Atmos. Chem. Phys., 8, 5477–5487, 2008
www.atmos-chem-phys.net/8/5477/2008/

Atmospheric Chemistry and Physics

Technical Note: Review of methods for linear least-squares fitting of data and application to atmospheric chemistry problems

C. A. Cantrell

National Center for Atmospheric Research Atmospheric Chemistry Division 1850 Table Mesa Drive Boulder,
CO 80305, USA

Received: 13 February 2008 – Published in Atmos. Chem. Phys. Discuss.: 1 April 2008
Revised: 11 August 2008 – Accepted: 15 August 2008 – Published: 12 September 2008

Abstract. The representation of data, whether geophysical observations, numerical model output or laboratory results, by a best fit straight line is a routine practice in the geosciences and other fields. While the literature is full of detailed analyses of procedures for fitting straight lines to values with uncertainties, a surprising number of scientists blindly use the standard least-squares method, such as found on calculators and in spreadsheet programs, that assumes no uncertainties in the x values. Here, the available procedures for estimating the best fit straight line to data, including those applicable to situations for uncertainties present in both the x and y variables, are reviewed. Representative methods that are presented in the literature for bivariate weighted fits are compared using several sample data sets, and guidance is presented as to when the somewhat more involved iterative methods are required, or when the standard least-squares procedure would be expected to be satisfactory. A spreadsheet-based template is made available that employs one method for bivariate fitting.

1 Introduction

Representation of the relationship between x (independent) and y (dependent) variables by a straight line (or other function) is a routine process in scientific and other disciplines. Often the parameters (slope and y-intercept) of such a fitted line can be related to fundamental physical quantities. It is therefore very important that the parameters accurately represent the data collected, and that uncertainties in the parameters are estimated and applied correctly or the results of the fitting process and thus the scientific study could be misinterpreted.

The approaches to fitting straight lines to collections of $x-y$ data pairs can be broadly grouped into two categories: the "standard" least-squares methods in which the distances between the fitted line and the data in the y-direction are minimized, and the "bivariate" least-squares methods in which the perpendicular distances between the fitted line and the data are minimized. A third method, similar to the second but less commonly employed, involves minimization of the areas of the right triangles formed by the data point and the line. In all of these methods, weights may be also applied to the data to account for the differing uncertainties in the individual points. In "standard" least-squares, the weighting pertains to the y-variables only, whereas in "bivariate" methods, weights can be assigned for the x- and y-variables independently. There is widely varying terminology for these procedures in the literature that can be confusing to the non-expert. Authors have used terms such as major axis regression, reduced major axis regression, ordinary least-squares, maximum likelihood, errors in variables, rigorous least-squares, orthogonal regression and total least-squares. Herein, the terms "standard" and "bivariate" will be used to denote these two categories of fitting methods. This paper does, however, present a detailed reference list of available methods and applications presented in the literature.

For demonstration and testing purposes, two data sets from the literature were employed. First, the well-known data of Pearson (1901) with weights suggested by York (1966) were used (see Table 1 and Fig. 1). The data values are similar to those that might be encountered in a laboratory study or acquired in atmospheric measurements, but with rather extreme weights that range 3 orders of magnitude as the data ranges about a factor of five. This data set has the advantage that the exact results of the bivariate fit are known and reported in the literature, and one that is frequently used as a test for new fitting methods.

Correspondence to: C. A. Cantrell
(cantrell@ucar.edu)

Published by Copernicus Publications on behalf of the European Geosciences Union.

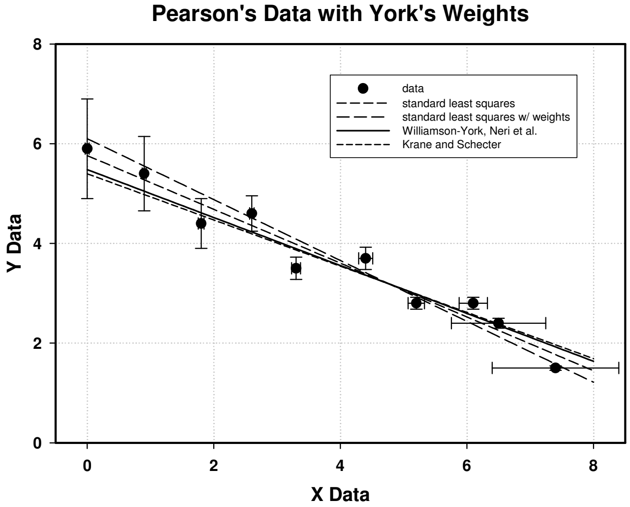

Fig. 1. Linear fits to the data of Pearson [1901] with weights suggested by York [1966] ("Pearson-York" data set, shown in Table 1). The weights have been plotted as σ values (w_i=1/σ_i^2). Fit parameters are shown in Table 2.

Table 1. Example data "Pearson's data with York's weights" for comparison of fitting procedures described in the text.

	x	w_x	y	w_y
1	0.0	1000.0	5.9	1.0
2	0.9	1000.0	5.4	1.8
3	1.8	500.0	4.4	4.0
4	2.6	800.0	4.6	8.0
5	3.3	200.0	3.5	20.0
6	4.4	80.0	3.7	20.0
7	5.2	60.0	2.8	70.0
8	6.1	20.0	2.8	70.0
9	6.5	1.8	2.4	100.0
10	7.4	1.0	1.5	500.0

A second data set was created by selecting random numbers from Gaussian distributions and adding them to base values, which were numbers 1 through 100 (see Fig. 2). Initially, the Gaussian distributions were set with means of zero, and standard deviations of 10 units plus 30% of the base value, but other tests were performed with different amounts of constant and proportional uncertainty. These data were meant to represent those that would result from an intercomparison of two instruments measuring the same quantity, which have baseline noise of 10 units (1 sigma), measurement uncertainties that are well above the baseline of 30% (1 sigma), and nominal "true" values from 1 to 100 units. This data set has the characteristic that in the absence of noise, or if the noise is properly dealt with, the best fit line should have a slope of one, and an intercept of zero.

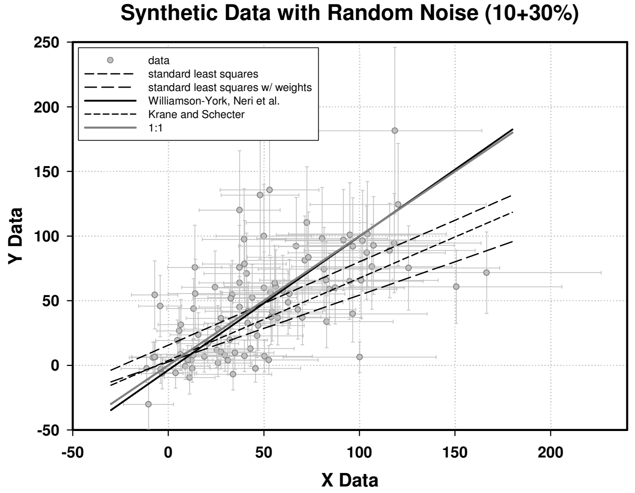

Fig. 2. Linear fits to data generated by sampling a Gaussian function with standard deviation of 10 units plus 30%, and adding the noise to the numbers 1 through 100. Fit parameters are shown in Table 3.

