https://doi.org/10.26565/2220-637X-2022-38-01 ISSN 2220-637X
УДК: 519.237.5

РОБАСТНЕ ОЦІНЮВАННЯ ПАРАМЕТРІВ РЕГРЕСІЙ. ТЕОРІЯ НЕЧІТКОСТІ ТА ІНШІ МОДЕЛІ

А. В. Пантелеймонов[a], Д. О. Анохін[b], В. В. Іванов[c]

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

a) panteleimonov@karazin.ua https://orcid.org/0000-0003-0265-1264
b) dmitriy25102002@gmail.com https://orcid.org/0000-0002-4958-2692
c) vivanov@karazin.ua https://orcid.org/0000-0003-2297-9048

Представлене співставлення результатів розрахунків параметрів лінійної регресії на основі теорії нечіткості та інших статистичних підходів. Запропоновано алгоритм простого зваженого методу найменших квадратів, що не спирається на апріорну інформацію щодо розподілу похибок вимірювань. Роботу алгоритму перевірено на модельних даних, його адекватність підтверджено на основі застосування широковживаних критеріїв. Алгоритм реалізований як окрема комп'ютерна програма на мові Python. Розроблено та верифіковано метод розрахунку розкиду нечіткої залежної змінної навколо її медіанного значення та знаходження верхньої та нижньої меж нечіткого регресійного рівняння. Доведено, що запропоновані методи можуть виступати альтернативою найвідомішим методам побудови лінійної регресії, які постулюють нормальний розподіл похибок.

***Ключові слова**: регресійний аналіз, метод найменших квадратів, метод найменших модулів, теорія нечіткості.*

Вступ

На сучасному етапі розвитку аналітичної хімії в блоці методів аналізу все більшого значення набувають інструментальні методи, а в загальнотеоретичному блоці – засоби забезпечення якості вимірювань хімічного складу. Інструментальні методи надають аналітику великі числові масиви, які необхідно зберігати, порівнювати з наявними в базах даних аналогами, обробляти, спираючись на методи інформатики та теорії аналізу даних. Забезпечення якості вимірювань висуває на авансцену аналітичної хімії метрологічну проблематику, у зв'язку з чим особливо актуальними стають проблеми розширення обсягу, підвищення точності та достовірності інформації, що вилучається з результатів вимірювань. Очевидно, що при розв'язанні цього завдання також не обійтися без інтенсивного застосування комп'ютерно-орієнтованих математичних методів. Не буде великим перебільшенням сказати, що наприкінці XX – початку XXI століття лейтмотивом застосування різних аналітичних методів стало перетворення масивів результатів вимірювань у аналітичні висновки із застосуванням теорії аналізу даних та інформаційних технологій. В умовах, коли об'єкти хімічного аналізу дуже відрізняються, а теоретичні основи використовуваних методів часто належать різним науковим дисциплінам, впроваджені в аналітичну хімію інформаційні технології та методи теорії аналізу даних залишаються, поряд з теорією пробовідбору та пробопідготовки, чи не єдиними елементами, які об'єднують розрізнені розділи аналітичної хімії.

Завдання оцінювання параметрів лінійних регресій залишається однією з найважливіших проблем сучасної статистики та хемоінформатики (хемометрії). Це пов'язано з тим, що вихідні дані, які отримуються в хімічному експерименті, часто мають значний розкид, вимірювання величин предикторів і залежної змінної не рівноточні, зустрічаються викиди (промахи) а також, доволі часто, відсутні теоретичні міркування щодо виду функціональної залежності. Особливу проблему становить також дослідження вибірок з недостатньою кількістю спостережень, а також ситуації, коли важко зробити висновки про закон розподілу похибок. Все це веде до необхідності розгляду низки альтернативних підходів до побудови рівнянь регресій.

Представлена стаття є складовою частиною циклу робіт, в яких описані різні підходи до побудови (полі)лінійних регресійних моделей, зокрема моделей QSAR (Quantitative Structure-Activity Relationship), [1-5].

Слід зазначити, що серед великої кількості підходів, останнім часом, у зв'язку з активізацією робіт в області машинного навчання, зокрема методів на основі нейронних мереж, значна увага приділяється підходам, які засновані на теорії «нечіткості» (fuzzy logic, fuzzy set) [6,7]. На відміну від традиційних теоретико-ймовірнісних підходів, в «нечітких» теоріях розглядається так звана теорія можливостей [8]. При цьому певний розкид вихідних даних вважається закладеним в самій основі розрахункової схеми. У ряді робіт реалізовано багато типових статистичних задач оцінювання та дискримінації на основі теорії нечітких даних. Значну увагу приділено і регресійним моделям. В представленій роботі розглядається альтернативний варіант побудови зваженої нечіткої задачі лінійної регресії, що орієнтована на робастне оцінювання даних.

