JOURNAL OF CHEMICAL EDUCATION

Article

Cite This: J. Chem. Educ. 2018, 95, 970–977

pubs.acs.org/jchemeduc

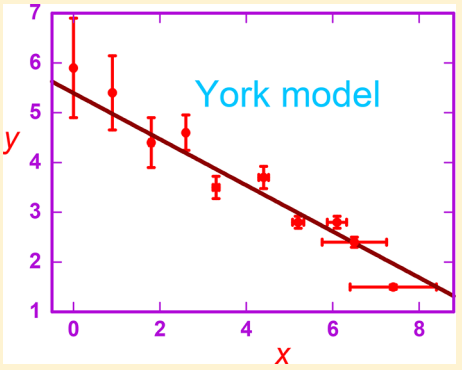
Least-Squares Analysis of Data with Uncertainty in *y* and *x*: Algorithms in Excel and KaleidaGraph

Joel Tellinghuisen*

Department of Chemistry, Vanderbilt University, Nashville, Tennessee 37235, United States

Supporting Information

ABSTRACT: For the least-squares analysis of data having multiple uncertain variables, the generally accepted best solution comes from minimizing the sum of weighted squared residuals over all uncertain variables, with, for example, weights in x_i taken as inversely proportional to the variance σ_{xi}^2. A complication in this method, here called "total variance" (TV), is the need to calculate residuals δ_i in every uncertain variable. In *x*–*y* problems, that means adjustments must be obtained for *x* as well as for the customary *y*. However, for the straight-line fit model, there is a simpler procedure, a version of effective variance (EV) methods, that requires only the residuals in *y* and agrees exactly with the TV method. Furthermore, Monte Carlo calculations have shown that this EV_2 method is statistically comparable to the TV method for many common nonlinear fit models. This method is easy to code for computation in Excel and programs like KaleidaGraph, as is illustrated here for several examples, including implicit nonlinear models. The algorithms yield estimates of both the parameters and their standard errors and can be used as well for more traditional problems requiring weighting in *y* only.



KEYWORDS: *Upper-Division Undergraduate, Graduate Education/Research, Physical Chemistry, Laboratory Instruction, Analytical Chemistry, Problem Solving/Decision Making, Chemometrics*

The straight line is surely one of the most widely used mathematical models for the analysis of data, and its far-and-away most common quantitative implementation is the method of least-squares (LS), going back over 200 years to Gauss (probably[1]). The properties of such linear LS (LLS) solutions are well-known and frequently stated. They include importantly that the standard errors (SE) of the parameter estimates are exactly predictable if the error structure of the data is known; the estimates are normally distributed if the data error is normal, and even if it is not, in the limit of a large number of points, where the central limit theorem ensures normality.[2] I have maintained that these and other properties of LLS solutions are best appreciated through Monte Carlo (MC) simulations.[3]

An important premise of LLS is that all the statistical error resides in a single dependent variable, commonly taken to be *y* in *x*–*y* problems. If this assumption fails only weakly, with relative error in *x* being much less than that in *y*, the LLS results remain adequate for most purposes.[4] However, there are situations where the relative errors in *x* and *y* are comparable, and then more complex methods are needed. Going back at least to Deming's work,[5] the "best" solution has been assumed to be that which minimizes

$$S_{TV} = \sum w_{xi}\delta_{xi}^2 + w_{yi}\delta_{yi}^2 + \ldots \quad (1)$$

where the δ_i's are residuals in the uncertain variables, here just *x* and *y* but readily extended to more than two variables. A complication in specifying and minimizing S_{TV} is the need for "calculated" or "adjusted" values (designated Y_i and X_i) for both *y* and *x*, giving residuals $Y_i - y_i$ and $X_i - x_i$, where capital letters represent the adjusted values. In LLS, $Y_i = y_{calc}(x_i)$ and $X_i = x_i$; when both variables are uncertain, the calculated points are displaced with respect to both y_i and x_i. Minimum-variance estimates of the model parameters are assumed to result from the use of weights inversely proportional to variance, $w_{xi} = C\sigma_{xi}^{-2}$, $w_{yi} = C\sigma_{yi}^{-2}$, and so on (*C* a single constant for all variables), as holds rigorously for LLS (where only *y* is uncertain).

The subscript TV in eq 1 stands for "total variance", which unfortunately has another meaning in the writings of some statisticians, who have included under this label methods like "orthogonal regression", which pay no attention to differences in the relative precisions of *x* and *y*. This means that results can change with simple changes in scale, among other deficiencies. Statisticians have also used terminology like "errors in variables (EIV) models" and "Model II regression", but both of these also seem inadequately descriptive. Physical scientists have occasionally used "generalized least squares" for cases with uncertainty in multiple variables, but that term means something altogether different among statisticians, namely, the simultaneous estimation of the variance function and the response function for a data set.[6] At the same time, "Deming regression" has incorrectly come to refer to just the straight-line model with two uncertain variables, when in fact Deming's treatment was completely

Received: February 2, 2018
Revised: March 29, 2018
Published: April 19, 2018

DOI: 10.1021/acs.jchemed.8b00069
J. Chem. Educ. 2018, 95, 970–977

general for nonlinear LS (NLS). With some minor corrections to Deming's expressions for the variable adjustments,[7−10] the solutions that minimize S_{TV} are invariant with respect to changes in scale and changes in the manner in which the fit model is expressed: for example, $a + bx - y = 0$ or $(a - y)/b - x = 0$ for the straight-line model. These points are discussed more comprehensively elsewhere.[11]

