Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2015, Article ID 360752, 11 pages
http://dx.doi.org/10.1155/2015/360752

Research Article

Inside of the Linear Relation between Dependent and Independent Variables

Lorentz Jäntschi,[1,2,3] Lavinia L. Pruteanu,[2] Alina C. Cozma,[3,4] and Sorana D. Bolboacă[4]

[1] *Institute for Doctoral Studies, Technical University of Cluj-Napoca, Muncii Boulevard 103-105, 400641 Cluj-Napoca, Romania*
[2] *Institute for Doctoral Studies, Babeş-Bolyai University, Kogălniceanu Street No. 1, 400084 Cluj-Napoca, Romania*
[3] *Department of Chemistry, University of Oradea, Universităţii Street No. 1, 410087 Oradea, Romania*
[4] *Department of Medical Informatics and Biostatistics, Iuliu Haţieganu University of Medicine and Pharmacy, Louis Pasteur Street No. 6, 400349 Cluj-Napoca, Romania*

Correspondence should be addressed to Sorana D. Bolboacă; sbolboaca@umfcluj.ro

Received 23 March 2015; Revised 21 April 2015; Accepted 21 April 2015

Academic Editor: Irini Doytchinova

Simple and multiple linear regression analyses are statistical methods used to investigate the link between activity/property of active compounds and the structural chemical features. One assumption of the linear regression is that the errors follow a normal distribution. This paper introduced a new approach to solving the simple linear regression in which no assumptions about the distribution of the errors are made. The proposed approach maximizes the probability of observing the event according to the random error. The use of the proposed approach is illustrated in ten classes of compounds with different activities or properties. The proposed method proved reliable and was showed to fit properly the observed data compared to the convenient approach of normal distribution of the errors.

1. Introduction

The quantitative structure activity/property relationships (QSARs/QSPRs) are computational techniques that quantitatively relate chemical feature (such as descriptors) to a biological activity or property [1]. Linear regression is one of the earliest methods [2] used to link the activity/property with structural information and is frequently used due to the relative easy interpretation [3]. Sometimes, linear regression is misuse due to the application without investigation of its assumptions (such as linearity, independence of the errors, normality, homoscedasticity, and absence of multicollinearity [4]).

The error, "a measure of the estimated difference between the observed or calculated value of a quantity and its true value" [5], was first used in mathematics/statistics in 1726 in *Astronomiae Physicae & Geometricae Elementa* [6]. In the late 1800's, Adcock [7, 8] suggested that the errors must pass through the centroid of the data. The method proposed by Adcock, named orthogonal regression, explores the distance between a point and the line in a perpendicular direction to the line [7, 8]. Kummell [9] investigated other than perpendicular directions between the points and line. The regression slope ("r") was described by Galton in 1894 based on an experiment of sweet pea seeds [10]. Two years later, Pearson generalized the errors in the variable and published a rigorous description of correlation and regression analysis [11] (Pearson recognized the contribution of Bravais [12] to mathematical formula of correlation). Due to the ability to produce best linear unbiased parameters [13], the coefficients in simple linear regression (SLR) models are estimated by minimizing the sum of squared deviations (least squares estimation, method introduced by Legendre in 1805 [14] and used/applied by Gauss in 1809 [15]). Furthermore, Fisher introduced the concept of maximum likelihood within linear models [16, 17].

The generic equation of simple linear regression (1) between observed dependent variable Y and observed independent variable X is:

$$Y \sim \widehat{Y} = a \cdot X + b, \tag{1}$$

where a and b are unknown constant values (estimators of statistics parameters of simple linear regression), $\widehat{Y}$ is the value of the dependent variable estimated by the model, Y is the observed value of dependent variable, and X is the observed value of the predictor variable.

The array use to estimate the residuals is given by $(Y_i - a \cdot X_i - b)^q$ formula, where i is the ith observation in the sample ($1 \leq i \leq n$, when n = sample size) and q is an unknown coefficient. The unknown q coefficient is an estimator of the power of the errors on simple linear regression.

In the SLR-LS (simple linear regression least squares), residuals ($S_i = Y_i - aX_i - b$, where S = residual) follow the Gauss-Laplace distribution with μ, σ, and q being unknown statistical parameters:

$$\mathrm{GL}(s;\mu,\sigma,q) = \frac{q}{2\sigma}\frac{\Gamma^{1/2}(3/q)}{\Gamma^{3/2}(1/q)}\exp\left(-\frac{|(s-\mu)/\sigma|^q}{(\Gamma(1/q)/\Gamma(3/q))^{q/2}}\right), \quad (2)$$

where μ is population mean, σ is population standard deviation, q is power of the errors, Γ is gamma function, and s is sample standard deviation.

Gauss-Laplace distribution is symmetrical and has three statistical parameters (population mean, population standard deviation, and power of the errors) [15, 18] and two main particular cases. First particular case is Gauss distribution [15] often observed on arrays of biochemical data [19–21] while the second particular case is Laplace distribution (with mean of zero and variance σ^2) [22, 23] commonly seen on astrophysical data [24, 25].

The problem of estimating the parameters of the SLR (1) for the first particular case (Gauss distribution) considers $q = 2$ residuals (where q is the power of the errors related with experimental errors). The coefficients of regression for this particular case are obtained by solving the system of linear equations under the assumption that $\sum {S_i}^2 = \min$ [26] ($\sum {S_i}^2 = \sum (Y_i - a \cdot X_i - b)^2$, where a and b are unknown parameters).

The second particular case is $q = 1$ when residuals follow the Laplace distribution. In view of the fact that $\sum |S_i| = \sum |Y_i - a \cdot X_i - b|$ "is not differentiable everywhere" [27], the solution in more difficult to be obtained for this particular case.

One question can be asked: "what is the proper value of q that should be used in the simple linear regression analysis (1)?" A previous study showed that, for different sets of biological active compounds, the distribution of the dependent variable (Y) can be approximated by Gauss distribution ($q = 2$) just in a relatively small number of cases when the whole Gauss-Laplace family is investigated [28]. Based on this result, the aim of the present study was to formulate the problem of solving the simple linear regression equation (1) without making any assumptions about the power of the errors (q).

