Journal of Mathematical Chemistry, Vol. 39, Nos. 3/4, May 2006 (© 2006)
DOI: 10.1007/s10910-005-9035-z

An overlooked property of plot methods

Emili Besalú*

Institute of Computational Chemistry, Universitat de Girona, Facultat de Ciències, Avda. Montilivi s/n, 17071 Girona, Spain
E-mail: emili@iqc.udg.es

J. Vicente de Julián-Ortiz

Unidad de Investigación en Diseño de Fármacos y Conectividad Molecular, Facultad de Farmacia, Dep. Química Física and Red de Investigación de Centros de Enfermedades Tropicales, Dep. Biología Celular y Parasitología. Av. V. Andrés Estellés s/n, 46100 Burjassot, València, Spain

Monica Iglesias

Department of Chemistry, Universitat de Girona, Facultat de Ciències, Avda. Montilivi s/n, 17071 Girona, Spain

Lionello Pogliani

Dipartimento di Chimica, Università della Calabria, 87030 Rende (CS), Italy

Received 20 July 2005; revised 15 August 2005 / Published online: 6 January 2006

An interesting and often overlooked property of the MLR least-squares method is discussed. Here it is shown how the observed vs. calculated and calculated vs. observed plots, which are usually presented as equivalent, in fact they are not. Due to the inherent properties of the MLR procedure, it is shown that the slope of the calculated vs. observed plot is r^2, as the slope of the experimental vs. calculated plot is just 1. This behaviour also has direct consequences to the corresponding residual plots.

KEY WORDS: multilinear regression properties, plot properties, error plots, calibration, MLR theorem

1. Introduction

Recently, two of the present authors published a critical letter of some model studies with wrong graphics or wrong calculations [1]. Now, the present article focuses even more on this problem, which seems based on either deeper misunderstandings or superficial oversight about the true character of a regression. Some misconceptions seem to be present and practiced by many chemists

*Corresponding author.

concerned with linear models (QSAR, quantum, analytical, etc.). The present subject, from recent correspondence we received, seems not at all evident to a significative amount of researchers. In fact, while some think that the present purpose is completely wrong others think that it is trivial. The problem is not closed, and precisely some literature [2–16] revealed that the covered topic in this article, despite being simple, is not trivial.

The main part of experimental studies are based on linear regressions and on their graphical representation, which take the form of calculated vs. observed activity or property plot, sometimes completed with the corresponding residual plots. Recently, many studies have been published showing the main information with the aid of tables only, a trend which has nasty drawbacks and which should always be avoided, as has recently been underlined [1], in fact, plot methods should always be used in any model study, as discussed in reference [1]. Many linear modelling studies assume that: (a) the plot of experimental vs. calculated values is centered around the first and third quadrants bisector, (b) the plot of calculated vs. experimental values is a mirror image of the former with respect to the bisector which, in an ideal case, must act as the bidimensional data fitting line, and (c) the residuals always ought to show Gaussian distribution with respect to the zero line. Here it is pointed out that the last two out of these three assumptions are not always fulfilled. The following alternative mathematical formulation will try to shed light in the problem. We will focus on some properties of the representation in a plane of experimental values against the adjusted ones by a multilinear model. If the calculated values are given by the equation

$$\hat{y}_i = \sum_{j=0}^{m} \beta_j x_{ij}, \quad i = 1, 2, \ldots, n. \tag{1}$$

where x_{ij} is a generic descriptor of the experimental dependent property y, then two representations are possible: the y vs. $\hat{y}$ representation (which we will generically denote here as $y|\hat{y}$) and the other way round, i.e., the $\hat{y}$ vs. y (or $\hat{y}|y$) representation. This means that if the plotted relations in the two cases are (the unidirectional symbol $\cong$ expresses the fitting line equation)

$$\hat{y} \cong a + by \tag{2}$$

and

$$y \cong c + d\hat{y} \tag{3}$$

then it is in practice assumed that points, either in $y|\hat{y}$ or in $\hat{y}|y$ representation, under good conditions of fitting must be randomly distributed around the 45° line, i.e., it is believed that equations (2), and (3) coincide and are $y \cong \hat{y}$ and $\hat{y} \cong y$. However, things are far more complex, as it will be discussed next.