Next, the methods were applied to two examples of authentic data to demonstrate specifically the value of bivariate methods, and to point out how and when they should be applied.

This review and recommendation does not attempt to be mathematically nor statistically rigorous. The reader is referred to the referenced literature for such details. The purpose here is to provide operational information for the scientific user of these routines, and to provide guidance for the choice of routine to be utilized.

Note that there is not universal agreement in the uses of symbols for the measured x and y values and the calculated slope and intercept that appear in the literature. The reader is cautioned in this regard. In this paper, x_i and y_i (lower case italics) refer to the measured x and y values, m refers to the slope of the best fit line, and b is the y-axis intercept. Other symbols are defined throughout the paper.

2 Standard least-squares

The equations for a line that best describes $x-y$ data pairs when all of the measurement error may be assumed to reside in the y-variable (i.e. the x values are exact or nearly so) is readily available and easily derived (e.g. Bevington, 1969). The fitted line then becomes a "predicted" value for y given a value for x. The usual method involves minimizing the sum of squares of the differences between the fitted line and the data points in the y-direction (although minimization of other quantities has been used). The slope, m, and y-intercept, b, of this best-fit line can be represented in terms of summations of computations performed on the n measured

Table 2. Comparison of fit parameters using various weighting and fitting procedures for Pearson's data with York's weights (reproduced in Table 1).

Reference	order	Slope	Std err Slope	% diff	Intercept	Std err Intcpt	% diff
Std Least-Squares	$y-x$	–0.53958	0.0421	12.3	5.7612	0.189	5.1
	$x-y$	–0.56589	0.0442	17.8	5.8617	0.216	7.0
Std Lst-Sqrs w/wgts	$y-x$	–0.61081	–	27.1	6.1001	–	11.3
	$x-y$	–0.66171	–	37.7	6.4411	–	17.5
Williamson-York	$y-x$	–0.48053	0.0706	0	5.4799	0.359	0
	$x-y$	–0.48053	0.0706	0	5.4799	0.359	0
Neri et al.	$y-x$	–0.48053	–	0	5.4799	–	0
	$x-y$	–0.48053	–	0	5.4799	–	0
Reed	$y-x$	–0.48053	0.0706	1×10^{-7}	5.4799	0.359	6×10^{-8}
	$x-y$	–0.48053	–	3×10^{-7}	5.4799	–	1×10^{-7}
Macdonald	$y-x$	–0.48053	–	5×10^{-6}	5.4799	–	2×10^{-6}
	$x-y$	–0.48053	–	5×10^{-5}	5.4799	–	8×10^{-6}
Lybanon	$y-x$	–0.48053	–	2×10^{-6}	5.4799	–	5×10^{-7}
	$x-y$	–	–	–	–	–	–
Krane and Schecter	$y-x$	–0.46345	–	3.6	5.3960	–	1.5
	$x-y$	–0.55049	–	14.6	5.8163	–	6.1

data pairs, $x_1, y_1, x_2, y_2, \ldots, x_n, y_n$.

$$m = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

$$b = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} \qquad (1)$$

The Σ symbols refer to the summation of the quantity over all n values, and the subscript, i, denotes the individual measured x and y values. The uncertainties in the slope and intercept can also be calculated.

$$\sigma_m = \frac{\sqrt{\frac{\sum y_i^2 - b\sum y_i - m\sum x_i y_i}{n-2}}}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2}} \sigma_b = \sigma_m \sqrt{\frac{\sum x_i^2}{n}} \qquad (2)$$

Another useful quantity is the correlation coefficient (also called the Pearson Correlation Coefficient), which provides an index of the degree of correlation between the x and y data.

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{\left(n\sum x_i^2 - \left(\sum x_i\right)^2\right)\left(n\sum y_i^2 - \left(\sum y_i\right)^2\right)}} \qquad (3)$$

It is usually the case that not all the data points have the same uncertainty. Thus, it is desired that data with least uncertainty have the greatest influence on the slope and intercept of the fitted line. This is accomplished by weighting each of the points with a factor, w_i, which is often assumed (and demonstrated mathematically to yield the best unbiased linear fit parameters, if set) equal to the inverse of the variance of the y-values (σ_{yi}^2). It could include estimates of all sources of uncertainty in the y-values. Other weighting procedures are also possible. The formulas for the slope and intercept are modified as shown to include data weights.

$$m = \frac{\sum w_i \sum w_i x_i y_i - \sum w_i x_i \sum w_i y_i}{\sum w_i \sum w_i x_i^2 - \left(\sum w_i x_i\right)^2}$$

$$b = \frac{\sum w_i x_i^2 \sum w_i y_i - \sum w_i x_i \sum w_i x_i y_i}{\sum w_i \sum w_i x_i^2 - \left(\sum w_i x_i\right)^2} \qquad (4)$$

These formulas are readily programmed, or exist as available spreadsheet or calculator functions, and can be routinely applied to fitting of straight lines to $x-y$ data sets.

The standard least-squares method was applied with and without weights, using Eqs. (1) and (4), to the two test data sets ("Pearson-York" and "synthetic data") for comparison with the bivariate methods (see Tables 2 and 3). Note that when there are significant x and y errors, that standard least-squares yields erroneous slopes. For the "synthetic data", the slope was usually too small, whereas for the "Pearson-York" data, the slope was too large (compared to the Williamson-York and Neri et al. methods, discussed below).

3 Methods when both x and y have errors

The application of fitting procedures that account for uncertainties in both the x- and y- variables is somewhat more complex. This is because minimization of the distance between data points and a fitted line in the x- and y-directions has not yielded to analytical solutions. Iterative approaches are therefore required. Several equation forms have been

Table 3. Comparison of fit parameters using various weighting and fitting procedures for synthetic data with random errors (see text).

