https://orcid.org/0000-0002-9399-191X

УДК 519.237.5

С.М. ЛАПАЧ[*]

ВИЗНАЧЕННЯ ВИКИДІВ У КОРЕЛЯЦІЙНОМУ І ОДНОВИМІРНОМУ РЕГРЕСІЙНОМУ АНАЛІЗІ

[*]Національний технічний університет України «КПІ ім. Ігоря Сікорського», м. Київ, Україна

Анотація. *У роботі розглядається метод визначення викидів при попередній обробці даних для кореляційного та одновимірного регресійного аналізу. Пропонується класифікація викидів відповідно до їх природи. Розроблений для їх визначення алгоритм базується на послідовному застосуванні методу складного ножа та елементів розрахунків відстані за LSD-критерієм з використанням замість середніх значень коефіцієнтів кореляції і довірчих інтервалів замість критичних значень відстані. Застосування алгоритму показано на прикладах для різних умов його використання. Розглянутий весь спектр можливих умов застосування: від таких, в яких метод безпомилково визначає викиди, до таких, в яких він не працює або результати його роботи неоднозначні. Встановлено, що описаний у роботі метод дозволяє автоматично визначати викиди незалежно від їх природи при лінійному зв'язку між змінними. При нелінійному зв'язку між змінними викиди визначаються у випадку лінеаризації змінних (лінеаризація в таких ситуаціях є необхідною умовою аналізу даних). У сумнівних і невизначених випадках і при складній залежності алгоритм визначає як викиди ті експерименти, видалення яких приводить до покращення лінійної регресійної моделі (a priori, до її побудови). Метод не працює у випадку викидів, які компенсують одне одного, але в таких ситуаціях наявність викидів не приводить до суттєвого зміщення коефіцієнтів моделі. Побудовані регресійні моделі, які показують зміну характеристик і значень регресійних коефіцієнтів моделі під впливом викидів. Використання методу дозволяє визначити сумнівні точки експериментів для подальшого прийняття рішень. Метод може бути використаний у системах автоматизованої обробки інформації, оскільки гарантовано автоматично визначає наявність викидів або покращує лінійну регресійну модель.*

Ключові слова: *кореляція, одновимірна регресія, метод складного ножа, LSD-критерій, викиди, аномальні спостереження, неоднорідність факторного простору.*

Аннотация. *В работе рассматривается метод определения выбросов при предварительной обработке данных для корреляционного и одномерного регрессионного анализа. Предлагается классификация выбросов в соответствии с их природой. Разработанный для их определения алгоритм базируется на последовательном использовании метода складного ножа и элементах расчета расстояний по LSD-критерию с использованием вместо средних значений коэффициентов корреляции и доверительных интервалов вместо критического расстояния. Использование алгоритма показано на примерах для различных условий его применения. Рассмотрен весь спектр возможных условий использования: от таких, в которых метод безошибочно определяет выбросы, до таких, в которых он не работает или результаты работы неоднозначны. Установлено, что описанный в работе метод позволяет автоматически определять выбросы независимо от их природы при линейной связи между переменными. При нелинейной связи между переменными выбросы определяются в случае линеаризации переменных (линеаризация в таких ситуациях является необходимой частью анализа). В сомнительных и неопределенных ситуациях и при сложных зависимостях алгоритм определяет в качестве выбросов те эксперименты, удаление которых приводит к улучшению линейной регрессионной модели (a priori, до ее построения). Метод не работает в случае выбросов, компенсирующих друг друга, но в таких ситуациях их наличие не приводит к существенным изменениям коэффициентов модели. Построены регрессионные модели, показывающие изменение характеристик и значений регрессионных коэффициентов под влиянием выбросов. Использование метода позволяет определить сомнительные точки эксперимента для дальнейшего принятия решений. Метод может быть употреблен в автоматических системах обработки информации, поскольку гарантированно определяет наличие выбросов или улучшает линейную модель.*

Ключевые слова: *корреляция, одномерная регрессия, метод складного ножа, LSD-критерий, выбросы, аномальные наблюдения, неоднородность факторного пространства.*

Abstract. *In this paper, it is considered the method of determination of emissions in case of preliminary processing of data for correlation and one-dimensional regression analysis. The classification of emissions according to their nature is proposed. The algorithm developed on their determination is based on the consistent application of the jack-knife technique and the elements of distance calculations using the LSD-criterion, using instead of the mean values of the correlation coefficients and confidence intervals instead of the critical distances. The application of the algorithm is shown in the examples for different conditions of its use. The whole spectrum of possible conditions of application is considered: from those in which the method determines the emissions accurately, to those in which it does not work or the results of its work are ambiguous. It is established that described in the paper method allows to automatically determine the emissions, regardless of their nature, in the linear relationship between the variables. In the nonlinear relationship between variable emissions, they are determined in the case of linearization of variables (linearization in such situations is a prerequisite for data analysis). In questionable and uncertain cases, and with complex dependence, the algorithm defines outliers as those experiments, the removal of which leads to an improvement of the linear regression model (a priori, prior to its construction). The method does not work in the case of outliers that compensate each other, but in such situations, the presence of outliers does not lead to a significant shift of the model coefficients. The regression models have been constructed, which show changes in the characteristics and values of regression coefficients of the model under the influence of outliers. Using the method allows to identify questionable experiment points for further decision making. The method can be used in automated information processing systems, since it automatically guarantees the presence of outliers or improves the linear regression model.*
Keywords: *correlation, one-dimensional regression, jack-knife technique, LSD-criterion, emissions, abnormal observations, heterogeneity of factor space.*

DOI: 10.34121/1028-9763-2019-4-126-138

1. Вступ

При використанні в дослідженнях парної кореляції і одновимірного регресійного аналізу існує проблема визначення так званих викидів, яку ще називають проблемою неоднорідності простору або аномальних значень. Хоча, по своїй природі, причини названих явищ різні, з точки зору пересічного дослідника вони мало відрізняються у випадку, коли вказаних елементів мало (один – три). Він повинен визначити ці експерименти і прийняти рішення про подальшу роботу: залишити їх у вибірці чи відкинути. Від правильності дій на даному етапі цілком залежить правильність прийняття рішень після обробки результатів експерименту.