Usually, model studies show the following characteristics: (i) only a type of plot is given and the $\hat{y}|y$ plot is much more frequent than the $y|\hat{y}$ one, even if the reason of the choice is not evident (despite that, references [2–16] show the later option and in reference [15] both plots are presented), (ii) the data in $\hat{y}|y$ plots are displayed around the bisector of the first and third quadrants, (iii) residual plots are seldom given [17] and on the rare occasions they are given they are displayed either as residuals vs. y or residuals vs. $\hat{y}$ with no reason for this choice either, (iv) original and residual plots are never further analysed, (v) it is tacitly assumed that points are randomly distributed around any bisector line or any zero line in residual plots, and (vi) in many cases no tables of experimental-calculated values are given, hampering the checking of the results. Actually, this last case parallels the other case where only tables are given, and no plots. The present study of the problem is coming out at a moment where there is a revival of statistical methods in model and validation methodologies [17–24]. The results presented here try to revise many multilinear model studies, which are based on a complete symmetry between $\hat{y}|y$ and $y|\hat{y}$ plots around the bisecting line of the first and third quadrants. The origin of the asymmetry of the two plots resides in the fact that the ordinary one-dimensional or multidimensional least-squares methods rely on the minimisation of the ordinates of the data only. However, the asymmetry found in the fitting lines in $\hat{y}|y$ and $y|\hat{y}$ plots disappears in the case of orthogonal regression, when the minimising function is the sum of perpendicular distances between the observation points and the regression line [25, 26].

2. Results and discussion

A fundamental and general theorem [27] regarding reversed linear regressions, as those of equations (2) and (3), states that $bd = r^2$ (and not $b = d = 1$). Furthermore, regarding again equations (2) and (3), an important result is usually overlooked for the particular case when the couple of variables treated are the experimental ones and the corresponding adjusted by a multilinear model. Such a result states that in this case always $d = 1$, as it will be demonstrated below, with the consequence that: $0 \leqslant b = r^2 \leqslant 1$, and $a \neq 0$. Then, equation (2) becomes

$$\hat{y} \cong a + r^2 y \tag{4}$$

This means that a $\hat{y}|y$ plot has a slope of r^2 and the intercept is a. In fact, as the regression line passes across the point $(\bar{y}, \hat{y})$ and $\bar{y} = \hat{y}$ (see proof in the appendix), it is straightforward to see that

$$a = (1 - r^2)\bar{y}. \tag{5}$$

In general, a compact form of equation (2) can be written:

$$\hat{y} \cong \bar{y} + r^2 (y - \bar{y}) . \tag{6}$$

Hence, as it appears trivially, only in the case of a perfect correlation this equation reads $\hat{y} \cong y$, and the intercept becomes zero.
On the other hand, equation (3), with $d = 1$ becomes

$$y \cong \hat{y} \tag{7}$$

as c is zero (see appendix). This means that a $y|\hat{y}$ plot always disposes the points randomly around the bisector of the first and third quadrants.
From equation (6) we easily obtain an additional fitting relationship for the residuals, D:

$$D = \hat{y} - y \cong -\left(1 - r^2\right)(y - \bar{y}) = -r_s^2 (y - \bar{y}) . \tag{8}$$