Reference	order	Slope	Std err Slope	% diff	Intercept	Std err Intcpt	% diff
Std Least-Squares	$y-x$	0.64455	0.0802	37.7	15.5840	5.068	526
	$x-y$	1.62395	0.2022	57.0	–33.2653	10.818	810
Std Lst-Sqrs w/wgts	$y-x$	0.51688	–	50.0	3.12330	–	185
	$x-y$	1.45084	–	40.3	–25.7369	–	604
Williamson-York	$y-x$	1.03409	0.1004	0	–3.65745	3.369	0
	$x-y$	1.03409	0.1004	0	–3.65745	3.369	0
Neri et al.	$y-x$	1.03409	–	0	–3.65745	–	0
	$x-y$	1.03409	–	0	–3.65745	–	0
Reed	$y-x$	1.03409	–	0	–3.65745	–	0
	$x-y$	1.03409	–	0	–3.65745	–	1×10^{-9}
Krane and Schecter	$y-x$	0.63716	–	38.4	3.73640	–	202
	$x-y$	1.69288	–	63.7	–16.2079	–	343

proposed and discussed (Barker and Diana, 1974; Borcherds and Sheth, 1995; Brauers and Finlayson-Pitts, 1997; Bruzzone and Moreno, 1998; Chong, 1991, 1994; Christian and Tucker, 1984; Christian et al., 1986; Gonzalez et al., 1992; Irwin and Quickenden, 1983; Jones, 1979; Kalantar, 1990, 1991; Krane and Schecter, 1982; Leduc, 1987; Lybanon, 1984ab, 1985; Macdonald and Thompson, 1992; MacTaggart and Farwell, 1992; Markovsky and Van Huffel, 2007; Moreno, 1996; Neri et al., 1990, 1991; Orear, 1982; Pasachoff, 1980; Pearson, 1901; Press et al., 1992a,b; Reed, 1990; Riu and Rius, 1995; Squire et al., 1990; Williamson, 1968; York, 1966, 1969; York et al., 2004). This list is large to provide a comprehensive reference for the reader. While these approaches are not as convenient as the straightforward equations applicable to standard least-squares, they can easily be programmed using standard languages or spreadsheet program routines.

In assessing the impacts of errors on linear fits, normal (Gaussian) distributions of the errors are assumed. This is a reasonable assumption for most real-world situations, but it should be recognized that formulations for error estimates of the slope and intercept of the fits will be different for other error distributions.

Some representative examples of exact and approximate procedures (discussed below) from the literature were applied to the sample data sets, and the results of the fits are shown in Tables 2 and 3. In each case, slopes and intercepts were derived by fitting y on x, and by exchanging the x and y variables, thus fitting x on y. The slopes and intercepts for the latter case were made comparable to those of the former case by calculating the equivalent values for $y{=}mx{+}b$ (since $x{=}y/m{-}b/m$, then m'=1/m and $b'{=}-b/m$). For methods that properly account for errors in both variables, the fit parameters by these two approaches should be identical (i.e. m' from fitting x on y should equal m from fitting y on x, and similarly for b' and b). This is termed invariance to exchange of x and y. Proper fitting methods should also be invariant to change of scale (i.e. fit parameters do not depend on the choice of units for x and y). Several numerical digits are shown in Tables 2 and 3, not all significant, so that the results from the various methods can be accurately compared.

The method described by York (1966; 1968) and York et al. (2004) was applied to the sample data sets. This involves iteratively solving the following equations (Eq. 5). This method allows for correlation between the x and y errors, indicated by r_i (different than the r_{xy} in Eq. 3), which is set to zero in the present case (i.e. errors are assumed to be uncorrelated).

$$\begin{aligned}
&b = \overline{y} - m\ \overline{x} \quad m{=}\frac{\sum W_i\beta_i V_i}{\sum W_i\beta_i U_i}\\
&\overline{x} = \sum W_i x_i \Big/ \sum W_i \quad \overline{y}{=}\sum W_i y_i \Big/ \sum W_i\\
&U_i = x_i - \overline{x} \quad V_i{=}y_i{-}\overline{y} \quad W_i{=}\frac{w_{xi}w_{yi}}{w_{xi}{+}m^2 w_{yi}{-}2mr_i\alpha_i}\\
&\beta_i = W_i\left[\frac{U_i}{w_{yi}} + \frac{mV_i}{w_{xi}} - (mU_i+V_i)\frac{r_i}{\alpha_i}\right] \quad \alpha_i = \sqrt{w_{xi}w_{yi}}
\end{aligned} \tag{5}$$

The procedure is to assume a starting value for m, calculate W_i, U_i, V_i, α_i, and β_i, and then calculate a revised value for m. This process is repeated until m changes by some small increment according to the accuracy desired. This is a simpler implementation of an earlier method of York (1966), which was described in York (1969) and York et al. (2004), and is the same as the method of Williamson (1968), if the x and y errors are uncorrelated (i.e. r_i=0). The method of Williamson (1968) has been praised in the literature (MacTaggert and Farwell, 1992; Kalantar, 1990) as being efficiently able to converge to the correct answer. Other approaches (including the earlier York method), may not always converge or may be slow to do so, depending on the specific data set. As with standard least-squares, one can perform bivariate fits without weighting. This is done by making all the weights the same (e.g. 1).

The uncertainties in the slope and intercept can also be calculated. Among various methods discussed in the literature

(Cecchi, 1991; Kalantar, 1992; Kalantar et al., 1995; Moreno and Bruzzone, 1993; Reed, 1990, 1992; Sheth et al., 1996; Williamson, 1968; York et al., 2004), the following forms appear to lead to correct estimates of the fit parameter uncertainties (after York et al. (2004) with some algebraic manipulation).

$$\sigma_b^2 = \frac{1}{\sum W_i} + \left(\overline{x} + \overline{\beta}\right)^2 \sigma_m^2 \qquad \sigma_m^2 = \frac{1}{\sum W_i \left(\beta_i - \overline{\beta}\right)^2}$$
$$\text{std err } b = \sqrt{\sigma_b^2}\sqrt{\frac{S}{n-2}} \qquad \text{std err } m = \sqrt{\sigma_m^2}\sqrt{\frac{S}{n-2}} \tag{6}$$
$$\overline{\beta} = \sum W_i \beta_i \Big/ \sum W_i$$
$$S = \sum \left[y_i - (mx_i + b)\right]^2$$

The quantity $\sqrt{S\big/(n-2)}$ is a "goodness of fit" parameter. Its expected value is unity. Its deviation from unity can be used to adjust the weighting factors (in a global sense), but the bivariate slope and intercept will not be affected.

Another straightforward method is that of Neri et al. (1989). This involves minimization of the shortest distance between the fitted line and that data points, and assumes the x and y errors are uncorrelated. The following equations are utilized.

$$\sum W_i x_i (mx_i + b - y_i) - \sum \frac{W_i^2 m (mx_i + b - y_i)^2}{w_{xi}} = 0$$
$$b = \sum W_i \left(y_i - mx_i\right) \Big/ \sum W_i \tag{7}$$
$$W_i = \frac{w_{xi} w_{yi}}{w_{xi} + m^2 w_{yi}}$$

In this method, an initial m is guessed (such as from standard least-squares or by inspection), b is calculated (second equation in Eq. 7), and then m is adjusted to minimize the left hand side of the first equation in Eq. (7). The process is repeated until the left side of the first equation in Eq. (7) is satisfactorily close to zero. The Williamson-York and Neri et al. methods give identical results for the slope and intercept of the two test data sets.