The added complexity of adjusting multiple uncertain variables for implementation of eq 1 has led to "effective variance" (EV) treatments, in which one variable (or the response function, see below) is designated as "dependent" (here y) and the variances in the other uncertain variables are converted into effective contributions to that in y. With just two variables with independent random error, this treatment yields

$$w_{\text{eff}^{-1}} \propto \text{var}(y)_{\text{eff}} = \sigma_y^2 + (\text{d}y/\text{d}x)^2\sigma_x^2 \tag{2}$$

For $y = f(x) = a + bx$, $\sigma_{y,\text{eff}}{}^2 = \sigma_y{}^2 + b^2\sigma_x{}^2$, yielding $w_{i,\text{eff}} \propto (\sigma_y{}^2 + b^2\sigma_x{}^2)^{-1}$. The minimization target is now

$$S_{\text{EV}} = \Sigma w_{i,\text{eff}}\delta_{yi}^2 \tag{3}$$

in which x uncertainty is accommodated through the weights, but only the y residuals, $\delta_{yi} = Y_i - y_i$, need be evaluated, with x_i treated as error-free in the computations. As for the straight line, the weights $w_{i,\text{eff}}$ generally depend on the parameters. Typically, (i) S_{EV} is minimized with respect to the parameters in the response function $f(x)$, with the $w_{i,\text{eff}}$ treated as constant, followed by (ii) adjustment of the $w_{i,\text{eff}}$ using the results from step (i). This cycle is repeated until there are no further changes in the parameters, usually in 5−15 iterations. However, there is an alternative approach, in which $w_{i,\text{eff}}$ is included directly in the minimization, which I have labeled EV_2.[11] Define $F(x, y; \beta)$ as

$$F(x, y; \beta) = (w_{\text{eff}})^{1/2}\delta_y \tag{4}$$

where β represents the adjustable parameters. The minimization target becomes

$$S_{\text{EV}}^2 = \Sigma F(x_i, y_i; \beta)^2 \tag{5}$$

which now has the form of the minimization target for an unweighted fit. For the straight line, $\delta_{yi} = a + bx_i - y_i$, and $(w_{i,\text{eff}})^{1/2} = \sigma_{i,\text{eff}}{}^{-1}$. Note that since the dependence on β in the weights is included in the minimization, there is no required consistency iteration, as in the EV method. Also, neither this nor the EV method is restricted to functions that are explicit in y; they just require that F be expressed such that $F = 0$ for exactly fitting data. Then, the residuals δy are replaced by δF. In fact, this is the manner in which NLS algorithms are normally implemented.[11]

If $w_{i,\text{eff}}$ contains a dependence on the parameters, the TV and EV_2 methods require NLS for any response function; through the iteration procedure described above, the EV method is also effectively nonlinear. In ref 11 I used MC simulations to compare the performance of the three methods on a number of common fitting problems, and I concluded that the performance differences will rarely be significant compared with the difficulty of obtaining reliable information about the statistical uncertainties in the variables. The EV_2 method is arguably the easiest to use, and in fact Williamson long ago showed that, for the straight line, the EV_2 method is equivalent to the TV method.[12] While Williamson's method has been used occasionally,[13,14] there still seems to be little awareness of this TV−EV_2 equivalence, and especially the ease with which the EV_2 method can be coded for computation.[15] In the present work, I address this shortcoming, through simple modifications of existing Excel algorithms that also permit straightforward evaluation of the parameter SEs using de Levie's SolverAid routine.[16,17] By extension, this approach works for *any* weighted linear or nonlinear fitting problem. I also illustrate the KaleidaGraph[18] solutions to several of these problems. Further, the algorithms provide an easy way to appreciate the differences between the EV and EV_2 methods.

COMPUTATIONAL NOTES

For Excel illustrations, I will draw on examples in the literature and assume the reader knows enough about Excel to follow the procedures in those examples. Many readers will be less familiar with KaleidaGraph (KG), so I will provide more procedural explanation. I also refer readers to my earlier papers in this *Journal* on the use of KG.[18−20] As regards the SEs, one point requires special attention, namely, the difference between *a priori* and *a posteriori* values.[3] The former are appropriate when the data errors are considered known in an absolute sense, the latter when they are known in only a relative sense. (Minimum-variance estimates require that all errors in x and y be known apart from a single scaling constant.) KG provides the *a priori* values any time a weighted fit is executed; de Levie's SolverAid normally provides *a posteriori* SEs.

The difference between the prior (*a priori*) and post (*a posteriori*) covariance matrices is just the factor χ^2/ν (the reduced χ^2), where χ^2 is the sum of weighted squared residuals that appears first in eq 1, and $\nu = n - p$ is the number of statistical degrees of freedom for n data points and p adjustable parameters. Accordingly, the prior and post SEs differ by the square root of this quantity. For absolutely known weights, χ^2 has statistical expectation value ν, so χ^2/ν has expectation value 1. The post values are supplied for unweighted KG fits; if they are desired for a weighted fit, the required χ^2 is the quantity Chisq in the KG output. SolverAid requests the address of a cell containing the minimization target ($M), which for a weighted fit will be the same as the Chisq output from KG. It takes the square root of ($M/$\nu$), provides this quantity in the output, and includes it as a scale factor for the SEs, thus producing the post values. To obtain the prior SEs, one need only direct SolverAid to a cell containing the value of ν in place of the minimization target, making this scale factor 1.0. Note that, for unweighted fits, the sum of squares divided by ν becomes an estimate of the variance in y (constant by assumption), while for weighted fits of unknown scale it is the variance in y for data having weight unity.