2. Materials and Methods

2.1. Mathematical Approach. The problem of regression (1) is transformed into a problem of estimation if the residuals ($S_i = Y_i - a \cdot X_i - b$) are introduced in (2) with a slight modification: in the quantity $(Y_i - a \cdot X_i - b) - \mu$ the constants b and μ are equivalent and just one (b) will be further used. Gauss-Laplace distribution is symmetrical and the observed mean is an unbiased estimator of the population mean ($\mu = b$). This could be expressed in terms of (1) as presented in

$$M(Y) \sim M\left(\widehat{Y}\right) = a \cdot M(X) + b, \quad (3)$$

where b is the population mean of the Gauss-Laplace quantity $Y - a \cdot X$ (2), Y is observed/measured dependent variable, $\widehat{Y}$ is dependent variable estimated by the regression model, X is independent/predictor variable, and M is mean operator. For certain arrays of paired observations (X, Y), the problem of regression expressed in (1) is transformed to a problem of estimating the parameters of the bidimensional Gauss-Laplace distribution as presented in

$$\mathrm{GL}(x, y;\sigma,q,a,b) = \frac{q}{2\sigma}\frac{\Gamma^{1/2}(3/q)}{\Gamma^{3/2}(1/q)}\exp\left(-\frac{|((y - a\cdot x) - b)/\sigma|^q}{(\Gamma(1/q)/\Gamma(3/q))^{q/2}}\right). \quad (4)$$

An efficient instrument to solve (4) is maximum likelihood estimation (MLE), method proposed by Fisher [16, 17]. The main assumption of the MLE is that the (X, Y) array has been observed due to its higher chance to be observed (simultaneously and independent). This could be translated as $\mathrm{GL}(X_i, Y_i;\sigma,q,a,b) = \max$, and thus $\log(\Pi\mathrm{GL}(X_i, Y_i;\sigma,q,a,b)) = \max$, which lead to the expression in

$$\sum \log\left(\mathrm{GL}(X_i, Y_i;\sigma,q,a,b)\right) = \max. \quad (5)$$

By including (4) in (5) and using the natural logarithm, the problem presented in (1) became a problem of optimization:

$$\sum_{i=1}^{N} \ln\left(\mathrm{GL}(X_i, Y_i;\sigma,q,a,b)\right) = N \cdot \ln\left(\frac{q}{2\sigma}\frac{\Gamma^{1/2}(3/q)}{\Gamma^{3/2}(1/q)}\right) - \frac{\sum_{i=1}^{N}|(Y_i - a\cdot X_i) - b|^q}{\sigma^q((\Gamma\cdot(1/q))/(\Gamma\cdot(3/q)))^{q/2}} = \max., \quad (6)$$

where N is number of (X, Y) pairs.

The optimization problem presented in (5) could be iteratively solved if the start point is a good initial solution (situated near the optimal solution). In this research, the start

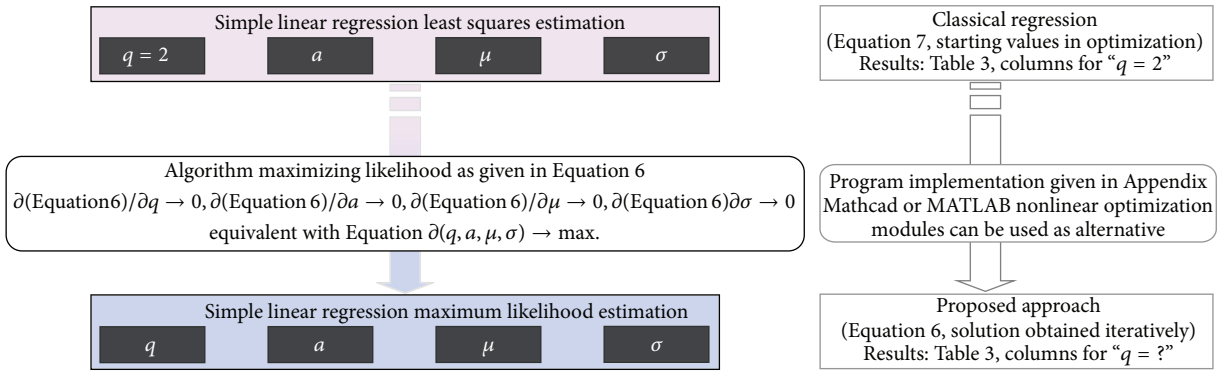

Figure 1: Flowchart of the implemented method. The starting values of the "a" (coefficient of the independent variable), "μ" (population mean), and "σ" (population standard deviation) coefficients are those obtained by least squares estimation method while the imposed value of power of the errors is equal to 2. The algorithm that maximizes likelihood finds optimal solution for "q," "a," "μ," and "σ" that satisfy (6).

point in the optimization was the solution of a particular case of (6) as presented in

$$\begin{aligned} q &= 2; \\ a &= \frac{(M(XY) - M(X) \cdot M(Y))}{D^2(X)} \\ \mu &= M(Y) - a \cdot M(X) \\ \sigma &= \left(\frac{D^2(Y) - (M(XY) - M(X) \cdot M(Y))^2}{D^2(X)} \right)^{1/2}, \end{aligned} \tag{7}$$

where q is power of the errors, μ is population mean, σ is population standard deviation, M is average (central tendency operator), and D^2 is variance (dispersion operator).

2.2. Algorithm Implementation. The classical simple linear regression uses least squares method to estimate a, μ, and σ coefficients in (7) using the fixed values of 2 for the power of the errors ($q = 2$). In our approach, starting with the optimal solutions for a, μ, and σ coefficients obtained by (7), the optimal solution of (6) was iteratively obtained by making small changes to the values of the coefficients and selecting the coefficients that make the MLE value higher. The implemented weights of changes were more or less arbitrary, and the selected ones are a compromise of convergence speed in the convergence space.

The flowchart of the proposed approach is presented in Figure 1.

A PHP program was developed to find the optimal solution for (6). As the input data, the implemented program needs a $*$.txt file with three columns (file named as mol-X-Y, where *mol* is the identification of the molecule and could be text or number, X is the independent variable, and Y is dependent variable). The program generates the output file as specified by the user (a $*$.txt file could be used) that contains for each iteration the data for the following coefficients: q, a, μ, σ, and *MLE*.

The source code of the implemented algorithm is free to be used and is presented in the Supplementary Material available online at http://dx.doi.org/10.1155/2015/360752. The full program can be obtained upon request from the authors.

2.3. Data Sets. Ten classes of previously investigated compounds were used to assess the proposed method. The class of compounds, the activity/property of interest along with the number of compounds in the dataset and the reference to the paper from where the independent and dependent variables were collected are given in Table 1.

Simple linear regression (SLR) models under the assumption of linear relationship between structural descriptors and activity/property of chemical compounds were identified using the values of descriptors previously published in the literature (see reference in Table 1). The characteristics of the models with the highest goodness-of-fit for each class of compounds are presented in Table 2.

3. Results and Discussion

The proposed solution for solving the simple linear regression without making any assumptions about the power of the errors has been successfully implemented and reliable solutions were obtained.

The developed algorithm was successfully tested on ten different data sets. The number of iteration needed to find the optimal solution varied from 9 (set10) to 185 (set4b) and seems not related with the number of compounds in the sample when the same class of compounds is investigated (63 iterations (set1a), 51 iterations (set1b), and 86 iterations (set1c)). The number of iterations needed to obtain the optimal solution was equal to 173 for the smallest dataset (set2) and 86 for the dataset with the highest number of compounds (set1c). Accordingly, the maximum number of iterations was almost 21 times more than the minimum number of iterations.

