Methods and objects of chemical analysis, 2024, Vol. 19, No. 3, 123–131

Linear Regression Models and Adequacy Parameters for Scattered Data with Outliers

A. V. Panteleimonov*, D. O. Anokhin, A. B. Zakharov, I. V. Khristenko, A. I. Korobov, V. V. Ivanov

Materials Chemistry Department, V. N. Karazin Kharkiv National University, 4 Svoboda sq., Kharkiv 61022, Ukraine; *e-mail: panteleimonov@karazin.ua

Received: February 28, 2024; Accepted: May 15, 2024
DOI: 10.17721/moca.2024.123-131

In the present paper, several test samples with scattered data and outliers were examined by means of different methods for building linear regression equations. This includes ordinary least squares, least absolute deviation, orthogonal distance regression, and the least absolute deviation of orthogonal distance. Also, a variant of the weighted least squares approach was investigated in the problem of identification of outliers. New indices have been proposed to describe the outliers, which can be viewed as weighted coefficients of determination.

Keywords: linear regression, least squares, determination coefficient, outliers

Introduction

The linear regression models of various properties of molecules and, more broadly, chemical objects inevitably appear in a large number of investigations. Just name the description of results of chemical analysis and construction of quantitative structure-activity relationship (QSAR) models. The regression QSAR models of the biological activity of organic compounds especially important in description of drug effects, toxicity, mutagenity, carcinogenity, *etc*. The QSAR description of physico-chemical properties of molecular systems includes linear regression models for lipophilicity, boiling and melting points, impact factors, octane numbers, critical and thermodynamic parameters of systems, and many other characteristics of molecules and chemical processes.

An essential problem of the regression analysis relates to the scatter of input data, sometimes rather significant. A special type of scattered data in the input samples is outliers. Informally, the outliers can be treated as points which deviate significantly from the bulk of the observations or regression line. As a more rigorous definition, it can be assumed that the outliers belong to another statistical distribution than the majority of points. However, some clarification should be made. Any distribution function (e.g. normal distribution) allows unlimited deviation from the mean value. As an example, one can check the data from a standard Python package numpy [1] which we are using in the present investigation. Up to 4.6% of normally distributed random numbers deviate from the average value by two standard deviations (2σ), and more than 0.27% deviates by 3σ. Random numbers with large deviations may considerably influence the results of statistical estimations. In this respect they should also be qualified as outliers. Thus, we are faced with the issue of distinguishing “natives” and “strangers” among outliers.

In general, it was observed that high contamination is inherent not only in the physicochemical data, but also in any experimental arrays, which always contains anomalous outliers of *ca* 5-10% [2,3]. It is essentially true in the problem of building QSAR models of different biochemical parameters.

When we deal with scattered data, including outliers, the use of standard statistical approaches (correlation analysis, regression analysis, *etc*) encounters certain difficulties. Usually they are designated as “non robust” behavior of statistical methods. In particular, the standard ordinary least squares (OLS) method demonstrates incorrect behavior.

Thus, investigation of capabilities of statistical methods in the case of samples with scattered data and outliers is an important problem of contemporary chemoinformatic. After a short survey of outlier detection methods and regression analysis approaches we analyze the numerical results of suggested linear regression models which do not need the pre-tuning procedure. The main idea of this paper is to use modified (weighted) determination coefficients as an outlier detection parameter that is inherently consistent with weighted regression models. The proposed quantity also can be treated as an internal validation parameter for weighted regression models and gives the possibility to evaluate the quality of the model. In particular, it is shown that the least absolute deviation and weighted least squares methods are convenient for the purposes of identification and correct handling of outliers. To realize all the calculations within the QUASAR chemoinformatic software suite under development, appropriate modules in the Python3 scripting language have been created.

A brief review of outlier identification methods

Several approaches were proposed for correct diagnostic and interpretation of outliers, especially in the field of linear regression analysis (see e.g. [4,5] for the general outlook). Here we will briefly describe only

some most important of them.

In what follows, a linear regression is assumed in the analytical form

$$y_i = \beta_0 + \beta_1 x_i + \dots . \quad (1)$$

or, in compact matrix form:

$$Y = X\beta . \quad (2)$$

Where y_i are input values of the dependent variable (response), x_i are independent variables (explanatory variables, predictors, descriptors). In eq. (2) the "design matrix" X contains all the values x_i in columns while Y and β are vectors in which y_i and (m + 1) regression coefficients β_0, β_1,... are collected.

