THE JOURNAL OF PHYSICAL CHEMISTRY B

Article

Weighting Formulas for the Least-Squares Analysis of Binding Phenomena Data

Joel Tellinghuisen, and Carl H. Bolster

J. Phys. Chem. B, **2009**, 113 (17), 6151-6157• DOI: 10.1021/jp8112039 • Publication Date (Web): 07 April 2009

Weighting Formulas for the Least-Squares Analysis of Binding Phenomena Data

Joel Tellinghuisen*

Department of Chemistry, Vanderbilt University, Nashville, Tennessee 37235

Carl H. Bolster

U.S. Department of Agriculture—Agricultural Research Service, 230 Bennett Lane, Bowling Green, Kentucky 42104

Received: December 18, 2008; Revised Manuscript Received: February 20, 2009

The rectangular hyperbola, $y = abx/(1 + bx)$, is widely used as a fit model in the analysis of data obtained in studies of complexation, sorption, fluorescence quenching, and enzyme kinetics. Frequently, the "independent variable" x is actually a directly measured quantity, and y may be a simply computed function of x, like $y = x_0 - x$. These circumstances violate one of the fundamental tenets of most least-squares methods—that the independent variable be error-free—and they lead to fully correlated error in x and y. Using an effective variance approach, we treat this problem to derive weighting formulas for the least-squares analysis of such data by the given equation and by all of its common linearized versions: the double reciprocal, y-reciprocal, and x-reciprocal forms. We verify the correctness of these expressions by computing the nonlinear least-squares parameter standard errors for exactly fitting data, and we confirm their utility through Monte Carlo simulations. The latter confirm a problem with inversion methods when the inverted data are moderately uncertain (~30%), leading to the recommendation that the reciprocal methods not be used for such data. For benchmark tests, results are presented for specific data sets having error in x alone and in both x and x_0. The actual estimates of a and b and their standard errors vary somewhat with the choice of fit model, with one important exception: the Deming–Lybanon algorithm treats multiple uncertain variables equivalently and returns a single set of parameters and standard errors independent of the manner in which the fit model is expressed.

Introduction

The rectangular hyperbola

$$y = \frac{abx}{1 + bx} \tag{1}$$

is one of the most frequently encountered nonlinear functional relationships in data analysis, occurring in studies of 1:1 binding,[1–4] sorption,[5–9] fluorescence quenching,[10–13] and enzyme kinetics.[14–17] Historically, to facilitate analysis, this relation has been rendered linear through inversion, to yield the double reciprocal

$$\frac{1}{y} = \frac{1}{abx} + \frac{1}{a} \equiv Bx' + A \tag{2}$$

or the y-reciprocal form

$$\frac{x}{y} = \frac{1}{ab} + \frac{x}{a} = B + Ax \tag{3}$$

A third linear variation is the x-reciprocal form

$$\frac{y}{x} = ab - by \tag{4}$$

These equations have come to be known by different names in different fields. Thus, eq 1 is called simply the binding curve in complexation studies, but is the Langmuir isotherm in sorption work and (with $b = 1/K_m$) the Michaelis–Menten equation of enzyme kinetics. Equation 2 is a version of the Lineweaver–Burk display of enzyme kinetics data, the Benesi–Hildebrand plot of complexation studies, and the Stern–Volmer relation of fluorescence quenching. Equation 3 is the Hanes–Woolf plot of enzyme kinetics. Equation 4 is the Scatchard equation of ligand–protein binding; with simple rearrangement it becomes the Eadie–Hofstee plot of enzyme kinetics.

It is well understood that the transformations from eq 1 to eqs 2–4 alter the relative significance of the data in different ranges, requiring attention to weights in least-squares (LS) analysis.[1,14,16] A subtler but still recognized problem is the fact that often x is a measured property, subject to experimental error, in violation of the usual assumptions of both linear and nonlinear LS. This problem is even more evident in the version of eq 4, where the variable normally considered dependent defines the abscissa of the graphical plot. Some workers have handled these problems through more sophisticated LS methods that allow for uncertainty in both x and y.[7,18,19] However, in many cases y and x are obtained from the same measurement, making them fully correlated. In some such cases it has been possible to reexpress the functional relations in terms of a truly independent variable to satisfy the assumptions of the standard linear and nonlinear LS methods.[6,9,20–22] However, to our knowledge there has been no general treatment of the correlated error problem to permit correct analysis of such data in terms of the preferred

10.1021/jp8112039 CCC: $40.75
Published on Web 04/07/2009

forms of eqs 1−4. In fact, the problem of correlated error in x and y may have been treated correctly only for the case of a linear relationship between the two variables.[23,24]

Here we present a treatment of this correlated error problem that renders statistically equivalent analyses by these four forms in the unusual situation where the uncertain quantity is, in effect, being fitted to a function of itself. Our treatment is in the framework of "effective variance" (EV) methods, in which the uncertainty in x is transformed into an effective contribution to the variance in y.[25−27] This contribution is calculated using the rules of error propagation, giving $\sigma_{\text{eff}}^2 = (\mathrm{d}y/\mathrm{d}x)^2\sigma_x^2$. This term is added to the existing variance in y to give an effective total, $\sigma_{y,\text{tot}}^2 = \sigma_{\text{eff}}^2 + \sigma_y^2$, which is then used to assign weights to the data, $w_i \propto 1/\sigma_{y,\text{tot}}^2$. However, in the LS fitting, x is still treated as error-free. Although the EV method has some obvious limitations, for which it has been criticized,[28−30] the flaws have mostly been discussed in the context of specific data sets, where the EV approach fails to minimize the target quantity—the sum of weighted squared residuals in x and y—whereas methods that treat both x and y as uncertain do achieve such minimization. However, such anecdotal comparisons fail to properly assess the statistical properties of the method. Here, we do that through Monte Carlo simulations, and we show that the statistical limitations of the EV approach are of little practical consequence in analyzing binding constant data.

