MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **55** (2006) 281-286

ISSN 0340 - 6253

Some Plots Are not that Equivalent

Emili Besalú[*], J. Vicente de Julián-Ortiz[a], Lionello Pogliani[b]

[*] Institute of Computational Chemistry, Universitat de Girona, Facultat de Ciències, Av. Montilivi s/n, 17071 Girona, Spain, emili@iqc.udg.es.
[a] Red de Investigación de Enfermedades Tropicales. Dept. de Biología Celular y Parasitología, and Unidad de Investigación en Diseño de Fármacos y Conectividad Molecular, Dept. de Química Física; Facultad de Farmacia, Av. V. A. Estellés s/n, 46100 Burjassot (València), Spain, julian@goya.combios.es
[b] Dipartimento di Chimica, Università della Calabria, 87030 Rende (CS), Italy, lionp@unical.it.

(Received October 10, 2005)

Abstract

In many scientific and chemistry-related fields it is very common to represent in a bidimensional plot calculated and observed data in many scientific and chemistry-related fields. If calculated values are obtained via a linear or multilinear regression procedure it will be shown how the two representation choices, fitted vs. observed and observed vs. fitted values, are not equivalent. The slopes of the bidimensional regression lines in both plots bear distinct properties: the former representation exhibits a regression line with a slope always equal to r^2 and the later line coincides exactly with the bisector of the first and third quadrants representation. The general proof of this problem is here exemplified by the aid of a simple numerical example. An alternative method for obtaining a graphical 'symmetric' representation is exposed, which relies on the minimization of the sum or quadratic orthogonal distances.

Introduction

The use of plot methods obtained by the aid of the least-squares methodology is quite common in correspondence of calibration and model studies of any type, and X-ray crystallographic studies. A good mastering of statistics is a good help either in theoretical or experimental chemistry. The fact that a good deal of spreadsheets and statistical packages are at disposition either for the student or for the trained researcher does not always mean that the person who is using them does clearly understand what is going on. Usually, at the level of handling plot methods based on the multilinear least-squares methodology some misconceptions continue to be detected. During these last years, several articles dealing with statistical methods, both theoretical and applicative have been published in chemical

literature. [1-5] The subject treated in this paper, independently rediscovered by the first author of the present article and extensively reviewed in a recent work [6], has been considered only once in the specialized literature and in an indirect form, [7] for what we know, and it seems that it went totally unnoticed. Furthermore, from plots published in the literature and from considerations collected by the present authors, it is evident that contradictory views on the subject are, actually, entertained: some believe that the subject is trivial while others believe that it is just wrong. [8, 9, and references in the cited papers] Thus, it is not at all unworthy to reconsider it, as the fundamental principle that interchanging the dependent and independent variables results in unequal regression plots does not at all seem obvious. The actual reconsideration will also allow correcting and completing the conclusions of two recent published works on the importance of plot methods in model studies, which without the support of the present concepts will not tell the whole story. [4, 5]

A lot of graphics can be found in the literature in which the plot of observed (y) versus calculated or fitted ($\hat{y} \equiv y_{calc}$) data (we will denote this plot as $y \mid \hat{y}$), are normally displayed in a quadrant where the bisector line, with unit slope and zero intercept, represents the ideal case. This representation is very useful in multivariate analysis and quite common in Quantitative Structure-Property Activity Studies. Heuristically, most people accept that the reversed representation, i. e., the $\hat{y} \mid y$ plot constitutes an equivalent representation of the same data. In fact, both representations are similar, but not equivalent. For instance, the linear regression fitting in each bidimensional plot returns two distinct equations:

$$y \cong a\hat{y} + b \tag{1}$$

and

$$\hat{y} \cong cy + d \tag{2}$$

From a theorem of linear regressions, [10] it is well-known that the slopes of eqs. 1 and 2 fulfill the condition $ac = r^2$. The question that now arises deals with the presumed equivalence between the two plots: are the slopes a and c equal? The answer is no. In fact, for fitted data with ordinary multilinear least-squares, $a = 1$ always, and in consequence, $c = r^2$. Even more, in eq. 1 $b = 0$ and this regression line coincides exactly with the first and third quadrants bisector, thus warranting that in the $y \mid \hat{y}$ representation the depicted points will lie around the bisector line representing the ideal case. The other regression line 2 will coincide with the former only in the very rare and 'trivial' case of $r^2 = 1$. Note that the correlation factor r^2

coincides with the determination coefficient of the observed data with respect to the parameters employed in the linear or multilinear model that defined the calculated values. For a detailed mathematical derivation of the concepts here discussed, the reader is referred to ref. [6].