All the approaches to the outlier problem can be divided into two types. The first one is the identification with subsequent elimination of outliers [6]. Usually, the identification of outliers starts with the calculation of the residuals $e_i = y_i - \hat{y}_i$ for each point, y_i are calculated according to the regression model (1, 2). Then internally studentized (rs_i) and externally studentized (rt_i) residuals have to be calculated [7,8].

$$rs_i = \frac{e_i}{s\sqrt{1-h_{ii}}}, \qquad rt_i = \frac{e_i}{s_{(i)}\sqrt{1-h_{ii}}}, \quad (3)$$

where s is the usual standard deviation. In the externally studentized residuals (rt_i) s_i is the standard deviation calculated using a regression equation obtained for a sample with N-1 points yet without i^{th} point, N is the number of points (observations). Such a regular procedure is usually called leave-one-out (LOO). It is why rt_i sometimes are termed the cross-validatory or jackknife residuals [9]. The leverage for the i^{th} observation, h_{ii}, are diagonal matrix elements of the projection onto regressor matrix (diagonal elements of so called hat-matrix, or influence matrix):

$$H = X(X^+X)^{-1}X^+ \quad (4)$$

Here the superscript "+" denotes the matrix transposition operation. The external studentized residuals (rt_i) have a Student's t-distribution with the number of degrees of freedom N-m-1, where m+1 is the number of parameters in a regression model. It allows the use of Student's t-distribution as a convenient criterion to decide whether the residuals are too large. For a small number of regression parameters and rather large number of observations the critical values of the t-distribution for the 5% significance level are close to two. Hence the inequalities $|rs_i| > 2$ and $|rt_i| > 2$ can serve as a simple outlier criterion. See however [10] where the upper bound of (3) is discussed.

Also there is a group of approaches which are connected to the influence statistics. They allows to describe how an i^{th} observation affects the regression equation[1].

One of them uses the squared Mahalanobis distance [11,12,13]:

$$MD_i^2 = (u_i - C)^+ \Sigma^{-1} (u_i - C), \quad (5)$$

where $u_i = (x_{i1}, x_{i2,\dots}, y_i)$ is the point for testing, C are coordinates of centroid and Σ is a covariation matrix. Also, as an alternative to (5), there is the Cook's distance (CD) which can be calculated as follows [14,15]:

$$CD_i = \frac{\sum_j^n \left(\hat{y}_j - \hat{y}_{j/i}\right)^2}{(m+1)s^2} . \quad (6)$$

In eq. (6) $y_{j/i}$ is the evaluation of y_j by means of the regression equation obtained from a sample where i^{th} observation is omitted. Hence the CD index describes the influence of i^{th} observation on the predicted values of y. The critical values for these distances can be obtained from F-distribution [16]. Alternatively one can use a simple *rule of thumb:* point i is the outlier if

$$CD_i \geq \frac{4}{n-(m+1)} .$$

Two other diagnostics based on the distance are DFFITS and DFBETAS. The DFFITS is a measure for how strongly the i^{th} observation influences the calculated value y_i, while DFBETAS characterizes the influences of the outlier on the values of regression coefficients [17,18]:

$$DFFITS_i = \frac{e_i}{s(i)\sqrt{1-h_{ii}}}\sqrt{\left(\frac{h_{ii}}{1-h_{ii}}\right)}, \quad (7)$$

$$DFBETAS_{ij} = \frac{\beta_j - \beta_j(i)}{s(i)\sqrt{\left(X^+X\right)^{-1}_{jj}}} . \quad (8)$$

The influence of i^{th} observation on variance is measured as [18]

$$COVRATIO_i = \frac{\det\left[s^2(i)\left(X^+_{(i)}X_{(i)}\right)^{-1}\right]}{\det\left[s^2\left(X^+X\right)^{-1}\right]} . \quad (9)$$

And again index (i) corresponds to calculations with the sample where point i is omitted.

It should be emphasized that if the outliers are simply recording or experimental mistakes, they definitely have to be removed. Alternatively, if they just cannot be explained by the chosen model, but contain definite chemical information, they have to be included in the statistical study. Given that these two types of outliers are indistinguishable, the more appropriate and more general approach is to include all the points (including all outliers) in the regression analysis.

Several robust regression models have been

[1] Points with a large effect on the regression equation are called "influential points".

developed earlier for this purpose. The general description of theoretical aspects of the robust regression can be found in [14,17,19]. Here we will only mention a few most important approaches from practical point of view.

Among them is the M-estimator (maximum likelihood estimator) which has been proposed by Huber [20, 21].

$$\beta_{M-Huber} = \arg\min_{\beta} \sum \rho(t_i), \quad t_i = y_i - \beta_0 - \beta_1 x_i - \ldots, \quad (10)$$

where ρ in one of many possible implementations is defined as follows:

$$\rho(t) = \frac{1}{2}t^2, \quad \text{if } |t| < c_{out} \quad (11)$$

$$\rho(t) = c_{out}|t| - \frac{1}{2}c_{out}^2, \quad \text{if } |t| > c_{out}$$

Here the first equation for ρ works for small residuals and corresponds to standard OLS, while for larger values of residuals the second equation for ρ corresponds to least absolute deviation (LAD) method. Parameter c_{out} is a constant which depends on the intensity of outliers. The value of 1.345 is given for it as typical [22]. Various variants of M-estimators and model distributions of outlier data are discussed in [2, 22]. Some alternatives to M-estimator are Ramsay, Andrews, Turkey and other weighted regressions (see [23, 24] for review).

