Linear regressions to which the standard formulas do not apply

Elena Mashkina¹ · Keith B. Oldham²

Received: 23 March 2015 / Accepted: 20 April 2015 / Published online: 22 May 2015
© Springer International Publishing 2015

Abstract Linear regression is, perhaps, the statistical technique most widely used by chemists. Regrettably, it is frequently misused. The many nuances in the procedure are commonly overlooked, leading to frequent misapplication of the traditional formulas. Here we discuss a number of alternatives and the circumstances under which each should be employed.

Keywords Linear regression · Best straight line · Least squares · Errors-in-variables model · Constrained regression

Introduction

Linear regression, otherwise known as ‘least squares’, or ‘putting the best straight line through the points’ is employed when a measured quantity y , derived from some measurement or observation, depends on the value of another quantity x , and one has a set of data in the form of n pairs of values $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. These data may be represented as n points in a Cartesian plane, as in Fig. 1. If it is believed, or speculated, that there is a linear relationship

$$y = y(x) = I + Sx, \quad \text{where } I = y(0) \text{ and } S = \frac{dy}{dx} \quad (1.1)$$

between y and x , but the points do not lie on this straight line because of uncertainties in the measurements, then

linear regression is a method of finding the most satisfactory straight line to represent the data. The problem devolves into determining the most appropriate values of the I and S quantities that, on account of their graphical interpretation, are commonly called the *intercept* and *slope* of the sought relationship.

The solution to this problem gave rise to one of the most intense animosities in mathematical history. Legendre was the first – in 1806 – to publish the method of least squares. However, in his 1809 treatise *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*, Gauss claimed that, as early as 1795 when he was barely eighteen, he had anticipated Legendre and told his friend Olbers about his discovery. Legendre was sceptical and indignant, and even Laplace was drawn into the bitter quarrel [1].

The originators of the procedure, and the mathematical scientists who eagerly adopted it, called the method by the appropriate name of ‘least squares’, but the alternative designation ‘linear regression’ has since come into vogue. This unfortunate name was adopted by social scientists following Galton’s study on human characteristics and his false hypothesis of *regression towards mediocrity* [2], which in modern terms is a *regression to the mean* [3, 4]. Galton suggested that later generations are going to exhibit less variability – literally more mediocrity – than earlier ones, but that is not the case. It is a purely statistical phenomenon.

Our discussion is restricted to so-called ‘linear’ regression; see elsewhere (for example Chapter 4 of Ref. [3]) for non-linear cases. Not only is the ‘regression’ part of the name unfortunate, but the adjective ‘linear’ is also misleading; it appears that we are locked into an inappropriate nomenclature.

During the two centuries since the method’s discovery, linear regression has been thoroughly assimilated into the

✉ Elena Mashkina
elena.mashkina@monash.edu

¹ Monash University, Clayton, VIC 3800, Australia

² Trent University, Peterborough, ON K9J 7B8, Canada

Fig. 1 Example of a linear regression plot. D_j is the *residual* for j th point. I is the intercept

repertoire of the physical, chemical and social sciences and the method is discussed in a great many publications [see, for example, 3–9]. Linear least squares regression is the workhorse in the chemical laboratory which is used not only for determination of the slope and intercept of the regressed line, but, also, used to demonstrate a technique of removing bad data points from the fit when the residuals are analysed. However, with the coming of ‘one-button’ implementation of standard linear regression, it has often come to be used in inappropriate circumstances. When a problem in electrochemistry [10] demonstrated the inadequacy of conventional linear regression, we decided to look into the standard method, its failings, and how those failings might be rectified.

The standard regression

Because random uncertainties generally follow a Gaussian distribution [5], the optimal line is the one that minimizes the squares of the departures of the n points from the line, which is therefore the ‘best’ line in this sense. Figure 1 shows several points and a straight line obeying the equation $y = y(x) = I + Sx$. The deviation of the j th point from the line is given by the absolute-value expression

$$D_j = |I + Sx_j - y_j| \quad (2.1)$$

and there are $n-1$ similar deviations. Figure 1¹ interprets these deviations, which are often called *residuals*,² as the

lengths of the vertical straight-line segments connecting the points to the line. The square of the j th deviation is