The results of simulation study obtained for the convenient solution ($q = 2$, residual follows the Gaussian distribution) and for solution that satisfies (6) are presented in Table 3. The values of calculated coefficients (a, b, and σ) are provided with three decimals; equal values for $q = 2$ and

Table 1: Characteristics of the investigated classes of compounds.

Set	n	Class	Activity/property, expressed as	Reference
1a	35	Phenols	Toxicity on *Tetrahymena pyriformis*, $\log(1/\mathrm{IGC}_{50})$	[29–31]
1b	126			
1c	250			
2	24	Organic compounds	Solubility, $\log P$	[32, 33]
3	73	Alkanes	Boiling point, BP	[34]
4a	40	Flavonoids	Solubility, $\log P$	[35]
4b	30		Lethal Dose 50%, $\ln(\mathrm{LD}_{50})$	
5	132	Estrogen receptor (ER)	Binding affinities, log(RBA)	[36]
6	80	Pyrrolo-pyrimidine derivatives	c-Src tyrosine kinase inhibitory activity, $\mathrm{pIC}_{50} = -\log_{10}(\mathrm{IC}_{50})$	[37]
7	47	Substituted aromatic sulfonamides	Inhibition activity on carbonic anhydrase II, $\log K_I$	[38]
8	37	Carboquinone derivatives	Molar concentration, log(1/MC)	[39]
9	47	Dipeptides	ACE (angiotensin converting enzyme) inhibitory activity, ACE	[40]
10	60	Mycotoxins compounds	Retention time, ln(RT)	[41]

Table 2: Characteristics of the SLR-LS models used in the optimization study.

Set	SLR model	R^2	s	F	n
1a	$\log(1/\mathrm{IGC}_{50}) = +0.677 \cdot \log P - 1.38$	0.90	0.22	287	35
1b	$\log(1/\mathrm{IGC}_{50}) = +0.647 \cdot \log P - 1.05$	0.84	0.30	666	126
1c	$\log(1/\mathrm{IGC}_{50}) = -0.443 \cdot \log P + 0.509$	0.53	0.57	276	250
2	$\log P = -0.004 \cdot \mathrm{ISDRTHg}^* + 2.09$	0.53	0.43	25	24
3	$\mathrm{BP} = +188.40 \cdot \mathrm{lbMdsHg}^* - 507.95$	0.99	3.81	8050	73
4a	$\log P = +0.99998 \cdot \mathrm{SD} + 5.232$	0.71	0.32	92	40
4b	$\ln(\mathrm{LD}_{50}) = +0.0018 \cdot \mathrm{SD} - 61.168$	0.41	0.98	19	30
5	$\log \mathrm{RBA} = +0.026 \cdot \mathrm{TIC1} - 4.145$	0.36	1.44	72	132
6	$\mathrm{pIC}_{50} = +0.255 \cdot \mathrm{DCW} - 1.216$	0.71	0.57	191	80
7	$\log K_I = -0.578 \cdot N\text{-rings} + 2.646$	0.49	0.37	43	47
8	$\log(1/\mathrm{MC}) = -4.129 \cdot \mathrm{TEuIFFDL}^* + 5.789$	0.65	0.38	64	37
9	$\mathrm{ACE} = 47.5480 \cdot \mathrm{IHMdpMg}^* - 0.1687$	0.74	0.33	128	47
10	$\ln(\mathrm{RT}) = 0.348 \cdot \log P + 1.711$	0.56	0.50	75	60

SLR = simple linear regression.
$\log(1/\mathrm{IGC}_{50})$ = concentrations (expressed as mM) producing a 50% growth inhibition on *T. pyriformis*.
*MDF descriptors [33, 39, 40, 42].
SD = global correlation descriptor [35]; TIC1 = total information content index (neighborhood symmetry of 1-order).
DCW = flexible (activity dependent) descriptor.
std_dim3 = the square root of the third largest eigenvalue of the covariance matrix of the atomic coordinates [43].
R^2 = determination coefficient; s = standard error of the estimate.
F = Fisher's statistic of the regression model; n = sample size.

optimal q were obtained as follows: a, coefficient in set1b, set3, and set6; b, coefficient in set3, set6, set8, and set10; and σ, coefficient in the following sets: 1b, 1c, 3, 4a, 5, 6, 8, 9, and 10.

The analysis of the obtained coefficient presented in Table 3 revealed the following.

(i) In 9 out of 13 cases, at least one coefficient (a, b, or σ) proved equal for convenience; $q = 2$ and q is determined to satisfy (6).

(ii) In 6 out of 13 cases, the power of the errors obtained by MLE proved significantly higher than 2. The difference varied from 0.8099 (set4a) to 7.5176 (set1a).

(iii) Just in one case, the difference between powers of the errors proved not statistically different (set3, $P = 0.0693$).

(iv) In 6 out of 13 cases, the difference between power of the errors (SLR-LS and SLR-MLE) proved lower than 1.

(v) The smallest distance between the powers of the errors (from SLR-LS and SLR-MLE) was of 0.2613 (set10) and was identified as being statistically significant ($P < 0.0001$).

(vi) Two classes of compounds (set3 and set6) proved identical values of a, b, and σ unconcerned with the method used in the regression analysis (SLR-LS and SLR-MLE).

Table 3: Optimization results: $q = 2$ versus q determined to satisfy (6).

set	n	$q = 2$			$q = ?$				P value (H_o: $q = 2$)
		a	$b = \mu$	σ	q	a	$b = \mu$	σ	
1a	35	0.678	−1.386	0.218	9.52	0.638	−1.181	0.222	$4.20 \cdot 10^{-54}$
1b	126	0.647	−1.050	0.298	4.36	0.647	−1.029	0.298	$3.07 \cdot 10^{-115}$
1c	250	0.509	−0.443	0.596	1.29	0.563	−0.623	0.569	$2.42 \cdot 10^{-53}$
2	24	−0.004	2.095	0.414	0.61	−0.005	2.270	0.516	$1.76 \cdot 10^{-12}$
3	73	188.408	−507.959	3.762	1.34	188.408	−507.959	3.762	$6.93 \cdot 10^{-2}$
4a	40	1.000	5.232	0.308	2.81	1.041	5.338	0.308	$1.30 \cdot 10^{-19}$
4b	30	0.002	−61.168	0.945	0.67	0.002	−64.950	0.964	$1.16 \cdot 10^{-8}$
5	132	0.024	−3.812	1.374	1.70	0.026	−3.967	1.374	$7.33 \cdot 10^{-3}$
6	80	0.255	−1.216	0.558	2.87	0.255	−1.216	0.558	$3.39 \cdot 10^{-23}$
7	47	−0.578	2.646	0.360	3.43	−0.555	2.594	0.353	$1.06 \cdot 10^{-30}$
8	37	−4.129	5.789	0.372	1.29	−4.297	5.789	0.372	$4.75 \cdot 10^{-14}$
9	47	47.561	−0.169	0.319	3.17	49.502	−0.279	0.319	$9.01 \cdot 10^{-29}$
10	60	0.348	1.711	0.492	1.74	0.355	1.711	0.492	$6.09 \cdot 10^{-5}$

q = power of the errors; a, b = coefficients in the simple linear model.
μ = population mean; σ = population standard deviation.