In this expression, $r_s^2 = 1 - r^2$ corresponds to the correlation coefficient between the residuals and the dependent variable y [28]. Equation (8) shows that a $D|y$ plot bears a regression line which passes across the mass centre of the data with slope $-r_s^2$, and this means that the residuals are not randomly distributed around the zero line. A similar interpretation is assigned to equation (6).
Alternatively, expressing the regression line in the $D|\hat{y}$ plot we obtain, according to equation (7),

$$D = \hat{y} - y \cong \hat{y} - \hat{y} = 0\hat{y}. \tag{9}$$

Thus, the slope and the intercept for this linear fit is equal to zero. This means that the residuals are not correlated with the fitted values and they are randomly scattered around the zero baseline. The asymmetry of the two starting plots $y|\hat{y}$ and $\hat{y}|y$, is also reflected in the asymmetry of the derived $D|\hat{y}$ and $D|y$ plots. It has to be noted that, when either modelling or calibrating with simple linear regressions, it is usually represented a $D|x$ plot, that is, the residuals vs. the original independent descriptor variable. As the variables $\hat{y}$ and x are related by a linear relation, the main characteristics of the plot $D|\hat{y}$ are inherited by the $D|x$ one: the residuals are also randomly scattered around the zero line.

Linear regression methods and their validation with plot methods have assumed a paramount importance in a continuously growing number of model studies of any type. From what has been proved and shown it is evident that particular 'asymmetry' properties regarding the bisector lines should always be expected in the $y|\hat{y}$ and the $\hat{y}|y$ plots, and that these properties are shared by their corresponding residual plots. This has the interesting consequence that a critical use of the two types of plots can help to detect anomalies, whenever throughout the $y|\hat{y}$ and the $D|\hat{y}$ plots the points are no longer random around the bisector and the zero line respectively, while the $\hat{y}|y$ and $D|y$ plots should maintain a trend around these two lines. As it is, for example, shown in reference [2]. Here, one is able to detect the error in the plot: the "Calculated" label must be "Experimental" and the "Predicted" label must read "Fitted" or "Calculated".

3. Example

A simple numerical example will help us to illustrate the present scheme. Since the treatment of experimental errors is beyond the scope of the present article, all calculations were performed with the highest precision available. Figures lower than 10^{-5} in absolute value were discarded.

Let $\boldsymbol{x}_d$ be the arbitrary vector of independent variables (1–10), and $\boldsymbol{y}$ the vector of experimental values (1.7, 2, 4, 5, 10, 10, 11, 14, 15, 20), joined by the fitting equation,

$$y_{\text{calc}} \equiv \hat{y} \cong 1.96x_d - 1.52,\ r^2 = 0.9605,\ F = 195,\ s = 1.3,\ n = 10.$$

Now, the plot $y|\hat{y}$ shown in figure 1 Top is scattered around the bisector of the quadrant with $c = 0$, and $d = 1$. The plot $\hat{y}|y$, shown in figure 2 Top, is clearly non-symmetric, with $a = 0.3657$ and $b = r^2 = 0.9605$, as expected. The $D|y$ plot