$$D_j^2 = I^2 + S^2 x_j^2 + 2ISx_j - 2Iy_j - 2Sx_j y_j + y_j^2 \quad (2.2)$$

and, the total of the squared deviations is

$$\sum D^2 = nI^2 + S^2 \sum x^2 + 2IS \sum x - 2I \sum y - 2S \sum xy + \sum y^2 \quad (2.3)$$

where we use an abbreviated notation exemplified by

$$\sum (y^2) = \sum_{j=1}^n y_j^2 \quad (2.4)$$

The optimal values of S and I are those that minimize this quantity and they may be found by setting

$$\frac{\partial}{\partial S} \sum D^2 = 0 \quad \text{and} \quad \frac{\partial}{\partial I} \sum D^2 = 0 \quad (2.5)$$

On performing the partial differentiations, one finds

$$\begin{aligned} 2S \sum x^2 + 2I \sum x - 2 \sum xy &= 0 \quad \text{and} \\ 2nI + 2S \sum x - 2 \sum y &= 0 \end{aligned} \quad (2.6)$$

These two equations solve simultaneously to give the well-known “best straight line” formulas

$$\begin{aligned} S &= \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2} \quad \text{and} \\ I &= \frac{\sum x^2 \sum y - \sum x \sum xy}{n \sum x^2 - (\sum x)^2} = \frac{\sum y - S \sum x}{n} \end{aligned} \quad (2.7)$$

that are nowadays built into many calculators and computer programs. Often, other statistics, such as ‘standard errors’ and ‘correlation coefficients’ are also generated by these devices, but ancillary quantities such as these will not concern us here.

Another, and perhaps simpler, way of looking at the least-squares formulas in (2.7) is in terms of average values. Thus,

$$S = \frac{(xy)_{ave} - x_{ave} y_{ave}}{(x_{rms})^2 - (x_{ave})^2} \quad \text{and} \quad I = y_{ave} - S x_{ave} \quad (2.8)$$

where $x_{ave} = (x_1 + x_2 + x_3 + \dots + x_n)/n$ is the average of x ’s, and x_{rms} is the root-mean-square of the x ’s, equal to $\sqrt{(x_1^2 + x_2^2 + x_3^2 + \dots + x_n^2)/n}$. Notice from the second formula in (2.8) that the ‘best straight line’ passes through a point (akin to a centre of gravity) which represents the averages of all the data.

Perhaps because of the convenience and ubiquity of linear-regression technology, the ‘best straight line’ formulas (2.7) are frequently applied in circumstances where *better* straight-line formulas exist. Here, we address several such scenarios.

¹ The ordinal data for Fig. 1 and all the following figures were produced using a pseudo-random number generator implemented in Matlab.

² Though the same name is also applied to the signed value.

Weighted regressions

The formulas just derived treat all data points even-handedly. If the j th measurement and its uncertainty³ is represented as $(y_j \pm \delta y)$ then the same value of δy was taken to apply to every other y . For a variety of reasons, this is not always realistic; the uncertainty $\pm \delta y_j$ associated with point j may differ from that associated with other points. Those points with lower uncertainty deserve greater weight. If we associate the weight w_j with the j th point, and, instead of Eq. (2.5), we substitute

$$\frac{\partial}{\partial S} \sum w D^2 = 0 \quad \text{and} \quad \frac{\partial}{\partial I} \sum w D^2 = 0, \quad (3.1)$$

where $\sum w D^2 = \sum_{j=1}^n w_j D_j^2$

then a repetition of the minimization procedure of “The standard regression” leads to the formulas

$$S = \frac{\sum w \sum wxy - \sum wx \sum wy}{\sum w \sum wx^2 - (\sum wx)^2} \quad \text{and} \quad (3.2)$$

$$I = \frac{\sum wy - S \sum wx}{\sum w}$$

The absolute values of the weights are immaterial; only relative weights are significant.

How may weights be assigned? Sometimes ad hoc assignments might be made. Perhaps some of the measurements were made by a more precise instrument or a more proficient observer, warranting a greater weight. The significance of the weights, and how they might be assigned methodically, will be more evident after reading “Repeated measurements” and “Regression with explicit uncertainties”.