KaleidaGraph and some other desktop data analysis programs are tailored for 2-dimensional problems, for which at most 3 columns of data can be accessed in conventional fitting: x, y, and the weights (a σ value for each y_i in KG). The data must first be plotted in KG before a fit can be invoked. In the EV method, the weights are estimated separately from the target minimization, so one can display y versus x and execute an EV fit in the usual way, with manual iteration to convergence, through reassessment of the weights and refitting. However, the EV_2 method normally requires access to at least 4 columns of data: x, y, and their weights or σ values. To accomplish this in the manner indicated by eqs 4 and 5, I use KG's cell(row, column) function, where both indices are absolute, starting with 0.[19,20] The "x" value becomes the row entry, enumerating the points in the data set and normally starting with 1 in the second row. A column of 0s becomes the formal "y" for plotting, and designated columns contain the actual x, y, and weighting data. A column of 1's can be selected as formal weights to obtain the prior SEs; leaving the Weight Data box unchecked in the Define Fit box yields post SEs.

DOI: 10.1021/acs.jchemed.8b00069
J. Chem. Educ. 2018, 95, 970−977

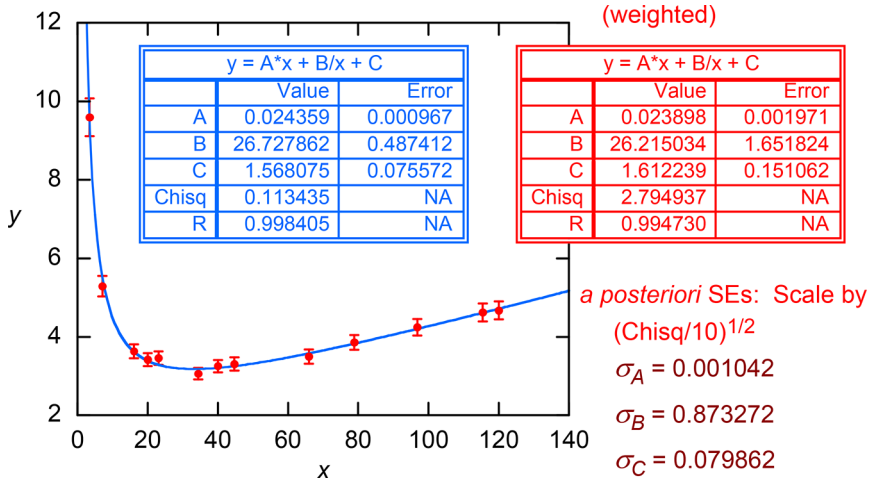


Figure 1. KaleidaGraph LS analyses of the problem treated by Harris[21] for gas chromatographic data from Moody,[22] using the van Deemter model of eq 6. The results box to the left is for unweighted LS; to the right for weighting using the ~5% uncertainties given by Harris (also used for the data error bars). Both fits use KG's General routine, which is required for parameter SEs. The *a priori* SEs ("Error") are converted to *a posteriori* as indicated below the weighted results box. The Chisq values are the sums of weighted squared residuals, with weights taken as unity in the unweighted fit, as σ_{yi}^{-2} in the weighted fit.

	A	B	C	D	E	F	G	H
1	x	y	σyi	ycalc	resid/σ	(resid/σ)^2		
2	3.4	9.59	0.48	9.40380	0.38792	0.150484		
3	7.1	5.29	0.26	5.47418	-0.70837	0.501785		
12	96.8	4.24	0.21	4.19642	0.20751	0.043061		
13	115.4	4.62	0.23	4.59728	0.09877	0.009755		
14	120.0	4.67	0.23	4.69851	-0.12395	0.015363		
15							SolverAid	
16						sumSQ=	2.7949361	0.5286716
17								
18	in D2:	=(B2-D2)/C2				A=	0.0238984	0.0010421
19	in E2:	=(B2-D2)/C2				B=	26.2150333	0.8732726
20	in F2:	=(E2)^2				C=	1.6122385	0.079862
21	(cells D2-F2 are copied to D3:D14,							
22		E3:E14, and F3:F14,			CM:	0.000001	0.0005409	-7.328E-05
23		respectively.)				0.000541	0.7626050	-0.054385
24	in $G$16:	=SUM(F2:F14)				-0.000073	-0.0543850	0.0063779

Figure 2. Excel worksheet for the weighted analysis in Figure 1. The computational instructions for columns D−F and for the minimization target, cell $G$16, are shown at the bottom. For SolverAid, the requested calculated values are the scaled residuals in E.

Since there is no plotting associated with minimizing a target with Solver in Excel, and since the cell instructions can reference any other cells, there are no such unconventional operations required for EV_2 fitting in Excel. However, manual updating of the weights and refitting to convergence are still required in the EV method.

■ ILLUSTRATIONS

Harris' van Deemter Model

In a frequently cited paper,[21] Harris showed how to use Excel's Solver to fit gas chromatography data[22] to the equation

$$y = Ax + B/x + C \quad (6)$$

both without and with weighting. This is actually a linear LS problem in both cases, even though the response function is nonlinear in x. (It is linear in the 3 adjustable parameters.) A quick check indicates that the "error" values (σ_{yi}) supplied in column C of Figure 5 in ref 21 are about 5% of the respective y values; if we instead take the errors to be 5% of the *calculated* y values, this *does* become a nonlinear LS problem.