Лінійні регресійні моделі

Спочатку розглянемо систему лінійних рівнянь вигляду

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m \tag{1}$$

де $\beta_0, \beta_1, \beta_2, \ldots$ – регресійні коефіцієнти, які підлягають оцінюванню. У матричному формулюванні задача (1) має вигляд:

$$Y = X\beta \tag{2}$$

де X – матриця значень аргументів – x_i, Y – вектор-стовпчик відгуків експериментальної системи з матричними елементами y_i, а β –вектор-стовпчик оцінюваних параметрів:

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & \cdots & x_{2m} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{N1} & x_{N2} & \cdots & x_{Nm} \end{pmatrix}, \qquad y = \begin{pmatrix} y_1 \\ y_2 \\ \ldots \\ y_N \end{pmatrix}, \qquad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \ldots \\ \beta_m \end{pmatrix}, \tag{3}$$

N – обсяг експериментальної вибірки. Оцінка параметрів такої регресії має відому матричну форму методу найменших квадратів (Ordinary Least Squares, **OLS**):

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

де символ «+» відповідає операції транспонування матриці. Традиційно нерівноточність вимірювань компенсується застосуванням зваженого методу найменших квадратів (Weighted Least Squares, **WLS**) шляхом призначення статистичних ваг, що означає внесення додаткової апріорної інформації (наприклад, про закон розподілу похибок) в досліджувану систему. В цьому випадку система (2) модифікується в такий спосіб:

$$WY = WX\beta \tag{5}$$

де W – квадратна діагональна матриця:

$$W = \begin{pmatrix} w_1 & 0 & \ldots & 0 \\ 0 & w_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & w_N \end{pmatrix} \tag{6}$$

Розв’язок рівняння (5) має вигляд:

$$\beta_{WLS} = (X^+ WX)^{-1} X^+ WY \tag{7}$$

Альтернативним підходом може виступати метод найменших модулів (Least Absolute Deviation, **LAD**) [9], який є завідомо робастним. Основою методу **LAD** є мінімізація функції «модуль» (U) за параметрами регресії

$$U_{LAD}(\beta) = \left\| Y - X\beta \right\|_1 = \sum_{i=1}^{N} \left| y_i - \beta_0 - \beta_1 x_1 - \ldots - \beta_m x_m \right| \tag{8}$$

Легко показати, що (8) можна представити як зважений за допомогою функції «модуль» метод **WLS**:

$$U_{LAD}(\beta) = \left\| Y - X\beta \right\|_l = (Y^+ - \beta^+ X^+)W(Y - X\beta) \tag{9}$$

При цьому ваговий фактор має наступне діагональне представлення:

$$W = diag(1 / \left| y_i - \beta_0 - \beta_1 x_1 - ... - \beta_m x_m \right|), i = 1,...,N \tag{10}$$

В рамках методу варіаційно-зважених квадратичних наближень [9] задача мінімізації (8) розв'язується ітераційно.

Теорія нечітких множин, як і відповідна арифметика, в простішому випадку, ґрунтується на «трикутному» нечіткому поданні числа $\tilde{Z}$ у вигляді трійки чисел $\tilde{Z} \equiv (Z,a,b)$. Тут Z – мода або чітке значення трикутного числа, a – ліва, а b – права границі числа $\tilde{Z}$.

$$a \le Z \le b \tag{11}$$

Такі величини іноді називають LR-числом (лівим-правим числом). При цьому вводиться так звана функція належності, яка дозволяє інтерпретувати будь-яке дійсне число (скажімо X) по відношенню до числа $\tilde{Z}$:

$$\mu_{\tilde{Z}}(X) = \begin{cases} \dfrac{X - a}{Z - a}, X \in [a,Z] \\ \dfrac{X - b}{Z - b}, X \in [Z,b] \\ 0, \quad Z \notin [a,b] \end{cases} \tag{12}$$

Для (12) зазвичай використовується лінгвістичне трактування. Наприклад, якщо $\mu_{\tilde{Z}}(X) \sim 1$, то X – «майже Z». Якщо X близько до границь $X \sim a$ або $X \sim b$, то X «майже не Z». У центрі інтервалу (наприклад $X \sim |Z + a|/2$): X чи то Z, чи то не Z.

Як приклад, зобразимо (Рис. 1) нечітке трикутне число $\tilde{Z} \equiv (3,0,8)$.

