https://doi.org/10.26565/2220-637X-2018-30-01

УДК 544.169+519.237.5

METHODS FOR BUILDING LINEAR REGRESSION EQUATIONS IN THE "STRUCTURE-PROPERTY" PROBLEMS

M.I. Berdnyk[*,a], M.O. Onizhuk[*,b], V.V. Ivanov[*,c]

** V.N. Karazin Kharkiv National University, School of Chemistry, Department of Materials Chemistry, 4 Svobody sqr., 61022 Kharkiv, Ukraine*

a. e-mail: berdnikm@i.ua, ORCID: 0000-0002-0609-6088
b. e-mail: foxfixfaxono@gmail.com, ORCID: 0000-0003-0434-4575
c. e-mail: vivanov@karazin.ua, ORCID: 0000-0003-2297-9048

The application of different alternative approaches for building linear regression equations in tasks which are connected with description of physicochemical parameters of molecules has been described. The Ordinary Least Squares, the Least Absolute Deviation, and the Orthogonal Distances methods are among the chosen approaches. In tasks, connected with multicollinearity of predictor sets, the principle component regression and L_2-regularization have been applied. The special attention has been given to those approaches that made possible to reduce the number of predictors (the L_1-regularization, the Least Angles methods). In case of data with noticeable errors in both dependent and independent variables, the orthogonal distance method has been examined as an alternative to the least square approach. The adequacy of previously investigated least absolute deviation of orthogonal distances (LADOD) method has been demonstrated.

Keywords: The Least Squares Method, Least Absolute Deviations method, L_1–, L_2– regularization, The Principle Component Regression, Orthogonal Distance method, Physical-Chemistry molecular properties.

Introduction

More than two hundred years ago the *ordinary lest squares* (**OLS**) method, which is cornerstone of contemporary experimental investigations, has been developed in works of Gauss and Legendre (in the present article we treated the **OLS** as a simplest approach for building regression equation). Later, profound statistical justification of the **OLS** in conjunction with huge amount of experimental data demonstrated great significance of the **OLS** in descriptive and predictive tasks. The wide application of **OLS** in chemical science made it possible to construct a set of both purely phenomenological (correlational) and theoretically justified equations (*e.g.* [1]). The regression analysis plays a significant role in the construction of QSAR (*Quantitative structure-activity relationship*) equations. Such dependences allow to describe and predict the important physical-chemical characteristics and biological effect of molecular systems. A lot of regression equations which describe biological activity can be found for instance, in [2].

Of course, if a) the required equation is theoretically justified, b) the data contains set of linearly independent descriptors, c) the equation calibrated with the "sufficiently" sized training sample, and e) there is no significant "noise" in the data, then using the **OLS** provides an unambiguous solution of the regression analysis problem. However, in practice, there are much more data sets with a wide spread. In addition, a typical QSAR problem does not provide any reason to how many and which descriptors should be included in the desired equation. Thus, we have to deal with a redundant (multicollinear) descriptor set.

It should be noted that for the present day the statistical science offers alternatives to **OLS** approach. They are focused on robust estimations – stability in relation to outliers and multicollinearity. There are also the regression methods which aim to shrink the set of descriptors.

Some of these approaches are known for a long while. For instance, the *least absolute deviation* (**LAD**) first appeared in 1755, 50 years prior to **OLS** [3]! But it is surprising that even in present-day most of calculations of regression equations in chemistry are performed only with the **OLS** method. In addition, many of these approaches are not implemented in common statistical packages at all ! Thus, the possibilities of alternative models for regression equations constructing are still outside of the scope of chemists.

© Berdnyk M.I., Onizhuk M.O., Ivanov V.V., 2018

With this in mind, a package of computer programs with various approaches to construct regression equations was developed in the present work. We used the programming languages **FORTRAN** and **Python** for effective implementation of different methods. The calculations of different molecular parameters have been performed for illustrative purposes.

Methods for calculations of linear regression equations

In this section, we give a brief description of the methods used in the article. Detailed information can be found in original works (see references in the text). In general, the goal of constructing linear regressions is to find the coefficients of the following equation (β_k):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_m x_m \quad (1)$$

where x_1, x_2, … x_m are independent variables (predictors, descriptors), y is a single dependent variable (property, system's response). It is assumed that the equation (1) is calibrated according to the training (N-size) sample.

$$Y = \{y_i\}; \quad X = \left\{x_{i,1}, x_{i,2}, x_{i,3}, ..., x_{i,m}\right\}, \quad i = 1, ..., N \quad (2)$$

In the standard **OLS** method, task of finding the β_k coefficients is associated with minimization problem:

$$\beta_{OLS} = \arg\min_\beta \left\|Y - X\beta\right\|_2^2. \quad (3)$$

In this expression and below with symbol $\|\cdot\|_2$ we denote the Euclidean (L_2) norm. Expression (3) can be transformed to the well-known matrix representation (see, for example [4, 5]):

$$\beta_{OLS} = (X^+X)^{-1}X^+Y. \quad (4)$$

In the eq. (4) X^+ designates the transposition of matrix of predictors X.

The least absolute deviation method, **LAD**, is a more robust approach.

$$\beta_{LAD} = \arg\min_\beta \left\|Y - X\beta\right\|_1. \quad (5)$$

Here $\|\cdot\|_1$ is an absolute value (L_1-norm). The feature of **LAD** is an "automatic" adjustment of weights for certain data points. Thus, **LAD** can be interpreted as a "weighted" **OLS** method, but without the use of a priori information about data errors. Several algorithms for solving problem (5) are described in the literature [6].