One more robust regression model uses the least "trimmed" squares [25] defined as follows:

$$\beta_{LTS} = \arg\min_{\beta} \sum_{i=1}^{h} e_{[i]}^2(\beta) \quad (12)$$

where the sequence of squared residuals

$$\left\{e_{[i]}^2\right\} = \left\{\left(y_i - \hat{y}_i(\beta)\right)^2\right\}$$

is arranged in ascending order of absolute values:

$$e_{[1]}^2 < e_{[2]}^2 < e_{[3]}^2 < \ldots < e_{[h]}^2 < \ldots < e_{[n]}^2. \quad (13)$$

In contrast to OLS the sum in (5) is restricted by inequality:

$$\frac{n}{2} < h \le n. \quad (14)$$

The use of a preliminary defined h-parameter makes it possible to exclude influential points from the sample.

All the above methods used for constructing regression equations have a common bottleneck, a rather artificial introduction of specific parameters which needs *a priori* evaluation. We give preference to another type of methods, black-box methods, more appropriate in our opinion. They do not require pre-tuning of the calculation parameters with the use of information about statistical distributions of errors or any heuristic parameters. An essential peculiarity of these "*closed*" methods, the subject of the present study, is an "automatic" adjustment to scatters and outliers in the data. Historically first robust approach of this type has been proposed by R. J. Boscovich in 1775 year [26], even before OLS method! It is a variant of the least absolute deviation (LAD)[2] method. However, during next two hundred years, LAD was practically not used in statistical studies because of the high computational costs. Only in 1887 the great Irish economist and statistician F. Y Edgeworth pointed out the robust properties of LAD [27]; see also [28] and references therein. For a history of the early years of robust estimations see [29].

The interest to LAD revived with increased computing power. Several algorithms were proposed for its effective implementations [30,31]. Also, in order to decrease the number of predictors, the LAD regularized with L_1-penalty (LASSO, least absolute selection and shrinkage operator [32]) has been suggested [33].

In general, it should be noted that the interest to L_1-norm and L_1-regularization in statistical science as well as in machine learning techniques has significantly grown during the last decades. For example, the approach to feature selection (LASSO) is widely used in regression analysis and in pattern recognition (see [34,35] for review). The applications of the L_1-norm in QSAR problem can be found in [36], in quantum chemistry in [37, 38].

Mention also the Theil-Sen estimator [39], in some aspect similar to the LAD approach, in which the regression line is constructed as the median of the slopes determined over all pairs of sample points.

This group of "*closed*" methods also includes the approaches where weight of points can be calculated directly from input data. One option of these weighted methods has been proposed in the present paper. There are also several publications on the orthogonal distance regression (ODR) method (admitted errors in both dependent and independent variables) [40].

In this context, an important aspect is that the asymptotic theory of robustness of LAD, ODR, and weighted LS methods are significantly more difficult in comparison with OLS. While the breakdown point (BP) for OLS can be evaluated as 1/n, very general assumptions give for LAD the estimation n/2. For other complicated "weighted" regression models it can be even more difficult. This determines the value of numerical experiments which can provide additional information about the practical usefulness of the method and its behavior in processing scattered data with outliers.

[2] In contemporary statistical literature it is sometimes called L_1-estimator or median regression.

The purpose of this paper is to study "*closed*" linear regression models in the problem of describing scattered data with outliers. We also investigated appropriate indicators of outliers, which are inherently related to the method of constructing regression models.

Selected linear regression models in outlier problem

In the present paper we examine different "*closed*" approaches for building linear regression for data with outliers. Among them the OLS, weighted least squares (WLS), LAD, ODR and recently proposed in our works Least Absolute Deviation of Orthogonal Distances (LADOD) [4, 20, 21]. We will not discuss detailed expressions, but restrict ourselves to brief description of minimization problems in matrix terms:

$$\beta_{OLS} = \arg\min_{\beta} \left\| Y - X\beta \right\|_2^2, \quad (15)$$

$$\beta_{WLS} = \arg\min_{\beta} \left\| W^{1/2}Y - W^{1/2}X\beta \right\|_2^2 \quad (16)$$

$$\beta_{LAD} = \arg\min_{\beta} \left\| Y - X\beta \right\|_1, \quad (17)$$

$$\beta_{ODR} = \arg\min_{\beta} \left\{ \left\| Y - X\beta \right\|_2^2 / \left(1 + \left\| \beta \right\|_2^2\right) \right\}, \quad (18)$$

$$\beta_{LADOD} = \arg\min_{\beta} \left\{ \left\| Y - X\beta \right\|_1 / \sqrt{1 + \left\| \beta \right\|_2^2} \right\} \quad (19)$$

Here designation of vector-matrix norm $\|.....\|_2$ corresponds to Euclidean norm, while $\|.....\|_1$ is absolute value.

