MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **63** (2010) 573-583

ISSN 0340 - 6253

Ordinary and Orthogonal Regressions in QSAR/QSPR and Chemistry-Related Studies

Emili Besalú*, Jesus V. de Julián-Ortiz+, Lionello Pogliani^,

* Institute of Computational Chemistry, Universitat de Girona, Facultat de Ciències, Av. Montilivi s/n, 17071 Girona, Spain, emili.besalu@udg.es

+ MOLware SL and Fundación Investigación e Innovación para el Desarrollo Social, C/Burriana 36, 3, 46005, Valencia, Spain, jejuor@uv.es

^Dipartimento di Chimica, Università della Calabria, 87030 Rende (CS), Italy, lionp@unical.it

(Received January 27, 2010)

Abstract

A critical examination of different least squares orthogonal methods (OR1-OR4) and of the ordinary least squares (LS) method, which is normally used in QSAR and QSPR studies, and in many scientific and chemistry-related fields, reveals that not always the orthogonal regression methods perform better than LS in the aforementioned fields. Nonetheless the OR methods, whose use in statistics and economics are considered superior in most cases to LS, relying on the minimization of the sum or quadratic orthogonal distances offer an interesting alternative method for obtaining a graphical 'symmetric' representation, which is not rendered by the LS method.

Introduction

A critical analysis of plot methods [1-4, and references therein] in QSAR/QSPR and chemistry-related studies, has brought us to center the attention into some improper uses of plots resulting from the ordinary least squares (LS) regressions and to stress the importance of a no well-known characteristics of orthogonal least squares regression (OR) and of the corresponding plots [5-7]. Plot methods in QSAR/QSPR and in many other scientific and chemistry-related studies vehicle essential information in a compact and easy way, and should no more be considered an optional by practitioners of these fields, even if papers without plots but burdened of Tables with observed/calculated data continue to be published. Plots methods can illustrate and detect violation of assumptions; that is, values should show random fluctuations around the main diagonal of the observed vs. calculated plot, while the corresponding residual plots (residual vs. calculated values) should show random fluctuation

around a value of zero. Furthermore, it is well-known that to base a model on statistical parameters only can be quite misleading. Orthogonal regressions avoid the pitfall of unsymmetrical observed vs. calculated plots and unsymmetrical residual plots, nevertheless statisticians have proposed different types of orthogonal regressions, based on different assumptions [5-9].

In this paper we will center our attention on some more recent and intriguing results on orthogonal regressions which, even if they are well-known to the statisticians [10, 11], biologists [12], bioinformatics [13], economists [14] and to the physicists [15], are practically ignored by the chemistry community, and especially by computational chemists and QSAR/QSPR practitioners.

Orthogonal regression analysis is required by ISO 16140. The ISO 16140 (2003) standard describes the technical protocol for the validation of alternative methods within the framework of the microbiology of food and animal feeding stuffs [16].

In the present paper we will review the different orthogonal least squares methods, with special attention to the seminal and widely cited paper of Dissanaike and Wang [11] on the subject and compare their validity with the ordinary least squares procedure, as this is a topic that the computational chemistry and the analytical chemistry community, should, at least, not overlook.

Method

Le us assume that the two variables y and x are linearly related, that is,

$$y + \varepsilon_y = a + b(x + \varepsilon_x) + u \tag{1}$$

Here, a is the intercept, b is the slope, u is the equation error with zero mean, ε_y is the measurement error for y and ε_x is the measurement errors for x, both with zero mean. The usual way to estimate a and b in chemistry-related studies is to use the LS method, which minimizes the vertical distance between the observations and the fitted line. In this method, with the assumptions $\sigma^2_{\varepsilon x} = 0$, $\sigma^2_{\varepsilon y} = 0$, and $Cov(u, x) = 0$, for the slope and the intercept we have: $b = \sigma_{xy} / \sigma^2_x$ and $a = \langle y \rangle - b\langle x \rangle$, where $\langle y \rangle$ and $\langle x \rangle$ are the sample means of y and x respectively [8-10]. Here σ^2 is the variance of the corresponding subscript, when subscripts are different (i.e., xy, σ^2 is their covariance). In the following lines are displayed four

orthogonal regression methods proposed until now. The figures the reader will found along the text are obtained by the distinct proposed procedures and will be explained with more detail in the simulation section.