In the present article we are using an iterative method called "variational-weighted quadratic approximations" [7,8], which is implemented in the matrix form:

$$\beta_{LAD} = \arg\min_\beta \left(Y^+ - \beta^+X^+\right)S^{-1}(\beta)\left(Y - X\beta\right), \quad (6)$$

where $S^{-1}(\beta)$ – pseudoinverse diagonal matrix.

$$S(\beta)_{ij} = \delta_{ij}\left|\beta_0 + \sum_{k=1}^{m}\beta_k x_{ik} - y_i\right|. \quad (7)$$

Obviously, the strict reason for applying the **LAD** method is the Laplace distribution of data errors. An important feature of **LAD** is the robustness of the method. However, it is necessary to acknowledge the drawbacks of the method. There are cases when multiple and degenerate solutions of LAD exist.

In the situations where the initial set of descriptors is deliberately redundant, Tikhonov's regularization (also known as **Ridge**-regression) can be used [9,10]. A special feature of the method is the presence of an additional factor in eq. (3) in the form of an L_2-norm $\|\beta\|_2^2 = \beta^+\beta$ (we will designate the method as **L_2-OLS**):

$$\beta(\lambda)_{L_2-OLS} = \arg\min_\beta \left\{\left\|Y - X\beta\right\|_2^2 + \lambda\left\|\beta\right\|_2^2\right\}. \quad (8)$$

The "strength" of the regularizing factor in (8) is determined by the parameter $\lambda \geq 0$. In this method, the problem of explicit (or not explicit) inversion of the matrix (X^+X) (4) is solved, even in the case when it is ill-conditioned or even degenerate. **L_2-OLS** approach makes it possible to obtain a closed expression for regression coefficients:

$$\beta(\lambda)_{L_2-OLS} = (X^+X + \lambda I)^{-1}X^+Y \quad (9)$$

In (9) I is an unity matrix.

An analog of **L_2-OLS** is the *principal component regression* (**PCR**) method [11]. Formally, the **PCR** is described with the same expression as the **OLS** (3,4). But the inversion of the matrix (X^+X) is

performed by using a *singular value decomposition* (SVD) of the matrix X. In these matrix manipulations we take into account only "sufficiently large" singular numbers of X (pseudoinversion). The **PCR** approach does not attempt to reduce the set of descriptors. In practice, in **PCR**, as well as in **L_2-OLS**, a "long", not easily visualized (and therefore difficult to analyze) equation is usually obtained. This equation can include thousands of terms in the form (1), which turns the method into a "black box" approach.

Also, we should note the distinctive features of **PCR** and **L_2-OLS**. In the **L_2-OLS**, a smooth deviation from the solutions of eq. (3) occurs with increase of the regularization parameter λ. In the **PCR**, the solution of eq. (3) changes discretely with removal of terms of the SVD of matrix X. The most common implementation of **PCR** ideology is the *partial least squares* (**PLS**) method [11,12,13]. Sometimes this abbreviation interpreted as *projection to latent structures*. The **PLS** takes into account the joint factor structure $\{X, Y\}$.

The **LASSO** (*Least Absolute Selection and Shrinkage Operator*) method [14] is an opposite to **PCR**.

$$\beta(\lambda)_{\text{LASSO}} = \arg\min_{\beta} \left\{ \|Y - X\beta\|_2^2 + \lambda\|\beta\|_1 \right\}. \qquad (10)$$

Function (10) is similar to (8), however here the regularization factor is an absolute value of regression parameters β, $\|\beta\|_1 = |\beta| = \beta^{+}\text{sign}(\beta)$. Such a regularization guarantees the shrinkage of descriptor set, when $\lambda > 0$. Detailed description of the **LASSO** and discussion on how and why such shrinkage can be achieved can be found in [15].

In the *elastic net*, **EN**, both (8) and (10) regularization factors have to be included to the minimization function [16]. This variant of regression is characterized by numerical stability in the initial stages of calculation, when the set of descriptors is still large and can be multicollinear.

$$\beta(\lambda,\alpha)_{\text{EN}} = \arg\min_{\beta} \left\{ \|Y - X\beta\|_2^2 + \lambda\left(\alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2^2\right) \right\} \qquad (11)$$

The parameters $\lambda > 0$ and $0 \le \alpha \le 1$ give control of the relative contributions of both L_1- and L_2- norms in function (11).

The *least angle regression and shrinkage* (**LARS**) [17,18] is a variant of forward stepwise regression [19]. The classical stepwise regression is a kind of so-called "greedy" algorithms which have several essential drawbacks. For instance, it cannot include several correlated variables to the regression. In general, the simple stepwise regression poorly takes into account the factorial structure of the problem. In the **LARS** method new predictors are included sequentially (step by step, starting from the simplest equation $y = \beta_0$), and these new predictors should be correlated with the remainder ($Y - X\beta$) to the same degree as those variables that have already been included in the regression. According to [17], the **LARS** algorithm does not lose in computational costs to **OLS**. The most important peculiarity of modified **LARS** is a possibility to obtain compact **LASSO**-solution. For this, an additional condition is included into the algorithm. While moving to the next predictor, if one of the coefficients already included to the model (say β_ℓ) changes its sign, the movement in this direction is canceled, β_ℓ is equated to zero, and the ℓ-th descriptor is excluded from the model (for the details see [17,18]). In the present article we are using this modified variant of **LARS**.

It should be noted that in all the above-mentioned regression models (including **OLS**) it is assumed that X is error free matrix of predictors. It is common when theoretical indices are used as the predictors and their values are absolutely determined. However, in the situations when both dependent and independent parameters are obtained from the experimental measurements (*Errors in Variables*, **EIV**), made with certain error, it is essential to use different specialized approaches.

One among them is the *total least squares* (**TLS**) which is general case of *orthogonal distances regression* (**ODR**) method. In the **ODR** method, the desired regression equation can be found by minimizing the sum of the Euclidean distances from the given points to the hyperplane determined by the regression equation (Fig. 1).