To give this problem specificity, we consider 1:1 binding of ligand (L) to substrate (S), using the notation of Connors.[1] At equilibrium the concentrations are related by

$$K = \frac{[\mathrm{SL}]}{[\mathrm{S}][\mathrm{L}]} \tag{5}$$

where SL is the complex and K is the equilibrium constant in concentration units. The equilibrium concentrations of S and L obey the mass balance equations

$$S_t = [\mathrm{S}] + [\mathrm{SL}]; \qquad L_t = [\mathrm{L}] + [\mathrm{SL}] \tag{6}$$

where S_t and L_t are the total concentrations. Taking the complexed fraction of S as $f_{11} = [\mathrm{SL}]/S_t$, we can reexpress eq 5 as the isotherm

$$f_{11} = \frac{K[\mathrm{L}]}{1 + K[\mathrm{L}]} \tag{7}$$

according to which f_{11} is studied as a function of the varying equilibrium ligand concentration. Equation 7 contains only one parameter (K), but in typical experiments either S_t is not known (sorption) or the measured quantity is related to [SL] or its equivalent through another parameter (molar absorptivity in spectrophotometric studies), giving two adjustable parameters in the relationship of eq 1. Sometimes, especially for very weak binding, [L] ≫ [SL] at all times, and the approximation [L] ≈ L_t makes [L] in eq 7 and in its double- and y-reciprocal forms effectively error-free, in satisfaction of the LS assumptions. However, this assumption is not required in the era of modern data analysis methods, and it is possible to reexpress the fit relation in terms of L_t and S_t by solving the quadratic

$$(L_t - [\mathrm{SL}])(S_t - [\mathrm{SL}]) = [\mathrm{SL}]/K \tag{8}$$

for [SL] and back substituting to yield the measured quantity, usually ∝[S] or [SL], in terms of L_t (and sometimes S_t), known by assumption from preparation. When there is only a single uncertain measured quantity, this approach is fully correct statistically, requiring only that the fitted quantities be properly weighted inversely as their variances, $w_i \propto \sigma_i^{-2}$. The fit relation is nonlinear in the adjustable parameters, but such problems are routinely handled by programming methods,[1,18,22,31,32] and commercial data analysis programs like Origin and KaleidaGraph.[33] However, the resulting fit relation loses the visual appeal of eq 1 and its linear transformations.

In the present work, we show how weights w_i can be derived to yield statistically correct analysis of binding constant data in terms of eqs 1−4 when the "dependent" variable is directly obtained from measured [SL] or [L], and [L] cannot be approximated by L_t. Formally identical results are obtained for analysis by all of eqs 1−4 and by the transformed fit relation expressed in terms of L_t as just described above. These results are extended to the case where there is uncertainty in both variables, and both cases are confirmed by Monte Carlo simulations, which are also used to investigate the magnitude of the losses when weights are neglected in the popular linearized versions.

Although these five fit relations are formally identical and statistically identical for suitably small data error, they do not yield identical results for the two adjustable parameters when applied to individual data sets. On the other hand, there is a nonlinear LS algorithm that does yield such agreement for all ways of expressing the fit relationship and for any representation of the data uncertainty in both x and y—the algorithm based on Deming's treatment[34] and implemented in iterative form by Jefferys and Lybanon.[29,35] This point does not appear to be widely appreciated, so we devote some effort to its illustration.

Theoretical Considerations

Least Squares. The properties of linear and nonlinear least-squares fitting methods have been covered thoroughly in several of the already cited works,[1,18,31,32] and have been examined through Monte Carlo simulations in application to selected important fit models in earlier works in this journal by one of us.[36,37] Here we need emphasize only several important points:

• Minimum-variance estimation of the adjustable parameters requires that the data be weighted inversely as their variances

$$w_i \propto \sigma_i^{-2} \tag{9}$$

• The variances for the estimated parameters are the diagonal elements of the variance−covariance matrix, of which we distinguish two versions, $\mathbf{V}_{\text{prior}}$ and $\mathbf{V}_{\text{post}}$, both proportional to

$$\mathbf{A}^{-1} = (\mathbf{X}^{\mathbf{T}}\mathbf{W}\mathbf{X})^{-1} \tag{10}$$

where the design matrix $\mathbf{X}$ is as given in the earlier works and the weight matrix $\mathbf{W}$ is diagonal, with elements $W_{ii} = w_i$.

• If the data variances are known a priori, we take $w_i = \sigma_i^{-2}$ and obtain $\mathbf{V}_{\text{prior}} = \mathbf{A}^{-1}$. $\mathbf{V}_{\text{prior}}$ is exact for linear LS (LLS) and exact in the limit of small data error for nonlinear LS (NLS). If the data are normally distributed, the estimated parameters will be normally distributed for LLS and normal in the small data error limit for NLS.

• In LLS $\mathbf{V}_{prior}$ depends on only the x-structure and the error structure of the data; in NLS it may depend also on the values of the y_i and the fit parameters.

• If the data variances are not known absolutely, we use $\mathbf{V}_{post} = s_y^2\,\mathbf{A}^{-1}$, where s_y^2 is the estimated variance for data of unit weight and is calculated from the fit residuals δ_i using

$$s_y^2 = \frac{\sum w_i \delta_i^2}{n - p} = \frac{S}{\nu} \tag{11}$$

with the number of statistical degrees of freedom ν being the difference between the numbers of data points and adjustable parameters. The parameter variances are now estimated quantities having the statistical properties of s_y^2, which is a scaled χ^2 variate. From these properties, $\mathbf{V}_{post}$-based parameter standard error estimates have relative standard error $(2\nu)^{-1/2}$ in LLS; in NLS this variability is added to that inherent in $\mathbf{V}_{prior}$.

