КІБЕРНЕТИКА та КОМП'ЮТЕРНІ ТЕХНОЛОГІЇ

УДК 519.85 DOI:10.34229/2707-451X.20.3.2

В.О. СТОВБА

МЕТОД ЕЛІПСОЇДІВ ДЛЯ ЗНАХОДЖЕННЯ ПАРАМЕТРІВ ЛІНІЙНОЇ РЕГРЕСІЇ

Описано задачу визначення параметрів лінійної регресії у формі задачі мінімізації негладкої функції, що являє собою L_p-норму вектора-нев'язки системи лінійних рівнянь. Наведено загальна схема алгоритма методу еліпсоїдів для мінімізації цієї функції при довільному значенні параметра $p \geq 1$. Описано спосіб запису задачі апроксимації спостережень квадратичною функцією як задачі визначення параметрів лінійної регресії. Проаналізовано результати обчислювальних експериментів для двох прикладів апроксимації спостережень лінійною та квадратичною функціями з використанням алгоритму методу еліпсоїдів.

Ключові слова: *метод еліпсоїдів, лінійна регресія, аномальні спостереження.*

Вступ. Лінійна регресія – один з найпоширеніших інструментів регресійного аналізу в багатьох областях науки. Вона використовується для аналізу експериментальних даних в економіці, психології, соціології, фізиці, хімії, геології тощо. Наприклад, в економіці на її основі будуються багатофакторні моделі продуктивності праці та функцій попиту, виробничих функцій та економіко-статистичних моделей [1].

Варто зазначити, що лінійна регресія найчастіше використовується та є найбільш досконало вивченою в економетриці. З іншого боку, ця модель є типовим представником алгоритмів машинного навчання, підрозділу штучного інтелекту. Як і задача класифікації, вона відноситься до класу задач навчання з учителем, де за заданим набором даних про об'єкт, що спостерігається, необхідно спрогнозувати певну цільову змінну.

Задачу визначення параметрів лінійної регресії можна сформулювати як задачу мінімізації негладкої функції. В основі такої оптимізаційної задачі стоїть спосіб знаходження вектора невідомих, який мінімізує L_p-норму вектора-нев'язки системи лінійних рівнянь, що задається за допомогою матриці плану регресії. Для її розв'язання можна використовувати методи мінімізації негладких функцій, зокрема, субградієнтні методи, такі як метод еліпсоїдів, *r*-алгоритми тощо.

В статті розглянемо задачу мінімізації негладкої функції для визначення параметрів лінійної регресії, та її розв'язання за допомогою методу еліпсоїдів при довільному значенні параметра $p \geq 1$. Проаналізуємо результати обчислювальних експериментів для двох прикладів апроксимації спостережень лінійною та квадратичною функціями. Наведемо також спосіб запису задачі апроксимації спостережень квадратичною функцією як задачі визначення параметрів лінійної регресії.

1. Постановка задачі. Розглянемо регресійну модель

$$y = f(x, b) + \varepsilon, \qquad (1)$$

де (x_i, y_i), $i = \overline{1,n}$ – вибірка спостережень, причому $x_i \in \mathbb{R}^m$, $y_i \in \mathbb{R}$. Також $b \in \mathbb{R}^m$ – вектор параметрів

(коефіцієнтів) регресії, $\varepsilon \in \mathbb{R}^n$ – вектор похибок спостережень (математичне сподівання $E(\varepsilon_i) = 0$, $i = \overline{1,n}$). Якщо функція $f(x,b)$ є лінійною, тобто

$$f(x,b) = b_1 x_1 + ... + b_m x_m, \tag{2}$$

модель (1) називається *лінійною* регресією. Тут $x_j \in \mathbb{R}^n$, $j = \overline{1,m}$ – регресори або фактори моделі. Необхідно знайти такий вектор параметрів b, при якому вектор похибок $\varepsilon = y - f(x,b) = (\varepsilon_1,...,\varepsilon_n)^T$ буде мінімальним.