У випадку, коли в таких дослідженнях відсутні дублюючі, тобто проведені в номінально однакових умовах експерименти, то завдання визначення викидів ускладнюється. Як правило, їх рекомендують визначати візуально, за діаграмою розсіяння експериментальних даних [1–3].

У літературі описується велика кількість показників, які повинні допомагати визначити такі спостереження для регресійного аналізу, а саме видалений залишок, показник розбалансування h_i, показник розбалансування, центрований h^*_i, видалений стьюдентизований залишок, зміну відгуку при видаленні спостереження DFFIT (і DFFITS стандартизована), зміну коефіцієнтів регресії при видаленні залишку DFBETA (і DFBETAS стандартизована). Всі ці показники вимагають багаторазового (по кількості спостережень) перерахування моделі регресії і її аналізу. Крім того, пропонують використання відстаней Махаланобіса і Кука, а також коваріаційного відношення з зауваженням, що вони можуть виділити тільки підозрілі спостереження з подальшим аналізом [4–8].

Тобто, всі показники і методи вимагають, по-суті, не просто прийняття рішень, а і аналіз, часто неоднозначний.

При реальному використанні, як правило, діаграми розсіяння для візуального аналізу ніхто не будує. Виконання перерахунків моделі також не виконується. А якщо врахувати, що в різних системах обробки даних, як правило, тільки вказуються умови відбору даних і отримується результат, то зрозуміло, що є нагальна потреба у процедурах формаль-

ного визначення наявності викидів у вибірці. Оскільки як кореляція, так і парна регресія стали просто інструментом масового використання користувачами без достатньої підготовки в галузі математичної статистики, то потрібен інструмент виявлення викидів для масового застосування. Фактично випадки з так званими викидами можна розділити на три категорії. До першої категорії відносяться власне викиди, тобто результати, отримані через помилки персоналу, несправність обладнання, недотримання умов експерименту і безперечно не можуть бути включені в навчальну вибірку. Викиди необхідно видалити з вибірки і обробляти дані без них. До другої категорії відноситься ситуація, при якій у вибірці присутні дані із двох генеральних сукупностей, тобто випадок неоднорідності факторного простору. При неоднорідності простору його необхідно розбити на однорідні підвибірки і далі проводити обробку в кожній підвибірці окремо [9]. Останній випадок – аномальні спостереження. Аномальні значення є частиною досліджуваної генеральної сукупності, а назву аномальності носять у зв'язку з низькою ймовірністю їх появи. Наприклад, випробування на міцність до зруйнування завжди має деяку частку значень, які дуже сильно відрізняються від середнього в той чи інший бік. У випадку аномальних значень обробку необхідно виконувати з їх урахуванням. У першій і другій категоріях викиди належать до іншої генеральної сукупності, а в останньому – до тієї ж самої. При малій кількості викидів першу і другу категорії можливо розрізнити між собою тільки за рахунок смислового аналізу.

Таким чином, проблема полягає у відсутності формального методу, який дозволяв би виділити «викиди» для кореляційного аналізу і побудови одновимірної регресії. Такий метод дозволяв би автоматизувати процес визначення викидів і забезпечив зменшення прийняття неправильних рішень.

Розглядається ситуація з обмеженою кількістю аномальних експериментів невизначеної природи.

Отже, *метою статті* є розробка методу, який дозволяє формалізовано визначати наявність викидів та визначення меж його застосування.

2. Опис методу визначення викидів

Ідея запропонованого методу у використанні елементів методу складного ножа [10] і множинного LSD-критерію [11, 12] для виділення тих експериментів, які відрізняються від більшості елементів вибірки. При наявності викидів коефіцієнт кореляції в підвибірці з і без викиду повинен суттєво розрізнятись. Розділення на групи значень дозволяє виконати LSD-критерій.

Пропонується такий алгоритм визначення.

1. Розрахунок кореляцій методом складного ножа. Послідовно відкидається по одному експерименту і обчислюються коефіцієнти кореляції.

2. Впорядкування вектора кореляцій за абсолютною величиною в порядку зменшення.

3. Знаходження різниці між сусідніми значеннями.

4. Для масиву різниць розраховуються середнє, середньоквадратичне і довірчий інтервал.

5. Ті елементи, для яких значення різниці виходить за межі довірчого інтервалу в бік перевищення, вважаються викидами.

Обмеження методу теоретичне: лінійна залежність між змінними або ж така, в якій домінуючою складовою є лінійна.

3. Приклади застосування

Для ілюстрації використання методу розглянемо приклади, в яких він використовується в різних умовах з різними результатами.

Ми можемо виділити чотири варіанти умов, в яких може застосовуватись метод:

1) умови, в яких метод працює без обмежень;

2) ситуації, при яких метод працює при певних передумовах;

3) умови, де результати роботи вимагають прийняття вольових рішень;

4) обставини, в яких метод не працює.

Для ілюстрації використано квартет Енскомба (табл. 1) [13], дані з роботи Дрейпера і Сміта (табл. 6) [4] і спеціально створені приклади.

Таблиця 1 – Дані квартету Енскомба

Квартет Енскомба (№)								
№ експ.	I		II		III		IV	
	x	y	X	y	x	y	x	y
1	10	8,04	10	9,14	10	7,46	8	6,58
2	8	6,95	8	8,14	8	6,77	8	5,76
3	13	7,58	13	8,74	13	12,74	8	7,71
4	9	8,81	9	8,77	9	7,11	8	8,84
5	11	8,33	11	9,26	11	7,81	8	8,47
6	14	9,96	14	8,1	14	8,84	8	7,04
7	6	7,24	6	6,13	6	6,08	8	5,25
8	4	4,26	4	3,1	4	5,39	19	12,5
9	12	10,84	12	9,13	12	8,15	8	5,56
10	7	4,82	7	7,26	7	6,42	8	7,91
11	5	5,68	5	4,74	5	5,73	8	6,89
Дисперсія	11	4,127269	11	4,127629	11	4,12262	11	4,123249
Середнє	9	7,500909	9	7,500909	9	7,5	9	7,500909

3.1. Приклади, де метод без сумніву працює, включаючи вироджений

Розрахунки за трьома прикладами (штучний і два Енскомба) приведені в табл. 2.