• There are two consequences of violating the weighting prescribed by eq 9: (a) the parameter estimates are no longer minimum variance; and (b) the $\mathbf{V}$-based parameter error estimates are unreliable.

• The main violation of eq 9 is the use of unweighted LS to analyze heteroscedastic data. The precision losses and $\mathbf{V}$ errors depend mainly on the range of the true σ_i over the data set, being great for strong heteroscedasticity and vanishing when the σ_i become constant.

• Since $\mathbf{V}_{prior}$ is exact for NLS in the small-error limit, we can use error-free data to predict the parameter errors. In general, nonlinear estimators are biased and nonnormal, and the bias and deviation from normality increase with the data error. Still, the $\mathbf{V}_{prior}$-based parameter standard errors remain reliable as long as they are smaller than 1/10 the magnitude of the parameters, whereupon they are typically within 10% of true, a behavior we refer to as the "10% rule of thumb".[36]

A fundamental assumption of LLS is that the independent variable or variables are error-free and there is a single uncertain dependent variable. Then S in eq 11 is the LS minimization target. That assumption carries over into most NLS algorithms but it is not correct when more than one variable has statistical uncertainty. In Deming's formalism,[34] all variables are treated on equal terms, and the target of the minimization procedure is the expanded sum of squared residuals

$$S = \sum w_{xi}\delta_{xi}^2 + w_{yi}\delta_{yi}^2 + \ldots \tag{12}$$

over all uncertain variables. Here, all variables are adjusted iteratively in the fit, and each residual is the difference between measured and adjusted value, e.g., $\delta_{xi} = x_{adj,i} - x_i$. If the weights are again taken as the inverse variances in x, y,..., the resulting $\mathbf{V}_{prior}$ is again exact in the small-error limit. Furthermore, in the iterative implementation of Deming's approach in the algorithms of Jeffreys[35] and Lybanon,[29] the LS outcome is independent of the manner in which the fit model is expressed, meaning that analysis by all of eqs 1−4 yields identical results for any data set.

From the foregoing, the choice of a fit relationship for the analysis of binding constant data becomes moot if the Deming approach is taken. However, familiarity with that approach is still limited and so in the present work we show how with suitably defined weights, such data can be analyzed with the user's choice of eqs 1−4, with confidence that the results will be at least statistically equivalent, though not numerically identical in all cases. In demonstrating this, we consider the two common extremes of data error: constant σ_i and σ_i proportional to the measured quantity.

Weighting Expressions. We treat a single case, motivated by the 1:1 binding considerations in eqs 5−8. Let the equilibrium ligand concentration [L] be a measured quantity x that is subject to experimental error. The concentration of complex [SL] is then $y = x_0 - x$, where x_0 is the prepared initial concentration (L_t) of ligand and for now is assumed to be error-free. The parameter a in eqs 1−4 then becomes S_t and $b = K$, both taken as adjustable parameters to be determined from the fits.

We first consider treatment via what we call the direct approach, in which the true dependent variable x is expressed as a function of the error-free independent variable x_0 by solving the quadratic eq 8, which with the current substitutions reads

$$x^2 + x(a - x_0 + 1/b) - x_0/b = 0 \tag{13}$$

This approach is fully consistent with the fundamental tenet of LLS noted above. The fit relation is nonlinear in the adjustable parameters, but with proper weighting it still must yield correct results for a and b and their standard errors in the small-error limit. Then the only question is the extent to which these predictions lose validity as the data error increases. Weighting is simple for our two data error structures, since the weights are just σ_{xi}^{-2} in both cases. However, consideration of proportional data error for real data does force a decision between using the observed x_i or its LS prediction in computing $\sigma_{xi} = cx_i$. MC tests generally show better results for the latter choice, though this makes the adjustment of the weights a part of the iterative computation (which is true also for LLS with proportional error).

Next consider analysis using eq 1, which we will call the binding curve (BC). Consistent with the premise of fitting an uncertain y to a function of an error-free x, we seek an effective σ_y that correctly accounts for the actual uncertainty in x. From its definition as $y = x_0 - x$, there is the obvious direct contribution, $\sigma_{y,dir} = \sigma_x$. There is a second, indirect contribution $\sigma_{y,ind}$ that arises as follows. Let a point on the true curve be subject to an error ε_x in x. This produces a direct error $\varepsilon_y = -\varepsilon_x$; it also produces an effective or indirect error $-(dy/dx)\varepsilon_x$, through its displacement of the fit function to $(x + \varepsilon_x)$. The two errors are fully correlated, and they lead to a total effective σ_y that is a sum of two contributions,

$$\sigma_{eff,1} = \sigma_x\left[1 + \frac{ab}{(1 + bx)^2}\right] \tag{14}$$

leading straightforwardly to the needed weights w_i. [Mathematically, the two contributions to $\sigma_{eff,1}$ appear to have opposite signs, but the perfect anticorrelation between y and x makes them add in magnitude.]

As a variation on the BC fit, we consider fitting y/x to the appropriately modified form of eq 1. In deriving eq 14 we have already taken into account both contributions to the effective error in y, so simple error propagation yields for this modified BC fit, $\sigma'_{eff,1} = \sigma_{eff,1}/x$. Alternatively, we can obtain the same result by again computing the direct and indirect contributions to the effective error in $y/x = x_0/x - 1$. Noting that the two contributions now add with opposite signs (because of the negative slope of y/x vs x), we again obtain $\sigma'_{eff,1} = \sigma_{eff,1}/x$.