Якщо в моделі наявний лише один фактор (без урахування константи), говорять про *парну* регресію, причому модель (1) набуває вигляду

$$y_i = b_0 + b_1 x_{1i} + \varepsilon_i, \quad i = \overline{1,n}, \tag{3}$$

де x_{1i} – значення першого (і єдиного) фактору в i-му спостереженні. Цю модель можна представити також у матричному вигляді:

$$y = Xb + \varepsilon, \tag{4}$$

де $y = (y_1,...,y_n)^T$, $b = (b_0, b_1)^T$, X – $n \times m$-матриця факторів, яка має вигляд

$$X = \begin{pmatrix} 1 & x_{11} \\ 1 & x_{12} \\ \vdots & \vdots \\ 1 & x_{1n} \end{pmatrix}.$$

Задачу (3) можна представити як задачу знаходження «найкращої» L_p-норми вектора-нев'язки ε для системи (4) [2 – 4]. Їй відповідає задача мінімізації негладкої функції

$$f_p(b) = \|Xb - y\|_p = \|\varepsilon\|_p = \left(\sum_{i=1}^{n} |\varepsilon_i|^p \right)^{1/p}, \tag{5}$$

тобто знаходження

$$b_p^* = \arg\min_{b \in \mathbb{R}^m} f_p(b). \tag{6}$$

Тут $p \in \mathbb{R}$ – скалярний параметр, який при $p \geq 1$ забезпечує опуклість функції $f_p(b)$. Задача (5), (6) завжди має розв'язок, а якщо $n > m$ та матриця X має повний ранг – цей розв'язок єдиний. В загальному випадку розв'язок не обов'язково єдиний і точка b_p^* належить множині оптимальних розв'язків, яким відповідає мінімальне значення функції f_p^* в задачі (5), (6).

Для розв'язання цієї задачі доцільно використовувати методи мінімізації негладких опуклих функцій і такі спроби неодноразово виконувались. Наприклад, в роботі [4] запропоновано доволі загальний підхід для оцінки невідомих параметрів у класичній лінійній моделі, що базується на модифікації *r*-алгоритму Шора – Журбенко. В роботі [5] на основі методу еліпсоїдів [6] розроблено алгоритми та їх програмні реалізації для знаходження точки мінімуму $f_p(\cdot)$ при двосторонніх обмеженнях на змінні.

2. Мінімізація функції (5) при великих значеннях параметра *p*. Задачу (5), (6) можна розв'язувати при різних значеннях параметра p. Наприклад, випадок $p = 1$ відповідає методу найменших модулів (МНМ), випадок $p = 2$ – методу найменших квадратів (МНК), а випадок $p = \infty$ –

мінімаксному методу Чебишева. Кожен з них дає свій розв'язок, який може бути більш якісним та стійким залежно від типу розподілу випадкової величини, що характеризує похибку результатів спостережень. В роботі [7] задачу (5), (6) розв'язано з використанням методу еліпсоїдів для значень параметра $1 \le p \le 2$. При довільних значеннях параметра p ця задача є загальною задачею безумовної мінімізації опуклої негладкої функції $F_p(b)$. Для її розв'язання можна використовувати методи мінімізації опуклих функцій, наприклад, метод еліпсоїдів або *r*-алгоритми. Їх використання потребує обчислення значення функції $f_p(b)$ та її субградієнта $g_f(b)$ в точці b; перше можна обчислити за формулою (5), друге – за формулою

$$g_f(b) = \|Xb - y\|_p^{1-p} \sum_{i=1}^{n} \operatorname{sgn}(x_i b - y_i) \cdot |x_i b - y_i|^{p-1} x_i^T . \qquad (7)$$

