Науковий вісник НЛТУ України
Scientific Bulletin of UNFU
https://doi.org/10.36930/40330212
Article received 21.03.2023
Article accepted 25.04.2023
UDC 004.048
Correspondence author: N. I. Boyko nataliya.i.boyko@lpnu.ua

ISSN 1994-7836 (print)
ISSN 2519-2477 (online)

Н. І. Бойко[1], К. П. Газдюк[2]

[1] Національний університет "Львівська політехніка", м. Львів, Україна
[2] Чернівецький національний університет ім. Юрія Федьковича, м. Чернівці, Україна

ПОРІВНЯННЯ РЕГРЕСІЙНИХ МОДЕЛЕЙ ЗА НАЯВНОСТІ ВИКИДІВ У НАБОРІ РІЗНОТИПОВИХ ДАНИХ

У дослідженні зосереджено увагу на надійній статистиці, обґрунтовано вплив надійної регресії на подолання обмежень традиційного регресійного аналізу. Закцентовано на регресійному аналізі, який моделює зв'язок між однією чи кількома незалежними змінними та залежною змінною. Описано стандартні типи регресії, такі як звичайний метод найменших квадратів, що мають сприятливі властивості. Наведено приклади оцінювання за методом найменших квадратів для регресійних моделей. Проаналізовано критерії моделей, які чутливі до викидів. Розглянуто викиди із подвійною величиною помилки, аніж типове спостереження, та з більшою величиною, що впливає на квадратичну втрату помилки, і тому має більше важелів впливу на оцінки регресії. Розглянуто аналіз лінійних моделей за оцінками параметрів за методом найменших квадратів завжди виявлялися найкращими лінійними незміщеними оцінками. Наведено порівняння властивостей цих методів, що здійснюється за допомогою моделювання. Обґрунтовано критерії порівняння їх ефективності. Досліджено критерії для M-estimators, які можуть бути вразливими до спостережень із високим важелем. Робота зосереджена на даних, які містять викиди. Досліджено їх вплив на оцінки методом найменших квадратів. Обґрунтовано застосування функції втрат Хубера, яка є надійною альтернативою стандартним квадратичним втратам помилок, та зменшує кількість викидів у квадратичні втрати помилок. Розглядається випадки втрати помилок, які обмежують їхній вплив на оцінки регресії. Досліджено алгоритм Random Sample Consensus (RANSAC) для надійної підгонки моделей. Показано його надійність у процесі аналізу викидів у експериментальних даних. Проаналізовано критерії, за яких алгоритм здатний інтерпретувати та згладжувати дані, які містять значний відсоток грубих помилок. Обґрунтовано процес генерування статистики, який покладається на звичайний метод найменших квадратів (МНК) у моделі лінійної регресії завдяки його оптимальним властивостям і простоті обчислень. Обґрунтовано МНК, який дає незміщену та мінімальну дисперсію серед усіх незміщених лінійних оцінок, коли помилки є незалежними, однаково та нормально розподіленими із середнім значенням 0 та постійною дисперсією 2σ. Показано однорідності дисперсій помилок (гомоскедастичність), що є важливим припущенням у лінійній регресії, для якої оцінки методом найменших квадратів мають властивість мінімальної дисперсії. Проведено порівняльний аналіз таких регресійних моделей: лінійна регресія (англ. Linear Regression, not Robust); регресія Губера (англ. Huber Regression); RANSA (англ. RANdom SAmple Consensus); оцінююча функція Тейла-Сена (англ. Theil-Sen Regression). У роботі проведено дослідження на вибірках із різними показниками викидів для чотирьох моделей регресій, зокрема першої не надійної (LR). Оцінено точність отриманих моделей для даних із викидами та без. Наведено дослідження, які демонструють важливість аналізу викидів у наборі даних та вибору правильного методу регресії. Розглянуто різні алгоритми, що по-різному пріоритезують важливість елементів вибірки та дають результати різної точності залежно від кількості викидів та однорідності даних.

Ключові слова: регресія; моделювання; викид; надійна регресія; *M*-оцінювання; лінійна регресія; регресія Губера.

Вступ / Introduction

Однією із задач машинного навчання є пошук прихованої залежності для вирішення задач регресійного аналізу, використовуючи навчання з наглядом (англ. *Supervised Learning*). Алгоритми мають бути навчені розуміти зв'язок між незалежними змінними та результатом або залежною змінною. Потім таку модель можна використати для прогнозування результатів нових вхідних даних або для заповнення прогалини в даних, яких не вистачає [2, 14].

Одним з найважливіших статистичних інструментів для багатьох галузей є аналіз лінійної регресії (англ. *Linear Regression*, LR). Майже весь регресійний аналіз ґрунтується на методі найменших квадратів для оцінки параметрів моделі. Проблема, з якою надзвичайно часто стикаються під час застосування регресії – це наявність викиду (outlier) або викидів у наборі даних [5].

Викид (англ. *Outlier*) – це спостереження з великим залишком. Іншими словами, це спостереження, значення залежної змінної якого є незвичайним, враховуючи його значення для передбачуваних змінних. Викид може вказувати на особливість зразка або може вказувати на помилку введення даних чи іншу проблему.

Інформація про авторів:
Бойко Наталя Іванівна, канд. екон. наук, доцент, кафедра систем штучного інтелекту. **Email:** nataliya.i.boyko@lpnu.ua; https://orcid.org/0000-0002-6962-9363
Газдюк Катерина Петрівна, д-р філософії за спеціальністю 121, асистент, кафедра програмного забезпечення комп'ютерних систем. **Email:** kateryna.gazdyik@gmail.com; https://orcid.org/0000-0002-7568-4422

Цитування за ДСТУ: Бойко Н. І., Газдюк К. П. Порівняння регресійних моделей за наявності викидів у наборі різнотипових даних. Науковий вісник НЛТУ України. 2023, т. 33, № 2. С. 84–91.
Citation APA: Boyko, N. I., & Hazdiuk, K. P. (2023). Comparison of regression models for the presence of outbreaks in a set of heterogeneous data. *Scientific Bulletin of UNFU*, 33(2), 84–91. https://doi.org/10.36930/40330212

Викиди можуть утворитись унаслідок різних причин – від простої операційної помилки до включення невеликої вибірки з іншої сукупності, і вони мають серйозні наслідки для статистичного висновку. Навіть одне віддалене спостереження може зруйнувати оцінку за методом найменших квадратів, що призведе до оцінок параметрів, які не надають корисної інформації для більшості даних. Аналіз на підставі Надійної регресії (англ. *Robust Regression*, RR) розроблено для покращення оцінки найменших квадратів за наявності викидів, щоб надати інформацію про те, що є істинним спостереженням і чи варто ці результати відкинути [1, 2, 4]. Основною метою RR-аналізу є відповідність моделі, яка представляє інформацію в більшості даних.

