Inside of the Linear Relation between Dependent and Independent Variables

Lorentz JÄNTSCHI[1,2,3], Lavinia L. PRUTEANU[2], Alina C. COZMA[3,4], Sorana D. BOLBOACĂ[4,*]

[1] Technical University of Cluj-Napoca, Department of Physics and Chemistry, Muncii Bvd. 103-105, 400641 Cluj-Napoca, Romania. E-mail: lorentz.jantschi@gmail.com
[2] Babeş-Bolyai University, Institute for Doctoral Studies, Kogălniceanu Street no. 1,400084 Cluj-Napoca, Romania
[3] University of Oradea, Department of Chemistry, Universității Street no. 1, 410087 Oradea, Romania. E-mail: acozma@uoradea.ro
[4] Iuliu Haţieganu University of Medicine and Pharmacy, Department of Medical Informatics and Biostatistics, Louis Pasteur Street no. 6, 400349 Cluj-Napoca, Romania. E-mail: sbolboaca@umfcluj.ro

* Corresponding author: Sorana D. Bolboacă. E-mail: sbolboaca@umfcluj.ro; Phone: +40750774506

Abstract
Simple and multiple linear regression analysis are statistical methods used to investigate the link between activity/property of active compounds and the structural chemical characteristics. The linear regression models assume a normal distribution of the errors and the analysis is correctly conducted when this assumption is verified. The paper introduces a new approach of solving the simple linear regression without making any assumptions about the distribution of the errors. The proposed approach maximizes the probability of observing the event according to the random error. The use of the proposed approach is illustrated on ten classes of compounds with different activity or property. The proposed method proved reliable and showed to fit properly the observed data compared to the convenient approach of normal distribution of the errors.

Keywords: maximum likelihood estimate (MLE); simple linear regression (SLR); regression parameter; robust approach

Introduction

The quantitative structure activity/property relationships (QSAR/QSPR) are computational techniques that quantitatively relates chemical feature (such as descriptors) to a biological activity or property [1]. Linear regression is one of the earliest method [2] used to link the activity/property with structural information and is frequently used due to the easiness interpretation [3]. 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 multicolinearity [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* (Oxford, 1702; 2nd ed., Geneva, 1726; English title *The Elements of Physical and Geometrical Astronomy* – London, 1715, 2nd ed., 1726). In the late 1800's, Adcock [6,7] suggested that the errors must pass through the centroid of the data. The method proposed by Adcock, named orthogonal regression, explore the distance between a point and the line in a perpendicular direction to the line [6,7]. Kummel [8] investigated other than perpendicular directions between the points and line. Galton described in 1894 the regression slope ('r') based on an experiment of sweet pea seeds [9]. Two years later, Pearson generalized the errors in the variable and published a rigorous description of correlation and regression analysis [10] (Pearson recognized the contribution of Bravais [11] to the mathematical formula for correlation). Due to the ability to produced best linear unbiased parameters [12], the coefficients in simple linear regression (SLR)

model are usually estimated by minimizing the sum of squared deviations (least squares estimation, method introduced by Legendre in 1805 [13], and used/applied by Gauss in 1809 [14]). Fisher has introduced the concept of maximum likelihood within linear models [15,16].
The generic equation of simple linear regression (Eq1) between observed dependent variable *Y* and observed independent variable *X* is:

$$Y \sim \hat{Y} = a \cdot X + b \quad (1)$$

where *a* and *b* are unknown constant values (estimators of statistics parameters of simple linear regression), $\hat{Y}$ is the value of the dependent variable estimated by the model, *Y* is the observed value of dependent variable, 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 i^{th} observation in the sample ($1 \leq i \leq n$, when n = sample size), and *q* is an unknown coefficient. The unknown *q* parameter is an estimator, the power of the errors on simple linear regression.
In the general case, residuals ($S_i = Y_i - aX_i - b$, where S = residual) follow the Gauss-Laplace distribution with μ, σ and *p* unknown statistical parameters:

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

where μ = population mean, σ = population standard deviation, *q* = power of the errors, and Γ = gamma function, *s* = sample standard deviation.
Gauss-Laplace distribution is symmetrical, has three statistical parameters (population mean, population standard deviation, and power of the errors) [14,17] and two main particular cases. First particular case is Gauss distribution [14] often observed on arrays of biochemical data [18,19,20] while the second particular case is Laplace distribution (with mean of zero and variance σ^2) [21,22] commonly seen on astrophysical data [23,24].
The problem of estimating the parameters of the SLR (Eq1) for the first particular case (Gauss distribution) consider $q = 2$ residuals (power of the errors related with experimental errors). The simplest ways to obtain the coefficients of regression for this particular case is by solving the system of linear equations under the assumption that $\Sigma S_i^2 = \min$. [25] ($\Sigma S_i^2 = \Sigma(Y_i - a \cdot X_i - b)^2$, where *a* and *b* are unknown parameters).
Another particular case is $q = 1$ when residuals follows the Laplace distribution. In view of the fact that $\Sigma|S_i| = \Sigma|Y_i - a \cdot X_i - b|$ is 'is not differentiable everywhere' [26], the solution in more difficult to be obtain for this particular case.
One question that can be ask is "what is the proper value of *q* that should be used in the simple linear regression analysis (Eq1)?". A previously published 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 distribution family is investigated [27]. Based on this result, the aim of the present study was to formulate the problem of solving the simple linear regression equation (Eq1) without making any assumptions about the power of the errors (*q*).

2. Material and Methods

2.1. Mathematical approach

The problem of regression (Eq1) is transformed into a problem of estimation if the residuals $S_i = Y_i - a \cdot X_i - b$ are introduced in Eq2 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 Eq1 as presented in Eq3:

$$M(Y) \sim M(\hat{Y}) = a \cdot M(X) + b \quad (3)$$

where *b* is the population mean of the Gauss-Laplace quantity $Y - a \cdot X$ (Eq2), *Y* = observed / measured dependent variable, $\hat{Y}$ = dependent variable estimated by the regression model, X =

independent/predictor variable, M = mean operator. For certain arrays of paired observations (X,Y), the problem of regression expressed in Eq1 is transformed in a problem of estimating the parameters of the bi-dimensional Gauss-Laplace distribution as presented in Eq4:

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

An efficient instrument to solve the estimation problem presented in Eq4 is maximum likelihood estimation (MLE), method proposed by Fisher [15,16]. 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 $\Pi GL(X_i,Y_i;\sigma,q,a,b)$ = max., thus $\log(\Pi GL(X_i,Y_i;\sigma,q,a,b))$=max., which led to the expression in Eq5:

$$\Sigma\log(GL(X_i,Y_i;\sigma,q,a,b)) = \max. \quad (5)$$

By including Eq4 in Eq5 and using the natural logarithm, the problem presented in Eq1 became a problem of optimization (Eq6):

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

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

The optimization problem presented in Eq5 could be solved iteratively if the start point is a good initial solution (situated near the optimal solution). In this research, the start point in the optimization was the solution of a particular case of Eq6 as presented in Eq7 ():

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

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

2.2. Data sets

Ten classes of previously investigated compounds were used as input data in our analysis. The class of compounds, the activity/property of interest along the number of compounds in the set and the reference to the paper from where the descriptors were collected are given in Table 1.