При великих значеннях параметра p розрахунок цих величин є доволі трудомісткою задачею. В такому випадку мінімізація функції (5) еквівалентна мінімізації функції $f_p(b) = \max\limits_{i=1,n} |x_i b - y_i|$, що відповідає випадку $p = \infty$. Для спрощення обчислення та його коректності вираз (5) для обчислення значення функції пропонується змінити таким чином:

$$f_p(b) = \left(\sum_{i=1}^{n} |\varepsilon_i|^p \right)^{1/p} = \frac{|\varepsilon_{max}|}{|\varepsilon_{max}|} \left(\sum_{i=1}^{n} |\varepsilon_i|^p \right)^{1/p} = \frac{|\varepsilon_{max}|}{\sqrt[p]{|\varepsilon_{max}|^p}} \left(\sum_{i=1}^{n} |\varepsilon_i|^p \right)^{1/p} = |\varepsilon_{max}| \left(\sum_{i=1}^{n} \frac{|\varepsilon_i|^p}{|\varepsilon_{max}|^p} \right)^{1/p} =$$

$$= |\varepsilon_{max}| \left(\sum_{i=1}^{n} \left| \frac{\varepsilon_i}{\varepsilon_{max}} \right|^p \right)^{1/p} = |\varepsilon_{max}| \left(\sum_{i=1}^{n} |z_i|^p \right)^{1/p} = |\varepsilon_{max}| \cdot \|z\|_p , \text{ де } \varepsilon_{max} = \max_{i=1,n} |\varepsilon_i|, \; z_i = \frac{\varepsilon_i}{\varepsilon_{max}} \le 1, \; i = \overline{1,n}. \qquad (8)$$

Формула (7) для обчислення субградієнта функції модифікується так:

$$g_f(b) = \|Xb - y\|_p^{1-p} \sum_{i=1}^{n} \operatorname{sgn}(x_i b - y_i) \cdot |x_i b - y_i|^{p-1} x_i^T = \|\varepsilon\|_p^{1-p} \sum_{i=1}^{n} \operatorname{sgn}(\varepsilon_i) \cdot |\varepsilon_i|^{p-1} x_i^T =$$

$$= \|\varepsilon\|_p^{1-p} \frac{|\varepsilon_{max}|^{p-1}}{|\varepsilon_{max}|^{p-1}} \sum_{i=1}^{n} \operatorname{sgn}(\varepsilon_i) \cdot |\varepsilon_i|^{p-1} x_i^T = \|\varepsilon\|_p^{1-p} |\varepsilon_{max}|^{p-1} \sum_{i=1}^{n} \operatorname{sgn}(\varepsilon_i) \cdot |z_i|^{p-1} x_i^T =$$

$$= \left(\frac{\|\varepsilon\|_p}{|\varepsilon_{max}|} \right)^{1-p} \sum_{i=1}^{n} \operatorname{sgn}(\varepsilon_i) \cdot |z_i|^{p-1} x_i^T = \left(\left(\sum_{i=1}^{n} |\varepsilon_i|^p \right)^{1/p} \Big/ |\varepsilon_{max}| \right)^{1-p} \sum_{i=1}^{n} \operatorname{sgn}(\varepsilon_i) \cdot |z_i|^{p-1} x_i^T =$$

$$= \left(\left(\sum_{i=1}^{n} \left| \frac{\varepsilon_i}{\varepsilon_{max}} \right|^p \right)^{1/p} \right)^{1-p} \sum_{i=1}^{n} \operatorname{sgn}(\varepsilon_i) \cdot |z_i|^{p-1} x_i^T = \|z\|_p^{1-p} \sum_{i=1}^{n} \operatorname{sgn}(\varepsilon_i) \cdot |z_i|^{p-1} x_i^T , \qquad (9)$$

де x_i – вектор-рядок матриці X з номером $i = \overline{1,n}$.

Якщо $\varepsilon_{max} \ne 0$ всі перетворення є коректними. Умова $\varepsilon_{max} = 0$ коректно опрацьовується в програмній реалізації методу за допомогою перевірки відповідної умови.

