Biochimica et Biophysica Acta, 567 (1979) 43—48

BBA 68683

A NEW PROCEDURE TO DERIVE WEIGHTING FACTORS FOR NONLINEAR REGRESSION ANALYSIS APPLIED TO ENZYME KINETIC DATA

BENGT MANNERVIK, INGA JAKOBSON and MARGARETA WARHOLM

Department of Biochemistry, Arrhenius Laboratory, University of Stockholm, S-10691 Stockholm, and Section of Occupational Toxicology, National Board of Occupational Safety and Health, S-10026 Stockholm 34 (Sweden)

(Received August 28th, 1978)

Key words: Regression analysis; Weighting factor; Variance; Kinetics

Summary

The experimental variance of enzymic steady-state kinetic experiments depends on velocity as approximated by a power function $\mathrm{Var}(v) = K_1 \cdot v^{\alpha}$ (Askelöf, P., Korsfeldt, M. and Mannervik, B. (1976) Eur. J. Biochem. 69, 61—67). The values of the constants (K_1, α) can be estimated by making replicate measurements of velocity, and the inverse of the function can then be used as a weighting factor. In order to avoid measurement of a large number of replicates to establish the error structure of a kinetic data set, a different approach was tested. After a preliminary regression using a 'good model', which satisfies reasonable goodness-of-fit criteria, the residuals were taken to represent the experimental error. The neighbouring residuals were grouped together and the sum of their mean squared values was used as a measure of the variance in the neighbourhood of the corresponding measurements. The values of the constants obtained in this way agreed with those obtained by replicates.

Introduction

The experimental variance of empirical data has to be considered in the mathematical modelling and quantitative analysis of a chemical or physical system. For example, in the evaluation of goodness-of-fit of a given model or in the discrimination between alternative models it is essential to know if the random error is dependent on the magnitude of the measured variable. More specifically, if the analysis of data involves regression analysis it may be necessary to introduce weighting factors to obtain unbiased results. The present

paper introduces a method to derive weights for regression analysis in cases which require weighting. The procedure is quite general but the examples discussed will be analyses of enzyme kinetic data, from which our experience derives.

Previous attempts to derive weighting factors for regression analysis

In a previous paper we have shown that for the enzyme, glutathione *S*-transferase A, the experimental variance of the initial velocity of the enzyme-catalyzed reaction increases with the measured values [1]. Similar observations have been made with other enzymes [2–6]. However, it has also been found that with some systems no significant dependence of experimental error on velocity can be demonstrated, even if similar experimental techniques have been used as in the cases in which such a dependence can be demonstrated (unpublished data). Therefore, every experiment may have its own particular error structure and it is obviously an unjustified oversimplification to consider only two types of error, viz. constant absolute error or constant relative error [7]. The importance of using weighting factors for least-squares regression analysis when the experimental error is heteroscedastic (nonconstant) has been pointed out earlier in connection with analysis of enzyme kinetic data [8,9].

It has been proposed that the experimental variance should be determined as a function of the measured velocity (v) by making numerous replicates in the whole range of the measurements [1]. The estimated local variances are then used to define the constants of an empirical function, which describes the experimental variance:

$$\mathrm{Var}(v) = K_1 \cdot v^{\alpha} \qquad (1)$$

The variance can then be predicted from Eqn. 1 for each experimental point and the inverse of the variance can be used as a weighting factor in the regression analysis [1]. A similar procedure was independently developed [6]. However, for a rigorous analysis of the error structure many replicates in various points have to be measured and comparatively little information about other properties of the system investigated will be obtained in relation to the number of measurements. Further, it is also possible that replicates can be made with high precision and that the estimated variance from such experiments may be too small and not be representative of the error of normal kinetic experiments lacking replicates.

