A spreadsheet for linear least-squares fitting with errors in both coordinates

2010 Phys. Educ. 45 93

(http://iopscience.iop.org/0031-9120/45/1/011)

A spreadsheet for linear least-squares fitting with errors in both coordinates

B Cameron Reed

Department of Physics, Alma College, Alma, MI 48801, USA

E-mail: reed@alma.edu

Abstract
A freely available, easy-to-use Excel spreadsheet for performing linear least-squares fits for (x, y) data with errors in both coordinates is described.

Introduction

It is almost certain that every student and practitioner of physical science will at some point face the task of determining the slope and intercept of the 'best' straight line fit through a set of data in the (x, y) plane. With modern calculators and spreadsheet software the task of performing such a 'linear least-squares regression' when the data points are *unweighted* is practically trivial. If the data carry uncertainties in the values of the dependent quantities, that is, in the y-direction, the usual procedure is to weight each point as the inverse square of its uncertainty $\delta(y_i)$, and the appropriate algorithm in this case is developed in many standard texts [1]. A much trickier business, however, is that of the more realistic situation where there are uncertainties in both coordinates. While the literature on this situation goes back to Carl Friedrich Gauss, it was not until 1966 that a general solution to this problem was developed by Canadian geophysicist Derek York [2]. The expression for the slope derived by York is highly non-linear and requires an iterative solution. Unfortunately, the implementation proposed by York—in the days before easy access to personal computing—could lead to incorrect results in some circumstances. The present author corrected and clarified York's algorithm in a pair of papers [3, 4], and York *et al* subsequently published a summary paper on this issue [5]. It should be pointed out that references [2–5] contain a number of references to other contributions on this issue, about which hundreds of papers have been published over the years.

At the time of my work on this problem I developed a FORTRAN program for carrying out the 'fit-with-errors-in-both-coordinates' computation. Now, however, thanks to the availability of powerful spreadsheet software, these computations can be quickly run by users who need no particular training in the underlying mathematics of the method or familiarity with any particular computer language. The purpose of this paper is to describe and make available to the physics teaching community a Microsoft Excel spreadsheet for carrying out this computation in a few simple steps. As treatment of experimental uncertainties and their analysis enters even into secondary-school circles, such a spreadsheet should prove a useful tool at a variety of levels.

The least-squares residual

The fundamental premise of least-squares fitting is that one desires to determine the slope m and intercept c of the straight line

$$Y = mX + c, \tag{1}$$

0031-9120/10/010093+04$30.00

where X and Y are the *predicted* (also known as *calculated* or *adjusted*) values of the data points, that is, their values after 'adjustment' to the best-fit line. The 'best fit' is defined as minimizing the sum of the weighted squared residuals (differences between observed and calculated values), commonly designated as S:

$$S = \sum_{i=1}^{N} [W(x_i)(x_i - X_i)^2 + W(y_i)(y_i - Y_i)^2], \quad (2)$$

where $(x_i,\ y_i)$ are the observed values of the data points, which are numbered $i = 1$ to N. The $W(x_i)$ and $W(y_i)$ are the x- and y-direction weights of each point; for reasons discussed in [4], these are conventionally assumed to be assigned as the inverse squares of the uncertainties in the data values.

Using a variational method, York [2] found that the best-fit slope is given by the root of a function which we can designate as $g(m)$, that is, by the value of m which renders $g(m) = 0$; the intercept c follows from m via an expression involving the values of the observed points and the best-fit value of m. York expressed this function in pseudo-cubic form, and other authors have developed equivalent pseudo-quadratic and pseudo-linear versions. 'Pseudo-' here means that the coefficients of the cubic, quadratic, or linear constraint equation are actually functions of m, so one is forced to consider an iterative solution. The usual approach is to start with an initial guess for m (such as that from a simple no-errors solution), and then use some root-finding or exploratory algorithm until $g(m) = 0$ to a desired level of accuracy.