Simple error propagation suffices also to obtain correct weights for fits to eqs 2 and 3. Thus, $\sigma_{1/y} = \sigma_y/y^2$, and substitution for y using eq 1 yields

$$\sigma_{\mathrm{eff},2} = \sigma_x \frac{(1 + bx)^2 + ab}{(abx)^2} \tag{15}$$

Again using error propagation, $\sigma_{\mathrm{eff},3} = x\sigma_{\mathrm{eff},2}$. For eq 4, the pseudoindependent variable is now y; and similar considerations yield

$$\sigma_{\mathrm{eff},4} = [1 + b(a - y)]\frac{\sigma_x}{x} + b\sigma_x \tag{16}$$

Now suppose we allow for independent random error in x_0. Considering first the direct model of eq 13, we treat x_0 as a source of indirect error in x, in the same way that we treated x with respect to y in fitting with eqs 1−4. This indirect contribution is

$$\sigma_{\mathrm{ind},x_0} = \frac{\mathrm{d}x}{\mathrm{d}x_0}\sigma_{x_0} = \sigma_{x_0}\frac{x + 1/b}{2x + a - x_0 + 1/b} \tag{17}$$

Since the two contributions are assumed to be independent, their variances add, giving a total that is $\sigma_x^2 + \sigma_{\mathrm{ind},x_0}^2$, for any assumption about the natures of σ_x and σ_{x_0} individually.

For fitting to the binding curve (eq 1), x_0 occurs only in y. From its assumed independence, we need only add $\sigma_{x_0}^2$ to $\sigma_{\mathrm{eff},1}^2$ from eq 14 to obtain the effective total variance. This result then leads directly to those for fitting to the modified BC (y/x), $1/y$, and x/y by simple error propagation, just as was used to derive the results given above for the case of error in x only. Equation 4, however, requires more care, because x_0 occurs on both sides of the equation and so makes both direct and indirect contributions to the effective error. The result is of the same form as eq 16

$$\sigma_{\mathrm{eff},4,x_0} = [1 + b(a - y)]\frac{\sigma_{x_0}}{x_0} + b\sigma_{x_0} \tag{18}$$

The total variance is then $\sigma_{\mathrm{eff},4}^2 + \sigma_{\mathrm{eff},4,x_0}^2$.

The above results bear some resemblance to expressions presented by Connors[1] in his Table 3.1; however, he did not consider correlated effects. While our treatment correctly accounts for the indirect contributions from uncertainty in x in the EV framework, the *direct* contributions are specific for our assumption that $y = x_0 - x$. Thus, if this relation is not correct, the results will need to be modified. For example, suppose that $y = (x_0 - x)\,C$, where C is a capacity factor (as, e.g., in sorption work[8,9]). Then $\sigma_{y,\mathrm{dir}} = \sigma_x C$ and in eq 14, the first 1 is replaced by C. If there is an independent contribution from uncertainty in x_0, it becomes $C^2\sigma_{x_0}^2$ in place of $\sigma_{x_0}^2$. Similar changes are required for the linearized fit relations, the results for all of which are summarized in Table 1.

We can compute the exact $\mathbf{V}_{\mathrm{prior}}$-based σ_a and σ_b for all of these models by using exactly fitting data of selected $x_{0,i}$ for adopted values of a and b; indeed just such calculations have been used to verify the results given in Table 1. However, in the analysis of actual data we must adjust the weights iteratively in conjunction with the changing values of a and b in the iteration. All of the expressions in Table 1 are in forms that facilitate such adjustment. In this regard, note that eqs 1−4 are all nonlinear LS models when expressed in terms of the adjustable parameters a and b, in which form the analyses yield directly the parameter error estimates. Analysis of the linearized forms with LLS yields identical estimates of a and b but does not directly yield correct error estimates of b in eqs 2 and 3 or a in eq 4, because A and B (and a and b) are correlated.[38]

TABLE 1: Summary of Effective-Variance-Based Weighting Expressions for the Least-Squares Analysis of Data Following the Equation $y = C(x_0 - x) = abx(1 + bx)^{-1}$ [a]

relation (eqn)	σ_{eff}	$\sigma_{\mathrm{eff},x0}$
direct (13)[b]	σ_x	$\sigma_{x0}\frac{x + 1/b}{2x + a/C - x_0 + 1/b}$
BC (1)	$\sigma_x\left[C + \frac{ab}{(1+bx)^2}\right]$	$\sigma_{x0}\,C$
BC/x	$\frac{\sigma_x}{x}\left[C + \frac{ab}{(1+bx)^2}\right]$	$\frac{\sigma_{x0}}{x}\,C$
double reciprocal (2)	$\sigma_x\frac{C(1+bx)^2 + ab}{(abx)^2}$	$\sigma_{x0}\,C\frac{(1+bx)^2}{(abx)^2}$
y-reciprocal (3)	$\sigma_x\,x\frac{C(1+bx)^2 + ab}{(abx)^2}$	$\sigma_{x0}\,x\,C\frac{(1+bx)^2}{(abx)^2}$
x-reciprocal (4)	$\frac{\sigma_x}{x}\,[C + b(a-y)] + \sigma_x\,bC$	$\frac{\sigma_{x0}}{x_0}\,[C + b(a-y)] + \sigma_{x0}\,bC$

[a] C is presumed to be a known constant of negligible uncertainty. Weights are $w = \sigma_{\mathrm{tot}}^{-2}$, where $\sigma_{\mathrm{tot}}^2 = \sigma_{\mathrm{eff}}^2 + \sigma_{\mathrm{eff},x_0}^2$ and quantities are evaluated for each point using the relevant x_i and $x_{0,i}$ values. Errors in x and x_0 are assumed to be independent. [b] For $y = C\,(x_0 - x)$, replace a in eq 13 by a/C.

