observed that analysis of drug-protein bind-

ing data using a non-linear approach (as in equation 1) ensures more reproducible results than any of the linear transformations, and our work is in support of this contention. However, as both equations 2 and 3 are still widely used in the literature, it seemed of value to demonstrate quantitatively their limitations, and to propose a more reliable procedure.

*Theory*

In drug-protein binding studies, the free and bound drug data are the important variables from a statistical viewpoint since they are generally the most uncertain. Other variables usually have a much higher degree of confidence attached to them; e.g. determination of protein concentration is not subject to the same experimental difficulties and assumptions.

When fitting data to a one class model of equation 1 where the primary concern is to obtain the most reliable estimates of $n_1$ $K_1$, then the dependent variable in any regression should be that with the largest associated error. This means the investigator must critically examine the experimental technique. For example, with equilibrium dialysis and strongly bound drugs, it is possible for the free drug estimate to be in error by 50% while the corresponding bound drug ($D_b$) might only change from 99·1% to 99% if there is uncertainty on the proper equilibration time. Under these circumstances, it would be better to regress $D_f$ on $\bar{\nu}$ for maximum stability of $n_1$ and $K_1$. Alternatively, since equilibrium dialysis is a subtractive technique, $\bar{\nu}$ could be more uncertain than $D_f$. If the amount of protein is small, $D_f$ may be estimated by subtraction of two large numbers i.e. total drug in the system and assayed free drug. Regressing $\bar{\nu}$ on to $D_f$ would then be applicable.

To show the significance of these considerations, we examined a theoretical model with one class of sites where $n_1 = 3{\cdot}0$ and $K_1 = 1 \times 10^6$ M$^{-1}$. Using equation 1, a series of $\bar{\nu}$ can be calculated for typical values of $D_f$ (Table 1). Normally distributed noise can then be introduced to these data using a pseudo-random number generator on a computer (Table 2). By varying the proportion and amount of noise in each of $D_f$ and $\bar{\nu}$, a wide range of cases can be studied some of which will lie in the region of

Table 1. *The first two columns show theoretical data for a one class model with three sites and average association constant of 1 × 10⁶ M⁻¹.* A typical scattered data set is shown in columns 3 and 4 where the standard deviations were $\pm 1 \times 10^{-7}$ and $\pm$ 0·003 respectively.

| $D_f$ (x $10^{-7}$M) | $\bar{\nu}$ | $D_f \pm \sigma_a$(x $10^{-7}$M) | $\bar{\nu} \pm \sigma_b$ |
|---|---|---|---|
| 2·0 | ·5000 | 2·157 | ·5005 |
| 3·0 | ·6923 | 1·332 | ·6943 |
| 5·0 | 1·0000 | 4·612 | 1·0071 |
| 6·0 | 1·1250 | 6·883 | 1·1251 |
| 9·0 | 1·4210 | 8·740 | 1·4244 |
| 10·7 | 1·5507 | 9·567 | 1·5412 |
| 11·5 | 1·6046 | 11·471 | 1·6000 |
| 12·8 | 1·6842 | 12·430 | 1·6894 |
| 13·3 | 1·7124 | 13·978 | 1·7116 |
| 15·6 | 1·8281 | 17·069 | 1·8269 |
| 18·2 | 1·9361 | 18·795 | 1·9381 |
| 20·3 | 2·0099 | 20·681 | 2·0101 |
| 22·2 | 2·0683 | 20·891 | 2·9699 |
| 24·2 | 2·1228 | 25·547 | 2·1222 |

experimentally attainable precision. With these data sets, equations 1, 2, 3 and 4 can be compared for reliability in estimating n & K.

$$D_f = \frac{\bar{v}}{K(n-\bar{v})} \tag{4}$$

Westlake (1971) has used a similar technique in commenting on pharmacokinetic analyses.

