JOURNAL OF CHEMICAL EDUCATION

pubs.acs.org/jchemeduc

Article

Misuse of Linear Regression Technique in Analytical Chemistry?

Manuel Aboal-Somoza* and Rosa M. Crujeiras

Cite This: *J. Chem. Educ.* 2024, 101, 1062−1070

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Nowadays, frontiers among different sciences are revealed as diffuse, and as a consequence, research must necessarily be faced from an interdisciplinary approach. Similarly, teaching certain topics in Chemistry requires the consideration of developments in other sciences (Mathematics, Biology, Physics, etc.). For instance, the estimation of the parameters of calibration lines in Analytical Chemistry (via linear regression) exemplifies this mandatory interaction since the comprehension of the regression tools should also consider how estimation is regarded from a statistical perspective. This work focuses on how to overcome the contradictions that undergraduate Chemistry students may encounter between the chemists' and statisticians' perspectives, when they are lectured on least-squares linear regression. A mixed analytical chemistry−statistics approach is proposed to present a method to get over the discordant issues between both scientific viewpoints.

KEYWORDS: *Calibration, Linear Regression, Least-Squares Method, Statistics, Analytical Chemistry*

INTRODUCTION

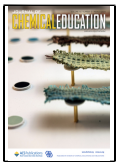
In Analytical Chemistry, the term "calibration" refers to the operation that determines the relationship between the signal values measured (in an analytical instrument) and the amount of analyte.[1−3] Indeed, determining the calibration function, which relates the expected value of the signal to the analyte amount,[4] as well as the close observation of the graphical display of such a function jointly (i.e., in the same graph) with the experimental data, are nowadays ordinary tasks in every analytical laboratory, and are mandatory for valid, quantitative results (for example, analyte concentration in samples) to be obtained. The requirement for this "methodological" or "analytical" calibration (colloquially, "calibration"), arises from the fact that every quantitative result obtained from an analytical instrument is based on a comparison between the signal recorded on the sample and the signals recorded on several standard solutions. This is a consequence of the current lack of absolute methods for routine instrumental analysis.

Certainly, the knowledge of the calibration process in Analytical Chemistry has a major importance in the chemistry curriculum, as proved by its inclusion in the Anchoring Concepts Content Maps (ACCM), recently published for Analytical Chemistry by the Division of Chemical Education of the American Chemical Society.[5] Similarly, other topics and sciences are usually included in chemistry degree curricula: chemists are expected to work in interdisciplinary environments requiring (at least basic) knowledge of other sciences. Particularly, Statistics has been included in chemistry degree curricula for years[6−9] not only to help "traditional, analogic" data handling but also as an essential support to assist the analysis of the huge amounts of data resulting from the development of analytical instrumentation and computer hardware and software[7] (i.e., the part of Analytical Chemistry called "Chemometrics"[10]). Moreover, a recent survey conducted by Kovarik et al.[9] revealed that for the more than three hundred undergraduate Analytical Chemistry instructors polled, the top ten (out of 38) course topics sorted for importance included "Standardization/Calibration methods", "Data handling", and "Statistical analysis" (in positions 1, 3, and 7, respectively). Indeed, these views show the relevance that Statistics has for the training of future chemists and confirm the statements in this *Journal* almost 60 years ago by Wentworth[11] and last month by McCluskey[12] about the necessity of giving [chemistry] students an adequate background in statistical analysis. Unfortunately, such an importance seems not to be noticed at the secondary school level and, as a result, students perceive Statistics as an ability to count and use formulas with an appropriate software, and consequently misconceptions arise, as described lately by J.M. Sanchez.[13,14]

In analytical chemistry, (straight) calibration lines are often calculated using the least-squares method (usually known as "linear least-squares"). Bearing in mind the interdisciplinary context mentioned above, this work deals with a proposal for explaining the least-squares method from a combined analytical chemistry−statistics point of view. It should be noted that the authors' aim is purely didactical, without seeking for a formal statistical approach to calibration (which can be found in the literature, for example in the book by E. Mullins[15]), keeping in mind that students may confront calibration tasks with a quite scarce statistical knowledge.

Line fitting has been extensively studied for many years. For example, a literature search in the Web of Science database on the topics "curve fitting" or "least-squares" produces more than 250,000 articles to date. Among those works, only 36 deal with

Received: October 9, 2023
Revised: February 8, 2024
Accepted: February 12, 2024
Published: February 29, 2024



https://doi.org/10.1021/acs.jchemed.3c01042
J. Chem. Educ. 2024, 101, 1062−1070

the topic "linear least-squares" and have been published in this *Journal*, but to the best of the authors' knowledge, none of them address the least-squares method from the perspective described in this work. Moreover, to name a few, the use of graphic aids for least-squares,[16,17] the proper application of weighting factors when needed,[18−20] the comparison with other fitting methods (the least absolutes method or the median method),[23] or the implementation of variations of the method to improve the quality of the fitting[11,21,22] are some of the matters investigated in those works. In addition, readers interested in the least-squares method can be referred to the works published some years ago by Kim and Kim,[23] and by Raposo,[24] whose tutorial review covers a good deal of ground about the use of the least-squares method in instrumental analysis techniques.

Besides, and without being exhaustive, many works have been published in this *Journal* dealing either with the application of the least-squares method to nonlinear (i.e., not aligned on a straight line) data[11,25−32] or with the use of computer programs and packages. Regarding this latter topic, although decades ago some authors proposed the use of programs written in BASIC (Beginners' All-purpose Symbolic Instruction Code)[26,33−35] or FORTRAN[36] languages, nowadays the current widespread fad is to use Excel spreadsheets (by Microsoft),[22,23,27,29,32,35,37−42] whereas the use of other statistical software such as BMPD,[25] Mathcad,[43] or KaleidaGraph[28,30,32] is rather scarce. Also, in recent years, open-source software projects such as JAMOVI,[44] JASP (supported by the University of Amsterdam),[45] or R[46] seem to be more and more accepted and used by the scientific community.

It is pertinent to mention that the authors of the published works regularly express their concern about the fact that the drawbacks of the least-squares method are frequently overlooked in many chemical courses and research papers,[21] perhaps because some details about the method are poorly understood,[19] particularly among chemists (as commented by Chong[47] some years ago). The obvious consequence of such a misunderstanding of the basics of the method is its indiscriminate, error-prone application. This topic will be dealt with later in this work.