For reference, Figure 1 shows the KG results for the two fits described by Harris. The parameter values agree completely, but the SEs differ significantly from those estimated with the jackknife (JK) procedure (on the unweighted fit) in ref 21. There appears to be an error in the final expression for the SEs from this procedure in Figure 6 of ref 21: the prefactor should be sqrt((n − 1)/n), making the corrected SEs close to the observed standard deviations (SD). For a correct model and properly weighted data, the parameter SDs for repeated realizations of the model should approximate the parametric SEs from a single fit,[2,3] and this correction factor compensates for the use of 1 fewer points in

DOI: 10.1021/acs.jchemed.8b00069
J. Chem. Educ. 2018, 95, 970−977

	A	B	C	D	E	F	G	H	I
1	x	y	σyi	ycalc	σx,eff^2	σeff	resid/σ	(resid/σ)^2	
2	3.4	9.59	0.48	9.42761	0.05252	0.29676	0.54722	0.29945	
3	7.1	5.29	0.26	5.49249	0.01120	0.15255	-1.32736	1.76188	
14	120.0	4.67	0.23	4.69047	0.00620	0.12249	-0.16708	0.02792	
15								SolverAid	
16							sumSQ=	13.94419	1.180855
17							ν=	10.00000	1.000000
18									
19	in E2:	(0.03*A2*($H$20-$H$21/A2^2))^2							
20	in F2:	(E2 + (0.02*D2)^2)^0.5					A=	0.02370	0.000977
21		(both copied to subsequent cells)					B=	26.24614	0.956405
22							C=	1.62757	0.073729
23									
24						CM:	9.55E-07	0.00057	-0.000064
25							0.000569	0.91471	-0.057313
26							-6.4E-05	-0.05731	0.005436

Figure 3. Excel worksheet modified for EV_2 treatment of 2% error in *y* and 3% in *x*. The minimization target is now cell $H$16, and the SolverAid target is set to $H$17, in order to obtain the prior parameter SEs.

	0 in	1 x	2 y
0			
1	1	3.4	9.59
2	2	7.1	5.29
3	3	16.1	3.63
4	4	20.0	3.42
5	5	23.1	3.46
6	6	34.4	3.06
7	7	40.0	3.25
8	8	44.7	3.31
9	9	65.9	3.50
10	10	78.9	3.86
11	11	96.8	4.24
12	12	115.4	4.62
13	13	120.0	4.67
14			

```
xx = cell(x,1);
yy = cell(x,2);
sigx = (0.03*xx);
calc = (a*xx + b/xx + c);
sigy = (0.02*calc);
sigt = sqrt(sigy^2 + (a-b/xx^2)^2*sigx^2);
fitf = ((calc - yy)/sigt);
```

y = fitf(x)	Value	Error
a	0.023702	0.00097792
b	26.246	0.95703
c	1.6276	0.073777
Chisq	13.944	NA
R	1.0000	NA

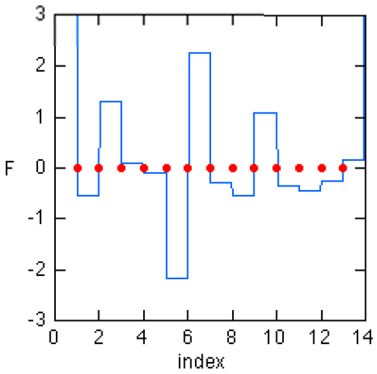


Figure 4. KaleidaGraph version of weighted analysis shown in Figure 3. The computational instructions are entered in the Macro Library. Then, a General curve fit is selected, and fitf and initial parameter values are entered in the Define··· box. In the latter, the weight data box is checked; then, when the column of 0 values is selected for *y*, a column containing 1's is selected in response to the prompt for weights.

each such realization of the model in the JK procedure. The appropriateness of the JK method is also questionable, since the estimation precision can depend strongly on the data structure. Thus, deleting the first point (accomplished by masking out that line in the data sheet in KG) almost doubles the SE for *B*. In any event, since this is a linear fit, the parametric SEs are completely reliable, provided the data error structure is known, and exact if the errors are known absolutely. Thus, the reliability of the SEs hinges entirely on knowledge of the data error structure: constant (unweighted analysis), proportional to signal, or something else.

The Chisq value for the weighted fit in Figure 1 occurs less than 2% of the time for a correct model with correct weighting for $\nu = 10$.[23] This suggests that the assumed 5% error in *y* is pessimistic by a factor of ~2. In that case, the post SEs are more reasonable. These are obtained as described earlier and provided below the fit results box. They are now reasonably close to the unweighted SE estimates for *A* and *C*. For *B*, we need more information about the data error to decide which SE is better. Most instrumental measurements are dominated by constant error in the low-signal limit but display proportional error at high signal.[24] Assuming that this holds for these chromatographic data, we still need to know where in the measurement range these data fall. There is also the issue of uncertainty in *x*, which here is the flow rate of helium. I address that below.