The well-known expression [20] allows one to obtain the form of a minimized **ODR**-function. In general, **ODR** can be implemented both within the frameworks of least squares (**ODR** as such)[*)]:

$$\beta_{\text{ODR}} = \arg\min_{\beta} \left\{ \|Y - X\beta\|_2^2 / \left(1 + \|\beta\|_2^2\right) \right\} \qquad (12)$$

[*)] *here data is autoscaled,* $\beta_0 = 0$.

and in least absolute deviation (*Least Absolute Deviation of Orthogonal Distances*, **LADOD)**:

$$\beta_{LADOD} = \arg\min_{\beta} \left\{ \left\| Y - X\beta \right\|_1 / \sqrt{1 + \left\| \beta \right\|_2^2} \right\} \tag{13}$$

The latter case, **LADOD**, has been investigated by us in [21].

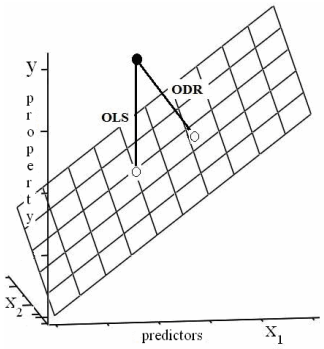

Figure. 1. Geometrical interpretation of difference between **OLS** and **ODR** (the figure corresponds to the equation $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$).

A remarkable peculiarity of **ODR** and **LADOD** is the presence of only one equation which connects the dependent and all independent variables. Unlike **ODR** and **LADOD** in the **OLS** for the regression (1), additionally to itself, it is possible to obtain *m* additional linear equations where the corresponding predictors take place of the dependent variable.

To evaluate the predictive ability of the obtained equations, we used the well-known formulas for the determination coefficients (for the discussion see, for instance, ref. [22]):

$$R^2 = 1 - \sum_i (y_i - \hat{y}_i)^2 \Big/ \sum_i (y_i - \bar{y})^2 \tag{14}$$

$$Q^2 = 1 - \sum_i (y_i - \hat{y}_{i/i})^2 \Big/ \sum_i (y_i - \bar{y})^2 \tag{15}$$

$$\theta = R^2 - Q^2 \tag{16}$$

where y_i – approximated values, $\bar{y}$ – mean value for sample $\{y_i\}$, $\hat{y}_i$ – calculated values which were obtained for training sample, $\hat{y}_{i/i}$ – «predicted» by *leave-one-out cross validation* (LOO-CV) procedure. Determination coefficient obtained by LOO-CV (Q^2), and θ are important parameters of predictive ability of regression model. Namely, the model is treated as successful when $Q^2 > 0.5$ and $\theta < 0.3$ [23]. For the detailed discussion of predictive ability of QSAR models see refs. [24,25].

Numerical Results

In the present article, before construction of the descriptor set, we optimized the geometry of the corresponding molecules (with semiempirical method AM1 from **GAMESS** package [26]). Next, a number of descriptors was calculated with the **PaDEL-Descriptor** program [27].

Ionization constants of carbonic acids

This problem has been considered as a first test case. To find the equation for pK_a ($pK_a = -\log K_a$, where K_a – acidity constant at equilibrium) as a function of structural parameters, 15 saturated carboxylic acids were selected:

HCOOH	CH_3COOH	C_2H_5COOH	C_3H_7COOH	$(CH_3)_2CHCOOH$
$CH_3(CH_2)_3COOH$	$(CH_3)_2CHCH_2COOH$	$(CH_3)_3CCOOH$	CH_2FCOOH	$CH_2ClCOOH$
$CH_2BrCOOH$	CH_2ICOOH	$CHCl_2COOH$	CCl_3COOH	CF_3COOH

Experimental values for pK_a (25°C) were taken from [28]. We selected 9 parameters as molecular descriptors: the charges on oxygen of the carbonyl (x_1, a.u.) and hydroxyl (x_2, a.u.) groups, on the hydrogen of hydroxyl group (x_3, a.u.), the surface area of the molecule (x_4, Å^2), its volume (x_5, Å^3), molar refraction (x_6, Å^3), polarizability (x_7, Å^3), Randic index (x_8) and informational index of routes in the graph of the molecule (x_9). Hence the equation for pK_a should be obtained from the most general expression:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_9 x_9 \quad (17)$$

The selection of the necessary descriptors, from these nine could be done manually from structural and chemical considerations. Let's see, however, how the **L_2-OLS** and **LASSO** approaches behave in this task.

By changing the parameter λ in the expressions (8) and (10), we obtain the dependences (Fig. 2) and (Fig. 3), respectively, which describe the changes of the regression coefficients.

For the sake of comparability of the **L_2-OLS** and **LASSO** data in both cases we show the dependence of the regression coefficients β_k on the norm $\|\beta\|_1 = |\beta|$. As one can see, with fairly strict limitations ($\|\beta\|_1 < 0.7$) in the **LASSO** method, only three descriptors out of nine survive – x_1, x_2, x_3 (Fig. 3). Here $\beta_1 \approx \beta_3$ and $|\beta_2| > |\beta_3|$.

Further increasing of λ, in the **LASSO** regression, leads to elimination of all but one parameter – x_2 (charge on oxygen of the hydroxyl group). Unlike **LASSO**, in the **L_2-OLS** method the values of all coefficients β_k decrease monotonically (Fig. 2). Obviously, the nature of changes β_k in the **L_2-OLS** method does not allow to make conclusion about the significance of a particular descriptor.