Table 1. Characteristics of the investigated classes of compounds

Set	n	Class	Activity/Property – expressed as	Ref
1a	35	phenols	toxicity on *Tetrahymena pyriformis* – $\log(1/IGC_{50})$	[28,29,30]
1b	126			
1c	250			
2	24	organic compounds	Solubility – logP	[31,32]
3	73	alkanes	Boiling point – BP	[33]
4a	40	flavonoids	Solubility – logP	[34]
4b	30		Lethal Dose 50% – $\ln(LD_{50})$	
5	132	estrogen receptor (ER)	binding affinities – log(RBA)	[35]
6	80	pyrrolo-pyrimidine derivatives	c-Src Tyrosine Kinase inhibitory activity – $pIC_{50} = -\log_{10}(IC_{50})$	[36]
7	47	substituted aromatic sulfonamides	inhibition activity on carbonic anhydrase II – $\log K_I$	[37]
8	37	carboquinone derivatives	molar concentration – log(1/MC)	[38]
9	47	dipeptides	ACE (angiotensin converting enzyme) inhibitory activity – ACE	[39]
10	60	mycotoxins compounds	retention time – ln(RT)	[40]

Simple linear regression models under the assumption of linear relationship between structural descriptors and activity/property of active compounds were identified starting with the values of descriptors previously published in the literature. The models with highest goodness-of-fit for each class of compounds are presented in Table 2.

Table 2. Characteristics of sets used in the optimization study

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

SLR = simple linear regression;
$\log(1/IGC_{50})$ = concentrations (expressed as mM) producing a 50% growth inhibition on *T. pyriformis*;
* MDF descriptors [32, 38, 39, 41]
SD = global correlation descriptor [34];
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 [42]
R^2 = determination coefficient; s = standard error of the estimate;
F = Fisher's statistic of the regression model; n = sample size

3. Algorithm implementation

An object was created to solve the problem:

```
class ERegression{
  function DCopy(&x,&y); //initialize with data
  function DStat(); //do the analysis
  function m01(&z); //compute the average of the array
  function m02(&z,&w); //compute the average of the arrays product
  function GLMLE(&it); //calculate MLE of GL for given parameters
  function Steps(); //iterates the convergence to the optimal parameters
  function ERegression(&u,&v){DCopy(u,v); DStat();} //constructor of the object
}
function ERegression.m01(&z){s=0.0;for(i=0;i<m;i++)s+=z[i];s/=m; return(s);}
function ERegression.m02(&z,&w){s=0.0;for(i=0;i<m;i++)s+=z[i]*w[i];s/=m; return(s);}
function ERegression.DCopy(&u,&v){y=&u;x=&v;m=count(y);}
function ERegression.DStat(){
    my1=m01(y);my2=m02(y,y);dy2=my2-pow(my1,2);
    mx1=m01(x);mx2=m02(x,x);dx2=mx2-pow(mx1,2);
    mxy=m02(x,y);cxy=mxy-mx1*my1;
    guess=array(
      "p" => 2,
      "a" => cxy/dx2,
      "m" => my1-guess["a"]*mx1,
      "s" => pow(dy2-pow(cxy,2)/dx2,0.5),
      "MLE" => 0
    );
    guess["MLE"]=GLMLE(guess);
    stepx=(stepn-stepn%2)/2;
    stepv=array();for(i=0;i<stepn;i++)stepv[i]=exp((i-stepx)/50.0);
}
function ERegression.GLMLE(&it){
    g1=gamma1p(it["p"]);
    g3=gamma3p(it["p"]);
    t1=m*log(it["p"]*pow(g3,0.5)/pow(g1,1.5)/it["s"]/2.0);
```

```
    t2=pow(g1/g3,it["p"]/2.0)*pow(it["s"],it["p"]);
    t3=0.0;
    for(i=0;i<m;i++){t3+=pow(abs(y[i]-it["a"]*x[i]-it["m"]),it["p"]);}
    return(t1-t3/t2);
}
```

The optimal solution of Eq6 is iteratively obtained from the optimal solution Eq7 by making small changes to the actual values of the coefficients and selecting the new ones which makes the MLE value greater. The weights of changes are more or less arbitrary, and the selected ones are a compromise of convergence speed in the convergence space.

```
function ERegression.Steps(){
  itera=array();bestf=guess;
  for(;;){
    for(i1=0;i1<stepn;i1++){ //for p
      itera["p"]=guess["p"]*stepv[i1];
      for(i2=0;i2<stepn;i2++){ //for a
        itera["a"]=guess["a"]*stepv[i2];
          for(i3=0;i3<stepn;i3++){//for m
            itera["m"]=guess["m"]*stepv[i3];
            for(i4=0;i4<stepn;i4++){//for s
              itera["s"]=guess["s"]*stepv[i4];
              itera["MLE"]=GLMLE(itera);
              if(itera["MLE"]>guess["MLE"])bestf=itera;
            }
          }
        }
      }
      itera=bestf;
      if(abs(itera["MLE"]-guess["MLE"])<stope)break;
      guess=itera;
    }
  }
}
```

The algorithm was implemented (in PHP). The program find of the solutions of Eq6 starting with the initial solutions identified by applying Eq7.
The values of (Y) and independent (X) variables are read from a given *.*txt* file and the array of X and Y is returned through the function named **get_data**:

```
function get_data($text){
  $a=explode("\r\n",file_get_contents($text)); //get the data from the text file
  array_shift($a); //chop the headers
  $x=array(); $y=array(); //initialize with empty the arrays of observed values
  for($i=0; $i<count($a); $i++){ //for each line containing data
    $b=explode("\t",$a[$i]); //split the line into the pair of x and y values
    $x[]=$b[1]; $y[]=$b[2]; //collect x and y values
  }
  return(array($x,$y)); //return arrays of values
}
```

The main part of the program calls the ERegression object to find the solution to the given data:

```
$xy=get_data("some_file_name.txt"); //return the data as an array of two arrays
$reg_xy = new ERegression($xy[0],$xy[1]); //instantiate the ERegression object
$reg_xy->Steps(); //iterate the optimal solution
```

The source code of the implemented algorithm is freely available on request from the authors.

4. Results

The developed algorithm was tested on ten different data sets. The number of iteration needed to find the optimal solution varied from 9 (set10) to 185 (set4b). The number of iteration needed to reach the optimal solution seems not being related with the number of compounds in the sample

when the same class of compounds is investigated (63 iteration – set1a, 51 iteration – set1b, 86 iteration – set1c). The number of iteration needed to obtained the optimal solution was equal with 173 for the smallest dataset (set2) and 86 for the dataset with the highest number of compounds (set1c).