Four other methods give results that are reasonably close to the above results, but are not exactly the same, and do not always give the same slope on exchange of the x and y variables. These approximate methods may be satisfactory for many applications.

Reed (1992) suggests finding roots of the following quadratic expression.

$$g(m) = Am^2 + Bm + C = 0$$
$$A = \sum \frac{W_i^2 U_i V_i}{w_{xi}} \; B = \sum W_i^2 \left(\frac{U_i^2}{w_{yi}} - \frac{V_i^2}{w_{xi}}\right) \; C = -\sum \frac{W_i^2 U_i V_i}{w_{yi}} \tag{8}$$

This equation is solved for m by the quadratic formula, $m = \left(-B \pm \sqrt{B^2 - 4AC}\right)\Big/2A$, where the choice of roots is refined by comparison with standard least-squares or by inspection.

Macdonald and Thompson (1992) describe a number of cases for which their method is applicable. They have made available a FORTRAN program that applies their procedures, which provides nearly exact results for the Pearson-York data set. Similarly, Lybanon (1984) presents a detailed method that also yields results very close to those of the "exact" methods. Krane and Schecter (1982) put forward a method proposed by Barker and Diana (1974) and discussed by others (Irwin and Quickenden, 1983; Orear, 1984; Lybanon, 1984b) that is called "effective variance". One begins with Eq. (4), but the weights, w_i, are adjusted to the following form.

$$w_i = \frac{w_{xi} w_{yi}}{w_{xi} + m^2 w_{yi}} \tag{9}$$

This is same as York's W_i value with uncorrelated errors. Since m appears in the formula for the weight, an iterative process is required, in which an initial m value is guessed, w_i is calculated, followed by calculation of a revised m. The result differs from the "exact" methods for the Pearson-York data set by a few percent, but it is more accurate than the standard least-squares. This method does not retrieve the same slope and intercept when the x- and y-variables are exchanged. The errors are larger with the "synthetic" data set.

Press et al. (1992a,b) present a method called "maximum likelihood estimation" and include a routine written in C or FORTRAN for its implementation. This is also discussed by Titterington and Halliday (1979). York et al. (2004) demonstrate that their method and "maximum likelihood estimation" are mathematically identical. Brauers and Finlayson-Pitts (1997) applied the Press et al. method to analysis of kinetic data.

The methods of Williamson (1968), York (1969), York et al. (2004) and Neri et al. (1989) all agree and appear to provide the exact answer to the best fit for the Pearson-York data set. The approaches of Reed (1992), Macdonald and Thompson (1992), and Lybanon (1984) provide results very close to the exact ones. The "effective variance" method performs reasonably well for the Pearson-York data set, but poorer for the synthetic data. Because of this variability in performance, it should be used with caution.

4 Comparing the methods

A more detailed examination of the behavior of bivariate and standard least-squares as a function of the random noise added in the "synthetic data" was performed. The purpose here is to advise the reader when the more involved bivariate methods should be used or when the standard least-squares are expected to provide satisfactory values for the fit parameters. A series of calculations was performed in which random noise was sampled from Gaussian distributions with varying constant and proportional standard deviations (like the second test data set used above). Standard least-squares (without

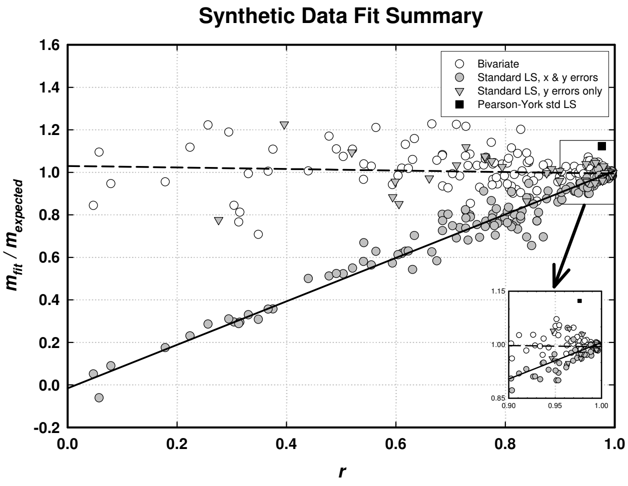

Fig. 3. Ratio of fitted to expected slopes ($m_{\text{fit}}/m_{\text{expected}}$) from standard least-squares and the Williamson-York bivariate method versus r-values from Eq. (3). Errors in both the x and y variables lead to systematic errors in the slope from standard least-squares. Slopes from the bivariate method show no such systematic variation with r.

weights) were applied to the data sets, as was the method of Williamson-York. The values of r (Eq. 3) were also calculated. This test has the advantage that the "correct" slope and intercept are known (1 and 0, respectively). Note that the errors are normally distributed, which may not necessarily be the case in "real" data sets.

Figure 3 shows $m_{\text{fit}}/m_{\text{expected}}$ versus r of the best fit lines using standard least-squares when proportional uncertainties of zero to 50% and/or constant uncertainties of up to 50 units were applied to the x-data, the y-data, or both. Standard least-squares performs well by retrieving slopes close to unity (the expected value) when errors are applied to the y-data only. However, when errors are added to the x-variable either alone or with errors added to the y-variable, the slopes of the best fit lines from standard least-squares are significantly less than unity. The ratio of the fitted slope to that expected is approximately equal to $|r|$. This is true even when r values are very small. The normal interpretation of these small r values is that the data are uncorrelated and cannot be represented by a linear relationship. In this case, however, we know that there is a linear relationship between x and y because of the way the data were constructed. The F statistics for the standard fits indicate that they are statistically significant at the 95% confidence level for all but those corresponding to the 3 smallest r values.

Applying the Williamson-York bivariate method to the same data sets, leads to slopes within about 20% of the expected value of unity. Note that this is the case even when the data are very noisy and thus correlation coefficients are small. Values much closer to the expected value are retrieved when the data is less noisy (see inset in Fig. 3). These fits were performed with 100 data points. If the sizes of the data sets are increased, the error (scatter) in the slope decreases accordingly. As an example, for a constant error of 28 units, the average error in the slope (5 repetitions) decreases from 19% to 6% to less than 1% as the number of data points goes from 100 to 1000 to 10 000 (an approximate $\sqrt{n}$ relationship).