New procedure to derive weighting factors

Assume that a mathematical model (rate equation) has been fitted without bias to a set of experimental data by a suitable regression program. If the model is adequate, the residuals of the regression analysis should reflect the experimental error only. Therefore, the set of residuals resulting from the regression analysis contains information which can be used to estimate the error structure of the original data set. The local variance in the neighbourhood of an experimental point (velocity) can be estimated as the mean of the squared residuals:

$$(\sum_{i=1}^{m} q_i^2)/m$$

where q_i ($i = 1, \ldots, m$) are the m residuals in a region considered. The dependence of the variance on velocity can then be defined by relating the values of the estimated local variances to the corresponding velocities by use of Eqn. 1. Considering the difficulties of estimating variance from a small number of replicates [5,10], we suggest that at least 5 points ($m \geqslant 5$) should be used for the estimates. The median of the velocity in a region corresponding to the estimated local variance was used as the independent variable of Eqn. 1, but the geometric or arithmetic mean may also be chosen. The suitability of the choice is dependent on the spacing of the points.

A prerequisite for the application of the new procedure is that the residuals have been obtained from an unbiased fit of the data. If any uncertainty exists as to the adequacy of the equation fitted to the data set, it is better to choose a possibly overdetermined equation with redundant parameters than to use a simpler equation, which might introduce bias in the resulting set of residuals. The choice of an overdetermined equation will always give a good fit of the data and the residuals will reflect only the experimental error. For example, in the analysis of the kinetics of butyrylcholinesterase [11] a 1 : 1 function gives biased residuals and cannot be used (Fig. 2a in ref. 11), a 2 : 2 function appears adequate (Fig. 2b in ref. 11), but the use of a 3 : 3 function would be a safeguard against undetected bias in the fit of the 2 : 2 function. Further, an unbiased fit will only be expected if proper weighting factors have been used in the regression analysis. Therefore, an iterative procedure has to be used in which provisional weights are used in the first regression. New weighting factors can be obtained from the residuals of the first regression analysis, and a new run can then be made. If necessary, the regression and the derivation of weighting factors should be repeated to obtain convergence, but generally the regression will not have to be repeated more than once, provided that good provisional weights have been used. The initial weighting factors can normally be taken from Eqn. 1 by assuming that α equals zero (constant absolute error) or 2 (constant relative error). The choice between the two alternatives can be based on preliminary information. We have demonstrated with data obtained with butyrylcholinesterase that the different initial assumptions ($\alpha = 0$ and $\alpha = 2$) gave the same result ($\alpha = 1.2$ and 1.1, respectively) after analysis of the residuals of the two regressions, and this result was in agreement with the α value determined on the basis of replicates (see below).

Results

In order to test the new procedure to derive weighting factors, the enzymic reactions catalyzed by glutathione *S*-transferase A from rat liver, glutathione reductase from yeast, and butyrylcholinesterase from horse plasma were examined. The kinetics of the enzymes have been studied previously [1,4,11, 12], and the dependence of the experimental error on velocity was established for each of the enzymes by replicates ($n = 10$) over the ranges of velocities covered in the kinetic experiments as previously described [1]. The estimates of the α value based on replicates were 1.6 ± 0.3 for glutathione *S*-transferase A [1], 0.7 ± 0.2 for glutathione reductase, and 1.0 ± 0.2 for butyrylcholinesterase. The α value was then estimated independently by use of residuals

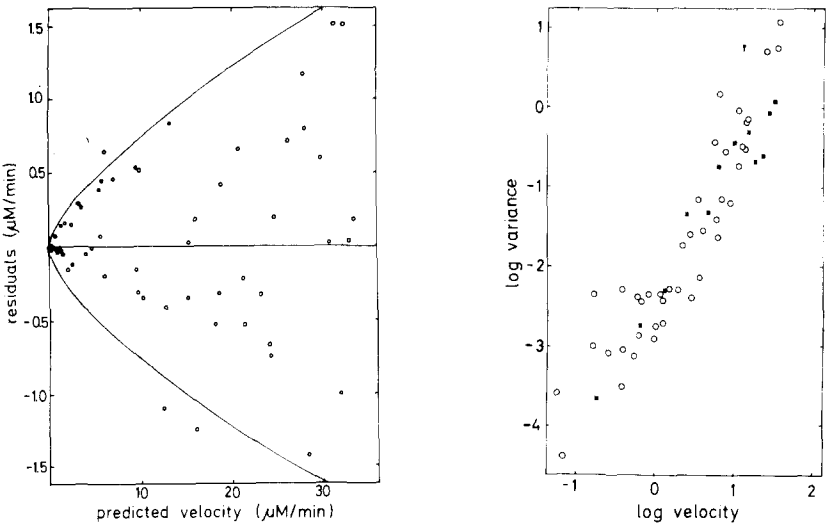

Fig. 1. Plot of residuals versus predicted velocity after fitting

$$v = \frac{K_1[G] + K_2[G]^2 + K_3[G][I]}{K_4 + [G] + K_5[G]^2 + K_6[I] + K_7[I]^2 + K_8[G][I]}$$