3.1.1. Штучний приклад для двох викидів

Розглядається лінійна залежність із двома викидами (рис. 1). Розрахунки містяться в перших трьох колонках табл. 2. За розрахунками викиди чітко виділяються, що видно з рис. 2, на якому приведені значення різниць.

3.1.2. Квартет Енскомба III

Даний випадок з одним викидом тривіальний і викид легко виділяється (колонки 4–6) в табл. 2.

3.1.3. Вироджений випадок, квартет Енскомба IV

Цей приклад цікавий тим, що він вироджений, тобто після відкидання аномального спостереження залишається один експеримент, повторений кілька разів. Як видно з табл. 2, метод працює і в цьому випадку.

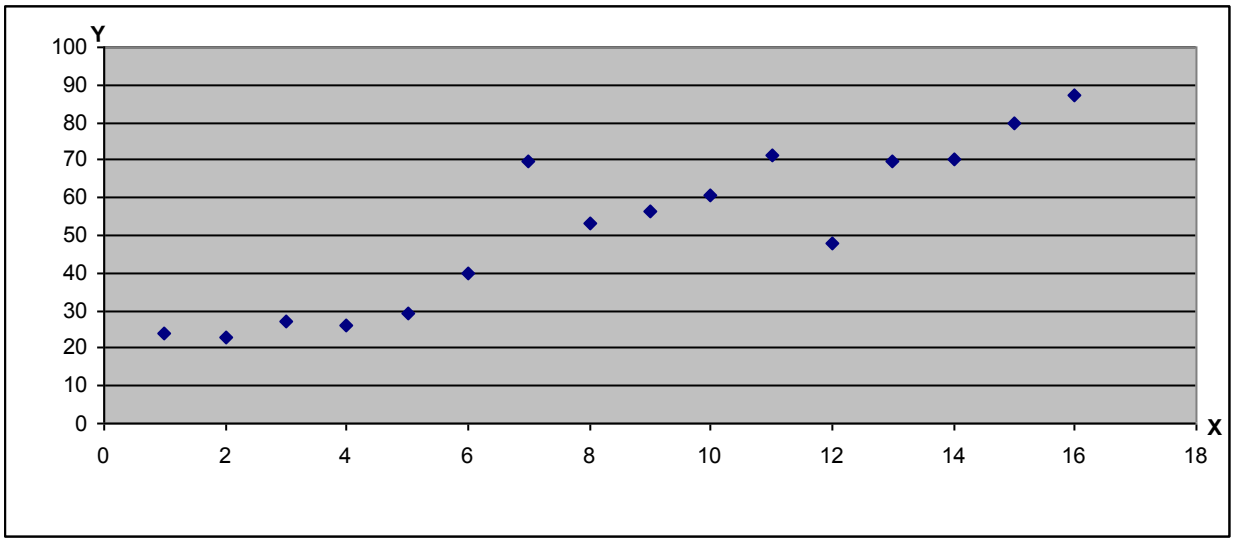

Рисунок 1 – Наявність двох викидів. Штучний приклад

Таблиця 2 – Розрахунки для визначення викидів при виконанні передумови лінійності

Штучний приклад для двох викидів			Енскомба III			Енскомба IV		
№ виключеного експерименту	Коефіцієнт кореляції	Різниця	№ виключеного експерименту	Коефіцієнт кореляції	Різниця	№ виключеного експерименту	Коефіцієнт кореляції	Різниця
7	0,955014		7	-0,45576		4	0,863525	
12	0,942007	0,013007	10	-0,2296	0,226163	5	0,847224	0,016301
11	0,913386	0,028621	11	-0,22458	0,005017	7	0,842704	0,00452
8	0,912719	0,000667	5	-0,18557	0,039009	9	0,831671	0,011033
9	0,912137	0,000582	3	-0,18512	0,000446	10	0,829204	0,002468
10	0,911456	0,000681	1	-0,17098	0,014142	2	0,82616	0,003044
5	0,910544	0,000912	6	-0,16888	0,0021	3	0,824548	0,001612
6	0,910208	0,000336	4	-0,1662	0,002686	6	0,815383	0,009166
14	0,909095	0,001113	9	-0,12895	0,037246	1	0,814788	0,000595
13	0,907648	0,001447	2	-0,04945	0,079501	11	0,814672	0,000115
4	0,906914	0,000734	8	-0,03523	0,014221	8	0	0,814672
3	0,90286	0,004054						
1	0,900511	0,002349						
15	0,900115	0,000396						
2	0,898729	0,001386						
16	0,892543	0,006186						
Середнє		0,004165			0,042053			0,086352
Середньоквадратичне відхилення		0,007543			0,040911			0,255959
Напівширина довірчого інтервалу		0,003696			0,082964			0,158642
Верхня межа довірчого інтервалу		0,007861			0,040911			0,244994

3.2. Ситуації, в яких метод працює при певних умовах

Виходячи з основних положень методу, він не буде працювати для нелінійних залежностей.