Historical Background

To contextualize the procedure often followed to carry out calibration, a chronological look on how calibration has been carried out since the dawn of instrumental analysis (ca. second quarter of 20th century) shall be helpful.

Calibration was initially performed through the measurement of the signal produced by just two standard solutions containing known analyte concentrations (recall that a standard solution is commonly defined as a solution prepared for calibration that contains a known amount of analyte[48]). Both standard solutions were prepared based on previous knowledge about the approximate concentration of the analyte in the sample(s) to be analyzed. Therefore, one standard solution contained an analyte concentration lower than that expected in the sample, whereas the other standard solution contained a higher analyte concentration than that expected in the sample. The graphical depiction of the data (signal vs analyte concentration) and the straight line that linked both points, would eventually require an interpolation on that line to calculate the concentration of analyte in the sample. Note that the calculations involving dilutions, which are occasionally needed to figure the concentration of the analyte in the sample, are assumed in this and the following discussions. Sometimes this graphic-based procedure is named "graphical solution".

When the approximate content of analyte was unknown, or when different levels of analyte concentration were expected in the sample(s) to be analyzed, the obvious and reasonable approach was adopted: instead of just two standard solutions, the preparation of several standard solutions containing different analyte levels was carried out, for a wide-enough range of concentrations to be covered. Then, as in the two-standard solutions problem, the concentration of the samples was obtained by interpolation in the graphical display (signal vs analyte concentration), as depicted in Figure 1.

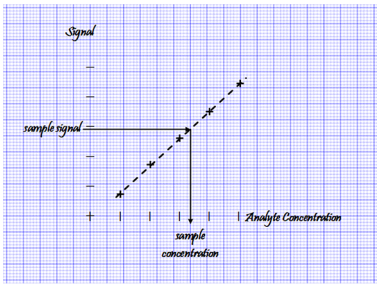


Figure 1. Graphical display of the calibration curve (a straight line) including the graphical procedure to obtain the value of concentration from a given value of signal.

This eyeball-and-ruler,[49] graphic-based calibration procedure involved the drawing of the best straight line that passed over as many points as possible (or, at least, as close as possible to those points)[50] and, therefore, presented the drawback of being highly dependent on both the drawing skills and the visual capacity of the analyst. In addition, in cases when the same graph was used to obtain the concentration of many samples, the paper would soon become damaged due to the number of lines traced on it. And this would probably hinder the achievement of valid results for concentration.

The problems exposed above made chemists focus on Mathematics and Statistics, looking for any procedure that allowed the parameters defining the calibration line to be obtained and to avoid the graphic-based calibration. Thus, the regression techniques used long ago in Statistics were the solution, not only for eluding the evident limitations of the graphical procedure described above but also because such techniques certainly enabled the validation of the calibration process. Among the different regression methodologies available, the method of (classical) least-squares for linear regression was adopted by the Analytical Chemistry community, and nowadays it tends to be the method of choice for estimating the parameters of the calibration curve.[2] This widespread use of least-squares estimation can be justified by its availability in daily use devices and tools such as calculators and standard spreadsheets. Moreover, although, strictly speaking, what analytical chemists usually calculate are straight lines by least-squares regression (first-order linear regression), it is quite common to refer to such calculi (even in the literature) as a "linear regression" or "regression line". This work, though applicable to general (possibly nonlinear) least-squares regression, is focused on the simplest linear case, the most usual in practice.

As a historical comment, a great controversy raised at the beginning of the 19th century between two mathematicians

https://doi.org/10.1021/acs.jchemed.3c01042
J. Chem. Educ. 2024, 101, 1062−1070

(A.M. Legendre and C.F. Gauss) regarding the authorship of the least-squares method. According to the literature, where further information about this dispute can be found,[51,52] though both authors probably independently discovered the method, Legendre is recognized as the first who published the method (in 1805[53]), whereas Gauss is considered the first who used it (from 1795, as claimed by Gauss himself in 1809[54]). Nevertheless, the invention of the method is usually attributed to Gauss.

Teaching Least-Squares Method to Undergraduate Chemistry Students

As it is well-known, the least-squares method is designed to minimize the (squared) differences between the observed values of the response and the ones predicted by the fitted model. In order to explain this idea, the lecturer may easily find two or three approaches to illustrate the least-squares method to undergraduate Chemistry students. On the one hand, the least-squares method could be described from a purely statistical point of view, disregarding its application to the calibration process and to the achievement of quantitative results in chemical analysis. In fact, the authors do not know any lecturer who uses this strategy on undergraduate Chemistry students.

A second possible path involves the explanation of the basis of the method, the resulting equations for the slope, the intercept, and the correlation and determination coefficients, and how the method can be validated as well as how all these calculations can be made on a calculator, a tablet, or a PC. Probably this genuinely "chemical, application-based" route is the most common one, focusing on the use of a series of equations to achieve the desired calibration line and presumably without much or any remark at all on the statistical background of the method.

Finally, an alternative methodology involves taking into account both the statistical basis and the chemical application of least-squares estimation. But this approach uncovers, in principle, remarkable discrepancies between the chemical use and the statistical rigor, particularly regarding the theoretical foundations of the method, as it is going to be commented below.

First, regression models are devised to model the relation between a response (or dependent) variable Y and an explanatory variable X (or many of them) trying to answer the question of which is the expected value of Y for a given value of X. The simplest form of such a relation is given by linear regression, which can be written as eq 1 (model equation), where β_0 and β_1 are the parameters of the model (respectively, the intercept and slope of the true—but unknown—regression line) and ε represents an error term, which tries to capture the random difference between the response variable and the linear model depending on X.

$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon \tag{1}$$

This error term is not observed (and in fact would be difficult to discover since it changes for each observation of Y[55]), but it can be approximated in practice by the difference between the observed values of the response and the ones predicted by the fitted line for the corresponding (observed) values of X (the so-called "residuals"). Therefore, eq 1 represents the (theoretical) linear regression model of Y over X.