RR є чудовою альтернативою LR. Коли дані засмічені викидами або впливовими спостереженнями. Їх також можна використовувати для виявлення вагомих спостережень.

RR можна використовувати в будь-якій ситуації, коли використовується LR. При застосуванні LR ми можемо виявити деякі викиди або дані, що значно відрізняються від інших. Проте може виявитись, що такі записи не є помилками введення і не належать до іншої популяції, ніж решта даних. Тому немає вагомих причин виключати їх з аналізу. RR може бути хорошою стратегією, оскільки вона є компромісом між повним виключенням цих точок з аналізу та включенням усіх точок даних і однаковою обробкою їх у регресії OLS. Ідея RR полягає в тому, щоб по-різному зважувати спостереження на підставі того, наскільки добре ці спостереження поводяться [7]. Тобто це форма зваженої та переваженої регресії найменших квадратів.

Об'єкт дослідження – процес покращення оцінки найменших квадратів за наявності викидів. Аналіз викидів вхідної вибірки, створення моделі регресії, яка першочергово представлена інформацією з надійних даних.

Предмет дослідження – засоби обґрунтування точності та об'єктивності моделей RR та їх стійкість до великої кількості викидів у вибірці даних.

Мета роботи – порівняти регресійні моделі за наявності викидів під час аналізу різнотипових даних.

Для досягнення зазначеної мети визначено такі *основні завдання дослідження*:

- дослідити такі моделі RR, як регресії Губера (HR), Тейла-Сена (Theil-Sen Regression) та алгоритм RANSA (RANdom SAmple Consensus);
- порівняти та проаналізувати ефективність методів відносно один одного та відносно лінійної регресії (LR);
- порівняти точність передбачень із моделлю Linear Regression і визначити основні переваги та недоліки використання зазначених методів;
- зробити висновки про об'єктивність використання методів RR та LR для різних наборів даних.

Аналіз останніх досліджень та публікацій. Проаналізувавши роботи відомих авторів у сфері RR, можна зробити висновки, що, незважаючи на кращу продуктивність цих методів порівняно з оцінкою найменших квадратів, у багатьох ситуаціях вони все ще не використовуються широко. Кілька причин можуть допомогти пояснити їхню непопулярність. Одна з можливих причин полягає в тому, що існує кілька конкуруючих методів, і ця сфера дала багато фальстартів. Окрім цього, обчислення надійних оцінок більш затратна в обчисленнях, ніж оцінка за методом найменших квадратів; однак останніми роками це заперечення стало менш актуальним, оскільки обчислювальна потужність значно зросла. Іншою причиною може бути те, що деякі популярні пакети статистичних програм не змогли реалізувати методи [9, 13].

Хоча впровадження надійних методів було повільним, сучасні основні підручники зі статистики часто містять обговорення цих методів. До прикладу роботи Себера [3] і Лі [14] та Фаравея [1] подають загальний опис методів надійної регресії. Окрім цього, дослідження Venables [12] і Ripley [2], а також Maronna [11] описують сучасні пакети статистичного програмного забезпечення, такі як R, Statsmodels, Stata і S-PLUS та розглядають значну функціональність для надійної оцінки.

Результати дослідження та їх обговорення / Research results and their discussion

Лінійні оцінки за методом найменших квадратів можуть вести себе погано, якщо розподіл помилок не є нормальним, особливо коли помилки є heavy-tailed. Одним із засобів є видалення впливових спостережень із підгонки найменших квадратів. Інший підхід, який називають RR, полягає у використанні відповідного критерію, який не настільки вразливий, як найменші квадрати, до незвичайних даних. Найпоширенішим загальним методом RR є M-Estimaton [6]. Адже відомо, у статистиці M-оцінки (англ. *M-Estimaton*) – це широкий клас екстремальних оцінок, для яких цільова функція є вибірковим середнім. І нелінійні найменші квадрати, і оцінка максимальної правдоподібності є окремими випадками M-оцінок. Їх визначення було мотивовано надійною статистикою, яка внесла нові типи M-оцінок. Статистична процедура оцінювання M-оцінки на наборі даних називається M-оцінкою.

Загалом, M-оцінку можна визначити як нуль оцінювальної функції. Ця функція оцінювання часто є похідною іншої статистичної функції. Наприклад, оцінка максимальної правдоподібності – це точка, де похідна функції правдоподібності за відношенням до параметра дорівнює нулю; отже, оцінка максимальної правдоподібності є критичною точкою функції оцінювання. У багатьох застосуваннях такі M-оцінки можна розглядати як оцінювальні характеристики сукупності.

Цей клас оцінок можна розглядати як узагальнення оцінки максимальної правдоподібності, звідси й термін "M-оцінка". Ми розглядаємо тільки лінійну модель:

$$y_i = \alpha + \beta_i x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i = x_i^{'}\beta + \varepsilon_i, \quad (1)$$

де: y_i – змінна відповіді на i-те спостереження; $\beta_i, \beta_2, \beta_k$ – параметри. x_i – значення незалежної змінної на i-му спостереженні, ε_i – нормально розподілена випадкова величина.

Дано оцінку b для β, тоді відповідна модель має такий вигляд:

$$\hat{y}_i = \alpha + b_1 x_{i1} + b_2 x_{i2} + ... + b_k x_{ik} + e_i = x_i^{'} b, \quad (2)$$

а залишки задаються за допомогою такої формули:

$$e_i = y_i - \hat{y}_i. \quad (3)$$

За допомогою M-Estimaton запропоновано узагальнювальну оцінку максимальної правдоподібності до мінімізації. Відповідно, оцінки b визначаються шляхом мінімізації певної цільової функції за всіма b:

$$\sum_{i=1}^{n} \rho(e_i) = \sum_{i=1}^{n} \rho(y_i - x_i^{'} b), \quad (4)$$

де функція ρ дає співвідношення кожного залишку до цільової функції.