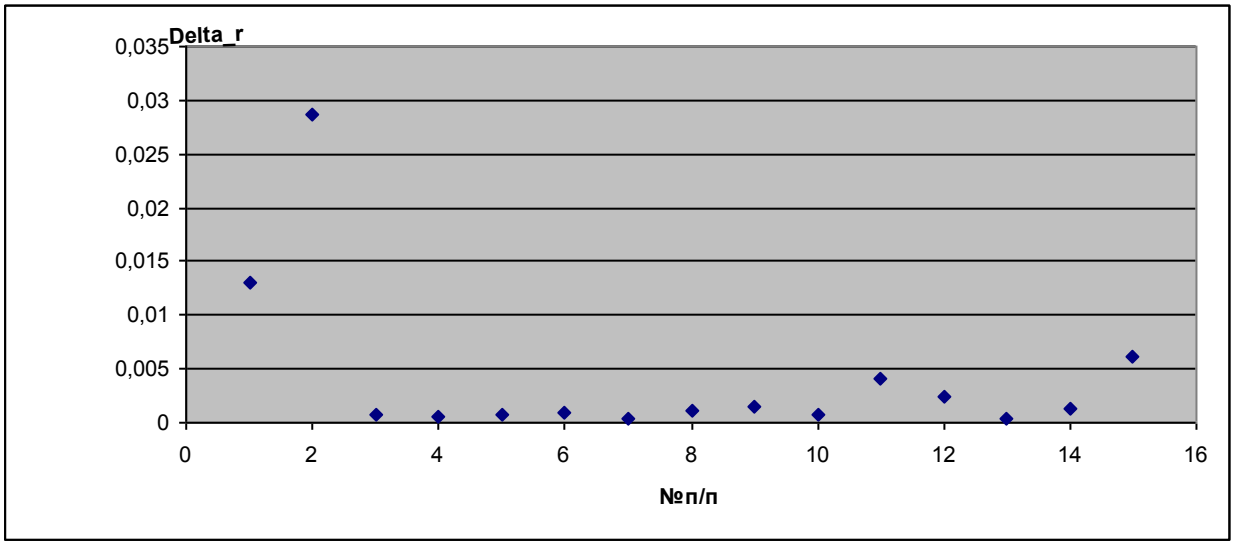

Рисунок 2 – Відхилення значень коефіцієнтів кореляції для випадку із двома викидами

Розглянемо штучний приклад параболічної залежності (рис. 3) з одним викидом. У разі розрахунків за даними, як вони є, викид не виділяється. Якщо виконати лінеаризацію заміною $x' = x^2$, викид впевнено виділяється (табл. 3, колонки 1–3). Зауважимо, що в таких ситуаціях без лінеаризації виявлення наявності залежності за допомогою кореляційного аналізу неможливе.

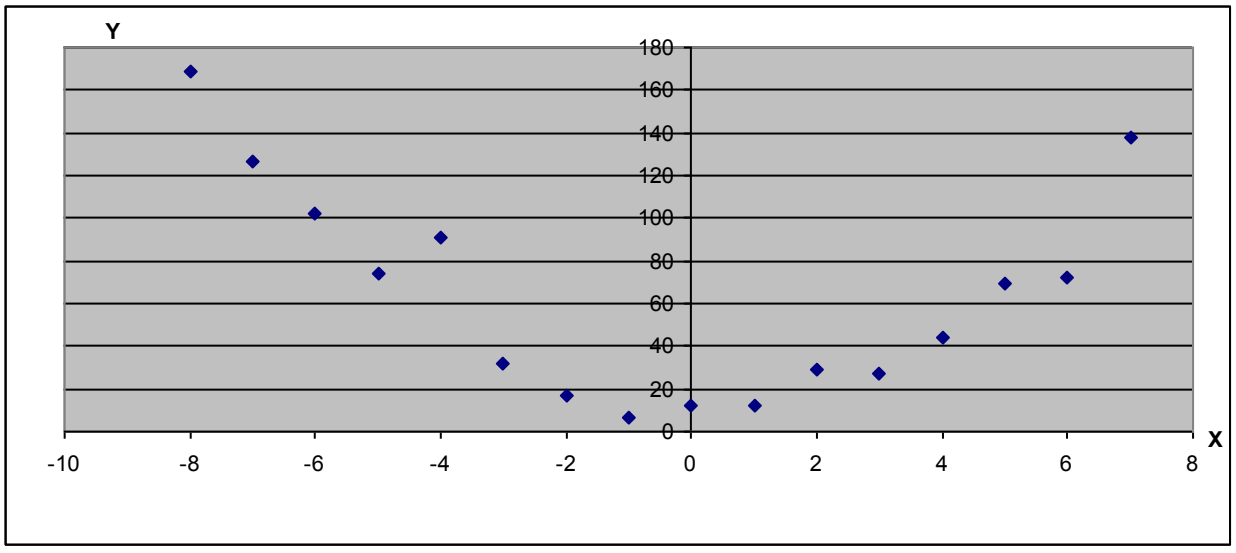

Рисунок 3 – Нелінійна залежність з одним викидом

Таблиця 3 – Розрахунки визначення викидів

Параболічна залежність з одним викидом			Енскомб (складний випадок)		
№ виключеного експерименту	Коефіцієнт кореляції	Різниця	№ виключеного експерименту	Коефіцієнт кореляції	Різниця
5	0,987381		7	-0,45576	
15	0,97461	0,012772	10	-0,2296	-0,22616
13	0,964515	0,010095	11	-0,22458	-0,00502
14	0,964296	0,000219	5	-0,18557	-0,03901
4	0,964159	0,000137	3	-0,18512	-0,00045
11	0,96388	0,000279	1	-0,17098	-0,01414
12	0,96346	0,00042	6	-0,16888	-0,0021
6	0,963187	0,000274	4	-0,1662	-0,00269
3	0,962822	0,000365	9	-0,12895	-0,03725
7	0,962199	0,000624	2	-0,04945	-0,0795
8	0,961629	0,00057	8	-0,03523	-0,01422
10	0,961334	0,000294			
9	0,961246	8,85E-05			
2	0,95974	0,001506			
16	0,958755	0,000985			
1	0,94786	0,010896			
Середнє		0,002635			-0,04205
Середньоквадратичне відхилення		0,004505			0,069229
Напівширина довірчого інтервалу		0,002208			0,040911
Верхня межа довірчого інтервалу		0,004842			0,082964

У табл. 4 і 5 порівнюються рівняння регресії при параболічній залежності, побудовані як з включенням у навчальну вибірку викиду, так і з виключенням його. Відкидання викидів не тільки підвищує точність і інформативність, але й призводить до суттєвих змін у значеннях коефіцієнтів регресії.