Ідея перетворення виразів (5) і (7) така. При великих значеннях параметра p обчислення L_p-норми вектора ε може призвести до накопичення обчислювальної похибки при виконанні операцій піднесення до степеню та додавання великих чисел. В модифікованих формулах (8) і (9)

такі операції проводяться над елементами вектора z_i, які або рівні одиниці (їм відповідають максимальні елементи вектора ε_i), або менші одиниці. Якщо $z_i = 1$ вищенаведені операції не змінюють його; якщо ж $z_i < 1$, то при великих значеннях параметра p таке значення прямує до нуля та інтерпретується як нуль після досягнення певної границі. Таким чином операції виконуються коректно, а значення функції та її субградієнта обчислюються досить точно навіть при великих значеннях параметра p.

3. Алгоритм методу еліпсоїдів. Ініціалізація. Вибираємо стартові точку $x_0 \in \mathbb{R}^m$, радіус $r_0 \in \mathbb{R}$ та $m \times m$-матрицю B_0, яку покладемо рівною одиничній матриці I_m. Перейдемо до першої ітерації зі значеннями x_0, r_0 та B_0.

Нехай на k-й ітерації отримали значення x_k, r_k та B_k. Для переходу на $(k+1)$-у ітерацію виконаємо таку послідовність кроків.

Крок 1. Обчислимо $f_p(x_k)$. Якщо $f_p(x_k) = 0$, тоді STOP: $k^* = k$ та $x_p^* = x_k$. Інакше обчислимо $g_p(x_k)$. Якщо $f_p(x_k) < +\infty$ та $\left\| B_k^T g_p(x_k) \right\| r_k \le \varepsilon_f$, тоді STOP: $k^* = k$ та $x_p^* = x_k$. Інакше переходимо до кроку 2.

Крок 2. Покладемо $\xi_k = \dfrac{B_k^T g_p(x_k)}{\left\| B_k^T g_p(x_k) \right\|}$.

Крок 3. Обчислимо наступну точку $x_{k+1} = x_k - h_k B_k \xi_k$, де $h_k = \dfrac{1}{n+1} r_k$.

Крок 4. Обчислимо $B_{k+1} = B_k + \left(\sqrt{\dfrac{n-1}{n+1}} - 1 \right) (B_k \xi_k) \xi_k^T$ та $r_{k+1} = r_k \dfrac{n}{\sqrt{n^2 - 1}}$.

Крок 5. Переходимо до $(k+1)$-ї ітерації зі значеннями x_{k+1}, r_{k+1} та B_{k+1}.

Теорема [8]. Послідовність точок $\{x_k\}_{k=0}^{k^*}$ задовольняє нерівності

$$\left\| B_k^{-1} \left(x_k - x_p^* \right) \right\| \le r_k, \; k = \overline{0, k^*}.$$

На кожній ітерації $k > 0$ величина зменшення об'єму еліпсоїда $E_k = \left\{ x \in \mathbb{R}^n : \left\| B_k^{-1} (x_k - x) \right\| \le r_k \right\}$, який локалізує x_p^*, є сталою і рівною

$$q = \frac{vol(E_k)}{vol(E_{k-1})} = \frac{n}{n+1} \left(\frac{n}{\sqrt{n^2 - 1}} \right)^{n-1} < \exp\left\{ -\frac{1}{2n} \right\} < 1.$$

4. Приклад 1: лінійна функція. В роботах [3, 7, 9] наводився приклад апроксимації набору з 6 спостережень, що містив аномальне, лінійною функцією. Така задача є прикладом задачі визначення параметрів лінійної регресії для функції однієї змінної. В роботі [3] показано, що перевагу варто віддавати методу найменших модулів (МНМ), адже він ігнорує аномальне спостереження і точно відновлює лінійну функцію. Для визначення стійкості класичних методів, що відповідають значенням $p = 1, 2, \infty$, розглянемо аналогічний приклад з більшою кількістю спостережень та вищим відсотком аномалій і проведемо серію експериментів з ним.