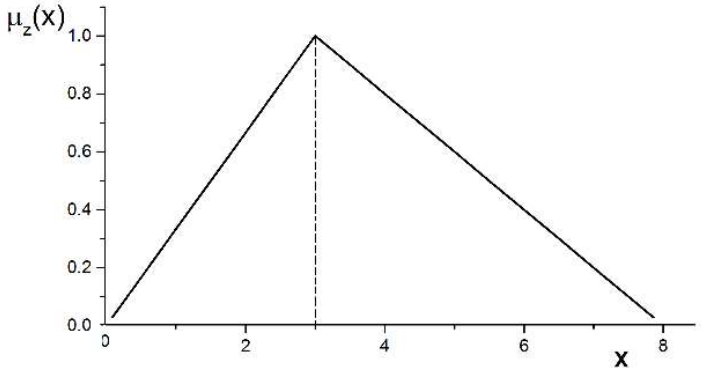


Рис. 1. Графік функції належності для числа $\tilde{Z} \equiv (3,0,8)$

Fig. 1. Graph of the membership function for the $\tilde{Z} \equiv (3,0,8)$ number

Арифметика, що реалізує операції з нечіткими числами, описана в ряді робіт (див. наприклад [10]).

Побудова відповідних рівнянь нечіткої регресії на основі методу найменших квадратів (Fuzzy Least Squares, **FLS**) вперше було описано в рамках лінійного програмування [11]. Пізніше, в роботі [12], було запропоновано розв'язок **FLS** в термінах класичного методу найменших квадратів.

У самій постановці завдання нечіткої регресії можна виявити кілька варіантів (Табл. 1). Загальний огляд різних підходів наведено, наприклад, в [13,14].

Таблиця 1. Основні підходи щодо реалізації нечіткої регресії
Table 1. Basic approaches to implementing fuzzy regression

	Залежна змінна, Y	Незалежні змінні, X	Коефіцієнти регресії, β
1	нечітка,	нечіткі,	нечіткі,
2	нечітка,	нечіткі,	чіткі,
3	нечітка,	чіткі,	нечіткі,

У представленій роботі ми розглядаємо третій варіант реалізації теорії як найбільш типовий у хімічній науці. Тобто, предиктори X є чіткими, а відгук – нечітким LR-числом $\tilde{Y} \equiv (Y, \lambda, \eta)$. Очевидно, що при цьому параметри регресії також повинні бути нечіткими. Таким чином, апроксимація, за аналогією (2), зв'язується з рішенням лінійної задачі

$$\tilde{Y} = X\tilde{\beta} \tag{13}$$

Для знаходження розв'язків за допомогою методу **OLS** (наприклад [12,15]) використовується узагальнене представлення функції, що мінімізується

$$D^2 = \left\| X\beta - Y \right\|_2^2 + \left\| Xb - \lambda \right\|_2^2 + \left\| Xc - \eta \right\|_2^2 \tag{14}$$

Тут $\left\| ..... \right\|_2^2$ – квадрат евклідової норми вектора. Цілком реалістично припустити, що відгук представляє собою симетричне нечітке число $\tilde{Y} \equiv (Y, (Y - \lambda), (Y + \lambda))$, і, як наслідок, коефіцієнти регресії також є симетричними нечіткими числами, $\tilde{\beta} \equiv (\beta, (\beta - b), (\beta + b))$. Відповідна (подібна **OLS**) система рівнянь **FLS**:

$$\frac{\partial D^2}{\partial \beta^+} = 0, \quad \frac{\partial D^2}{\partial b^+} = 0 \tag{15}$$

дозволяє знайти шукані величини аналогічно (4).

Треба відзначити, що хоч певний розкид даних і закладено у **FLS** від самого початку, тим не менш, ця теорія, строго кажучи, не є робастною, оскільки спирається на стандартний метод найменших квадратів. У даній роботі розглянуто зважений варіант **FLS** (weighted FLS, **WFLS**). В цьому випадку узагальнений вираз (14) на основі (12), для симетричних нечітких даних, може виглядати наступним чином:

$$\begin{aligned} D_W^2 &= (\beta^+ X^+ - Y^+) W (X\beta - Y) + \\ &+ ((\beta - b)^+ X^+ - (Y - \lambda)^+) W (X(\beta - b) - (Y - \lambda)) + \\ &+ ((\beta + b)^+ X^+ - (Y + \lambda)^+) W (X(\beta + b) - (Y + \lambda)) \end{aligned} \tag{16}$$

Очевидно, що за аналогією з (9), виходячи з (14), можна реалізувати також і нечіткий метод найменших модулів (Fuzzy Least Absolute Deviation, **FLAD**). В цьому випадку вагова функція має вигляд (10).