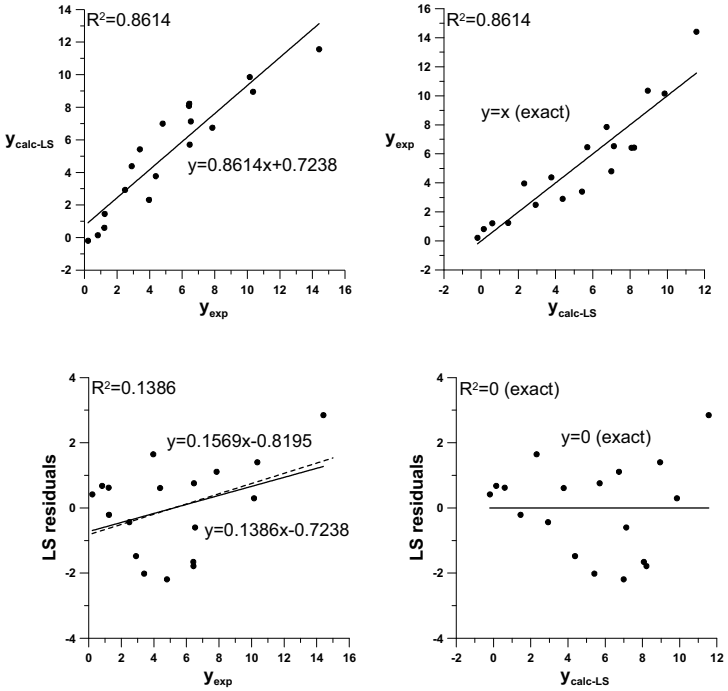

Fig. 1. LS method: calculated vs. observed values with their residuals (left), observed vs. calculated values with their residuals (right).

Classical Orthogonal Regression (OR1)

OR1 has recently been suggested [5, and reviewed in 7] to solve some problems inherent to the LS method. This method considers a least squares criteria which minimizes the (orthogonal) distance between the observations and the fitted line. OR1 assumes no errors whatsoever, i.e., $\sigma^2_{\varepsilon x} = 0$, $\sigma^2_{\varepsilon y} = 0$, $\sigma^2_{u} = 0$, and its slope is,

$$b_{OR1} = \frac{\sigma_y^2 - \sigma_x^2 \pm \sqrt{\left(\sigma_y^2 - \sigma_x^2\right)^2 + 4\sigma_{xy}^2}}{2\sigma_{xy}} \qquad (2)$$

As already told, the intercept is calculated by the same way for all models: $a = \langle y \rangle - b\langle x \rangle$. OR1 method presents a special characteristic: the obtained fitting line defines the first Principal Component of the data [7], i.e., the direction in space for which orthogonal projection of the data points gives the maximal data variance. The other fitting line arising from the alternative sign in equation (2) defines the second Principal Component.