Knowing that the bivariate methods are an improvement over standard least-squares when there are errors in the x-variable is a start, but can the information gathered be used to indicate when the extra trouble of the bivariate fit is called for, versus when standard least-squares will suffice. Figure 3 shows that there is a rather robust relationship between the systematic error in the slope from standard least-squares and the absolute value of the correlation coefficient (as expected, comparing Eqs. 1 and 3). For errors in both variables, the fractional error in the standard least-squares slope is approximately 1–$|r|$. Thus, a quick calculation of the correlation coefficient can give a rough indication of the error in the derived standard least-squares slope for data with comparable errors in both variables. If this error in the slope is outside the needs of the task at hand, then a bivariate approach should be employed. For unusual weighting situations (such as the Pearson-York data), it is probably best to always use robust bivariate methods, since the impact of such weights on the fit parameters is not intuitive (although in this specific case, the standard least-squares slope is only in error by 12%). When the error in the y-variable is much greater than the error in x-variable, then standard least-squares performs better than indicated by the calculated r value.

5 Application to actual observations

Two authentic sets of data from the TRACE-P campaign (TRansport And Chemistry Experiment – Pacific) were selected for application of these fitting procedures. TRACE-P involved two aircraft (the NASA DC-8 and P3-B) as platforms for observations primarily in the western Pacific Ocean basic. The observations used here are gas-phase formaldehyde (CH_2O) concentrations collected by Alan Fried and colleagues aboard the NASA DC-8 aircraft (Fried et al., 2003), and peroxy radical concentrations (HO_2+RO_2) collected by the author and colleagues aboard the NASA P-3B aircraft (Cantrell et al., 2003). These data represent very typical situations that might require the fitting procedures discussed here.

The details of the measurement techniques and the modeling approaches can be found in the references cited above. Briefly, CH_2O was measured in the NASA DC-8 aircraft in a low-pressure cell with multi-pass optics (100 m path total optical path) using a tunable lead salt diode infrared laser as the source. A spectral line near 2831.6 cm^{-1} was scanned and the second harmonic spectrum (after subtraction of the background) was related to the ambient concentration through addition of known mixtures of CH_2O in zero air to the instrument inlet. The measurements were corrected for a small interference from methanol. The estimated uncertainty of the measurements was 15%, and detection limits typically ranged from 50 to 80 pptv (parts per trillion by volume). One minute average retrieved concentrations ranged from –47 to 10 665 pptv. Concentrations measured below the detection limit were used as observed in the fits described here.

HO_2+RO_2 concentrations were measured on the NASA P-3B aircraft and were determined by conversion to gas-phase sulfuric acid through the addition of reagent gases NO and SO_2 to the instrument inlet. The sulfuric acid product was ionized by reaction with negatively charged nitrate ions. The product and reagent ions were quantified by quadrupole mass spectrometry. Calibrations were performed using quantitative photolysis of water vapor at 184.9 nm. The estimated uncertainty for these data was 17% and the detection limits were 2–5 pptv. Concentrations below the detection limit were used as observed in the fits described here.

CH_2O and HO_2+RO_2 concentrations were estimated by a photochemical box model with inputs of key parameters constrained by the observations (Crawford et al., 1999; Olson et al., 2004). The time-dependent model is run for several days to diurnal steady state. Monte Carlo calculations yielded uncertainty estimates of 20% for modeled CH_2O and 30% for HO_2+RO_2.

Figure 4 shows the measured CH_2O concentrations versus those estimated by the constrained box model on linear scales (4466 data pairs). The inset plots show the high range of concentrations (>500 pptv, lower right) and the data plotted on logarithmic scales (upper left). The lines represent different methods of fitting the data. The solid line is a weighted bivariate fit to all of the data with the measurements weighted using a variance of the square of 15% of the concentration plus 50 pptv, and the model results using a variance of the square of 20% of the concentration. The slope is near unity (1.054, standard error=0.0144) and the y-intercept is small (1.283, standard error=2.046), in agreement with assessments by Fried et al. (2003) and Olson et al. (2004). The correlation coefficient squared, r^2 is 0.856. The long dashed line is a standard unweighted least-squares fit which yields a slope of 1.462 (standard error=0.0090) and a y-intercept of –44.6 (standard error 3.16). It appears that the line is being unduly weighted by the handful of points at high concentrations in which the model systematically underestimates the observations, leading to a larger slope than the bivariate method. The medium dashed line is a weighted least-squares fit (Eq. 4), with weights calculated using the "effective variance" method (Eq. 9). Its slope is 0.873 and the y-intercept is 20.1. Finally, the short dashed line is a weighted least-squares fit (equation 4) with weights in the y-direction only (i.e. $w_i=w_{yi}$). The slope for this fit is 0.811 (std err=0.012) and the y-intercept is 22.4 (std err=2.14). These fits mostly have slopes of unity within the combined measurement-model uncertainties (0.25, 1σ), with the exception of the standard unweighted least-squares fit. The intercepts are all within the detection limit of the measurements (around 50 pptv). The large slope retrieved with the standard unweighted approach could lead one to make the assessment that there are missing processes in the model, errors in the measurements, or both. While it does appear that there are statistically significant differences between the measurements and the model at high concentrations, the small number of outliers should not significantly change the fit of the entire data set. Eliminating data pairs with measurements greater than 4000 pptv, results in bivariate fit slope and y-intercept values of 1.041 and 2.476, respectively. The weighted standard fits change by small amounts as well. The unweighted standard fit, though, yields slope and y-intercept values of 1.248 and –5.744, respectively. This is a significant change and shows how susceptible the standard fit is to a small number of outliers (the term outlier is used here to mean data that are not described well by the bivariate fit line).

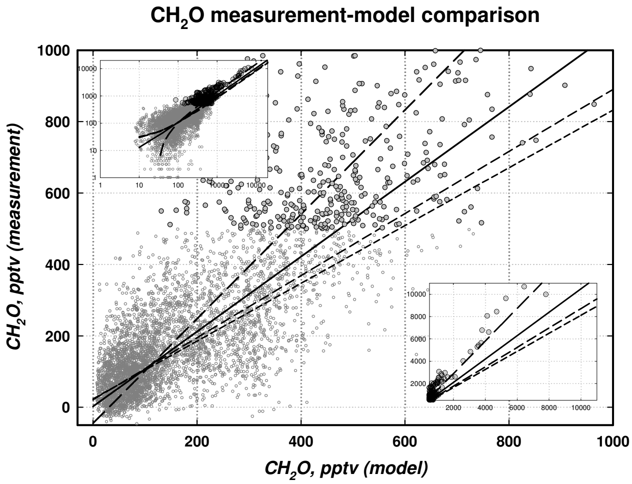

Fig. 4. Comparison of measured formaldehyde concentrations with those estimated from a constrained box model during the TRACE-P campaign (after Fried et al., 2003; Olson et al., 2004). The data points are divided into two groups: those corresponding to measurements below 500 pptv (small points), and those for measurements above 500 pptv (large points). The main window (on linear scales) shows results of linear fits using four approaches: solid line, bivariate weighted fit to all data; long dash, standard unweighted least-squares fit; medium dash, fit using weighted standard least-squares (Eq. 4) with weights calculated using effective variance; and short dash, fit using weighted standard least-squares with weights in the y-direction only. The lower right inset shows the fit lines and data on expanded x- and y-scales (linear). The upper left inset shows the full range of data on logarithmic scales. See text for fit parameters and discussion.