Repeated measurements

It is often convenient to employ a range of x values, but to take several y readings at a particular value of x . There are two ways of looking at the resulting data. All the measurements may be regarded as individual, the regression approach of “The standard regression” being followed. Alternatively, one may lump together all measurements made at a common value – say x_i – of x , treating $(y_i)_{ave}$, the group-averaged y_i , as the pertinent y datum. There might be several such groups, not necessarily of the same size, along with some lone measurements. Clearly, even if there is uniformity in the uncertainty of each individual measurement, that uniformity will be destroyed by group-

averaging, because the averaging of repeats diminishes random uncertainties.

To make the regression results of group-averaging match those obtained by treating each measurement in the standard fashion, it is necessary to employ the formulas

$$S = \frac{\sum m \sum mxy - \sum mx \sum my}{\sum m \sum mx^2 - (\sum mx)^2} \quad \text{and} \quad (4.1)$$

$$I = \frac{\sum my - S \sum mx}{\sum m}$$

where, for example,

$$\sum mxy = \sum_{i=1}^g m_i x_i y_i \quad (4.2)$$

Here m_i is the number of samples in group i , all measured at x_i and y_i is the average of those measurements. Of course $\sum m$ means the number of measurements in *all* groups, equal to n , the total number of measurements overall. The g is the number of groups, some of which could have as few as a single member.

The similarity of the equations in (4.1) to those in (3.2) provides insight into what ‘weight’ means quantitatively. To give a particular measurement a weight of 3, all the others being unweighted (that is, having $w = 1$) is equivalent to repeating that special measurement three times. Of course, there are reasons other than repetitions for weighting points unevenly. Generally, unlike the m ’s, the w ’s need not be integers.

Regression with explicit uncertainties

Averaging n repeat measurements decreases the random uncertainty by a factor of $\sqrt{n}$ if Gaussian behaviour applies. Thus we can associate the uncertainty $\pm \delta y_j$ in a measurement with $1/\sqrt{n_j}$ and hence with $1/\sqrt{w_j}$. In consequence, w_j can be replaced by $1/(\delta y_j)^2$ in the regression analysis, and Eq. (3.1) rewritten as

$$S = \frac{\sum \frac{1}{(\delta y)^2} \sum \frac{xy}{(\delta y)^2} - \sum \frac{x}{(\delta y)^2} \sum \frac{y}{(\delta y)^2}}{\sum \frac{1}{(\delta y)^2} \sum \frac{x^2}{(\delta y)^2} - (\sum \frac{x}{(\delta y)^2})^2} \quad \text{and} \quad (5.1)$$

$$I = \frac{\sum \frac{y}{(\delta y)^2} - S \sum \frac{x}{(\delta y)^2}}{\sum \frac{1}{(\delta y)^2}}$$

Of course, these equations reduce to those in (2.7) when the uncertainties are uniform. However, when this is not so, Eq. (5.1) can be used to adjust the weighting. For example, imagine carrying out a long series of measurements, $j = 1, 2, 3, \dots, n$. After J of these, however, one improves the instrumentation, which one judges to halve the uncertainty in subsequent runs. This implies a need to weight subsequent measurements by a factor of 4 and this is taken into

³ The quantity that we represent by δy and refer to as an ‘uncertainty’ in y is often called an ‘error’. The latter name is eschewed because it can also refer to departures that arise from phenomena that are not random.

account in formula (5.1) by adjusting each summation. Thus $\sum \left(\frac{xy}{(\delta y)^2} \right)$, which abbreviates $\sum_{j=1}^n \left(\frac{x_j y_j}{(\delta y_j)^2} \right)$, is replaced by $\sum_{j=1}^J \left(\frac{x_j y_j}{(\delta y_j)^2} \right) + \sum_{j=J+1}^n \left(\frac{4x_j y_j}{(\delta y_j)^2} \right)$ with similar changes in all the other sums. Notice that δy_j which here represents the uncertainty in the earlier measurements, cancels out on forming the quotients in (5.1), and could have been omitted.

Though the magnitude of the individual δy terms is unimportant, the most revealing interpretation of $a \pm \delta y$ value is, perhaps, as a ‘probable uncertainty’⁴. The significance of this quantity is that if a measurement of y_j , subject to random uncertainty, is repeated a huge number of times, then virtually one-half of those y_j measurements will lie within the range $(y - \delta y) < y_j < (y + \delta y)$, where y is the true⁵ value, and one-half outside.

