Talanta 116 (2013) 45–49

Contents lists available at SciVerse ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta

Short communication

The connection between inverse and classical calibration

Emili Besalú *

Department of Chemistry and Institute of Computational Chemistry and Catalysis, Universitat de Girona, Av. Montilivi s/n, 17071 Girona, Spain

ARTICLE INFO

Article history:
Received 18 May 2012
Received in revised form
17 April 2013
Accepted 24 April 2013
Available online 2 May 2013

Keywords:
Classical calibration
Inverse calibration
Linear equation
Regression towards the mean

ABSTRACT

Within the context of the simple classical linear calibration procedure (regression of y on x), here it is shown how a distinction between the distributions of the observed dependent variable (y_{obs}) and the calculated (fitted) one (y_{calc}) leads to the following counterintuitive approach: in order to get the independent x values with lesser systematic deviations, do not identify as direct inputs in the classical calibration equation the new y observed ones (experimentally acquired), but instead the transformed ones by means of a regression towards the mean effect correction. It is shown how the conjunction of both steps, i.e., first the transformation of observed values and then the ulterior use in the classical calibration equation, corresponds to an operation totally equivalent to the direct implementation of the inverse calibration equation (regression of x on y in a single step). The reasoning given here explains in a simple manner why the inverse calibration numerically performs usually better for predictions of interpolated x values. Results are accompanied with the analysis of both theoretical and experimental data.

1. Introduction

In classical calibration, the dependent variable y_{obs} (for instance, an experimentally acquired absorbance or signal) is regressed on the independent one x_{obs} (usually a concentration). It is assumed that the dependent variable has a statistical uncertainty and that the errors are normally distributed around the true values with constant variance, whereas x_{obs} is error-free or, at least, it has a substantial lesser error than the dependent variable. The regression of n (x_{obs},y_{obs}) data points provides with the classical linear model useful to obtain the expected value for y_{obs} variable from the knowledge of x_{obs}

$$E(y_{obs}|x_{obs}) = y_{calc} = a + b\,x_{obs} \tag{1}$$

being a and b the estimators of the true unknown parameters α and β which express the underlying relationship among the x and y variables (e.g., Beer–Lambert law). The model is also characterized by the determination coefficient r^2 indicating the fraction of data variance explained by Eq. (1). It has not to be understood that previous Eq. (1) stands for the relationship $y_{obs}=a+bx_{obs}$. The observed y_{obs} variable and the calculated one, y_{calc}, are not interchangeable entities. This article concerns about this distinction.

On the other side, the linear inverse calibration consists into regress the x_{obs} variable on y_{obs}, usually disregarding which one bears errors or which one is a random independent error-free variable, thus violating basic assumptions underlying least squares regression [1]. The inverse model obtained with the same previous set of n available points is of the form

$$E(x_{obs}|y_{obs}) = x_{calc} = a' + b'y_{obs} \tag{2}$$

Models (1) and (2) are not homologous nor interchangeable, but are related in such a way that the knowledge of a model will automatically lead to the other equation (see next paragraph). The two equations share the same determination coefficient and, as much as the determination coefficient tends to 1, both equations will tend to be the same, minimizing the difference between the classical and inverse procedures [2].

The numerical relationship between models (1) and (2) obtained from the same set of n points is very simple [3]: the product of the slopes b and b' is r^2, and both equations are satisfied by the data points center of mass $(\bar{x},\bar{y})$. Note that, for the n points data set, $\overline{y_{calc}} = \overline{y_{obs}} \equiv \bar{y}$, as it is also well known. Consequently, once the regression parameters a, b and r are known from the classical fitting (1), the ones appearing in (2) can be obtained either by the inverse regression procedure or directly by the following relationships:

$$\begin{cases} a' = \frac{(1-r^2)\bar{y}-a}{b} \\ b' = \frac{r^2}{b} \end{cases} \tag{3}$$

Conversely, a similar transformation exists in order to infer the equation parameters of (1) from the previous knowledge of those of (2). We will call here *conjugated* a couple of regression equations (the classical and the inverse) arising from the same set of fitted points and linked by the aforementioned rule (3).

* Tel.: +34 972 41 8875; fax: +34 972 41 8356.
E-mail address: emili.besalu@udg.edu

0039-9140/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.talanta.2013.04.054

Within the classical approach to calibration, Eq. (1) serves to estimate the x of an unknown sample from its measured y_{obs} value, the one provided by the experiment. Along this process, the acquired variable y_{obs} is implicitly identified as being the calculated variable y_{calc}. That's a mistake because doing that it is *incorrectly* assumed that, once one instance of y_{obs} is known, the corresponding expected value for y_{calc} is y_{obs} itself. Below it will be shown how the expression $E(y_{calc}|y_{obs})=y_{obs}$ does *not* hold in general. In spite of this, the two-step process usually followed to obtain x from the knowledge of y_{obs} is described by the codification

$$\begin{cases} 1. \quad y_{obs} \rightarrow y_{calc} \\ 2. \ (y_{calc}-a)/b \rightarrow x_{obs} \end{cases} \tag{4}$$

where in step (4.1) the erroneous identification $E(y_{calc}|y_{obs})=y_{obs}$ is being assumed, and then x is derived isolating from (1) by means of step (4.2).