Weights for our WLS method were obtained from the following considerations. If all the points lie exactly on the same line, slopes of lines connecting each pair of points are equal. Therefore, the weights can be interpreted as a measure of proximity to the average value of slopes. Outliers do not lie on the common line. As a result, lines connecting outliers with other points have significantly different slopes compared to lines between regular points. Consequently, corresponding samples of slopes have a high standard deviation. Hence, weights of these points should be small. These considerations are implemented in our algorithm. For each point, slopes of lines towards all other points are calculated. Then average value and standard deviation of these samples of slopes are calculated and afterwards average value and standard deviation of averages are also calculated. Weights are calculated from these values as follows:

$$w_i = \frac{1}{\left(\sigma_i \cdot \exp\left(\left((\overline{g}_i - \overline{g}) / \sigma_{\overline{g}}\right)^m\right)\right)^t} \quad (20)$$

Here $\overline{g}_i$ is the average of slopes for i-th point; σ_i is the standard deviation for i-th point; $\overline{g}$ is the average of averages; $\sigma_{\overline{g}}$ is the standard deviation of averages. Parameter **m** in our following calculations is equal to 4. In order to minimize errors of calculation for too small standard deviations, parameter **t** is set to $1 + 0.05 \lg(\sigma_i)$. The regular solution of WLS (16) in matrix form is well known:

$$\beta_{WLS} = \left(X^+ W X\right)^{-1} X^+ W Y. \quad (21)$$

In the present study weights are also calculated in another way, with the use of the Mahalanobis distances (5) [11,12]. Namely, the inverse of distance from i^{th} point to centroid is used as weight w_i:

$$w_i = 1 / \sqrt{MD_i^2} \quad (22)$$

This method we will designate as weighted Mahalanobis distance least squares (WMDLS).

Quality of obtained equations is characterized by the well-known determination coefficients

$$R^2 = 1 - \frac{\sum_i \left(y_i - \hat{y}_i\right)^2}{\sum_i \left(y_i - \overline{y}\right)^2}, \quad (23)$$

where y_i are input values of dependent variable, $\overline{y}$ is the average value for sample $\{y_i\}$. For the WLS and LAD/LADOD methods we propose to calculate weighted determination coefficient R_W^2.

$$R_W^2 = 1 - \frac{\sum_i w_i \left(y_i - \hat{y}_i\right)^2}{\sum_i w_i \left(y_i - \overline{y}_w\right)^2}, \quad (24)$$

$$\overline{y}_w = \sum_i w_i y_i / \sum_i w_i$$

where w_i are the weights we obtained within the corresponding method. For the LAD calculation

$$w_i = 1 / \left| y_i - \beta_0 - \beta_1 x_i - ... \right| \quad (25)$$

In the R_W^2 calculations the quantities (20) and (22) were used for WLS and WMDLS correspondingly.

Also, for the LAD approach a parameter is used, which is rather natural in the case when the minimization function is the absolute value of residue:

$$R^1 = 1 - \left(\frac{\sum_i \left| y_i - \hat{y}_i \right|}{\sum_i \left| y_i - \tilde{y} \right|} \right)^2, \quad \tilde{y} = \mathrm{median}(y) \quad (26)$$

Its weighted variant reads

$$R_w^1 = 1 - \left(\frac{\sum_i w_i \left| y_i - \hat{y}_i \right|}{\sum_i w_i \left| y_i - \tilde{y} \right|} \right)^2. \quad (27)$$

Finally, the relative difference between weighted and unweighted determination coefficients is calculated as follows:

$$\Delta R^2 = \left|R^2_w - R^2\right| / R^2_w, \quad \Delta R^1 = \left|R^1_w - R^1\right| / R^1_w \qquad (28)$$

Numerical results

The model problem in the present study is a rather simple linear equation

$$y = 1 + 0.5 \cdot x + N(0,\sigma^2) \qquad (29)$$

where normally distributed errors $N(0,\sigma^2)$ from the Python numpy package [1] with zero mean and standard deviation σ were added to the dependent variable y. The samples of 25 points each have been generated for σ = 0.1 and σ = 0.25 (designated as $S_{0.1}$, $S_{0.25}$ respectively). The independent variable x starts from 0.1, the constant step is 0.1. For each value of σ we generated two samples. The first one (designated as $S_{0.1}$(**a**), $S_{0.25}$(**a**)) corresponds to data without outliers, while the second ($S_{0.1}$(**b**), $S_{0.25}$(**b**)) contains one outlier each. In the latter case the outliers were created in S_0(**a**) samples by changing the y value for the last point.

Table 1 presents the dependence of different outlier identification parameters (see equations 3,5,9,11,13) on the distance (ℓ) between the y-coordinates for the last point of sample $S_{0.1}$, primarily generated and modified:

$$(x_0, y_0) = (2.5,\ 2.3199\text{-}\ell) \qquad (30)$$

Note that at a distance ℓ ~ 0.32 all the parameters indicate the outlier point (or at least close to critical value). The peculiarity of the LAD method is the ability to calculate the weights of the corresponding points (w_i). This makes it possible, based on a ready-made LAD solution, to estimate the R^1_w and R^2_w values from formulas (24-27) and, further, to estimate the values ΔR^2 and ΔR^1 (see eqs. 28). The proposed parameters ΔR^2 and ΔR^1 which inherently connected with the LAD regression model demonstrates the correspondence with above parameters (Table 1). This can be seen especially clearly in Fig. 1.