Реалізація методу WFLS для функції однієї змінної

У представленій роботі діагональні елементи матриці функцій належності, що характеризують нечіткість даних, а також діагональні елементи матриці ваг методу WFLS розраховуються виходячи з наступних міркувань. Припускаючи належність у і х до одного регресійного рівняння, при наявності лінійної функціональної залежності між цими змінними, нахил прямої повинен бути «незмінним» при переході від точки до точки. Таким чином, величини похибок вимірювань можна охарактеризувати як коливання g_{ij} навколо середнього значення. Для оцінювання цих коливань ми розраховуємо нахили прямих, що з'єднують кожну пару точок.

$$g_{ij} = \frac{y_i - y_j}{x_i - x_j};\ i = 1, 2, ..., N;\ j = 1, 2, ..., N, i \neq j \tag{17}$$

Таким чином, для кожної точки формується вибірка з N–1 нахилів. Вибірки коригуються видаленням нахилів, які найбільше відрізняються від інших. Таке коригування введено для того щоб ідентифікувати такі пари точок-викидів, які вносять похибку у визначенні ваги точки. Для кожної точки знаходиться середнє значення ($\overline{g}_i$) та стандартне відхилення (σ_{g_i}) відповідної вибірки нахилів. Також знаходиться середнє значення ($\overline{g}$) та стандартне відхилення ($\sigma_{\overline{g}}$) середніх значень нахилів. В ідеальному випадку, коли усі точки лежать на лінії, середні значення нахилів однакові для усіх точок, а стандартні відхилення усіх вибірок дорівнюють нулю. Отже, найбільші значення стандартних відхилень та найбільші відхилення від середнього відповідають точкам-викидам.

Вагові коефіцієнти (w_i) визначаються на основі стандартних відхилень нахилів. Для знаходження вагових коефіцієнтів використовуємо формулу

$$w_i = \frac{1}{\sigma_i \cdot \exp\left(\left((\overline{g}_i - \overline{g}) / \sigma_{\overline{g}}\right)^m\right)} \tag{18}$$

У випадку, коли σ_i є надто малою величиною, для зменшення впливу похибок у її визначені використовуємо адаптовану вагову функцію:

$$w_i = \frac{1}{\left(\sigma_i \cdot \exp\left(\left((\overline{g}_i - \overline{g}) / \sigma_{\overline{g}}\right)^m\right)\right)^{1+0.05lg(\sigma_i)}} \tag{19}$$

В формулах (18,19) параметр m = 2, 4; також визначає вплив відхилення нахилу від середнього на вагову функцію. В представлених нижче прикладах нами обрано m = 4.

Отримані ваги нормуються ($\sum_i w_i = 1$), а знайдені значення ваг використовуємо для визначення функції належності μ (12). При цьому припускаємо, що чим більше значення ваги точки, тим точніше вона визначається, і відповідна область її нечіткості є вужчою.

Визначення виду функції нечіткості зазвичай є досить вільним. У нашій роботі ми розглядали наступну функцію μ як величину, що залежить від ваги точки та різниці між експериментальним і теоретичним значенням залежної величини (Δy):

$$\mu_i = \frac{|\Delta y|}{2} + \frac{a}{bNw_i + 1}\sigma \tag{20}$$

В такому разі вибір параметрів a та b визначає поведінку μ у граничних випадках. Також граничні значення, за нашими представленнями, мають бути орієнтовані на стандартне відхилення регресії σ. Так, з формули (20) випливає, що при виборі параметрів a = 3 та b = 5, для рівноточного випадку маємо $Nw_i = 1$ і функція належності $\mu_i = \frac{\Delta y + \sigma}{2}$, а у випадку, коли $w_i \to 0$ – максимальна $\mu_i \to \frac{\Delta y}{2} + 3\sigma$.

Отримані матриці μ та W, відповідно до рівняння (16), використовуємо для знаходження коефіцієнтів регресії. Таким чином отримуємо три ряди залежних змінних: *Y, (Y + μ), і (Y - μ)*, і один ряд незалежних змінних X. За цими вибірками, використовуючи метод **WLS**, згідно (6) і (7) будуємо три регресійних рівняння **FWLS**:

$$\begin{aligned} y &= b_0 + xb_1 \\ (y + \mu) &= b_0^{max} + xb_1^{max} \\ (y - \mu) &= b_0^{min} + xb_1^{min} \end{aligned} \tag{21}$$

Перше рівняння є медіанним рівнянням регресії, а лінії, задані двома іншими рівняннями, обмежують смугу нечіткості лінійної моделі. Для врахування точок-викидів при побудові рівнянь-границь вагові коефіцієнти зводяться до ступеню 0.5 у порівняні з (18) та (19). Функція належності для точок до регресії є максимальною для точок медіани, а по мірі віддалення від медіани – спадає.

Описані алгоритми реалізовано нами на скриптовій комп'ютерній мові Python.