The process of inverse calibration, i.e., to rely on Eq. (2), in some circumstances performs better for predictions of x of new known experimental y values [4–9]. This constitutes a statistical trend but not a general situation because the performance depends on the kind of data we are manipulating, on the distribution of x variable, on the number of replicates, on the distance of the predicted points from the mean x and y values and, among others, on the followed criteria to define superiority of a method. Here, by better we mean that the sum of quadratic errors (mean squared errors, MSE) between obtained x values and the actual ones is lesser. In many practical situations, inverse approach performs better for interpolations of x values within the explored interval [4]. This feature was noticed early and discussed in the context of Montecarlo simulations [5,6] and the interest continued during the time until now (see for instance references in [8]). Investigations have been carried on dealing with many aspects of this situation, as for instance the asymptotic performance of the inverse calibration for $n\rightarrow\infty$ [1,9] or showing that the inverse procedure if favored respect to the classical one even if the number of data point samples is small [8]. Shukla [7] shows that, despite one method is not in all the cases superior to the other, the inverse approach gives lesser MSE when a single value of y is available and the inferred value of x lies near the mean of the population sample (e.g. for interpolations). Modern approaches take into account a balance of practical factors and favor the inverse approach respect to the classical one [8].

This paper exposes an elemental relationship which links inverse and classical procedures and that explains in a simple manner why many times inverse calibration performs better than the classical one when inferring interpolated values of x. The reasoning deals with the distinction which has to be made between calculated y values (y_{calc}) and observed ones (y_{obs}). The key idea is that the knowledge of a particular value of the variable y_{obs} has not to be directly identified with the same numerical value of the variable y_{calc}. This seems counterintuitive because, according to (4), the usual classical approach consists into get y_{obs} values from the population sample and identify them with y_{calc} ones by plugging the former numerical value directly into Eq. (1) to get x.

2. Results

The exposition below is based on a general theorem described within the multiple linear regression framework [10,11] and is related to the regression towards the mean effect [12]. The theorem applies for any arbitrary and finite data set disregarding particular statistical distributions of x and y variables and their correlation values. The exposition will be initially illustrated using an arbitrary artificial data set. In order to improve the visualization of some features, the artificial set shows a small correlation value despite this is not the case in general calibration or analytical purposes. Afterwards, data coming from real experiments will be also analyzed.

2.1. Artificial data set

An artificial large set of $n=3000$ (x_{obs},y_{obs}) sample points has been generated exhibiting a low correlation value ($r^2=0.80$). Without loose of generality, the low correlation value and other data parameters have been chosen in order to made graphically evident the explored features. The variable x is normally distributed ($\mu={}^{9}/_{2}$, $\sigma={}^{7}/_{6}$) and the y data points obey to the inner (true) linear relationship $y=1+2x$ (i.e., $\alpha=1$ and $\beta=2$), but modified by a Gaussian noise error function having mean zero and a fixed arbitrary variance (1.3535) which leads to the aforementioned coefficient of determination. Once the classical linear regression equation is calculated, the fitting line $y_{calc}=0.995+2.00x_{obs}$ is obtained, which stands for Eq. (1). Fig. 1 shows the representation of the calculated values (y_{calc}) against the observed ones (y_{obs}). The diagonal solid line corresponds to the equation bisector $y=x$. It has *not* to be assumed that the cloud of points depicted in Fig. 1 is symmetrically distributed along the bisector equation. Sometimes that's the erroneous underlying idea which leads to apply the procedure (4). The assumption that the numerical values of the variables y_{obs} and y_{calc} are interchangeable is incorrect.

Fig. 1 reveals the evidence of a regression towards the mean effect [12] due to an artifact of the linear model construction (in fact, due to the asymmetry in x and y variables treatment). As it can be seen, the point cloud is not symmetrically distributed along the bisector line but rotated around the point cloud center of mass. As a consequence, for a given experimental value, for instance the observation $y_{obs}=15$ depicted in Figure 1, the expected corresponding y_{calc} one *is not* the same number in general (unless for the particular cases for which $r^2=1$ or y_{obs} coincides with the sample mean value $\bar{y}$). In Fig. 1 it is also shown the distribution of y_{calc} values attached to the particular observed result $y_{obs}=15$. It is graphically revealed how from the knowledge that the value $y_{obs}=15$ has been experimentally acquired, the corresponding mean (expected) value of variable y_{calc} is *distinct* than 15. This artifact is overlooked many times. The expected y_{calc} value is depicted at the center of the distribution curve and lies in the diagonal dashed line. Fig. 1 reveals that, within the context of

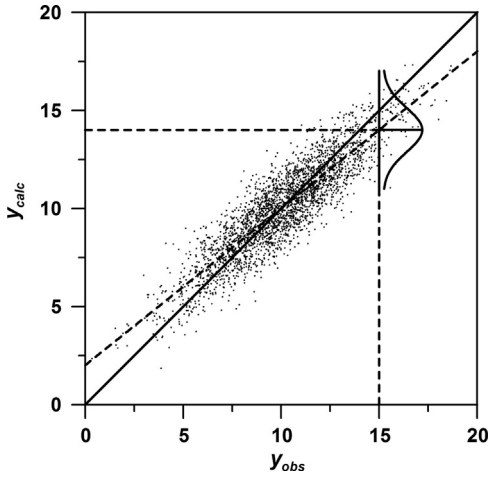

Fig. 1. Representation of calculated y values (y_{calc}) by means of the classical regression Eq. (1) against the original observed (experimental) ones (y_{obs}). Solid diagonal line is the quadrant bisector whereas the dashed diagonal one corresponds to the linear regression of calculated values on the observed ones. The other dashed lines and the Gaussian show how the expected y_{calc} value attached to an observed one ($y_{obs}=15$) are not coincident.