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:

\[
Y \sim \tilde{Y} = a \cdot X + b, \tag{1}
\]
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, and $X$ is the observed value of the predictor variable.

The array used 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 $q$ coefficient is an estimator of the power of the errors on simple linear regression.

In the SLR-LS (simple linear regression least squares), residuals $(S_i = Y_i - a \cdot X_i - b)$, where $S$ = residual follow the Gauss-Laplace distribution with $\mu$, $\sigma$, $q$, and $a$, $b$ being unknown statistical parameters:

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

where $\mu$ is population mean, $\sigma$ is population standard deviation, $q$ is power of the errors, $\Gamma$ 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 $\sigma^2$) [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 $\sum S^2_i = \min [26]$ $\sum S^2_i = \sum (Y_i - a \cdot X_i - b)^2$, 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 $\mu$ 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

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

where $b$ is the population mean of the Gauss-Laplace quantity $Y - a \cdot X$ (2), $Y$ is observed/measured dependent variable, $\hat{Y}$ 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

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

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 GL$(X, Y; \sigma, q, a, b) = \max$, and thus $\log(\Pi GL(X, Y; \sigma, q, a, b)) = \max$, which lead to the expression in

$$\sum \log(GL(X, Y; \sigma, q, a, b)) = \max. \quad (5)$$

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

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

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
Simple linear regression least squares estimation

\[ q = 2 \]

\[ a \]

\[ \mu \]

\[ \sigma \]

Algorithm maximizing likelihood as given in Equation 6

\[ \frac{\partial (\text{Equation 6})}{\partial q} \rightarrow 0, \frac{\partial (\text{Equation 6})}{\partial a} \rightarrow 0, \frac{\partial (\text{Equation 6})}{\partial \mu} \rightarrow 0, \frac{\partial (\text{Equation 6})}{\partial \sigma} \rightarrow 0 \]

equivalent with Equation \[ \frac{\partial q, a, \mu, \sigma}{\partial} \rightarrow \text{max} \]

Simple linear regression maximum likelihood estimation

\[ q \]

\[ a \]

\[ \mu \]

\[ \sigma \]

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).

2.2. Algorithm Implementation. The classical simple linear regression uses least squares method to estimate \( a, \mu, \) and \( \sigma \) 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 \( a, \mu, \) and \( \sigma \) 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: \( q, a, \mu, \sigma, \) 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 (set1a) 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 (seta), 51 iterations (setb), and 86 iterations (setc)). 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 (setlc). 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 \( \sigma) \) are provided with three decimals; equal values for \( q = 2 \) and
Table 1: Characteristics of the investigated classes of compounds.

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

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

<table>
<thead>
<tr>
<th>Set</th>
<th>SLR model</th>
<th>( R^2 )</th>
<th>s</th>
<th>F</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a</td>
<td>( \log(1/IGC_{50}) = +0.677 \cdot \log P − 1.38 )</td>
<td>0.90</td>
<td>0.22</td>
<td>287</td>
<td>35</td>
</tr>
<tr>
<td>1b</td>
<td>( \log(1/IGC_{50}) = +0.647 \cdot \log P − 1.05 )</td>
<td>0.84</td>
<td>0.30</td>
<td>666</td>
<td>126</td>
</tr>
<tr>
<td>1c</td>
<td>( \log(1/IGC_{50}) = -0.443 \cdot \log P + 0.509 )</td>
<td>0.53</td>
<td>0.57</td>
<td>276</td>
<td>250</td>
</tr>
<tr>
<td>2</td>
<td>( \log P = -0.004 \cdot ISDRTHg^* + 2.09 )</td>
<td>0.53</td>
<td>0.43</td>
<td>25</td>
<td>24</td>
</tr>
<tr>
<td>3</td>
<td>( \text{BP} = +188.40 \cdot \text{lbMdsHg}^* − 507.95 )</td>
<td>0.79</td>
<td>3.81</td>
<td>8050</td>
<td>73</td>
</tr>
<tr>
<td>4a</td>
<td>( \log P = +0.99998 \cdot \text{SD} + 5.232 )</td>
<td>0.90</td>
<td>0.32</td>
<td>92</td>
<td>40</td>
</tr>
<tr>
<td>4b</td>
<td>( \ln(\text{LD}_{50}) = +0.0018 \cdot \text{SD} − 61.168 )</td>
<td>0.41</td>
<td>0.98</td>
<td>19</td>
<td>30</td>
</tr>
<tr>
<td>5</td>
<td>( \log \text{RBA} = +0.026 \cdot \text{TIC1} − 4.145 )</td>
<td>0.36</td>
<td>1.44</td>
<td>72</td>
<td>132</td>
</tr>
<tr>
<td>6</td>
<td>( \text{pIC}_{50} = +0.255 \cdot \text{DCW} − 1.216 )</td>
<td>0.71</td>
<td>0.57</td>
<td>191</td>
<td>80</td>
</tr>
<tr>
<td>7</td>
<td>( \log K_I = -0.578 \cdot \text{N-rings} + 2.646 )</td>
<td>0.49</td>
<td>0.37</td>
<td>43</td>
<td>47</td>
</tr>
<tr>
<td>8</td>
<td>( \log(1/MC) = −4.129 \cdot \text{TEuIFFDL}^* + 5.789 )</td>
<td>0.65</td>
<td>0.38</td>
<td>64</td>
<td>37</td>
</tr>
<tr>
<td>9</td>
<td>( \text{ACE} = 47.5480 \cdot \text{IHMdphg}^* − 0.1687 )</td>
<td>0.74</td>
<td>0.33</td>
<td>128</td>
<td>47</td>
</tr>
<tr>
<td>10</td>
<td>( \ln(\text{RT}) = 0.348 \cdot \log P + 1.711 )</td>
<td>0.56</td>
<td>0.50</td>
<td>75</td>
<td>60</td>
</tr>
</tbody>
</table>

SLR = simple linear regression.