The mathematics of York's method is challenging, but one does not need to understand it in order to operate the spreadsheets described here. Two versions of the best-fit spreadsheet are freely available for downloading [6]; both are based on expressions given in [4]. In one, LLS(SIGMAS).xls, the user enters the $(x_i,\ y_i)$ values and their uncertainties, and the spreadsheet computes the corresponding inverse-square weights. In the other, LLS(WEIGHTS).xls, the user enters the $(x_i,\ y_i)$ values and the $W(x_i)$ and $W(y_i)$ weights directly. Otherwise, the two spreadsheets operate identically. Both are set up to accommodate up to 1000 data points. The built-in unweighted least-squares fitting routine gives an estimate of m which can be used as a seed value for evaluating $g(m)$. The Excel 'Goal seek' function is then run on $g(m)$ by adjusting m until $g(m) = 0$ to a desired level of accuracy.

The spreadsheet also computes uncertainties in the slope and intercept, δ_m and δ_c. This requires a brief description of a subtle technical issue. The expression for calculating δ_m is

$$\delta_m = \sqrt{\frac{S}{N-2}\sum_{i=1}^{N}\left[\frac{1}{W(x_i)}\left(\frac{\partial m}{\partial x_i}\right)^2 + \frac{1}{W(y_i)}\left(\frac{\partial m}{\partial y_i}\right)^2\right]}, \quad (3)$$

where S is the sum of squared residuals in equation (2). An explicit analytic form for δ_m appears in [4]; that for δ_c is identical except that the derivatives are of c as opposed to m. Some statistical authorities do not include the factor $S/(N-2)$ here, although the vast majority do so, as do my spreadsheets. The subtlety is that there are two ways of calculating δ_m and δ_c: the derivatives can be evaluated either at the *observed* data points $(x_i,\ y_i)$ or at the *adjusted* data points $(X_i,\ Y_i)$. If the data are well correlated, the difference between these two approaches should be slight. Strictly, one should evaluate the derivatives at the $(X_i,\ Y_i)$ as equation (3) fundamentally derives from a Taylor-series expansion about the calculated points. Equation (3) (and the corresponding expression for δ_c) derive from retaining only first-order terms in the Taylor expansion, and as such are only approximate expressions. My spreadsheets are set up to report δ_m and δ_c as evaluated at both the observed and calculated data points by having the user adjust a corresponding variable that acts as a 'switch' between the two situations.

An application: Magellanic cloud ionized hydrogen regions

As an example of a linear fit with errors in both coordinates I consider some astronomical data: electron temperatures of ionized hydrogen regions in the Magellanic clouds as derived from abundances of doubly ionized oxygen and sulfur [7]. Derived temperatures and their uncertainties for 14 regions are listed in table 1 and plotted in figure 1; all numbers are in units of 10^3 K. This is an interesting case to analyse

Table 1. Magellanic cloud electron temperatures derived from observations of ionized oxygen and sulfur (10^3 K).

T_e [O III]	T_e [S III]	δ [O III]	δ [S III]
13.32	13.96	0.17	1.15
14.99	14.83	0.24	1.17
12.30	13.75	0.50	2.80
10.64	9.00	0.70	1.00
10.10	9.80	0.29	0.70
11.27	9.80	0.70	1.00
9.69	10.40	0.20	1.40
11.30	8.00	0.90	1.50
10.99	10.47	0.60	1.50
10.64	10.73	0.50	1.90
10.57	9.97	0.50	1.00
10.97	10.86	0.50	1.60
9.97	9.23	0.20	0.60
9.71	9.14	0.18	0.60