The impact of outliers on the various fit methods is demonstrated further. To the full data set are added numbers of data pairs (up to 1000) for which x is 50 and y is 5000. A second trial added data pairs with x values of 5000 x and y values

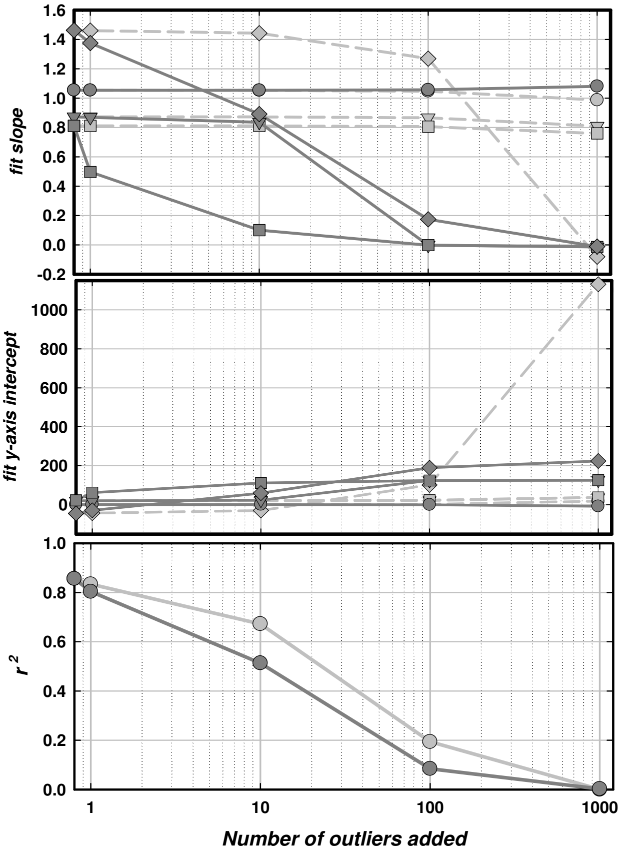

Fig. 5. Impacts of added data outliers to the formaldehyde dataset presented in Fig. 4. Shown are slopes (top panel), intercepts (middle panel), and correlation coefficients (bottom panel) of various fits as impacted by adding extra points, in amounts indicated on the x-axis, to the dataset that are clearly outliers. Eight collections of fit parameters are shown for 1, 10, 100, and 1000 outliers added. Four collections had outliers equal to x=50, y=5000 (dark gray); the other four had outliers equal to x=5000, y=50 (light gray). The circles in the top two panels represent parameters derived from weighted bivariate fits; the downward pointing triangles represent parameters derived from Eq. (4) using effective variance; the squares represent parameters derived from Eq. (4) with weights in the y-direction only; and the diamonds represent parameters derived from unweighted standard least-squares. The values on the y-axis (corresponding to x=0.8) are those derived from the original formaldehyde data with no added outliers.

of 50. These results are summarized in Fig. 5. It can be seen that outliers above the fit line have little impact on the bivariate and the other weighted fit slopes, even when the number of outliers approaches 20% of the data. The standard unweighted least-squares fit is affected moderately by outliers above the fit line. Outliers below the fit line impact all of the fits greatly except the bivariate. In fact, as shown before, the bivariate fit procedure continues to perform well even when the r^2 parameter indicates that the x and y data are completely uncorrelated. While there have been various techniques put forward to eliminate outliers (e.g. the Q-test, Dean and Dixon, 1951) that can applied, these exercises show that the bivariate fit method is relatively insensitive to outliers.

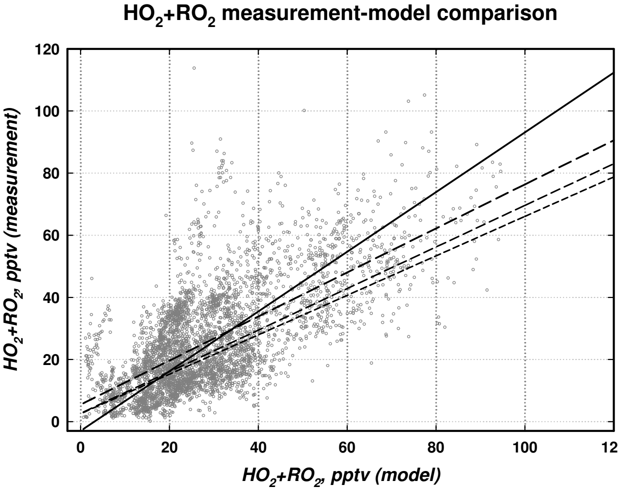

Fig. 6. Fits of HO_2+RO_2 measurements versus constrained box model estimates. The lines are four different fit approaches: solid line, bivariate weighted fit to all data; long dash, standard unweighted least-squares fit; medium dash, fit using weighted standard least-squares (Eq. 4) with weights calculated using effective variance; and short dash, fit using weighted standard least-squares with weights in the y-direction only.

As mentioned earlier, and discussed by Fried et al. (2003), there appears to be a change in the ratio of measurement to model values from near unity at lower concentrations to well above unity at higher concentrations. As one approach, the data were separated into two groups for measured values below and above 500 pptv, and each group was fit separately. The bivariate slope of the low concentration group is 0.789, while the bivariate slope of the high concentration group is 1.403. An alternate method is to fit the ratio of measurement to model ($[CH_2O]_{meas}/[CH_2O]_{model}$) versus measurement value. Separating into two groups as before leads to a bivariate slope of 0.00607 for the low concentration group (i.e. moderate dependence of the ratio on the concentration) and an intercept of 0.797 (the ratio at the limit of zero concentration). The slope for the high concentration group is 0.000679 and the intercept is 1.290. It seems that there could be atmospheric processes missing from the model or instrumental issues affecting the measurements in the high concentration regime that need to be addressed (in agreement with Fried et al., 2003).

Fits of measured versus modeled HO_2+RO_2 are shown along with the data in Fig. 6. The solid line is a bivariate fit weighted using variances for the measurements that are the square of 20% of the concentration plus 5, and using variances for the model results that are the square of 30% of model values. Its slope is 0.961 (std err=0.015) and the intercept is –2.96 (std err=0.35). The correlation coefficient squared, r^2, is 0.437. The other methods (effective variance, y-weighting only, and no weighting) yield smaller slopes (0.63 to .71). There are some noticeable outliers in which the measured concentrations are systematically higher than the modeled ones at low modeled concentrations. Elimination of these data does not greatly affect the bivariate fit. A fit of the measured to model ratios versus measured values yields a moderate slope (–0.00864) and an intercept near unity (1.053).