\( \log(1/IGC_{50}) = \) 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.

\( \text{std}_{\text{dim}3} = \) 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 \( \sigma \), 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 \( \sigma \)) 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 \( \sigma \) uncerned 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).

<table>
<thead>
<tr>
<th>set</th>
<th>$n$</th>
<th>$a$</th>
<th>$b = \mu$</th>
<th>$\sigma$</th>
<th>$q$</th>
<th>$q = ?$</th>
<th>$\sigma$</th>
<th>$P$ value ($H_q; q = 2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a</td>
<td>35</td>
<td>0.678</td>
<td>-1.386</td>
<td>0.218</td>
<td>9.52</td>
<td>0.638</td>
<td>-1.181</td>
<td>0.222</td>
</tr>
<tr>
<td>1b</td>
<td>126</td>
<td>0.647</td>
<td>-1.050</td>
<td>0.298</td>
<td>4.36</td>
<td>0.647</td>
<td>-1.029</td>
<td>0.298</td>
</tr>
<tr>
<td>1c</td>
<td>250</td>
<td>0.509</td>
<td>-0.443</td>
<td>0.596</td>
<td>1.29</td>
<td>0.563</td>
<td>-0.623</td>
<td>0.569</td>
</tr>
<tr>
<td>2</td>
<td>24</td>
<td>-0.004</td>
<td>2.095</td>
<td>0.414</td>
<td>0.61</td>
<td>-0.005</td>
<td>2.270</td>
<td>0.516</td>
</tr>
<tr>
<td>3</td>
<td>73</td>
<td>188.408</td>
<td>-507.959</td>
<td>3.762</td>
<td>1.34</td>
<td>188.408</td>
<td>-507.959</td>
<td>3.762</td>
</tr>
<tr>
<td>4a</td>
<td>40</td>
<td>1.000</td>
<td>5.232</td>
<td>0.308</td>
<td>2.81</td>
<td>1.041</td>
<td>5.338</td>
<td>0.308</td>
</tr>
<tr>
<td>4b</td>
<td>30</td>
<td>0.002</td>
<td>-61.168</td>
<td>0.945</td>
<td>0.67</td>
<td>0.002</td>
<td>-64.950</td>
<td>0.964</td>
</tr>
<tr>
<td>5</td>
<td>132</td>
<td>0.024</td>
<td>-3.812</td>
<td>1.374</td>
<td>1.70</td>
<td>0.026</td>
<td>-3.967</td>
<td>1.374</td>
</tr>
<tr>
<td>6</td>
<td>80</td>
<td>0.255</td>
<td>-1.216</td>
<td>0.558</td>
<td>2.87</td>
<td>0.255</td>
<td>-1.216</td>
<td>0.558</td>
</tr>
<tr>
<td>7</td>
<td>47</td>
<td>-0.578</td>
<td>2.646</td>
<td>0.360</td>
<td>3.43</td>
<td>-0.555</td>
<td>2.594</td>
<td>0.335</td>
</tr>
<tr>
<td>8</td>
<td>37</td>
<td>-4.129</td>
<td>5.789</td>
<td>0.372</td>
<td>1.29</td>
<td>-4.297</td>
<td>5.789</td>
<td>0.372</td>
</tr>
<tr>
<td>9</td>
<td>47</td>
<td>47.561</td>
<td>-0.169</td>
<td>0.319</td>
<td>3.17</td>
<td>49.502</td>
<td>-0.279</td>
<td>0.319</td>
</tr>
<tr>
<td>10</td>
<td>60</td>
<td>0.348</td>
<td>1.711</td>
<td>0.492</td>
<td>1.74</td>
<td>0.355</td>
<td>1.711</td>
<td>0.492</td>
</tr>
</tbody>
</table>

$q = \text{power of the errors}; a, b = \text{coefficients in the simple linear model}; 
\mu = \text{population mean}; \sigma = \text{population standard deviation}.$
Figure 2: Continued.
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 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).
Calculated structural descriptor

\[ y = 188.41x - 507.96 \]

\[ R^2 = 0.9913 \]

Estimated boiling point

\[ y = 0.5091x - 0.4427 \]

\[ R^2 = 0.5271 \]

Calculated global correlation descriptor

\[ y = x + 5.2324 \]

\[ R^2 = 0.7077 \]

Observed total information content index

\[ y = 0.677x - 1.3821 \]

\[ R^2 = 0.8968 \]

\[ y = 0.6466x - 1.0496 \]

\[ R^2 = 0.843 \]

Figure 3: Continued.
The relation presented in (6) may be also extended to the multiple linear regression \( Y \sim a_0 + \sum b_j X_j \) when the expression \( s_i = y_i - a x_i \) becomes \( s_i = y_i - \sum b_j x_{ij} \). 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 \((\hat{Y})\) 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 (6) did not change the correlation coefficient between \(Y\) and \(\hat{Y}\) 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 \((\phi)\) (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 \((\pm\text{influential compounds})\) 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 outliers have been identified in the dependent variable on \(X\) 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 \((\phi)\) (this led to decrease by one unit of the degrees of freedom in the analysis of variance in the regression model).

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/I38776 entitled “Model colaborativ instituțional pentru translațarea cercetării științifice biomedicale în practica clinică, TRANSCEnt”

References