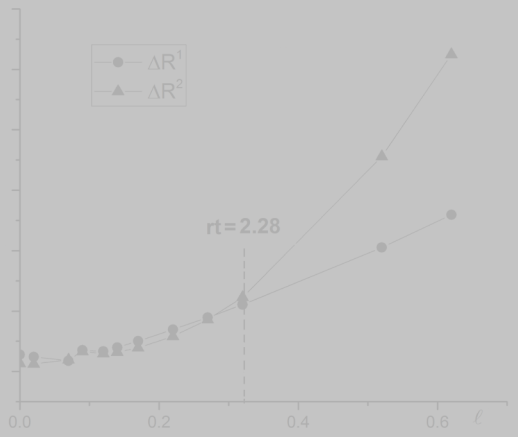


Figure 1. Relative ΔR^2 and ΔR^1 (LAD) *versus* distance (ℓ) of outlier from primarily generated point. Sample $S_{0.1}$.

In the case of $S_{0.1}$ sample with ℓ ~ 0.32 there is a definite splitting which evidences a significant difference of weighted R^2_w and unweighted R^2 determination coefficients.

Results of calculations for different regression models are presented in Fig. 2, and the corresponding numerical data are collected in Table 2. In the last column of the table, the angles (°) between lines for samples without (**a**) and with (**b**) outliers are presented. In the latter case (**b**) we use very large value of ℓ = 1.55. This guarantees samples with a significant pronounced outlier.

Table 1. Outliers identification parameters near threshold for sample $S_{0.1}$.

Parameter	Distance, ℓ				Critical value
	0.2	**0.32**	0.55	0.8	
rs_i	1.087	2.094	3.328	3.979	$t_{[0.05, df=n-2]}$ = 2.0687
rt_i	1.091	2.276	4.521	6.970	$t_{[0.05, df=n-2]}$ = 2.0687
$DFFITS_i$	0.460	0.959	1.905	2.937	$2\sqrt{(n+1)/n}$ = 0.57
$DFBETAS_0$	0.230	0.479	0.952	1.468	0.57
$DFBETAS_1$	0.394	0.822	1.633	2.517	0.57
CD_i	0.105	0.389	0.983	1.405	$F_{[\alpha=0.5,\ df_1=2,\ df_2=n-2]}$ = 0.7145
MD^2	3.705	6.543	12.474	16.687	5.49
COVRATIO	1.158	0.843	0.346	0.125	>> 1, or << 1
ΔR^1 (LAD)	0.052	0.062	0.084	0.107	
ΔR^2 (LAD)	0.050	0.065	0.121	0.221	

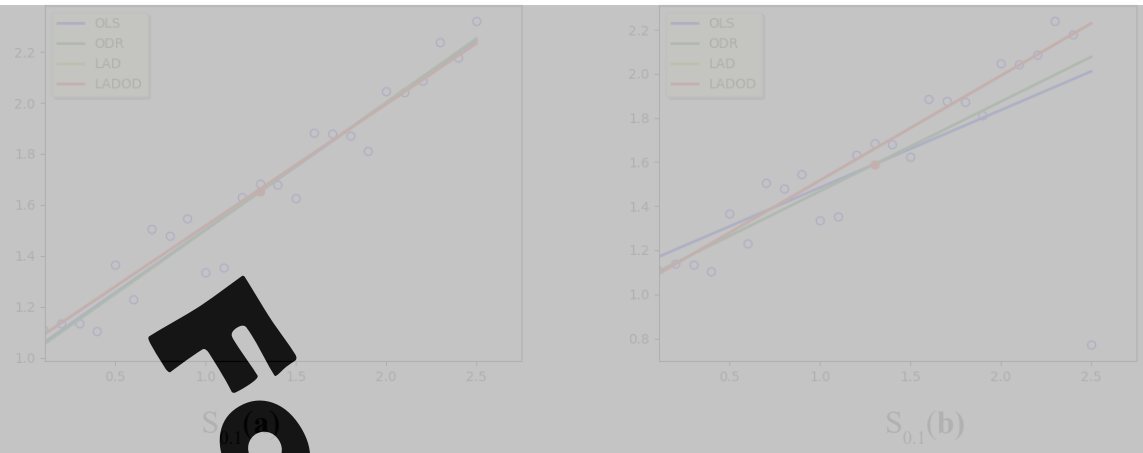


Figure 2. Different Linear regression models for data without outlier ($S_{0.1}$(**a**)) and with outlier ($S_{0.1}$(**b**)) according to Eqs. (29-30).

Figure 2 shows that in the absence of outliers all methods give closely situated regression lines. When an outlier appears ($S_{0.1}$(**b**)), this is clearly manifested by the splitting of regression lines. The OLS line is significantly shifted towards the outlier point. The ODR line is slightly less sensitive. LAD and LADOD give identical results. Corresponding equations presented in the Table 2 provide additional quantities illustrations of obtained results. Coefficients of the OLS equation undergoes a noticeable change in contrast to the LAD and LADOD methods. The angle of rotation with respect to the sample without outliers ($S_{0.1}$(**a**)) in the case of OLS is 7 degrees, while for the LAD/LADOD is only 0.1 degrees.

The specific behavior of the calculation methods for a sample with outliers is best illustrated by the comparison of the standard determination coefficient (R^2) and the weighted coefficient (R_w^2) . As long as there is no outlier in the sample, these quantities take close values for each method. For the sample with outliers the significant differences appear (0.62 for LAD, 0.64 for WLS). Therefore, it is based on R_w^2 rather than R^2 we can conclude about the adequacy of one or another regression method.