Тестові розрахунки

Результати тестових розрахунків представлено для ряду модельних даних, що включають певні викиди. В якості характеристик регресійних моделей, крім стандартних відхилень (standard deviation, SD), нами розраховано коефіцієнти детермінації R^2 що характеризують розраховані значення залежної змінної і аналогічні величини отримані за процедурою «Leave-One-Out» LOO – Q^2 [16,17].

$$R^2 = 1 - \frac{\sum_i \left(y_i - \hat{y}_i\right)^2}{\sum_i \left(y_i - \overline{y}\right)^2}, \qquad Q^2 = 1 - \frac{\sum_i \left(y_i - \hat{y}_{i/i}\right)^2}{\sum_i \left(y_i - \overline{y}\right)^2} \tag{22}$$

В формулах (22) y_i – задані (вхідні) значення залежної змінної, $\hat{y}_i$ – розраховані за отриманим лінійним рівнянням відповідні величини, $\hat{y}_{i/i}$ – величини розраховані за процедурою LOO, а $\overline{y}$ – середне значення для вибірки $\{y_i\}$. Для методів **WLS** та **LAD** нами розраховано також зважений коефіцієнт кореляції

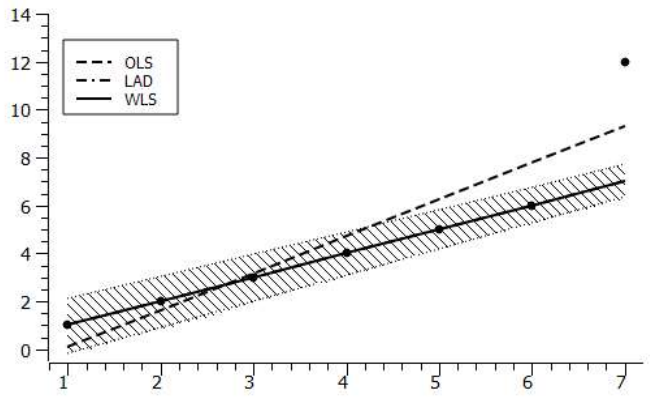
$$R_W^2 = 1 - \frac{\sum_i w_i \left(y_i - \hat{y}_i\right)^2}{\sum_i w_i \left(y_i - \overline{y}_w\right)^2}, \qquad \overline{y}_w = \sum_i w_i y_i \Big/ \sum_i w_i, \tag{23}$$

де w_i – ваги що були отримані в результаті розрахунку методом **WLS**. В методі **LAD** в якості ваг ми будемо використовувати величини

$$w_i = 1 / \left| y_i - a_0 - a_1 x_i \right| \tag{24}$$

При цьому, в методі **LAD**, в формулі (23), при розрахунках суми, будемо пропускати ті точки, через які проходить лінія регресії.

Проілюструємо чисельні дані. Спочатку розглянемо модельні ситуації з явними викидами (рис. 2).



а

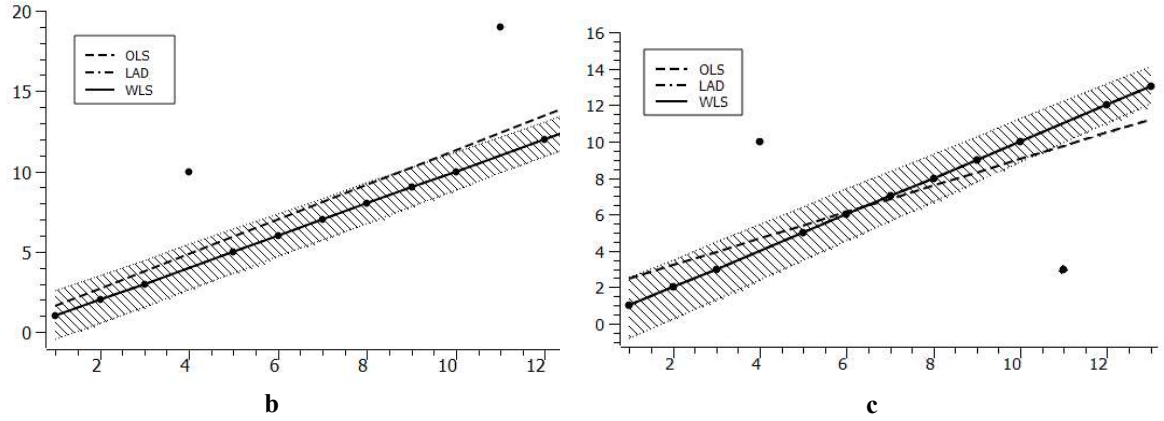


Рис. 2. Графіки-смуги регресійних моделей для вибірок з одним (**a**) та двома (**b**, **c**) викидами. В усіх трьох графіках лінії WLS та LAD співпадають.
Fig. 2. Bar graphs of regression models for samples with one (**a**) and two (**b**, **c**) outliers. In all three graphs, the WLS and LAD lines coincide.

Тут представлено результат апроксимації даних, коли одна точка (рис. 2-**a**), та дві точки (рис. 2-**b**, 2-**c**) суттєво віддалені від основної маси. При цьому основна маса точок точно лежить на лінії. Звісно, що не робастний метод **OLS** в такій ситуації дає рівняння, які значно «розгорнуті» в сторону викидів. Цікаво, що при наявності двох «випадаючих» точок з однієї сторони (рис. 2-**b**), **OLS** хоч помітно зсуває лінію регресії відносно основної маси точок, однак в цілому дає не такий поганий результат як у випадку з однією «випадаючою» точкою (рис. 2-**a**). Відзначимо, що результати отримані методом **WLS** близькі до результатів методу **LAD** і на представлених графіках ці лінії співпадають. Два викиди по різні сторони від основної густини точок (рис. 2-**c**) є особливо складною проблемою для методу **OLS,** тоді як **WLS** та **LAD** знову дають близькі (і найкращі за даної ситуації) результати.