(vii) The q obtained by SLR-MLE proved significantly different by convenient value ($q = 2$) with one exception represented by set3.

The most probable distribution of the power of the error obtained by MLE is Fatigue Life or Birnbaum-Saunders distribution [44] (Kolmogorov-Smirnov statistics = 0.1245, $P = 0.9728$; Anderson-Darling statistics = 0.2753 $P = 0.9509$; P value associated with Anderson-Darling statistics was calculated taking into account the values of the statistics and the sample size [45]). The Fatigue Life distribution of the power of the errors is characterized by two parameters represented by continuous shape parameter ($\alpha = 0.7777$) and continuous scale parameter ($\beta = 2.0599$). The median of the power of the errors is closed to the convenient values of 2, with a mean of 2.68. Nevertheless, the normal distribution of the obtained power of the errors could not be rejected at a significance level of 5% (Kolmogorov-Smirnov statistics = 0.278, $P = 0.2229$; Anderson-Darling statistics = 1.178, $P = 0.2731$).

The evolution of value of power of the errors according to iteration was in both directions and, as expected, never achieved negative values (see Figure 2). The analysis of the evolution of the power of the errors as function of iteration revealed that even if identical values of q are obtained in the first 29 iterations for the first two related samples (set1a and set1b, Figure 2), the pattern is not representative for the class of the compounds. Thus, the pattern from 1c is significantly different by those observed on subsets of the whole class of compounds (1a and 1b). Opposite behavior is also observed for the other two related samples (set4a and set4b), and the value of q increased until a maximum (iteration 10 for set4a) and decreased after this value while the value of q decreases in steps for set4b.

Overall, two distinct patterns are observed in Figure 1. In the first pattern, the values of power of the error increase with iteration until a peak and after that the value decreases (sometimes with a decrease in steps (set6, set7, and set9)); see set1a, set1b, set4a, set6, and set9 (Figure 2). In the second pattern, the power of the error decreases in steps with the increase of iteration as for set1c, set2, set3, set4b, set5, set8, and set10 (see Figure 2).

The plot of both regression lines (simple linear regression and associated 95% confidence interval and MLE regression) for each investigated data sets is presented in Figure 3.

The analysis of the regression lines presented in Figure 2 revealed that, in one case represented by set7, the assumption of the linearity of $\log K_I$ with *n-rings* is breached and, for this dataset, the simple linear regression is not the proper analysis. In 4 out of 13 cases, the SLR-MLE line is partly outside the 95% confidence boundaries of the SLR-LS line (set1a, set1c, set2, and set4b; Figure 3). Accordingly, it could be considered in all these cases that the SLR-MLE model is significantly different by the SLR-LS model. The overlapping of SLR-MLE and SLR-LS line is observed for the set3, without being possible to make a visual distinction between them (Figure 3). For this set, the q obtained by SLR-MLE was equal to 1.34 and proved not significantly different by convenient value of 2 (see Table 3). For all other sets, the SLR-MLE line is within the boundaries of 95% confidence intervals of SLR-LS line and thus even if the powers of the errors proved significantly different by the convenient value of 2, these SLR-MLE models could not be considered significantly different by the SLR models.

To conclude, it is certain that the proposed approach of maximizing the probability of observing the event according to the random error fits well the observed data and frequently the power of the errors (q) is significantly different by the convenient value ($q = 2$). However, no pattern could be identified between iteration and sample size on the investigated sets of (X, Y) pairs. It is expected that the recognized behavior of the power of the errors is to be identified on other (X, Y) pairs, analysis which is currently conducted by our team. The relation presented in (6) thereby defines a new

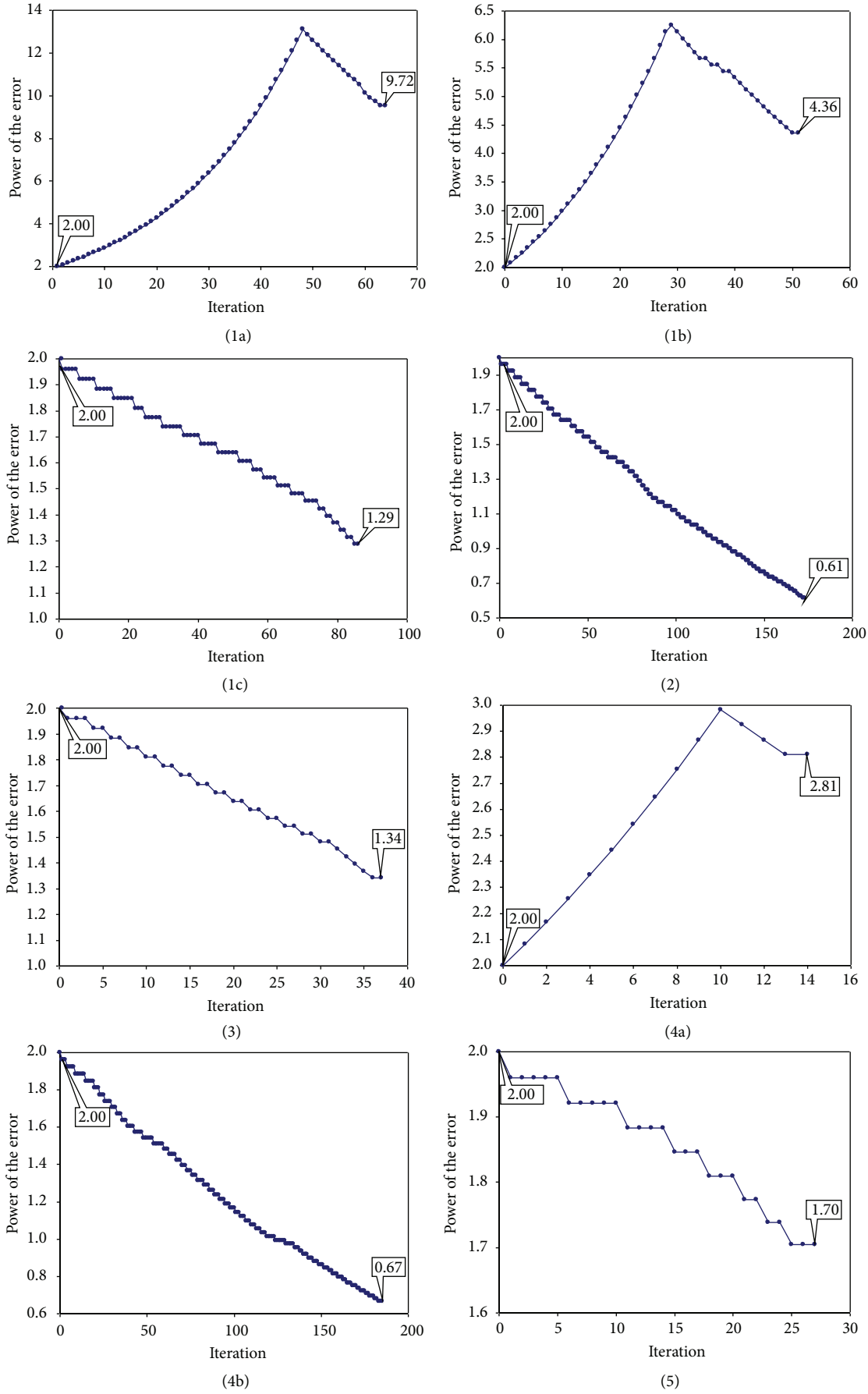

Figure 2: Continued.