In instrumental techniques, analytical chemists estimate this model by preparing a series of standard solutions each containing a known concentration of analyte (variable X) and by recording the signal (variable Y) for each standard solution. The number of standard solutions must be enough to define the response profile and, therefore, a minimum n value of 5−6 calibration standards is usually recommended.[24,56] Observe that, although evenly distributed or equidistant levels of concentration across the calibration range are the ideal option,[24,57] the published literature shows that, in practice, for wide calibration ranges and, most importantly, depending on the particular instrumental technique used (as, for example, happens in ICP-OES), partial arithmetic series (where the concentrations of the upper standards differ by a constant amount, not by a constant factor), are the option of choice.

The recommended next step is the visual inspection, whenever possible, of the scatter plot of the experimental data obtained,[58] before carrying out regression calculations, in order to detect outliers or points of influence[59] or curvature[60] in the data, because that visual inspection allows to check whether a straight line relationship (between both variables) is realistic for the experiment in question.[61] In addition, there are standard techniques based on residual and leverage analysis for identifying outliers and influential points (recall that a residual is the difference between the observed value of the response and the predicted value, and leverage is the weight that an observation has on its own prediction).[62] This visual examination is crucial because only if the experimental data are really consistent with a linear model will the adoption of such a model eventually make sense. Otherwise, the linear model must be discarded, of course.

It is desirable, when facing a calibration procedure, to adopt a flexible attitude because, as described in the previous paragraphs, for several issues concerning calibration there are no general, universal guidelines and sometimes it is a matter of degree. For example, in view of a scatter plot, one can see if data are close to a straight line, but the problem is to state how close or whether it is close enough.[56]

Finally, with the resulting n pairs of data (x_i, y_i), b_0 and b_1 (the corresponding estimates of β_0 and β_1, respectively) are calculated, and the equation of the calibration line (usually, a straight line) is obtained. This is called the regression line (eq 2).

$$Y = b_0 + b_1 \cdot X \tag{2}$$

Note that b_0 and b_1 are obtained with eqs 3 and 4 (respectively), that have been widely discussed (among other relevant topics on calibration and regression) in the literature from long ago[1,50,63−65] and are currently included in calculators, spreadsheets and other devices. In addition, the formula to calculate the determination coefficient r^2 (a key parameter to check the suitability for the model to fit experimental data), eq 5, is also included as it will be mentioned later.

$$b_0 = \bar{y} - b_1 \cdot \bar{x} \tag{3}$$

$$b_1 = \frac{\sum_{i=1}^{n} [(x_i - \bar{x})(y_i - \bar{y})]}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{4}$$

$$r^2 = \frac{\left(\sum_{i=1}^{n} [(x_i - \bar{x})(y_i - \bar{y})]\right)^2}{\left(\sum_{i=1}^{n} (x_i - \bar{x})^2\right)\left(\sum_{i=1}^{n} (y_i - \bar{y})^2\right)} \tag{5}$$

(n is the number of pairs of experimental data used for calculations, and, as it is well-known, the correlation

https://doi.org/10.1021/acs.jchemed.3c01042
J. Chem. Educ. 2024, 101, 1062−1070

coefficient, r, is the square root of the determination coefficient). Of course, in the literature, other equations to calculate b_0, b_1 and r^2 are available,[22] but indeed such expressions are equivalent to the ones above (eqs 3−5). Also, it is worth showing eqs 3−5 here to underline the simplicity of the calculi needed to obtain the estimates (just additions, subtractions and multiplications), which allows even its manual calculation!

Afterward, the signal(s) for the sample(s) is/are recorded, and (in the simplest case study) by replacing (in eq 2) "Y" for each signal, the corresponding values of concentration "X" can be calculated. Therefore, in analytical laboratories, the final aim of the regression analysis is then to use the mathematical expression (which relates the signal and the concentration) to predict the concentration of unknown samples,[66] i.e., to predict values for X starting from Y values (a procedure known, at times, as "inverse regression",[60,67] which may be a confusing term, since other authors use it as the regression of X over Y[68,69]). However, every statistician would absolutely state that this use of a regression model is completely wrong because the model described in eq 1 (whose corresponding estimation is presented in eq 2), can only be used for predicting values for the response variable ("Y") from values of the explanatory or predictive variable ("X"). Furthermore, by any means, this model could never be applied to predict X values on the basis of Y values. It should be noted that the least-squares estimation of β_1 (namely, b_1 in eq 2) is obtained considering the covariance between the response and the covariate divided by the sample variance of the covariate, that is, the sample variance of X (eq 4).

Therefore, the disagreement between the Analytical Chemistry and the Statistics points of view is really evident: analytical chemists use the regression model for the only thing that, according to statisticians, such a model cannot be used at all! Then, one could wonder which of both perspectives (if any) is the correct one or whether this controversy means that instrumental analysis, which is mainly based on regression lines, is baseless... The authors think that, luckily, there undoubtedly is a meeting place for both positions. Moreover, the bidisciplinary (Analytical Chemistry−Statistics) standpoint is the only valid one for the students to actually understand the role played by each discipline in this topic.

Bidisciplinary Approach to the Least-Squares Method

As an example, for the following discussion, the real experimental data in Table 1 will be used. Those data are the concentrations of Ti of a series of standard solutions and their corresponding Ti emission intensities (expressed in cps, i.e., counts per second) at 334.945 nm, obtained by ICP-OES (Inductively Coupled Plama-Optical Emission Spectrometry). This calibration set was prepared to determine Ti in water samples, without dilution. For three of the water samples analyzed, the emission intensities recorded were 4190.691 cps (sample 1), 5562.674 cps (sample 2), and 5802.361 cps (sample 3). Note that in the discussions described in the following paragraphs, some aspects related to quantitative analysis (such as limit of detection, limit of quantification, linear range, etc.) have been omitted, since they are tangential to the point of this work. In addition, as is usual in instrumental analysis, the signal recorded for the blank was subtracted from itself (this explains the value "0.000" in Table 1) and from each emission intensity recorded for standard solutions and samples, and the working range was reduced to the interval 0.0−94.9 μg/L (in order to use evenly distributed concentration levels along the working range), although the whole raw data are included in the Supporting Information.

Table 1. Data for a Calibration Curve for the Determination of Ti in Water Samples by ICP-OES[a]

Concentration of Ti (X, μg/L)	Emission intensity (Y, cps)
0.0	0.000
9.5	3114.369
23.7	10,892.702
47.5	23,051.704
74.3	34,534.168
94.9	44,575.015

[a]Emission intensities recorded at a wavelength of 334.945 nm.

