УДК 519.816:519.24:519.854.2

О.А. Павлов, А.В. Кущ

ЗАДАЧА ЛІНІЙНОЇ РЕГРЕСІЇ З ЛІНІЙНО ЗАЛЕЖНИМИ КОЕФІЦІЄНТАМИ

Анотація: Розглядається задача знаходження оцінок коефіцієнтів багатовимірної лінійної регресії за результатами активного чи пасивного експерименту для випадку, коли невідомі значення коефіцієнтів повинні задовольняти певним лінійним обмеженням. В якості критерію оптимальності для знаходження оцінок невідомих коефіцієнтів пропонується замість мінімуму суми квадратів обрати мінімум суми модулів відповідних різниць, що дозволяє замість нелінійної комбінаторної задачі квадратичного програмування, знаходити розв’язок задачі лінійного програмування, яка гарантовано знаходиться точно. Наводяться приклади постановки задач, коли на невідомі коефіцієнти накладаються лінійні обмеження. На завершення приводяться результати статистичних досліджень точності знаходження оцінок коефіцієнтів лінійної регресії методом найменших квадратів та методом мінімуму суми модулів відповідних різниць для випадку незалежних невідомих коефіцієнтів. На думку авторів, наведені результати можуть бути корисними, так як хоча оцінки, отримані методом мінімуму суми модулів є лінійними, але, як показано, на них не поширюється теорема Маркова про те, що оцінки отримані методом найменших квадратів є ефективними в класі незміщених лінійних оцінок.

Ключові слова: метод найменших квадратів, лінійне програмування, задача регресії, квадратичне програмування, побудова регресії, статистичне імітаційне моделювання.

1. Вступ

Побудова багатовимірної регресії, як засвідчують сучасні дослідження [1–9], є актуальною задачею як у теоретичному, так і прикладному вимірах. Найбільш поширена постановка задачі оцінок невідомих коефіцієнтів багатовимірної лінійної регресії є наступною:

$$Y(\bar{x}) = \theta_0 + \theta_1 x_1 + \cdots + \theta_r x_r + E, \tag{1}$$

де $x_i, i = 1,\bar{r}$ – детерміновані вхідні змінні, $\theta_j, j = \overline{0,r}$, – невідомі коефіцієнти, E – випадкова величина, $ME = 0$, $DE = \sigma^2 < \infty$. За результатами активного чи пасивного експерименту $\bar{x_i} \to y_i, i = 1,\bar{n}$, де $y_i = \theta_0 + \theta_1 x_{1i} + \cdots + \theta_r x_{ri} + \delta_i$ (δ_i – це реалізації випадкової величини E), треба знайти оцінки $\widehat{\theta_j}, j = \overline{0,r}$, невідомих значень коефіцієнтів $\theta_j, j = 0,\bar{r}$. Найчастіше задача розв’язується методом найменших квадратів, а саме:

$$\widehat{\theta_j} = (A^T A)^{-1} A^T y, \text{ де}$$

$$A = \begin{pmatrix} 1 & x_{11} & \cdots & x_{r1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & \cdots & x_{rn} \end{pmatrix}, \hat{\theta} = \begin{pmatrix} \widehat{\theta_1} \\ \vdots \\ \widehat{\theta_r} \end{pmatrix}, y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}. \quad (2)$$

Практичні та теоретичні проблеми виникають тоді, коли кількість випробувань $n > r + 1$ недостатньо велика та модель (1) є надлишковою, тобто може містити вхідні змінні, що несуттєво чи взагалі не впливають на вихідну змінну. Ефективні методи розв’язання задач в такій постановці викладені в [8, 9]. Задача суттєво ускладнюється, якщо практична постановка задачі регресії вимагає виконання додаткових умов, а саме:

$$B\theta \le b, \quad (3)$$

де B – матриця, $\theta = (\theta_0, \theta_1, \dots, \theta_r)^T$, b – вектор. Саме знаходження оцінок $\widehat{\theta_j}, j = \overline{0,r}$, в такій постановці задачі регресії і присвячене дане наукове дослідження.