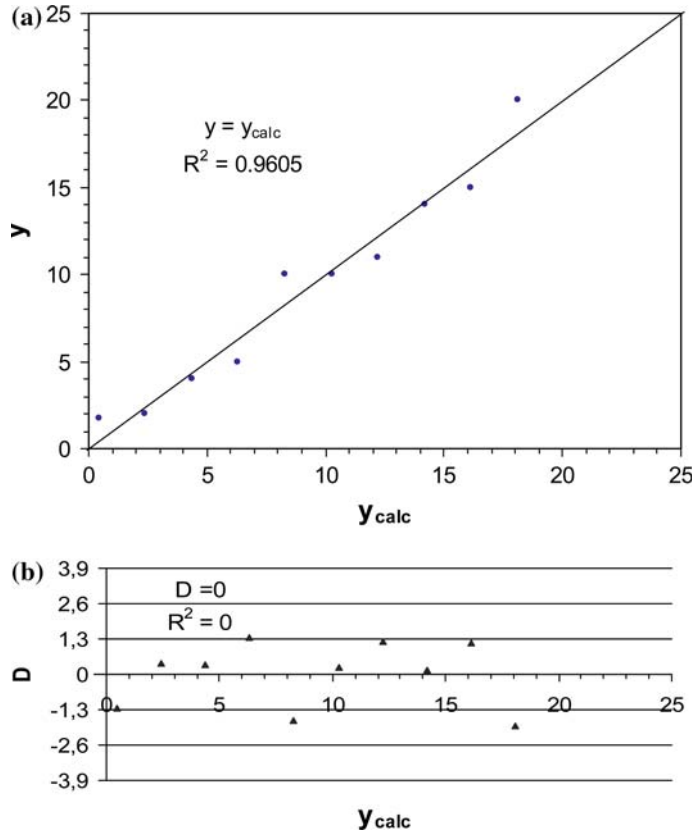

Figure 1. Top: Plot of y vs $\hat{y} \equiv y_{\text{calc}}$. Fitted line coincides with the first and third quandrants bisector. Bottom: plot of the corresponding residuals for ten casually chosen values.

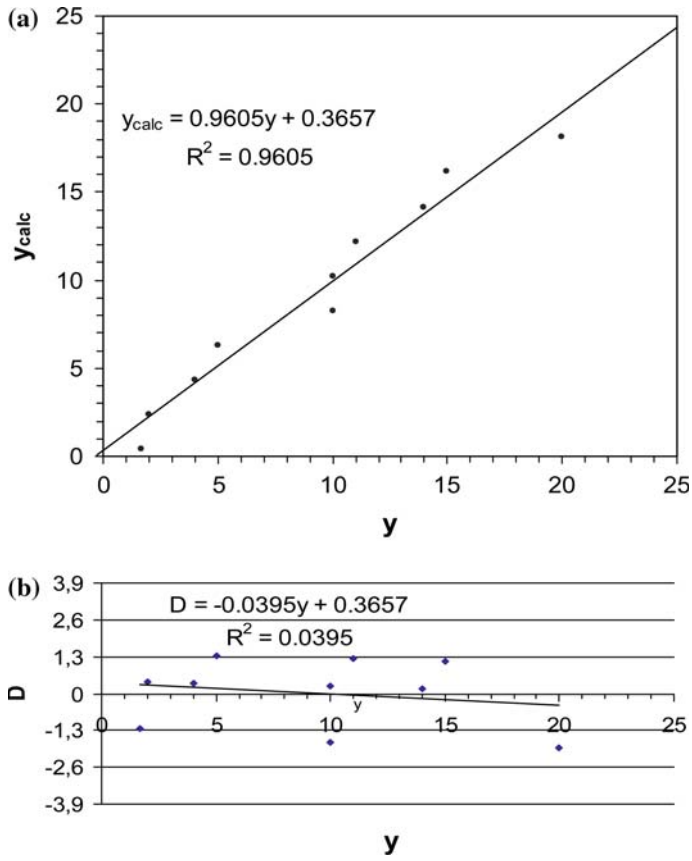

Figure 2. Top: Plot of $\hat{y} \equiv y_{\text{calc}}$ vs y. Note how the fitted line does not coincide with the bisector. Bottom: plot of the corresponding residuals for ten casually chosen values.

is shown in figure 2 Bottom, where the regions $\pm s$, $\pm 2s$, $\pm 3s$ have been highlighted by horizontal lines. This plot is, as expected, patterned, as it is detected by the fitting line, $D \cong 0.37 - 0.0395y$, with slope $r_s^2 = 0.0395 (= 1 - r^2)$ instead of zero. Additionally, the $D|\hat{y}$ plot shown in figure 1 Bottom has $r_s^2 = c = 0$, i.e., it is *more random* than the previous residual plot.

4. Conclusions

The proof which hides behind equation (4) has been presented in order to peruse its consequences, both at the level of the $\hat{y}|y$ plot as well as at the level of the residual plot. The consequences that such findings have on the description of a model have been underlined. It has been shown how gratuitous symmetry properties around the bisector line do not have to be tacitly assumed. This

becomes specially relevant in many studies, as in QSAR and QSPR fields, where this kind of graphs are usually represented.

The properties here outlined prompt to consider the observed vs. calculated plot a very first validation method, which should not exclude the further use of other important external validation methods. Of course, in order to avoid the effects of regression, $D|\hat{y}$ plot must be considered instead of $D|y$ one.

