

ІТЕРАЦІЙНИЙ АЛГОРИТМ АДАПТИВНОГО МЕТОДУ ПОБУДОВИ БАГАТОВИМІРНОЇ ЛІНІЙНОЇ РЕГРЕСІЇ З ВИКОРИСТАННЯМ КРИТЕРІЮ МІНІМІЗАЦІЇ СУМИ МОДУЛІВ

Анотація: Дана публікація продовжує серію наукових досліджень авторів в області побудови багатовимірних регресій з використанням критерію мінімуму суми модулів різниць, що використовуються в загальній процедурі методу найменших квадратів. Реалізація цього критерію зводиться до розв'язання відповідної задачі лінійного програмування. Оптимізаційна модель лінійного програмування, на відміну від методу найменших квадратів, дозволяє вводити довільні, лінійні відносно невідомих коефіцієнтів регресійної моделі, обмеження, що використовують результати статистичних випробувань і реалізують додаткові властивості шуканої багатовимірної регресії. Ця ідея дозволила створити новий ітераційний алгоритм побудови багатовимірної лінійної регресії по обмеженому об'єму випробувань (≥ 90) в діапазоні значень дисперсії випадкового фактору від 1 до 150. На кожній ітерації алгоритму розв'язується відповідна задача лінійного програмування, параметри якої змінюються з кожною наступною ітерацією алгоритму. Логіка ітераційного алгоритму є наслідком запропонованої евристики, яка в роботі обґрунтовується як на якісному рівні, так і результатами імітаційного статистичного моделювання. Ефективність ітераційного алгоритму побудови багатовимірної лінійної регресії по невеликому об'єму статистичних даних досліджувалась для нормального розподілу випадкового фактору регресійної моделі при відомих і не відомих значеннях математичного сподівання і дисперсії. Було статистично визначено обмеження на мінімальну кількість випробувань (≥ 90), яке дозволяє запропонувати критерій обґрунтованості знайдених оцінок невідомих коефіцієнтів багатовимірної лінійної регресії. Запропонований алгоритм модифіковано для випадку, коли багатовимірна лінійна регресія задана надлишковим описом. Наведені теоретично обґрунтовані практичні рекомендації по дослідженню фізично існуючої регресійної моделі, які дозволяють використовувати наведену в даній роботі математичну модель.

Ключові слова: регресійний аналіз, багатовимірна лінійна регресія, задача лінійного програмування, мінімізація суми модулів, ітераційний алгоритм, критерій χ^2 .

1. Вступ

Задача побудови багатовимірних лінійних регресій по невеликому об'єму експериментальних даних, очевидно, і досі є актуальною як в теоретичному, так і в прикладному аспектах [1–5].

Результати останніх досліджень в області побудови багатовимірних поліноміальних регресій [6] показали, що в багатьох випадках в умовах обмеженого активного експерименту можливо знаходити оцінки коефіцієнтів при нелінійних членах БПР з наперед заданою точністю. Проблематика точності оцінок коефіцієнтів виникає для лінійних членів БПР.

В [6] вона розв'язувалась за допомогою запропонованої модифікації евристичного методу групового урахування аргументів, який в практичному плані ефективний при не великих значеннях дисперсії випадкового фактора при обмеженому об'ємі експериментальних даних. Тому задача знаходження статистично обґрунтованих оцінок коефіцієнтів багатовимірної лінійної регресії для обмеженої кількості експериментів (≥ 90) і діапазону значень дисперсії 1–150 нормально розподіленого випадкового фактора є і наразі актуальною. Наведений ітераційний алгоритм є ефективною модифікацією відповідного алгоритму адаптивного методу побудови багатовимірних регресій [7] і використовує задачу лінійного програмування (ЗЛП) із змінними параметрами і критерій мінімуму суми модулів різниць, що використовується в МНК.

2. Постановка задачі багатовимірної лінійної регресії

БЛР задана у вигляді

$$Y(\bar{x}) = \theta_0 + \sum_{j=1}^m \theta_j x_j + E, \quad (1)$$

де $\forall x_j$ – всі детерміновані вхідні змінні, значення коефіцієнтів невідомі, випадкова величина E має нормальний розподіл з обмеженими параметрами $v = ME$, $\sigma^2 = DE$. Необхідно при невеликому об'ємі експериментальних даних $\bar{x}_i \rightarrow y_i, i = \overline{1, n}, n \geq 90$ обґрунтовано оцінити невідомі коефіцієнти $\theta_j, j = \overline{0, m}$.