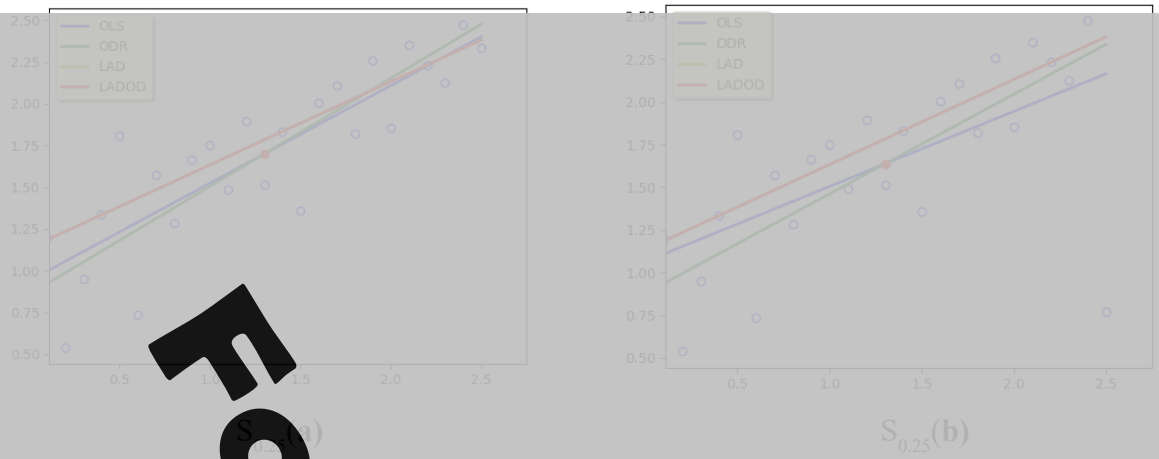
In the case of sample $S_{0.25}$(**a**) data are significantly scattered (Fig. 3, Table 3). Figure 3 shows that even in the absence of outliers the regression lines significantly differ.

The differences between R^2 and R_w^2 are large for each method even for data without outlier $S_{0.25}$(**a**). When one outlier is introduced, these differences increase markedly. And again LAD, LADOD and WLS methods demonstrate high robust properties. The angles between lines for samples without and with an outlier are equal to zero in these methods. Slightly worse results are obtained for ODR and WMDLS. Anyway, all the methods based on "automatic" weight estimates have high predictive ability that can be evaluated with the R_w^2 index.

It is logical to further increase the level of scattering of the data (Table 4). With this level of scattering all values of R^2 and R_w^2 are small, both for data with and without outlier. This can be considered as the evidence of the absolute inadequacy of the regression models. Also it can be seen that LADOD solution does not differ from LAD solution in the presented data (Table 2, Table 3) except the results obtained for large scattered data (Table 4). This result differs from our previous study data [41, 42] where specific LADOD solutions have been obtained for a number of examples.

Table 2. Regression equations parameters for model problem (29) with σ = 0.1 (see also Fig. 2).

Method	$S_{0.1}$(a)			$S_{0.1}$(b), ℓ = 1.5			angle°
	Equation	R^2	R_w^2	Equation	R^2	R_w^2	
OLS	1.008+0.495x	0.941	–	1.132+0.351x	0.444	–	7.0
WLS	1.020+0.490x	0.940	0.977	1.031+0.484x	0.345	0.957	0.3
WMDLS	1.019+0.492x	0.941	0.945	1.064+0.434x	0.410	0.726	2.7
LAD	1.042+0.477x	0.939	0.980	1.042+0.475x	0.356	0.969	0.1
ODR	1.000+0.501x	0.940	–	1.061+0.406x	0.434	–	4.5
LADOD	1.042+0.477x	0.939	0. 980	1.042+0.475x	0.356	0.969	0.1



$S_{0.25}$(**a**) $S_{0.25}$(**b**)

Figure 3. Different Linear regression models for data without outlier ($S_{0.25}$(**a**)) and with outlier ($S_{0.25}$(**b**)).

Table 3. Regression equation parameters for model problem (29) with σ = 0.25 (see also Fig. 3).

Method	$S_{0.25}$(a)			$S_{0.25}$(b), ℓ = 1.55			angle°
	Equation	R^2	R^2_w	Equation	R^2	R^2_w	
OLS	0.939+0.584x	0.718	–	1.064+0.440x	0.385	–	6.5
WLS	1.081+0.534x	0.689	0.893	1.080+0.534x	0.294	0.871	0.0
WMDLS	0.9671+0.578x	0.716	0.729	1.0268+0.501x	0.371	0.522	3.4
LAD	1.135+0.499x	0.674	0.894	1.135+0.499x	0.296	0.849	0.0
ODR	0.857+0.647x	0.710	–	0.877+0.584x	0.344	–	2.6
LADOD	1.135+0.499x	0.674	0.894	1.135+0.499x	0.296	0.849	0.0

Conclusions

In the present paper, we propose new indices based on the coefficient of determination to identify scattered data with outliers. These rather simple quantities, inextricably linked to the construction of regression equations, can be used as outlier detectors and also to assess the adequacy of the regression model as a whole. In our work we advocate an approach that we qualified as “pragmatic”. At the present level of computer technology, it is not difficult to use a number of regression approaches simultaneously. Among them we primarily emphasize “closed” approaches: OLS, ODR, LAD, and WLS. If obtained equations significantly differ, this means that the sample under study is still questionable and requires additional efforts to justify the model selection. The proposed indices expand the capabilities of the method in this direction.