Proportional uncertainties

In “The standard regression” it was assumed that each individual measurement is tainted by the same uncertainty. The assumption is clarified in Fig. 2. The two dashed lines that have been added represent probable-uncertainty limits; that is, any one point is as likely to lie outside these lines as inside. The lines are parallel to the $y = I + Sx$ line because of the uniform uncertainty postulate.

This postulate is often inappropriate, however. Another possible scenario is illustrated in Fig. 3, in which the regression line is bordered by two *diverging* probable-uncertainty lines, because the uncertainty $\pm \delta y_j$ is proportional to the measured value y_j . Instances to which this behaviour applies include cases in which y results from a counting, or some equivalent procedure. The bigger the count, the larger the absolute uncertainty in that count will be. In such cases it is the *fractional* (or percentage) uncertainty that is uniform, not the uncertainty itself. Because of the proportionality of $\pm \delta y_j$ to y_j an appropriate weight in such instances is $w_j = 1/y_j^2$. When such a weighting is introduced into formula (3.2), one finds

$$S = \frac{\sum \left(\frac{1}{y^2} \right) \sum \left(\frac{xy}{y^2} \right) - \sum \left(\frac{x}{y^2} \right) \sum \left(\frac{y}{y^2} \right)}{\sum \left(\frac{1}{y^2} \right) \sum \left(\frac{x^2}{y^2} \right) - \left(\sum \left(\frac{x}{y^2} \right) \right)^2} \quad \text{and} \quad (6.1)$$

$$I = \frac{\sum \left(\frac{y}{y^2} \right) - S \sum \left(\frac{x}{y^2} \right)}{\sum \left(\frac{1}{y^2} \right)}$$

as the parameters of the best straight line.

⁴ The probable uncertainty is 0.6745σ , where σ is statistician’s ‘standard error’.

⁵ More precisely, the population mean.

Fig. 2 Example of linear relation $y(x) = I + Sx$. The same uncertainty is assumed on each value. The solid line is the linear regression and the dashed lines represent probable uncertainty limits

Fig. 3 Example of linear relation $y(x) = I + Sx$. The uncertainties are proportional to the y values. The red solid line is the linear regression and the dashed lines represent probable uncertainty limits

The same results follow from (5.1) by assigning an uncertainty proportional to $1/y_j$. Proportional uncertainties may arise in other ways. An example will be found in the next section.

Inherent uncertainties

Often, the nature of the problem demands a particular assignment of weights. For instance, consider launching a number of projectiles and measuring (exactly) the distance l travelled and (with less precision) the time of flight t . You hypothesize that the two variables are linked by the equation $t^2 = Sl + I$, where S and I are unknown constants, and

you propose to test this postulate by a linear regression equivalent to plotting t^2 versus l . To use the standard equations from “[The standard regression](#)” would be inappropriate in this case, because those formulas place equal weight on all the points, whereas the data for long flights are more precise (they have less fractional uncertainty) than those for shorter flights. The variation in precision can be quantified, at least approximately. If the timing uncertainty is $\pm\delta t$, a constant, then, because $(t \pm \delta t)^2 = t^2 \pm 2t\delta t + (\delta t)^2$, the uncertainty in t^2 is approximately $\pm 2t\delta t$. Accordingly, the equations in (3.2) should be used, with w proportional to $1/t^2$. Setting $x_j = l_j$, $y_j = t_j^2$ and $w_j = k/t_j^2$, where k is an arbitrary constant, these equations become

$$S = \frac{\sum \frac{1}{t^2} \sum (l) - \sum \frac{l}{t^2}}{\sum \frac{1}{t^2} \sum \frac{l}{t^2} - (\sum \frac{l}{t^2})^2} \quad \text{and} \quad I = \frac{1 - S \sum \frac{l}{t^2}}{\sum \frac{1}{t^2}} \quad (7.1)$$

Note that, as expected, the k 's have cancelled and this factor need not have been included.

The formulas in (7.1) are specific to the conditions associated with this projectile problem. This is a common circumstance in regression analysis. The standard formulation is likely to be unsuitable for your particular conditions. It will give an answer, but it will not be the best answer.