The results of simulation study obtained for the ideal solution (q = 2 – residual follows the Gaussian distribution) and for solution that satisfy the Eq6 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 optimal q were obtained for:

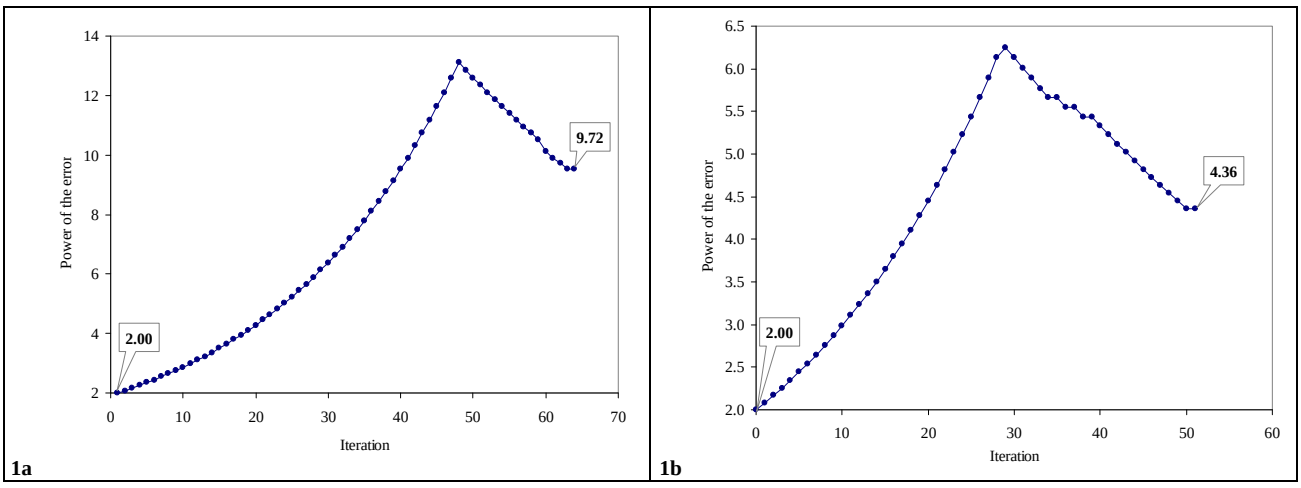
- a coefficient – set1b, set3, set6
- b coefficient – set3, set6, set8, set10;
- σ – set 1b, set1c, set3, set4a, set5, set6, set8, set9, set10.

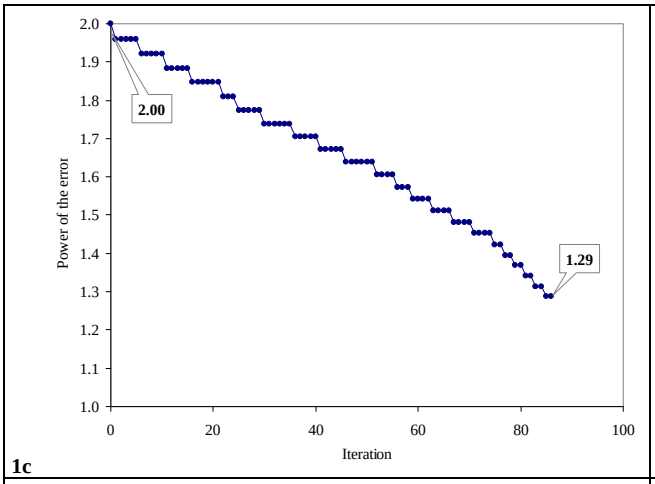
Table 3. Optimization results: q=2 vs. q determined to satisfy Eq6

set	n	q=2			q=?			
		a	$b=\mu$	σ	p	a	$b=\mu$	σ
1a	35	0.678	-1.386	0.218	9.52	0.638	-1.181	0.222
1b	126	0.647	-1.050	0.298	4.36	0.647	-1.029	0.298
1c	250	0.509	-0.443	0.596	1.29	0.563	-0.623	0.569
2	24	-0.004	2.095	0.414	0.61	-0.005	2.270	0.516
3	73	188.408	-507.959	3.762	1.34	188.408	-507.959	3.762
4a	40	1.000	5.232	0.308	2.81	1.041	5.338	0.308
4b	30	0.002	-61.168	0.945	0.67	0.002	-64.950	0.964
5	132	0.024	-3.812	1.374	1.70	0.026	-3.967	1.374
6	80	0.255	-1.216	0.558	2.87	0.255	-1.216	0.558
7	47	-0.578	2.646	0.360	3.43	-0.555	2.594	0.353
8	37	-4.129	5.789	0.372	1.29	-4.297	5.789	0.372
9	47	47.561	-0.169	0.319	3.17	49.502	-0.279	0.319
10	60	0.348	1.711	0.492	1.74	0.355	1.711	0.492

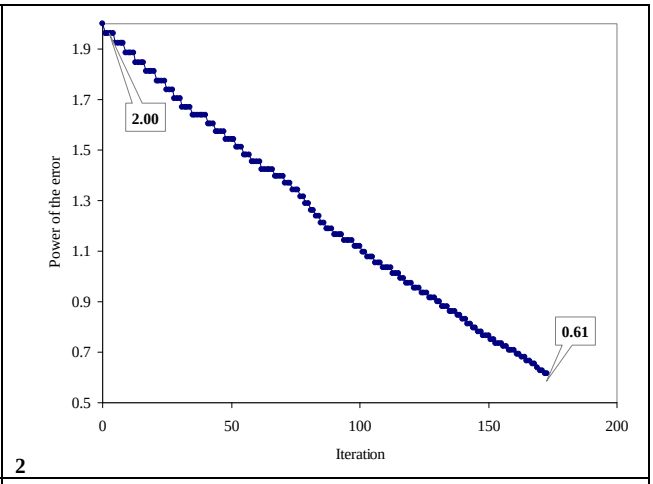
q = power of the errors; a, b = coefficients in the simple linear model;
μ = population mean; σ = population standard deviation

The evolution of power of the errors obtained in the simulation study according to iteration is presented in Figure 1.