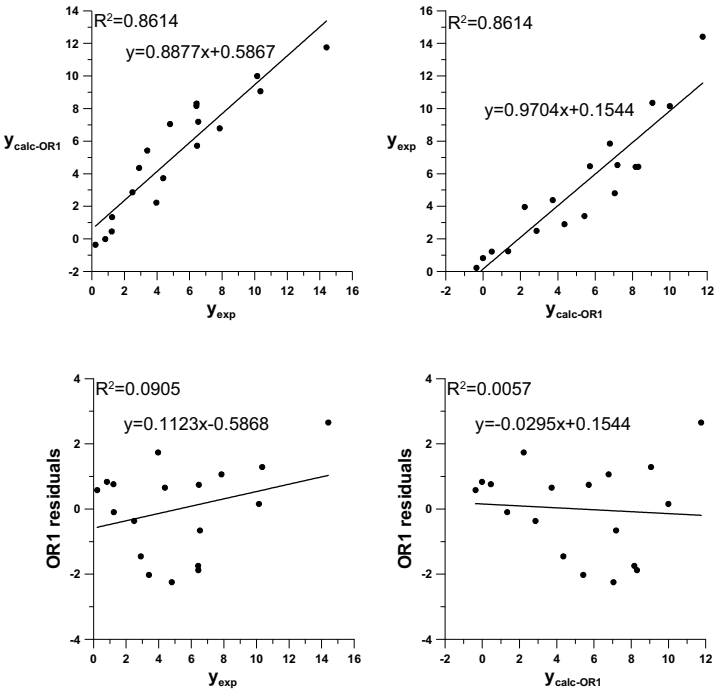

Fig. 2. OR1 method: calculated vs. observed values with their residuals (left), observed vs. calculated values with their residuals (right).

Ref. 11 (and references therein) describes three more orthogonal methods (OR2-OR4). Each one of them is based on different assumptions. Let us first review these three orthogonal methods (the detailed calculations are in the appendix section of ref. 11).

Orthogonal regression method 2 (OR2)

OR2 is a classical Orthogonal Regression with measurement errors in variables and no equation error term. The assumptions are $\sigma^2_{\varepsilon y} / \sigma^2_{\varepsilon x} = \lambda$, $\sigma^2_u = 0$, $Cov(u, x) = 0$; with these assumptions the slope is given by :

$$b_{OR2} = \frac{s_y^2 - \lambda s_x^2 \pm \sqrt{\left(s_y^2 - \lambda s_x^2\right)^2 + 4\lambda s_{xy}^2}}{2 s_{xy}} \tag{3}$$

were

$$s_x^2 = \sigma_x^2 + \sigma_{\varepsilon x}^2, \quad s_y^2 = \sigma_y^2 + \sigma_{\varepsilon y}^2, \quad b = \frac{\sigma_y}{\sigma_x}, \quad s_{xy} = \sigma_x \sigma_y$$

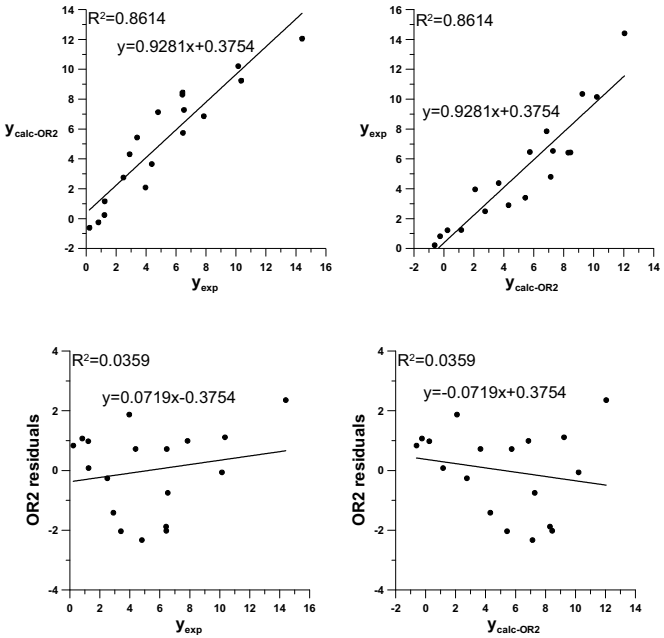

Fig. 3. OR2 method: calculated vs. observed values with their residuals (left), observed vs. calculated values with their residuals (right).

Orthogonal regression method 3 (OR3)

OR3 constitutes a classical Orthogonal Regression with the effect of the equation error term, but no measurement error in the variables. Assumptions are $\sigma^2_{\varepsilon y} = \sigma^2_{\varepsilon x} = 0$, $Cov(u, x) = 0$, but $\sigma^2_u \neq 0$, here the slope now is,

$$b_{OR3} = \frac{\sigma_y^2 - \sigma_x^2 - \sigma_u^2 \pm \sqrt{\left(\sigma_y^2 - \sigma_x^2 - \sigma_u^2\right)^2 + 4\sigma_{xy}^2}}{2\sigma_{xy}} \tag{4}$$

In practice, since σ^2_u is unknown, we use OR1 to estimate σ^2_u as the quadratic average of its residuals, and then run OR3 and OR4 procedures.