Appendix: Proof

Here it is demonstrated that in any case $d=1$ always holds in equation (3). Let us focus on fitted values obtained by the ordinary MLR procedure [28] involving one or more independent parameters. In this context, as it is expressed by equation (1), n fitted values are given by a linear combination of m descriptors or independent terms (plus a constant term) taken from an $n\times(m+1)$ matrix, $\boldsymbol{X}=x_{ij}$, where x_{ij} is the numerical value for the j-th descriptor attached to the datum number i. In this notation the β_j terms are the linear coefficients, being β_0 the independent term. The matrix and vector notation of equation (1) can be obtained:

$$\hat{y} = \sum_{j=0}^{m} \beta_j \boldsymbol{x}_j = \boldsymbol{X}\boldsymbol{\beta}, \tag{10}$$

if the following definitions are taken into account: the vector of coefficients is $\boldsymbol{\beta}=(\beta_0, \beta_1, \beta_2, \ldots, \beta_m)^T$ and the j-th column of matrix $\boldsymbol{X}$ is denoted as $\boldsymbol{x}_j=(x_{1j}, x_{2j}, \ldots, x_{nj})^T$. Introducing the n-dimensional unity vector, $\boldsymbol{1}=(1, 1, \ldots, 1)^T$, the first column of matrix $\boldsymbol{X}$ also coincides with it: $\boldsymbol{x}_0=\boldsymbol{1}$.

Regarding the $y|\hat{y}$ plot, the mono-dimensional fitted line corresponds to equation 3 and its slope is calculated as

$$d = \frac{\sum_{i=1}^{n} \left(\hat{y}_i - \bar{\hat{y}}\right)(y_i - \bar{y})}{\sum_{i=1}^{n} \left(\hat{y}_i - \bar{\hat{y}}\right)^2} = \frac{\left(\hat{\boldsymbol{y}} - \boldsymbol{1}\bar{\hat{y}}\right)^T \left(\boldsymbol{y} - \boldsymbol{1}\bar{y}\right)}{\left(\hat{\boldsymbol{y}} - \boldsymbol{1}\bar{\hat{y}}\right)^T \left(\hat{\boldsymbol{y}} - \boldsymbol{1}\bar{\hat{y}}\right)}, \tag{11}$$

where

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{\boldsymbol{1}^T\boldsymbol{y}}{n} = \frac{\boldsymbol{y}^T\boldsymbol{1}}{n}, \tag{12}$$

and

$$\bar{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i = \frac{\boldsymbol{1}^T\hat{\boldsymbol{y}}}{n} = \frac{\hat{\boldsymbol{y}}^T\boldsymbol{1}}{n} \tag{13}$$

are mean values. The $\boldsymbol{y}$ vector collects the dependent terms. The revisited equations are adapted from the generic ones which can be found in many books of statistics [28]. Then, it follows that

$$d = \frac{\hat{\boldsymbol{y}}^T \boldsymbol{y} - \bar{y}\hat{\boldsymbol{y}}^T \boldsymbol{1} - \bar{\hat{y}}\boldsymbol{1}^T \boldsymbol{y} + \bar{\hat{y}}\bar{y}\boldsymbol{1}^T \boldsymbol{1}}{\hat{\boldsymbol{y}}^T \hat{\boldsymbol{y}} - \bar{\hat{y}}\hat{\boldsymbol{y}}^T \boldsymbol{1} - \bar{\hat{y}}\boldsymbol{1}^T \hat{\boldsymbol{y}} + \left(\bar{\hat{y}}\right)^2 \boldsymbol{1}^T \boldsymbol{1}} = \frac{\hat{\boldsymbol{y}}^T \boldsymbol{y} - \bar{y}\boldsymbol{1}^T \hat{\boldsymbol{y}} - \bar{\hat{y}}\boldsymbol{1}^T \boldsymbol{y} + n\bar{\hat{y}}\bar{y}}{\hat{\boldsymbol{y}}^T \hat{\boldsymbol{y}} - 2\bar{\hat{y}}\boldsymbol{1}^T \hat{\boldsymbol{y}} + n\left(\bar{\hat{y}}\right)^2}, \tag{14}$$