Розглянемо набір спостережень (x_i, y_i), де $x_i = y_i = i-1$, $i = \overline{1,20}$, які необхідно апроксимувати лінійною функцією $y = cx + d$, де c та d – невідомі. Для імітації 5 % аномальних спостережень змінимо перше спостереження таким чином: $(x_1, y_1) = (0,19)$. Для досягнення 10 % аномалій окрім першого спостереження змінимо додатково друге: $(x_2, y_2) = (1,19)$, для 15 % – третє: $(x_3, y_3) = (2,19)$. Таким чином утворилась група аномалій ліворуч. Для побудови правосторонньої групи аномалій змінимо такі спостереження: $(x_{20}, y_{20}) = (19,0)$, $(x_{19}, y_{19}) = (18,0)$, $(x_{18}, y_{18}) = (17,0)$. Додаючи по одній аномалії, отримаємо 5, 10 та 15 % викидів відповідно. Для апроксимації збурених спостережень використаємо алгоритм методу еліпсоїдів зі значеннями параметра $p = 1,2,\infty$. Результати роботи алгоритму для збурених спостережень ліворуч наведено в табл. 1, праворуч – в табл. 2.

ТАБЛИЦЯ 1. Апроксимація збурених ліворуч спостережень (x_i, y_i) лінійною функцією за допомогою алгоритму методу еліпсоїдів при різних кількостях аномалій та значеннях параметра p

Відсоток аномалій	p	$\overline{c^*}$	$\overline{d^*}$	$f\left(\overline{c^*}, \overline{d^*}\right)$
5 %	1	1.0000	0.0000	19.000
	2	0.7285	3.5285	17.145
	∞	0.0000	3.0000	16.000
10 %	1	1.0000	0.0000	37.000
	2	0.4985	6.6142	21.196
	∞	0.0000	3.0000	16.000
15 %	1	1.0000	0.0000	54.000
	2	0.3067	9.2857	22.552
	∞	0.0000	3.0000	16.000

ТАБЛИЦЯ 2. Апроксимація збурених праворуч спостережень (x_i, y_i) лінійною функцією за допомогою алгоритму методу еліпсоїдів при різних кількостях аномалій та значеннях параметра p

Відсоток аномалій	p	$\overline{c^*}$	$\overline{d^*}$	$f\left(\overline{c^*}, \overline{d^*}\right)$
5 %	1	1.0000	– 7.5265e – 15	19.000
	2	0.72857	1.6286	17.145
	∞	5.8453e – 14	9.0000	9.0000
10 %	1	1.0000	3.3573e – 15	37.000
	2	0.4985	2.9143	21.197
	∞	5.5649e – 14	8.5000	8.500
15 %	1	1.0000	2.6091e – 14	54.000
	2	0.3067	3.8857	22.553
	∞	6.0668e – 14	8.0000	8.0000

У першій колонці табл. 1 і 2 наведено процентне відношення аномалій до загального числа спостережень, у колонці 2 – значення параметра p, в колонках 3–5 – знайдені невідомі параметри та значення функції для них. Як бачимо, метод найменших модулів $(p=1)$ точно відновлює лінійну функцію при будь-якому відсотку аномалій: параметр $\overline{c^*}$ рівний одиниці, а $\overline{d^*}$ рівний нулю при аномаліях ліворуч та прямує до нуля при аномаліях праворуч, що відповідає функції $y=x$. Метод найменших квадратів $(p=2)$ відхиляється від оптимальної лінійної функції у бік групи аномалій, причому при збільшенні відсотка викидів це відхилення росте. Мінімаксний метод $(p=\infty)$ апроксимує спостереження майже горизонтальною лінією, y-компонента якої залишається сталою при зростанні проценту аномалій ліворуч і змінюється, якщо аномалії розташовані праворуч. Таким чином отримані результати демонструють стійкість методу найменших модулів до появи аномалій різного характеру, яка не спостерігається при використанні МНК або мінімаксного методу.