## METHODS

The method of least squares was used in all cases to find the best parameter estimates. Linear equations were solved analytically by double precision matrix inversion using MINV*. The Adaptive Simplex approach to function minimization of Nelder & Mead (1965) was programmed as a subroutine FUNMIN and used for solution of the non-linear equations. Parameter standard deviations were calculated from the diagonal elements in the inverse matrix of the sums of cross products of the partial derivatives (Kendall & Stuart, 1961). Normally distributed numbers were generated by summing 12 uniformly distributed random numbers. The standard deviation of the second parameter from the linear treatments was calculated using equation 5 (Wilson, 1952).

If $z = f(x,y)$

then $$S_z = \sqrt{\left(S_x \cdot \frac{\partial z}{\partial x}\right)^2 + \left(S_y \cdot \frac{\partial z}{\partial y}\right)^2} \tag{5}$$

where $S_i$ is the uncertainty in the ith parameter.

The computer was a CDC 6600 and all programming was in FORTRAN IV.

## RESULTS

We examined many different situations of varying proportions of noise and a selection is presented in Table 2.

Cases 1 and 2 represent probably the more common situation for clinical levels of many drugs in the range of the first class of binding sites i.e. where there is a larger relative uncertainty in $D_f$ than in $\bar{v}$. For very precise data (Case 1) there is a difference in parameter estimates between alternative treatments, but it is probably of no consequence from a biological viewpoint since there are other variables in the system with comparable uncertainties. The molecular weight of albumin is probably not known to better than 2–3% (Phelps & Putnam, 1960).

This does not apply for more realistic data of less precision (Case 2). Notice that regressing $D_f$ onto $\bar{v}$ (4) gives the most accurate estimates of n & K, with $\bar{v}$ on $D_f$ (3) next best, followed by the Scatchard plot (1) and lastly the Reciprocal plot (2). Hence, if an investigator had correctly assessed that the data would be most reliable in $\bar{v}$ and subsequently used only equation 4 he could confidently hope to obtain the best estimates.

The differences between the four estimates in Case 2 are not insignificant. There is some controversy at present whether n estimates should be allowed to be non integral

* IBM System/360 Scientific Subroutine Package (1967) p55.

Table 2. *Different situations with varying proportions of noise.* Cases 1 and 2 illustrate the situation where the relative uncertainty in $D_f$ is greater than that in 3 and 4 the inverse and 5 and 6 where the uncertainty in each variable is comparable. The standard deviations of the artificially introduced noise are given as approximate percentages of the mean. The replicate results in each case correspond to treatments by the respective equations and parameter uncertainties are expressed as standard deviations. Case 2 results were derived from the scattered data of Table 1.

| Case | Standard deviation of artificial noise | | Derived parameters | |
|---|---|---|---|---|
| | $D_f$ (x $10^{-7}$M) | $\bar{v}$ | n | K(x $10^6$ $M^{-1}$) |
| 1 | ·1 (·81%) | ·0002 (·013%) | (3) 2·97 ± ·07 | 1·03 ± ·02 |
| | | | (2) 2·98 ± ·06 | 1·02 ± ·02 |
| | | | (1) 2·98 ± ·02 | 1·02 ± ·01 |
| | | | (4) 2·99 ± ·01 | 1·00 ± ·01 |
| 2 | 1·0 (8·1%) | ·003 (·2%) | (3) 2·52 ± ·72 | 1·67 ± ·40 |
| | | | (2) 2·06 ± ·35 | 2·62 ± ·65 |
| | | | (1) 2·79 ± ·17 | 1·21 ± ·19 |
| | | | (4) 2·94 ± ·10 | 1·04 ± ·10 |
| 3 | ·0125 (·1%) | ·04 (2·62%) | (3) 2·96 ± ·14 | 1·06 ± ·04 |
| | | | (2) 2·89 ± ·06 | 1·12 ± ·02 |
| | | | (1) 3·03 ± ·08 | 1·00 ± ·06 |
| | | | (4) 3·19 ± ·09 | ·88 ± ·07 |
| 4 | ·05 (·4%) | ·16 (10·5%) | (3) 3·44 ± ·65 | ·68 ± ·12 |
| | | | (2) 3·75 ± ·73 | ·59 ± ·12 |
| | | | (1) 3·19 ± ·27 | ·78 ± ·14 |
| | | | (4) 3·72 ± ·54 | ·57 ± ·17 |
| 5 | ·25 (2%) | ·025 (1·65%) | (3) 2·87 ± ·16 | 1·12 ± ·05 |
| | | | (2) 2·74 ± ·09 | 1·23 ± ·05 |
| | | | (1) 2·97 ± ·06 | 1·02 ± ·05 |
| | | | (4) 3·01 ± ·04 | ·99 ± ·04 |
| 6 | 1·5 (12·1%) | ·15 (9·9%) | (3) 2·60 ± ·83 | 1·75 ± ·47 |
| | | | (2) 1·93 ± ·18 | 3·92 ± ·55 |
| | | | (1) 2·91 ± ·44 | 1·18 ± ·46 |
| | | | (4) 4·82 ± 1·96 | ·42 ± ·29 |