Таблиця 4 – Характеристики регресійного рівняння для параболічної залежності

Статистична характеристика	Регресія з викидами	Регресія без викидів
Множинний коефіцієнт кореляції R	0,968424	0,988389
R^2	0,937846	0,976912
Середньоквадратичне відхилення для відгуку	13,38975	8,406809
F_R	98,07886	253,8807

Таблиця 5 – Коефіцієнти регресійного рівняння для параболічної залежності

Коефіцієнт регресії	Регресія з викидами	Регресія без викидів
Вільний член (b_0)	12,98053	9,28355
Коефіцієнт при X (b_1)	-0,97454	-0,48933
Коефіцієнт при X^2 (b_2)	2,345385	2,409793

Тобто, у випадку нелінійної залежності необхідні попередні перетворення, без яких викиди не будуть розпізнані. При цьому слід зауважити, що без таких перетворень не буде успішною обробка даних.

3.3. Складні випадки. Виділяється лінійна складова

Розглянемо складні випадки, в яких однозначне формальне прийняття рішень неможливе.

3.3.1. Квартет Енскомба (Г) – складна модель

Для даних Енскомба представлена складна залежність. На рис. 4 показані дані і найкращий варіант опису їх за допомогою кусково-неперервної регресії з двох парабол. Якщо цей варіант вважати правильним, то можливим викидом буде експеримент (7; 12,5). Але метод, як і очікувалось, орієнтується на лінійну залежність і визначає як викид ту точку (6; 6,89), видалення якої максимально наблизить залежність до лінійної. Насправді ми маємо недостатньо даних, щоб прийняти правильне рішення: чи то залежність лінійна, а всі відхилення випадкові, чи залежність насправді складна, зі зміною елементів моделі. Авторові у практичній діяльності неодноразово зустрічались як один, так і другий випадок.

3.3.2. Можливість неоднорідності простору

Розглядається набір даних із залежністю між товщиною і висотою дерев (табл. 6 і рис. 5) [4].

Таблиця 6 – Залежність між товщиною і висотою дерев

№п/п	Діаметр (дюйми)	Висота (фути)	№п/п	Діаметр (дюйми)	Висота (фути)
1	5,5	58	14	10,1	75
2	5,7	60	15	10,2	72
3	5,8	42	16	10,4	78
4	6,5	64	17	10,6	65
5	6,6	60	18	10,6	80
6	6,7	65	19	10,8	82
7	6,9	56	20	11,30	70
8	7	57	21	11,3	74
9	7,3	70	22	11,6	68
10	8,3	68	23	11,6	68
11	8,6	65	24	13	82
12	9,5	70	25	18	88
13	10	63			