DOI: 10.1021/acs.jchemed.8b00069
J. Chem. Educ. 2018, 95, 970−977

	A	B	C	D	E	F	G
3			a=	5.47991022	0.35924652		
4			b=	-0.48053341	0.07062027		
5							
6		1.028325E+00	CM:	0.12905806	-0.02443363		
7		SolverAid		-0.02443363	0.00498722		
8		11.86635319	1.2179056				
9		SolverAid					
10	ν =	8		b=	-4.805334E-01		
11							
12							
13	x	y	wx	wy			
14							
15	0	5.9	1000	1	1.000230912	-4.20090E-01	-4.200413E-01
16	0.9	5.4	1000	1.8	0.555786468	-3.52570E-01	-4.729238E-01
17	1.8	4.4	500	4	0.250461825	2.14950E-01	4.295037E-01
18	2.6	4.6	800	8	0.12528864	-3.69477E-01	-1.043833E+00
19	3.3	3.5	200	20	0.051154562	3.94150E-01	1.742687E+00
20	4.4	3.7	80	20	0.052886404	-3.34437E-01	-1.454260E+00
21	5.2	2.8	60	70	0.018134254	1.81137E-01	1.345105E+00
22	6.1	2.8	20	70	0.025831332	-2.51344E-01	-1.563847E+00
23	6.5	2.4	1.8	100	0.138284642	-4.35569E-02	-1.171306E-01
24	7.4	1.5	1	500	0.232912356	4.23963E-01	8.784796E-01
25							
26	in E15: =(1/D15 + $E$10^2/C15)				in B8: = SUMSQ(G15:G24)		
27	in F15: =$D$3 + $D$4*A15 - B15				in E10: = $D$4		
28	in G15: =($D$3 + $D$4*A15 - B15)/sqrt(E15)						

Figure 5. Excel worksheet for EV_2 (=TV) analysis of York model. The instructions in E15:G15 are copied to subsequent cells. Cell $B$8 is the Solver target and is also used here by SolverAid to yield the post SEs. *b* in E10 is used in the weight computations in column E; here, it is also set equal to $D$4, which with $D$3 is varied by Solver to minimize $B$8. In the call to SolverAid, the calculated values are those in G15:G24.

The Excel worksheet for the weighted analysis is shown in Figure 2. Note the agreement with both parameters and post SEs from the weighted fit in Figure 1. Now suppose we would like to include uncertainty in *x*. Recognizing the too-small χ^2 value for the weighted fit, let us reduce the error in *y* to 2% and take that in *x* to be 3%. From eq 2, the effective contribution to the variance in *y* from that in *x* is

$$\mathrm{var}(y)_{\mathrm{eff},x} = (\mathrm{d}y/\mathrm{d}x)^2\sigma_{x^2} = (\mathrm{A} - \mathrm{B}/x^2)^2(0.03x)^2 \qquad (7)$$

and is computed in column E in the worksheet in Figure 3. σ_{eff} is then computed in column F, with the rest of the worksheet as in Figure 2. The minimization target for Solver is again the sum of weighted squared residuals, but for SolverAid we now use H17, which contains the number of statistical degrees of freedom. In this way, we obtain the *a priori* SEs. It is worth noting that now the χ^2 value (13.944) is reasonable for $\nu = 10$; but this cannot be taken as verification of our assumption of proportional error in both *x* and *y*.

Figure 4 shows the data sheet, instructions, and results for the KG counterpart of Figure 3, in which I have also chosen weighted analysis with weights = 1 to obtain the prior SEs. Recall that to generate a fit in KG, we must first plot the data, here just 0 for each value of the index, 1−13. While this display is not very informative, the fit results are more so, being the scaled residuals, $\delta_{yi}/\sigma_{i,\mathrm{eff}}$ (=fitf).

It is worth emphasizing the importance of parentheses in Library definitions in KG. From the manner in which the program parses out complex instructions by simply substituting these expressions, omitted parentheses can lead to incorrect results. For example, "bb = −3" in the Library gives −9 for bb^2 ((bb)^2 does give 9).

Orear treated a model very similar to the present van Deemter model, with a sign change for the second term and missing the constant term.[25] He found agreement to four significant figures in the results from the TV and EV_2 methods (see his erratum), but I later confirmed that there were real though very small differences.[11] The same comparison for the proportional error model of Figures 3 and 4, obtained from FORTRAN programs, shows greater but still practically insignificant differences: slightly higher χ^2 for TV (by 0.12%) and differences <0.2% in the parameters and <0.5% in the SEs.

The York Straight Line

In an example that has become a touchstone in *x*−*y* error tests, York[26] took a 10-point data set from Pearson[27] and added weights strongly favoring *x* at small *x* and *y* at large *x*, spanning ranges of 1000 for *x* and 500 for *y*. This was, in fact, the example treated by Williamson.[12] The results from many efforts have been summarized by Riu and Rius[28] and Reed.[29] While most of these works employed tedious algebraic expressions, I have noted that the purely numerical approach is good to at least 10 digits for the parameters, 7 for the SEs, and 12 for S_{TV}.[11] Thus, there is little

DOI: 10.1021/acs.jchemed.8b00069
J. Chem. Educ. 2018, 95, 970−977

	A	B	C	D	E	F	G
1							
2						iterations	
3			a=	5.39605209	0.36145838	11.77858	-0.463260
4			b=	-0.46344888	0.07066273	11.95847	-0.463452
5						11.95642	-0.463449
6		1.18274691	CM:	0.13065216	-0.02462112	11.95645	-0.463449
7		SolverAid		-0.02462112	0.00499322		
8		11.95644785	1.2225203				
9		SolverAid					
10				b=	-0.463449		

Figure 6. EV analysis of the York model. Now b in E10 is decoupled from the optimization variable in $D$4, so it must be entered manually for each cycle. The iterations (to right) show changes in the sum of weighted squares and b for 4 cycles.