or not. Conceptually, many workers find it difficult to talk of two and a half sites per molecule, and round off results to the nearest integer. It is not difficult to envisage 2·94 ±0·10 as providing evidence of 3 sites and possibly even 2·79 ± 0·17. But 2·52 can equally well be rounded up to 3 or down to 2 by those who maintain n must be integral, and so dissension becomes possible. Those who accept a non-integral estimate for n of $2\frac{1}{2}$ have to conceive of, for example, the presence of half a mole of endogenous fatty acid or microheterogeneity. If equation 2 was used then n equal to 2 would be assigned without question. In fact, we know that all such interpretations are unfounded and the possibility of arriving at either troublesome or erroneous results was a direct result of linear data treatment.

The differences between treatments can be further illustrated graphically. Fig. 1 presents the data according to equation 3. If the second point to the left is rejected completely then this will bias the n estimate too high whereas when left in the regression it has an undue influence on depressing n. Clearly this rearrangement places an incorrect weighting on the low $D_f$ values, so much so that the whole regression might be rejected with what is in fact valid, normally distributed data.

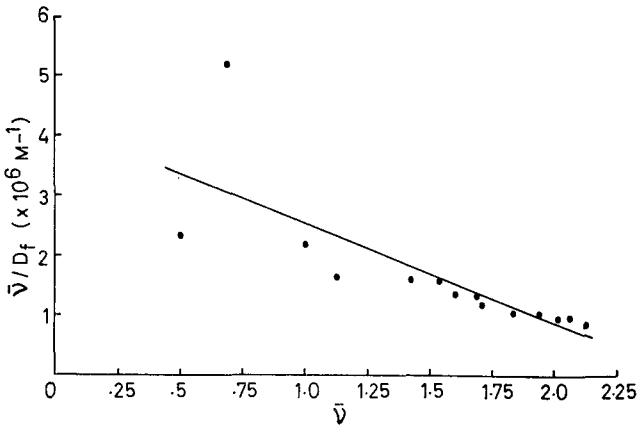


FIG. 1. Scatchard plot of the scattered data shown in Table 1 corresponding to treatment 1, Case 2 in Table 2. The points represent the experimental data and the line that given as the best least squares fit.

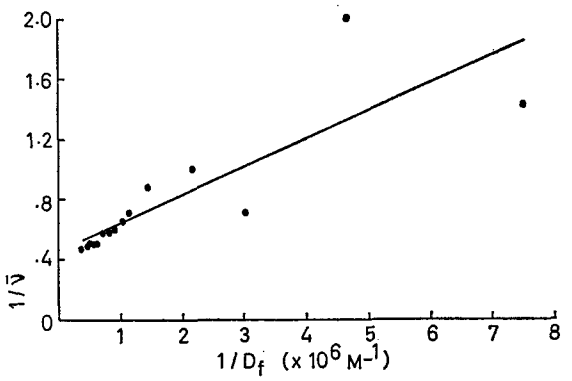


FIG. 2. Reciprocal plot of the scattered data shown in Table 1 corresponding to treatment 2 Case 2 in Table 2. Points are experimental and the line gives the best least squares fit.