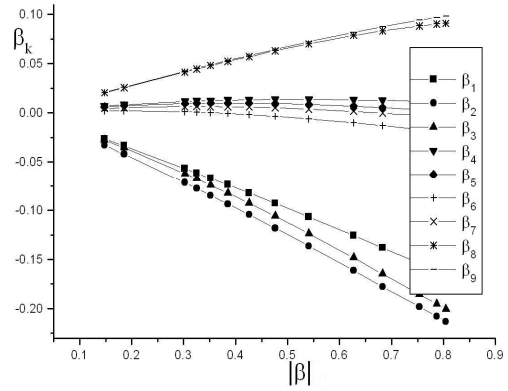

Figure. 2. Regression coefficients of **L_2-OLS** method in the problem of pK_a of carboxylic acids

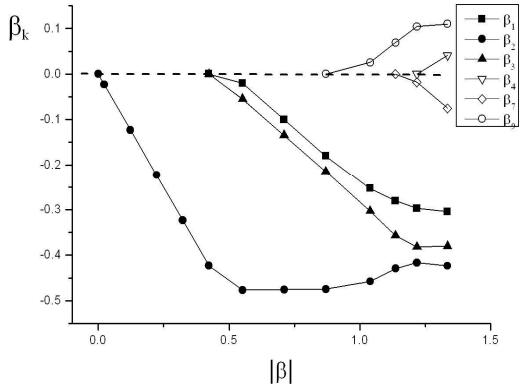

Figure. 3. Regression coefficients of **LASSO** method in the problem of pK_a of carboxylic acids

Thus, according to the **LASSO** the most important descriptor is x_2. The required equation using the **OLS** method has the form:

$$pK_a = -24.44 - 91.35x_2, R^2 = 0.852,\ s = 0.27,\ Q^2 = 0.805,\ \theta \approx 0.05, \quad (18)$$

while in the **LAD**:

$$pK_a = -20.53 - 79.06x_2,\ R^2 = 0.839,\ s = 0.28,\ Q^2 = 0.798,\ \theta \approx 0.04. \quad (19)$$

The equations (18) and (19) can be considered as satisfactory and consistent with each other, including proximity of standard deviations, s.

Let's check the equations which include three descriptors selected by **LASSO** (at $\|\beta\|_1 \approx 0.7$, see Fig. 3).

OLS: $$pK_a = -1.08 - 19.93x_1 - 46.80x_2 - 67.61x_3,\ R^2 = 0.971,\ s = 0.62,\ Q^2 = 0.746, \quad (20)$$

LAD: $$pK_a = -4.28 - 17.22x_1 - 55.12x_2 - 61.17x_3,\ R^2 = 0.969,\ s = 0.67,\ Q^2 = 0.201. \quad (21)$$

As we can see, although **OLS** is characterized by a rather good value of R^2, the predictive ability is noticeably worse than one of (18) with $\theta \approx 0.23$. At the same time, the **LAD** ($Q^2 = 0.201$) equation is completely inadequate. The poor quality of the **LAD** approach in this case requires additional research.

It is usual to see an increase in value of R^2 as the number of parameters increases. However, it is not associated with an enhance of the predictive ability of the equation. In the present example the one-parameter equation based on **OLS** (18), or **LAD** (19), should be considered as the best.

The ideology of **PCR** does not assume an explicit selection of descriptors. Instead of definite selection of the descriptors, in the **PCR** adjustable parameter is the number of singular values (n_s)

included in the **SVD** expansion. The results of **PCR** calculations for different n_s are presented in Fig. 4. It is clear from the picture that at $n_s = 1$ the **PCR** equation does not allow reliable estimates ($R^2 \approx 0.2,\ Q^2 \approx 0.1$). With increasing n_s to two, the predictive ability of the method is significantly enhanced. Further increase of n_s does not lead to the significant increase in value of Q^2.

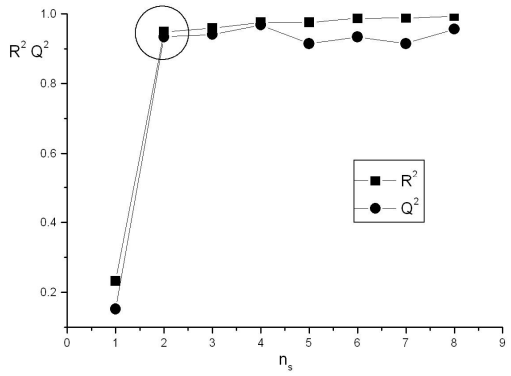

Figure. 4. pK_a of organic acids. The R^2 and Q^2 as a function of singular values numbers, n_s, in **PCR**.

Thus, the PCR method with only two singular numbers ($n_s = 2$, **PCR** (2)) gives the best regression equation. We are not presenting here the complicated **PCR** equation which includes 9 terms in the expansion. In this example, the merit of **LASSO** analysis is obvious simplicity of the resulting regression equation.

The boiling points of organic sulfides (thioethers)

For these calculations we used the training sample with 43 molecules of organic sulfides [29]. As the training sample contains the same type of molecules with different aliphatic residues, it can be assumed that to describe the *boiling point* (**BP**), only two-dimensional (2D) descriptors would be sufficient.

These quantities describe the order of bonding of atoms in a molecule – "molecular topology". We removed descriptors which have same values for all the molecules of the training sample. After removing the constant descriptors, a set of 501 descriptors was used in the calculations. L_1-regularized methods (**LASSO**, **LARS**, **EN**) allowed us to select the most statistically important values. Based on these descriptors, models were built within the frameworks of **OLS** and **LAD**. Our calculations showed that the use of single descriptor, namely MLFER_L (Solute gas-hexadecane partition coefficient) [30], is sufficient to describe **BP**. The corresponding equations have the form.