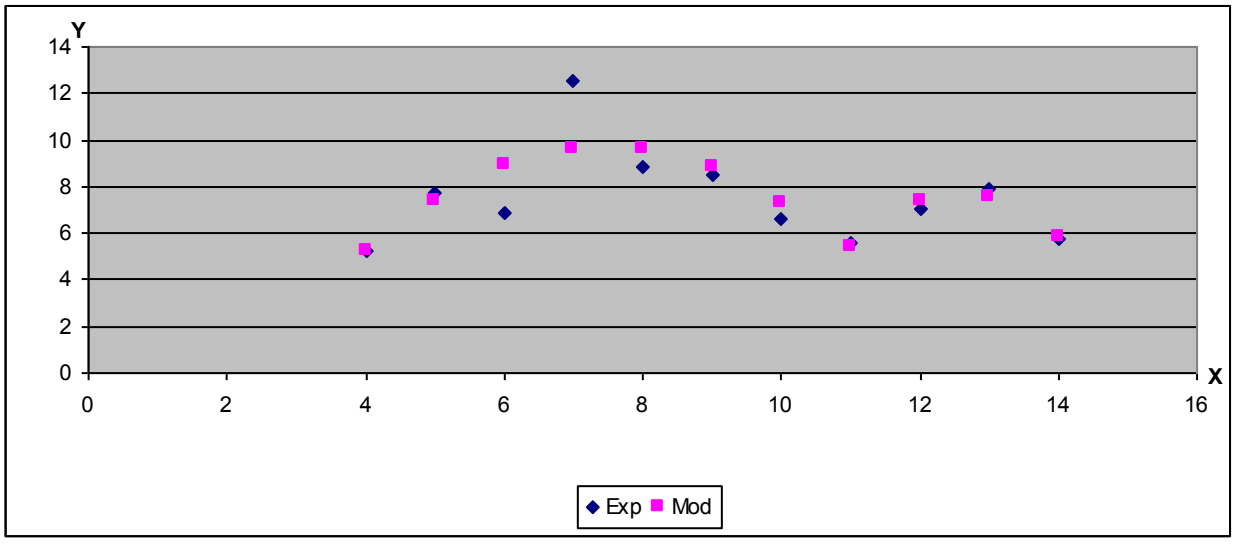

Рисунок 4 – Квартет Енскомба – складна залежність

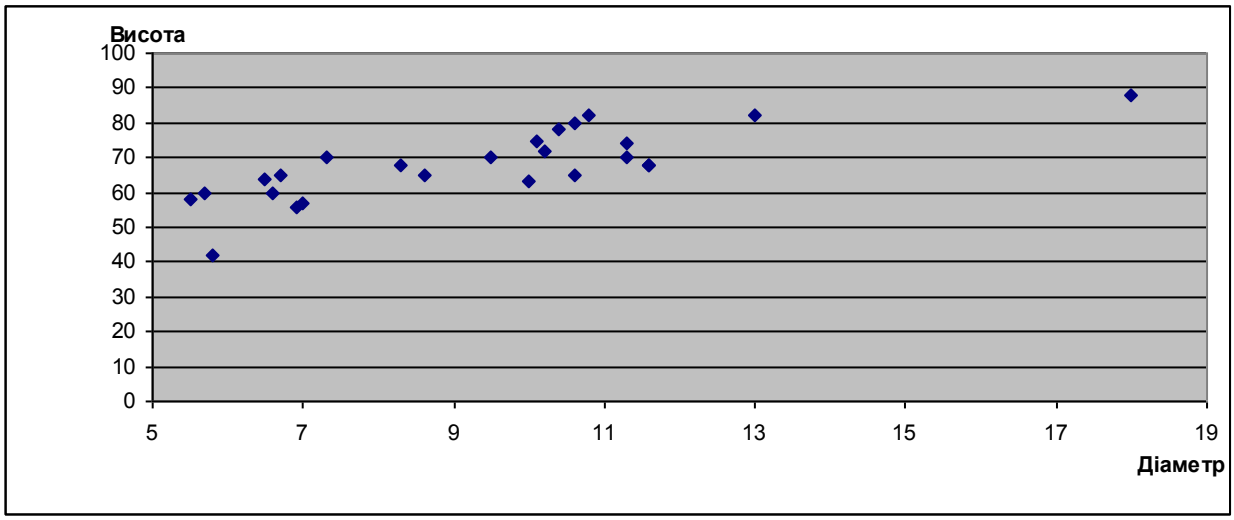

Рисунок 5 – Дані щодо відповідності між діаметром і висотою дерев

Візуально автори виділяють експерименти № 3, 25 як можливі викиди; розраховується статистика Кука, за якою вплив обох точок є суттєвим. Будується звичайна регресія і робастна. Висновок: побудовані моделі мало розрізняються, але ж суттєвим аналізом відкинуті як такі, що належать іншій генеральній сукупності. Скидається на вольове рішення, бо аналіз належності до інших генеральних сукупностей чисто умоглядний.

Застосування розробленого методу виділяє як викиди експерименти № 3, 24, 25 (рис. 6, враховуючи, що верхнє значення довірчого інтервалу рівне 0,004303).

Порівняємо коефіцієнти кореляції, отримані в різних умовах (табл. 7). Як видно, вони відрізняються мало.

Таблиця 7 – Коефіцієнти регресії

Коефіцієнт	Для всієї вибірки	Без експериментів 3, 25	Робастна	Без експериментів 3, 24, 25
b_0	41,956	44,353	42,872	45,635
b_1	2,784	2,617	2,704	2,456

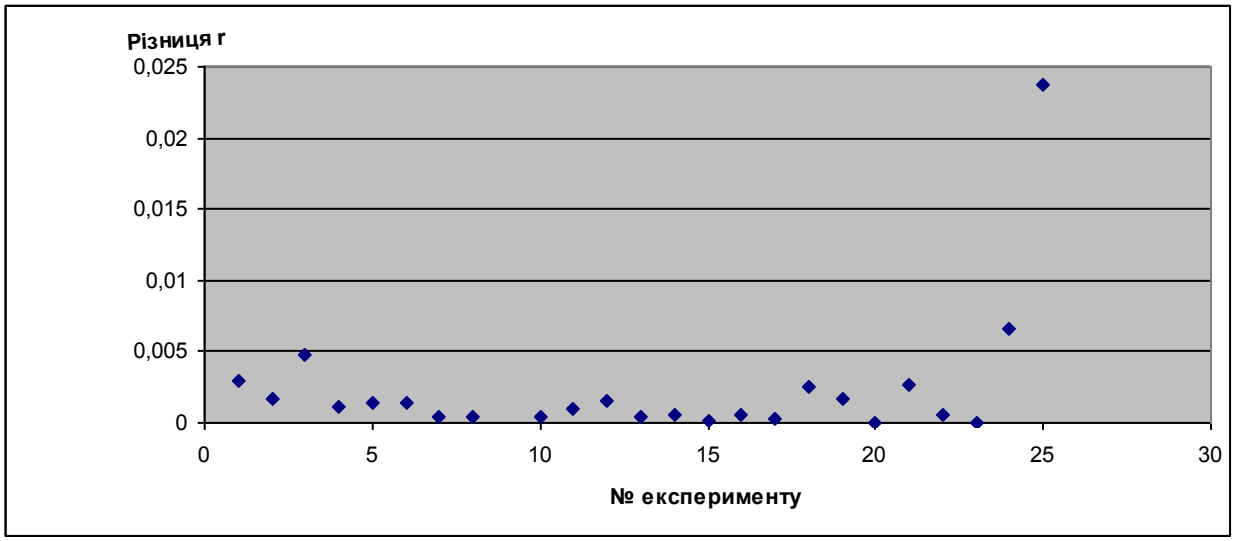

Рисунок 6 – Значення різниць для залежності між висотою і товщиною дерев

Якщо ж ми виконаємо прогнозування значення висоти для дерев товщиною 18 при тому, що при побудові моделі експеримент № 25 не буде входити в навчальну вибірку, то теж отримаємо близькі результати (табл. 8), але при цьому знайдений ближчий до експерименту результат буде за моделлю, отриманою при формуванні вибірки запропонованим методом.

Таблиця 8 – Експериментальне та прогнозоване значення

Діаметр	Висота				
	Експеримент	Для всієї вибірки	Без експериментів № №3, 25	Без експериментів № №3, 24, 25	Робастна
18	88	92,0617	91,4632	89,83878	91,5494

3.3.3. Штучний приклад із розмитою формою

Розглянемо приклад із невиразною формою, наближеною до лінійної залежності (рис. 7). Для цієї вибірки запропонований метод виділяє як викид експеримент (6; 2,24). Як видно з табл. 9 і 10, видалення даного спостереження приводить до покращення характеристик прямолінійного рівняння регресії, побудованого на вказаних даних.

Таблиця 9 – Статистичні характеристики рівняння регресії

Статистична характеристика	Регресія з викидами	Регресія без викидів
Множинний коефіцієнт кореляції R	0,817894	0,855875
R^2	0,668951	0,732522
Середньоквадратичне відхилення для відгуку	1,565279	1,052715
F_R	18,18629	19,17042

Таблиця 10 – Значення коефіцієнтів регресії

Коефіцієнт регресії	Регресія з викидами	Регресія без викидів
Вільний член b_0	1,318273	2,833998
Коефіцієнт при X b_1	0,636455	0,480543

У складних випадках метод вибирає точки, видалення яких приводить до покращення статистичних характеристик прямолінійної моделі.