Dedication

This paper is a continuation of the earlier publication “Alternative methods for constructing linear regressions” [41]. It was conceived in 2022 as a paper in the memory of professor Yu. V. Kholin (1962-2017), the main author of [41], on the occasion of his 60th birthday. Though considerably postponed because of the war, it is now, after all, published as a gratitude to the Leader and Great Man who encouraged us to deal with this issue and who has been involved in it for many years. In his extensive bibliography, a significant place is devoted to the development and implementation of computer-oriented methods, new methodology of exploring experimental data, including robust estimation, cross-validation and regularization of incorrectly posed problems [2, 43-46]. Also, as a pragmatist and authoritative specialist in the fields of physical and analytical chemistry he clearly understood and convincingly explained in his publications the applied aspects of these studies.

Conflicts of Interest

All authors declare no conflict of interest.

Acknowledgment

The work was partially supported by Ministry of Education and Science of Ukraine (grant BF/32-2021, state registration number 0121U112886).

References

1. Harris, C. R.; Millman, K. J.; van der Walt S. J. *et al. Array programming with NumPy. Nature*. **2020**, *585*, 357–362.
2. Kholin, Yu.V. Quantitative physicochemical analysis of complex formation in solutions and on the surface of chemically modified silicas: meaningful models, mathematical methods and their applications; Folio, Kharkiv, 2000, 288 p. [in Rus].
3. Huber, P. *Robust Statistics*, J. Wiley and Sons, New York, 1981, 308 p.
4. Clarke B. R. *Linear Models. The Theory and Application of Analysis of Variance*. J. Wiley & Sons, 2008, 241 p.
5. Burnett, V.; Lewis, T. *Outliers in statistical data*. J. Wiley and Sons, New York, 1978, 365 p.
6. Hawkins, D. M. *Identification of Outliers*. Chapman & Hal, 1980, 188 p.
7. Pope, A. J. *The statistics of residuals and the detection of outliers*. NOAA Technical Report NOS 65 NGS1. 1976. (https://www.ngs.noaa.gov/PUBS_LIB/TRNOS65NGS1.pdf)
8. Fox J. *Regression Diagnostics*. Sage publications, London, 1991, 92 p.
9. *Stone, M.; Cross-Validatory Choice and Assessment of Statistical Predictions. Journal of the Royal Statistical Society, Series B.* **1974**, *36 (2)*, 111-147.
10. Gray, J.B.; Woodall, W.H. The Maximum Size of standardized and Internally Studentized Residuals in Regression Analysis. *The American Statistician*. **1994**, *48 (2)*, 111-113.
11. Mahalanobis, P. C. On the Generalised Distance in Statistic. *Proceedings of the National Institute of Sciences of India*. **1936**, *2 (1)*, 49-55.
12. Maesschalck, R. De; Jouan-Rimbaud D.; Massart, D.L. The Mahalanobis distance. *Chemometrics and Intelligent Laboratory System*. **2000**, *50*, 1–18.
13. Li, X.; Deng, S.; Li, L.; Jiang Y. Outlier Detection Based on Robust Mahalanobis Distance and Its Application. *Open Journal of Statistics*. **2019**, *9*, 15-26.
14. Cook, R. D. Detection of influential observations in linear regression. *Technometrics*. **1977**, *19(1)*, 15-18.
15. Cook, R. D. *Regression Graphics*. J. Wiley and Sons, New York, 1998, 349 p.
16. Johnson, N. L.; Kotz, S.; Balakrishnan, N. *Continuous univariate distributions*. Wiley-Interscience, 1995, 732 p.
17. Rousseeuw, P. J.; Leroy, A. M. *Robust Regression and Outlier Detection*. J. Wiley and Sons, New York, 1987, 329 p.
18. *Rawlings,* J. O.; *Pantula,* S.G.; *Dickey,* D.A. *Applied Regression Analysis. A Research Tool.* 2nd Edition, *Springer*, 1998, 678 p.
19. Belsley, D. A.; Kuh, E.; Welsch, R. E. *Regression Diagnostics. Identifying Influential Data and Sources of Collinearity*. J. Wiley and Sons, New York, 1980, 292 p.
20. Huber, P. J. *Robust Statistic Procedures*. Society industrial and Applied Mathematics, Philadelphia, 1996, 67 p.
21. Huber, P.J. Robust Statistics. Maximum likelihood type estimates (M-estimates), P. 43-55. J. Wiley and Sons, 1981, New York.
22. Wang, L. ; Zheng, C. ; Zhou, W.; *et. al.* A new principle for tuning-free Huber regression. *Statistica Sinica*, **2021**, *31 (4)*, 2153-2177.
23. Yu C., Yao W. Robust linear regression: A review and comparison. *Communications in Statistics - Simulation and Computation*. **2017**, *46 (8)*, 6261-6282.
24. Yohai, V. J. High breakdown-point and high efficiency robust estimates for regression. *Annals of Statistics*. **1987**, *15(20)*, 642-656.
25. Rousseeuw, P. J.; Driessen, K. V. Computing LTS Regression for Large Data Sets. *Data Mining and Knowledge Discovery*. **2006**, *12*, 29–45.
26. Heyde, C. C.; Seneta, E. (Ed.) *Statisticians of the centuries*. Springer-Verlag, New York. 2001, 500 p.
27. Edgeworth, F. Y. On Observations Relating to Several Quantities. *Hermathena*. **1887**, *6*, 279-285.
28. Branham, R. L. Jr. Alternatives to least squares. *The Astronomical Journal*. **1982**, *87(6)*, 928-937.
29. Stigler, S. M. Newcomb S., Percy D. and the History of Robust Estimation 1885–1920. *Journal of the American Statistical Association*. **1973**, *68*, 872-879.
30. Li, Y.; Arce, G. R. A Maximum Likelihood Approach to Least Absolute Deviation Regression. *EURASIP Journal on Applied Signal Processing*. **2004**, *12*, 1762–1769.
31. Bloomfield, P.; Steiger, W. L. Least Absolute Deviations: Theory, Applications and Algorithms. *Progress in probability and statistics*. Birkhauser, **1983**, 240 p.
32. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. Roy. Statist. Soc.* **1996**, *58 (1)*, 267–288.
33. Gordon, M.; Zhu, J.; Wang, L. Regularized Least Absolute Deviations Regression and an Efficient Algorithm for Parameter Tuning. *Sixth International Conference on Data Mining (ICDM'06)*, Hong Kong, **2006**, 690-700.
34. Dodge, Y. (Ed.) Statistical Data Analysis Based on the L1-Norm and Related Methods. Springer Basel AG, **2002**, 454 p.
35. Farebrother, R. W. *L_1-Norm and L_∞-Norm Estimation: An Introduction to the Least Absolute Residuals, the Minimax Absolute Residual and Related Fitting Procedures.* Springer Berlin Heidelberg, **2013**, 58 p.
36. Berdnyk, M. I.; Zakharov, A. B.; Ivanov, V. V. Application of L_1-regularization approach in QSAR problem. Linear regression and artificial neural networks. *Methods and Objects of Chemical Analysis*. **2019**, *14 (2)*, 79-90.
37. Ivanov, V.V.; Berdnik, M. I.; Adamowicz, L.