In **OLS:** $$\mathbf{BP}(^\circ\mathrm{C}) = -52.04 + 48.58\mathrm{MLFER\_L},\ R^2 = 0.982,\ Q^2 = 0.979, \quad (22)$$

In **LAD**: $$\mathbf{BP}(^\circ\mathrm{C}) = -49.31 + 47.75\mathrm{MLFER\_L},\ R^2 = 0.981,\ Q^2 = 0.981. \quad (23)$$

As one can see the equations are almost identical. Predictive ability of both equations are close to each other.

The **PCR** method in this task needed several singular values to achieve the values R^2 and Q^2 close to ones obtained in the **OLS** (22) and **LAD** (23) methods (Table. 1). Also, the **PCR** method leads to equation which includes the 501 terms in the eq. (1).

Table. 1. Determination coefficients R^2 and Q^2 as a function of number of accounted singular values (n_s) PCR.

n_s	R^2	Q^2
1	0.919	0.926
2	0.979	0.978
3	0.981	0.976
4	0.981	0.976
5	0.989	0.985

It is obvious that the one-parameter equations (22,23) are not unique. In order to find alternative and rather simple equations, we excluded the “good” descriptor (MLFER_L) and repeated the calculations. In Fig. 5 it is shown profiles of β_i changes in the **LARS** method.

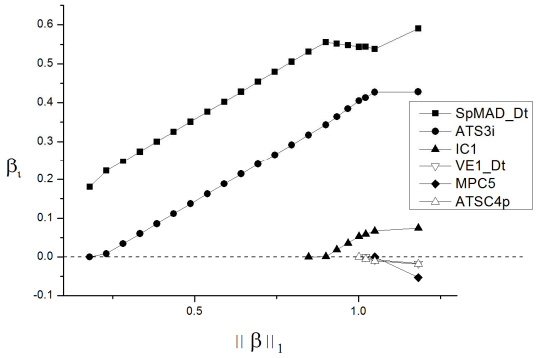

Figure. 5. **BP** thioethers. Profile of **LARS** regression coefficients.

As one can see the most important descriptors are SpMAD_Dt (*Spectral mean absolute deviation from detour matrix*), ATS3i (*Broto-Moreau autocorrelation - lag 3 / weighted by first ionization potential*) and IC_1 (*First-order Informational Contents Index*). Detailed information about these parameters can be found in **PaDEL-Descriptor** manual [27] and in the book [31].

The characteristic of the equations (**OLS** vs **LAD**) are presented in the Table. 2. As one can see the equations have a good predicting ability.

Table 2. Coefficients, R^2 and Q^2 values in alternative equations for thioethers’ **BP**. Methods **OLS** / **LAD,** m – number of descriptors in equation.

m	β_0	SpMAD_Dt	ATS3i	IC_1	R^2	Q^2
1	-13.02 / -17.54	33.83 / 34.48	–	–	0.961/0.958	0.956/0.955
2	1.23 / 2.01	20.98 / 21.99	$6.12\cdot10^{-3}$ / $5.27\cdot10^{-3}$	–	0.978/0.977	0.974/0.975
3	-35.26 / -35.71	18.47 / 19.14	$6.93\cdot10^{-3}$ / $6.77\cdot10^{-3}$	26.64 / 26.07	0.985/0.984	0.981/0.981

A graphical representation of the relationship “theory (LOO-CV) – experiment” for regression from **LAD** method (m = 3) is shown in Fig. 6. We don’t show corresponding plot for **OLS**, as it coincides with one for **LAD** (Fig. 6).

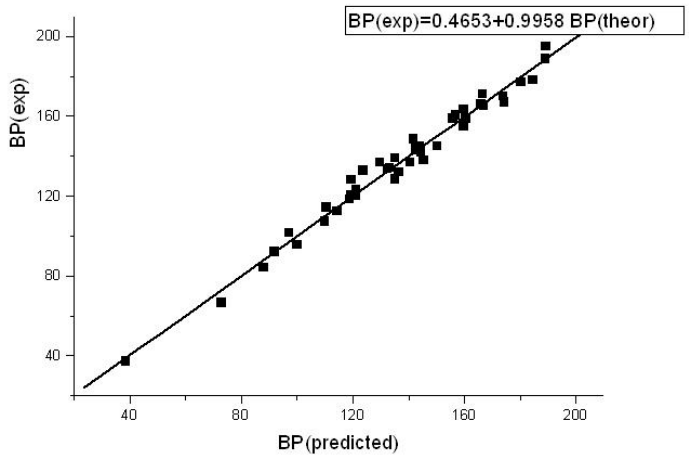

Figure. 6. Theoretical (**LAD**) and experimental values of thioethers’ **BP**.

To validate results obtained in calculation without descriptor MLER_L we selected test sample which consisted of 10 molecules, other 33 molecules were used as a training set to build models in PCR, OLS, and LAD methods.

We used equation (14) to calculate R^2 as the metric of external validation for the test sample. It appeared that in the worst case for PCR with the number of latent variables equal to one $R^2 = 0.875$. With higher amount of latent variables $R^2 \approx 0.97$. In OLS and LAD coefficient R^2 for the test sample appeared to be $R^2 \approx 0.9$ with one descriptor used in calculation. With increase of the number of descriptors used in calculation R^2 also tended to increase. Coefficients of internal validation almost did not change when we decreased the number of molecules in training set from 43 to 33.

Liquid viscosity and saturated vapor pressure of organic compounds

In this part we demonstrate application of orthogonal distance methods (**ODR** and **LADOD**). The correlation between two experimental parameters, viscosity ($\log\eta$) and saturated vapor pressure of organic compounds at the temperature 20 ºC is considered. Experimental data for 116 different organic molecules was taken from [32]. Brief analysis shows some level of correlation between these two parameters, and the biggest deviation from linear dependence is observed only in systems with high viscosity of the corresponding liquids. Of course, such simple dependences cannot be used to describe liquids with strong intermolecular interactions (*e.g.* containing strong hydrogen bonds). Nevertheless, chosen data shows weak linear relation between $\log\eta$ and $\log P$ (see Table 3 and Fig. 7). Equations from **LADOD** and **LAD** approaches are almost identical and notably different from ones from **OLS** and **ODR** (Fig. 7). These distinctions come from robustness of the **LADOD** and **LAD** methods. On the last note, even with weakly correlated data, **LADOD** shows stability towards LOO-CV procedure with $\theta \approx 0$.