2. Метод мінімізації суми модулів

Якщо для цього випадку використовувати очевидну модифікацію метода найменших квадратів, отримуємо наступну задачу квадратичного програмування

$$\min_{\theta_j, j=\overline{0,r}} \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_{1i} - \cdots - \theta_r x_{ri})^2, \quad (4)$$

$$A\theta \le b. \quad (5)$$

Як відомо у всіх випадках, якщо це можливо, в якості розв’язку бажано замість нелінійної моделі використовувати лінійну модель. Тому модель (4), (5) пропонується замінити наступною моделлю:

$$\min_{\theta_j, j=\overline{0,r}} \sum_{i=1}^{n} |y_i - \theta_0 - \theta_1 x_{1i} - \cdots - \theta_r x_{ri}|, \quad (6)$$

$$A\theta \le b. \quad (7)$$

Покажемо, що задача (6), (7) зводиться до наступної задачі лінійного програмування (ЗЛП):

$$\min_{\theta_j, j=\overline{0,r}} \sum_{i=1}^{n} (u_i^+ + u_i^-), \quad (8)$$

$$y_i - \theta_0 - \theta_1 x_{1i} - \cdots - \theta_r x_{ri} = u_i^+ - u_i^-, \quad (9)$$

$$u_i^+ \ge 0, u_i^- \ge 0, i = \overline{1,n},$$

$$A\theta \le b. \quad (10)$$

Змінними ЗЛП (8)–(10) є $\theta_j, j = \overline{0,r}$, u_i^+, u_i^-, $i = \overline{1,n}$. ЗЛП (8)–(10) повинна розв’язуватися лише симплекс-методом. Тільки в цьому випадку для $\forall i$ виконується $u_i^+ \cdot u_i^- \equiv 0$.

Примітка 1. Легко бачити, що ЗЛП (8)–(10) представлена в стандартній формі, містить меншу кількість рівнянь і змінних в порівнянні з ЗЛП

$$\min_{\theta_j, j=\overline{0,r}} \sum_{i=1}^{n} y_i, \quad (11)$$

$$-y_i \le y_i - \theta_0 - \theta_1 x_{1i} - \cdots - \theta_r x_{ri} \le y_i, \quad (12)$$

$$y_i \ge 0, i = \overline{1,n}, \; A\hat{\theta} \le b. \quad (13)$$

 ISSN 1560-8956

Очевидно, що ЗЛП (11)–(13) також еквівалентна задачі (6), (7).

Хоча задача побудови БЛР з залежними коефіцієнтами є узагальненням стандартної задачі побудови БЛР, виникає питання, чи має це узагальнення практичне значення. Наступні два приклади дають позитивну відповідь на це питання.

Приклад 1. Нехай вхідні змінні $x_j, j = \overline{1, r}$, трактуються як фактори, які впливають на значення вихідної змінної y, $x_j \in [a_j; b_j], a_j \geq 0,\ j = \overline{1, r}$. Причому якісний аналіз досліджуваного об’єкта дозволяє зробити обґрунтовані висновки про те, що деякі коефіцієнти $\theta_{je}, e = \overline{1, L}$, обов’язково додатні числа, а деякі $\theta_{jm}, m = \overline{1, M}$ – від’ємні числа. Покажемо, що якщо кількість випробувань n лише в рази (2…4) більше, ніж $r + 1$, регресійна модель (1) повинна мати додаткові обмеження $\theta_{je} \geq 0, e = \overline{1, L}, \theta_{jm} \leq 0, m = \overline{1, M}$.