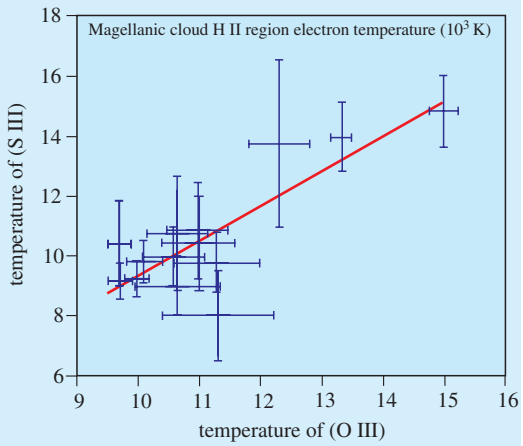

Figure 1. Linear fit for Magellanic-cloud ionized hydrogen region electron temperatures as derived from sulfur vs oxygen abundances. The data appear in table 1. The solid line shows the fit when the errors in both coordinates are accounted for.

since, as often happens with astronomical data, the data do not uniformly sample the range of the independent variable and the uncertainties are fairly large; a conventional unweighted least-squares analysis is not really appropriate. A simple unweighted fit gives $(m, c) = (1.146, -2.098)$ with a correlation coefficient of only $r^2 = 0.690$. With the uncertainties accounted for, the correct fit proves to have $(m, c) = (1.166, -2.313)$. While these values are close to the unweighted values, the agreement is somewhat illusory as the uncertainties in the slope and intercept are substantial: computed at the observed points they are $(\delta_m, \delta_c) = (0.155, 1.654)$; at the adjusted points they are $(\delta_m, \delta_c) = (0.148, 1.591)$. These correspond to uncertainties of $\sim$13% and 70% in the slope and intercept, respectively. The sum of squared residuals is $S = 6.035$. The spreadsheet fit is shown as a solid line in figure 1.

Users who download LLS(WEIGHT).xls will find the historical example of 'Pearson's data with York's weights' pre-loaded in the spreadsheet [2, 4]; in the case of LLS(SIGMAS).xls, the data are those of example I in [3]; users are encouraged to try running the example described above. The conventional 'errors-in-y' situation can be recovered with these spreadsheets by setting the $\delta(x_i)$ to be extremely small, or, correspondingly, the $W(x_i)$ to be extremely large.

Summary

This work makes available an easily used spreadsheet for performing linear least-squares fits when the data have errors in both coordinates. Given the frequency with which this situation arises in both teaching and research circles it is anticipated that this should be a useful contribution to the physical science community. The author would be grateful to receive any suggestions for improvements to the spreadsheets.

Acknowledgments

I am grateful to a reviewer whose careful reading of this paper led to a number of improvements. This paper is dedicated to the memory of B Murray Burgoyne.

Received 1 September 2009, in final form 3 November 2009
doi:10.1088/0031-9120/45/1/011

References

[1] Bevington P R 1969 *Data Reduction and Error Analysis for the Physical Sciences* (New York: McGraw-Hill) chapter 6
[2] York D 1966 Least-squares fitting of a straight line *Can J. Phys.* **44** 1079–86
[3] Reed B C 1989 Linear least-squares fits with errors in both coordinates *Am. J. Phys.* **57** 642–6
Reed B C 1990 Linear least-squares fits with errors in both coordinates *Am. J. Phys.* **58** 189 (erratum)

[4] Reed B C 1992 Linear least-squares fits with errors in both coordinates. II: comments on parameter variances *Am. J. Phys.* **60** 59–62
[5] York D, Evensen N M, Martinez M L and Delgado J D B 2004 Unified equations for the slope, intercept, and standard errors of the best straight line *Am. J. Phys.* **72** 367–75
[6] http://othello.alma.edu/∼reed/LLS(WEIGHTS).xls http://othello.alma.edu/∼reed/LLS(SIGMAS).xls
[7] Vermeij R and van der Hulst J M 2002 The physical structure of Magellanic cloud H II regions: II. Elemental abundances *Astron. Astrophys.* **391** 1081–95

Cameron Reed is a professor in the Department of Physics at Alma College, Alma, MI, USA. He holds a PhD in physics from the University of Waterloo, Canada and is a Fellow of the American Physical Society. His research interests include the study of hot, intrinsically luminous 'OB' stars within our home Milky Way galaxy and the history of the Manhattan Project.