Примітка 1. Обмеження на кількість випробувань обґрунтовується в п.4 даного дослідження. Для зручності покладено $\bar{x}_i = (1, x_{1i}, \dots, x_{mi})$.

Примітка 2. Вважається, що $ME = 0$, якщо це не так, запропонований ітераційний алгоритм оцінює замість θ_0 величину $\theta_0 + ME$. Тому вважаємо, що завжди $ME = 0$.

3. Базовий підалгоритм ітераційного алгоритму адаптивного методу для відомої дисперсії випадкового фактора $DE = \sigma^2$

Як показано в [8] критерій мінімуму суми модулів різниць, що використовується в МНК, зводиться до розв'язку ЗЛП. Це, в свою чергу, дозволяє в оптимізаційну модель знаходження оцінок $\theta_j, j = \overline{0, m}$ вводити додаткові лінійні обмеження, що є наслідком аналізу експериментальних даних ($\bar{x}_i \rightarrow y_i, i = \overline{1, n}$). В п. 3 приведено агрегований опис базового підалгоритму ітераційного алгоритму адаптивного методу побудови багатовимірної лінійної регресії.

3.1. Розв'язується ЗЛП

$$\min \sum_{i=1}^n z_i, \quad (2)$$

$$-z_i \leq y_i - \theta^T \bar{x}_i \leq z_i, z_i \geq 0, i = \overline{1, n}, \quad (3)$$

де $\theta^T = (\theta_0, \theta_1, \dots, \theta_m)$ змінними ЗЛП (2), (3) є $\theta_0, \theta_1, \dots, \theta_m, z_i, i = \overline{1, n}$.

Нехай $\hat{\theta} = (\hat{\theta}_0, \hat{\theta}_1, \dots, \hat{\theta}_m)$ – оптимальний розв'язок ЗЛП (2), (3).

3.2. Знаходяться оцінки реалізацій випадкової величини E

$$\hat{\varepsilon}_i = y_i - \hat{\theta}^T \bar{x}_i, i = \overline{1, n}. \quad (4)$$

Критерієм χ^2 перевіряється проста гіпотеза про те, що $\hat{\varepsilon}_i, i = \overline{1, n}$ належить нормальному розподілу з параметрами $0, \sigma^2$. Якщо реалізація критерію χ^2 належить допустимій області при $\alpha = 0,05$, кінець алгоритму, вектор $\hat{\theta}$ є обґрунтованим розв'язком задачі регресії. В протилежному випадку переходимо до наступного кроку базового підалгоритму.

3.3. Знаходимо число

$$z = \sum_{i=1}^n (y_i - \hat{\theta}^T \bar{x}_i). \quad (5)$$

Для заданих чисел $\Delta z > 0$, заданих натуральними числами p, l розв'язуємо послідовно наступні ЗЛП

$$\min \sum_{i=1}^n z_i, \quad (6)$$

$$-z_i \leq y_i - \theta^T \bar{x}_i \leq z_i, z_i \geq 0, i = \overline{1, n}, \quad (7)$$

$$-(|z| + p\Delta z - j\Delta z) \leq \sum_{i=1}^n (y_i - \theta^T \bar{x}_i) \leq |z| + p\Delta z - j\Delta z, l\Delta z < |z|, j = \overline{0, p+l}, \quad (8)$$

Змінними ЗЛП (6)–(8) є $\theta_j, j = \overline{0, m}, z_i, i = \overline{1, n}$.

Для кожного j для оптимального розв'язку $\hat{\theta}$ відповідної ЗЛП (6)–(8) реалізується пункт 3.2. Базовий підалгоритм завершує свою роботу в одному з двох випадків.

Перший – для мінімального j , при якому реалізація критерію χ^2 належить допустимій області, відповідний вектор $\hat{\theta}$ є обґрунтованим розв'язком задачі побудови багатовимірної поліноміальної регресії.

Другий – $\forall j = \overline{0, p+l}$ реалізації критерія χ^2 належать критичній області. Базовий підалгоритм реалізує $p+l+1$ ЗЛП (7), (8). В результаті статистично обґрунтованого розв'язку задачі регресії не було знайдено.

Примітка 3. Статистичне дослідження ефективності знаходження обґрунтованої оцінки $\hat{\theta}$ вектора θ наведено в п. 4.

Примітка 4. Величини $\Delta z, p, l$ задаються емпірично, їх можливі значення приведені в розділі 4.