A numerical example

Let us check the aforementioned aspects with a rather simple example: be the following vector of a-dimensional experimental data,

$$\boldsymbol{y} = (120, 166, 115, 105, 155, 257, 220, 100, 70, 55)$$

and the following vector of arbitrary descriptors,

$$\boldsymbol{x} = (147, 171, 134, 133, 159, 312, 313, 113, 103, 95)$$

By means of the least-squares method the model equation $\hat{y} = 0.766\ x + 7.653$ is obtained, giving the vector of calculated data:

$$\hat{\boldsymbol{y}} = (120._2, 138._6, 110._3, 109._5, 129._4, 246._6, 247._3, 94._2, 86._5, 80._4)$$

The elemental statistics are: number of data points $n = 10$, significance distribution ratio $F = 82$, coefficient of determination $r^2 = 0.912$, standard deviation $s = 20$. The simplest case of a least-squares model has been investigated here. If desired, the reader can also proceed with a multilinear equation involving additional series of descriptors. In Fig. 1 left and right are the corresponding two plots, $\hat{y} \mid y$ and $y \mid \hat{y}$, respectively, on which the corresponding linear regressions (eqs. 1 and 2) have been added (absolute values less than 10^{-4} have been rounded to zero). From the attached equations we see that $ac = r^2 = 0.912$, but we note also that it is the other way round, i.e., the ideal case is reproduced in the $y \mid \hat{y}$ plot with slope $a = 1$, and $b = 0$, while $c = r^2$, and $d \neq 0$.

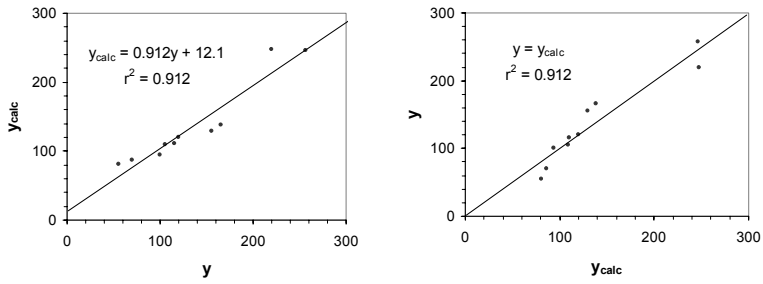

Figure 1. Bidimensional plots, $\hat{y} \mid y$ (left), and $y \mid \hat{y}$ (right) corresponding to the example discussed in the text and which shows the use of the linear regressions of eqs. 1 (left), and 2 (right).

A closer look into the corresponding residual D plots, with $D = \hat{y} - y$, i.e., into the $D \mid y$ and the $D \mid \hat{y}$ plots, a similitude with the previous two plots can be detected, as can be seen in Fig. 2 left, and right, respectively. Here the attached equations tell that only the $D \mid \hat{y}$ has random character, while the $D \mid y$ equation indicates that this plot is patterned and that its intercept equals the intercept of the $\hat{y} \mid y$ plot. The proofs about this D plots can also be found in ref. [6].

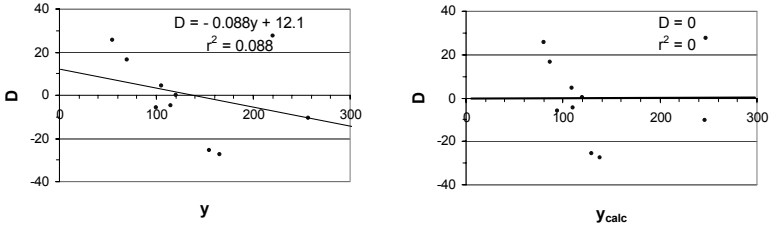

Figure 2. Residual plots, $D \mid y$ (left), and $D \mid \hat{y}$ (right) obtained from the example given in the text and related to the graphics shown in Figure 1.

A symmetric model

This asymmetry, which is present in the revisited plots, can be avoided if the least squares method considers orthogonal distances to the best fitting-line, instead of vertical distances from it. The method of orthogonal regression has a long history in statistics, chemometrics and economics. Although we present here a simplified treatment without experimental errors, it has been proposed as the correct algorithm when uncertainties in both dependent and independent variables are considered or when they are not predetermined. For this reason, it is sometimes called the errors-in-variables model. As stated by Anderson [11] "the method of orthogonal regression was discovered and rediscovered many times, often independently". However, this method was first published by Adcock. [12] In this section, a brief sketch of the method is presented.