	A	B	C	D	E	F	G
5				SolverAid			
6			SSQ=	2.41653494	0.7772604		
7							
8	P0=	363.9476	0.7732318	CM:	0.59788749	-0.32341672	0.00553264
9	k(x10^6)=	7.444115	0.849368		-0.3234167	0.72142562	-0.0164935
10	n=	1.976401	0.019633		0.00553264	-0.0164935	0.00038545
11							
12							
13	t	P	Pterm	seft	sefP	sigt	resid/sigt
14							
15	0	363	364.8951	-7.268E-06	-8.429E-06	1.113E-05	-0.71943463
16	42	397	330.8951	-7.268E-06	-1.023E-05	1.2546E-05	0.18874032
17	105	437	290.8951	-7.268E-06	-1.319E-05	1.5062E-05	0.59924553
18	242	497	230.8951	-7.268E-06	-2.082E-05	2.2056E-05	0.34485921
19	480	557	170.8951	-7.268E-06	-3.775E-05	3.8439E-05	-1.05138898
20	840	607	120.8951	-7.268E-06	-7.481E-05	7.5161E-05	-0.01006863
21	1440	647	80.8951	-7.268E-06	-1.655E-04	0.00016566	0.52894481
22							
23		NOTE: Increased k by factor 10^6 and then incorporated factors of 10^-6					
24			in columns D & G.				
25							
26	in C15: = (2*$B$8-B15)				in F15: =SQRT(D15^2+E15^2)		
27	in D15: = (1-$B$10)*$B$9*1.E-6				in G15: = (C15^(1-$B$10) - $B$8^(1-$B$10)		
28	in E15: = (1-$B$10)/C15^$B$10				+ (1-$B$10)*$B$9*A15*1E-6)/F15		

Figure 7. Excel worksheet for EV_2 analysis of the Wentworth kinetics model. The instructions in C15:G15 are copied to subsequent cells. Cell $D$6 is the Solver minimization target, with the instruction = SUMSQ(G15:G21). D6 is used also by SolverAid to yield the post SEs; prior SEs are smaller by the factor in E6. In the call to SolverAid, the calculated values are in G15:G21.

reason to use algebraic expressions, especially considering how easy the EV_2 method can be implemented numerically and its equivalence to TV for the straight line.

Figure 5 shows the Excel worksheet for this analysis. York provided weights rather than data SDs or variances, so the variances are taken as the reciprocals of York's weights in calculating $\sigma_{y,\text{eff}}^2$ in column E. The residuals are computed in column F and used to calculate $\delta_y/\sigma_{y,\text{eff}}$ in column G. Note that the calculations of $\sigma_{y,\text{eff}}^2$ in E reference the value of b in $E$10, while the residuals computations use $D$4 for b. By this procedure, we can do both the EV_2 and EV analyses with a small change. First, the instruction = $D$4 is put in $E$10, linking it to the value that is varied by Solver in the minimization. This gives the EV_2 (=TV) fit in Figure 5. By removing this instruction and simply entering a value for b, we can obtain the EV results, as shown in Figure 6, where convergence has been achieved in 4 iterations. Note that in the first iteration the sum of weighted squares in B8 actually drops, but it increases as the value of b is updated, finally achieving a value about 0.1 higher than that for the EV_2 analysis.

In Exercise 4.19.1 in his book,[30] de Levie used the same Pearson data with much more strongly varying weights, from

DOI: 10.1021/acs.jchemed.8b00069
J. Chem. Educ. 2018, 95, 970−977

10^{-5} for w_y and 3×10^{-5} for w_x for the first point, both increasing by a factor of 10 for each successive point. The expression for the weighted squared residuals given for his cell E20, though much more complex than used here in column G, can be shown to be equivalent (after the correction included in the citation of ref 30 below). Accordingly, the parameter estimates obtained using the present Solver algorithm with those weights agree with the values given in Figure 4.19.3 of ref 30. However, the SEs given there are not correct: they should be 0.24882 for a and 0.003403 for b.

Additional instructive computations can be done with this example. First, change all the weights for x to a very large number (say 1000) and those for y to 1. This approximates the conditions for ordinary unweighted LS, and the EV and EV_2 results both agree with those from an unweighted fit to $y = a + bx$. Next, reverse the weights for x and y. Now the EV_2 results agree with simple regression of x on y using $x = y/b - a/b$, as expected for this weighting. But the EV results remain the same as for the previous weighting, a basic flaw in the EV method that has been lodged against it. The EV_2 fit for the York weights can also be done using $x - y/b + a/b$ for the residuals and modifying σ_{eff} accordingly. All results remain unchanged. However, if the residuals are taken as $1/y - 1/(a + bx)$ and σ_{eff} is again changed to be consistent with this model, the EV_2 results are not the same, confirming that the TV-EV_2 equivalence holds only for the straight line. If this "inverse" model is fitted by the TV method, the results *do* remain unchanged.