Table 3. Regression coefficients and approximation criteria for dependence of $\log\eta$ (mPa·s) on $\log P$ (kPa) at T = 20 ºC.

Method	Regression coefficients		R^2	Q^2	θ
OLS	β_0	-0.0043	0.677	0.663	0.014
	β_1	-0.300			
LAD	β_0	-0.09	0.629	0.611	0.019
	β_1	-0.257			
ODR	β_0	0.0004	0.676	0.661	0.015
	β_1	-0.312			
LADOD	β_0	-0.0870	0.634	0.634	0.000
	β_1	-0.262			

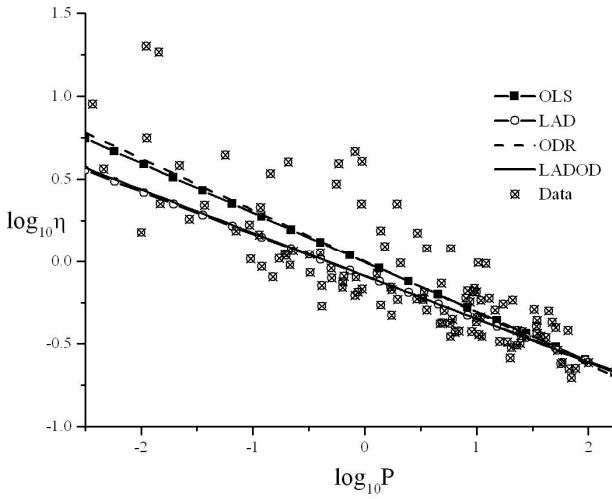

Figure. 7. Dependence of $\log\eta$ (mPa·s) on $\log P$ (kPa) at T = 20 ºC.

Evaluation of quality of nonempirical computations of phenols' pK_a

Standard approach to demonstrate accuracy of theoretical model is to graphically represent relation "theory-experiment" and show corresponding equations. In this part we show correlation between nonempirical computations of phenols' pK_a with experimental data. Theoretical and experimental data was taken from [33]. We chose two methods of pK_a estimations with different basis functions used in the quantum chemical computations:

Version	Neutral Molecule	Anion
a	CPCM/HF/6-31G(d)	CPCM/HF/6-31+G(d)
b	CPCM/HF/6-31+G(d)	CPCM/HF/6-31+G(d)

Regression model's computations are presented in Table 4, Table 5 and Fig. 8. Obviously, best "theory-experiment" relation corresponds to equation with intercept, equal to zero, and slope, equal to one:

$$y^{(\text{theor})} = y^{(\text{exp})} \tag{24}$$

Nonzero value of intercept tells about presence of systematic error, and deviation of slope from one characterizes discrepancy in quality of pK_a calculations of different molecules. From our calculations, **LADOD** shows high evaluation of accuracy in pK_a calculations, compared to other linear regression models. **LADOD** has minimal intercept β_0 and slope $\beta_1 = 1$, highest value of Q^2 and $\theta \approx 0$. **LAD** method, compared to **LADOD**, lowers accuracy of theoretical calculations of pK_a. **OLS** and **ODR** also hint on lower predicting ability of *ab initio* pK_a calculations.

Table 4. Regression coefficients and approximation criteria for dependence "theory-experiment" relation of pK_a values (version a).

Method	Regression coefficients		R^2	Q^2	θ
OLS	β_0	0.312	0.860	0.816	0.044
	β_1	0.970			
LAD	β_0	1.001	0.855	0.842	0.013
	β_1	0.898			
ODR	β_0	-0.423	0.855	0.800	0.055
	β_1	1.049			
LADOD	β_0	0.050	0.859	0.859	0.000
	β_1	1.000			

Table 5. Regression coefficients and approximation criteria for dependence "theory-experiment" relation of pK_a values (version b).

Method	Regression coefficients		R^2	Q^2	θ
OLS	β_0	0.318	0.877	0.833	0.043
	β_1	0.987			
LAD	β_0	0.504	0.867	0.864	0.003
	β_1	0.977			
ODR	β_0	-0.339	0.872	0.821	0.051
	β_1	1.058			
LADOD	β_0	0.290	0.868	0.867	0.000
	β_1	1.000			

Conclusion

In conclusion, it is worth to emphasize several important points concerning the regression analysis. Confronted with an abundance of approaches for construction of regression models, a naturally occurring problem is choice of appropriate model. Of course, this choice can be made based on a statistical investigation of the nature of the errors in the particular problem. Subsequent assessments of the significance of the regression coefficients and the equation as a whole (*t*-statistics, *F*-statistics), calculations of the studentized residuals, estimates of possible outliers, confidence intervals, *etc*., give

the most complete description of regression dependence. However, the reality of modern QSAR calculations suggests that a detailed analysis of the nature of the errors is usually impossible due to the limited data. In addition, the main criterion, characterizing prognostic ability of the equation, is the adequacy of the calculation's results with the training and, most importantly, test samples. Therefore, the methods of sample generation are intensively discussed in modern literature, *e.g.* LOO-CV, *Jack-knife*, *bootstrap* [34]. Following discussion on predictors selection, it is important to mention the criteria, based on theoretical-informational interpretation of statistical data. Among them are the information indices **AIC** (*Akaike Information Criterion*) and **BIC** (*Bayesian Information Criterion*) [35].