Функція ρ повинна мати такі властивості [5, 8]:

- бути додатною: $\rho(e) \geq 0$;
- дорівнювати 0, коли аргумент дорівнює 0: $\rho(0) = 0$;
- бути симетричною: $\rho(e) = \rho(-e)$;
- бути монотонною в $|e_i|, \rho(e_i) \geq \rho(e_{i'})$ для $|e_i| \geq |e_{i'}|$.

До прикладу, ρ – функція найменших квадратів $\rho(e_i) = e_i^2$ задовольняє ці вимоги аналогічно, як і багато інших функцій.

Нехай ψ є похідною від функції ρ. Похідну ψ називають *кривою впливу*. Продиференціювавши цільову функцію за коефіцієнтами *b* і прирівнявши часткові похідні до 0, отримаємо з *k*+1 оцінювальних коефіцієнтів:

$$\sum_{i=1}^{n} \psi(y_i - x_i' b) x_i' = 0. \quad (5)$$

Визначимо вагову функцію $\omega(e) = \psi(e)/e$. Нехай $\omega_i = \omega(e_i)$. Оцінювальні рівняння можна записати так:

$$\sum_{i=1}^{n} \omega_i (y_i - x_i' b) x_i' = 0. \quad (6)$$

Розв'язування цих рівнянь для оцінки еквівалентно зваженій задачі найменших квадратів, мінімізуючи $\sum \omega_i^2 e_i^2$. Однак ваги залежать від залишків. Залишки, своєю чергою, залежать від оцінених коефіцієнтів, а оцінені коефіцієнти – від ваг. Тому потрібне ітераційне рішення (так зване ітеративно перезваженими найменшими квадратами, *Iteratively Reweighted Least-Squares*, IRLS) [10]:

1. Обрати початкові оцінки $b^{(0)}$ як оцінки найменшими квадратами.
2. На кожній ітерації *t* обчислити залишки e_i^{t-1} та відповідні ваги $\omega_i^{t-1} = \omega[e_i^{t-1}]$ із попередніх ітерацій.
3. Розв'язати для нових зважених оцінок за методом найменших квадратів:

$$b^0 = \left[X' W^{t-1} X\right]^{-1} X' W^{t-1} y, \quad (7)$$

де X – модель матриці із рядками x_i' та $W^{t-1} = diag\{\omega^{t-1}\}$ – поточна матриця ваг.

Кроки 2 та 3 повторюються доти, поки оцінені коефіцієнти не збіжаться. Асимптотична коваріаційна матриця *b* має такий вигляд:

$$\upsilon(b) = \frac{E(\psi^2)}{\left[E(\psi')\right]^2} (X'X)^{-1}. \quad (8)$$

Використовуючи $\sum[\psi(e_i)]^2$, щоб оцінити $E(\psi^2)$ та $\sum[\psi'(e_i)/n]^2$. Для оцінки $\left[E(\psi')\right]^2$ утвориться оцінена асимптотична коваріаційна матриця – $\hat{\upsilon}(b)$, яка не є надійною для малих вибірок.

На рис. 1 порівнюються цільові функції та відповідні ψ і вагові функції для трьох M-Estimators: оцінки найменших квадратів; оцінювач Губера (the Huber estimator); і біквадрат (або двовагову) оцінку Тьюкі (the Tukey bisquare (or biweight) estimator).

З рис. 1 видно, що цільові функції найменших квадратів та відповідні ψ і вагові функції для трьох M-Estimators: знайомої оцінки найменших квадратів; оцінювач Губера (the Huber estimator); і біквадрат (або двовагову) оцінку Тьюкі (the Tukey bisquare (or biweight) estimator). Цільові та вагові функції для трьох оцінювачів також наведено в табл. 1.

І цільові функції найменших квадратів та цільові функції Губера безмежно збільшуються, оскільки залишок *e* відхиляється від 0, але цільова функція найменших квадратів зростає швидше. Навпаки, рівні біквадратної цільової функції зрештою вирівнюються (для $|e| > k$). Випадок одновимірної цільові функції часто трапляється на практиці; окрім цього, до нього зводяться значно складніші задачі багатовимірної оптимізації.

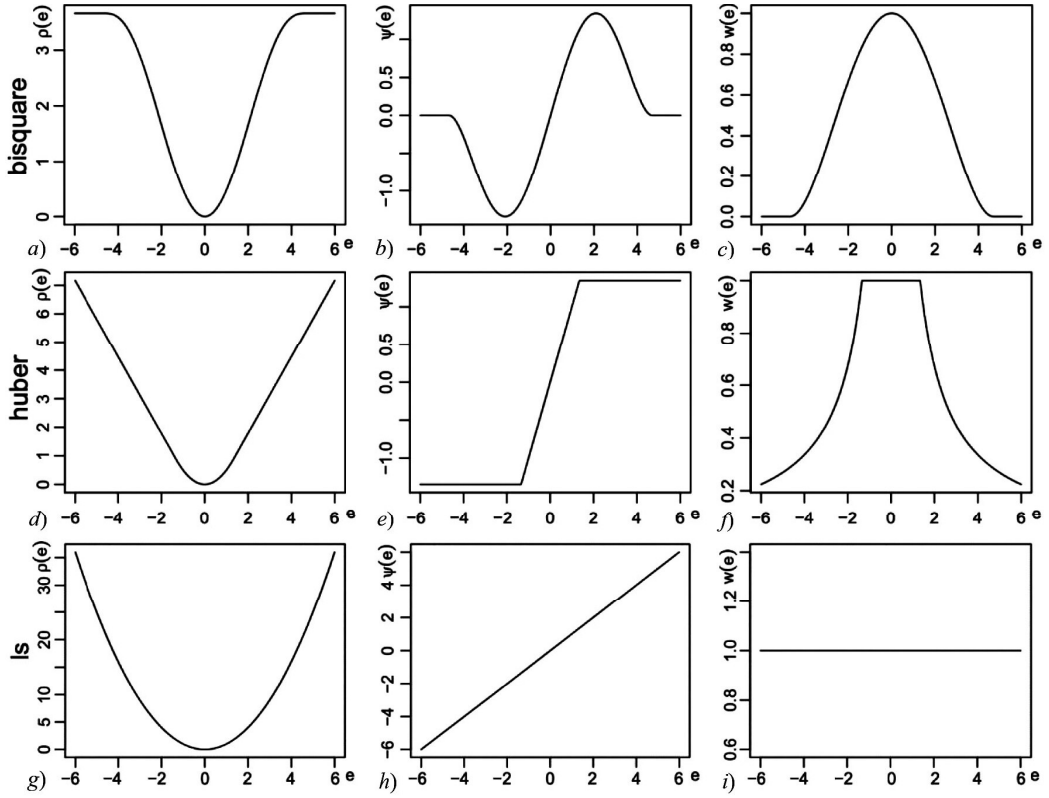

Рис. 1. Цільові функції та відповідні ψ і вагові функції / Objective functions and corresponding ψ and weight functions