Примітка 5. Евристика, яка була покладена в основу базового підалгоритму полягає в наступному припущенні – при послідовному зменшенні допустимої області, якій належить невідоме число

$$\sum_{i=1}^n (y_i - \theta^T \bar{x}_i), \quad (9)$$

де θ^T – вектор невідомих коефіцієнтів багатовимірної лінійної регресії, можна отримати вектор оцінок невідомих коефіцієнтів $\hat{\theta}$, для якого критерієм χ^2 приймається гіпотеза про нормальний розподіл з параметрами $ME = 0$ та відомою дисперсією σ^2 , для вибірки

$$\hat{\varepsilon}_i = y_i - \hat{\theta}^T \bar{x}_i, i = \overline{1, n}. \quad (10)$$

Евристика також припускає, що існує множина, елементами якої є фіксовані значення m, n, σ^2 , для кожного з яких суттєвим є як середній відсоток індивідуальних задач регресії, для кожної з яких базовий підалгоритм знаходить обґрунтовану оцінку вектора θ , так і середній відсоток індивідуальних задач регресії з цієї множини, для кожної з яких МНК не знаходить обґрунтованого рішення.

В наступному розділі аналіз результатів імітаційного моделювання підтвердив справедливість висунутої авторами евристики.

4. Властивості базового підалгоритму, що є наслідком аналізу результатів імітаційного моделювання

4.1. Вхідні параметри імітаційного моделювання

Для оцінювання ефективності базового підалгоритму були використані наступні параметри імітаційного моделювання:

- Кількість випробувань була обрана для кожного значення DE окремо (дана стратегія описана в п. 4.2)
- Кількість оцінюваних коефіцієнтів є найбільшим цілим від ділення кожного фіксованого значення кількості випробувань послідовно на числа 5, 7, 10, 12, 15 та 20. Область значень оцінюваних коефіцієнтів – величини коефіцієнтів генеруються випадковим чином з використанням рівномірного розподілу на інтервалі $[1; 10]$ за модулем, із незалежним випадковим вибором знаку для кожного значення ($P = 0.5$).
- Область значень детермінованих вхідних змінних – $[1; 10]$.
- Дисперсія випадкової величини E – від 5 до 100.
- Кількість повторень кожного експерименту – 100.

4.2. Знаходження кількості експериментів для кожного значення дисперсії

Для значень дисперсії у діапазоні від 5 до 100 було визначено а) мінімальну кількість реалізацій випадкової величини E , при якій критерій χ^2 приймає гіпотезу про нормальний розподіл з відповідними числовими параметрами, в середньому не менше ніж у 91% випадків, при кількості повторень експерименту 200, б) мінімальну кількість реалізацій випадкової величини E , при якій критерій χ^2 приймає гіпотезу про нормальний розподіл з відповідними числовими параметрами, в середньому не менше ніж у 97% випадків, при кількості повторень експерименту 200. Виявилось, що в першому випадку кількість реалізацій випадкової величини E дорівнює 90, в другому – 120.

4.3. Формування базового значення параметра Z та вибір діапазону його варіацій

Параметром імітаційного моделювання Z зветься величина, що задається виразом

$$Z = \sum_{i=1}^n (y_i - \hat{\theta}^T \bar{x}_i), \quad (11)$$

де вектор оцінюваних коефіцієнтів $\hat{\theta}$ задається оптимальним розв'язком ЗЛП (2), (3).

У процесі досліджень було виявлено, що значення параметра Z , розраховане на коефіцієнтах вектора θ регресійної моделі, не завжди по модулю є меншим, ніж значення модуля Z (11).

З метою підвищення ефективності базового підалгоритму було прийнято рішення формувати масив можливих значень параметра Z не лише за принципом поетапного зменшення модуля базового значення, але й за рахунок його попереднього збільшення на певний відсоток. Для цього було здійснено двоетапну генерацію послідовності: спочатку створюється серія значень із поступовим збільшенням модуля базового Z на заданий відсоток (у даному випадку 2% на кожному кроці протягом 20 кроків), після чого формується низка значень зі зменшенням модуля у тому ж відсотковому співвідношенні.

Примітка 6. В якості початкового базового параметра Z можна вибрати $Z = \frac{1}{n} \sum_{i=1}^n \varepsilon_i$, де ε_i це штучно згенеровані реалізації випадкової величини E .

4.4. Оцінка ефективності параметрів моделі

для досягнення точного та обґрунтованого розв'язку

У рамках проведеного моделювання визначався відсоток випадків, коли для кожного проміжного фіксованого значення параметра Z було знайдено розв'язок відповідної ЗЛП (6)–(8), що є обґрунтованою оцінкою $\hat{\theta}$ вектора θ .