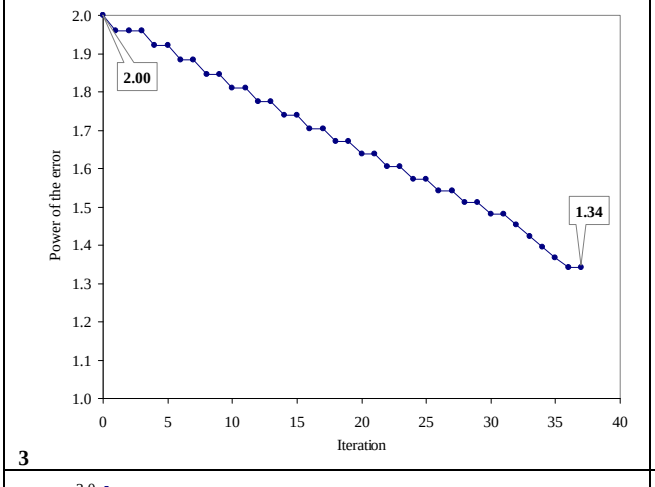




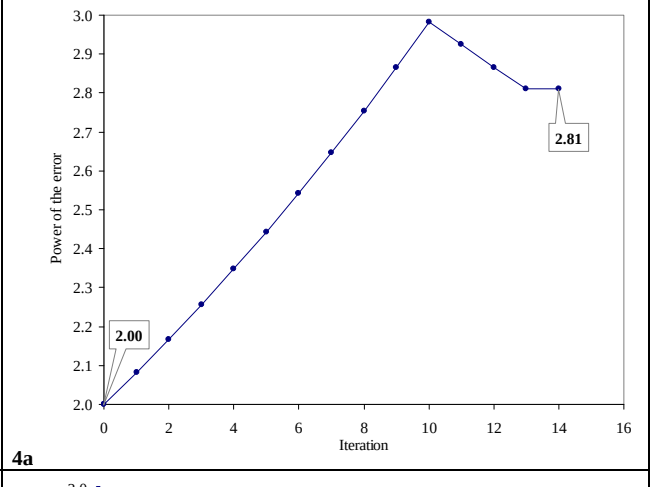
1c



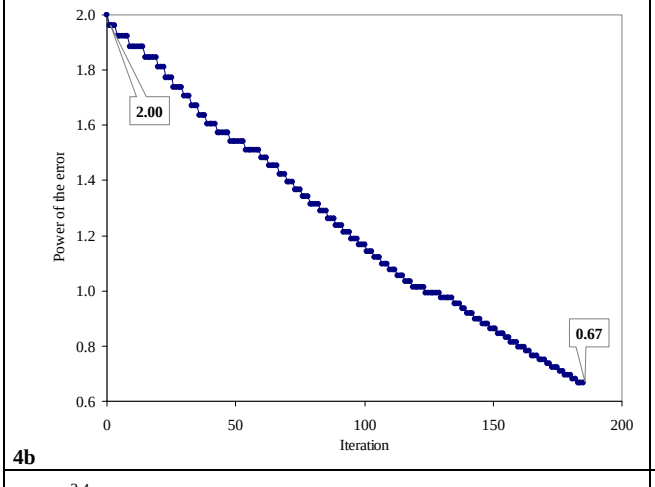
2



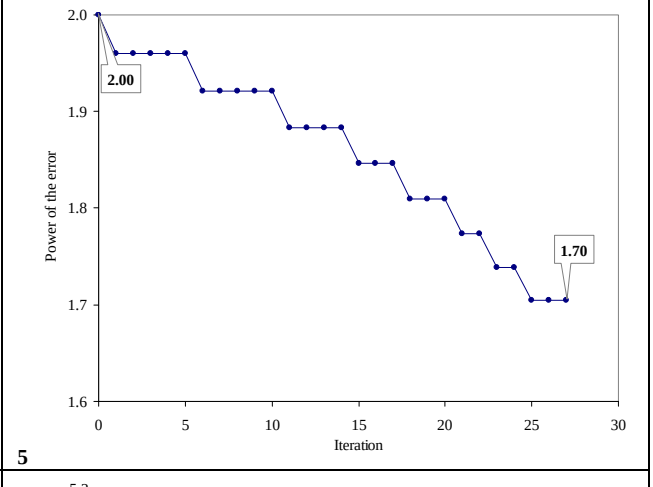
3



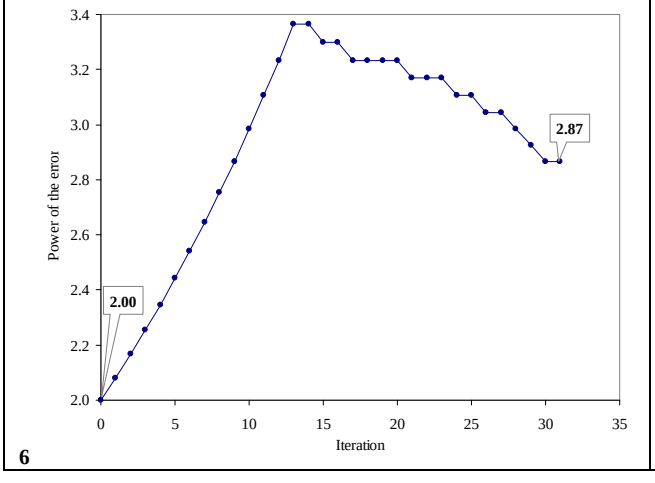
4a



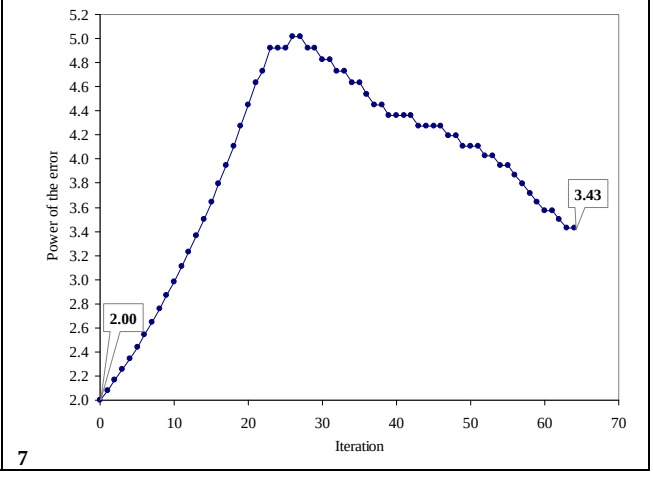
4b



5



6



7

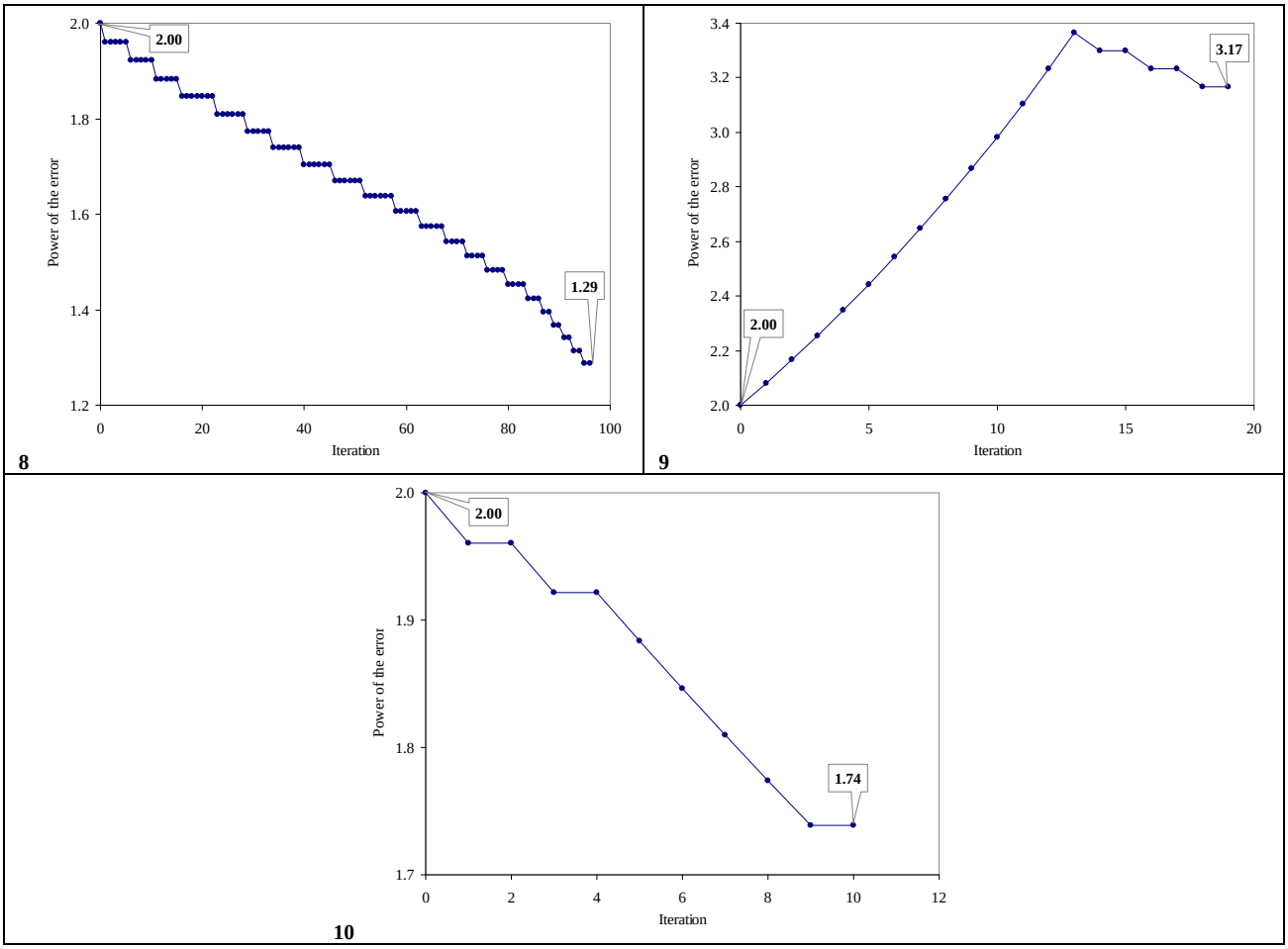


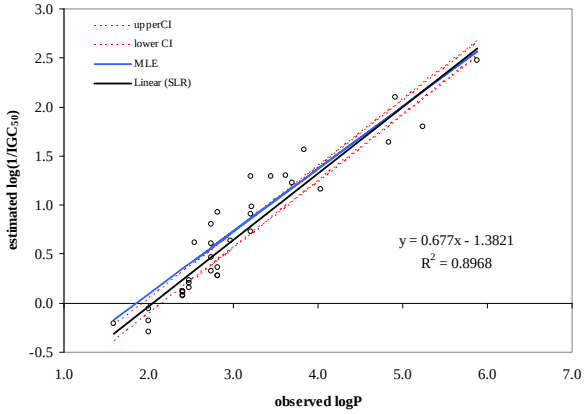
Figure 1. The pattern distribution of value of power of the errors according with iteration

The evolution of value of power of the errors was in both directions starting with the second iteration and as expected never achieved negative values (see Figure 1) and in most of cases the q from the MLE proved significantly different by q=2 (Table 4).