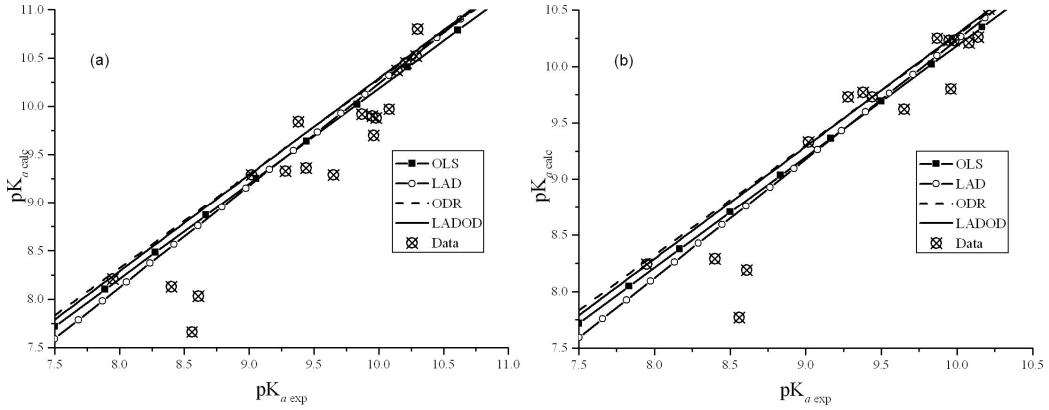

Figure. 8. Linear relationship between theoretical estimations of pKa and experimental values for two version of computations are shown.

Additionally, it should be noted that we did not feature all possible regression approaches that exist today, but included in the article only those which, in our opinion, constitute certain "reference points". The methods such as *Genetic Algorithms*, *Quasi Least Squares*, *Support Vector Machine Regression*, *Recursive Least Squares*, *Alternating Least Squares,* etc. have been out of consideration in the present article.

Speaking about the problem of choosing regression model in QSAR, we propose a pragmatic approach, partially demonstrated in this paper. Our approach is based on the fact that today's level of computer technology allows implementation and usage of different regression models simultaneously with low computational cost. Registration of significant discrepancies for the test sample calculations can serve as an indicator of necessity of an additional research of the problem. On the other hand, the identical results (within the limits of statistical significance) of analyses with different models indicate the effectiveness of the proposed equation.

Acknowledgement

The work has been carried out within the framework of the research project No. 0118U002025 (Ministry of Education and Science of Ukraine).

References

1. Reinhard M., Drefahl A. Handbook for Estimating Physicochemical Properties of Organic Compounds / New York, John Wiley & sons, inc. – 1999. – 238 p.
2. Kubinyi H. QSAR: Hansch Analysis and Related Approaches / New York, VCH. – 1993. – 240 p.
3. Statisticians of the centuries (eds. Heyde C. C., Seneta E.) / New York, Springer-Verlag. – 2001. – 500 p.
4. Demidenko E. Z. Lineynaya i nelineynaya regressii / M., Finansy' i statistika. – 1981. – 301 s. [in Russian]
5. Louson CH., Henson R. Chislennoe reshenie zadach metoda naimen'shih kvadratov / M., Nauka. - 1986. - 232 s. [in Russian]

6. Bloomfield P., Steiger W. L. Least Absolute Deviations. Theory, Applications and Algorithms / Boston, Birkhäuser. – 1983. – 349 p.
7. Mudrov V. I., Kushko V. L. Metod naimen'shih moduley / M., Znanie. - 1971. - 59 s. [in Russian]
8. Mudrov V. I., Kushko V. L. Metody' obrabotki izmereniy / M., Sovetskoe radio. - 1976. - 190 s. [in Russian]
9. Tikhonov A. N., Arsenin V. Y. Solutions of ill-posed problems / New York, John Wiley & Sons. – 1977. – 270 p.
10. Morozov V. A. Regulation Methods for ill-posed problems / New York, CRC Press. – 1993. – 273 p.
11. Geladi P., Kowalski B. R. // Analytica Chimica Acta. – 1986. – V. 185. – P. 1-17.
12. Handbook of Partial Least Squares (eds. Vinzi V. E., Chin W. W.) / New York, Springer-Verlag. – 2010. – 798 p.
13. New Perspectives in Partial Least Squares and Related Methods (eds. Abdi H., Chin W. W., Vinzi V. E., et. al.) / New York, Springer. – 2013. – 344 p.
14. Tibshirani R. // J. Roy. Statist. Soc. 1996. – B58, № 1. – P. 267–288.
15. Hastie T., Tibshirani R., Wainwright M. Statistical Learning with Sparsity. The Lasso and Generalizations / L., CRC Press. – 2015. – 335 p.
16. Zou H., Hastie T.// J. R. Statist. Soc. B. – 2005. – V. 67, Part 2. – P. 301–320.
17. Efron B., Hastie T., Johnstone I., Tibshirani R. // The Annals of Statistics. – 2004. –V. 32, № 2. – P. 407–451.
18. Tibshirani R. J. // Electronic Journal of Statistics. –2013. – V. 7. – P. 1456–1490.
19. Miller A., Subset Selection in Regression / New York, Chapman & Hall CRC. – 2002. – 234 p.
20. Rozenfel'd B. A. Mnogomerny'e prostranstva / M., Nauka. - 1966. - 547 s. [in Russian]
21. Onijuk N. O., Ivanov V. V., Panteleymonov A. V., Holin YU. V. // Methods and Objects of Chemical Analysis.- 2017.- V. 12, №. 3. - P. 105-111. [in Russian]
22. Ryan T. P. Modern Regression Methods / New York, Wiley. – 2008. – 672 p.
23. Veerasamy R., Rajak H., Jain A., Sivadasan S. // Int. J. Drug Design and Discovery. – 2011. – V. 2, № 3. – P. 511-519.
24. Golbraikh A., Tropsha A. // J. Mol. Graph. And Mod. – 2002. – V. 20. – P. 269-276.
25. Consonni V., Ballabio D., Todeschini R. // J. Chem. Inf. Model. – 2009. – V. 49. – P. 1669-1678.
26. Schmidt M. W., Baldridge K. K., Boatz J. A., Elbert S. T., et al. // J. Comput. Chem. – 1993. – V. 14, № 1. – P. 1347-1363.
27. Yap C. W. // Comput. Chem. – 2011. – V. 32, № 7. – P. 1466–1474.
28. Albert A., Serjeant E. P. The Determination of the Ionization Constants. A Laboratory Manual / London, Chapman & Hall. – 1984. – 218 p.
29. Zefirov N. S. and Palyulin V. A. // J. Chem. Inf. Comput. Sci. – 2001. – V. 41. – P. 1022-1027.
30. Platts J. A., Butina D., Abraham M. H., and Hersey A. // J. Chem. Inf. Comput. Sci. – 1999. – V. 39. – P. 835-845.
31. Todeschini R., Consonni V. Handbook of Molecular Descriptors / New York, Wiley-VCH Verlag. – 2000. – 667 p.
32. Suzuki T., Ohtaguchi K., Koide K. // Computers Chem. Eng. – 1996. – V. 20, № 2. – P. 161-173.
33. Liptak M. D., Gross K. C., Seybold P. G., et al // J. Am. Chem. Soc. –2002. – 124, P. 6421-6427.
34. Efron B., Tibshirani R. J., An Introduction to the Bootstrap / New York, Chapman & Hall CRC, 1993, 436 p.
35. Burnham K. P., Anderson D. R. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach / New York, Springer. – 1998. – 488 p.