3.4. Не працює, але на модель ці викиди слабко впливають

Розглянемо ситуації, в яких запропонований підхід не працює.

3.4.1. Нелінійна симетрична. Штучний приклад

Штучно розроблений приклад із двома симетричними викидами (рис. 8). Для цієї ситуації навіть після лінеаризації метод не визначає викидів (розрахунки опущені). В табл. 11, 12 приведені статистичні характеристики регресійних моделей, розраховані як з включенням викидів у навчальну матрицю, так і з виключенням їх із неї. Як видно, і статистичні характеристики, і коефіцієнти регресії практично не розрізняються. Це пояснюється тим фактом, що зміщення, викликане викидами, приблизно однакове і направлене у протилежні напрямки, що приводить до компенсації.

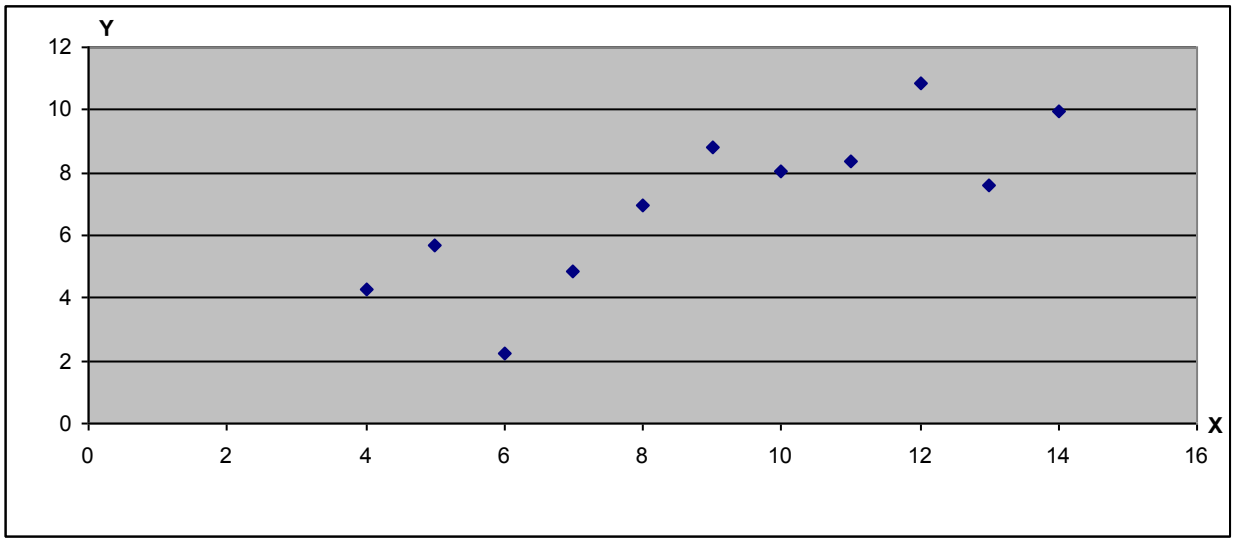

Рисунок 7 – Штучний приклад із розмитою формою

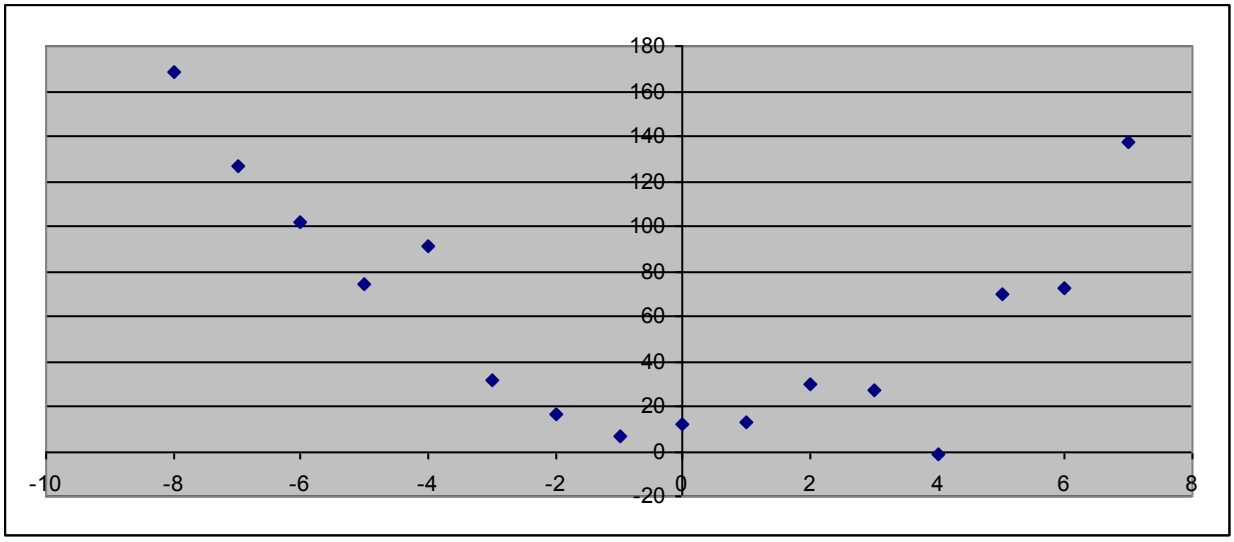

Рисунок 8 – Параболічна залежність із симетричними викидами

Таблиця 11 – Статистичні характеристики регресійних моделей

Статистична характеристика	Регресія з викидами	Регресія без викидів
Множинний коефіцієнт кореляції R	0,988263	0,988344
R^2	0,976663	0,976825
Середньоквадратичне відхилення для відгуку	272,0249	8,754036
F_R	66,39806	231,8204

У випадку парних викидів, коли вони компенсують один одного, метод не дозволяє їх визначити. Зауважимо, що викиди такого типу слабко впливають на характеристики і властивості регресійної моделі.