Оцінка частки повторів, у яких знаходилася обґрунтована оцінка $\hat{\theta}$ вектора θ , засвідчує стабільно високий рівень ефективності базового підалгоритму. Зокрема, при середніх та високих дисперсіях випадкової величини E (у межах від 5 до 100) та для відношень кількості випробувань до кількості невідомих коефіцієнтів від 5 до 20, цей показник становив у середньому 76.79%, тобто середній відсоток отримання обґрунтованого розв'язку задачі регресії (отримання вектору $\hat{\theta}$, для якого реалізація критерія χ^2 потрапляє в допустиму область).

Примітка 7. Порівняння різних обґрунтованих розв'язків індивідуальної задачі регресії, отриманих при різних значеннях параметра моделювання Z , досліджується в розділі 4.5.

Водночас було підтверджено, що зі зростанням дисперсії випадкової величини E середній відсоток отримання обґрунтованих розв'язків має тенденцію до зниження. Наприклад, при фіксованому співвідношенні кількості випробувань до кількості коефіцієнтів, що дорівнює 5, середній відсоток отримання обґрунтованого розв'язку для малих дисперсій (5–15) становив 37.81%, тоді як для великих дисперсій (60–100) – 23.64%. Подібна тенденція спостерігається і для співвідношення 7: при менших дисперсіях відсоток становив 83.28%, тоді як при більших – 65.67%.

Виявлено також важливу закономірність щодо впливу кількості оцінюваних коефіцієнтів на результативність методу. Зокрема, при фіксованому співвідношенні між кількістю випробувань і кількістю коефіцієнтів (наприклад, співвідношення 10 з кількістю експериментів 120 та кількості оцінюваних коефіцієнтів 12), збільшення абсолютної кількості оцінюваних коефіцієнтів лінійної регресії супроводжується зниженням частки отримання обґрунтованого розв'язку. Для дисперсій 60–100 це зниження (кількість оцінюваних коефіцієнтів дорівнює 16, кількість випробувань 160) становить від 88.82% до 58.17%. Така динаміка свідчить про доцільність, по можливості, зменшення кількості

коефіцієнтів, що оцінюються, наприклад, шляхом виключення вхідних змінних, що не впливають або слабо впливають на значення багатовимірної лінійної регресії.

4.5. Стратегія вибору обґрунтованого розв'язку

Під час експериментальних досліджень було встановлено, що у багатьох випадках існувало декілька значень параметра моделювання Z , для яких було знайдено обґрунтований розв'язок. У зв'язку з цим виникла необхідність обґрунтувати стратегію вибору одного вектору оцінок невідомих коефіцієнтів з множини допустимих розв'язків. Для цього було здійснено оцінку значень міри порівняння, що характеризують якість кожного із знайдених рішень. В якості міри порівнянь (близькості) двох векторів вибрано наступну:

$$\|\theta - \hat{\theta}\| = \left\| \frac{\theta}{\|\theta\|} - \frac{\hat{\theta}}{\|\hat{\theta}\|} \right\|, \quad (12)$$

де $\|\theta\| = +\sqrt{\sum_{j=0}^m \theta_j^2}$.

У табл. 1 наведено узагальнені показники подібності обґрунтованих розв'язків для кожного повтору експериментів. Зокрема, розраховано середнє мінімальне значення міри порівняння з ідеальним вектором коефіцієнтів, середнє відхилення векторів інших обґрунтованих розв'язків від мінімальної міри порівняння.

Середнє відхилення кожного із знайдених обґрунтованих розв'язків від мінімального розв'язку за мірою порівняння з вектором θ є вкрай малим – у більшості випадків не перевищує 0.03. Це свідчить про те, що всі обґрунтовані рішення в межах однієї задачі є майже рівнозначними за точністю.

У відсотковому відношенні для кількості випробувань в межах 90–120 середнє відхилення модулів значень компонентів вектору $\hat{\theta}$, що є обґрунтованою оцінкою вектору θ , від його відповідних компонентів для дисперсій випадкової величини E у межах від 5 до 100 становить 13.49%, що свідчить про високу точність оцінювання параметрів моделі. Крім того, середнє значення міри порівняння вектору обґрунтованих оцінок $\hat{\theta}$ з вектором θ становить 0.19.

Крім оцінки відхилень модулів компонентів вектору обґрунтованих оцінок $\hat{\theta}$ з вектором θ , було також проаналізовано структурну подібність між векторами з точки зору збереження знаків компонентів. Зокрема, досліджувався відсоток компонентів, знак яких у векторі обґрунтованого розв'язку збігався зі знаком відповідного елемента в векторі θ . За результатами експериментів, для всіх значень дисперсій випадкової величини E у межах 5–100 та відповідних комбінацій параметрів задачі регресії, відсоток збігу знаків компонентів векторів $\hat{\theta}$, θ варіювався від 95.4% до 100%, а середнє значення становило 97.82%, що засвідчує високий рівень відповідності знаків між оціненими та ідеальними коефіцієнтами.