The appropriate linear regression formula, in circumstances such as this, depends on the mathematical linkage between what is measured and what is plotted. Consider the following chemical application. Small additions of acetic acid are made to a solution of sodium acetate. The Law of Chemical Equilibrium predicts that the hydrogen ion concentration will vary linearly with the amount of added acid, and thus the plan is to carry out a linear regression with $y = [\text{H}^+]$ and $x = m_{\text{CH}_3\text{COOH}}$. However, the hydrogen ion concentration is being measured by a pH meter that can only discriminate to 0.05 of a pH unit. The relation between the regression variable and the measurement variable, in this case, is

$$y = [\text{H}^+] = 10^{-\text{pH}} = e^{-2.303\text{pH}} \quad (7.2)$$

and therefore

$$\begin{aligned} y \pm \delta y &= e^{-2.303[\text{pH} \pm \delta(\text{pH})]} = e^{-2.303[\text{pH} \pm 0.05]} \\ &= e^{-2.303\text{pH}} e^{\mp 0.1151} \approx y(1 \pm 0.12) = y \pm 0.12y \end{aligned} \quad (7.3)$$

The number ± 0.12 is unimportant as far as the regression is concerned (though not, of course, to the precision associated with the S and I values calculated via the regression). The vital conclusion is that δy is proportional to y ; the uncertainty in the hydrogen ion concentration is proportional to the concentration itself. This is exactly the circumstance addressed in “[Proportional uncertainties](#)” and the formulas developed there apply.

Regressions with abscissal error

Hitherto, all the formulas in this document presuppose uncertainty being present in the y values. In some scenarios, however, the measurement of the y variable may be exact, whereas uncertainty exists in the value of the independent variable x that elicited the measured response. For example, one might be able to measure accurately a sequence of gas pressures, while being somewhat uncertain of the precise temperature to which the measured values relate.

When uncertainty is present solely in the x measurements, the deviations are now represented by *horizontal* lines, as in Fig. 4. The deviation of the j th point is given by

$$D_j = \left| \frac{y_j - I}{S} - x_j \right| \quad (8.1)$$

and the total sum of the squared deviations by

$$\sum D^2 = \frac{\sum y^2}{S^2} - \frac{2I \sum y}{S^2} + \frac{nI^2}{S^2} - \frac{2 \sum xy}{S} + \frac{2I \sum x}{S} + \sum x^2 \quad (8.2)$$

When one goes through the same exercise as in “[The standard regression](#)” – partially differentiating with respect to S and to I , setting the derivatives to zero, and solving the resulting simultaneous equations – one finds that

$$I = \frac{\sum xy \sum y - \sum x \sum y^2}{n \sum xy - \sum x \sum y} \quad \text{and} \quad S = \frac{\sum y - nI}{\sum x} \quad (8.3)$$

Though not, of course, identical, the formulas in (8.3) mirror those in (2.7). Analogues of the results in “[Weighted regressions](#)”, “[Repeated measurements](#)”, “[Regression with explicit uncertainties](#)”, “[Proportional uncertainties](#)”, and “[Inherent uncertainties](#)” may be developed to cater to uncertain abscissal values.

Fig. 4 Linear regression plot with uncertainties in x measurements. D_j is the residual for j th point. I is the intercept

Regressions with ordinal and abscissal uncertainty

Not uncommonly neither the y data nor the x data are free from uncertainty, and one needs to perform a regression that takes account of both a δy and a δx . Occasionally, as in intercalibrating two thermometers, the ordinal and the abscissal units will be identical, but more usually x and y will have different physical dimensions. To overcome the difficulties that this would otherwise cause, we will develop a linear regression by a procedure inspired by the writing of Worthing and Geffner [7]. In this section δx and δy are assumed constant.

Hitherto, we have thought of the regression in terms of fitting data to the equation $y = I + Sx$ in (x, y) Cartesian space. Let us now, however, view the problem as one in a different Cartesian space in which the coordinates are

$$X = x/\delta x \quad \text{and} \quad Y = y/\delta y \quad (9.1)$$

Each original point (x_j, y_j) transforms to a point (X_j, Y_j) in the new space and the $y = I + Sx$ line transforms into

$$Y = i + sX \quad \text{where} \quad i = I/\delta y \quad \text{and} \quad s = S\delta x/\delta y. \quad (9.2)$$