Computational Methods

The Monte Carlo simulations were carried out on a laptop PC using programs coded in Microsoft FORTRAN. The procedures were otherwise as described in the earlier papers.[36−38] For testing the weighting formulas given above, we used the KaleidaGraph program (Synergy Software)[33] to fit exact data. This program has a General nonlinear LS routine that employs the Marquardt algorithm and provides $\mathbf{V}_{\mathrm{prior}}$-based parameter errors for all weighted fits, so is a valuable asset in such situations.

In all of the simulations for a single uncertain variable, random error is added only to x, consistent with the premise of this study. When x_0 is also taken as uncertain, error is added to it using a second random number, consistent with the assumption that the two are independent. Data are then analyzed for each simulated data set using one of the several test models with weights as specified in Table 1. In selected cases, results are also accumulated for unweighted analysis to assess the magnitude of the precision losses for neglect of weights.

As has been noted already, NLS fitting yields nonnormal parameter distributions and biased parameter estimates. From the earlier studies,[36,37] the parameter biases increase with increasing parameter uncertainty, initially scaling with the parameter variance. Since parameter variances scale with the data variance for small data error, the biases also go as σ_x^2 for small error. The actual parameter standard errors exceed the $\mathbf{V}_{\mathrm{prior}}$ predictions increasingly with increasing data error, but seldom by more than 10% when the parameter standard error is less than 1/10 the magnitude of the parameter, behavior referred to earlier as the 10% rule of thumb.

TABLE 2: Synthetic Binding Constant Data for $x_0 - x = abx(1 + bx)^{-1}$, with $a = b = 1$ and Uncertainty in x Alone[a]

x_0	x_{const}	x_{prop}
0.1	0.02	0.0535
0.2	0.10	0.108
0.5	0.31	0.273
1.0	0.56	0.63
2.0	1.49	1.43
5.0	4.21	4.47
10.0	9.04	9.28

[a] Data generated for constant uncertainty in x ($\sigma_x = 0.04$) and proportional ($\sigma_x = 0.04x_{true}$). "Exact" standard errors for this model: $\sigma_{a,const} = 0.051\,056$, $\sigma_{b,const} = 0.197\,92$, $\sigma_{a,prop} = 0.14159$, $\sigma_{b,prop} = 0.181\,39$.

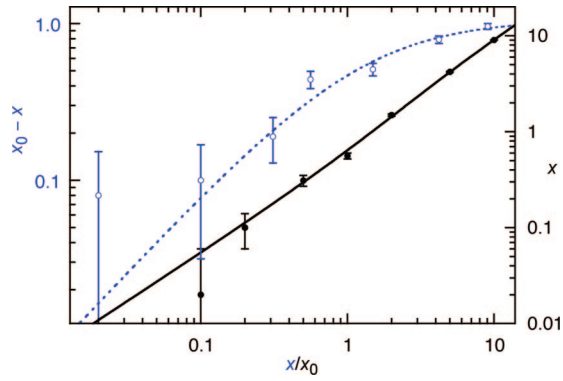

Figure 1. Illustration of the constant-error data set in Table 2 as analyzed using the direct relation, $x = x(x_0)$ (solid points and curve, ordinate scale to right) and the binding constant relation of eq 1. Logarithmic axis scales are used for clarity of display of the chosen geometric structure of x_0. Error bars represent σ_x (direct) and $\sigma_{eff,1}$ (BC).

Results and Discussion

We have verified the weighting expressions derived above for several different choices of the adjustable parameters, using MC computations in the small-error limit to check the $\mathbf{V}_{prior}$-based parameter standard errors. For initial illustrations here, we consider a model having $a = b = 1$ and a 7-point geometric data structure given in Table 2. We include in this table one specific synthetic data set for each of our two data error structures. Some results obtained by analyzing these data are illustrated in Figure 1, and complete results are presented in Table 3 with excess precision to facilitate numerical benchmark testing. Note that identical results are returned by several of the fit models when weights are properly employed. Also, all the results for each data structure are reasonably consistent when the parameter estimates are referenced to their standard errors, with the exception of the anomalous χ^2 values for the reciprocal models in both cases (see below). To the extent that there are any "correct" results, they would be those for the direct method, since these are identical to those from the Deming−Lybanon algorithm, which yields identical results for all correct ways of expressing the fit relation as $f(x,x_0) = 0$, including the implicit form of eq 13.

The illustrations are extended to uncertainty in both x and x_0 in Tables 4 and 5, where the same data error structures are adopted for both variables but with different values of a and b. The comparsions are similar to those that were made in discussion of Table 2. In particular, no method stands out as particularly good or bad, with the exception again of the Deming−Lybanon algorithm, which renders a single set of results for all ways of expressing the fit relation. Again, the key to success in all these methods is the use of correct data weights.

TABLE 3: Results from Analyzing Test Data from Table 2 with the Several Fit Models[a]

fit model	a	b	χ^2
constant data error			
direct[b]	1.05978 (5721)	0.80020 (15474)	7.97154
BC[c]	1.06352 (5847)	0.78056 (15458)	7.87286
reciprocal[d]	1.08395 (6202)	0.70510 (14027)	4.56618
x-reciprocal	1.08838 (5977)	0.71367 (13363)	8.68918
proportional data error[e]			
direct[b]	0.90078 (13016)	1.08274 (20182)	3.76362
BC[c]	0.90299 (13169)	1.08247 (20320)	3.72268
reciprocal[d]	0.87856 (12811)	1.11371 (20955)	4.63662
x-reciprocal	0.95020 (15333)	1.01256 (20914)	3.72136

[a] Figures in parentheses are a priori standard errors, in terms of final digits; a posteriori values obtained by multiplying by $(\chi^2/5)^{1/2}$. The results are thought to be numerically valid to the stated numbers of digits. [b] Use of eq 13, with x represented as an explicit function of x_0; identical results obtained fitting to all models using Deming−Lybanon algorithm with stated uncertainty in x. [c] Binding constant model of eq 1. Identical results obtained by fitting y/x values to appropriately modified form of this equation. [d] Identical results obtained using the double- and y-reciprocal forms of eqs 2 and 3. [e] Data error taken as $\sigma_{xi} = 0.04x_i$ for all but direct model, where $\sigma_{xi} = 0.04x_{calc}$ was used, with iterative adjustment.