Зафарбована площина навколо медіанної лінії функції відповідає множині точок, що належать до області нечіткості регресійної моделі (відповідає моделі **FWLS**). Можна бачити, що смуга нечіткості досить вузька і мало потерпає від точки викиду.

Отже, у випадку, коли у вибірці наявні один або більше викидів метод **WLS** (як і **LAD),** на відміну від методу **OLS**, зберігає свою робастність.

Типовий приклад регресійних залежностей із розкидом в усіх точках представлено на рис. 3. Результати розрахунків наведено в табл. 2 (для методу **WLS** наведене медіанне рівняння).

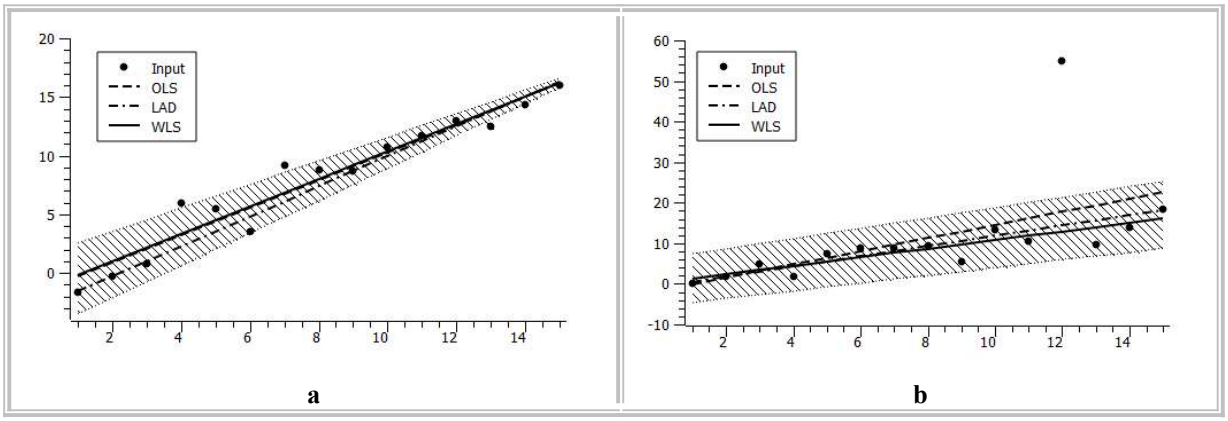


Рис. 3. Графік-смуга регресійних моделей із розкидом (**a**) без викиду та (**b**) з викидом
Fig. 3. Scatter plot of regression models **(a**) without outliers and (**b**) with outliers

Таблиця 2. Основні характеристики регресійних моделей для даних, наведених на рис. 3.
Table 2. The main characteristics of regression models for the data shown in fig. 3

Рис. 3-**a**	Рівняння $y=\beta_0+\beta_1 x$	R^2	Q^2	R^2-Q^2	R_w^2	SD
OLS	$y=-1.517+1.178\cdot x$	0.9349	0.9154	0.02	–	1.39
LAD	$y=-2.857+1.279\cdot x$	0.9177	0.8618	0.06	0.9896	1.56
WLS	$y=-1.327+1.167\cdot x$	0.9345	0.8812	0.05	0.9630	1.40
Рис. 3-**b**						
OLS	$y=-1.675+1.619\cdot x$	0.3072	0.0899	0.2	–	10.9
LAD	$y=-0.774+1.267\cdot x$	0.2695	0.1872	0.08	0.9040	11.2
WLS	$y=0.2567+1.054\cdot x$	0.2277	0.2319	-0.004	0.8587	11.5

На рис.3-**a** представлено модельні дані із розкидом, але без вираженого викиду. Тут можна бачити, що кожен метод поставляє «своє власне» рівняння, відмінне від інших. Якість рівняння, яка визначається коефіцієнтами детермінації (R^2, Q^2) та SD приблизно однакова для усіх трьох регресійних моделей. Втім різниця $R^2 - Q^2$, яка характеризує прогностичну здатність рівнянь, трохи гірша для **LAD** та **WLS** ніж для **OLS**. Смуга **FWLS** натурально звужена в області із меншим розкидом.