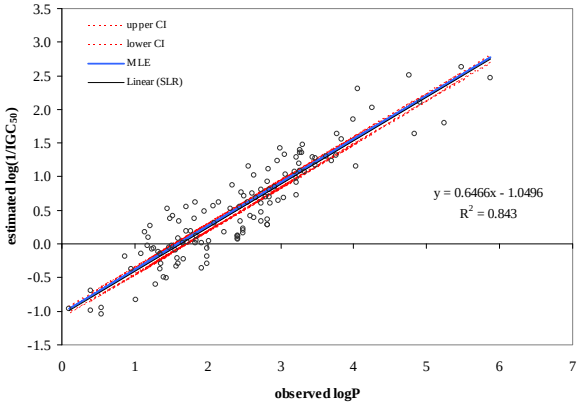
Table 4. The p-value associated to the difference between q=2 and q resulted from MLE

Set	p-value
1a	$4.20 \cdot 10^{-54}$
1b	$3.07 \cdot 10^{-115}$
1c	$2.42 \cdot 10^{-53}$
2	$1.76 \cdot 10^{-12}$
3	$6.93 \cdot 10^{-2}$
4a	$1.30 \cdot 10^{-19}$
4b	$1.16 \cdot 10^{-8}$
5	$7.33 \cdot 10^{-3}$
6	$3.39 \cdot 10^{-23}$
7	$1.06 \cdot 10^{-30}$
8	$4.75 \cdot 10^{-14}$
9	$9.01 \cdot 10^{-29}$
10	$6.09 \cdot 10^{-5}$

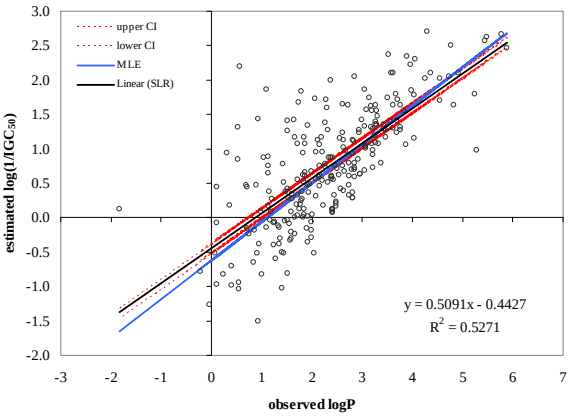
The plot of both regression lines (simple linear regression and associated 95% confidence interval and MLE regression) for each investigated data sets are presented in Figure 2.