Для виконання наступного дослідження область задання ідеальних оцінюваних коефіцієнтів було обрано [0, 10), а знаки коефіцієнтів було обрано випадковим чином. Кількість випробувань у дослідженні перевищувала кількість оцінюваних коефіцієнтів спочатку у 2 рази, тобто становила 10, 20, 30, 40, 50, 60 та 70 (кількість оцінюваних коефіцієнтів відповідно становила 5, 10, 15, 20, 25, 30, 35). Було додано обмеження до задачі лінійного програмування на знаки оцінюваних коефіцієнтів. Згідно цієї додаткової інформації отримано результати, які відображені у табл. 1, що містить середній відсоток випадків, у яких метод суми модулів (МСМ) з додатковими обмеженнями побудував кращу регресію, ніж метод найменших квадратів (МНК) без додаткових обмежень, а також середні показники абсолютних значень міри порівнянь для МНК та МСМ (відповідно avg.ls та avg.sa). Останній рядок avg.% містить середнє значення відсотків по відповідному стовпцю. У якості інтегральної міри порівнянь (МП) компонент двох векторів θ і $\hat{\theta}$ була обрана міра

$$\|\theta - \hat{\theta}\| = \left\| \frac{\theta}{\|\theta\|} - \frac{\hat{\theta}}{\|\theta\|} \right\|, \tag{14}$$

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

З табл. 1 випливає, що навіть при найбільшому значенні дисперсії випадкової величини E 100 – середній відсоток випадків, у яких МСМ краще побудував регресію, ніж МНК, становить 65.86%. Як показано в [10], допустимій точності оцінки коефіцієнтів відповідає значення міри, що належить відрізку [0, 0.12]. При збільшенні кількості випробувань – до рівня відношення до кількості оцінюваних коефіцієнтів у 3 та 4 рази – відсоток випадків, у яких модель з додатковими обмеженнями краща – зменшується до значення 62.86% та 62.14% відповідно. З іншого боку, при збільшенні кількості випробувань, міри порівнянь МСМ та МНК зменшуються у своїх значеннях – у середньому на 0.041 та 0.085 відповідно.

Таблиця 1

Результати статистичних досліджень

Сер.кв відх. / **К-ть коеф.**	**1**	**2**	**3**	**4**	**5**	**10**
5	75% avg.ls:0.106, avg.sa:0.114	60% avg.ls:0.184, avg.sa:0.196	52% avg.ls:0.258, avg.sa:0.248	63% avg.ls:0.355, avg.sa:0.277	61% avg.ls:0.428, avg.sa:0.335	73% avg.ls:0.693, avg.sa:0.414
10	81% avg.ls:0.072, avg.sa:0.079	62% avg.ls:0.151, avg.sa:0.143	53% avg.ls:0.205, avg.sa:0.207	63% avg.ls:0.295, avg.sa:0.223	63% avg.ls:0.358, avg.sa:0.288	71% avg.ls:0.587, avg.sa:0.367
15	81% avg.ls:0.071, avg.sa:0.08	64% avg.ls:0.146, avg.sa:0.138	62% avg.ls:0.211, avg.sa:0.193	60% avg.ls:0.253, avg.sa:0.205	65% avg.ls:0.292, avg.sa:0.233	67% avg.ls:0.486, avg.sa:0.311
20	91% avg.ls:0.059, avg.sa:0.059	77% avg.ls:0.098, avg.sa:0.098	64% avg.ls:0.169, avg.sa:0.163	72% avg.ls:0.23, avg.sa:0.171	65% avg.ls:0.273, avg.sa:0.222	68% avg.ls:0.441, avg.sa:0.305
25	94% avg.ls:0.045, avg.sa:0.056	71% avg.ls:0.103, avg.sa:0.108	72% avg.ls:0.15, avg.sa:0.139	54% avg.ls:0.185, avg.sa:0.19	66% avg.ls:0.242, avg.sa:0.186	57% avg.ls:0.39, avg.sa:0.31
30	97% avg.ls:0.042, avg.sa:0.05	79% avg.ls:0.083, avg.sa:0.092	72% avg.ls:0.137, avg.sa:0.137	72% avg.ls:0.195, avg.sa:0.145	63% avg.ls:0.229, avg.sa:0.194	56% avg.ls:0.353, avg.sa:0.305
35	100% avg.ls:0.077, avg.sa:0.071	84% avg.ls:0.113, avg.sa:0.092	68% avg.ls:0.143, avg.sa:0.123	59% avg.ls:0.158, avg.sa:0.143	67% avg.ls:0.176, avg.sa:0.158	69% avg.ls:0.356, avg.sa:0.247
avg.%	88.43%	71%	63.29%	63.29%	64.29%	65.86%