as $\boldsymbol{1}^T \boldsymbol{1} = n$ and the scalar product between vectors is commutative. That is, due to equations (12) and (13), the following general expression is found:

$$d = \frac{\hat{\boldsymbol{y}}^T \boldsymbol{y} - n\bar{y}\bar{\hat{y}}}{\hat{\boldsymbol{y}}^T \hat{\boldsymbol{y}} - n\left(\bar{\hat{y}}\right)^2}. \tag{15}$$

On the other hand, beta coefficients appearing in equation (10) are obtained from the normal equation:

$$\boldsymbol{X}^T \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^T \boldsymbol{y}. \tag{16}$$

Combining equations (10) and (16) a very useful identity arises:

$$\boldsymbol{X}^T \hat{\boldsymbol{y}} = \boldsymbol{X}^T \boldsymbol{y}. \tag{17}$$

Using again equation (10) and equation (17), it can be written

$$\hat{\boldsymbol{y}}^T \hat{\boldsymbol{y}} = (\boldsymbol{X}\boldsymbol{\beta})^T \hat{\boldsymbol{y}} = \boldsymbol{\beta}^T \boldsymbol{X}^T \hat{\boldsymbol{y}} = \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{y} = (\boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{y} = \hat{\boldsymbol{y}}^T \boldsymbol{y}. \tag{18}$$

Additionally, from equation (17) and looking at the scalar products between the first row of matrix $\boldsymbol{X}^T$ and the $\hat{\boldsymbol{y}}$ and $\boldsymbol{y}$ vectors, the relation

$$\boldsymbol{1}^T \hat{\boldsymbol{y}} = \boldsymbol{1}^T \boldsymbol{y} \tag{19}$$

holds. This relation, together with the consideration of equations (12) and (13) leads to another well known result: in a MLR procedure, the mean values of the dependent variable and the fitted ones coincide,

$$\bar{\hat{y}} = \bar{y}. \tag{20}$$

In fact, it corresponds to the intersection value between equation (2) and the bisector.

Consequently, combining equations (18–20), the numerator and the denominator in 15 cancel and the expression becomes identically equal to 1

$$d = 1. \tag{21}$$

This equality holds except for ill-conditioned numerical data for which the fitted values cannot be properly calculated by model (1), for instance, when linear

dependencies arise during the calculations. Finally, as the mean values appearing in equation (20) satisfy the regression equation (3), its intercept at the origin is zero: $c = \bar{y} - d\bar{\hat{y}} = \bar{y} - \bar{\hat{y}} = 0$. This concludes the demonstration: equation (3) *always* coincides with the bisector of the first and third quadrants. As a consequence, equation (2) becomes equation (4), as $bd = r^2$.

Acknowledgments

E. B. acknowledges the financial support of the grant number BQU2003-07420-C05-01 of the 'Ministerio de Ciencia y Tecnología' within the Spanish Plan Nacional I+D. This grant also allowed visiting the University of Valencia, the place where this work was started. J. V. de J.-O. thanks the *'Red de Investigación de Centros de Enfermedades Tropicales'*, Ministry of Health, Spain. L. P. thanks Professors J. Gálvez and R. García-Domenech of the University of Valencia, Spain, for their kind hospitality and help. Professor R. Carbó-Dorca is also deeply acknowledged for suggesting very valuable comments during the preparation of this manuscript.