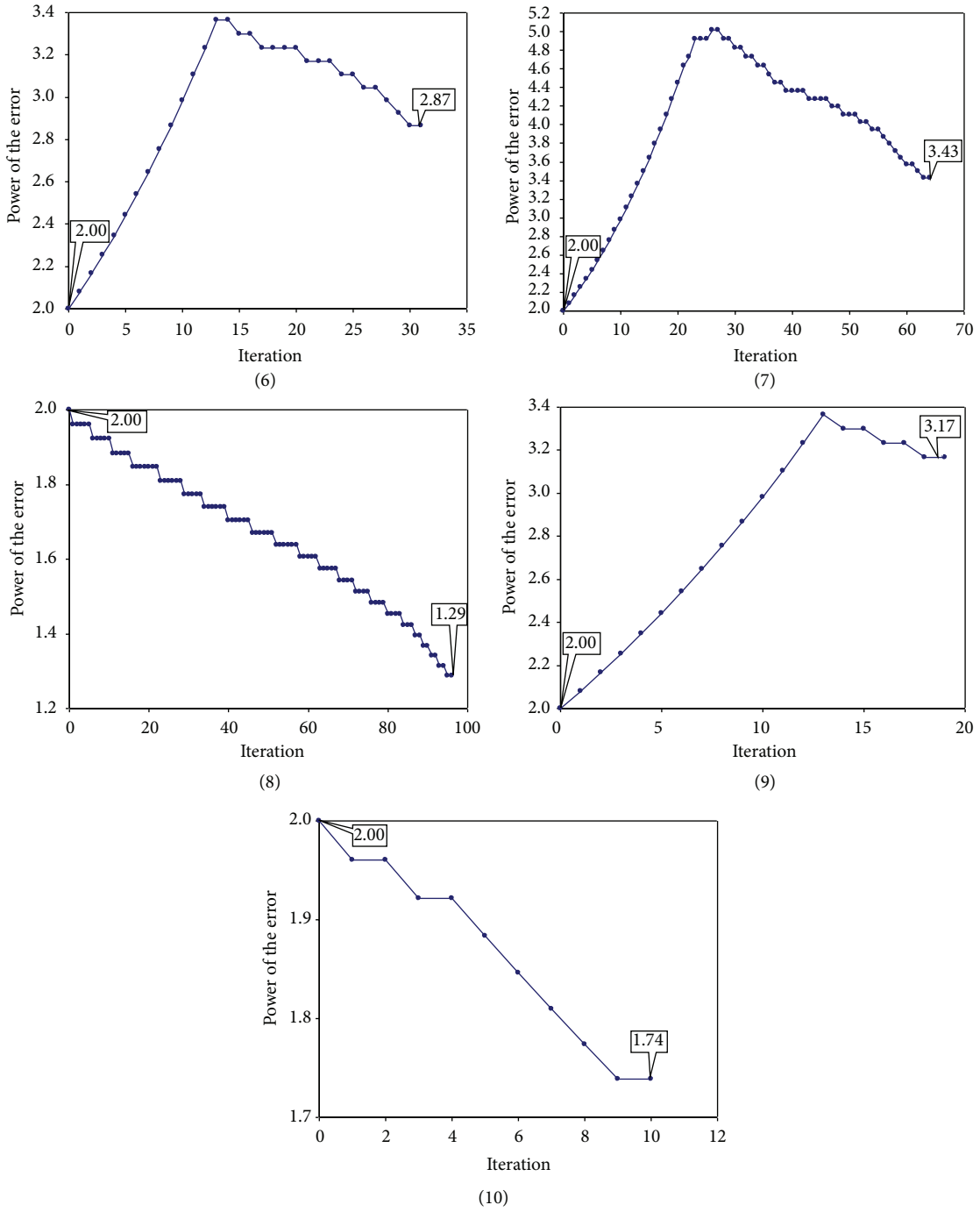

Figure 2: Distribution of power of the errors according to iteration: investigation of phenols set (35 compounds (1a) and 126 compounds (1b), resp.). Distribution of power of the errors according to iteration: phenols (1c), organic compounds (2), alkanes (3), flavonoids (4a and 4b), estrogen receptor (5), pyrrolo-pyrimidine derivatives (6), and substituted aromatic sulfonamides (7). Distribution of power of the errors according to iteration: behavior on carboquinone derivatives (8), dipeptides (9), and mycotoxins compounds (10).

general approach to treat the relationships. Practically, the expression $S_i = Y_i - aX_i$ could be replaced with any expression of dependency (not just linear), such as

(i) exponential: $S_i = Y_i - a_1 \cdot \exp(-X_i/a_2)$ for $Y \sim a_0 + a_1 \cdot \exp(-X/a_2)$;

(ii) double exponential: $S_i = Y_i - a_1 \cdot \exp(-X_i/a_2) - a_3 \cdot \exp(-X_i/a_4)$ for $Y \sim a_0 + a_1 \cdot \exp(-X/a_2) + a_3 \cdot \exp(-X/a_4)$;

(iii) power: $S_i = Y_i - a_1 \cdot \text{pow}(X_i, a_2)$ for $Y \sim a_0 + a_1 \cdot \text{pow}(X, a_2)$;

(iv) inversed: $S_i = Y_i - a_1/(X_i - a_2)$ for $Y \sim a_0 + a_1/(X - a_2)$.