Отже, навіть у найгіршому випадку при дисперсії випадкової величини E – 100, у середньому відсоток випадків, у яких МСМ краще побудував регресію, ніж МНК, за критерієм становить більше 65%. При збільшенні кількості випробувань даний відсоток зменшується. Даний результат дозволяє впевнено використовувати МСМ з додатковими обмеженнями при дисперсії випадкової величини E від 1 до 100 та невеликій кількості випробувань (в 2–4 рази більше кількості оцінюваних коефіцієнтів).

Приклад 2. Нехай вхідні змінні $x_j, j = \overline{1,r}$ $(x_j \in [a,b], a > 0)$, трактуються як фактори, які впливають на вихідну змінну об'єкта дослідження в якості зваженої лінійної згортки $\sum_{j=1}^{n} \theta_j x_j$, $\forall_j \theta_j \geq 0$.

Примітка 2. В цьому випадку в регресійній моделі (1) константа θ_j може бути відсутня, а випадкова величина E трактується як помилка вимірювання, дисперсія якої не дозволяє їй виключити її з моделі (1).

Тоді логічно припустити, що чим більше j-й фактор впливає на величину вихідної змінної, тим більше значення приймає невідомий коефіцієнт θ_j.

Тобто, якщо експертним шляхом обґрунтовано знайдено множину

$$\{x_{j_e} \succ x_{j_m}, e \neq m\}, \quad (15)$$

де $x_{j_e} \succ x_{j_m}, e \neq m$ означає, що вхідна змінна x_{j_e} більше впливає на вихідну змінну, ніж вхідна змінна x_{j_m}, а множина пар (15) задовольняє умові транзитивності, то для випадку, коли кількість випробувань n лише в рази (2–4) більше кількості невідомих коефіцієнтів в регресійній моделі слушно використати додаткову інформацію, що породжується множиною (15), а саме додати множину лінійних обмежень $\{\theta_{j_e} \geq \theta_{j_m}, e \neq m\}$.

Розглянемо частковий випадок, коли існують P експертів, кожний з яких спроможний суб’єктивно розташувати вхідні фактори в повному порядку.

Для знаходження залежностей типу (15) можна запропонувати метод, що використовує ідеї правила колективного парного порівняння (ПКПП).

Означення. ПКПП – зветься правило, відповідно до якого змінна x_j більше впливає на вихідну змінну, ніж вхідна змінна x_i ($x_j \succ x_i$), якщо кількість експертів, що вважають, що $x_j \succ x_i$ більше, ніж кількість експертів, що вважають ($x_i \succ x_j$). Якщо ця кількість однакова, то ($x_j \sim x_i$).

Безпосередньо ПКПП використовувати не можна, бо в загальному випадку для нього не виконується умова транзитивності. Тобто, наприклад, можливо отримати $x_{i_1} \succ x_{i_2}$, $x_{i_2} \succ x_{i_3}$, $x_{i_3} \succ x_{i_1}$.

Тому пропонується наступна очевидна модифікація ПКПП для знаходження нерівностей, що накладаються на коефіцієнти $\theta_j, j = \overline{1,r}$.

Нехай індивідуальні переваги експертів задаються повними порядками. Тобто виключаємо можливість, коли експерт вважає $x_j \sim x_i \Rightarrow \theta_j = \theta_i$. Знаходимо множину всіх пар вхідних змінних $\{(x_i, x_j)\}$, для яких виконується

$$\{x_i \succ_P x_j\}, \quad (16)$$

причому для кожної такої пари вхідних змінних рішення приймається одноголосно. Якщо отримані колективні парні переваги не порушують умови транзитивності, то вони породжують множину нерівностей

$$\{\theta_i \geq \theta_j\}. \quad (17)$$

В протилежному випадку інформацію експертів вважаємо не достатньо достовірною, і множина (16) не породжує жодної нерівності з множини (17).