All physical dimensions have now been removed from the problem, which may be portrayed as in Fig. 5, with the new coordinates and equal scaling on each axis. As illustrated, the deviation of the j th point from the line equals the distance from that point to the nearest location on the line, which trigonometric reasoning shows to be

$$D_j = \frac{|Y_j - i - sX_j|}{\sqrt{1 + s^2}} \quad (9.3)$$

Hence

Fig. 5 Linear regression plot using normalized coordinates X and Y , to remove the physical dimensions from regression problem. As before, D_j is the residual for j th point

$$\sum D^2 = \frac{\sum Y^2 + ni^2 + s^2 \sum X^2 - 2i \sum Y - 2s \sum XY + 2is \sum X}{1 + s^2} \quad (9.4)$$

We follow the now-familiar path of partial differentiation of expression (9.4), first with respect to i , and then with respect to s , each partial derivation being subsequently equated to zero. The first partial differentiation is easy:

$$\frac{\partial}{\partial i} \sum D^2 = \frac{2ni - \sum Y + 2s \sum X}{1 + s^2} = 0 \quad (9.5)$$

which leads to

$$i = \frac{\sum Y - s \sum X}{n} \quad (9.6)$$

Before attempting the more elaborate second partial differentiation, it is convenient to use result (9.6) to consolidate Eq. (9.4) into

$$\sum D^2 = \frac{A - 2sB + s^2C}{n(1 + s^2)} \quad (9.7)$$

where

$$\begin{aligned} A &= n \sum Y^2 - \left(\sum Y \right)^2, \quad B \\ &= n \sum XY - \sum X \sum Y, \quad \text{and} \quad C \\ &= n \sum X^2 - \left(\sum X \right)^2 \end{aligned} \quad (9.8)$$

Lengthy algebra establishes that the partial derivative with respect to s is

$$\frac{\partial}{\partial s} \sum D^2 = \frac{-sA + (s^2 - 1)B + sC}{\frac{1}{2}n(1 + s^2)^2} \quad (9.9)$$

and setting this quantity to zero creates a quadratic equation in s , of which the solution is

$$s = \frac{A - C}{2B} \pm \sqrt{1 + \left(\frac{A - C}{2B} \right)^2} \quad (9.10)$$

It now remains to transform the solution we have found in X, Y space back to the original x, y space. In the original variables, the quantity that appears twice in Eq. (9.10) is

$$\frac{A - C}{2B} = \frac{(\delta x)^2 [n \sum y^2 - (\sum y)^2] - (\delta y)^2 [n \sum x^2 - (\sum x)^2]}{2(\delta x)(\delta y)[n \sum xy - \sum x \sum y]} \quad (9.11)$$

and so s may be found, via Eqs. (9.11) and (9.10), from the original data. The sought slope and intercept then follow from

$$S = \frac{s\delta y}{\delta x} \quad \text{and} \quad I = \frac{\sum y}{n} - \frac{s(\delta y) \sum x}{n(\delta x)} \quad (9.12)$$

As Worthing and Geffner note [7], circumstances justifying the approach of this section occur quite often, but the method appears to find infrequent application.

Constrained regressions

Inherent in all the foregoing is the assumption that the slope S and the intercept I are entirely independent of each other; that is, the problem has *two* degrees of freedom. But the need also arises to locate the best straight line when the regression is “constrained” in some way, so that there is only *one* degree of freedom. Here we shall discuss three distinct types of constraint, all of which occur frequently in scientific applications.

The first type of constraint occurs when one of the $\{S, I\}$ pair is constrained to have some specified value; this case is addressed in “[Constant-parameter constraints](#)”. Other scenarios are considered in later sections.

Constant-parameter constraints

Here either S or I is required to have some specified value. It might be that the slope must adopt the predetermined value, K . Then the first equality in (2.4) becomes meaningless and, in consequence, the first item in (2.6) is invalid. The second item in (2.6) shows the intercept to be

$$I = \frac{\sum y - K \sum x}{n} \quad (11.1)$$

in the case of equal weights, or, when uncertainties vary

$$I = \frac{\sum \frac{y}{(\delta y)^2} - K \sum \frac{x}{(\delta y)^2}}{\sum \frac{1}{(\delta y)^2}} \quad (11.2)$$

An instance of constrained regression that arises frequently is when the intercept is required to be zero; the regression line must pass through the origin of the graph. The solution in the unweighted case

$$S = \frac{\sum xy}{\sum x^2}, \quad (11.3)$$

follows directly from the first item in (2.6). Once more, we are treating δy as constant. You will not need to look far in the scientific literature to find instances in which standard formula (2.7) was used, when the simpler (11.3) was appropriate. The answers can be quite different.