Табл. 1. Цільові функції та функції ваг для методу найменших квадратів, Губера та Біквадратного оцінювачів / Objective and weight functions for least squares, Huber and Biquadratic estimators

Метод	Цільова функція	Функція Ваг
Найменших квадратів	$\rho_{LS}e = e^2$	$\omega_{LS}(e) = 1$
Губер	$\rho_H(e) = \begin{cases} \frac{1}{2}e^2 & \text{для} \quad \lvert e\rvert \le k; \\ k\lvert e\rvert - \frac{1}{2}k^2 & \text{для} \quad \lvert e\rvert > k, \end{cases}$	$\omega_H(e) = \begin{cases} 1 & \text{для} \quad \lvert e\rvert \le k; \\ \frac{k}{\lvert e\rvert} & \text{для} \quad \lvert e\rvert > k, \end{cases}$
Біквадрат	$\rho_B(e) = \begin{cases} \frac{k^2}{6}\left\{1-\left[1-\left(\frac{e}{k}\right)^2\right]^3\right\} & \text{для} \quad \lvert e\rvert \le k; \\ \frac{k^2}{6} & \text{для} \quad \lvert e\rvert > k. \end{cases}$	$\omega_B(e) = \begin{cases} \left[1-\left(\frac{e}{k}\right)^2\right]^2 & \text{для} \quad \lvert e\rvert \le k; \\ 0 & \text{для} \quad \lvert e\rvert > k. \end{cases}$

Метод найменших квадратів присвоює кожному спостереженню однакову вагу; вагові коефіцієнти для оцінки Губера зменшуються, коли $|e| > k$; і коефіцієнти для біквадрата зменшуються, як тільки e відхиляється від 0, і дорівнюють 0 для $|e| > k$.

Для оцінок Губера та біквадрата значення k називають константою налаштування (tuning constant); менші значення k створюють більшу стійкість до викидів, але за рахунок меншої ефективності, коли помилки нормально розподіляються. Константа налаштування зазвичай вибирається для забезпечення досить високої ефективності в нормальному випадку. Зокрема, k = 1,345σ – для Губера та k = 4,685σ – для біквадрата, де σ – стандартне відхилення похибок. Воно забезпечує 95-відсоткову ефективність, коли помилки є нормальними і все ще забезпечують захист від викидів [3, 11].

У програмі нам потрібна оцінка стандартного відхилення помилок, щоб використовувати ці результати. Зазвичай надійна міра поширення використовується замість стандартного відхилення залишків. Наприклад, звичайний підхід полягає в тому, щоб прийняти $\hat{\sigma} = \frac{MAR}{0.6745}$, де *MAR* (англ. *Median Absolute Residual*) – середній абсолютний залишок.

На рис. 1 наведено цільові, ψ та функції ваг для методу найменших квадратів, Губера та Біквадратного оцінювачів. Константи налаштування для цих графіків k = 1,345 – для Губера та k = 4,685 – для біквадрата. У цьому разі стандартне відхилення помилок σ =1.

Регресія обмеженого впливу (Bounded-Influence Regression). За певних обставин M-Estimators можуть бути вразливими до спостережень із високим важелем. Ключовою концепцією в оцінці впливу є точка розбивки оцінювача: точка розбивки – це частка "поганих" даних, яку оцінювач може переносити без впливу як завгодно великої міри. Наприклад, у контексті оцінки центру розподілу середнє має точку розбивки 0, оскільки навіть одне погане спостереження може змінити середнє значення на довільну величину; навпаки, медіана має точку розбивки 50 %. Були запропоновані дуже високі оцінки розбивки для регресії, і тут представлені функції R для них. Однак треба уникати дуже високих оцінок розбивки, якщо ви не впевнені, що модель, яку ви підбираєте, правильна, оскільки дуже високі оцінки розбивки не дають змоги діагностувати неправильну специфікацію моделі [3, 8].

Одним з оцінок обмеженого впливу є регресія найменших обрізаних квадратів *LTS* (*least-trimmed squares*). Потрібно упорядкувати квадрати залишків від найменшого до найбільшого так:

$$(e^2)_{(1)}, (e^2)_{(2)}, \ldots, (e^2)_{(n)}. \tag{9}$$

Оцінювач *LTS* вибирає коефіцієнти регресії b, щоб мінімізувати суму найменших обрізних квадратів залишків m:

$$LTS(b) = \sum_{i=1}^{m} e^2, \tag{10}$$

де, як правило, $m = \lfloor n/2 \rfloor + \lfloor (k+2)/2 \rfloor$ трохи більше, ніж для половини спостережень. Хоча критерій LTS легко описується, механіка підгонки оцінки LTS є складною. Більше того, оцінки з обмеженим впливом можуть давати необґрунтовані результати за певних обставин і не існує простої формули для коефіцієнта стандартної помилки.

Random Sample Consensus. Алгоритм Random Sample Consensus (RANSAC) – це алгоритм надійної підгонки моделей. Він надійний у сенсі хорошої толерантності до викидів у експериментальних даних. Він здатний інтерпретувати та згладжувати дані, що містять значний відсоток грубих помилок. Оцінка є правильною тільки з певною ймовірністю, оскільки RANSAC є рандомізованою оцінкою. Алгоритм був 14 застосований до широкого кола проблем оцінки параметрів моделі в комп'ютерному баченні, таких як відповідність ознак, реєстрація або виявлення геометричних примітивів [11].

Підвибірка вхідних даних. Структура алгоритму RANSAC проста, але потужна. Спочатку вибірки відбираються рівномірно та випадково з набору вхідних даних. Кожна точка має однакову ймовірність відбору (рівномірна точкова вибірка). Для кожного зразка будується гіпотеза моделі шляхом обчислення параметрів моделі з використанням даних вибірки. Розмір вибірки залежить від моделі, яку потрібно знайти. Отже, це найменший розмір, достатній для визначення параметрів моделі. Наприклад, щоб знайти кола в наборі даних, потрібно намалювати три точки, оскільки для визначення параметрів кола потрібні три точки. Намалювати більше, ніж мінімальна кількість точок вибірки, є неефективним, оскільки ймовірність вибору вибірки, що складається тільки з внутрішніх точок даних (тобто всіх точок даних, що належать до однієї моделі), яка дає хорошу оцінку випадковим чином, зменшується відносно збільшення розміру вибірки. Отже, мінімальний вибірко-

вий набір максимізує ймовірність вибору набору інлайнерів, з яких пізніше буде обчислена хороша оцінка [7].