TABLE 4: Synthetic Binding Constant Data for $a = 2.1$, $b = 1.2$, and Uncertainty in Both x and x_0[a]

$\sigma_x = \sigma_{x0} = 0.04$		$\sigma_x = 0.04x$; $\sigma_{x_0} = 0.04x_0$	
x_0	x	x_0	x
0.09	0.05	0.103	0.029
0.17	0.08	0.192	0.065
0.50	0.14	0.536	0.157
1.00	0.36	1.00	0.364
2.06	0.98	2.03	0.90
5.01	3.37	5.05	3.64
10.04	8.03	9.52	7.90

[a] "Exact" standard errors for this model: $\sigma_{a,const} = 0.074\,402$, $\sigma_{b,const} = 0.18190$, $\sigma_{a,prop} = 0.251\,24$, $\sigma_{b,prop} = 0.182\,26$.

Since binding constant data have typically been analyzed with complete neglect of weights, it is of interest to ask what is lost thereby. A guide to the potential losses is the range of weights w_i across the data set, since when these weights become constant the analysis becomes de facto unweighted. Table 6 shows the range of $\sigma_{eff,i}$ for the data model of Table 2 in the different fit models. In general, if the w_i span a range of less than a factor of 10, the losses from neglect of weights should be modest.[39] On this basis, we would expect significant loss of precision in all cases of unweighted analysis except the direct and BC fits of constant-σ_x data and the BC/x analysis of proportional data error. The latter result shows how the fit model can be chosen to neutralize the data error in some cases; however, this assumes the data error is known, in which case the analysis should anyway be weighted.

We illustrate the effects of different weighing assumptions in Figures 2 and 3, which display the biases in b (K) and in its standard error from the MC computations. Parameters are inherently biased in NLS, though with proper weighting the bias is generally much smaller than the parameter standard error. Neglect of weights produces larger parameter biases, and systematic errors for all σ_x in the parameter standard errors (Figure 3). The actual precision loss is only 5% for unweighted analysis with the BC fit, but the analyst would not know this,

TABLE 5: Results from Analyzing Test Data from Table 4 with the Several Fit Models[a]

fit model[b]	a	b	χ^2
constant data error			
D-L[c]	2.22362 (8319)	0.96840 (14278)	3.29775
direct	2.22304 (8315)	0.96938 (14238)	3.30576
BC	2.22876 (8422)	0.95527 (14295)	3.28884
reciprocal	2.24249 (8594)	0.92098 (13786)	4.57076
x-reciprocal	2.24040 (8289)	0.93507 (13353)	3.35703
proportional data error[d]			
D-L[c]	1.99857 (23916)	1.27221 (19432)	9.66586
direct	2.00781 (23828)	1.25599 (19055)	10.07405
BC	1.98980 (23941)	1.27165 (19437)	9.66765
reciprocal	1.97988 (24108)	1.25839 (19442)	10.65092
x-reciprocal	2.06054 (26775)	1.22891 (20327)	9.35353

[a] Figures in parentheses are a priori standard errors, in terms of final digits; a posteriori values obtained by multiplying by $(\chi^2/5)^{1/2}$. [b] As defined in Table 3; identical results obtained for same corresponding models. [c] Deming−Lybanon algorithm; identical results obtained for all ways of expressing fit relation in this algorithm. [d] Weights based on $\sigma_{xi} = 0.04x_i$ and $\sigma_{x0i} = 0.04x_{0i}$ for all but D-L and direct models; iteratively adjusted values of x_i and x_{0i} used in D-L model, adjusted x_i used in direct model.

TABLE 6: Ranges (Maximum/Minimum) of σ_{eff} for Data Model of Table 2 in Different Fit Models[a]

fit model	constant σ_x	$\sigma_x \propto x$
direct	1.0	180
BC	1.9	94
BC/x	340	1.9
1/y	640	8.5
x/y	8.5	49
x-reciprocal	35	5.1

[a] From eqs 14−16 and their simple modifications for all but direct model, where $\sigma_{eff} = \sigma_x$.

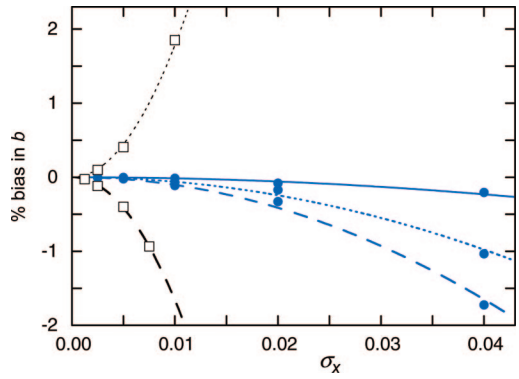

Figure 2. Parameter bias from MC computations (10^5 data sets each point) for analysis of constant-σ_x data in model of Table 2. Solid points, analysis with BC fit; open squares, x/y fit. Curves are fits of points to $c\sigma_x^2$ (c adjustable): solid, theoretical w_i; fine dash, w_i adjusted in each fit in accord with estimates of a and b; broad dash, $w_i = 1$. For the x/y fit, both biases at $\sigma_x = 0.01$ are ~40% of σ_b.