Similar considerations would apply to the reciprocal plot shown in Fig. 2. It is easy to see how confidence in the precision of the intercept is imagined and hence that the estimate for n will be accurate. Numerically, the method produces what is not wanted. If an estimate is to suffer in accuracy, then at least the confidence intervals should enclose the true value. Again, a result of 2·06 $\pm$ ·35 would probably be taken as good evidence for $n = 2$.

The four treatments are compared on a common set of axes in Fig. 3. Obviously the assumptions made in setting up regression equations 1–3 cannot be regarded as unequivocably correct.

Cases 3 and 4 confirm the belief that linear rearrangements should not be used at all and that the correct non-linear treatment can be chosen from a consideration of the relative errors in $\bar{\nu}$ and $D_f$. For the uncertainty in $\bar{\nu}$ greater than $D_f$, equation 3 should be used. On the other hand, if no information on uncertainties in $\bar{\nu}$ and $D_f$ is available then inspection of the predicted uncertainties from both non-linear treatments will enable selection of the most accurate estimates.

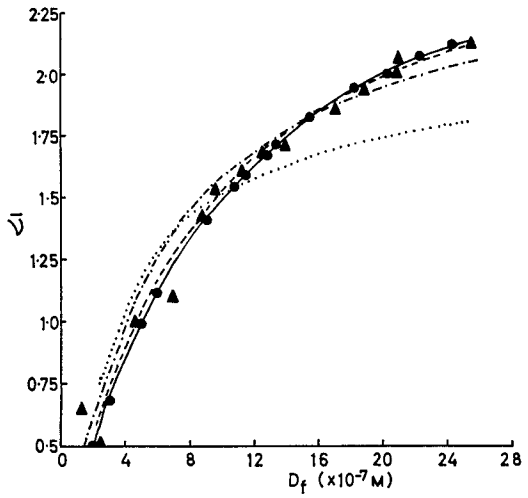


FIG. 3. Reduction of the four regressions to a common pair of axes $\bar{\nu}$ and $D_f$. Key — ● theoretical data; ▲ the primary scattered data used in all four treatments; ——— curve of best fit by regressing $D_f$ on $\bar{\nu}$; — — — — regression of $\bar{\nu}$ on $D_f$; — · — · — curve obtained from Scatchard plot (equation 3); ..... curve from a reciprocal plot (equation 2).

When the relative errors in $\bar{\nu}$ and $D_f$ are thought to be comparable (cases 5 and 6) there is no alternative but to use both equations 3 and 4 and to choose the parameters with smallest uncertainty. For example treatment (4) would be selected in case 5, and treatment (3) in case 6. Which method will be best depends on the particular distribution and scatter of each experimental set and cannot be predicted before numerical analysis.

## DISCUSSION

It is necessary to be concerned about the comparative reliability of different treatments when there is limited information available and significant uncertainty in the measured variables. Obviously, linear transformations in protein binding studies have simplicity to recommend them and should only be discarded when there is some likelihood of bias or false confidence. It would seem from our results that this possibility becomes real and biologically important when uncertainty levels in the primary data are of the order $\pm 5$—10%. Unfortunately in drug-protein binding studies, particularly in the clinical range, precision is worse than this more often than not and the best that can be done is to use procedures which are reliable. Non-linear processing is slightly more complicated than linear, but if the proper degree of sophistication is entertained, i.e. the method of least squares and estimation of parameter uncertainties, then the extra calculations are inconsequential on a computer.

Musulin (1973) has shown in viscosity work that linear and non-linear estimates of a given problem may differ, but that these converge with increasing precision of the primary data, in accordance with our results. The implications of this general rule are at variance with a not uncommon experimental attitude. Many workers, reasoning that their results will be subject to many variables, consider it not worthwhile to use a more sophisticated procedure and opt for a simple linear form. Yet these are precisely the type of data which should not be rearranged. In summary, Musulin (1973) states "the use of modern computers affords ..... an opportunity,

which should not be disregarded, to emphasize the effect of the experimenter's interaction upon the final quantitative results".