Оцінка гіпотез. На наступному кроці якість гіпотетичних моделей оцінюють за повним набором даних. Функція вартості обчислює якість моделі. Загальною функцією є підрахунок кількості внутрішніх елементів (тобто точок даних, які узгоджуються з моделлю в межах допуску помилок). Гіпотеза, яка отримує найбільшу підтримку з набору даних, дає найкращу оцінку. Як правило, параметри моделі, оцінені RANSAC, не дуже точні [1]. Тому оцінені параметри моделі перераховуються, наприклад, за методом найменших квадратів, що відповідає підмножині даних, яка підтримує найкращу оцінку. Вхідні дані можуть підтримувати кілька різних моделей. У цьому разі оцінюються параметри для першої моделі, точки даних, які підтримують модель, видаляються із вхідних даних, і алгоритм просто повторюється з рештою набору даних, щоб знайти наступну найкращу модель. Сильна позиція алгоритму полягає в тому, що він, ймовірно, 15 намалює принаймні один набір точок, який складається тільки із внутрішніх даних, і дає хорошу оцінку параметрів моделі.

Змінні процесу. Методика RANSAC використовує три змінні для управління процесом оцінки моделі. Перший визначає, чи узгоджується точка даних з моделлю. Як правило, це певний допуск до помилок, який визначає об'єм, в який повинні потрапляти всі сумісні точки. Другою змінною є кількість створених гіпотез моделі. Це залежить від ймовірності намалювати вибірку, включаючи тільки внутрішні точки даних [4]. Оскільки частка викидів і мінімальний розмір вибіркової сукупності збільшуються, кількість гіпотез моделі необхідно збільшити, щоб отримати хорошу оцінку параметрів моделі. Частка викидів залежить від рівня шуму та від того, скільки моделей підтримує набір даних. Окрім цього, потрібна одна змінна допуску, щоб визначити, чи була знайдена правильна модель. Вилучена модель вважається дійсною, якщо для цієї моделі є достатня підтримка точок даних. У даних знайдено дійсне коло, якщо, наприклад, знайдено принаймні 20 точок даних, які лежать досить близько до кола. У разі кількох моделей у даних витягується більше моделей, доки не буде недостатньо підтримувати інші моделі.

Покращення тривалості виконання. Обчислювальну ефективність алгоритму можна значно підвищити кількома способами. Швидкість залежить від двох факторів: по-перше, кількості вибірок, які необхідно відібрати, щоб гарантувати певну впевненість для отримання хорошої оцінки; по-друге, час, витрачений на оцінку якості кожної гіпотетичної моделі. Останній пропорційний розміру набору даних. Як правило, дуже велика кількість гіпотез створюється із забруднених зразків (тобто зразків, що містять викиди) [6]. Такі моделі узгоджуються тільки з невеликою частиною даних. Оцінку моделей можна оптимізувати обчислювально шляхом рандомізації оцінки. Будь-яка гіпотетична модель спочатку перевіряється тільки з невеликою кількістю випадкових точок даних із набору даних. Якщо модель не отримує достатньої підтримки від цього випадкового набору точок, тоді можна з високою впевненістю припустити, що модель не є хорошою оцінкою. Моделі, які пройшли рандомізоване оцінювання, потім оцінюються на підставі повного набору даних. Продуктивність алгоритму погіршується зі збільшенням розміру вибірки або у випадку, якщо кілька моделей підтримуються даними через зменшення ймовірності вибірки набору, який повністю складається з інлайнерів. Поширеним спостереженням є те, що викиди мають дифузний розподіл. Навпаки, внутрішні дані мають тенденцію розташовуватися близько один до одного. Отже, рівномірна вибірка точок замінюється відбором вибіркових наборів на підставі близькості з урахуванням просторових відносин. Перша початкова точка вибірки вибирається випадковим чином. Решта точок є випадковими точками, що лежать у межах гіперсфери з центром у першій точці. Вибір вибіркових наборів суміжних точок може значно підвищити ймовірність вибору набору внутрішніх точок і, в такий спосіб, різко зменшити кількість вибірок, необхідних для пошуку хорошої оцінки моделі.

Результати дослідження та їх обговорення / Research results and their discussion

Проведемо порівняльний аналіз таких регресійних моделей:

- Лінійна регресія (Linear Regression, not Robust);
- Регресія Губера (Huber Regression);
- RANSA (RANdom SAmple Consensus);
- оцінювальна функція Тейла-Сена (Theil-Sen Regression).

Проведемо навчання моделей за допомогою вибірок із різними співвідношеннями звичайних даних та викидами. Обчислимо такі метрики точності передбачень для даних із викидами та без:

- коефіцієнт детермінації передбачення (Score);
- абсолютна середня похибка (MAE, Mean Absolute Error);
- середня квадратична похибка (MSE, Mean Squared Error);
- стандартне відхилення (STD, Standard deviation).

Перед початком експериментів згенеруємо датасет для задачі регресії на 1000 елементів. Розділимо отриманий датасет на тренувальну і тестові вибірки у пропорції 8:2

Експеримент № 1. Згенеруємо 100 штучних викидів для тренувальної вибірки із випадковими дисперсіями ($3 \le \sigma \le 5$) (рис. 2).

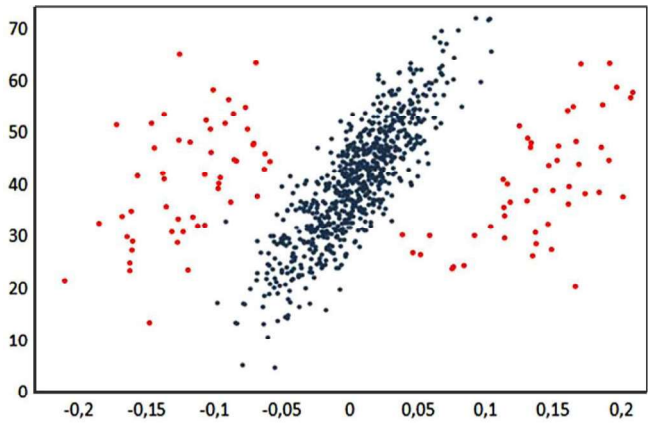

Рис. 2. Тренувальна вибірка із 100 викидами / Training sample with 100 outliers

Із наведеної вище візуалізації даних (див. рис. 2) можна легко розрізнити викиди від основних даних та помітити їх тісну лінійну залежність.