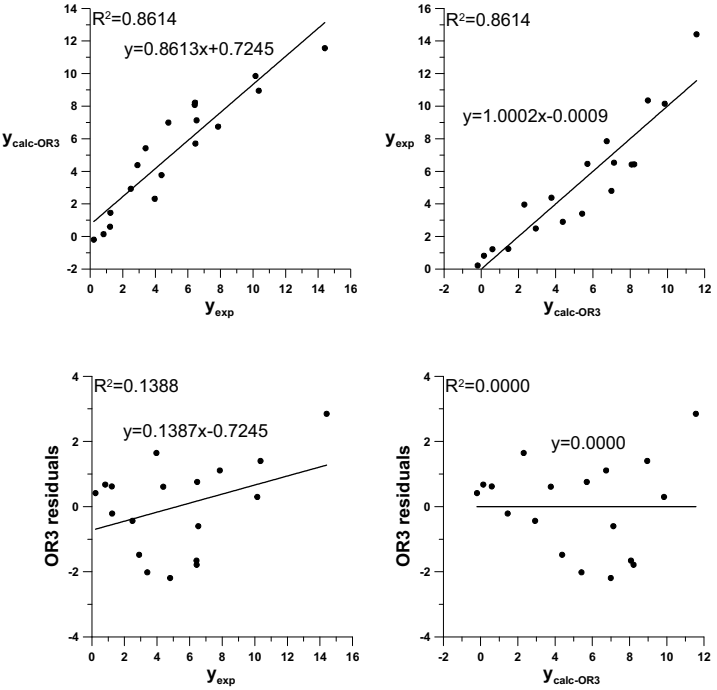

Fig. 4. OR3 method: calculated vs. observed values with their residuals (left), observed vs. calculated values with their residuals (right).

Orthogonal regression method 4 (OR4)

OR4 is the adjusted Classical Orthogonal Regression with measurement errors in the variables and equation errors. Now the statistical assumptions are $\sigma^2_{\varepsilon y} / \sigma^2_{\varepsilon x} = \lambda$, and $\sigma^2_u \neq 0$. The slope here is,

$$b_{OR4} = \frac{s_y^2 - \lambda s_x^2 - \sigma_u^2 \pm \sqrt{\left(s_y^2 - \lambda s_x^2 - \sigma_u^2\right)^2 + 4\lambda s_{xy}^2}}{2 s_{xy}} \tag{5}$$

providing a general expression from which the other slope formulas presented above can be deduced.

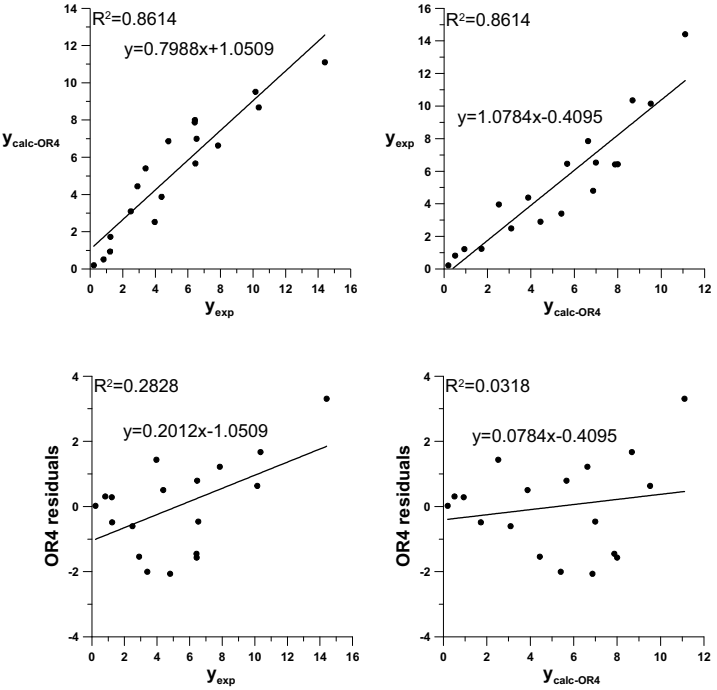

Fig. 5. OR4 method: calculated vs. observed values with their residuals (left), observed vs. calculated values with their residuals (right).