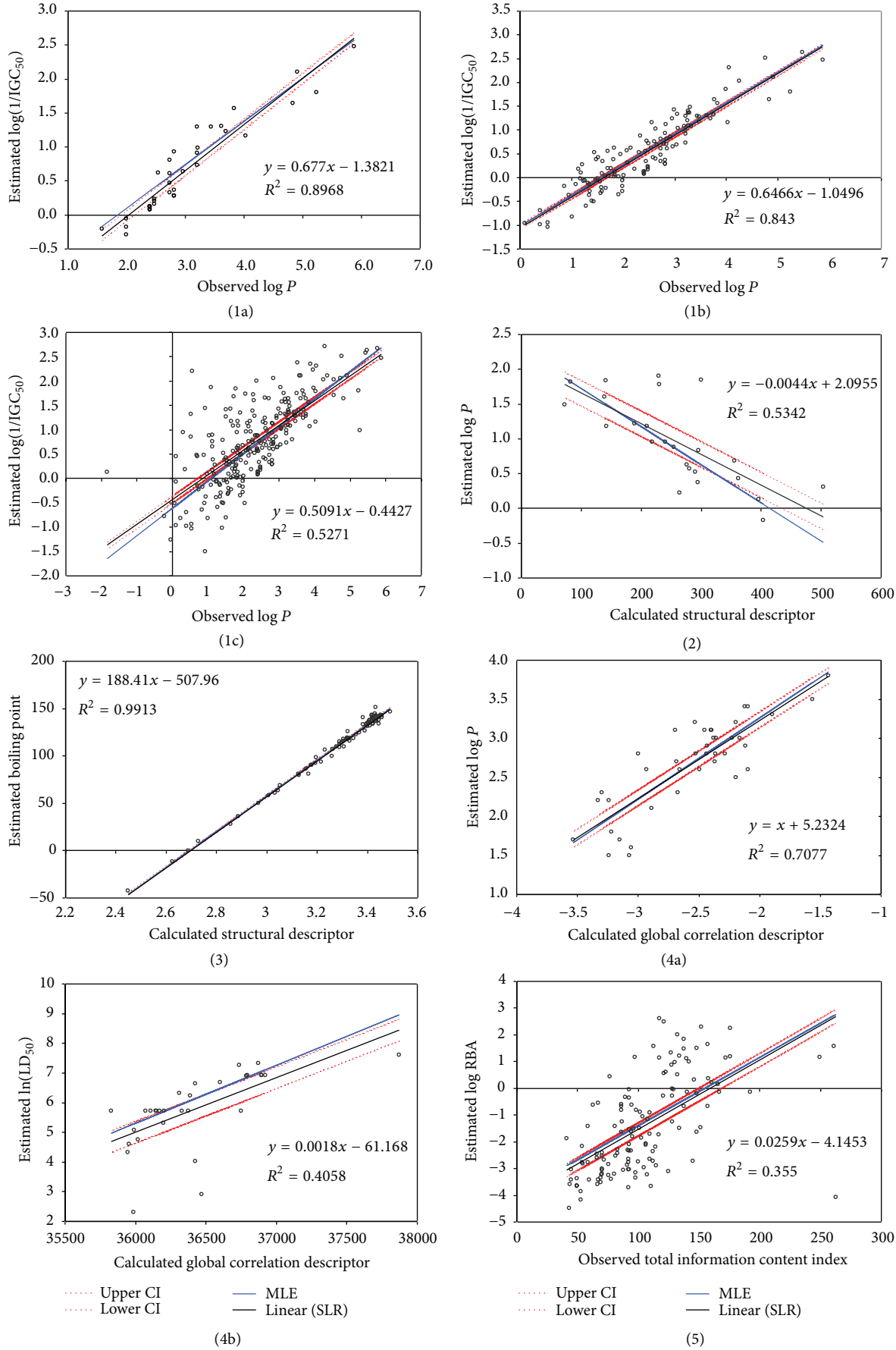

Figure 3: Continued.

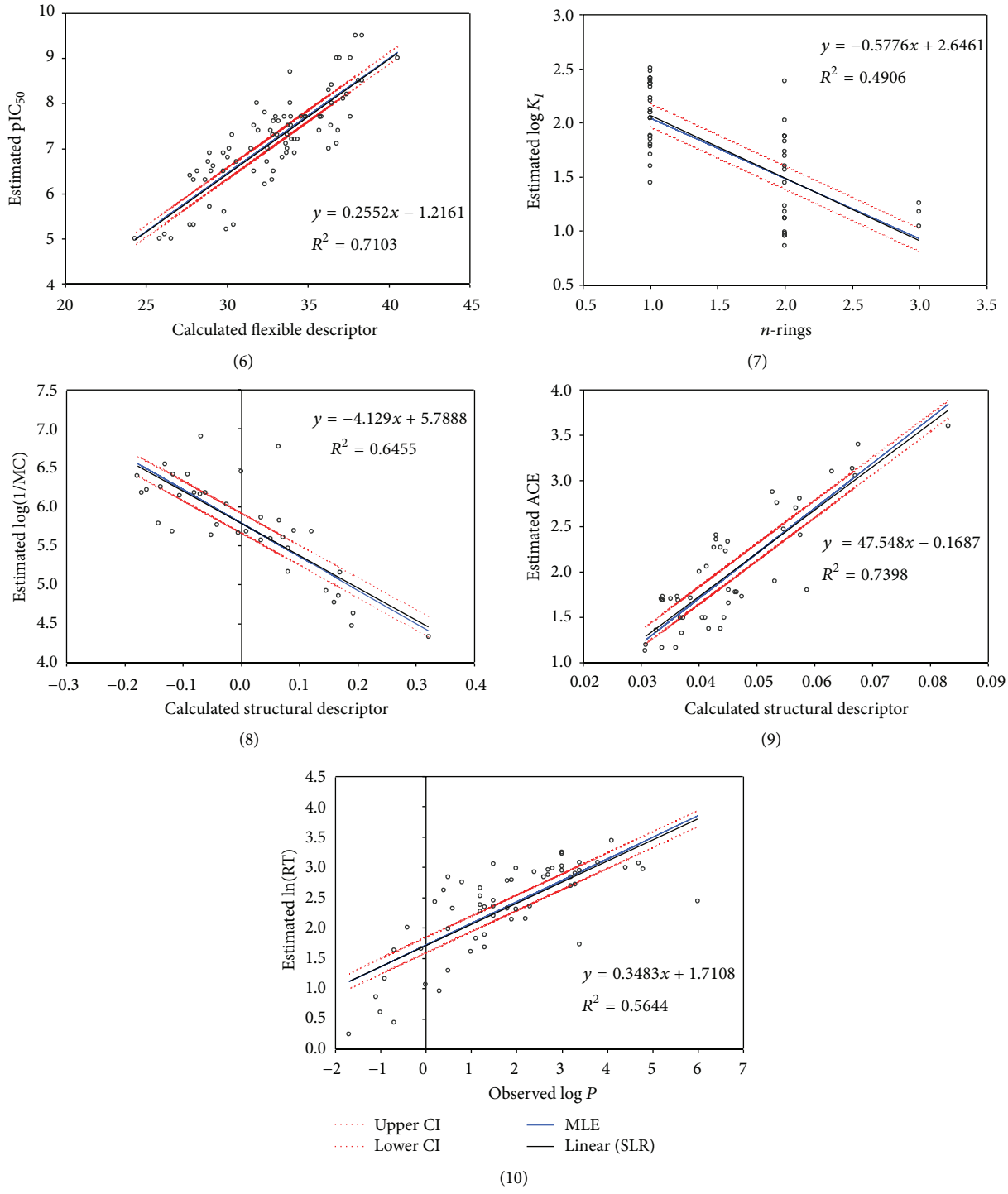

FIGURE 3: The line of SLR-LS ($q = 2$) and SLR-MLE (q determined to satisfy (6)): investigation of phenols set (35 compounds (1a) and 126 compounds (1b), resp.). Phenols (1c), organic compounds (2), alkanes (3), flavonoids (4a and 4b), estrogen receptor (5), pyrrolo-pyrimidine derivatives (6), and substituted aromatic sulfonamides (7). Carboquinone derivatives (8), dipeptides (9), and mycotoxins compounds (10).

