Least squares for different experimental cases

Héctor René Vega Carrillo

Centro Regional de Estudios Nucleares,
Universidad Autónoma de Zacatecas,
Apartado postal 495, 98000 Zacatecas, Zac., México

(Recibido el 1 de marzo de 1989; aceptado el 26 de julio de 1989)

Abstract. We discuss the problem of fitting an experimental data set into a linear curve when experimental uncertainties can not be overlooked. Difficulties with the standard least-squares method are pointed out. An alternative weighted least-squared method, depending on knowledge about the uncertainties for four different cases is presented, and we use it to analyze the data from a particular experiment. From this application it is shown that we get better results using the proper case.

PACS: 02.50.+s; 02.90.+p

1. Introduction

We want to point out a widespread error in the use of linear regression analysis when, in experimental situations, the experimentalist wants to determine the functional relationship between the experimental data.

The broad use of pocket programmable calculators and statistical software for microcomputers makes easier the heavy work necessary to obtain such relationship between the experimental data. Unfortunately, the most common analysis technique available in the least square fitting of a line is not always the most proper.

Experiments in physics made to determine parameters through the functional relationship between values of x and y involve a series of experimental measurements of x and the corresponding y. In several cases there are not only measurement errors in y_i, but also there are measurement errors in x_i. Many experimentalists apply the standard least-squares method, which implicitly assumes errors neither in x_i nor in y_i. Accordingly, this procedure affects the unknown parameters to be obtained from the functional relationship, and it gives estimates of errors that are smaller than the true errors.

This paper presents a review of standard least-squares method as applied to a straight line, and the weighted least-squares method, where the weighting factor is related to the experimental data precision. This gives place to four different experimental cases; each one of these cases are applied to the same experimental data set and the results of this weighted least-square fitting method of a straight line are discussed as well.

2. Review of the standard least-squares method

Suppose that in an experimental set of data points (x_i, y_i), where $i = 1, 2, \ldots, n$, it is assumed that there are not experimental uncertainties.

The standard least square method [1–4] requires that we minimize the quadratic sum, Q, of the ordinate differences between the experimental points, y_i and the required line $Y(x_i)$

$$Q = \sum [y_i - Y(x_i)]^2. \tag{1}$$

Here $\sum$ denotes the sum from $i = 1$ to n.

For a linear fit, $Y(x_i) = mx + b$,

$$Q = \sum [y_i - mx_i - b]^2 \tag{2}$$

must be a minimum. Using the standard differential calculus technique to find the minimum of Q,

$$\frac{\partial Q}{\partial m} = 0, \qquad \frac{\partial Q}{\partial b} = 0, \tag{3}$$

we obtain a system of two equations in the two unknowns m and b. In order to find m and b we write the resulting equations in matrix form;

$$\mathbf{A} = \begin{pmatrix} b \\ m \end{pmatrix} = \mathbb{C}^{-1}\mathbb{S}, \tag{4}$$

where, $\mathbb{C}$ is a 2×2 matrix whose elements C_{jk} are

$$\sum x_i^{j+k-2}, \qquad \text{where } j, k = 1, 2, \ldots, n \tag{5}$$

and $\mathbb{S}$ is a 2×1 column vector whose elements S_{ki} are

$$\sum x_i^{k-1} y_i. \tag{6}$$

The fitted values of m and b are just the elements of the 2×1 column vector $\mathbf{A}$, that is,

$$m = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}, \tag{7}$$

$$b = \frac{\sum y_i \sum x_i^2 - \sum x_i \sum x_i y_i}{n \sum x_i^2 - (\sum x_i)^2}. \tag{8}$$

3. The weighted least square method

In actual experimental cases there are not only uncertainties, Sy_i, for each y_i, but also uncertainties, Sx_i, for each x_i, and these uncertainties do not have the same values. Many experimentalists merely apply the standard least-squares method which implicitly assumes that all the Sx_i's and the Sy_i's are zero. Such a procedure loses accuracy in the determination of the unknown parameters m and b and gives estimates of errors that are smaller than the true errors; this has been previously noted in the literature [5–7].

Now we are going to discuss a more general procedure [8,9], the well-known *weighted least squares method* which requires the minimization of

$$Q = \sum w_i(y_i - mx_i - b)^2. \tag{9}$$

Here w_i are the weights which are related to the experimental uncertainties.

Using the same procedure as in Section 2, one now finds that

$$m = \frac{\sum w_i \sum w_i x_i y_i - \sum w_i x_i \sum w_i y_i}{\sum w_i \sum w_i x_i^2 - \left(\sum w_i x_i\right)^2}. \tag{10}$$

$$b = \frac{\sum w_i x_i^2 \sum w_i y_i - \sum w_i x_i \sum w_i x_i y_i}{\sum w_i \sum w_i x_i^2 - \left(\sum w_i x_i\right)^2}. \tag{11}$$

Here the w_i values represent the square-inverse values of the experimental uncertainty, *i.e.* the variance S, and we can distinguish four cases, when

i) $w_i = 1$,
ii) $w_i = 1/Sy_i$,
iii) $w_i = 1/(Sx_i + Sy_i)$,
iv) $x_i = 1/(Sy_i + m^2 Sx_i)$.

Here Sx_i and Sy_i are the x_i and y_i experimental data variances.

The first case gives the standard method, and it is, implicitly, assumed that there is not errors on the experimental data, (x_i, y_i); while the second case must be applied when $Sy_i \gg Sx_i$ [10,11], and the Sx_i values can be overlooked even if the uncertainties Sy_i are not the same for all.

The third case must be applied if, in your experimental data, you have significant uncertainties for both x_i and y_i [7,10,11], and it is easy and simple to apply. Finally, the last case is the acquainted *effective variance method* [7,10,11] This method must be used when you have the same uncertainty conditions as in the third case, but you want to obtain an improved linear model adjustment.