Таблиця 1.

Результати статистичних досліджень

Кількість випробувань	Кількість оцінюваних коефіцієнтів	Відношення між випробуваннями та коефіцієнтами	Дисперсія	Середнє мінімальне значення міри порівняння з ідеальним вектором	Середнє відхилення мір порівняння розв'язків від найкращого
120	24	5	20	0.115281	0.025544
90	18	5	40	0.242458	0.014331
120	24	5	60	0.272887	0.028315
90	18	5	80	0.210405	0.025075
120	24	5	100	0.239977	0.002951
120	17	7	20	0.124277	0.016185
90	13	7	40	0.223547	0.018221
120	17	7	60	0.249472	0.018103
90	13	7	80	0.217466	0.031866
120	17	7	100	0.30041	0.022197
120	12	10	20	0.088593	0.014338
90	9	10	40	0.214607	0.021857
120	12	10	60	0.204662	0.019651
90	9	10	80	0.297437	0.035061
120	12	10	100	0.219004	0.027696
120	10	12	20	0.114584	0.012738
90	8	12	40	0.189889	0.027822
120	10	12	60	0.203715	0.021635
90	8	12	80	0.278997	0.048238
120	10	12	100	0.284418	0.032425
120	8	15	20	0.112872	0.015067
90	6	15	40	0.170356	0.026917
120	8	15	60	0.168672	0.01787
90	6	15	80	0.290362	0.032212
120	8	15	100	0.262156	0.03456
120	6	20	20	0.096884	0.01472
90	5	20	40	0.178471	0.029722
120	6	20	60	0.182158	0.028542
90	5	20	80	0.271672	0.043496
120	6	20	100	0.215626	0.032925

Таким чином, можна зробити висновок, що в якості остаточного розв'язку задачі регресії можна брати будь-яку обґрунтовану оцінку $\hat{\theta}$ вектора θ і, як наслідок, завершувати ітераційну процедуру базового підалгоритму після отримання першого обґрунтованого розв'язку.

4.6. Обґрунтування необхідності ітераційної процедури базового підалгоритму

Запропонована методика оцінювання коефіцієнтів із поетапним варіюванням параметра Z демонструє свою доцільність та ефективність на основі наступних отриманих статистичних показників, які були зібрані як середні для усіх дисперсій випадкової величини E та відношення між кількістю випробувань та кількістю оцінюваних коефіцієнтів 5, 7, 10. По-перше, згідно з проведеними експериментами, у 47.72% випадків гіпотеза про відповідність моделі (за критерієм χ^2) підтверджувалася не на першій ітерації, а на наступних. Це свідчить про те, що стратегія багатокрокового пошуку рішень є виправданою: вона дозволяє знаходити коректні оцінки параметрів у ситуаціях, коли початкове обмеження не дає можливості отримати статистично значущий результат. Таким чином, процедура поступової зміни діапазону допустимих значень Z підвищує ймовірність виявлення обґрунтованої оцінки $\hat{\theta}$ вектора θ .

По-друге, важливою характеристикою запропонованого базового підалгоритму є його можливість отримання обґрунтованої оцінки $\hat{\theta}$ вектору θ у випадках, коли МНК його не реалізує. Так, для дисперсій випадкової величини E в межах 1–15 середній відсоток дорівнює 3.57%, а для діапазону дисперсій 20–100 він дорівнює 6.18%. Це свідчить про підвищення середньої ефективності базового підалгоритму в умовах великих значень дисперсії.

Аналіз результатів імітаційного моделювання показав, що існує множина параметрів задачі регресії, при яких середній відсоток суттєво відрізняється від наведених вище відсотків (див. останній стовпець табл. 2).

4.7. Ілюстраційний приклад

Нижче наведено ілюстраційний приклад із зазначенням використаних параметрів моделі та отриманих результатів:

- Кількість випробувань становила 120,
- Дисперсія випадкової величини E – 40,
- Кількість оцінюваних коефіцієнтів – 17,
- Ідеальні значення оцінюваних коефіцієнтів: 3.387, 8.654, -2.157, 3.429, -2.027, -2.897, 6.888, -1.859, -8.871, 4.601, 2.827, 9.982, -4.500, -2.483, -5.260, -9.074, 9.484,
- Кількість ітерацій для отримання першого обґрунтованого розв'язку: 46,
- Отриманий вектор оцінок: -1.333, 8.764, -2.245, 3.056, -2.149, -2.694, 7.346, -2.033, -8.519, 4.963, 2.753, 9.666, -4.081, -2.347, -5.442, -9.207, 9.889,
- Міра порівнянь отриманого вектору обґрунтованого розв'язку $\hat{\theta}$ з вектором θ : 0.1982,

- Міра порівнянь вектору оцінок МНК з вектором θ : 0.3966. Відповідна реалізація критерія χ^2 для цього вектору оцінок належить критичній області, тобто отриманий вектор оцінок МНК є необґрунтованим.