Applying the least-squares method, according to the Analytical Chemistry practice, to obtain the equation for the calibration line, eq 6 is obtained with a determination coefficient (r^2) of 0.9987. Y represents the emission intensity (in cps, i.e., counts per second) and X the concentration of Ti (in μg/L).

$$Y = -439.276 + 475.405 \cdot X \quad (6)$$

Then, going on with Analytical Chemistry habit, the Ti concentration in samples 1−3 would be, respectively, 9.7, 12.6, and 13.1 μg/L (these results are calculated by placing each emission intensity instead of "Y" in eq 6 and obtaining the respective value for "X"). Therefore, the regression eq 6 has been used to predict values for X (against, recall, the Statistics theory).

So, at this point, and taking into account all of the exposed above, there are some important items that must be highlighted:

i) There are at least two agreements between Analytical Chemistry and Statistics: first, regarding the distribution of random error, both sciences consider that the random error of the explanatory variable (X) is negligible with respect to the random error of the response variable (Y). Actually, most statistical regression analysis are carried out assuming that the values of X are fixed by the practitioner. Second, conceptually, the regression of Y (response variable) over X (explanatory variable) is not the same as the regression of X over Y, although, as commented by de Julián-Ortiz et al.,[21] some researchers still think that both regressions are the same.

ii) Although there are some techniques and particular case studies where the line that best represents the relation between two variables is not a straight line (and for those situations, modified least-squares procedures are available), certainly (analytical) chemists prefer a straight line over other curves.[70] Besides the availability in calculators and other devices (already mentioned above), there are other reasons that explain this choice:[71] on the one hand, there is usually theoretical basis that justifies the proportionality between both variables (Beer−Lambert law,[72] which relates the absorption of radiation with the concentration of the absorbent species, and is the basis of quantitative analysis by UV−Vis Spectrophotometry, is a good example of this). On the other hand, the equation that represents a curved calibration line includes second and/or higher degree exponents, that cause difficulties for drawing the line, handling the equation, and deriving

https://doi.org/10.1021/acs.jchemed.3c01042
J. Chem. Educ. 2024, 101, 1062−1070

conclusions from it (for example, when a given value of one of the variables corresponds to more than one value of the other variable). Additionally, in instrumental analysis, a good alignment of the experimental data of a calibration is usually obtained, and so, a straight line generally fits well to data. This lining up also happens in the data of Table 1, which are plotted in Figure 2.

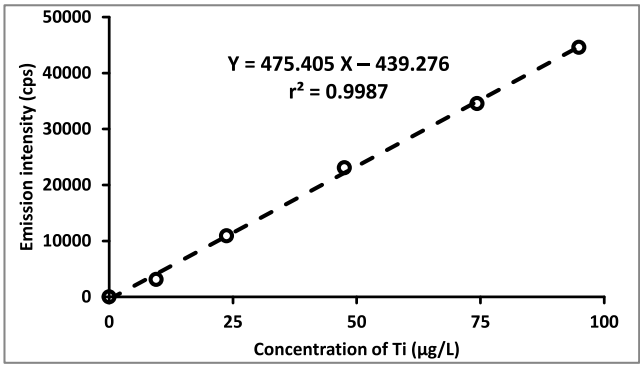


Figure 2. Plot of data in Table 1 and the linear regression (least-squares method) line obtained with those data, including the equation corresponding to that line (in the text, eq 6). "Y" is the emission intensity, and "X" the concentration of Ti.

iii) According to the previous point, on condition that the experimental data are aligned in a straight line, the equation of the regression line of Y over X is in practice the same as that of the regression line of X over Y): in our example (data in Table 1), the regression line of X over Y would be that of eq 7 (where X is assumed to be the response variable and Y the explanatory variable).

$$X = 0.977353 + 2.1007 \cdot 10^{-3} \cdot Y \quad (7)$$

When the terms in eq 7 are reordered, eq 8 is obtained:

$$Y = -465.251 + 476.032 \cdot X \quad (8)$$

A comparison of eqs 6 and 8 leads us to conclude that the slopes of both equations can, in fact, be considered the same and so can both equations (as can be seen in Figure 3), despite the difference observed between both intercepts, just required for shifting the scale. Moreover, although both slopes are not *exactly* the same (475.405 and 476.032 L/μg), they are quite similar, given that the product of both slopes equals the coefficient of determination r^2 (see Supporting Information for a more detailed explanation). Therefore, if both regression lines were exactly the same, the product of both slopes would be 1, but here, that product is not "1", but 0.9987, which is indeed close enough in practice.

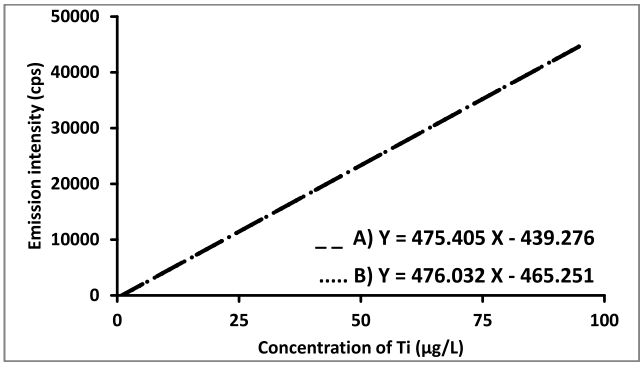


Figure 3. Plot of the two regression lines (fitted by least-squares) obtained with the data in Table 2. (A) Regression of Y over X (eq 6 in the text). (B) Regression of X over Y (eq 8 in the text).

iv) Furthermore, since in Analytical Chemistry the objective of the use of regression lines is to predict concentration values (i.e., values for variable "X"), Table 2 includes the concentration values calculated with both calibration curves (summarized in eqs 6 and 8). It must be recalled and emphasized that these signal data have been taken just as examples to show the concentrations that are obtained with both Equations, setting aside other features (precision of results, for example) that, though important, are absolutely peripheral in the context of this work.

Table 2. Predicted Concentrations of Ti in Water Samples, using eqs 6 and 8

		Concentration of Ti (X, μg/L)[a]		
Sample	Emission intensity (Y, cps)	eq 6	eq 8	Relative error (%)[b]
1	4190.691	9.7	9.8	1.0
2	5562.674	12.6	12.7	0.8
3	5802.361	13.1	13.2	0.8

[a]Calculated using eq 6 or eq 8. [b]Calculated taking the value calculated with eq 6 as the true value.