Оскільки запропонований алгоритм дозволяє розв'язувати вихідну задачу для довільного значення параметра p, розглянемо результати його роботи для цього ж прикладу з правосторонньою групою аномалій об'ємом 15 %. Отримані результати порівняємо з результатами роботи мінімаксного методу $(p=\infty)$, в якому необхідно мінімізувати функцію максимуму

$$f_\infty(b)=\max_{i=1,n}\left|x_i b-y_i\right|. \tag{10}$$

Результати тестувань наведено в табл. 3.

ТАБЛИЦЯ 3. Апроксимація збурених праворуч спостережень (x_i,y_i) лінійною функцією за допомогою алгоритму методу еліпсоїдів при великих значеннях параметра p

p	itn	$\overline{c_p^*}$	$\overline{d_p^*}$	$f_p\left(\overline{c_p^*},\overline{d_p^*}\right)$
10	112	4.4854e – 02	6.8507	9.3126
10^2	120	4.2751e – 03	7.8780	8.1090
10^3	130	4.2764e – 04	7.9878	8.0109
10^4	141	4.2763e – 05	7.9988	8.0011
10^5	151	4.2766e – 06	7.9999	8.0001
10^6	150	4.2789e – 07	8.0000	8.0000
∞	225	6.0668e – 14	8.0000	8.0000

З табл. 3 видно, що при зростанні p значення параметрів $\overline{c_p^*}$ та $\overline{d_p^*}$ прямують до значень 6.0668e-14 та 8.0000 відповідно. Така поведінка означає, що апроксимуюча пряма все більше наближається до горизонтальної прямої, що відповідає мінімаксному методу Чебишева та випадку $p=\infty$. Отже, обчислення за допомогою формул (8) та (9) виконуються коректно.

5. Приклад 2: квадратична функція. Цей приклад розглядався в роботі [3] та має таку постановку. Наявні 28 спостережень (u_i,f_i), $u_i\in\mathbb{R}^4$, $f_i\in\mathbb{R}$, взятих з анкети опитування [10], які необхідно «найкращим чином» наблизити квадратичною функцією $Q(u)=u^TAu+b^Tu+c$, що визначається такими параметрами: симетричною 4×4-матрицею A, вектором $b\in\mathbb{R}^4$ та скаляром $c\in\mathbb{R}$. Як аргумент функція $Q(u)$ приймає вектор $u\in\mathbb{R}^4$. Цю задачу можна сформулювати як задачу мінімізації негладкої функції: знайти

$$\left(A_p^*, b_p^*, c_p^*\right) = \arg\min_{A,b,c} F_p\left(A,b,c\right), \tag{11}$$

де

$$F_p\left(A,b,c\right) = \left(\sum_{i=1}^{28} \left| f_i - Q\left(u_i\right)\right|^p\right)^{1/p} = \left(\sum_{i=1}^{28} \left| f_i - u_i^T A u_i + b^T u_i + c\right|^p\right)^{1/p}. \tag{12}$$

Отже, необхідно знайти трійку параметрів A_p^*, b_p^*, c_p^*, що визначають квадратичну функцію $Q_p^*\left(u\right) = u^T A_p^* u + \left(b_p^*\right)^T u + c_p^*$ та мінімізують функцію $F_p\left(A,b,c\right)$. Оскільки матриця A є симетричною, загальна кількість параметрів функції $Q\left(u\right)$ рівна

$$N = \frac{4\left(4+1\right)}{2} + 4 + 1 = 15,$$

серед яких 10 параметрів задають компоненти матриці A, 4 параметри – компоненти вектора b та 1 параметр задає скаляр c.