Можна запропонувати наступний алгоритм перевірки виконання умови транзитивності на множині парних переваг (16).

Перший крок. По множині (16) будуємо орієнтований граф, у якого вершини – це всі вхідні змінні, що входять в множину (16), а кожній парі ($x_i \succ_P x_j$), що належить множині (16), відповідає орієнтоване ребро, що виходить з вершини x_i і входить у вершину x_j.

Другий крок. Перевірка візуально чи з використанням штучної нейронної мережі, чи існують в побудованому графі орієнтовані цикли. Якщо їх немає, то множина (16) задовольняє умови транзитивності; інакше, експертна інформація вважається недостовірною, і множина (16) не породжує лінійних обмежень в ЗЛП, що мінімізує суму модулів.

3. Результати емпіричного порівняльного аналізу критеріїв «мінімум суми квадратів» та «мінімум суми модулів»

Так як в даній роботі пропонується задачу квадратичного програмування замінити ЗЛП шляхом заміни критерія методу найменших квадратів на критерій мінімуму суми модулів відповідних різниць (МСМ), виникає проблема емпіричного порівняння ефективності цих критеріїв в регресійному аналізі.

Для цього досліджувались два методи оцінки коефіцієнтів моделі (1) – це метод найменших квадратів та розв’язок ЗЛП (8)–(10). Тобто без обмежень $B\theta \leq b$. Обидві оцінки є лінійними відносно вектора y, ($\hat{\theta} = Uy$, де в МНК $U = (A^T A)^{-1} A^T$), але лінійні оцінки, що є розв’язком ЗЛП (8)–(10) симплекс-методом, не задовольняють умовам теореми Маркова, так як матриця U залежить від значень компонент вектора y, і тому їх порівняльний аналіз ефективності не є тривіальним. Дійсно, відповідна матриця U є оптимальним базисом ЗЛП (8)–(10), заданої в стандартній формі, а він залежить від значень компонент вектора y.

Було проведено статистичне імітаційне дослідження, в рамках якого виконано порівняння двох методів оцінювання параметрів – методу найменших квадратів (МНК) та мінімуму суми модулів (МСМ). Дослідження здійснювалося на основі імітаційного моделювання, що дозволило оцінити точність та ефективність кожного з методів в умовах різних рівнів похибки та варіацій параметрів.

Для проведення експериментів було використано такі вхідні дані:

- Кількість повторень кожного експерименту – 100.
- Кількість оцінюваних коефіцієнтів – тестувалися такі значення: [3, 5, 10, 15, 20, 25, 30, 35]. Область задання оцінюваних коефіцієнтів – [–10, 10).
- Середньо квадратичне відхилення випадкової величини E, що має нормальний розподіл – приймало значення: [1, 2, 3, 4, 5, 10]. $ME = 0$.
- Кількість випробувань – обиралася пропорційно до кількості оцінюваних коефіцієнтів – значення кількості випробувань перевищувало кожне фіксоване значення оцінюваних коефіцієнтів у 2, 3, 4 та 5 разів.

Табл. 2 містить середній відсоток результатів експериментів, у яких МСМ побудував кращу регресію у порівнянні з МНК для різних комбінацій параметрів. В конкретному експерименті МСМ та МНК порівнюються з використання критерію на основі міри порівнянь (14), згідно якого МСМ вважається кращим за МНК: якщо обидва методи мають міру порівнянь, нижчу за поріг точності 0.12, – вважається, що точність обох методів практично еквівалентна, і в такому разі МСМ вважається прийнятним. В інших випадках МСМ вважається кращим, ніж МНК, якщо значення міри порівнянь МСМ є меншим, ніж МНК.

Структура таблиці:

- Перший рядок: значення середньо квадратичного відхилення випадкової величини E.
- Перший стовпець: значення кількості оцінюваних параметрів.
- Значення у комірках: середній відсоток випадків по всім випробуванням, у яких МСМ побудував кращу регресію, ніж МНК за наведеним раніше критерієм.