Regarding the LS method, the knowledge of a new numerical variable value (x or y) leads to the immediate computation of the fitted one (y_{fit} or x_{fit}, respectively) directly via the use of the regression equation and maintaining fixed the original value entered into the equation (x or y). On the other side, for OR1-4 methods, it must be noted that the knowledge of a new experimental point (x_{new},y_{new}) leads to the corresponding fitted values (x_{fit},y_{fit}) by projecting the (x_{new},y_{new}) point to the obtained regression line. This means that in these cases *both*, the x and y new values, are susceptible to be changed simultaneously to the pair of values (x_{fit},y_{fit}). In particular, for OR1 the projection must be orthogonal to the regression line [7] giving the result:

$$x_{fit} = \frac{x_{new} + b_{OR1} y_{new} - a b_{OR1}}{1 + b_{OR1}^{\ 2}} \quad \text{and} \quad y_{fit} = b_{OR1} x_{fit} + a \tag{6}$$

Simulations

We simulate now an ideal quantitative structure-property or structure-activity relationship and test the five different models LS, OR1, OR2, OR3, and OR4. Here x is a descriptor (a connectivity χ index) and y is a property or activity (P), where the choice of the data for simulation purposes has no effect on the quality of the simulation. Data are given in Table 1 and are taken from ref. 7. Actually a dependent variable y (=P) is generated based on a given x (=χ) according to pre-assigned slope values under the four different given assumptions. The distinct resulting regression equations are shown in Table 2. In figures 1-5 are shown the respective plots for the five methods LS and OR1-4.

In the LS method the graph of residuals versus calculated values gives a null correlation. The line fitting (again by LS) these points also presents a null slope. This is due to the inherent mathematical relations among the residuals and the fitted line in the graph which is another LS line (as in the other graphs). Due to a theorem [3,6], valid for LS and in general for Multiple Linear Regressions, if a new regression LS line is obtained for the experimental versus calculated points (see Figure 1 top-right) it has to be obtained the y=x equation (the bisector of the first and third quadrants). This goes accompanied by the additional rule being that for the calculated-LS versus experimental values (reversed graph, Figure 1 top-left) the bidimensinoal LS line has to have a slope coinciding exactly with the data determination coefficient, i.e., with the R^2 value.

Table 1. Experimental data used for simulations and taken from reference 7. R^2=0.8614 among the data.

x	y_{exp}
0.86	0.22
1.57	0.82
2.53	1.22
4.32	1.24
6.13	3.96
7.42	2.49
9.19	4.38
10.47	2.9
12.65	3.4
13.25	6.46
15.43	7.85
15.96	4.8
16.25	6.53
18.24	6.42
18.53	6.43
20.07	10.35
21.97	10.15
25.56	14.41

Table 2. Equations obtained with each method.

Method	Equation
LS	$y_{calc\text{-}LS} = 0.4760\,x - 0.6050$
OR1	$y_{calc\text{-}OR1} = 0.4905\,x - 0.7825$
OR2	$y_{calc\text{-}OR2} = 0.5129\,x - 1.0563$
OR3	$y_{calc\text{-}OR3} = 0.4760\,x - 0.6040$
OR4	$y_{calc\text{-}OR4} = 0.4414\,x - 0.1813$

It is well known that the LS equation (the first one in Table 2) is not 'reversible', i.e., the regression of x values over y ones has to be re-calculated (whereas there is a theorem allowing to compute it in a fast way). On the other side, the OR1 method is clearly symmetrical in the sense that both variables (x and y) are treated in the same terms (even more, the attached errors are equal for both variables: zero). This situation ensures that for this method the "regression" of x over y will give the equation arising from the isolation of the variable x of the second equation in Table 2: $x_{calc\text{-}OR1} = 2.0386\,y + 1.5951$.