The values calculated with both equations are quite similar (relative errors about 1%, far lower than 10%), which proves that for the purposes of Analytical Chemistry work, it would be unimportant to use any of the two regression lines (Y over X or X over Y). Also, from the statistical point of view, as it has been previously mentioned, when describing eq 6, the slope of the regression line of Y over X is given by the covariance between the two variables divided by the sample variance of X. In an analogous way, the slope for the regression line of X over Y (eq 7) is given by the covariance divided by the sample variance of Y. Hence, in eq 8, the slope is given by the inverse of this quotient (i.e., the variance of Y divided by the covariance between X and Y). Is there any general setting where the slope in eq 6 is equal (or very similar) to the slope in eq 8? Just checking how the two slopes are obtained, it is easy to see that these two quantities match when the coefficient of determination is equal (or very close) to one. Equivalently, eq 6 and eq 8 have the same slope when the correlation between X and Y is close to 1 or −1. This is the case in our example, where for both lines (eq 6 and eq 8) the correlation coefficient is equal to 0.9993 (and the corresponding coefficient of determination is 0.9987).

v) To illustrate the previous point, Figure 4 shows how the slope in eq 8 (obtained as the inverse of the slope in eq 7) approaches the slope in eq 6 (which is fixed to 1 in this simulated example), when the correlation coefficient

https://doi.org/10.1021/acs.jchemed.3c01042
J. Chem. Educ. 2024, 101, 1062−1070

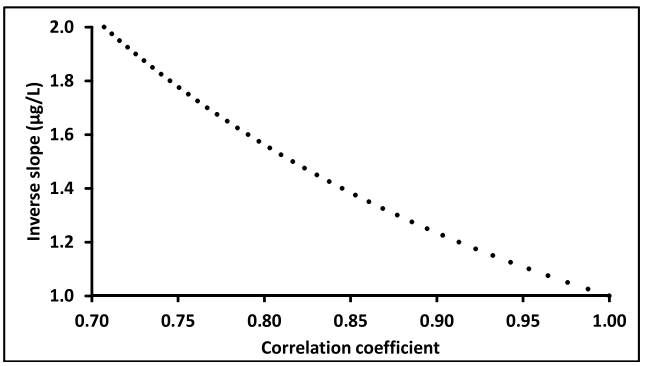


Figure 4. Plot of the change on the value of the inverse slope in eq 7 with the correlation coefficient.

between the response and the explanatory variables varies from 0.70 until 1. For instance, for r = 0.95, the inverse slope is 1.1 (0.1 higher than the real slope). This means that the more r approaches 1, the more both lines tend to become the same line.

EVALUATION OF THE PROPOSAL

In order to evaluate the adequacy of the proposal, the authors approached other university instructors, with a consolidated experience, through a brief questionnaire, with the final aim of getting critical considerations for the adoption of the proposal presented in this paper and their willingness to change their teaching strategy. A focus group was gathered, consisting of 18 university instructors, mainly from the fields of knowledge of Analytical Chemistry and Statistics, but also from other related fields, such as Applied Physics. The proposed teaching methodology was introduced to all the members of the focus, and then they were asked the following questions:

1. *What do you think about this strategy to teach least-squares method to undergraduate students of chemistry or other applied sciences? (i.e., is it useful?, does it make sense?, is it too complex?, etc.)*
2. *According to your own experience, do you think this approach would improve the learning process of the least-squares method by students?*
3. *Would you adopt this methodology to teach least-squares method?*
4. *Any additional comment/suggestion?*

About question 1, only two members of the focus group considered the strategy too complicated for first-year undergraduate students to understand, due to the lack of previous knowledge about Mathematics, Statistics, and/or Analytical Chemistry. Some instructors also pointed out the interest of this strategy because it fosters the necessary coordination between the lecturers of different courses and fields of knowledge that will eventually benefit the students.

Regarding question 2, most of the members of the focus group consider that this methodology poses some advantages with respect to the traditional option. Some members, as for question 1, insisted on the difficulties for first-year undergraduate students to follow the different points of view of Analytical Chemistry and Statistics, although others stated that being aware of such different views is good for students to grasp a global, unifying knowledge of the method. Moreover, as suggested by one lecturer, the stress should be put on the meeting points between Analytical Chemistry and Statistics rather than on their differences.

More than 60% of the members of the focus group would (and some of them "will" indeed) adopt the proposed methodology, and only one would not use it at all. There is agreement in the fact that the approach proposed here is more likely to be useful in the field of Analytical Chemistry than in Statistics courses, where the study of the least-squares method is more "mathematical" (as, of course, it is expected to be). Again, some comments by the members reflect the idea, very common in science, that sometimes the good choice is not the best or the most correct one. That is, from a practical perspective (and making science require large doses of practicality at times), what is not perfect (i.e., theoretically correct) can be good enough for a given situation (i.e., can help to obtain acceptable results). This topic is also interesting for students to know about.

The answers to the last question deal with a great variety of issues. Some comments underline the different uncertainties of both variables: whereas the concentration ("X") has a negligible random error, the signal ("Y") presents a given uncertainty. According to this, the only regression that really makes sense is Y over X (and not the regression of X over Y). Others indicate that it is essential to draw the regression line in the scatter plot of the experimental data, in order to have an immediate visual perception of the agreement between the model (straight line) and the data (of course, these graphs are complementary to other procedures aimed at checking the suitability of the model, such as residual analysis, lack of fit tests, homoscedasticity checks, etc.)[13] Finally, other suggestions aim at certain details concerning the methodology proposed, like the need for using real sets of data to teach the least-squares method or the necessity of holding meetings with instructors of the various courses involved in least-squares regression, in order to share experiences and coordinate the contents and topics covered. Certainly, the latter suggestion is mandatory if good learning results are to be achieved.