Interestingly, expression (11.3) for the slope may alternatively be derived from formula (3.2). Augment the data set by the point (0,0) and give this extra point a weighting that is massively greater than all the other weights, which

are equal. Equation (11.3) results from the S formula in (3.2).

The constraint treated analytically

The general instance of constrained linear regression occurs when S and I are interrelated in some specified way. Although the converse assignment is equally valid, here we treat S as a known function of I . Instead of having the two relations in (2.3), we have only a singular requirement. Partial derivatives are no longer appropriate; instead the regression criterion is

$$\frac{d}{dI} \sum D^2 = 0 \quad (12.1)$$

On applying this to Eq. (2.3), one finds

$$\frac{dS}{dI} = \frac{S \sum x + nI - \sum y}{\sum xy - S \sum x^2 - I \sum x} \quad (12.2)$$

This is the equation that governs constrained linear regression.

The simplest application of Eq. (12.2) is when the slope is proportional to the intercept, $S = kI$. Then the solution to Eq. (12.2) shows the intercept to be

$$I = \frac{\sum y + k \sum xy}{n + 2k \sum x + k^2 \sum x^2} \quad (12.3)$$

and, of course, S is just k times this.

Another simple case occurs when the line is required to pass through some fixed point (X, Y) . Then the requirement is that $S = (Y - I)/X$ and Eq. (12.2) leads to

$$I = \frac{XY \sum x - Y \sum x^2 + X \sum xy - X^2 \sum y}{2X \sum x - \sum x^2 - nX^2} \quad \text{with } S = \frac{Y \sum x - nXY - \sum xy + X \sum y}{2X \sum x - \sum x^2 - nX^2} \quad (12.4)$$

Of course, this result collapses to (11.3) when $X = Y = 0$.

A more elaborate instance of the applicability of Eq. (12.2) arose recently in an electrochemical context [10]. The constraint was that the slope of the line be proportional to the square of the line's intercept, the proportionality constant being known. Thus we have $S = kI^2$ and $dS/dI = 2kI$, so that (12.2) becomes

$$2kI = \frac{kI^2 \sum x + nI - \sum y}{\sum xy - kI^2 \sum x^2 - I \sum x} \quad (12.5)$$

It follows that the intercept is given by the solution to the cubic equation

$$2k^2 I^3 \sum x^2 + 3kI^2 \sum x + I(n - 2k \sum xy) - \sum y = 0 \quad (12.6)$$

and the slope is then accessible as kI^2 . A cubic equation has three solutions, but in the application we were able to reject two of the three solutions on inspection, because the intercept had to be positive and only one of the three solutions had a positive sign.

The constraint treated numerically

Cubic equations *can* be solved analytically, but that route is sufficiently onerous that they are nowadays usually solved numerically [8]. When the constraint is more complicated than the $S = kI^2$ case, it will usually be mandatory to solve Eq. (12.2) numerically. That being so, one might just as well tackle the problem numerically from the outset.

The slope S of an arbitrary straight line may adopt any value between $-\infty$ and $+\infty$. The intercept of that line has similar flexibility. Thus any straight line whatsoever may be represented as a point in a Cartesian plane that has I and S as its coordinate axes. An *unconstrained* regression line corresponds to a single point, say $\{I_0, S_0\}$, in this “parameter plane”. All *constrained* regression lines, on the other hand, are represented by a relationship – namely Eq. (12.2) – between S and I , and thus may be depicted as a line in the parameter plane. A pictorial representation of this concept is shown in Fig. 6a.