Here, in order to determine the unknowns m and b it is necessary to know m; to know this, you can use an iterative procedure where the first m value can be obtained using the standard least-squares method.

Comparing the third and fourth cases you find that the third case is easier to apply, while the fourth requires an iterative procedure.

It is quite common that in addition to the m and b values, you need to calculate some secondary parameters, for example in order to use the error propagation expression [12], it is necessary to know the m and b variances. These latter variances for the weighted least squares method [13] are

$$Sm = \left(\frac{\partial m}{\partial y}\right)^2 Sy = \frac{\sum w_i \sum w_i (y_i - Y)^2}{(n-2)\left(\sum w_i \sum w_i x_i^2 - (\sum w_i x_i)^2\right)}, \tag{12}$$

$$Sb = \left(\frac{\partial b}{\partial y}\right)^2 Sy = \frac{\sum w_i x_i \sum w_i (y_i - Y)^2}{(n-2)\left(\sum w_i \sum w_i x_i^2 - (\sum w_i x_i)^2\right)}, \tag{13}$$

Presenting this variances in matrix form, the variances in the elements of $\mathbb{A}$ are the corresponding elements of the inverse correlation matrix $\mathbb{C}^{-1}$

$$S\mathbb{A}_{jj} = \left[\mathbb{C}_{jj}^{-1}\right] Sy, \tag{14}$$

where Sy is defined by

$$Sy = \frac{(\sum (y_i - mx_i - b)^2)}{(n-2)}. \tag{15}$$

4. The experiment

We have used the previous four cases to analyze the data set from an experiment, made to propose a new method, in order to determine the *dead time*, τ, for a Geiger-Müller nuclear measurement system [14], where, by virtue of the linear adjustment of the counting rate by radioactive sample unit mass and the couting rate, we can obtain the dead time by dividing the slope and the intercept. In Table I it is shown the experimental data set of the mean counting rate and the mean counting rate by sample mass unit.

We found the linear regression for the experimental data using the four weighted least square cases discussed in Section 3; where in case *i* it was assumed that there is not error in the experimental data set; in case *ii* was only considered error in the y_i experimental data; for the cases *iii* and *iv* it was assumed that the errors in the x_i and y_i experimental data can be not overlooked. The results are shown in the Table II, where besides values for m, b, and τ the Q value is shown. The smallest value was obtained using the fourth case, where it was necessary to have an initial m value which was selected from case *i*. We repeated the fourth case once again but the Q value was the same and this is not shown in the table.

Mean counting rate (x_i)	Mean counting rate by mass unit (y_i)
14598. ± 123.	27912. ± 240.
13695. ± 102.	28006. ± 215.
13394. ± 226.	28020. ± 475.
11313. ± 71.	28211. ± 187.
8910. ± 47.	28375. ± 238.
5952. ± 51.	28615. ± 207.

TABLE I. Experimental data for the method used to get the Geiger-Müeller dead time.

Case	m	b	Q	τ [μsec]
i)	−0.0801	29096.2422	817.1462	165.1760
ii)	−0.0795	29093.1576	0.0189	163.9561
iii)	−0.0974	29090.8176	0.0161	163.7630
iv)	−0.0795	29091.3079	0.0145	163.9665

TABLE II. Results for the linear regression of the experimental data set from Table I, using all the weighted least squares methods.

5. Discussion and conclusions

If your objective is to determine exact values of some physical data, by means of the linear adjustment of your experimental data set, the standard least squares method is not always the most proper one.

As it was shown in the sample, you will have better data if you use the most proper case, which will be the one who fits better to the experimental data, take into account the experimental uncertainties and gives you the smallest value for Q.

Acknowledgments

Work partially supported by DGICSA-SEP, México under contract C88-01-0263, number DGICSA:880243.

References

1. B. Ostle, *Statistics in research.* Iowa State University Press (1965).
2. M.J. Moroney, *Facts form figures.* Penguin Books (1965).
3. C. Mack, *Essentials of statistics for scientists and technologist.* Plenum Press (1975).
4. A. Picot, *Am. J. Phys.* **48** (1980) 302.
5. H.R. Bacon, *Am. J. Phys.* **21** (1953) 428.
6. J.M. Pasachoff, *Am. J. Phys.* **48** (1980) 800.
7. J. Orear, *Am. J. Phys.* **50** (1982) 912.
8. A.W. Ross, *Am. J. Phys.* **58** (1980) 409.
9. S.L. Meyer, *Data analysis for scientists and engineers.* Wiley (1975) p. 75.
10. K.S. Krane and L. Schecter, *Am. J. Phys.* **50** (1982) 82.
11. D.R. Barker and L.M. Diana, *Am. J. Phys.* **42** (1974) 224.

12. H.R. Vega C., *Introducción al método científico experimental.* Ediciones SPAUAZ (1988).
13. D.C. Baird, *Experimentation: an introduction to measurements theory and experimental.* Prentice Hall (1962), pp. 186–188.
14. H.R. Vega C. *Rev. Soc. Quim. Mex.* **33** (1989) to be published.

Resumen. Se discute el problema de ajustar un conjunto de datos experimentales a una línea recta, cuando las incertidumbres experimentales no pueden ser despreciadas. Se señalan las dificultades que se tienen al usar el método estándar de la regresión lineal mediante los mínimos-cuadrados. Un procedimiento alternativo basado en los mínimos cuadrados ponderados, donde el factor de ponderación depende de las incertidumbres experimentales, es presentado para cuatro casos diferentes; estos casos son utilizados para analizar los datos de un experimento en particular. En esta aplicación se demuestra cómo se obtienen mejores resulados al utilizar el caso adecuado.