Assume that two variables, Y and X, are theoretically linearly related. That is,

$$Y = \alpha + \beta X + u \,,$$

where α is the intercept, β is the slope and u is the equation error, then, the orthogonal regression coefficients are (see [12] for the derivation)

$$\alpha = \overline{Y} - \beta\ \overline{X} \quad \text{and} \quad \beta = \frac{\sigma_Y^2 - \sigma_X^2 \pm \sqrt{\left(\sigma_Y^2 - \sigma_X^2\right)^2 + 4\sigma_{XY}^2}}{2\sigma_{XY}}$$

where σ_{XY} is the covariance between X and Y, and σ_X^2 and σ_Y^2 are the variances of X and Y, respectively. The slope β has two possible values: the two slopes give two perpendicular fitting lines coinciding with the principal components of the data. [12] One of such slopes gives the minimum sum of quadratic distances, the other the maximum for a line passing across the data center of mass.

In the present context, let us examine our example concerning the data vectors $\boldsymbol{y}$ and $\boldsymbol{\hat{y}}$ mentioned above. The straight line that minimizes the orthogonal distances is $\hat{y} = 0.953y + 6.453$. This line reproduces a mean squared point-line distance equal to 158.9 per point, and a mean distance of 10.0 per point. Note that the slope lies between those present in the regression lines in Figure 1, as the line of minimal quadratic orthogonal distances is always graphically found between the other two straight lines (all the three lines pass across the data center of mass). Nevertheless, it does not agree with the line that bisects the angle they form.

Conclusions

Representing calculated and observed data in the same bidimensional plot is very common in many scientific and chemistry-related fields. It has been shown that, if calculated values are obtained via a linear or multilinear regression procedure, the two representation choices, $\hat{y}$ vs. y and y vs. $\hat{y}$, are not exactly equivalent. Even more, in the last case the bidimensional regression line exactly coincides with the 'desired' bisector. As a consequence, the former representation exhibits a regression line with a slope equal to r^2. The properties in both graphs are inherited by the corresponding residuals plots. A simple numerical example dealing with linear regressions has been presented. While the general proof has been presented in ref [6], here, in the discussion section, has been exposed the basic formula in order to obtain the regression line which minimizes the sum or quadratic orthogonal distances.

Acknowledgements. The first two authors acknowledge financial help to grant number BQU2003-07420-C05 of the 'Ministerio de Ciencia y Tecnología' within the Spanish Plan Nacional I+D. This grant also allowed first author visiting the University of Valencia, the place where this work started. Furthermore, the second author acknowledges the 'Red

Tematica de Investigacion Cooperativa RICET (Red de Investigacion de Centros de Enfermedades Tropicales) of the Spanish Ministry of Health for financial support. The third author acknowledges the help and support of J. Gálvez and R. García-Domenech, during his sabbatical at the Unidad de Investigación en Diseño de Fármacos y Conectividad Molecular of the department of Química Física of the Facultad de Farmacia at the University of Valencia.

References

1) A. Golbraikh, and A. Tropsha, Beware of q^2 ! *J. Mol. Graph. Modell.* 20 (2002) 269-276.

2) H. Liu, R.G. Sadygov and J.R. Yates, III, A model for random sampling and estimation of relative protein abundance in Shotgun proteomics. *Anal. Chem.* 76 (2004) 4193-4201.

3) H.Z. Msimanga, P. Elkins, S.K. Tata and D.R. Smith, A Chemometrics Module for an undergraduate instrumental analysis chemistry course, *J. Chem. Educ.* 82 (2005) 415-424.

4) L. Pogliani and J.V. de Julián-Ortiz, Plot methods in quantitative structure-activity studies, *Chem. Phys. Lett.* 393 (2004) 327-330.

5) L. Pogliani and J.V. de Julián-Ortiz, Residual plots and the quality of a model, *MATCH Commun. Math. Comput. Chem.* 53 (2005) 175-180.

6) E. Besalú, J.V. de Julian-Ortiz, M. Iglesias, L. Pogliani, An Overlooked Property of Plot Methods. *J. Math. Chem*. in press.

7) N.R. Draper and H. Smith, *Applied Regression Analysis*, Wiley, New York, 1966.

8) J.T. Olesberg, M.A. Arnold, S.-Y. B. Hu, J.M. Wiencek, Temperature-Insensitive near-infrared method for determination of protein concentration during protein crystal growth. *Anal. Chem.* 72 (2000) 4985-4990.

9) Private letters in the hands of E. Besalú.

10) M.R. Spiegel, *Probability and Statistics*, McGraw-Hill, New York, 1975.

11) T.W. Anderson, Estimating Linear Statistical Relationships, *Ann. Statist.* 12 (1984) 1-45.

12) R.J. Adcock, A problem in least squares, *Analyst* 5 (1878) 53-54.