The authors are aware that another route for evaluation of the adequacy of the teaching proposal would consist of taking two groups of students and trying in one of them the "usual" method (from the "purely chemical" point of view), considering this new approach in the other group (taking this mixed chemical−statistical perspective), and afterward evaluating the learning outcomes. However, such an evaluation strategy presents some serious drawbacks. On the one hand, it is quite usual that most of the students of both groups (especially for those in their freshman or even sophomore years) have never heard or read about the least-squares method. As far as the authors' teaching experience is concerned, this means that those students lack capacity or basic knowledge to analyze critically what is being explained to them, and therefore, they are likely to just accept what is being exposed without questioning it. Only those (a minority of students, in the authors' experience) who have known about least-squares method previously can compare the new explanation with their prior background about the method. On the other hand, the authors consider it unethical in this case to take the students as kind of guinea pigs to experiment with them the results of different teaching strategies, first because the authors consider that the traditional approach used to teach least-squares method is not as correct as desirable (and thus, a deontological issue arises) and, second, because the subject (i.e., the least-squares method) is more complex

https://doi.org/10.1021/acs.jchemed.3c01042
J. Chem. Educ. 2024, 101, 1062−1070

than, for example, solving a mathematical or chemical problem. In fact, to describe different strategies to solve a mathematical problem is quite interesting and useful because, bearing in mind that the objective in such a situation is to help students understand what has to be calculated and why, some students may find a certain procedure easier to understand and follow than other procedures, but the least-squares method is much more complex than that (and to explain least-squares method by both approaches would positively produce nothing but confusion and misunderstanding in the students). In agreement with the statement by Lötz[73] and others[49] decades ago, authors think it is wiser (especially for first-year undergraduate students), to center the attention on consolidating basic issues rather than trying to harass students with theory as well as avoiding the oversimplification that the "purely chemical" perspective described above represents. Consequently, the proposed approach was introduced to other university instructors for them to consider its adoption rather than evaluate the suitability of the proposal on students.

CONCLUSIONS

From the exposed above, it can be concluded that in the field of Analytical Chemistry, Statistics is a tool that, of course, needs to be used appropriately. However, analytical chemists essentially apply the least-squares method just as a means to procure an equation that represents the relationship between two variables (i.e., usually, the instrumental signal and the concentration of analyte), without considering at all the statistical meaning of regression. This indicates that they [analytical chemists] oversee statistical rigor in favor of simplicity. Indeed, this is the key point of this bidisciplinary approach.

Therefore, the authors' proposal is that the least-squares method should be taught to undergraduate chemistry students in view of both the statistical meaning of regression and the tool-characteristic that such a regression method has for analytical chemists.

In addition, the methodology proposed in this work emphasizes the fact that nowadays the boundaries among the different fields of science are really diffuse and even fictitious at times. This means that future graduates in chemistry and other sciences must be prepared to work in a multidisciplinary scientific environment, and therefore, using this type of approach by the teaching staff can be eventually considered as a duty and integrity issue.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available at https://pubs.acs.org/doi/10.1021/acs.jchemed.3c01042.

Raw data obtained from the ICP-OES spectrometer, concentration predicted for Ti in samples, and relationship between the slopes of the regression lines of Y over X and X over Y (PDF, DOCX)

AUTHOR INFORMATION

Corresponding Author

Manuel Aboal-Somoza − *Group of Trace Elements, Speciation and Spectroscopy (GETEE) − Institute for Materials of the University of Santiago de Compostela (iMATUS), Department of Analytical Chemistry, Nutrition and Bromatology, Faculty of Chemistry, Universidade de Santiago de Compostela, E-15782 Santiago de Compostela, A Coruña, Spain*; orcid.org/0000-0002-0584-6289; Email: m.aboal@usc.es

Author

Rosa M. Crujeiras − *Galician Center for Mathematical Research and Technology, CITMAga, Department of Statistics, Mathematical Analysis and Optimization, Faculty of Mathematics, Universidade de Santiago de Compostela, E-15782 Santiago de Compostela, A Coruña, Spain*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jchemed.3c01042

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work is a consequence of the questions formulated three years ago by our undergraduate Chemistry student Ms. Lourdes Patricia Sanmiguel Vázquez, whose inspiring interest is gratefully thanked. Work by Rosa M. Crujeiras has been supported by project PID2020-116587GB-I00, from the Agencia Estatal de Investigación. The collaboration of our colleagues, members of the focus group, and their interesting suggestions are deeply acknowledged, as well as the valuable comments received from the Associate Editor and the Reviewers of this Journal. The authors also express their gratitude to Prof. Dr. Rafael Cela for his comments and, last but not least, the invaluable support and help received for years from their colleague Prof. Dr. José María Alonso-Meijide are deeply acknowledged too.

REFERENCES

(1) Danzer, K.; Currie, Ll. A. Guidelines for Calibration in Analytical Chemistry. Part 1. Fundamentals and Single Component Calibration (IUPAC Recommendations 1998). *Pure Appl. Chem.* **1998**, *70* (4), 993−1014.

(2) Hibbert, D. B. *Compendium of Terminology in Analytical Chemistry*, 4th ed. (the "Orange Book"); The Royal Society of Chemistry-IUPAC: Croydon, 2023; p 34.

(3) Massart, D. L.; Vandeginste, B. G. M.; Buydens, L. M. C.; de Jong, S.; Lewi, P. J.; Smeyers-Verbeke, J. *Handbook of Chemometrics and Qualimetrics: Part A*; Elsevier: Amsterdam, 1997; p 171.

(4) IUPAC. Calibration function. In *Compendium of Chemical Terminology*, 2nd ed. (the "Gold Book", compiled by McNaught, A. D., Wilkinson, A.); Blackwell Scientific Publications: Oxford, 1997; online version (2019−) created by Chalk, S. J. (accessed December 2023) .

(5) Holme, T. A.; Bauer, C.; Trate, J. M.; Reed, J. J.; Raker, J. R.; Murphy, K. L. The American Chemical Society Exams Institute Undergraduate Chemistry Anchoring Concepts Content Map V: Analytical Chemistry. *J. Chem. Educ.* **2020**, *97*, 1530−1535.

(6) Salzer, R. Eurocurriculum II for analytical chemistry approbed by the Division of Analytical Chemistry of FECS. *Anal. Bioanal. Chem.* **2004**, *378* (1), 28−32.

(7) Schlotter, N. E. A statistics curriculum for the undergraduate chemistry major. *J. Chem. Educ.* **2013**, *90* (1), 51−55.

(8) Davidian, M.; Kutal, C. Collaboration to meet the statistical needs in the chemistry curriculum. *J. Chem. Educ.* **2014**, *91* (1), 12.