since the $\mathbf{V}_{post}$ estimate is the only available result without the MC computations; and this value is too conservative by ~15%. Unweighted analysis with the x/y fit yields a factor of 2 increase in σ_b (factor of 4 loss in efficiency), but here at least the $\mathbf{V}_{post}$ estimate is close to true. By contrast, proper weighting in the BC analysis yields only a 1% excess in the MC σ_b estimate at $\sigma_x = 0.04$, and a 3% overshoot in the $\mathbf{V}_{post}$ estimate (results not shown). Since the $\mathbf{V}_{prior}$ value for this case is $\sigma_b = 0.198$ (Table 2), or ~20% of b, these overshoots are well within the guidelines of the 10% rule.[36]

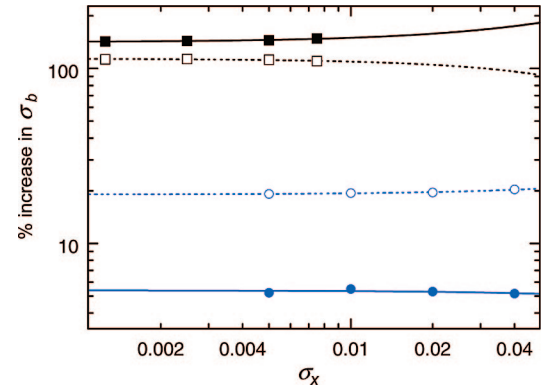

Figure 3. Loss of precision for estimating b from neglect of weights, from MC computations (10^5 data sets each point) for analysis of constant-σ_x data in model of Table 2. Solid points and curves, actual, from MC statistics; open points and dashed curves, apparent, from MC statistics of $\mathbf{V}_{post}$ estimates. Round points, BC analysis; squares, from x/y fit.

The $\mathbf{V}_{post}$-based estimates shown in Figure 3 are the MC rms values, which are statistically preferred to the means. It is worth noting that these also have large relative uncertainty of 37%, exceeding the predicted $(2\nu)^{-1/2}$ or 32%, from the properties of χ^2. In fact, even the $\mathbf{V}_{prior}$-based estimates show variability of 10% and 21%, respectively, for the BC estimates of the standard errors in a and b. This is from the inherent dependence of the $\mathbf{A}$ matrix on the specific values of the parameters and the y_i, and it accounts for differences between the parameter standard errors quoted in Tables 3 and 5 and their "exact" values given in Tables 2 and 4.

Figure 2 shows more pronounced biases for the x/y analysis, and in fact the relevant MC computations diverged rapidly beyond $\sigma_x = 0.01$, giving many nonconvergent data sets and parameter standard errors greatly in excess of the 10% rule of thumb. These anomalies are attributed to the pathological statistical properties of the x/y values for the first two points, where x and y are both occasionally ≈ 0, as we confirmed by repeating the MC simulations with deletion of these two points. This is more than an inconvenience in the MC computations; it is a real problem that argues against such inverse methods any time the data have large relative uncertainty.[2] In fact, if y is normally distributed, $1/y$ is not only nonnormal, it does not even have finite variance, violating one of the fundamental requirements of LS. With small relative error in y, such problems remain of little practical concern, but as the relative error increases, it results in divergent MC results and unreliable results for single data sets.[37] By contrast, no such problems occur in the fitting of x or y directly. However, they can occur in fit models that entail division by x, if the data set includes any x values uncertain by more than ~30%.

Avoiding anomalies concerning relatively uncertain data values, like those just discussed, the bias and precision loss from incorrect weighting depend mainly on the range of w_i over the data set. Thus, it is worth asking what is the effect of changing the selection of x_0 values to reduce the effect of the weights. For the constant-σ_x model of Table 2, dropping points at small x_0 and replicating those at large actually increases the precision. Thus, for example, dropping all points below $x_0 = 1$ and replicating those for $x_0 = 1-5$ reduces the parameter uncertainties by ~10% while also dropping the range of w_i to negligible levels for the x/y model (already negligible for direct and BC). For this reason and those discussed in the preceding paragraph, it is wise not to include very small x_0 points in the data set when σ_x is constant. For the proportional error structure,

dropping points reduces the parameter precision, but some choices give little loss while reducing weighting problems. Thus, dropping the two highest x_0 values and doubling the two lowest renders the x-reciprocal (Scatchard) model de facto almost constant w_i while increasing the parameter standard errors by only ~10%. Of course, to take advantage of such choices, one anyway needs to know the data error structure.

Conclusion

Using an effective variance approach, we have derived weighting formulas for the analysis of binding constant data in all of the least-squares fit models commonly used to treat such data. In most of these relationships, the "independent" variable x—the equilibrium concentration of ligand—is a measured quantity, subject to experimental error. At the same time, the "dependent" variable y is a simple function of this quantity, meaning x and y in the fit relation are fully correlated. Our treatment accounts for this correlation, yielding statistically identical results for all versions of the fit model in the small-error limit. Monte Carlo simulations verify these weighting expressions and quantify the effects of incorrect weighting. They also demonstrate the need to take care when using any of the reciprocal methods on data having large relative uncertainty, since inversion of such data is statistically unwise.

With the proviso just noted, one can trust the results of a properly weighted LS analysis of binding constant data carried out with any of the common fit models. The several models can be expected to yield different but statistically consistent numerical values for the constants. On the other hand, the Deming−Lybanon algorithm, which treats all variables on equal fitting, yields only a single set of parameters and parameter standard errors for all choices of the fit relationship. Thus the D-L results can in some sense be claimed as the "correct" values. However, when considered from the standpoint of the statistical properties of the resulting estimated LS parameters, the effective variance methods are essentially equivalent.