It has been reported (Faloona et al., 2000) that measured peroxy radical concentrations are systematically greater than model values at high NO_x concentrations. This is observed for TRACE-P HO_2+RO_2 data as well. For NO concentrations less than 500 pptv, the measured to modeled ratios are close to unity with no significant dependence on NO concentration. The bivariate fit yields a slope of –0.00145 and a y-intercept of 0.77. For NO concentrations greater than 500 pptv, there is a systematic dependence of the measured-modeled ratio on the NO concentration. The bivariate fit slope is 0.00317 and the y-intercept is –0.785. It has been suggested (Olson et al., 2006) that this phenomenon could be the result of short term large spikes in the NO concentration that impact the average NO concentration, but have little impact on the average peroxy radical concentration. Without high rate NO and peroxy radical data, we cannot rule out such an explanation. Alternatively, there could be unknown photochemical processes or instrumental issues that occur in the presence of high NO concentrations. The measurement and modeling communities continue to search for satisfactory explanations of these observations under high NO conditions.

Does the quality of fits obtained with the bivariate methods depend strongly on the selection of weights? This was examined using the CH_2O measurements and model results. The best estimate for the variance of the measurements is $(0.15\times[CH_2O]_{meas}+50)^2$, and for the model values is $(0.20\times[CH_2O]_{model})^2$. Varying the measured variance values from $(0.10\times[CH_2O]_{meas}+50)^2$ to $(0.30\times[CH_2O]_{meas}+200)^2$ results in bivariate fitted slopes ranging from 0.88 to 1.15. Thus, while there is some impact on the fit parameters by the choice of weights, the dependence is not strong. Obviously, every effort should be made to correctly estimate the weights, but small errors in these parameters are not likely to invalidate the fit results.

6 Summary

Scientists need to use care in applying fitting programs to derive parameters that summarize their observations. In the case of linear fits, significant errors in slopes and intercepts can result using standard least-squares methods if there are uncertainties in the x-values (as cautioned many times in the literature). If the x- and y-variable errors are comparable, $1-|r|$ may give an indication of the fractional error of the derived standard least-squares slope. If a more accurate slope is desired, then bivariate methods such as those reported by Williamson et al., York et al., or Neri et al. are recommended. For these methods, the accuracy of the slope improves with the number of data points (not so with the standard least-squares with significant errors in the x-variable).

7 Supplemental material

The Williamson-York method has been incorporated into a Microsoft Excel® spreadsheet available as supplemental material. http://www.atmos-chem-phys.net/8/5477/2008/acp-8-5477-2008-supplement.zip

Acknowledgements. The author acknowledges gratefully the help of Norm Evensen and Andi Andreae in the refinement of the spreadsheet tool. The National Center for Atmospheric Research is sponsored by the National Science Foundation. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author and do not necessarily reflect the views of the National Science Foundation.

Edited by: V. F. McNeill

References

Barker, D. R. and Diana, L. M.: Simple method for fitting data when both variables have uncertainties, Am. J. Phys., 42, 224–227, 1974.

Bevington, P. R.: Data reduction error analysis for the physical sciences, McGraw-Hill Book Company, New York, 1969.

Borcherds, P. H. and Sheth, C. V.: Least squares fitting of a straight line to a set of data points, Eur. J. Phys. 16, 1, 204–210, 1995.

Brauers, T. and Finlayson-Pitts, B. J.: Analysis of relative rate measurements, Int. J. Chem. Kin., 29, 665–672, 1997.

Bruzzone H. and Moreno, C.: When errors in both coordinates make a difference in the fitting of straight lines by least squares, Meas. Sci. Technol., 9, 2007–2011, 1998.

Cantrell, C. A., Edwards, G. D., Stephens, S., Mauldin, R. L., Zondlo, M. A., Kosciuch, E., Eisele, F. L., Shetter, R. E., Lefer, B. L., Hall, S., Flocke, F., Weinheimer, A. Fried, E. Apel, Y. Kondo, D. R. Blake, N. J. Blake, I. J. Simpson, A., Bandy, A. R., Thornton, D. C., Heikes, B. G., Singh, H. B., Brune, W. H., Harder, H., Martinez, M., Jacob, D. J., Avery, M. A., Barrick, J. D., Sachse, G. W., Olson, J. R., Crawford, J. H., and Clark, A. D.: Peroxy radical behavior during the Transport and Chemical Evolution over the Pacific (TRACE-P) campaign as measured aboard the NASA P-3B aircraft, J. Geophys. Res., 108(D20), 8797, doi:10.1029/2003JD003674, 2003.

Cecchi, G. C.: Error analysis of the parameters of a least-squares determined curve when both variables have uncertainties, Meas. Sci. Technol., 2, 1127–1128, 1991.

Chong, D. P.: Comment on “Linear least-squares fits with errors in both coordinates” by B. C. Reed, Am. J. Phys. 57, 642–646, 1989, Am. J. Phys., 59, 472–474, 1991.

Chong, D. P.: On the use of least squares to fit data in linear form, J. Chem. Ed., 71, 489–490, 1994.

Christian, S. D. and Tucker, E. E.: Analysis with the microcomputer. 5. General least-squares with variable weighting, Am. Lab., 16(2), 18–18, 1984.

Christian, S. D., Tucker, E. E., and Enwall, E.: Least-squares analysis: A primer, Am. Lab., 18, 41–49, 1986.

Crawford, J., Davis, D., Olson, J., Chen, G., Liu, S., Gregory, G., Barrick, J., Sachse, G., Sandholm, S., Heikes, B., Singh, H., and Blake, D.: Assessment of upper tropospheric HO_x sources over the tropical Pacific based on NASA GTE/PEM data: Net effect on HO_x and other photochemical parameters, J. Geophys. Res., 104(D13), 16 255–16 273, 1999.

Dean, R. B. and Dixon, W. J.: Simplified statistics for small numbers of observations, Anal. Chem., 23, 636–638, doi:10.1021/ac60052a025, 1951.

Faloona, I., Tan, D., Brune, W. H., Jaeglé, L., Jacob, D. J., Kondo, Y., Koike, M., Chatfield, R., Pueschel, R., Ferry, G., Sachse, G., Vay, S., Anderson, B., Hannon, J., and Fuelberg, H.: Observations of HO_x and its relationship with NO_x in the upper troposphere during SONEX, J. Geophys. Res., 105(D3), 3771–3783, 2000.