The relation presented in (6) may be also extended to the multiple linear regression ($Y \sim a_0 + \sum_{j>0} a_j X_j$) when the expression $S_i = Y_i - aX_i$ becomes $S_i = Y_i - \sum_{j>0} a_j X_{j,i}$. If in the case of multiple linear regressions the classical method (minimizing the squared error) maximizes the correlation coefficient, the proposed approach (6) maximizes the probability of observing the event according to the random error. In view of that, (6) has a significant advantage compared to the classical approach. The classical approach that maximizes the correlation coefficient is exposed to type I errors; a model

of regression could be accepted even if the model does not exist. On the contrary, the proposed approach that maximizes just the chance of observation (the approach has just one hypothesis: the error between the observation (Y) and the model ($\widehat{Y}$) must be random and its value does not depend on the size of the observed value) is not affected by a type I error. In the case of simple linear regression, application of (6) did not change the correlation coefficient between Y and $\widehat{Y}$ but offers a solution in regard to estimated valued of Y and of the unknown coefficients (estimators of the population coefficients) that enter the relation between X and Y. The relation proposed in this paper (6) introduced an additional parameter in the estimation, namely, the power of the errors of Gauss-Laplace distribution (q) (this led to decrease by one unit of the degrees of freedom in the analysis of variance in the regression model).

The MLE approach is frequently used in estimation of unknown parameters and it is known to be sensitive to outliers (±influential compounds) in the data [46–48]. No outliers have been identified in the dependent variable on set2 and set3 [42, 46, 47]. Therefore, on these two sets of compounds, it is a certainty that the proposed approach was not affected by the presence of outliers in the data. Evaluation of how the values in the investigated sets could lead to identification of outliers (±influential compounds [4, 31, 49]) was beyond the aim of the present study. The proposed approach proved its usefulness in estimation of SLR parameters and is now under evaluation by our team on different types of classes of compounds and relations to assess its behavior and robustness.

4. Conclusions

The proposed approach proved feasible for estimating the parameters of the simple linear regression, in the absence of the assumption that the errors are normally distributed, assumption replaced by a more general one that the errors are Gauss-Laplace distributed. The obtained results demonstrated that in 12 out of 13 investigated cases the power of the error is significantly different by the convenient values of two. However, the plot of SLR-MLE and SLR-LS lines showed that, just in 3 out of 12 cases, the models are significantly different. The proposed approach can be further extended from simple linear regressions to multiple linear regressions.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

Dr. Alina C. Cozma is a fellow of POSDRU Grant no. 159/1.5/S/138776 entitled "Model colaborativ instituţional pentru translatarea cercetării ştiinţifice biomedicale în practica clinică, TRANSCENT."

References

[1] M. Goodarzi, B. Dejaegher, and Y. V. Heyden, "Feature selection methods in QSAR studies," *Journal of AOAC International*, vol. 95, no. 3, pp. 636–651, 2012.

[2] L. P. Hammett, "Some relations between reaction rates and equilibrium constants," *Chemical Reviews*, vol. 17, no. 1, pp. 125–136, 1935.

[3] P. Liu and W. Long, "Current mathematical methods used in QSAR/QSPR studies," *International Journal of Molecular Sciences*, vol. 10, no. 5, pp. 1978–1998, 2009.

[4] S. D. Bolboacă and L. Jäntschi, "Quantitative structure-activity relationships: linear regression modelling and validation strategies by example," *International Journal on Mathematical Methods and Models in Biosciences*, vol. 2, no. 1, Article ID 1309089, 2013.

[5] Oxford University Press, *Oxford Dictionary*, 2014, http://www.oxforddictionaries.com/definition/english/error?q=error.

[6] D. Gregory, *Astronomiae Physicae at Geometricae Elementa*, Oxford, 2nd edition, 1702, english title *The Elements of Physical and Geometrical Astronomy*-London, 1715, 2nd edition, Geneva, 1726.

[7] R. J. Adcock, "Note on the method of least squares," *The Analyst*, vol. 4, no. 6, pp. 183–184, 1877.

[8] R. J. Adcock, "A problem in least squares," *The Analyst*, vol. 5, no. 2, pp. 53–54, 1878.

[9] C. H. Kummell, "Reduction of observation equations which contain more than one observed quantity," *The Analyst*, vol. 6, no. 4, pp. 97–105, 1879.

[10] F. Galton, *Natural Inheritance*, Macmillan, New York, NY, USA, 5th edition, 1894.

[11] K. Pearson, "Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia," *Philosophical Transactions of the Royal Society London*, vol. 187, pp. 253–318, 1896.

[12] A. Bravais, *Analyse Mathematique sur les Probabilites des Erreurs de Situation d'un Point*, vol. 9 of *Mémoires Présentés par Divers Savants*, 1846.

[13] R. C. Allen and J. H. Stone, "The Gauss Markov theorem: a pedagogical note," *The American Economist*, vol. 45, no. 1, pp. 92–94, 2001.

[14] A.-M. Legendre, *New Methods for the Determination of the Orbits of the Comets*, Courcier, Paris, France, 1805.

[15] C. F. Gauss, *Theory of the Motion of the Heavenly Bodies Moving about the Sun in Conic Sections*, 1809 (Latin).

[16] R. A. Fisher, "On the mathematical foundations of theoretical statistics," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 222, pp. 309–368, 1922.

[17] R. A. Fisher, "Theory of statistical estimation," *Philosophical Transactions of the Royal Society London*, vol. 22, pp. 700–725, 1925.

[18] P. S. Laplace, *Théorie analytique des probabilities*, Courcier, Paris, France, 1812.

[19] A. Lahti, P. Hyltoft Petersen, J. C. Boyd, C. G. Fraser, and N. Jørgensen, "Objective criteria for partitioning Gaussian-distributed reference values into subgroups," *Clinical Chemistry*, vol. 48, no. 2, pp. 338–352, 2002.

[20] M. Meloun, M. Hill, and D. Cibula, "Exploratory biochemical data analysis: a comparison of two sample means and diagnostic displays," *Clinical Chemistry and Laboratory Medicine*, vol. 39, no. 3, pp. 244–255, 2001.

[21] T. Kalliokoski, C. Kramer, A. Vulpetti, and P. Gedeck, "Comparability of mixed IC50 data—a statistical analysis," *PLoS ONE*, vol. 8, no. 4, Article ID e61007, 2013.