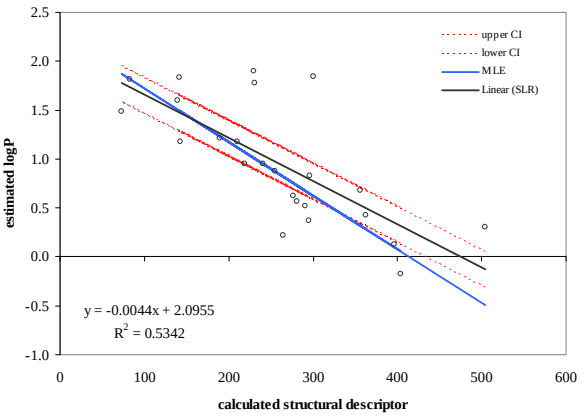
1a



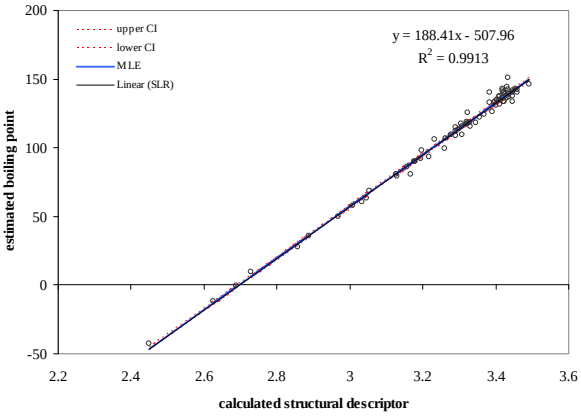
1b



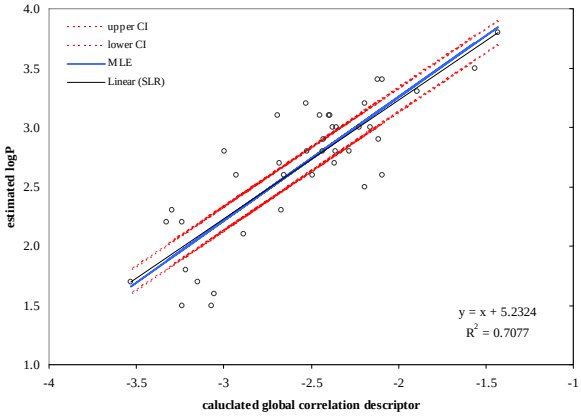
1c



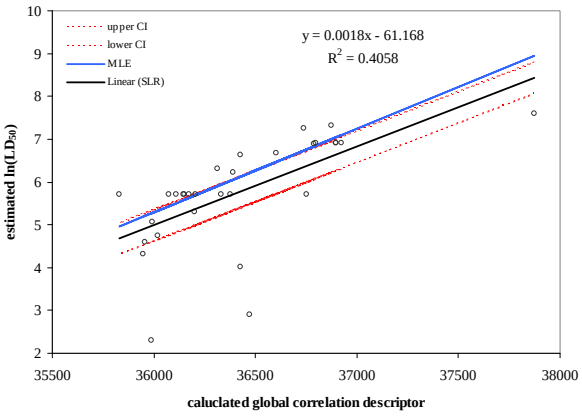
2



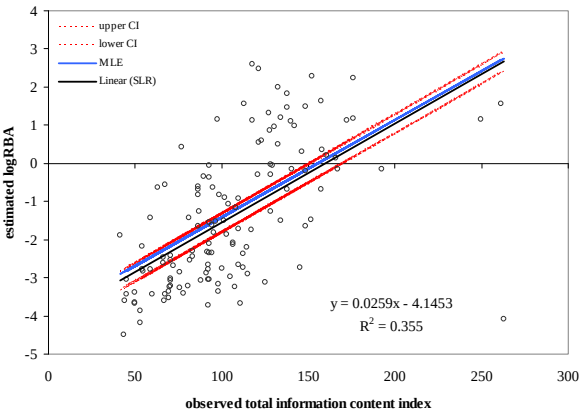
3



4a



4b



5

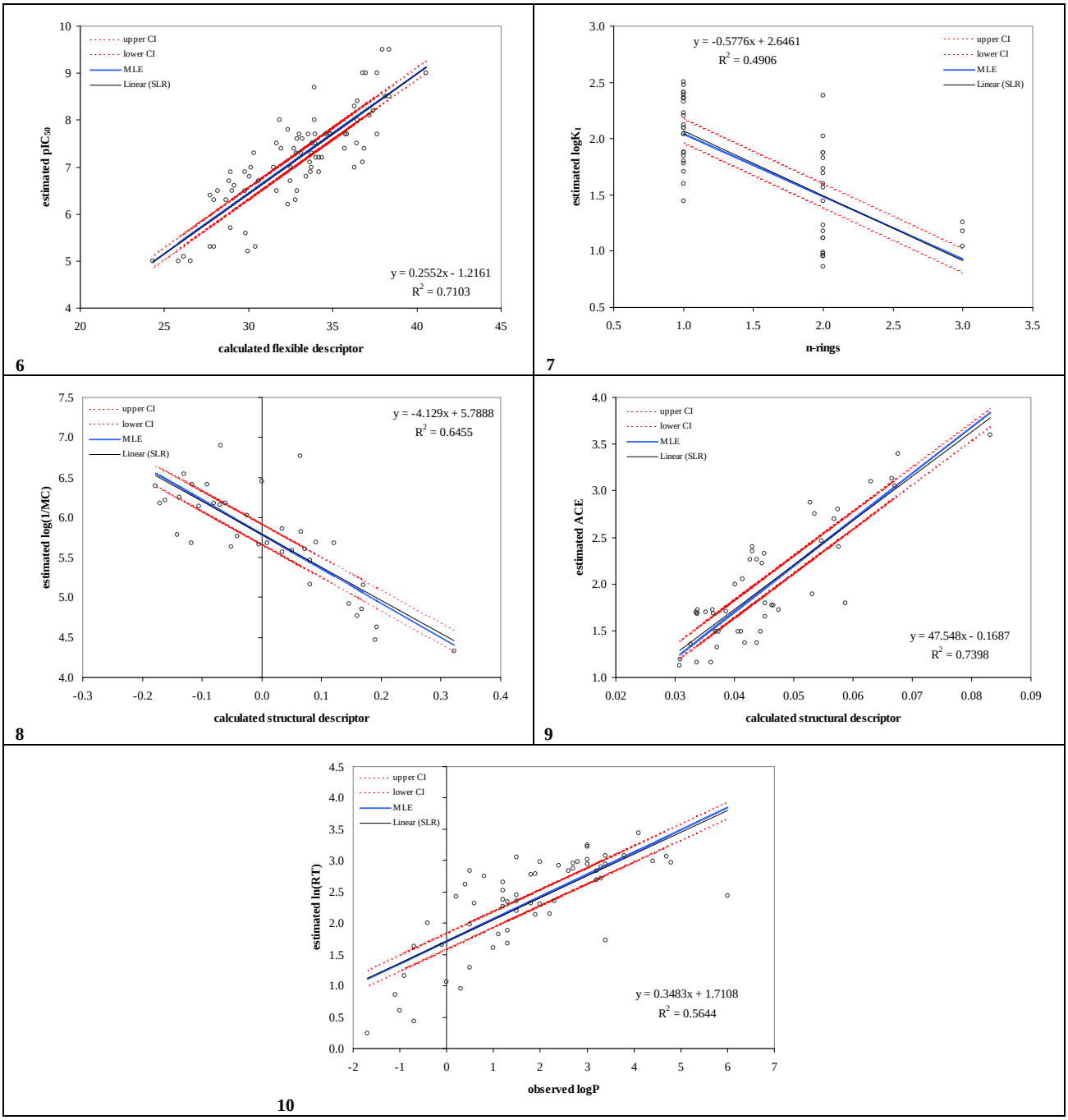


Figure 2. The line of SLR (q=2) and MLE (q determined to satisfy Eq6)

5. 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 number of iteration needed to reach the solution proved not in relation with the number of compounds in the sample, the maximum number of iterations being almost 21 times more than the minimum number of iterations.

The analysis of the obtained results revealed the following:

- In 9 out of 13 cases, at least one coefficient (a, b, σ) proved equal for q=2 and q determined to satisfy Eq6.

- In 6 out of 13 cases the power of the errors obtained by MLE proved significantly higher than 2 (see Table 2, and Table 4). The difference varied from 0.8099 (set4a) to 7.5176 (set1a) (see Table 2).
- Just in one case, the difference between powers of the errors proved not statistically different (set3, $p = 0.0693$).
- In 6 out of 13 cases the difference between power of the errors (SLR and MLE) proved lower than 1 (see Table 3)
- The smallest distance between the powers of the errors (from SLR and MLE) was of 0.2613 (set10). Note that the powers of the errors proved significantly different ($p = 6.09 \cdot 10^{-5}$, Table 4).
- Two classes of compounds (set3 and set6) proved identical values of a, b, and σ unconcerned the method used in the regression analysis (SLR or MLE, see Table 3).