The problem with rearrangements can be illustrated as follows. With perfect data, equations 6 and 7 will return identical estimates of m and c given x and y.

$$y = m \,.\, x + c \qquad (6)$$

$$x = \frac{1}{m} \,.\, y - \frac{c}{m} \qquad (7)$$

Once uncertainty in either or both of x and y exists, then estimates of m and c will invariably differ for the two treatments (Draper & Smith, 1966) since all the error is assumed to be in the dependent variable. If inversion or compounding of the measured variables also occurs as in equations 2 and 1 respectively, then parameter estimates may vary even more because of differing weights. Hence in Figs 1 and 2 the two highest points contribute the bulk of the total sum of squares, yet they are relatively the most uncertain. The problem could be overcome by divising a suitable weighting scheme, but it may be simpler to change to a non-linear approach.

Use of equation 2 in the unweighted form has been criticized many times (Klotz, Walker & Pivan, 1946; Scatchard, 1949; Meyer & Guttman, 1968), while equation 3 seems to have remained acceptable to many. Nevertheless, Weber & Anderson (1965) assert that all plots in which data are reduced to a linear form ought to be avoided, and we are in agreement with this. They stated "we would have been unable to detect many of the interesting regularities present had we resorted to reduction of the data to a linear form".

As a prediction from the work described in this report, it is probable that for studies where more than one class of sites is involved, selection of either $\bar{v}$ or $D_f$ in equation 1 as the dependent variable in a non-linear regression will become increasingly important for reliable n K estimates. Reports in the literature seem to have unanimously regressed $\bar{v}$ onto $D_f$ more from an ease of data treatment consideration than the relative uncertainty in the two variables.

## REFERENCES


CHIGNELL, C. F. (1970). *Mol. Pharmac.*, **6**, 1–12.

CHO, M. J., MITCHELL, A. G. & PERNAROWSKI, M. (1971). *J. pharm. Sci.*, **60**, 196–200.

DRAPER, N. R. & SMITH, H. (1966). *Applied Regression Analysis*, p. 5. Wiley: New York.

GARTEN, S. & WOSILAIT, W. D. (1971). *Comp. Prog. in Biomed.*, **1**, 281–285.

KENDALL, M. G. & STUART, A. (1961). *The Advanced Theory of Statistics*, Vol. 2, p. 75. London: Griffin.

KLOTZ, I. M., WALKER, F. M. & PIVAN, R. B. (1946). *J. Am. chem. Soc.*, **68**, 1486–1490.

KRUGER-THIEMER, E. (1967). *Pharmacologic Techniques in Drug Evaluation*, Vol 2, Editors: Siegler, P. E. and Moyer III, J. E. p. 225. Chicago: Year Book Medical Publishers.

MEYER, M. C. & GUTTMAN, D. E. (1968). *J. pharm. Sci.* **57**, 895–918.

MUSULIN, B. (1973). *J. Chem. Educ.*, **50**, 79.

NELDER, J. A. & MEAD, R. (1965). *Computer J.*, **7**, 308–313.

NOVAL, J. J. & CHAMPION, V. (1973). *Res. Comm. Chem. Path. Pharmac.*, **6**, 123–135.

O'REILLY, R. A. (1971). *Mol. Pharmac.*, **7**, 209–215.

PHELPS, R. A. & PUTNAM, F. W. (1960). *The Plasma Proteins* Vol. 1, Editor: Putnam, F. W. p. 168. New York: Academic Press.

SCATCHARD, G. (1949). *Ann. N. Y. Acad. Sci.*, **51**, 660–672.

WEBER, G. & ANDERSON, S. R. (1965). *Biochem.*, **4**, 1942–1947.

WESTLAKE, W. J. (1971). *J. pharm. Sci.*, **60**, 882–885.

WILSON, E. B. (1952). *An Introduction to Scientific Research*, p. 273. New York: McGraw-Hill.