In any case, for OR methods the user should be aware of data variable units: the ideal situation is the one for which the accounted magnitude and the units of x and y variables are the same. This allows the treatment of symmetrical or quasi-symmetrical cases, especially for the OR1 method.

In the figures, the differences found in leftmost and rightmost graphs are not due to a mathematical artifact. These differences are inherent to the least squares criteria considered by each method.

Conclusions

The given simulations of the five least squares methods, LS, OR1 – OR4, let us notice that the LS method is not such a bad model in some situations. In other situations specific OR models are advantageous if it is known that errors are present in several variables of model parameters, especially for the so-called 'independent' variables. A rather similar result was achieved by Dissanaike and Wang [11].

It should be noticed that with data, as in the analysis of atmospheric data [15], that does not lend itself to calling one variable independent and the other dependent the ordinary least squares approach, however, could be highly problematic. Furthermore, errors often exist for both measurements, and in both cases the use an orthogonal regression method to derive the slope estimator should be the preferred solution.

Acknowledgements

E. B. thanks project CTQ2009-09370 of the Spanish Ministerio de Ciencia e Innovación.

Supplemetary Materials

The interested reader can ask the Excel file with the complete calculations to E. Besalú: emili.besalu@udg.es

References

[1] L. Pogliani, J. V. de Julian–Ortiz, Plot methods in quantitative structure-property studies, *Chem. Phys. Lett.* **393** (2004) 327–330.

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

[3] E. Besalú, J. V. de Julian–Ortiz, M. Iglesias, L. Pogliani, An overlooked property of plot methods, *J. Math. Chem.* **39** (2006) 475–484.

[4] R. García–Domenech, J. Galvez, J. V. de Julian–Ortiz, L. Pogliani, some new trends in chemical graph theory, *Chem. Rev.* **108** (2008) 1127–1169.

[5] E. Besalú, J. V. de Julian–Ortiz, L. Pogliani, Some plots are not that equivalent, *MATCH Commun. Math. Comput. Chem.* **55** (2006) 281–286.

[6] E. Besalú, J. V. de Julian–Ortiz, L. Pogliani, Trends and plot methods in MLR studies, *J. Chem. Inf. Model.* **47** (2007) 751–760.

[7] E. Besalú, J. V. de Julian–Ortiz, L. Pogliani, Two–variable linear regression: Modeling with orthogonal least squares, *J. Chem. Educ.* in press.

[8] M. R. Spiegel, *Probability and Statistics,* McGraw–Hill, New York, 1975.

[9] N. R. Draper, H. Smith, *Applied Regression Analysis,* Wiley, New York, 1966, pp. 174.

[10] J. D. Jackson, J. A. Dunlevy, Orthogonal least squares and the interchangeability of alternative proxy variables in the social sciences, *The Statistician* **37** (1988) 7–14.

[11] G. Dissanaike, S. Wang, A critical examination of orthogonal regression, *Soc. Sci. Res. Net.* **407560** (2003) 1–39.

[12] K. A. Mooijman, M. Poelman, H. Stegeman, C. Warmerdam, P. F. M.Teunis, A. M. de Roda Husman, *Validation and Comparison of Methods for Enumeration of Faecal Coliforms and Escherichia Coli in Bivalve Molluscs*, RIVM report 330310001, 2006.

[13] E. Novikov, E. Barillot, An algorithm for automatic evaluation of the spot quality intwo–color DNA microarray experiments, *BMC Bioinformatics* **6** (2005) 293–311.

[14] E. Dobrescu, Double conditioned potential output, *28th General Conference of the International Association for Research in Income and Wealth Cork, Ireland, August 22–28, 2004.*

[15] L. Leng, T. Zhang, L. Kleinman, W. Zhu, Ordinary least square regression, orthogonal regression, geometric mean regression and their applications in aerosol science, *J. Phys. Conference Series* **78** (2007) 1–5.

[16] http://www.adria.tm.fr/vars/fichiers/ISO%2016140%20revision.pdf. Accessed on January 22, 2010.