Таблиця 2

Результати статистичних досліджень

Сер. кв.відх. / К-ть оцін. коеф.	1	2	3	4	5	10
3	79.25	61	48.75	50.5	43	42
5	81.75	63.5	51.25	48.75	46.25	42.25
10	89	66	54	48.5	45.75	42.5
15	91	71.75	58.5	54.5	42.75	42.25
20	93	78	61.25	57.5	53	45
25	96.5	77.75	62.25	60.25	53.5	41.25
30	90	83	71.5	58.75	50.75	44.5
35	96	82.75	67.5	62.25	59	44

Результати табл. 2 показують, що при низькому рівні середньо квадратичного відхилення – 1, відсоток кращих результатів перевищує 79.25% для всіх випадків. Для великих моделей (наприклад, кількості оцінюваних коефіцієнтів 25, 30, 35) цей показник наближається до 96%, що вказує на стабільну перевагу МСМ.

Зі збільшенням значення середньо квадратичного відхилення методи стають ближчими за ефективністю – при значенні 5 частка випробувань, у яких МСМ перевершує МНК, в середньому становить 49.25%. При значенні 10 показники знижуються, але все ж залишаються на рівні 43%, що свідчить про те, що МСМ не поступається МНК і в деяких випадках навіть є ефективнішим.

Для всіх значень середньо квадратичного відхилення видно, що при збільшенні кількості параметрів (від 3 до 35) МСМ частіше перевершує МНК – це свідчить про те, що метод краще адаптується до моделей з більшою кількістю параметрів.

Отже, з результатів табл. 2 можна зробити наступні висновки:

- Метод мінімуму суми модулів у значному відсотку випадків перевершує метод найменших квадратів, особливо при низькому рівні середньо квадратичного відхилення випадкової величини E.

- При збільшенні середньо квадратичного відхилення випадкової величини E методи стають ближчими, але навіть у найгіршому випадку метод мінімуму суми модулів дає кращі результати у 41–45% випадків, що є доволі високим показником.
- При більшій кількості оцінюваних коефіцієнтів метод мінімуму суми модулів демонструє стабільнішу перевагу над методом найменших квадратів.

Табл. 3 містить детальну статистику випадків, у яких МСМ побудував регресію гірше, ніж МНК, та відповідні значення міри порівнянь вийшли за межі порогу точності 0.12. Дані у таблиці відображають найгірший випадок (найбільшу різницю між МП МСМ та МП МНК) та розподіл випробувань за діапазонами відносно даного найгіршого випадку. Дані у таблиці згруповані за середньо квадратичним відхиленням випадкової величини E. Кожна комірка містить середні показники, що характеризують МНК та МСМ (причому усереднення береться як по значенням кількості оцінюваних коефіцієнтів так і по значенням кількості проведених випробувань, що були описані у параграфі вхідних даних). Іншими словами, таблиця відображає, наскільки значною була різниця між МНК та МСМ у випадках, коли метод мінімуму суми модулів показав гірші результати.

Структура таблиці:

- Перший рядок: значення середньо квадратичного відхилення випадкової величини E.
- Перший стовпець: позначення наступних показників: n.c. – значення кількості оцінюваних коефіцієнтів, при якому спостерігається найбільша різниця між МП МСМ та МП МНК; n.e. – значення кількості випробувань, при якому спостерігається найбільша різниця між МП МСМ та МП МНК; max.diff. – значення найбільшої різниці між МП МСМ та МП МНК; max.ls – значення міри порівнянь методу найменших квадратів, при якому спостерігається найбільша різниця між МП МСМ та МП МНК; max.sa – значення міри порівнянь методу мінімуму суми модулів, при якому спостерігається найбільша різниця між МП МСМ та МП МНК; рядки позначені r.d.i ($i = \overline{1,5}$) – діапазони значень різниці між МП МСМ та МП МНК, комірки яких містять наступні показники: відсоток випробувань, різниця між мірами порівнянь для кожного з яких, входить у даний діапазон; avg.ls – середнє значення МП МНК серед випробувань, які належать даному діапазону; avg.sa – середнє значення МП МСМ серед випробувань, які належать даному діапазону.