(9) Kovarik, M. L.; Galarreta, B. C.; Mahon, P. J.; McCurry, D. A.; Gerdon, A. E.; Collier, S. M.; Squires, M. E. Survey of the undergraduate analytical chemistry curriculum. *J. Chem. Educ.* **2022**, *99* (6), 2317−2326.

(10) IUPAC. Chemometrics. In *Compendium of Chemical Terminology*, 2nd ed. (the "Gold Book", compiled by McNaught, A. D., Wilkinson, A.); Blackwell Scientific Publications: Oxford, 1997; online version (2019−) created by Chalk, S. J. (accessed December 2023) .
(11) Wentworth, W. E. Rigorous least squares adjustment. Application to some non-linear equations, I. *J. Chem. Educ.* **1965**, *42* (2), 96−103.
(12) McCluskey, A. R. Is there still a place for linearization in the chemistry curriculum? *J. Chem. Educ.* **2023**, *100* (11), 4174−4176.
(13) Sanchez, J. M. Ordinary least squares with laboratory calibrations: a practical way to show students that this fitting model may easily yield biased results when used indiscriminately. *World J. Anal. Chem.* **2017**, *5* (1), 1−8.
(14) Sanchez, J. M. The need to reinforce the teaching of basic descriptive statistics required in reporting quantitative laboratory results: diagnose of common students' misconceptions. *J. Chem. Educ.* **2023**, *100* (7), 2713−2718.
(15) Mullins, E. *Statistics for the Quality Control Chemistry Laboratory*; Royal Society of Chemistry: Cambridge, 2003; pp 247−307.
(16) Christian, S. D. Graphical least squares analysis. *J. Chem. Educ.* **1965**, *42* (11), 604−607.
(17) Henderson, C. Lecture graphic aids for least-squares analysis. *J. Chem. Educ.* **1988**, *65* (11), 1001−1003.
(18) Sands, D. E. Weighting factors in least squares. *J. Chem. Educ.* **1974**, *51* (7), 473−474.
(19) Christian, S. D.; Lane, E. H.; Garland, F. Linear least-squares analysis. A caveat and a solution. *J. Chem. Educ.* **1974**, *51* (7), 475−476.
(20) de Levie, R. When, why, and how to use weighted least squares. *J. Chem. Educ.* **1986**, *63* (1), 10−15.
(21) de Julián-Ortiz, J. V.; Pogliani, L.; Besalú, E. Two-variable linear regression: modelling with orthogonal least-squares analysis. *J. Chem. Educ.* **2010**, *87* (9), 994−995.
(22) Patzer, A. B. C.; Bauer, H.; Chang, C.; Bolte, J.; Sülzle, D. Revisiting the scale-invariant, two-dimensional linear regression method. *J. Chem. Educ.* **2018**, *95* (6), 978−984.
(23) Kim, M.-H.; Kim, M. S. Interactive visual least absolutes method: comparison with the least squares and the median methods. *J. Chem. Educ.* **2016**, *93* (10), 1737−1743.
(24) Raposo, F. Evaluation of analytical calibration based on least-squares linear regression for instrumental techniques: a tutorial review. *Trends Anal. Chem.* **2016**, *77*, 167−185.
(25) Copeland, T. G. The use of non-linear least squares analysis. *J. Chem. Educ.* **1984**, *61* (9), 778−779.
(26) Kahley, M. J.; Novak, M. A practical procedure for determining rate constants in consecutive first-order systems. *J. Chem. Educ.* **1996**, *73* (4), 359−364.
(27) Harris, D. C. Nonlinear least-squares curve fitting with Microsoft Excel Solver. *J. Chem. Educ.* **1998**, *75* (1), 119−121.
(28) Tellinghuisen, J. Nonlinear least-squares using microcomputer data analysis programs: KaleidaGraph in the physical chemistry teaching laboratory. *J. Chem. Educ.* **2000**, *77* (9), 1233−1239.
(29) Barton, J. S. A comprehensive enzyme kinetic exercise for biochemistry. *J. Chem. Educ.* **2011**, *88* (9), 1336−1339.
(30) Tellinghuisen, J. Using least squares for error propagation. *J. Chem. Educ.* **2015**, *92* (5), 864−870.
(31) Perrin, C. L. Linear or nonlinear least-squares analysis of kinetic data? *J. Chem. Educ.* **2017**, *94* (6), 669−672.
(32) Tellinghuisen, J. Least-squares analysis of data with uncertainty in y and x: algorithms in Excel and KaleidaGraph. *J. Chem. Educ.* **2018**, *95* (6), 970−977.
(33) Christian, S. D.; Tucker, E. E. LINGEN-A general linear least squares program. *J. Chem. Educ.* **1984**, *61* (9), 788.
(34) O'Neill, R. T.; Flaspohler, D. C. Least-squares fitting of multilinear equations. *J. Chem. Educ.* **1990**, *67* (1), 40−42.
(35) 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* (4), A130−A131.
(36) Kim, H. Computer programming in physical chemistry laboratory. Least-squares analysis. *J. Chem. Educ.* **1970**, *47* (2), 120−122.
(37) de Levie, R. Estimating parameter precision in nonlinear least squares with Excel's Solver. *J. Chem. Educ.* **1999**, *76* (11), 1594−1598.
(38) Salter, C.; de Levie, R. Nonlinear fits of standard curves: a simple route to uncertainties in unknowns. *J. Chem. Educ.* **2002**, *79* (2), 268−270.
(39) Burnett, J.; Burns, W. A. Using a spreadsheet to fit experimental pH titration data to a theoretical expression: estimation of analyte concentration and K_a. *J. Chem. Educ.* **2006**, *83* (8), 1190−1193.
(40) de Levie, R. Nonisothermal analysis of solution kinetics by spreadsheet simulation. *J. Chem. Educ.* **2012**, *89* (1), 79−86.
(41) Dias, A. A.; Pinto, P. A.; Fraga, I.; Bezerra, R. M. F. Diagnosis of enzyme inhibition using Excel Solver: a combined dry and wet laboratory exercise. *J. Chem. Educ.* **2014**, *91* (7), 1017−1021.
(42) Evans, J. S. O.; Evans, I. R. Structure analysis from power diffraction data: Reitveld refinement in Excel. *J. Chem. Educ.* **2021**, *98* (2), 495−505.
(43) Zielinski, T. J.; Allendoerfer, R. D. Least squares fitting of nonlinear data in the undergraduate laboratory. *J. Chem. Educ.* **1997**, *74* (8), 1001−1007.
(44) *The Jamovi project*. https://www.jamovi.org/ (accessed December 2023).
(45) *JASP Team*. https://jasp-stats.org/ (accessed December 2023).
(46) *The R project for statistical computing*. https://www.r-project.org/ (accessed December 2023).
(47) Chong, D. P. On the use of least squares to fit data in linear form. *J. Chem. Educ.* **1994**, *71* (6), 489−490.
(48) IUPAC. Standard solution. *Compendium of Chemical Terminology*, 2nd ed. (the "Gold Book", compiled by McNaught, A. D., Wilkinson, A.); Blackwell Scientific Publications: Oxford, 1997; online version (2019−) created by Chalk, S. J. (accessed December 2023). .
(49) Pattengill, M. D.; Sands, D. E. Statistical significance of linear least-squares parameters. *J. Chem. Educ.* **1979**, *56* (4), 244−247.
(50) Hibbert, D. B. The uncertainty of a result from a linear calibration. *Analyst* **2006**, *131* (12), 1273−1278.
(51) Plackett, R. L. Studies in the history of probability and statistics. XXIX: the discovery of the method of least squares. *Biometrika* **1972**, *59* (2), 239−251.
(52) Harter, H. L. The method of least squares and dome alternatives-Part I. *Int. Stat. Rev.* **1974**, *42* (2), 147−174.
(53) Legendre, A. M. *Nouvelles Méthodes pour la Détermination des Orbites des Comètes*; Firmin-Didot: Paris, 1805.
(54) Gauss, C. F. *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientum (in English: Theory of the Motion of the Heavenly Bodies Moving about the Sun in Conic Sections)*; Frid. Perthes and I. H. Besser: Hamburg, 1809.
(55) Draper, N. R.; Smith, H. *Applied Regression Analysis*, 3rd ed.; John Wiley & Sons: New York, 1998; pp 22.
(56) Analytical Methods Committee of the Analytical Division of RSC, *AMCTB No. 3. Is my calibration linear?* https://www.rsc.org/ (accessed December 2023).
(57) Jurado, J. M.; Alcázar, A.; Muñiz-Valencia, R.; Ceballos-Magaña, S. G.; Raposo, F. Some practical considerations for linearity assessment of calibration curves as function of concentration levels according to the fitness-for-purpose approach. *Talanta* **2017**, *172*, 221−229.
(58) Massart, D. L.; Vandeginste, B. G. M.; Buydens, L. M. C.; de Jong, S.; Lewi, P. J.; Smeyers-Verbeke, J. *Handbook of Chemometrics and Qualimetrics: Part A*; Elsevier: Amsterdam, 1997; p 4.
(59) Ellison, S. L. R.; Barwick, V. J.; Farrant, T. J. D. *Practical Statistics for the Analytical Scientist. A Bench Guide*, 2nd ed.; RSC Publishing: Cambridge, 2009; p 93.
(60) Analytical Methods Committee of the Analytical Division of RSC. AMCTB No. 113. Avoiding some common mistakes in straight line regression. Part 1. *Anal. Methods* **2023**, *15*, 6105−6107.