На рис 3-**b** представлено дані в яких разом із розкидом присутній також певний (досить значний) викид. В такій ситуації коректні оцінки надано лише методами **LAD** та **WLS**. Звісно, що звичайний коефіцієнт детермінації R^2 не може бути використаний для характеристики отриманих рівнянь. Але запропонована величина R_w^2, яка враховує ваги точок, характеризує отримані рівняння як задовільні. Відзначимо втім, що рівняння **LAD** та **WLS** суттєво відрізняються вільним членом β_0.

В останньому прикладі використано класичні дані відомого своїми піонерськими роботами в області робастної статистики П. Х'юбера [18]. Аналіз цієї задачі див. в [19]. Вибірка Х'юбера включає шість точок які відповідають лінійній моделі

$$y=-2-x \tag{25}$$

В перші п'ять точок внесено випадкову похибку із нульовим середнім значенням і стандартним відхиленням SD = 0.6. В шосту точку внесено значно більшу похибку.

Результати розрахунків представлено в табл. 3 та на рис. 4.

Таблиця 3. Характеристики регресійних моделей для задачі Х'юбера.
Table 3. Characteristics of regression models for the Huber problem.

	Рівняння $y=\beta_0+\beta_1 x$	R^2	Q^2	R_w^2	SD
OLS	$y=0.068-0.081\cdot x$	0.0824	-13.42	–	1.39
LAD	$y=-0.0333+0.0033\cdot x$	-0.0128	-14.04	-0.0067	1.46
WLS	$y=-1.325-0.608\cdot x$	-4.470	-9.98	0.6610	3.38

Не дивно, що метод **OLS** показав неадекватні результати. Можна бачити, що навіть **LAD**, який зазвичай є робастним виявився неадекватним. Вочевидь це пов'язано із малою кількістю точок в лінійній частині даних і результуюча лінія **LAD** розгорнута до точки-викиду як і **OLS** (див. рис. 4). Найкращі, хоч і не «блискучі» ($R_w^2=0.6610$) результати виявив метод **WLS**. Отримане **WLS** регресійне рівняння якісно близьке до незбуреного первинного рівняння (25). В цьому прикладі точка, що відповідає викиду веде до значного розширення смуги невизначеності **FWLS**.

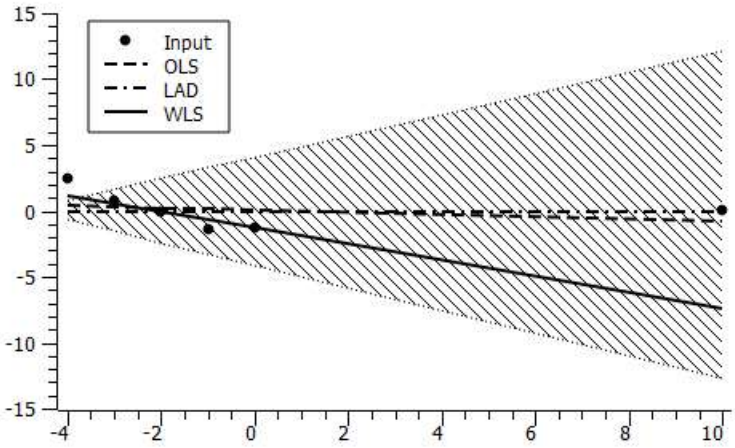


Рис. 4. Лінійні регресійні моделі для прикладу Х'юбера
Fig. 4. Linear regression models for Huber's example

Висновки

У представленій роботі запропоновано підхід для знаходження вагових параметрів зваженого методу найменших квадратів, а також максимальних і мінімальних можливих параметрів регресійного рівняння. Для призначення вагових коефіцієнтів використовується принцип рівності нахилів прямої лінії між кожною парою точок. Цей метод досить адекватно відсіює точки-викиди, призначаючи їм найменші значення ваг. Відповідно, методи **WLS** у розглянутих в даній роботі прикладах також показали доволі високу робастність. Їхньою перевагою, на відміну від методу найменших модулів, є неперервність та диференційованість функціоналу, який мінімізується. На підставі знайдених ваг запропоновано метод розрахунку розкиду нечіткої вимірюваної величини навколо її середнього значення та знаходження верхньої та нижньої меж нечіткого регресійного рівняння.

Зважаючи на вищезазначені переваги, ці методи можуть виступати альтернативою найвідомішим методам побудови регресії, які постулюють закони розподілу похибок Гауса або Лапласа. Тому їх можна запропонувати до застосування у випадках, коли ідентифікація похибки у вимірюваннях ускладнена та у разі складностей з припущенням про закон розподілу похибок.

Подяки

Робота виконувалась за часткової фінансової підтримки на виконання завдань перспективного плану розвитку наукового напрямку "Математичні науки та природничі науки" Харківського національного університету імені В. Н. Каразіна, № держреєстрації 0121U112886.