Fried, A., Crawford, J., Olson, J., Walega, J., Potter, W., Wert, B., Jordan, C., Anderson, B., Shetter, R., Lefer, B., Blake, D., Blake, N., Meinardi, S., Heikes, B., O'Sullivan, D., Snow, J., Fuelberg, H., Kiley, C. M., Sandholm, S., Tan, D., Sachse, G., Singh, H., Faloona, I., Harward, C. N., and Carmichael, G. R.: Airborne tunable diode laser measurements of formaldehyde during TRACE-P: Distributions and box model comparisons, J. Geophys. Res., 108(D20), 8798, doi:10/1029/2003JD003451, 2003.

Gonzalez, A. G., Marquez, A., and Fernandez Sanz, J.: An iterative algorithm for consistent and unbiased estimation of linear regression parameters when there are errors in both the x and y variables, Computers Chem., 16(1), 25–27, 1992.

Irwin, J. A. and Quickenden, T. I.: Linear least squares treatement when there are errors in both x and y, J. Chem. Ed., 60(5), 711–712, 1983.

Jones, T. A.: Fitting straight lines when both variables are subject to error. I. Maximum likelihood and least-squares estimation, Math. Geo., 11(1), 1–25, 1979.

Kalantar, A. H.: Weighted least squares evaluation of slope from data having errors in both axes, Trends in Anal. Chem., 9(1), 149–151, 1990.

Kalantar, A. H.: Kerrich's method for $y=\alpha x$ data when both y and x are uncertain, J. Chem Ed., 68(1), 368–370, 1991.

Kalantar, A. H.: Straight-line parameters' errors propagated from the errors in both coordinates, Meas. Sci. Technol., 3, 1113–1113, 1992.

Kalantar, A. H., Gelb, R. I., and Alper, J. S.: Biases in summary statistics of slopes and intercepts in linear regression with errors in both variables, Talanta, 42(4), 597–603, 1995.

Krane, K. S. and Schecter, L.: Regression line analysis, Am. J. Phys., 50(1), 82–84, 1982.

Leduc, D. J.: A comparative analysis of the reduced major axis technique of fitting lines to bivariate data, Can. J. For. Res., 17, 654–659, 1987.

Lybanon, M.: A better least-squares method when both variables have uncertainties, Am. J. Phys., 52(1), 22–26, 1984a.

Lybanon, M.: Comment on "Least squares when both variables have uncertainties", Am. J. Phys., 52(3), 276–278, 1984b.

Lybanon, M.: A simple generalized least-squares algorithm, Comp. Geosci., 11, 501–508, 1985.

Macdonald, J. R. and Thompson, W. J.: Least-squares fitting when both variables contain errors: Pitfalls and possibilities, Am. J. Phys., 60(1), 66–73, 1992.

MacTaggart, D. L. and Farwell, S. O.: Analytical use of linear regression, Part II: Statistical error in both variables, J. AOAC Intl., 75(4), 608–614, 1992.

Markovsky, I. and Van Huffel, S.: Overview of total least-squares methods, Sig. Proc., 87, 2283–2302, 2007.

Moreno, C., and Bruzzone, H.: Parameters' variances of a least-squares determined straight line with errors in both coordinates, Meas. Sci. Technol., 4, 635–636, 1993.

Moreno, C.: A least-squares-based method for determining the ratio between two measured quantities, Meas. Sci. Technol., 7, 137–141,1996.

Neri, F., G. Saitta, and S. Chiofalo: An accurate and straightforward approach to line regression analysis of error-affected experimental data, J. Phys. E: Sci. Inst., 22, 215–217, 1989.

Neri, F., Patanè, S., and Saitta, G.: Error-affected experimental data analysis: application to fitting procedure, Meas. Sci. Technol., 1, 1007–1010, 1990.

Olson, J. R., Crawford, J. H., Chen, G., Fried, A., Evans, M. J., Jordan, C. E., Sandholm, S. T., Davis, D. D., Anderson, B. E., Avery, M. A., Barrick, J. D., Black, D. R., Brune, W. H., Eisele, F. L., Flocke, F., Harder, H., Jacob, D. J., Kondo, Y., Lefer, B. L., Martinez, M., Mauldin, R. L., Sachse, G. W., Shetter, R. E., Singh, H. B., Talbot, R. W., and Tan, D.: Testing fast photochemical theory during TRACE-P based on measurements of OH, HO_2, and CH_2O, J. Geophys. Res., 109, D15S10, doi:10.1029/2003JD004278, 2004.

Olson, J. R., Crawford, J. H., Chen, G., Brune, W. H., Faloona, I. C., Tan, D., Harder, H., and Martinez, M.: A reevaluation of airborne HO_x observations from NASA field campaigns, J. Geophys. Res., 111, D10301, doi: 10.1029/2005/JD006617, 2006.

Orear, J.: Least squares when both variables have uncertainties, Am. J. Phys., 50, 912–916, 1982; Erratum, Am. J. Phys., 52, 278–278, 1984.

Pasachoff, J. M.: Applicability of least-squares formula, Am. J. Phys., 48(10), 800–800, 1980.

Pearson, K.: On lines and planes of closet fit to systems of points in space, Philos. Mag., 2(2), 559–572, 1901.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P.: Numerical Recipes in C, 2nd ed., Cambridge University Press, 1992a.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P.: Numerical Recipes in Fortran, 2nd ed., Cambridge University Press, 1992b.

Reed, B. C.: Linear least-squares fits with errors in both coordinates, Am. J. Phys., 57(3), 642–646, 1989; Erratum, Am J. Phys., 58(2), 189–189, 1990.

Reed, B. C.: Linear least-squares fits with errors in both coordinates. II: Comments on parameter variances, Am. J. Phys., 60(1), 59–62, 1992.

Riu, J. and Rius, F. X.: Univariate regression models with errors in both axes, J. Chemometrics, 9, 343–362, 1995.

Sheth, C. V. A.: Ngwengwe, and P. H. Borcherds, Least squares fitting of a straight line to a set of data points: II: Parameter variances, Eur. J. Phys., 17(2), 322–326, 1996.

Squire, P. T.: Comment on "Linear least-squares fits with errors in both coordinates” by B. C. Reed, Am. J. Phys. 57, 642–646, 1989, Am. J. Phys., 58(12), 1209–1209, 1990.

Titterington, D. M. and Halliday, A. N.: On the fitting of parallel isochrones and the method of maximum likelihood, Chem. Geol. 26, 183–195, 1979.

Williamson, J. H.: Least-squares fitting of a straight line, Can. J. Phys., 46, 1845–1847, 1968.

York, D.: Least-squares fitting of a straight line, Can. J. Phys., 44, 1079–1086, 1966.

York, D.: Least squares fitting of a straight line with correlated errors, Earth and Planet. Sci. Lett., 5, 320–324, 1969.

York, D., Evensen, N. M., López Martinez, M., and De Basabe Delgado, J.: Unified equations for the slope, intercept, and standard errors of the best straight line, Am J. Phys., 72(3), 367–375, 2004.