Для покращення візуального сприйняття усі викиди було перефарбовано у червоний колір (див. рис. 2).

Побудуємо та візуалізуємо досліджувані регресійні моделі (рис. 3).

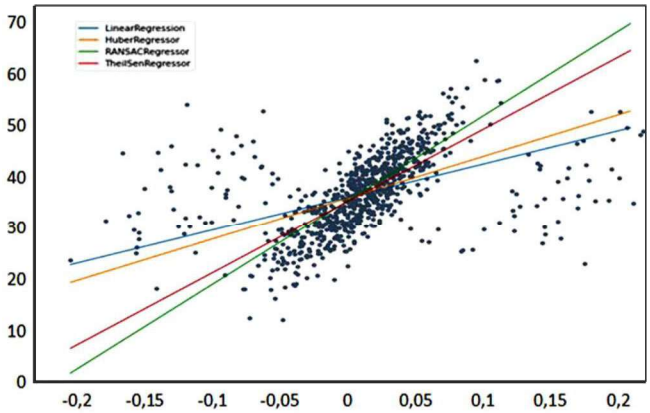

Рис. 3. Візуалізація ліній регресій досліджуваних моделей / Visualization of regression lines of the studied models

Із наведеної вище візуалізації регресійних моделей варто відзначити високу стійкість алгоритму RANSAC та регресії Тейла-Сена до викидів даних, які практично ідеально описують основну частину вибірки. Водночас, HR є достатньо наближеною до LR та більш залежною до викидів даних. У табл. 2 наведено оцінку точності передбачення. Для апріорної оцінці якості розробленого передбачення мають бути встановлені критерії, за якими він може бути оцінений. Як правило, в оцінці передбачення бере участь диференційний або інтегральний метод. При диференціальному методі визначають набори оцінок окремих складових якості передбачення, мають досить чіткий об'єктивний сенс. Зокрема, можуть використовуватися такі критерії, як ясність і чіткість завдання на передбачення, відповідність передбачення завданням, своєчасність розроблення передбачення, професійний рівень розроблення передбачення, надійність використаної інформації і т.д.

Табл. 2. Метрики регресійних моделей, натренованих на вибірці із 100 викидами / Metrics of regression models trained on a sample with 100 outliers

	Score/train	Score/test	MAE/train	MAE/test	MSE/train	MSE/tets	STD/train	STD/tets
Linear Regression	0,2206	0,3833	1,3279	1,2004	2,8320	2,1422	0,8952	0,5333
Huber Regressor	0,2067	0,4540	1,3197	1,1285	2,8823	1,8965	1,1193	0,6668
RANSAC Regressor	-0,3076	0,6528	1,4008	0,8819	4,7513	1,2060	2,2755	1,3556
TheilSen Regressor	-0,0799	0,6247	1,3517	0,9241	3,9236	1,3036	1,9400	1,1557

Із наведених метрик можна підбачити не надто високу точність передбачень LR, що зумовлено наявністю великої кількості викидів у даних та порівняно кращу ефективність методів RR.

Експеримент № 2. Проведемо аналогічні до Експерименту № 1 дії, проте згенеруємо тільки 10 викидів для тієї самої величини вибірки (рис. 4).

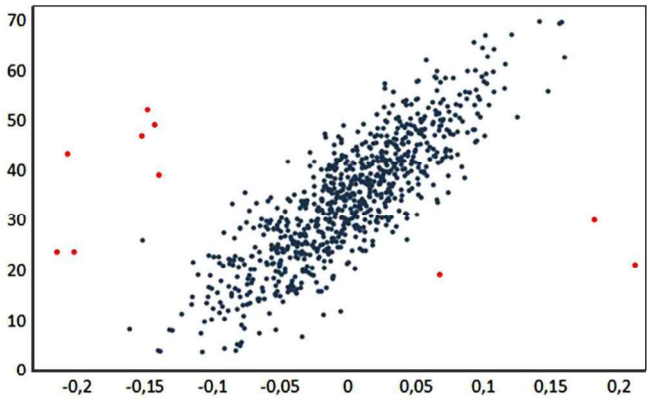

Рис. 4. Тренувальна вибірка із 10 викидами / Training sample with 10 outliers

На отриманій візуалізації нових тестових даних, можемо помітити значно меншу кількість викидів у загальній вибірці. Побудуємо та візуалізуємо відповідні досліджувані регресійні моделі (рис. 5).

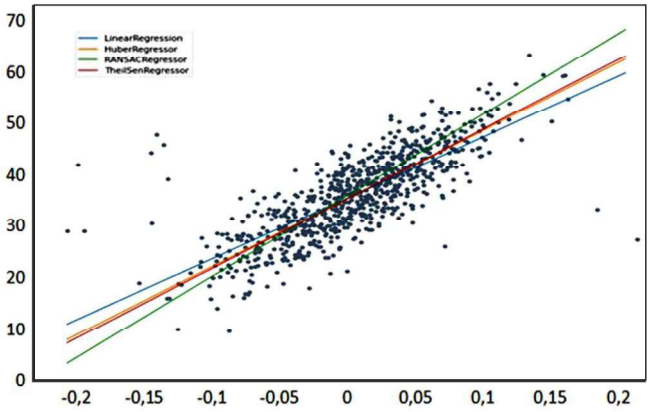

Рис. 5. Візуалізація ліній регресій досліджуваних моделей / Visualization of regression lines of the studied models

Надзвичайно важливим спостереженням буде те, що алгоритм RANSAC та регресія Тейла-Сена майже не змінили свого стану порівняно із зображеним на рис. 3, що зайвий раз підкреслює їхню стійкість. Натомість HR поводиться достатньо подібно LR. У табл. 3 наведено оцінку точності передбачення.