References

[1] L. Pogliani and J.V. de Julián-Ortiz, Chem. Phys. Lett. 393(4–6) (2004) 327.
[2] S. Yin, Z. Shuai and Y. Wang, J. Chem. Inf. Comput. Sci. 43 (2003) 970.
[3] M. Murcia-Soler, F. Pérez-Giménez, R. Nalda-Molina, M.T. Salabert-Salvador, F.J. García-March, R.A. Cercós-del-Pozo and T.M. Garrigues, J. Chem. Inf. Comput. Sci. 41 (2001) 1345.
[4] M.B. Kroeger Smith, B.M. Hose, A. Hawkins, J. Lipchock, D.W. Farnsworth, R.C. Rizzo, J. Tirado-Rives, E. Arnold, W. Zhang, S.H. Hughes, W.L. Jorgensen, C.J. Michejda and R.H. Smith, Jr., J. Med. Chem. 46 (2003) 1940.
[5] A. Micheli, A. Sperduti, A. Starita and A.M. Bianucci, J. Chem. Inf. Comput. Sci. 41 (2001) 202.
[6] C.A.S. Bergström, C.M. Wassvik, U. Norinder, K. Luthman and P. Artursson, J. Chem. Inf. Comput. Sci. 44 (2004) 1477.
[7] F. Ignatz-Hoover, R. Petrukhin, M. Karelson and A.R. Katritzky, J. Chem. Inf. Comput. Sci. 41 (2001) 295.
[8] D. Ostrovsky, M. Udier-Blagovic′ and W.L. Jorgensen, J. Med. Chem. 46 (2003) 5691.
[9] R. Guha and P.C. Jurs, J. Chem. Inf. Comput. Sci. 44 (2004) 1440.
[10] A.R. Katritzky, D.B. Tatham and U. Maran, J. Chem. Inf. Comput. Sci. 41 (2001) 1162.
[11] A. Linusson, J. Gottfries, T. Olsson, E. Örnskov, S. Folestad, B. Nordén and S. Wold, J. Med. Chem. 44 (2001) 3424.
[12] R. Liu and S.-S. So, J. Chem. Inf. Comput. Sci. 41 (2001) 1633.
[13] S.M. Muskal, S.K. Jha, M.P. Kishore and P. Tyagi, J. Chem. Inf. Comput. Sci. 43 (2003) 1673.
[14] A.C. Pierce and W.L. Jorgensen, J. Med. Chem. 44 (2001) 1043.
[15] J. Polanski, R. Gieleciak and M. Wyszomirski, J. Chem. Inf. Comput. Sci. 43 (2003) 1754.
[16] G. Tarzia, A. Duranti, A. Tontini, G. Piersanti, M. Mor, S. Rivara, P.V. Plazzi, C. Park, S. Kathuria and D. Piomelli, J. Med. Chem. 46 (2003) 2352.
[17] L. Pogliani and J.V. de Julián-Ortiz, MATCH Comm. Math. Comp. Chem. 53 (2005) 175.
[18] D.M. Hawkins, J. Chem. Inf. Comput. Sci. 44 (2004) 1.

[19] S.C. Peterangelo and P.G. Seybold, Int. J. Quantum Chem. 96 (2004) 1.
[20] A. Golbraikh, M. Shen, Z. Xiao, Y. Xiao, K.-H. Lee and A. Tropsha, J. Comput.-Aid. Mol. Des. 17 (2003) 241.
[21] D.M. Hawkins, S.C. Basak and D. Mills, J. Chem. Inf. Comput. Sci. 43 (2003) 579.
[22] J.V. de Julián-Ortiz, E. Besalú and R. García-Domenech, Ind. J. Chem. 42A (2003) 1392.
[23] A. Golbraikh and A. Tropsha, J. Mol. Graph. Mod. 20 (2002) 269.
[24] E. Besalú, J. Math. Chem. 29 (2001) 191.
[25] J.D. Jackson and J.A. Dunlevy, Statistician. 37 (1988) 7.
[26] E. Besalú, J.V. de Julián-Ortiz and L. Pogliani, MATCH Commun. Math. Comp. Chem. 55(2) (2006) 281.
[27] M.R. Spiegel, *Probability and Statistics* (McGraw-Hill, New York, 1975).
[28] N.R. Draper and H. Smith, *Applied Regression Analysis* (Wiley, New York, 1966) p. 174.