to a data set involving variable inhibitor (I, *S*-*n*-octylglutathione) and glutathione (G) concentrations obtained with glutathione *S*-transferase A (cf. Fig. 5 in [12]). The experimental variance was estimated as $\mathrm{Var}(v) = 0.0061 \cdot v^{1.37}$ from the analysis of residuals described in the present paper. The estimated local variances are plotted in Fig. 2 together with the estimates obtained by replicate measurements. The curved lines from the origin are the calculated 95% confidence contours obtained by multiplying the square root of the predicted variance with the appropriate *t*-statistic.

Fig. 2. Estimates of experimental variance as a function of velocity for glutathione *S*-transferase A. The variance was estimated by analysis of the residuals of Fig. 1 using the procedure described in the present paper (■) or by making replicates ($n = 10$) as described in [1] under similar experimental conditions (variable concentrations of one substrate and one inhibitor) (○). The velocity is expressed in μM/min and the variance in $\mu M^2/min^2$.

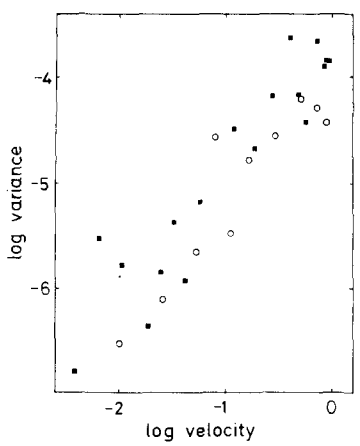

Fig. 3. Estimates of experimental variance as a function of velocity for butyrylcholinesterase. The variance was estimated by analysis of the residuals in Fig. 2b in [11] using the procedure described in the present paper (■) or by making replicates ($n = 10$) at different substrate concentrations (○). The velocity is expressed in $\Delta A_{412nm}/min$ and the variance in $\Delta A^2/min^2$.

from the best-fitting regressions of kinetic data sets by arranging residuals in groups of 5 to 6 and fitting the mean of their squares as a function of the median of the predicted velocities of the group. The residuals used for one such analysis of glutathione *S*-transferase A are plotted in Fig. 1; the residuals used for glutathione reductase are from the data set in Fig. 12 in [4]; and the residuals used for butyrylcholinesterase are those of Fig. 2b in [11]. The resulting α values were 1.4 ± 0.4, 0.6 ± 0.2, and 1.1 ± 0.3 for glutathione *S*-transferase A, glutathione reductase, and butyrylcholinesterase, respectively. The excellent correspondence between the two different estimators of experimental variance is shown for glutathione *S*-transferase A in Fig. 2 and for butyrylcholinesterase in Fig. 3.

Discussion

The information about the experimental error inherent in residual plots has previously been discussed in a qualitative way [1,5,13], but the quantitative analysis described in the present paper has to our knowledge not been applied before. A great advantage of the new method is the use of data which are collected for the primary purpose of the investigation, thereby often making extra measurements unnecessary to establish the error structure. Also, the variance estimation will be made under exactly the same conditions as the experiment proper. An important part of the procedure is the original fitting of data to give the residuals, which should be used for estimation of the variance. The fit must be good in all respects and fulfil reasonable goodness-of-fit criteria (cf. refs. 14–16). For this purpose it may be advantageous to make use of some preliminary replicates at high and low velocity values to support the choice of the initial weighting factors (for example, $\alpha = 0$ and $\alpha = 2$ in Eqn. 1 for constant absolute error and constant relative error, respectively) [13]. The iterative analysis of residuals will then start from good conditions for convergence to the 'best' α value. In the case that convergence is not obtained, the weighting factors of the original fitting may have been unsuitable and a new weighting scheme should be attempted with the original data.