Implicit Fit Models: Wentworth Kinetics Example

With the exception of the just-mentioned "inverse" model, the previous examples have all been formulated to be explicit in one variable: y in the Harris problem, both y and x in the York model. It is not always possible or convenient to make the most uncertain variable explicit in the fit model, but that is not a problem, as is illustrated using Wentworth's gas kinetics model,[31] in which pressure (P) and time (t) are both considered to have uncertainty 1 (in torr and s). The model can be expressed as explicit in either t or P, and has been treated as such in ref 11, along with the implicit model I consider here

$$F \equiv (2P_0 - P)^{1-n} - P_0^{1-n} + (1 - n)kt = 0 \tag{8}$$

The adjustable parameters are P_0, n, and k. To compute the effective variance, we need $\partial F/\partial P = (n - 1)/(2P_0 - P)^n$ and $\partial F/\partial t = (1 - n)k$. Figure 7 shows the Excel spreadsheet for the solution, which can be seen to be identical to the values in the fourth line of Table 6 in ref 11 (except the SEs, which are post here, hence smaller by the factor in E6). However, achieving these results did require dealing with an idiosyncrasy in Solver that I also encountered when using it to solve systems of equations in ref 20, namely, its difficulty in handling adjustable parameters that differ greatly in magnitude, particularly when one or more are very small. Thus, with k in its native s^{-1} units, Solver seemed to satisfy its convergence criteria long before it had actually converged on the solution. Proper convergence was achieved by scaling k up by a factor of 100 or more and adjusting the other quantities accordingly.

There is a different problem with KG on this example, but one that is likely to be important only in numerical tests. The program evidently uses too-large increments in estimating numerical derivatives, perhaps changes of about 0.1%. When high precision is desired, the adjustable parameters can be redefined in the Library to be much smaller quantities, for example, in this case, as P0 = $(363 + p)$, $k = (7.4 \times 10^{-6} + a)$, and $n = (1.97 + b)$, where p, a, and b are the adjustable parameters. This leads to higher precision in the numerical derivatives, giving agreement with the ref 11 results to the sixth or seventh digit, as compared with about 2 fewer digits without this redefinition.

Implicit fit models can be useful in situations where the nominal dependent variable seems to require solving a quadratic or higher-order equation, for example, in binding and equilibrium studies. The LS algorithm can solve these equations in the course of obtaining the minimum-variance estimates of the parameters, just as when it is used to solve systems of equations for exactly fitting data.[20] In some such studies, casting the equations in the commonly used linear forms makes the "dependent" variable a function of the measured and hence uncertain "independent" variable, making the two correlated. Both correlation and weighting are usually ignored in such work.[32] For example, in ligand binding studies where the concentration of free ligand [L] ($\equiv x$) is measured as a function of initial concentration x_0 for total substrate concentration S_t, the binding equation takes the form

$$x^2 + x(S_t - x_0 + 1/K) - x_0/K = 0 \tag{9}$$

where K is the binding constant. The effective variance contributions from uncertainty in x and x_0 are easy to evaluate, making eq 9 preferred as fit model over the commonly used rectangular hyperbolic and straight-line relationships.[32] (An example using eq 9 as fit model is included in the Supporting Information, together with Excel and KG files for the other examples treated here.)

CONCLUSION

The EV_2 method for least-squares fitting of data having error in more than one variable is easy to code for computations in Excel and KaleidaGraph, and by extension other LS desktop programs and program environments. It produces results that will rarely be statistically inferior to those produced by what is assumed to be "best" for such problems, the TV method. For the straight-line model, the EV_2 method is equivalent to the TV method. Its capabilities have been illustrated in the present work for both explicit and implicit fit models. The weighting procedures used here are also suitable for more traditional problems with varying uncertainty in y alone.

Correct use of the TV and EV methods does require that the user address the matter of uncertainties in both x and y, since minimum-variance estimation of the parameters requires that the weights in x and y be defined to within a single scaling constant and be proportional to the respective reciprocal variances. Those who are accustomed to analyzing data by blithely clicking a "Fit" button without considering their data error may find this an inconvenience. However, such a consideration should always be a part of data analysis, because neglected data heteroscedasticity leads to erroneous parametric SEs and nonoptimal parameter estimates. This is especially a problem when data are transformed to a more convenient relationship, usually the straight line, as such transformations invariably change the relative weights of the data.[2,32]

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available on the ACS Publications website at DOI: 10.1021/acs.jchemed.8b00069.

Overview of the Supporting Information files with guidance for use (PDF, DOCX)

KaleidaGraph plot files for the examples illustrated (ZIP)

DOI: 10.1021/acs.jchemed.8b00069
J. Chem. Educ. 2018, 95, 970−977

Excel worksheets for the examples illustrated (ZIP)

AUTHOR INFORMATION

Corresponding Author

*E-mail: joel.tellinghuisen@vanderbilt.edu.

ORCID

Joel Tellinghuisen: 0000-0002-4487-5197

Notes

The author declares no competing financial interest.

REFERENCES

(1) Stigler, S. M. Gauss and the Invention of Least Squares. *Ann. Statist.* **1981**, *9*, 465−474.

(2) Tellinghuisen, J. Can You Trust the Parametric Standard Errors in Nonlinear Least Squares? Yes, with Provisos. *Biochim. Biophys. Acta, Gen. Subj.* **2018**, *1862* (4), 886−894.

(3) Tellinghuisen, J. Understanding Least Squares through Monte Carlo Calculations. *J. Chem. Educ.* **2005**, *82*, 157−166.

(4) Tellinghuisen, J. Least Squares in Calibration: Dealing with Uncertainty in *x*. *Analyst* **2010**, *135*, 1961−1969.

(5) Deming, W. E. *Statistical Adjustment of Data*; Wiley: New York, 1938.

(6) Carroll, R. J.; Ruppert, D. *Transformation and Weighting in Regression*; Chapman and Hall: New York, 1988.