Таблиця 2.

Результати статистичних досліджень

Кількість випробувань	Кількість оцінюваних коефіцієнтів	Відношення між випробуваннями та коефіцієнтами	Дисперсія	Частка повторів з ≥ 1 успішною реалізацією χ^2	Частка повторів з успішною реалізацією χ^2 , у яких МНК не забезпечив успішного проходження критерію
90	18	5	15	50	12.1752
120	24	5	20	16	37.5576
120	10	12	20	94	12.7659
120	6	20	30	100	10.7216
120	17	7	40	60	13.3333
120	24	5	80	4	50.3891
120	17	7	80	58	17.2413
90	18	5	90	46	17.3913

5. Ітераційний алгоритм адаптивного методу побудови багатовимірної лінійної регресії за результатами обмеженого об'єму випробувань і відомої дисперсії

Розглядається для випадку обмеженої кількості випробувань, а саме $n \leq 120$.

Примітка 8. Достовірність числових характеристик ефективності ґрунтуються на результатах імітаційного моделювання, виходячи з того, що вибрані параметри, а саме $|\theta_i| \in [1; 10]$, область значень вхідних змінних $[1; 10]$, використання рівномірного розподілу для визначення знаку кожного коефіцієнта $\theta_j, j = \overline{0, m}$, приводить до того, що в середньому відношення

$$\frac{|\varepsilon_i|}{|\theta^T \bar{x}_i|}, \quad (13)$$

достатньо велике, щоб вважати, що наведені середні характеристики ефективності базового підалгоритму (п. 4) є нижніми границями при розв'язанні практичних задач. Далі наводяться агреговані кроки ітераційного базового підалгоритму.

5.1. Розв'язання задачі методом найменших квадратів

Задача регресії розв'язується МНК, а для отриманого вектору оцінок $\hat{\theta}$ реалізується п. 3.2. Якщо реалізація χ^2 належить допустимій області, отриманий вектор оцінок є обґрунтованим, в протилежному випадку ступінь достовірності розв'язку задачі оцінки коефіцієнтів БЛР є невідомою.

5.2. Другий етап ітераційного алгоритму

Реалізується, якщо на першому кроці (п. 5.1) отримані необґрунтовані оцінки коефіцієнтів $\theta_j, j = \overline{0, m}$ і виконуються наступні обмеження на параметри задачі регресії

$$n \geq 90, \frac{n}{m+1} \geq 6,5, \sigma^2 \in [5, 150], \quad (14)$$

полягає в реалізації базового підалгоритму ітераційного алгоритму (п. 3). Якщо на другому кроці обґрунтовані оцінки $\theta_j, j = \overline{0, m}$ не знайдені, розв'язком задачі регресії є оцінки, отримані на першому кроці ітераційного алгоритму (п. 5.1).

6. Узагальнення, допоміжні результати

6.1. Ітераційний алгоритм при невідомій дисперсії σ^2 випадкової величини E

Реалізується якщо значення невідомої дисперсії випадкової величини E задовольняє умові

$$\sigma^2 \in [a, b] \subset [1, 150]. \quad (15)$$

Полягає в тому, що ітераційний алгоритм, викладений в п. 3 послідовно реалізується для $\sigma_1^2 = a < \sigma_2^2 < \dots < \sigma_k^2 = b$, $\sigma_{i+1}^2 - \sigma_i^2 \leq \Delta\sigma_i^2$, $\Delta\sigma_i^2 \leq \Delta\sigma_{i+1}^2$, де $\Delta\sigma_{i+1}^2$ достатньо малі числа і задаються емпірично.

Ітераційний алгоритм завершує роботу в одному з наступних двох випадків.

Перший – знайдено таке мінімальне $l \leq 150$, при якому для σ_l^2 ітераційним алгоритмом отримані обґрунтовані оцінки $\hat{\theta}_j, j = \overline{0, m}$.

Другий – таке l не знайдене. Розв'язком задачі регресії беруться оцінки $\hat{\theta}_j, j = \overline{0, m}$, отримані загальною процедурою МНК.