Детальний аналіз табл. 3 показує, що найбільші розбіжності між мірами порівнянь МСМ та МНК (r.d.1) спостерігаються в середньому лише у 0.2% загальної кількості випробувань. У середньому 93.14% експериментів потрапляють у діапазон r.d.5, що означає, що МСМ дає дуже схожі результати з МНК у переважній більшості випробувань. Середні значення МП для обох методів свідчать про те, що навіть у випадках, коли метод МСМ поступається МНК, його значення похибки залишається порівнянним з МНК. Наприклад, при середньо квадратичному відхиленні випадкової величини E – 10 та діапазоні r.d.5, у 93.75% експериментів середня міра порівнянь МНК становить 0.360, а МСМ – 0.533, що показує, що навіть у крайніх випадках МСМ залишається у контрольованих межах.

Таблиця 3

Результати статистичних досліджень

σ / Метрика	1	2	3	4	5	10
n.c.	3	3	3	3	5	3
n.e.	12	6	6	12	15	12
max.diff.	0.340	1.125	1.488	1.413	1.229	1.706
max.ls	0.066	0.428	0.327	0.338	0.087	0.169
max.sa	0.406	1.553	1.815	1.751	1.316	1.875
r.d.1	[0.306, 0.340]: 0.692% avg.ls: 0.150, avg.sa: 0.484	[1.013, 1.125]: 0.116% avg.ls: 0.428, avg.sa: 1.553	[1.339, 1.488]: 0.077% avg.ls: 0.327, avg.sa: 1.815	[1.272, 1.413]: 0.070% avg.ls: 0.338, avg.sa: 1.751	[1.106, 1.229]: 0.062% avg.ls: 0.087, avg.sa: 1.316	[1.535, 1.706]: 0.164% avg.ls: 0.158, avg.sa: 1.823
r.d.2	[0.238, 0.306): 1.730% avg.ls: 0.116, avg.sa: 0.385	[0.788, 1.013): 0.231% avg.ls: 0.340, avg.sa: 1.181	[1.041, 1.339): 0.000% avg.ls: 0.000, avg.sa: 0.000	[0.989, 1.272): 0.070% avg.ls: 0.132, avg.sa: 1.132	[0.860, 1.106): 0.246% avg.ls: 0.349, avg.sa: 1.340	[1.194, 1.535): 0.438% avg.ls: 0.193, avg.sa: 1.537
r.d.3	[0.170, 0.238): 3.114% avg.ls: 0.214, avg.sa: 0.404	[0.563, 0.788): 0.116% avg.ls: 0.251, avg.sa: 0.851	[0.744, 1.041): 0.154% avg.ls: 0.199, avg.sa: 1.008	[0.707, 0.989): 0.279% avg.ls: 0.251, avg.sa: 1.065	[0.615, 0.860): 0.862% avg.ls: 0.333, avg.sa: 1.038	[0.853, 1.194): 0.767% avg.ls: 0.292, avg.sa: 1.285
r.d.4	[0.102, 0.170): 17.301% avg.ls: 0.072, avg.sa: 0.198	[0.338, 0.563): 1.618% avg.ls: 0.256, avg.sa: 0.683	[0.447, 0.744): 1.154% avg.ls: 0.254, avg.sa: 0.795	[0.424, 0.707): 2.298% avg.ls: 0.283, avg.sa: 0.799	[0.369, 0.615): 4.741% avg.ls: 0.286, avg.sa: 0.739	[0.512, 0.853): 4.877% avg.ls: 0.243, avg.sa: 0.876
r.d.5	[1.502e-05, 0.102): 77.163% avg.ls: 0.127, avg.sa: 0.178	[4.328e-04, 0.338): 97.919% avg.ls: 0.133, avg.sa: 0.221	[3.093e-04, 0.447): 98.615% avg.ls: 0.169, avg.sa: 0.279	[2.636e-04, 0.424): 97.284% avg.ls: 0.197, avg.sa: 0.315	[5.409e-04, 0.369): 94.089% avg.ls: 0.217, avg.sa: 0.341	[1.271e-04, 0.512): 93.753% avg.ls: 0.360, avg.sa: 0.533

Отже, аналіз даних з табл. 3 дозволяє сформулювати наступні висновки:

- При середньоквадратичному відхиленні випадкової величини E більше 1 лише 0.06–4.88% випадків мають суттєву різницю між мірами порівнянь МСМ та МНК (r.d.1 – r.d.4).
- У середньому 93.14% випадків характеризуються різницею 0.111 між значеннями МП МНК та МП МСМ (r.d.5), що доводить ефективність МСМ навіть у випадках гірше побудованої регресії, ніж МНК.