Табл. 3. Метрики регресійних моделей, натренованих на вибірці із 10 викидами / Metrics of regression models trained on a sample with 10 outliers

	Score/train	Score/test	MAE/train	MAE/test	MSE/train	MSE/tets	STD/train	STD/tets
Linear Regression	0,6008	0,6600	0,8843	0,8741	1,4503	1,1810	1,4776	1,4248
Huber Regressor	0,5924	0,6607	0,8702	0,8777	1,4810	1,1785	1,6520	1,5931
RANSAC Regressor	0,5300	0,6144	0,9081	0,9417	1,7079	1,3394	1,9644	1,8943
TheilSen Regressor	0,5883	0,6587	0,8700	0,8819	1,4959	1,1856	1,6900	1,6297

На відміну від табл. 2 із попереднього експерименту точність LR покращилась істотно із зменшенням кількості викидів. Також варто зауважити що попри високу ефективність LR у даному випадку, методи RR поводять себе досить гідно, що підтверджує тезу про можливе заміщення LR у всіх можливих випадках, без істотних втрат у точності.

Отже, в роботі проведено дослідження на вибірках із різними показниками викидів для чотирьох моделей регресій, зокрема першої не надійної (LR). А також оцінили точність отриманих моделей для даних із викидами та без. Як і очікувалось, стандартне відхилення (STD) для лінійної регресії завжди мінімальне порівняно зі трьома іншими моделями, оскільки вона рівноцін-

но враховує усі вхідні дані, і не розрізняє викиди. Із рис. 3 та 5 видно, що алгоритм RANSAC та регресія Тейла-Сена мають найбільшу стійкість до викидів, і керуються інформацією із більшості даних.

З табл. 1, де навчання та передбачення проводились на вибірці із співвідношенням викидів до інших даних як 1:8 (train), бачимо, що найкращий коефіцієнт детермінації, для моделі LR, що знову ж таки передбачувано із наведених вище причин. Достатньо близьким до LR показником володіє HR, та найгірший для алгоритму RANSAC. Така градація пояснюється ступенем залежності моделі від викидів.

З іншого боку, якщо поглянути на результати передбачень для вибірки без викидів (для тих же моделей, які навчені на вибірках із викидами), то побачимо абсолютно протилежну ситуацію, де для LR показник детермінації є найнижчим, а для алгоритму RANSAC, відповідно, найвищий. Аналогічні результати бачимо для обчислених похибок (MAE та MSE): LR натренована на даних із викидами, дає найбільші похибки передбачень для даних без викидів, попри те, що для вибірки із викидами результати доволі не погані, порівняно із надійними методами.

Варто проаналізувати результати табл. 2, де навчання моделей проводилось над даними із значно нижчим показником викидів, а саме 1:80. У цьому разі LR здобуває абсолютну першість практично за всіма показниками, як для вибірок із викидами так і без, адже їх частка у навчальних даних є незначною, що дозволяє надавати усім вхідним даним рівноважливість. Проте надійні методи все ще скеровуються значеннями більшості даних, пріоритезуючи їх, що й відображається на точності передбачень.

Обговорення результатів дослідження. Поширеною проблемою в регресійному аналізі є наявність викидів. За визначенням Барнетта та Льюїса [14]), викиди – це спостереження, які здаються несумісними з рештою даних. У цьому дослідженні розглянуто викиди, які можуть виникнути внаслідок незвичайних, але зрозумілих подій, таких як помилкове вимірювання, неправильний запис даних, несправність вимірювального приладу тощо. Тому наводимо різні підходи для їх аналізу.

Також розподіл із сильними хвостами зазвичай породжує викиди, і вони можуть мати помітний вплив на оцінки параметрів (Chatterjee and Hadi, [2]). Rousseeuw і Leroy [1]) для класифікації викидів. Тому порівняння робастних методів оцінки параметрів регресії у напрямі *y* проводиться за допомогою симуляційного дослідження та реального застосування. Дослідження присвячено критеріям оцінки та порівняння регресійних оцінок.

Також актуальним є дослідження середньоквадратичної помилки (MSE) та її інтерес для регресійного аналізу за наявності викидів. У роботі наводимо ефективність різних оцінок, що вивчаються за допомогою моделювання та реальних даних для викидів у напрямі *y*.

Отже, за результатами виконаної роботи можна сформулювати такі наукову новизну та практичну значущість результатів дослідження.

Наукова новизна отриманих результатів дослідження – удосконалено методики вибору алгоритмів RR для пріоритизації елементів вибірки для отримання результатів різної точності залежно від кількості викидів та однорідності даних.

Практична значущість результатів дослідження – можна використати методику вибору алгоритмів RR в навчальному процесі під час викладання дисципліни Машинне навчання. Адже у процесі проведених експериментів, було розроблено методику для покращення точності результатів регресійних задач із великою кількістю викидів через використання методів робасної регресії.

Роботу, в майбутньому, можна розширити для оброблення викидів у напрямі *x*; хороші очки кредитного плеча. Іншим можливим майбутнім напрямком є порівняння надійних методів, коли викиди є як у напрямі *y*, так і в напрямі *x* (погані точки кредитного плеча).

Висновки / Conclusions

У дослідженні здійснено порівняння різних підходів регресійних задач, в яких утворюється велика кількість викидів, та проаналізовано методи робасної регресії. Експериментально доведено, що при побудові регресійних моделей реальних систем і процесів вказані вище припущення виконуються не завжди. Досліджено, що здебільшого їх невиконання призводить до некоректності застосування процедур класичного регресійного аналізу і потребує застосування більш складних методів аналізу різнотипових даних.

З'ясовано, що проведені дослідження демонструють важливість аналізу викидів у наборі даних та вибору правильного методу регресії, зокрема тому, що різні алгоритми по-різному пріоритезують важливість елементів вибірки та дають результати різної точності залежно від кількості викидів та однорідності даних. Відповідно до проведеного дослідження, зроблено такі висновки:

З'ясовано, що якщо кількість викидів вхідної вибірки є незначна, то найоптимальнішим буде використання саме LR, хоча HR та регресія Тейла-Сена майже не поступаються їй.

Встановлено, що в разі, коли показник викидів перевищує норму, проте не є надто великим, найкращим вибором моделі стане саме HR, яка власне достатньо близька за всіма показниками до LR, проте враховує загальну частку викидів і показує більш точні передбачення.

Застосовано експерименти, внаслідок виконання яких спостерігається досить велика кількість викидів у навчальній вибірці. Зрозуміло, що найдоцільніше використовувати алгоритм RANSAC або ж регресію Тейла-Сена, які показали себе найбільш стійкими до значних відхилів у даних.