Примітка 9. Ітераційний алгоритм можна було б спростити, реалізувавши його один раз, перевіряючи критерієм χ^2 складну гіпотезу про нормальний розподіл. Але в цьому випадку прийнявши гіпотезу ми отримуємо в середньому гірші оцінки $\hat{\theta}_j, j = \overline{0, m}$, ніж у випадку, коли ітераційний алгоритм перевіряє і приймає просту гіпотезу про нормальний розподіл по вибірці оцінок реалізацій випадкової величини E .

6.2. Допоміжні результати

6.2.1. Умови застосування ітераційного алгоритму

Використання ітераційного алгоритму є можливим, якщо відомо, що випадкова величина E має нормальний розподіл, відоме значення її дисперсії σ^2 , чи числа $a < b$, для яких виконується $\sigma^2 \in [a; b]$ з достатньо високою ймовірністю, причому бажано, щоб $b - a$ було максимально можливо малим числом. Як показано в [7], ці умови реалізується, якщо існує таке значення $\bar{x}_i$ – вектору детермінованих вхідних змінних $\bar{x}$, яке в реальному експерименті може послідовно подаватися на вхід фізично існуючої регре-

сійної моделі достатню кількість (n_1) разів, що дозволяє критерію χ^2 обґрунтовано перевірити складну гіпотезу про нормальний розподіл. В цьому випадку отримаємо вибірку реалізацій випадкової величини $E + \hat{\theta}^T \bar{x}_i$, по якій можливо: а) перевірити критерієм χ^2 складну гіпотезу про нормальний розподіл випадкової величини $E + \hat{\theta}^T \bar{x}_i$, а значить і самої випадкової величини E , б) враховуючи, що $D(E + C) = DE$, де C – константа, побудувати з заданою ймовірністю α довірчий інтервал для $\sigma^2 = DE$:

$$\frac{nS_n^2}{x_2^2} \leq \delta^2 \leq \frac{nS_n^2}{x_1^2}, \quad (16)$$

де

$$S_n^2 = \frac{1}{n} \sum_{j=1}^n (E_j + \theta^T \bar{x}_i - \overline{(E + \theta^T \bar{x}_i)})^2, \quad (17)$$

де

$$\overline{E + \theta^T \bar{x}_i} = \frac{1}{n} \sum_{j=1}^n (E_j + \theta^T \bar{x}_i), \quad (18)$$

де $E_j + \theta^T \bar{x}_i$ – реалізації випадкової величини $E + \theta^T \bar{x}_i$ в j -тому повторі реального експерименту над регресійною моделлю $j = \overline{1, n_1}$. Числа χ_1^2, χ_2^2 задовольняють умовам

$$P(\chi^2 \leq \chi_1^2) = \frac{\alpha}{2}, P(\chi^2 \leq \chi_2^2) = \frac{\alpha}{2}, \quad (19)$$

де випадкова величина χ^2 має розподіл χ^2 з $n_1 - 1$ ступенями свободи.

Примітка 10. n_1 вибирається з умови мінімізації різниці $\chi_2^2 - \chi_1^2$.

6.2.2. Ітераційний алгоритм для моделей із надлишковим описом

Якщо виконуються умови п. 6.2.1, а саме існує можливість реалізації активного експерименту для достатньо великої кількості випробувань при деякому фіксованому значенні $\bar{x}_i$, запропонований ітераційний алгоритм можна поширити на випадок, коли БЛР задана надлишковим описом, тобто припускається, що деякі вхідні змінні можуть не впливати, чи практично не впливати на вихідну змінну лінійної багатовимірної регресії. В цьому випадку пропонується по даним активного чи пасивного експерименту $\bar{x}_i \rightarrow y_i, i = \overline{1, n}, n \geq 90$ спочатку реалізувати алгоритм виключення вхідних змінних з нульовими чи практично нульовими по модулю коефіцієнтами викладений в [7], п. 3. В результаті отримуємо нову багатовимірну лінійну регресію

$$Y(\bar{x}) = \theta_0 + \sum_{j=1}^{m_1} \theta_{l_j} x_{l_j} + E, m_1 < m. \quad (20)$$

Примітка 11. $\forall x_{l_j} \in \{x_1, \dots, x_m\}$, а також припускаємо, що в моделі (20) залишився коефіцієнт θ_0 .