L_1 – regularisation of the Coupled Cluster solutions. *Molecular Physics*. **2017**, *115(21-22)*, 2892-2902.

38. Ivanov, V.V. L_1-regularized solutions of coupled cluster theory equations. Test system F_2. *Kharkov University Bulletin. Chemical Series*. **2017**, *28(51)*, 30-34.

39. Fernandes, R.; Leblanc, S. G. Parametric (modified least squares) and non-parametric (Theil-Sen) linear regressions for predicting biophysical parameters in the presence of measurement errors. *Remote Sensing of Environment*. **2005**, *95*, 303–316.

40. Ahn, S. J. *Least Squares Orthogonal Distance Fitting of Curves and Surfaces in Space*. Springer-Verlag: Berlin, Heidelberg, **2004**, 127 p.

41. Onizhuk, M. O.; Ivanov, V. V.; Panteleimonov, A. V. *et al.* Alternative Methods for Constructing of Linear Regressions. *Methods and Objects of Chemical Analysis*. **2017**, *12(3)*, 105–111.

42. Berdnyk, M. I.; Onizhuk, M. O.; Ivanov, V. V. Methods for building linear regression equations in the "structure-property" problems. *Kharkov University Bulletin. Chemical Series*. **2018**, *30 (53)*, 6-17.

43. Bugaevsky, A. A.; Kholin, Yu. V. Computer-aided determination of the composition and stability of complex compounds in solutions with complicated equilibria. *Anal. Chim. Acta*. **1991**, *249 (2)*, 353–365.

44. Myerniy, S.A.; Konyaev, D. S.; Kholin, Yu.V. Robust parameter estimation in problems of quantitative physical and chemical analysis. *Kharkov University Bulletin. Chemical Series*. **1998**, *420 (2)*, 6–17. [in Rus].

45. Myerniy, S.; Varshal, G.; Kholin, Yu. Determination of affinity distributions: numerical algorithm and its application for estimating energetic heterogeneity of complexing silicas and humic substances. *Adsorption Science & Technology*. **2000**, *18 (3)*, 267–294.

46. Khristenko, I. V; Panteleimonov, A. V.; Iliashenko R. Yu.; *et al* Heterogeneous polarity and surface acidity of silica-organic materials with fixed 1-n-propyl-3-methylimidazolium chloride as probed by solvatochromic and fluorescent dyes. *Colloids and Surfaces A: Physicochemical and Engineering Aspects*. **2018**, *538,* 280-286.