References

1. Abeida, H. (2021). Singular Non-circular Complex Elliptically Symmetric Distributions: New Results and Applications. *Mathematics and Statistics, 9*(6), 1019–1033. https://doi.org/10.13189/ms.2021.090618
2. Ahmed, M. G., & Maha, E. Q. (2016). Regression Estimation in the Presence of Outliers: A Comparative Study. *International Journal of Probability and Statistics, 5*(3), 65–72. https://doi.org/10.5923/j.ijps.20160503.01
3. Besson, O., Abramovich, Y., & Johnson, B. (2016). Direction of arrival estimation in a mixture of K-distributed and Gaussian noise. *Signal Process,* 128, 512–520. https://doi.org/10.1016/j.sigpro.2016.05.027
4. Greco, M., & Gini, F. (2013). Cramer-Rao lower bounds on covariance matrix estimation for complex elliptically symmetric distributions. Trans. *Signal Process,* 21, 6401–6409. https://doi.org/10.1109/TSP.2013.2286114

5. Hippert-Ferrer, A., El Korso, M. N., & Breloy, A. (2021). Guillaume Ginolhac. Robust Mean and Covariance Matrix Estimation Under Heterogeneous Mixed-Effects Model. HAL Open science. HAL Id: version 1. URL: https://hal.science/hal-03156771
6. Ollier, V., El Korso, M. N., Boyer, R., Larzabal, P., & Pesavento, M. (2017). Robust calibration of radio interferometers in non-gaussian environment. Trans. *Signal Process,* 65, 5649–5660. https://doi.org/10.1109/TSP.2017.2733496
7. Ollila, E., & Koivunen, V. (2009). Influence function and asymptotic efficiency of scatter matrix based array processors: Case MVDR beamformer. *Transactions on Signal Processing, 57*(1), 247–259. https://doi.org/10.1109/TSP.2008.2007347
8. Ollila, E., Tyler, D., Koivunen, V., & Poor, H. (2012). Complex elliptically symmetric distributions: Survey, new results and applications. *Transactions on Signal Processing, 60*(11), 5597–5625. https://doi.org/10.1109/TSP.2012.2212433
9. Serenko, A. (2011). Student satisfaction with Canadian music programmes: the app.lication of the American Customer Satisfaction Model in higher education. *Assessment & Evaluation in Higher Education, 36*(3), 281–299. https://doi.org/10.1080/02602930903337612
10. Temizer, L., & Turkyilmaz, A. (2012). Implementation of Student Satisfaction Index Model in Higher Education Institutions. *Procedia – Social and Behavioral Sciences*, 46, 3802–3806. https://doi.org/10.1016/J.SBSPRO.2012.06.150
11. Tenenhaus, A., & Tenenhaus, M. (2011). Regularized Generalized Canonical Correlation Analysis. *Psychome-trika, 76*(2), 257–284. https://doi.org/10.1007/s11336-011-9206-8
12. Tenenhaus, A., & Tenenhaus, M. (2014). Regularized generalized canonical correlation analysis for multiblock or multigroup data analysis. *European Journal of Operational Research, 238*(2), 391–403. https://doi.org/10.1016/j.ejor.2014.01.008
13. Tenenhaus, M., Hanafi, M. (2010). A bridge between PLS path modeling and multi-block data analysis. Handbook of Partial Least Squares, 99–123. https://doi.org/10.1007/978-3-540-32827-8_5
14. Wijnholds, S., Van Der Tol, S., Nijboer, R., & Van der Veen, A. (2009). Calibration challenges for future radio telescopes. *Signal Processing Magazine*, 1, 30–42.

N. I. Boyko[1], K. P. Hazdiuk[2]

[1] Lviv Polytechnic National University, Lviv, Ukraine
[2] Yuriy Fedkovych Chernivtsi National University, Chernivtsi, Ukraine

COMPARISON OF REGRESSION MODELS FOR THE PRESENCE OF OUTBREAKS IN A SET OF HETEROGENEOUS DATA

The study focuses on reliable statistics investigating the impact of reliable regression on overcoming the limitations of traditional regression analysis. Emphasis is placed on regression analysis, which models the relationship between one or more independent variables and a dependent variable. Standard types of regression such as ordinary least squares are described, which have favourable properties. Examples of least squares estimation for regression models are given. The criteria of models that are sensitive to emissions are analyzed. Outliers with twice the error magnitude than the typical observation are considered, and with a larger magnitude that affects the squared error loss and therefore have more leverage on the regression estimates. The considered analysis of linear models according to parameter estimates by the method of least squares always turned out to be the best linear unbiased estimates. A comparison of the properties of these methods is given using simulation. The criteria for comparing their effectiveness are substantiated. Criteria for M-estimators that may be vulnerable to highly leveraged observations are investigated. The work is focused on data that contain emissions. Their influence on estimations by the method of least squares is investigated. The application of the Huber loss function, which is a reliable alternative to the standard squared error loss, and reduces the number of outliers in the squared error loss, is justified. Cases of error loss that limit their impact on regression estimates are considered. The Random Sample Consensus (RANSAC) algorithm for reliable model fitting is investigated. Its reliability in the process of analyzing emissions in experimental data is shown. The criteria under which the algorithm can interpret and smooth data containing a significant percentage of gross errors are analyzed. The process of generating statistics is analyzed, which relies on the ordinary least squares (OLS) method in a linear regression model due to its optimal properties and simplicity of calculations. The MNC is substantiated, which gives unbiased and minimum variance among all unbiased linear estimators when the errors are independent, identically and normally distributed with mean 0 and constant variance 2σ. Homogeneity of error variances (homoscedasticity) is shown, which is an important assumption in linear regression, for which least squares estimate have the property of minimum variance. A comparative analysis of the following regression models was carried out: linear regression (Linear Regression, not Robust); Huber regression (Huber Regression); RANSA (RANdom SAMple Consensus); the Theil-Sen regression function. In the work, a study was conducted on samples with different outliers for four regression models, in particular, the first non-reliable (LR). The accuracy of the obtained models was evaluated for data with and without outliers. Studies are presented that demonstrate the importance of analyzing outliers in a data set and choosing the right regression method. Different algorithms that prioritize the importance of sample elements in different ways and give results of different accuracy depending on the number of outliers and the homogeneity of the data are considered.

Keywords: regression; modeling; emission; reliable regression; *M*-evaluation; linear regression; Huber regression.