Коректність моделі (20) перевіряється наступним чином: на фізично існуючій регресійній моделі проводимо активний експеримент ($\bar{x}_{1i} \rightarrow y_j, j = \overline{1, n}$), де $\bar{x}_{1i} = (x_{l_1 i}, \dots, x_{l_{m_1} i})^T$. Тобто на вхід фізично існуючої регресійної моделі послідовно n разів подається вектор вхідних змінних $\bar{x}_{1i} = (x_{l_1 i}, \dots, x_{l_{m_1} i})^T$.

Примітка 12. Якщо це можливо, значення інших змінних $\forall x_j \in \{x_{l_1}, \dots, x_{l_m}\}$ на вхід фізично існуючої регресійної моделі не подаються. Якщо це неможливо, то в випробуваннях їх значення можуть бути довільні, достатньо великі по модулю.

Критерієм χ^2 перевіряється складна гіпотеза про те, що вибірка $y_j, j = \overline{1, n}$ належить нормальній генеральній сукупності. Якщо гіпотеза приймається, коефіцієнти багатовимірної лінійної регресії (20) оцінюються запропонованим ітераційним алгоритмом.

Примітка 13. Так як $m_1 < m$, то у випадку $\frac{n}{m_1+1} \geq 7$ ефективність ітераційного алгоритму суттєво підвищується.

Висновки

1. Запропоновано задачу оцінки значень невідомих коефіцієнтів лінійної багатовимірної регресії знаходити по критерію мінімуму суми модулів різниць, що використовується в загальній процедурі методу найменших квадратів.

2. Використання моделі лінійного програмування у випадку обмеженого об'єму експериментальних даних (≥ 90) на основі оригінальної евристики дозволило створити ітераційний алгоритм знаходження обґрунтованих оцінок коефіцієнтів багатовимірної лінійної регресії, на кожній ітерації якого розв'язується відповідна задача лінійного програмування.

3. Аналіз результатів імітаційного моделювання дозволив виявити область значень параметрів задачі регресії для якої використання запропонованого ітераційного алгоритму статистично значимо приводить до знаходження обґрунтованих оцінок невідомих коефіцієнтів багатовимірної лінійної регресії.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Singh, P., Adebajo, A., Shafiq, N. et al. Development of performance-based models for green concrete using multiple linear regression and artificial neural network. *International Journal on Interactive Design and Manufacturing (IJIDeM)*. 2024. Vol. 18. No. 5. P. 2945–2956. DOI: <https://doi.org/10.1007/s12008-023-01386-6>.

2. Gao C. Robust regression via multivariate regression depth. *Bernoulli*. 2020. Vol. 26. No. 2. P. 1139–1170. DOI: <https://doi.org/10.3150/19-BEJ1144>.

3. Haghighat P., Gandara D., Kang L., Anahideh H. Fair multivariate adaptive regression splines for ensuring equity and transparency. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2024. Vol. 38. No. 20. P. 22076–22086. DOI: <https://doi.org/10.1609/aaai.v38i20.30211>.

4. Thien Nguyen Luu Minh, Van Melkebeke M., Carrieri E., Hogie J., Poelman H., Michiels Y., Van Geem K.M., De Meester S. Multiple linear regression in adsorption capacity prediction: Application in plastic waste pyrolysis oil purification. *Separation and Purification Technology*. 2025. Vol. 378, Part 2. Article 134651. DOI: <https://doi.org/10.1016/j.seppur.2025.134651>.

5. Omer A., Sedeeq B., Ali T. A proposed hybrid method for multivariate linear regression model and multivariate wavelets (Simulation study). *Polytechnic Journal of Humanities and Social Sciences*. 2024. Vol. 5. No. 1. P. 112–124. URL: <https://www.researchgate.net/publication/377558021>, Access date: 16.04.2025.

6. Pavlov, A., Holovchenko, M., & Drozd, V. Modification of the decomposition method of constructing multivariate polynomial regression which is linear with respect to unknown coefficients. *Bulletin of National Technical University "KhPI". Series: System Analysis, Control and Information Technologies*. 2024. No. 2(12). P. 3–10. DOI: <https://doi.org/10.20998/2079-0023.2024.02.01>.

7. Pavlov, A., Holovchenko, M., & Drozd, V. An adaptive method for building a multivariate regression. *Bulletin of National Technical University "KhPI". Series: System Analysis, Control and Information Technologies*. 2024. No. 1(11). P. 3–8. DOI: <https://doi.org/10.20998/2079-0023.2024.01.01>.

8. Pavlov, A., Kushch, A. Formulation of a linear regression problem with linearly dependent coefficients. *Interdepartmental scientific technical journal «Adaptive systems of automatic control»*. 2025. No. 2(47). P. 1–8. DOI: <https://doi.org/10.20998/2079-0023.2024.01.01>.