The analysis of the evolution of the power of the errors as function of iteration revealed that even if identical values of p are obtained in the first 29 iterations for the two related samples (set1 and set2, Figure 1) the pattern is not representation for the class of the compounds. As can be seen from the distribution of power of errors in Figure 1, 1c the pattern is significantly different by those observed on subsets of the whole class of compounds (1a and 1b, Figure 1). Opposite behavior is also observed for the other two related samples (), the value of p increased until a maximum (iteration 29 for set2 and iteration 48 for set1) and decrease after this value (the decrease is sharper in set1 compared to set2). The power of the errors for set3 and set4 decreased in steps, with a sharper decreased in set3 compared to set4 (Figure 1). Overall, two distinct patterns can be observed in Figure 1:

- The values of power of the error increase with iteration until peaks and after the power of the error decrease (sometimes decreases in steps – set6, set7 and set9): set1a, set1b, set4a, set6, set9.
- The values of power of the error decrease in steps with the increase of iteration: set1c, set2, set3, set4b, set5, set8 and set10.

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 violated and for this dataset the simple linear regression is not the proper analysis. In 4 out of 13 cases, the MLE line is partly outside the 95% confidence boundaries of the SLR line (set1a, set1c, set2, and set4b Figure 2). Accordingly, in all these cases, it could be considered that the MLE linear model is significantly different by the SLR model. The overlapping of MLE linear line and SLR line is observed for the set3, without being possible to make a visual distinction between the two lines (Figure 2). For this set, the q determined to satisfy Eq6 was equal with 1.34 and proved not significantly different by convenient value of 2 (see Table 4). For all other sets of investigated compounds the MLE linear line is within the boundaries of 95% confidence intervals of SLR line and thus even if the powers of the errors proved significantly different by the convenient value of 2 (see Table 2), these MLE models could not be considered significantly different by SLR models.

To sum-up, it is certainly that the proposed approach of maximizing the probability of observing the event according to the random error fit the observed data and the q is significantly different by the convenient value (when q=2) in assessment of the linear relationship between one dependent and one independent variable. 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 to be identified on other (X,Y) pairs, analysis which is currently conducted by our team. The relation presented in Eq6 thereby defines a new 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:

- Exponential: $S_i = Y_i - a_1 \cdot \exp(-X_i/a_2)$ for $Y \sim a_0 + a_1 \cdot \exp(-X/a_2)$;
- Double exponential: $S_i = Y_i - a_1 \cdot \exp(-X_i/a_2) - a_3 \cdot \exp(-X_i/a_4)$ for $Y \sim a_0 + a_1 \cdot \exp(-X/a_2) + a_3 \cdot \exp(-X/a_4)$;
- Power: $S_i = Y_i - a_1 \cdot \mathrm{pow}(X_i, a_2)$ for $Y \sim a_0 + a_1 \cdot \mathrm{pow}(X, a_2)$;
- Inversed: $S_i = Y_i - a_1/(X_i - a_2)$ for $Y \sim a_0 + a_1/(X - a_2)$;

The relation presented in Eq6 may be also extended to the multiple linear regression ($Y \sim a_0 + \Sigma_{j>0}a_jX_j$) when the expression $S_i = Y_i - aX_i$ become $S_i = Y_i - \Sigma_{j>0}a_jX_{j,i}$. If in the case of multiple linear regressions the classical method (minimizing the squared error) maximizes the correlation coefficient, the proposed approach (Eq6) maximizes the probability of observing the event according to the random error. Accordingly, Eq6 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. Opposite, 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 ($\hat{Y}$) 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 Eq6 did not change the correlation coefficient between Y and $\hat{Y}$ but offer a solution in regards of estimated valued of Y and of the unknown coefficients (estimators of the population coefficients) that enter in a dependence relation between X and Y. The relation proposed in this manuscript (Eq6) introduced an additional parameter in the estimation, namely the power of the errors of Gauss-Laplace distribution (p) (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 [43,44,45]. No outliers have been identified in the dependent variable on set2 and set3 [41,43,44]. Therefore, on these two sets of compounds, 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 [46,47,48]) 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.

6. Conclusions

The approach proposed in this manuscript demonstrate 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 MLE and SLR lines showed that just in 3 out of 12 cases, the models are significantly different.

Acknowledgements

Alina C. Cozma is a fellow of POSDRU grant no. 159/1.5/S/138776, grant with title: “Model colaborativ instituțional pentru translatarea cercetării științifice biomedicale în practica clinică – TRANSCENT”. The funding source had no role in the study design, data collection and analysis, decision to publish, or in the preparation of the manuscript.

References

[1] Goodarzi, M., Dejaegher, B., Heyden, Y.V. Feature selection methods in QSAR studies. Journal of AOAC International Volume 95, Issue 3, May 2012, Pages 636-651

[2] Hammett LP. Some relations between reaction rates and equilibrium constants. Chem. Rew. 1935, 17, 125-136.

[3] P. Liu, W. Long. Current Mathematical Methods Used in QSAR/QSPR Studies. Int J Mol Sci. 2009; 10(5): 1978–1998.

[4] Sorana D. BOLBOACĂ, Lorentz JÄNTSCHI, 2013, Quantitative Structure-Activity Relationships: Linear Regression Modelling and Validation Strategies by Example, International Journal on Mathematical Methods and Models in Biosciences (BIOMATH), 2(1), 1309089

[5] Oxford dictionary. [online] © 2014 Oxford University Press. [Accessed March 15, 2014]. Available from: http://www.oxforddictionaries.com/definition/english/error?q=error

[6] R.J. Adcock, Note on the method of least squares, Analyst. 4 (1877) 183-184.

[7] R.J. Adcock, A problem in least squares, Analyst. 5 (1878) 53-54.