Поступила до редакції 5 квітня 2018 р.

М.И. Бердник*, Н.О. Онижук*, В.В. Иванов*. Методы построения уравнений линейной регрессии в задачах «структура-свойство».

* Харьковский национальный университет имени В.Н. Каразина, химический факультет, площадь Свободы, 4, Харьков, 61022, Украина

Продемонстрировано применение ряда альтернативных подходов к построению уравнений линейной регрессии в задачах описания физико-химических параметров молекул. Среди рассмотренных подходов стандартный метод наименьших квадратов, метод наименьших модулей, методы ортогональных расстояний. В задачах связанных с мультиколлинеарностью в наборе предикторов рассматриваются метод регрессии главных компонент и L_2–регуляризация. Особое внимание уделяется подходам позволяющим сократить количество предикторов (L_1–регуляризация, метод наименьших углов). Для данных, содержащих погрешность и в зависимых и в независимых переменных, в качестве альтернативы стандартному методу наименьших квадратов, рассматривается метод ортогональных расстояний. Продемонстрирована адекватность исследованного ранее метода наименьших модулей ортогональных расстояний (LADOD).

Ключевые слова: метод наименьших квадратов, метод наименьших модулей, L_1–, L_2–регуляризация, регрессия главных компонент, метод ортогональных расстояний, физико-химические свойства молекул.

М.І. Бердник*, М.О. Оніжук*, В.В. Іванов*. Методи побудови рівнянь лінійної регресії в задачах «структура-властивість».

* Харківський національний університет імені В.Н. Каразіна, хімічний факультет, майдан Свободи, 4, Харків, 61022, Україна

Представлено застосування ряду альтернативних підходів до побудови рівнянь лінійної регресії в задачах опису фізико-хімічних параметрів молекул. Серед розглянутих підходів стандартний метод найменших квадратів (Ordinary Least Squares, OLS) та метод найменших модулів (Least Absolute Deviation, LAD). У завданнях пов'язаних із мультиколлінеарністю даних в наборі предикторів розглядаються методи регресії головних компонент (Principal Component Regression, PCR) і L_2-регуляризація (Ridge Regression). Особливу увагу приділяється підходам які дозволяють скоротити кількість предикторів: L_1-регуляризація (Least Absolute Selection and Shrinkage Operator, LASSO) та метод найменших кутів (Least Angle Regression and Shrinkage, LARS). Для даних що містять похибку і в залежних і в незалежних змінних, в якості альтернативи стандартному методу найменших квадратів, розглядається метод ортогональних відстаней (Orthogonal Distance Regression, ODR). У статті дано скорочений опис перерахованих методів побудови регресійних рівнянь та особливості їх використання. На прикладі задачі опису pK_a органічних карбонових кислот наведено техніку розрахунку методом LASSO. Отримані найпростіші рівняння, що описують pK_a як функцію параметрів електронного розподілу. Дано порівняння прогностичної здатності рівнянь для pK_a, що отримані у рамках OLS, LAD та PCR. На прикладі задачі щодо побудови регресійного опису температури кипіння органічних сульфідів встановлено кілька найпростіших OLS та LAD рівнянь їх прогностичну здатність, було порівняно із результатами PCR. В якості прикладу побудови регресійних рівнянь, що пов’язують експериментально знайдені величини, було досліджено залежності в’язкості від тиску насиченого пару органічних сполук. Для знаходження шуканих рівнянь було використано метод ODR та досліджений раніше авторами метод найменших модулів ортогональних відстаней (Least Absolute Deviation Orthogonal Distances, LADOD). В перерахованих проблемах, а також а задачах оцінки адекватності неемпіричних розрахунків pKa органічних кислот, було продемонстровано результативність методу ODR та LADOD.

Ключові слова: метод найменших квадратів, метод найменших модулів, L_1–, L_2–регуляризація, регресія головних компонент, метод ортогональних відстаней, фізико-хімічні властивості молекул.