Таким чином, проведені дослідження дозволяють зробити висновок, що використання методу мінімуму суми модулів є більш громіздким у порівнянні з класичним методом найменших квадратів (замість реалізації формули $\widehat{\theta_J} = (A^T A)^{-1} A^T y$ треба розв’язати відповідну задачу лінійного програмування), проте демонструє в середньому вищу ефективність та стабільність результатів для симетричних відносно нуля областей значень вхідних змінних.

Висновки

1. Сформульована задача побудови багатовимірної лінійної регресії з залежними коефіцієнтами.

2. Запропоновано знаходити оцінки значень лінійно залежних коефіцієнтів багатовимірної лінійної регресії, мінімізуючи не суму квадратів різниць, а суму модулів різниць, розв’язуючи відповідну задачу лінійного програмування.

3. Наведено два приклади, що підтверджують практичне значення запропонованої задачі регресії з залежними коефіцієнтами та запропонований метод її розв’язання.

4. Наведені результати порівняльного емпіричного аналізу ефективності використання критеріїв мінімуму суми квадратів різниць та мінімуму суми модулів різниць для знаходження оцінок незалежних коефіцієнтів багатовимірної лінійній регресії. Вони показують, що однозначно жодному критерію не можна заздалегідь віддати перевагу, але для симетричних відносно нуля областей значень вхідних змінних МСМ демонструє в середньому вищу ефективність та стабільність результатів в порівнянні з МНК.

Список використаних джерел

1. Roustaei N. Application and interpretation of linear-regression analysis. Medical hypothesis discovery and innovation in ophthalmology. 2024. Vol. 13. No. 3. P. 151–159. DOI: https://doi.org/10.51329/mehdiophthal1506.

2. Liu J., Bellows B., Hu X.J., Wu J., Zhou Z., Soteros C., Wang L. A new time-varying coefficient regression approach for analyzing infectious disease data. Scientific Reports. 2023. Vol. 13. Article number 14687. DOI: https://doi.org/10.1038/s41598-023-41551-1.

3. 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.

4. 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.

5. Zhou Z., Ji J., Wang Y. Zhu Z., Chen J. Hybrid regression model via multivariate adaptive regression spline and online sequential extreme learning machine and its application in vision servo system. International Journal of Advanced Robotic Systems. 2022. Vol. 19. No. 3. P. 1–13. DOI: https://doi.org/10.1177/17298806221108603.

6. 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.

7. Abdulrahman A.T., Alshammari N.S. Factor analysis and regression analysis to find out the influencing factors that led to the countries’ debt crisis. Advances and Applications in Statistics. 2022. Vol. 78. P. 1–16. DOI: https://doi.org/10.17654/0972361722047.

8. Pavlov, A., Holovchenko, M., & Drozd, V. (2024). 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.

9. Pavlov A. A., Holovchenko M. N., Drozd V.V. Efficiency substantiation for a synthetical method of constructing a multivariate polynomial regression given by a redundant representation. Bulletin of National Technical University "KhPI". Series: System Analysis, Control and Information Technologies. 2024. No. 1(9). P. 3–9. DOI: https://doi.org/10.20998/2079-0023.2023.01.01.

10. Zgurovsky M.Z., Pavlov A.A. Combinatorial Optimization Problems in Planning and Decision Making: Theory and Applications, 1st edn. Studies in Systems, Decision and Control. Vol. 173 // Springer, 2019. DOI: https://doi.org/10.1007/978-3-319-98977-8.