It has been suggested in the procedure (see above) that a possibly overdetermined model could be fitted to the data whenever uncertainty exists concerning the proper model. A possible drawback of this practice might be that trends in the data due to statistical fluctuations may be fitted as if they were real expressions of the underlying physical events. In such a case the residuals after regression would be smaller than the values representative of the experimental error. If this condition is suspected it can be tested by comparison of the variance estimated by residuals with that estimated independently by replicative measurements. In case the former value is significantly lower than the latter estimate, a simpler model can be chosen. However, the use of overdetermined models in the derivation of weighting factors would in general not be expected to give biased results. Other problems generally encountered in least-squares methods, such as the influence of outliers, are more serious. The major difficulties concerning overdetermined models are instead directly related to the choice of the 'best' of alternative models [14,15].

Proposed procedure for weighting of data in regression analysis

(i) Collect some information about the error structure; if variance increases with velocity assume $\mathrm{Var}(v) \propto v^2$ and apply weighting factors $w \propto 1/\mathrm{Var}(v)$ in the fitting of an adequate model by regression analysis. In the case of constant variance use no weights ($\alpha = 0$).

(ii) Order the squares of the residuals of the regression analysis as a function of increasing predicted velocity (independent variable) and divide them evenly into groups of at least 5 values. Regress according to Eqn. 1 the mean of the squared residuals of each group on the median of the corresponding velocities of the group (using the reciprocal of the mean as a weighting factor).

(iii) If the α value obtained by the regression analysis of the mean of the squared residuals differs significantly from the value used in the initial fit of the experimental data set, start again from (i) using the new value of α for provisional weighting factors and proceed through (ii). The procedure is repeated until no significant change in the α value is obtained.

It should be noted that somewhat different α values in the weighting factors (e.g. $1.6 \leqslant \alpha \leqslant 2.0$ [1]) do not give significantly different results in the regression analysis of a set of kinetic data. Thus, the convergence criterion for the estimation of α can be set with wide limits and may often be fulfilled without iterations, if the original assumption about the error structure is approximately correct.

(iv) When the α value and thereby the weighting factors have been chosen, the original data set should be analyzed by a final definitive weighted regression analysis.

Acknowledgements

We thank Mrs. Kerstin Larson for expert assistance with some of the analyses. This work was supported by grants (to B.M.) from the Swedish Natural Science Research Council.

References

1 Askelöf, P., Korfsteldt, M. and Mannervik, B. (1976) Eur. J. Biochem. 69, 61—67
2 Haarhoff, K.N. (1969) J. Theoret. Biol. 22, 117—150
3 Morrison, J.F. (1969) in: Least Squares Methods in Data Analysis (Anderssen, R.S. and Osborne, M.R., eds.), pp. 63—69, Australian National University Computer Centre Publication CC 2/69, Canberra
4 Mannervik, B. (1975) BioSystems 7, 101—119
5 Storer, A.C., Darlison, M.G. and Cornish-Bowden, A. (1975) Biochem. J. 151, 361—367
6 Siano, D.B., Zyskind, J.W. and Fromm, H.J. (1975) Arch. Biochem. Biophys. 170, 587—600
7 Box, M.J. (1970) Technometrics 12, 219—229
8 Johansen, G. and Lumry, R. (1961) C.R. Trav. Lab. Carlsberg 32, 185—214
9 Wilkinson, G.N. (1961) Biochem. J. 80, 324—332
10 Jacquez, J.A., Mather, F.J. and Crawford, C.R. (1968) Biometrics 24, 607—626
11 Augustinsson, K.-B., Bartfai, T. and Mannervik, B. (1974) Biochem. J. 141, 825—834
12 Jakobson, I., Askelöf, P., Warholm, M. and Mannervik, B. (1977) Eur. J. Biochem. 77, 253—262
13 Endrenyi, L. (1974) Symp. Biol. Hung. 18, 11—30
14 Bartfai, T. and Mannervik, B. (1972) FEBS Lett. 26, 252—256
15 Mannervik, B. and Bartfai, T. (1973) Acta Biol. Med. Ger. 31, 203—215
16 Reich, J.G., Wangermann, G., Falck, M. and Rohde, K. (1972) Eur. J. Biochem. 26, 368—379