[8] C.H. Kummel, Reduction of observed equations which contain more than one observed quantity, Analyst. 6 (1879) 97-105.
[9] F. Galton, Natural Inheritance (5th ed.). New York: Macmillan and Company, 1894.
[10] K. Pearson, Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity and Panmixia, Philos. Trans. Roy. Soc. London 187 (1896) 253-318.
[11] A. Bravais, Analyse Mathematique sur les Probabilites des Erreurs de Situation d'un Point, Memoires par divers Savans 9 (1846) 255-332.
[12] R.C. Allen, J.H. Stone, The Gauss Markov Theorem: A Pedagogical Note, Am. Econ. 45 (2001) 92-94.
[13] A.-M. Legendre, Nouvelles méthodes pour la détermination des orbites des comètes. (New methods for the Determination of the Orbits of the Comets). Paris: Courcier, 1805.
[14] C.F. Gauss, Theoria motvs corporvm coelestivm in sectionibvs conicis Solem ambientivm [Theory of the Motion of the Heavenly Bodies Moving about the Sun in Conic Sections] (in Latin). 1809.
[15] R.A. Fisher, On the mathematical foundations of theoretical statistics, Philos. Trans. Roy. Soc. London Ser. A 222 (1922) 309-368.
[16] R.A. Fisher, Theory of statistical estimation, Proc. Cambridge Philos. Soc. 22 (1925) 700-725.
[17] P.S. Laplace, Théorie analytique des probabilities. Paris, 1812.
[18] A. Lahti, P. Hyltoft Petersen, J.C. Boyd, C.G. Fraser, N. Jørgensen, Objective criteria for partitioning gaussian-distributed reference values into subgroups, Clin. Chem. 48 (2002) 338-352.
[19] M. Meloun, M. Hill, D. Cibul, Exploratory Biochemical Data Analysis: a Comparison of Two Sample Means and Diagnostic Displays, Clin. Chem. Lab. Med. 39 (2001) 244-255.
[20] T. Kalliokoski, C. Kramer, A. Vulpetti, P. Gedeck, Comparability of Mixed IC50 Data - A Statistical Analysis, PLoS ONE 8 (2013) e61007.
[21] P.-S. Laplace, Mémoire sur la probabilité des causes par les évènements, Mémoires de l'Academie Royale des Sciences Presentés par Divers Savan 6 (1774) 621-656.
[22] P.S. Laplace, Theorie Analytique des Probabilites. Paris: Courcier, 1812.
[23] E.D. Feigelson, Statistics in Astronomy, in: S. Kotz and N.L. Johnson (eds.), Encyclopedia of Statistical Science, 1989, vol. 9.
[24] B.P. Kondratev, Potential theory: equigravitating line segments for axisymmetric bodies, Comp. Math. Math. Phys. 41 (2001) 247-259.
[25] L. Jänstchi, Distribution Fitting 1. Parameters Estimation under Assumption of Agreement between Observation and Model, Bulletin UASVM Horticulture 66 (2009) 684-690.
[26] K. Kuljus, S. Zwanzig, Asymptotic properties of a rank estimate in heteroscedastic linear regression, U.U.D.M. Report 2008:33. Available from: http://www2.math.uu.se/research/pub/Kuljus3.pdf
[27] L. Jäntschi, S.D. Bolboacă, Observation vs. Observable: Maximum Likelihood Estimations according to the Assumption of Generalized Gauss and Laplace Distributions, Leonardo El. J. Pract. Technol. 15 (2009) 81-104.
[28] Y.H. Zhao, X. Yuan, L.M. Su, W.C. Qin, M.H. Abraham, Classification of toxicity of phenols to Tetrahymena pyriformis and subsequent derivation of QSARs from hydrophobic, ionization and electronic parameters, Chemosphere 75 (2009) 866-871.
[29] M.T.D. Cronin, A.O. Aptula, J.C. Duffy, T.I. Netzeva, P.H. Rowe, I.V. Valkova, T.W. Schultz. Comparative assessment of methods to develop QSARs for the prediction of the toxicity of phenols to Tetrahymena pyriformis. Chemosphere, 49 (2002), pp. 1201–1221
[30] Sorana D. BOLBOACĂ, Lorentz JÄNTSCHI, 2014, Sensitivity, Specificity, and Accuracy of Predictive Models on Phenols Toxicity, Journal of Computational Science, 5(3):345–350, Elsevier (ISSN: 1877-7503, doi:10.1016/j.jocs.2013.10.003).
[31] M.H. Abraham, R. Kumarsingh, J.E. Cometto-Muniz, W.S. Cain, A Quantitative Structure-Activity Relationship (QSAR) for a Draize Eye Irritation Database, Toxicol. in Vitro 12 (1998) 201-207.
[32] S.D. Bolboacă, L. Jäntschi, From molecular structure to molecular design through the Molecular Descriptors Family Methodology, in: QSPR-QSAR Studies on Desired Properties for Drug Design (Ed.: Eduardo A. CASTRO). Research Signpost, Transworld Research Network, 2010, p. 117-166.
[33] A. Toropova, A. Toropova, T. Ismailov, D. Bonchev, 3D weighting of molecular descriptors for QSPWQSAR by the method of ideal symmetry (MIS). 1. Application to boiling points of alkanes, J. Mol. Struct. (THEOCHEM) 424 (1998) 237-247.
[34] Alexandra M. HARSA, Teodora E. HARSA, Sorana D. BOLBOACĂ, Mircea V. DIUDEA, 2014, Qsar In Flavonoids By Similarity Cluster Prediction, Current Computer-Aided Drug Design, 10:115-128.
[35] Mol Divers. 2010 Nov;14(4):687-96. doi: 10.1007/s11030-009-9212-2. Epub 2009 Nov 17. The importance of molecular structures, endpoints' values, and predictivity parameters in QSAR research: QSAR analysis of a series of estrogen receptor binders. Li J1, Gramatica P.
[36] Conformation-independent QSAR on c-Src tyrosine kinase inhibitors Nieves C. Comelli, Erlinda V. Ortiz, Magdalena Kolacz, Alla P. Toropova, Andrey A. Toropov, Pablo R. Duchowicz, Eduardo A. Castro. Chemometrics and Intelligent Laboratory Systems Volume 134, 15 May 2014, Pages 47–52
[37] Melagraki G1, Afantitis A, Sarimveis H, Igglessi-Markopoulou O, Supuran CT. QSAR study on para-substituted aromatic sulfonamides as carbonic anhydrase II inhibitors using topological information indices. Bioorg Med

Chem. 2006 Feb 15;14(4):1108-14. Epub 2005 Oct 5.
[38] Sorana D. BOLBOACĂ, Lorentz JÄNTSCHI, 2009, Comparison of QSAR Performances on Carboquinone Derivatives, TheScientificWorldJOURNAL, 9(10), 1148-1166
[39] Sorana D. BOLBOACĂ, Lorentz JÄNTSCHI, Mircea V. DIUDEA, 2013, Molecular Design and QSARs with Molecular Descriptors Family, Current Computer-Aided Drug Design, 9(2), 195-205, Bentham Science Publishers (ISSN: 1573-4099, E-ISSN: 1875-6697, doi:10.2174/1573409911309020005
[40] Palanivelu MANOJ KUMAR, Chandrabose KARTHIKEYAN, Narayana Subbiah HARI NARAYANA MOORTHY, Piyush TRIVEDI. Quantitative Structure–Activity Relationships of Selective Antagonists of
Glucagon Receptor Using QuaSAR Descriptors. Chem. Pharm. Bull. 54(11) 1586—1591 (2006)
[41] S.D. Bolboacă, D.D. Rosca, L. Jäntschi, Structure-Activity Relationships from Natural Evolution, MATCH-Commun. Math. Co. 71 (2014) 149-172.
[42] Molecular Operating Environment (MOE), 2013.08; Chemical Computing Group Inc., 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2015.
[43] N. Neykov, P. Filzmoser, R. Dimova, P. Neytchev, Robust fitting of mixtures using the trimmed likelihood estimator, Comput. Stat. Data An. 52 (2007) 299-308.
[44] N.D. Thang, L. Chen, C.K. Chan, Robust mixture model-based clustering with genetic algorithm approach, Intell. Data Anal. 15 (2011) 357-373.
[45] X. Bai, W. Yao, J.E. Boyer, Robust fitting of mixture regression models, Comput. Stat. Data An. 56 (2012) 2347-2359.
[46] S.D. Bolboacă, L. Jäntschi, Quantitative Structure-Activity Relationships: Linear Regression Modelling and Validation Strategies by Example, BIOMATH 2 (2013) 1309089. doi:10.11145/j.biomath.2013.09.089.
[47] S.D. Bolboacă, L. Jäntschi, The Effect of Leverage and/or Influential on Structure-Activity Relationships, Comb. Chem. High T. Scr. 16 (2013) 288-297.
[48] Sorana D. BOLBOACĂ, Lorentz JÄNTSCHI, 2014, Sensitivity, Specificity, and Accuracy of Predictive Models on Phenols Toxicity, Journal of Computational Science, 5(3):345–350.