A numerical approach can be suggested, as depicted in Fig. 6b, c. We start by finding the pair $\{I_0, S_0\}$ that solves the unconstrained linear regression. $\{I_0, S_0\}$ is the starting point in the parameter plane for the optimization procedure (Fig. 6b). The first step is to calculate the pair of parameters $\{I_1, S_1\}$ that satisfies the constraint $S_1 = f(I_1)$ (e.g. it may be the parabola $S = kI^2$) and is the closest to the $\{I_0, S_0\}$ pair. This is accomplished by minimizing the square distance D^2 of $\{I_0, S_0\}$ from the constraint line $S = f(I)$ in the parameter plane (Fig. 5b):

$$D^2 = (I_0 - I_1)^2 + (S_0 - S_1)^2 = (I_0 - I_1)^2 + (S_0 - f(I_1))^2 \quad (13.1)$$

This quantity is minimized by nullifying the derivative:

$$\begin{aligned} \frac{\partial D^2}{\partial I_1} &= 2(I_0 - I_1) + 2 \left[f(I_1) \frac{\partial f(I_1)}{\partial I_1} - S_0 \frac{\partial f(I_1)}{\partial I_1} \right] \\ &= 2(I_0 - I_1) + 2 \frac{\partial f(I_1)}{\partial I_1} (f(I_1) - S_0) \end{aligned} \quad (13.2)$$

In all cases for which an analytic solution does not exist, this nullified equation is solved numerically to find the value of I_1 and hence S_1 . The final step of the procedure is to find a pair $\{I_m, S_m\}$ which belongs to the curve and minimize the sum of square residuals:

Fig. 6 **a** The starting point $\{I_0, S_0\}$ in the parameter plane, the *solid line* being the constraint relation between I and S . **b** The minimal distance D of $\{I_0, S_0\}$ from the *solid line*. **c** A schematic drawing of the search procedure to identify the best fit that satisfies the constraints. The *crosses* represent subsequent steps in the search procedure

$$\sum_{i=1}^N (y_i - y_i^m)^2 \quad (13.3)$$

where $y_i^m = S_m x_i + I_m$. This is achieved by successively varying the coefficient I_m along the curve using I_1 as the starting point (Fig. 6c). The algorithm searches both directions sequentially⁶ and stops when a minimum is found. The precision of the method depends on the step size ΔI of the successive searches. Such a step can be chosen arbitrarily small. Typically, choosing $\frac{\Delta I}{I} \sim 0.1\%$ is sufficient for the precision of the constrained linear regression in an electrochemical context.

Let us re-examine the practical example introduced consider towards the end of “Constrained regressions”, where $S = kI^2$. Then Eq. (13.2) becomes

⁶ There are two directions to “walk”, but it will soon be obvious if you’re walking the wrong way.

$$\frac{\partial D^2}{\partial I_1} = 2(I_0 - I_1) + 4kI_1(kI_1^2 - S_0) = 0 \quad (13.4)$$

which is solved numerically and the iterative optimization procedure described above is implemented.

Summary

‘Linear regression’ or ‘least squares’ has a distinguished history and is one of the most widely used statistical techniques. Its popularity has burgeoned now that it can be implemented electronically with little effort or thought. We have demonstrated, however, that there are many circumstances under which the standard formulas of linear regression analysis are inappropriate. Replacement formulas are presented.

Acknowledgments Authors would like to thank Dr. A. Simonov and Prof. A. Bond from Monash University for earlier discussions.

Compliance with ethical standards The authors deny any financial, or other, conflict of interest in disseminating this material.

References

1. Newman JR (1956) *The World of Mathematics*. Simon & Schuster, New York
2. Galton F (1885) Regression towards mediocrity in hereditary stature. *Anthropol Miscellanea*: 246–263
3. de Levie R (2012) *Advanced Excel for scientific data analysis*, 3rd edn. Atlantic Academic, Harpswell
4. Brownlee KA (1965) *Statistical theory and methodology in science and engineering*, 2nd edn. Wiley, New York
5. Simon MK (2006) *Probability distributions involving Gaussian random variables. A handbook for engineers, scientists and mathematicians*. Springer-Verlag, New York
6. Kohman TP (1970) Least-squares fitting of Data with Large Errors. *J Chem Educ* 47:657–658
7. Worthing AG, Geffner J (1943) *Treatment of experimental data*. Wiley, New York
8. Press WH, Flannery BR, Teukolsky SA, Vetterling WT (1986) *Numerical Recipes, the art of scientific computing*. Cambridge University Press, Cambridge
9. Oldham K, Myland J, Spanier J (2000) *An atlas of functions*, 2nd edn. Springer, New York
10. Simonov AN, Mashkina E, Mahon PJ, Oldham KB, Bond AM (2015) Determination of diffusion coefficients from semiintegrated d.c. and a.c. voltammetric data: overcoming the edge effect at macrodisc electrodes. *J Electroanal Chem* 744:110–116