Таблиця 12 – Коефіцієнти регресії

Коефіцієнт регресії	Регресія з викидами	Регресія без викидів
Вільний член b_0	8,910682	9,463919
Коефіцієнт при X b1	-0,4404	-0,45833
Коефіцієнт при X^2 b_2	2,416289	2,409293

3.4.2. Лінійна залежність із симетричними викидами

На рис. 9 приведена лінійна залежність із двома симетричними викидами.
Характеристики регресійних моделей, побудовані на матриці з викидами і без, та відповідні регресійні коефіцієнти приведені в табл. 13 і 14 відповідно.

Таблиця 13 – Статистичні характеристики регресійних моделей

Статистична характеристика	Регресія з викидами	Регресія без викидів
Множинний коефіцієнт кореляції R	0,93546	0,9875
R^2	0,875086	0,975157
Середньоквадратичне відхилення для відгуку	19,17514	8,677729
F_R	98,0772	471,0249

Таблиця 14 – Коефіцієнти регресії

Коефіцієнт регресії	Регресія з викидами	Регресія без викидів
Вільний член b_0	8,598939	9,156087
Коефіцієнт при X b_1	2,44103	2,434858

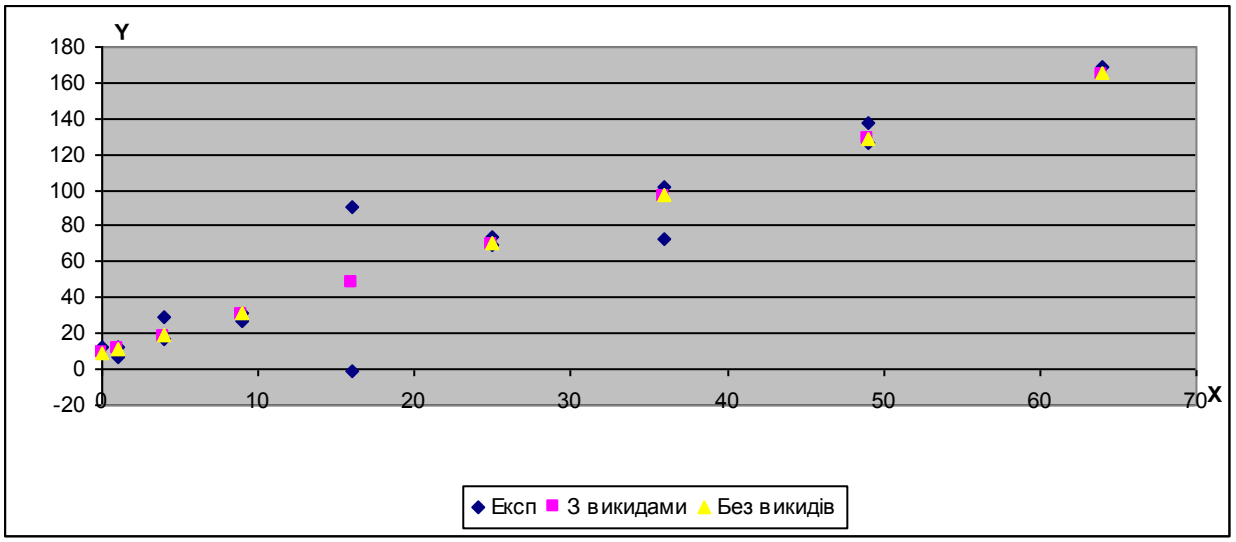

Рисунок 9 – Експериментальні і прогнозовані значення для лінійної залежності з симетричними викидами

4. Висновки

1. Запропоновано метод визначення викидів при кореляційному та одновимірному регресійному аналізі.
2. Запропонований метод дозволяє автоматично визначати викиди, незалежно від їх природи, при лінійному зв'язку між змінними.
3. При нелінійному зв'язку між змінними викиди визначаються у випадку лінеаризації змінних.
4. У сумнівних і невизначених випадках як викиди алгоритм визначає ті експерименти, видалення яких приводить до покращення лінійної моделі.
5. Метод не працює у випадку викидів, які компенсують одне одного, але в цих випадках викиди слабко впливають на отриману модель.
6. Може бути використаний у системах автоматизованої обробки інформації, оскільки гарантовано автоматично визначає наявність викидів або покращує лінійну регресійну модель.

СПИСОК ДЖЕРЕЛ

1. Лагутин М.Б. Наглядная математическая статистика. М.: Бином, 2007. 472 с.
2. Лапач С.Н., Чубенко А.В., Бабич П.Н. Статистика в науке и бизнесе. К.: Морион, 2002. 640 с.
3. Бизнес-статистика: учебник и практикум для академического бакалаврата / ред. И. Елисеевой. М.: Юрайт, 2018. 411 с.
4. Дрейпер Н., Смит Г. Прикладной регрессионный анализ. 3 изд. М.: Вильямс. 2007. 910 с.
5. Penny K.I. Appropriate critical values when testing for a single multivariate outlier by using the Mahalanobis distance. *Applied statistics*. 1996. Vol. 45, N 1. P. 73–81.
6. McDonald B. A Teaching Note on Cook's Distance – A Guideline. *Res. Lett. Inf. Math. Sci.* 2002. N 3. P. 127–128. URL: http://www.massey.ac.nz/~wwiims/research/letters/.
7. URL: http://st-atistics.narod.ru/presentation/linear_regression.pdf.
8. Лапач С.Н., Радченко С.Г. Регрессионный анализ в условиях неоднородности факторного пространства. *Математичні машини і системи*. 2016. № 3. С. 55–63.
9. Вероятность и математическая статистика: энциклопедия / ред. Прохорова Ю.В. М.: Большая российская энциклопедия, 1999. 910 с.
10. Закс Л. Статистическое оценивание. М.: Статистика, 1976. 600 с.
11. Лапач С.Н., Чубенко А.В., Бабич П.Н. Статистические методы в медико-биологических исследованиях с использованием Excel. Киев: Морион, 2001. 408 с.
12. URL: https://ru.wikipedia.org/wiki/Квартет_Энскомба.

Стаття надійшла до редакції 27.06.2019