Of course the key to proper use of any of these fit models is reliable knowledge about the data error structure, since for implementation, all of the weighting expressions require at least σ_x for each data point. Data variance functions can be derived straightforwardly for the instrumental contribution, by simply repeating measurements for analytes over a suitable range of concentration or amount.[40,41] However, this measurement variance may be a minor component of the total uncertainty for a particular method, and replication of entire experiments may be needed to fully characterize the method variance.

Finally, it is worth bearing in mind that even perfect characterization of the data variance cannot yield good results if the data sample a poorly chosen region of the binding phenomenon in question. The important matter of data coverage has only been touched on in the present study but has been considered comprehensively in many previous works, including refs 1 and 2. Although those studies did not consider the weighting problems addressed here, their conclusions about the dependence of precision on the data range should remain generally valid.

References and Notes

(1) Connors, K. A. *Binding Constants: The Measurement of Molecular Complex Stability*; Wiley: New York, 1987.
(2) Bowser, M. T.; Chen, D. D. Y. *J. Phys. Chem. A* **1998**, *102*, 8063–8071.
(3) Chaires, J. B. *Methods Enzymol.* **2001**, *340*, 3–22.
(4) Tanaka, Y.; Terabe, S. *J. Chromatogr. B* **2002**, *768*, 81–92.
(5) Persoff, P.; Thomas, J. F. *Soil Sci. Soc. Am. J.* **1988**, *52*, 886–889.
(6) Bothwell, M. K.; Walker, L. P. *Bioresour. Technol.* **1995**, *53*, 21–29.
(7) Schulthess, C. P.; Dey, D. K. *Soil Sci. Soc. Am. J.* **1996**, *60*, 433–442.
(8) Bolster, C. H.; Hornberger, G. M. *Soil Sci. Soc. Am. J.* **2007**, *71*, 1796–1806.
(9) Bolster, C. H. *J. Environ. Qual.* **2008**, *37*, 1986–1992.
(10) Laws, W. R.; Contino, P. B. *Methods Enzymol.* **1992**, *210*, 448–463.
(11) Di Marco, G.; Lanza, M.; Mamo, A.; Stefio, I.; Di Pietro, C.; Romeo, G.; Campagna, S. *Anal. Chem.* **1998**, *70*, 5019–5023.
(12) Clements, J. H.; Webber, S. E. *J. Phys. Chem. A* **1999**, *103*, 2513–2523.
(13) Amao, Y. *Microchim. Acta* **2003**, *143*, 1–12.
(14) Dowd, J. E.; Riggs, D. S. *J. Biol. Chem.* **1965**, *240*, 863–869.
(15) Johnson, M. L.; Faunt, L. M. *Methods Enzymol.* **1992**, *210*, 1–37.
(16) Di Cera, E. *Methods Enzymol.* **1992**, *210*, 68–87.
(17) Ritchie, R. J.; Prvan, T. *J. Theor. Biol.* **1996**, *178*, 239–254.
(18) Johnson, M. L. *Methods Enzymol.* **1992**, *210*, 106–117.
(19) Valsami, G.; Iliadis, A.; Macheras, P. *Biopharm. Drug Dispos.* **2000**, *21*, 7–14.
(20) Meinert, C. L.; McHugh, R. B. *Math. Biosci.* **1968**, *2*, 319–338.
(21) Feldman, H. A. *Anal. Biochem.* **1972**, *48*, 317–338.
(22) Munson, P. J.; Rodbard, D. *Anal. Biochem.* **1980**, *107*, 220–239.
(23) York, D. *Earth Planet. Sci. Lett.* **1969**, *5*, 320–324.
(24) York, D.; Evensen, N. M.; Martinez, M. L.; Delgado, J. D. B. *Am. J. Phys.* **2004**, *72*, 367–375.
(25) Clutton-Brock, M. *Technometrics* **1967**, *9*, 261–269.
(26) Barker, D. R.; Diana, L. M. *Am. J. Phys.* **1974**, *42*, 224–227.
(27) Orear, J. *Am. J. Phys.* **1982**, *50*, 912–916.
(28) Chandler, J. P. *Technometrics* **1972**, *14*, 71–76.
(29) Lybanon, M. *Am. J. Phys* **1984**, *52*, 22–26.
(30) Lybanon, M. *Am. J. Phys.* **1984**, *52*, 276–278.
(31) Bevington, P. R. *Data Reduction and Error Analysis for the Physical Sciences*; McGraw-Hill: New York, 1969.
(32) Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. *Numerical Recipes*; Cambridge Univ. Press: Cambridge, UK, 1986.
(33) Tellinghuisen, J. *J. Chem. Educ.* **2000**, *77*, 1233–1239.
(34) Deming, W. E. *Statistical Adjustment of Data*; Dover: New York, 1964.
(35) Jefferys, W. H. *Astron. J.* **1980**, *85*, 177–181.
(36) Tellinghuisen, J. *J. Phys. Chem. A* **2000**, *104*, 2834–2844.
(37) Tellinghuisen, J. *J. Phys. Chem. A* **2000**, *104*, 11829–11835.
(38) Tellinghuisen, J. *J. Phys. Chem. A* **2001**, *105*, 3917–3921.
(39) Tellinghuisen, J. *Analyst* **2007**, *132*, 536–543.
(40) Tellinghuisen, J. *Appl. Spectrosc.* **2000**, *54*, 431–437.
(41) Zeng, Q. C.; Zhang, E.; Dong, H.; Tellinghuisen, J. *J. Chromatogr. A* **2008**, *1206*, 147–152.

JP8112039