[22] P.-S. Laplace, "Mémoire sur la probabilité des causes par les évènements," *Mémoires de l'Académie Royale des Sciences Presentés par Divers Savans*, vol. 6, pp. 621–656, 1774.

[23] P. S. Laplace, *Théorie Analytique des Probabilités*, Courcier, Paris, France, 1812.

[24] E. D. Feigelson, "Statistics in astronomy," in *Encyclopedia of Statistical Science*, S. Kotz and N. L. Johnson, Eds., vol. 9, Wiley, 1989.

[25] B. P. Kondratev, "Potential theory: equigravitating line segments for axisymmetric bodies," *Computational Mathematics and Mathematical Physics*, vol. 41, no. 2, pp. 247–259, 2001.

[26] L. Jänstchi, "Distribution fitting 1. Parameters estimation under assumption of agreement between observation and model," *Bulletin of University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca, Horticulture*, vol. 66, pp. 684–690, 2009.

[27] K. Kuljus and S. Zwanzig, "Asymptotic properties of a rank estimate in heteroscedastic linear regression," U.U.D.M. Report 2008:33, Uppsala University, Uppsala, Sweden, 2008, http://www2.math.uu.se/research/pub/Kuljus3.pdf.

[28] L. Jäntschi and S. D. Bolboacă, "Observation vs. observable: maximum likelihood estimations according to the assumption of generalized gauss and laplace distributions," *Leonardo Electronic Journal of Practices and Technologies*, vol. 8, no. 15, pp. 81–104, 2009.

[29] Y. H. Zhao, X. Yuan, L. M. Su, W. C. Qin, and M. H. Abraham, "Classification of toxicity of phenols to *Tetrahymena pyriformis* and subsequent derivation of QSARs from hydrophobic, ionization and electronic parameters," *Chemosphere*, vol. 75, no. 7, pp. 866–871, 2009.

[30] M. T. D. Cronin, A. O. Aptula, J. C. Duffy et al., "Comparative assessment of methods to develop QSARs for the prediction of the toxicity of phenols to *Tetrahymena pyriformis*," *Chemosphere*, vol. 49, no. 10, pp. 1201–1221, 2002.

[31] S. D. Bolboacă and L. Jäntschi, "Sensitivity, specificity, and accuracy of predictive models on phenols toxicity," *Journal of Computational Science*, vol. 5, no. 3, pp. 345–350, 2014.

[32] M. H. Abraham, R. Kumarsingh, J. E. Cometto-Muniz, and W. S. Cain, "A quantitative structure-activity relationship (QSAR) for a Draize eye irritation database," *Toxicology in Vitro*, vol. 12, no. 3, pp. 201–207, 1998.

[33] S. D. Bolboacă and L. Jäntschi, "From molecular structure to molecular design through the Molecular Descriptors Family Methodology," in *QSPR-QSAR Studies on Desired Properties for Drug Design*, E. A. Castro, Ed., pp. 117–166, Research Signpost, Transworld Research Network, 2010.

[34] A. Toropov, A. Toropova, T. Ismailov, and D. Bonchev, "3D weighting of molecular descriptors for QSPR/QSAR by the method of ideal symmetry (MIS). 1. Application to boiling points of alkanes," *Journal of Molecular Structure: THEOCHEM*, vol. 424, no. 3, pp. 237–247, 1998.

[35] A. M. Harsa, T. E. Harsa, S. D. Bolboacă, and M. V. Diudea, "QSAR in flavonoids by similarity cluster prediction," *Current Computer Aided-Drug Design*, vol. 10, no. 2, pp. 115–128, 2014.

[36] J. Li and P. Gramatica, "The importance of molecular structures, endpoints' values, and predictivity parameters in QSAR research: QSAR analysis of a series of estrogen receptor binders," *Molecular Diversity*, vol. 14, no. 4, pp. 687–696, 2010.

[37] N. C. Comelli, E. V. Ortiz, M. Kolacz et al., "Conformation-independent QSAR on c-Src tyrosine kinase inhibitors," *Chemometrics and Intelligent Laboratory Systems*, vol. 134, pp. 47–52, 2014.

[38] G. Melagraki, A. Afantitis, H. Sarimveis, O. Igglessi-Markopoulou, and C. T. Supuran, "QSAR study on para-substituted aromatic sulfonamides as carbonic anhydrase II inhibitors using topological information indices," *Bioorganic & Medicinal Chemistry*, vol. 14, no. 4, pp. 1108–1114, 2006.

[39] S. D. Bolboacă and L. Jäntschi, "Comparison of QSAR performances on carboquinone derivatives," *TheScientificWorldJOURNAL*, vol. 9, no. 10, pp. 1148–1166, 2009.

[40] S. D. Bolboacă, L. Jäntschi, and M. V. Diudea, "Molecular design and QSARs/QSPRs with molecular descriptors family," *Current Computer-Aided Drug Design*, vol. 9, no. 2, pp. 195–205, 2013.

[41] P. Manoj Kumar, C. Karthikeyan, N. S. Hari Narayana Moorthy, and P. Trivedi, "Quantitative structure-activity relationships of selective antagonists of glucagon receptor using QuaSAR descriptors," *Chemical and Pharmaceutical Bulletin*, vol. 54, no. 11, pp. 1586–1591, 2006.

[42] S. D. Bolboacă, D. D. Roşca, and L. Jäntschi, "Structure-activity relationships from natural evolution," *MATCH: Communications in Mathematical and in Computer Chemistry*, vol. 71, pp. 149–172, 2014.

[43] Chemical Computing Group, *Molecular Operating Environment (MOE), 2013.08*, Chemical Computing Group, Montreal, Canada, 2015.

[44] Z. W. Birnbaum and S. C. Saunders, "A new family of life distributions," *Journal of Applied Probability*, vol. 6, no. 2, pp. 319–327, 1969.

[45] L. Jäntschi, *Anderson-Darling Statistic*, 2014, http://l.academicdirect.org/Statistics/tests/AD/.

[46] N. Neykov, P. Filzmoser, R. Dimova, and P. Neytchev, "Robust fitting of mixtures using the trimmed likelihood estimator," *Computational Statistics & Data Analysis*, vol. 52, no. 1, pp. 299–308, 2007.

[47] N. D. Thang, L. Chen, and C. K. Chan, "Robust mixture model-based clustering with genetic algorithm approach," *Intelligent Data Analysis*, vol. 15, no. 3, pp. 357–373, 2011.

[48] X. Bai, W. Yao, and J. E. Boyer, "Robust fitting of mixture regression models," *Computational Statistics & Data Analysis*, vol. 56, no. 7, pp. 2347–2359, 2012.

[49] S. D. Bolboacă and L. Jäntschi, "The effect of leverage and/or influential on structure-activity relationships," *Combinatorial Chemistry & High Throughput Screening*, vol. 16, no. 4, pp. 288–297, 2013.