Список літератури

1. Onizhuk M.O., Ivanov V.V., Panteleimonov A.V., Kholin Yu.V. Alternative Methods for Constructing of Linear Regressions. *Method and object Chemical Analysis*. 2017, 12(3), 105-111. https://doi.org/10.17721/moca.2017.105-111
2. 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. https://doi.org/10.26565/2220-637X-2018-30-01
3. Berdnyk M.I., Zakharov A.B., Ivanov V.V. Application of L_1-regularization approach in QSAR problem. Linear regression and artificial neural networks. *Method and Object Chemical Analysis*. 2019, 14(2), 79-90. https://doi.org/10.17721/moca.2019.79-90
4. Zakharov A.B., Dyachenko A.V., Ivanov V.V. Topological Characteristics of Iterated Line Graphs in QSAR Problem: Octane Numbers of Saturated Hydrocarbons. *Journal of Chemometrics*. 2019, 33 (9), e3169. https://doi.org/10.1002/cem.3169

5. Zakharov A. B., Tsarenko D. K., Ivanov V. V. Topological characteristics of iterated line graphs in the QSAR problem: a multigraph in the description of properties of unsaturated hydrocarbons. *Struct Chem.* 2021, 32, 1629-1639. https://doi.org/10.1002/cem.3169
6. Fuzzy logic in Chemistry. Rouvray D.H. ed. Academic Press, London, 1997, 364 p.
7. Zadeh L. A. The Concept of a Linguistic Variable and its Application to Approximate Reasoning-I. *Information Sciences*. 1975, 8, 199-249. https://doi.org/10.1016/0020-0255(75)90036-5
8. Dubois D., Prade H. Possibility Theory, Probability Theory and Multiple-Valued Logics: A Clarification. *Annals of Mathematics and Artificial Intelligence*. 2001, 32, 35–66. https://doi.org/10.1023/A:1016740830286
9. Bloomfield P., Steiger W.L. Least Absolute Deviations. Theory, Applications and Algorithms. Boston: Birkhäuser, 1983, 351 p.
10. Hanss M. Applied Fuzzy Arithmetic. An Introduction with Engineering Applications. Springer-Verlag Berlin Heidelberg 2005, 256 p.
11. Tanaka H., Uegima S., Asai K. Linear Regression Analysis with Fuzzy Model. *IEEE Trans. on Systems, Man and Cybernetics*. 1982,12, 903–907. http://dx.doi.org/10.1109/TSMC.1982.4308925
12. Diamond P. Fuzzy Least Squares. *Information Sciences*. 1988, 46, 141–157. https://doi.org/10.1016/0020-0255(88)90047-3
13. de Andrés-Sánchez J. Fuzzy Regression Analysis: An Actuarial Perspective in Fuzzy Statistical Decision-Making Theory and Applications. Springer. 2016. 173-201. https://doi.org/10.1007/978-3-319-39014-7_11
14. Conventional and fuzzy regression theory and engineering applications. Hrissanthou V., Spiliotis M. (eds). Nova Science Publishers Inc, New York. 2018, 332 p.
15. Haggag M.M.M. A New Fuzzy Regression Model by Mixing Fuzzy and Crisp Inputs. *American Review of Mathematics and Statistics*. 2018, 6(2), 9-25. https://doi.org/10.15640/arms.v6n2a2
16. Golbraikh A., Tropsha A. Beware of Q2! *Journal of Molecular Graphics and Modelling*. 2002, 20, 269-276. https://doi.org/10.1016/S1093-3263(01)00123-1
17. Alexander D.L.J., Tropsha A., Winkler D.A. Beware of R2: Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models. *J. Chem. Inf. Model*. 2015, 55 (7), 1316–1322. https://doi.org/10.1021/acs.jcim.5b00206
18. Huber P.J. Robust Statistics. J. Wiley and sons, New York. 1981, 308 p.
19. Kholin Yu. V. A quantitative physicochemical analysis of complexation in solutions and on the surface of complexing silicas: meaningful models, mathematical methods and their application. Kharkiv, Folio, 2000, 294 p. (in Rus).

Надійслано до редакції 27.06.2022

Прийнято до друку 12.09.2022

A. V. Panteleimonov, D. O. Anokhin, V. V. Ivanov. Robust evaluation of regression parameters. The fuzzy theory and other models.

V. N. Karazin Kharkiv National University, 4 Svobody sq., Kharkiv, 61022, Ukraine

Linear regression parameters based on fuzzy theory are compared with other statistical approaches. A new algorithm of a simple weighted least squares method, independent of a priori information, is proposed. The algorithm was verified on model data, and its adequacy was confirmed with the use of standard criteria. The algorithm has been implemented as Python language computer program. New method of calculation of the scatter of fuzzy dependent variable around its mediane value, as well as the upper and lower bonds of fuzzy regression equations have been developed and verified. Proposed methods are shown to be useful alternatives to the most popular methods for constructing linear regression, which assume a normal distribution of errors.

Keywords: *regression analysis, Least Square method, Least Absolute Deviation method, fuzzy theory*.