(7) Powell, D. R.; Macdonald, J. R. A Rapidly Convergent Iterative Method for the Solution of the Generalized Nonlinear Least Squares Problem. *Computer J.* **1972**, *15*, 148−155.

(8) Britt, H. I.; Luecke, R. H. The Estimation of Parameters in Nonlinear, Implicit Models. *Technometrics* **1973**, *15*, 233−247.

(9) Jefferys, W. H. On the Method of Least Squares. *Astron. J.* **1980**, *85*, 177−181.

(10) Lybanon, M. A Better Least-Squares Method When Both Variables Have Uncertainties. *Am. J. Phys.* **1984**, *52*, 22−26.

(11) Tellinghuisen, J. Least-Squares Analysis of Data with Uncertainty in *x* and *y*: A Monte Carlo Methods Comparison. *Chemom. Intell. Lab. Syst.* **2010**, *103*, 160−169.

(12) Williamson, J. H. Least-Squares Fitting of a Straight Line. *Can. J. Phys.* **1968**, *46*, 1845−1847. This treatment is essentially Exercise 3 in Chapter VIII in Deming's book (see ref 5); Deming cites an 1879 paper by C. H. Kummell for it.

(13) Christian, S. D.; Tucker, E. E. LINGEN—A General Linear Least Squares Program. *J. Chem. Educ.* **1984**, *61*, 788.

(14) Ogren, P. J.; Norton, J. R. Applying a Simple Linear Least-Squares Algorithm to Data with Uncertainties in Both Variables. *J. Chem. Educ.* **1992**, *69*, A130−A131. These authors used Williamson's approach in its algebraic form rather than in a numerical minimization approach targeting eq 5, which for the straight-line model becomes Williamson's 4th equation.

(15) Irvin, J. A.; Quickenden, T. I. Linear Least Squares Treatment When There Are Errors in Both *x* and. *J. Chem. Educ.* **1983**, *60*, 711−712. Referring to the need for a "complicated numerical approach" to minimize Williamson's 4th equation, these authors use the EV method.

(16) de Levie, R. Estimating Parameter Precision in Nonlinear Least Squares with Excel's Solver. *J. Chem. Educ.* **1999**, *76*, 1594−1598.

(17) Excellaneous: An ad-free, spyware-free web site for Excel users in the physical sciences. http://www.bowdoin.edu/~rdelevie/excellaneous (accessed Apr 2018).

(18) Tellinghuisen, J. Nonlinear Least-Squares Using Microcomputer Data Analysis Programs: KaleidaGraph in the Physical Chemistry Teaching Laboratory. *J. Chem. Educ.* **2000**, *77*, 1233−1239.

(19) Tellinghuisen, J. Using Least Squares for Error Propagation. *J. Chem. Educ.* **2015**, *92*, 864−870.

(20) Tellinghuisen, J. Using Least Squares To Solve Systems of Equations. *J. Chem. Educ.* **2016**, *93*, 1061−1067.

(21) Harris, D. C. Nonlinear Least-Squares Curve Fitting with Microsoft Excel Solver. *J. Chem. Educ.* **1998**, *75*, 119−121.

(22) Moody, H. W. The Evaluation of the Parameters in the van Deemter Equation. *J. Chem. Educ.* **1982**, *59*, 290−291.

(23) Bevington, P. R.; Robinson, D. K. *Data Reduction and Error Analysis for the Physical Sciences*, 2nd ed.; McGraw-Hill: New York, 1992. See Table C.4 for the χ^2 distribution.

(24) Zeng, Q. C.; Zhang, E.; Dong, H.; Tellinghuisen, J. Weighted Least Squares in Calibration: Estimating Data Variance Functions in High-Peformance Liquid Chromatography. *J. Chromatogr. A* **2008**, *1206*, 147−152.

(25) Orear, J. Least Squares When Both Variables Have Uncertainties. *Am. J. Phys.* **1982**, *50*, 912−916; *Am. J. Phys.* **1984**, *52*, 278.

(26) York, D. Least-Squares Fitting of a Straight Line. *Can. J. Phys.* **1966**, *44*, 1079−1086.

(27) Pearson, K. On Lines and Planes of Closest Fit to Systems of Points in Space. *Philos. Mag.* **1901**, *2*, 559−572.

(28) Riu, J.; Rius, F. X. Univariate Regression Models with Errors in Both Axes. *J. Chemom.* **1995**, *9*, 343−362.

(29) Reed, B. C. Linear Least-Squares Fits with Errors in Both Coordinates. II: Comments on Parameter Variances. *Am. J. Phys.* **1992**, *60*, 59−62. See especially Note 15.

(30) de Levie, R. *Advanced Excel for Scientific Data Analysis*, 3rd ed.; Atlantic Academic, LLC: Orrs Island, ME, 2012; available only from Amazon.com. In Figure 4.19.3, the instruction for cell E20 should be changed at the end from "))^2))" to ")^2)))".

(31) Wentworth, W. E. Rigorous Least Squares Adjustment: Application to Some Non-Linear Equations, II. *J. Chem. Educ.* **1965**, *42*, 162−167.

(32) Tellinghuisen, J.; Bolster, C. H. Weighting Formulas for the Least-Squares Analysis of Binding Phenomenon Data. *J. Phys. Chem. B* **2009**, *113*, 6151−6157.

DOI: 10.1021/acs.jchemed.8b00069
J. Chem. Educ. 2018, 95, 970−977