https://doi.org/10.1021/acs.jchemed.3c01042
J. Chem. Educ. 2024, 101, 1062−1070

(61) Miller, J. N., Miller, J. C., Miller, R. D. *Statistics and Chemometrics for Analytical Chemistry*, 7th ed.; Pearson: Harlow, 2018; p 129.
(62) Faraway, J. J. *Linear Models with R*, 2nd ed.; Chapman & Hall/CRC: Boca Raton, 2015; pp 73−98.
(63) Miller, J. N. Statistical Methods for Analytical Chemistry. Part 2. Calibration and Regression Methods. *Analyst* **1991**, *116* (1), 3−14.
(64) Currie, Ll. A.; Svehla, G. Nomenclature for the Presentation of Results of Chemical Analysis (IUPAC Recommendations 1994). *Pure Appl. Chem.* **1994**, *66* (3), 595−608.
(65) Miller, J. N., Miller, J. C., Miller, R. D. *Statistics and Chemometrics for Analytical Chemistry*, 7th ed.; Pearson: Harlow, 2018; pp 120−164.
(66) Massart, D. L.; Vandeginste, B. G. M.; Buydens, L. M. C.; de Jong, S.; Lewi, P. J.; Smeyers-Verbeke, J. *Handbook of Chemometrics and Qualimetrics: Part A*; Elsevier: Amsterdam, 1997; p 171, 197.
(67) Analytical Methods Committee of the Analytical Division of RSC. *AMCTB No. 22. Uncertainties in concentrations estimated from calibration experiments*. https://www.rsc.org/ (accessed December 2023).
(68) Massart, D. L.; Vandeginste, B. G. M.; Buydens, L. M. C.; de Jong, S.; Lewi, P. J.; Smeyers-Verbeke, J. *Handbook of Chemometrics and Qualimetrics: Part A*; Elsevier: Amsterdam, 1997; p 207.
(69) Delgado, R. Misuse of Beer-Lambert law and other calibration curves. *R. Soc. Open. Sci.* **2022**, *9*, No. 211103.
(70) Johnson, D. C. Guest Editorial. *Anal. Chim. Acta* **1988**, *204*, 1−5.
(71) Boqué, R.; Rius, F. X. Profundizando en la Calibración Lineal Univariante. In *Avances en Quimiometría Práctica*; Cela, R., Coord.; University of Santiago de Compostela: Santiago de Compostela, 1994; pp 157−187.
(72) IUPAC. Beer−Lambert law (Beer−Lambert−Bouguer law). In *Compendium of Chemical Terminology*, 2nd ed. (the "Gold Book", compiled by McNaught, A. D., Wilkinson, A.); Blackwell Scientific Publications: Oxford, 1997; online version (2019−) created by Chalk, S. J. (accessed December 2023) .
(73) Lötz, A. Statistics by computer simulation. *J. Chem. Educ.* **1995**, *72* (2), 128−129.

https://doi.org/10.1021/acs.jchemed.3c01042
J. Chem. Educ. 2024, 101, 1062−1070