Задача (5), (6) полягає у знаходженні «найкращої» L_p-норми вектора-нев'язки ε для системи (4), яка задається матрицею X, що містить першу компоненту спостережень $\left(x_i, y_i\right)$, та вектором параметрів b. Якщо 15 невідомих параметрів функції $Q\left(u\right)$ розмістити у векторі розмірності 15, а також спеціальним чином побудувати матрицю X, задачу (11), (12) можна переформулювати в термінах задачі (5), (6): знайти

$$\beta_p^* = \arg\min_{\beta \in \mathbb{R}^{15}} \left\| X\beta - f \right\|_p. \tag{13}$$

Тут $f \in \mathbb{R}^{28}$ – друга компонента спостережень $\left(u_i, f_i\right)$. Вектор параметрів $\beta \in \mathbb{R}^{15}$ має таку структуру:

$$\beta = \begin{bmatrix} \left(a_{11},...,a_{nn}\right) \left(a_{ij}\right)_{i<j} \left(b_1,...,b_n\right) c \end{bmatrix},$$

де $a_{11},...,a_{nn}$ та a_{ij}, $i<j$ – діагональні та наддіагональні елементи матриці A, $b_1,...,b_n$ – компоненти вектора b, c – невідомий скаляр. Матриця X розмірності 28×15 – блочна матриця та має такий вигляд:

$$X = \begin{bmatrix} U^2 & W & U & e \end{bmatrix}.$$

Тут $U = \left\{u_i\right\}_{i=1}^{28}$ – матриця розмірності 28×4, що складається з векторів-рядків u_i – першої компоненти спостережень $\left(u_i, f_i\right)$; U^2 – це матриця U, кожен елемент якої піднесений до квадрату. Матриця $W = \left\{ \left(2u_i^k u_i^l \right)_{k<l} \right\}_{i=1}^{28}$ побудована таким чином: її i-й рядок є набором впорядкованих попарних подвійних добутків компонент вектора u_i, причому $k<l$, $1\le k\le 3$, $2\le l\le 4$. Нарешті e – одиничний вектор-стовпчик з $\mathbb{R}^{28}$.

Для розв'язання задачі (13) скористаємось алгоритмом методу еліпсоїдів. Вхідними параметрами є початкова точка $x_0 = \left(0,...,0\right)^T$, радіус локалізації $r=1$, точність $\varepsilon_f = 10^{-3}$ та значення параметра $p = 1, 1.3, 1.6, 2, 5$. В табл. 4 наведені відхилення $f_i - Q_p^*\left(u_i\right)$ для всіх $i = \overline{1,28}$ при різних значеннях p, а також всі 28 спостережень $\left(u_i, f_i\right)$ з анкети опитування.

ТАБЛИЦЯ 4. Спостереження (u_i, f_i) та відхилення $f_i - Q_p^*(u_i)$ для значень $p = 1, 1.3, 1.6, 2, 5$

i	u_i	f_i	$f_i - Q_p^*(u_i)$				
			$p=1$	$p=1.3$	$p=1.6$	$p=2$	$p=5$
1	(2.0, 8.5, 2.5, 4.7)	0.5	0.000	0.000	0.000	0.000	– 0.001
2	(1.5, 9.0, 2.5, 4.7)	0.5	0.000	– 0.002	– 0.007	– 0.009	– 0.013
3	(1.5, 8.5, 2.5, 5.2)	0.5	– 0.001	0.000	– 0.003	– 0.006	– 0.010
4	(1.5, 8.5, 2.5, 5.2)	0.5	– 0.001	0.000	– 0.003	– 0.006	– 0.010
5	(2.0, 8.0, 3.0, 4.7)	0.5	0.000	0.004	0.007	0.009	0.011
6	(2.0, 8.0, 2.5, 5.2)	0.5	0.000	0.008	0.010	0.010	0.011
7	(2.0, 8.5, 2.0, 5.2)	0.5	– 0.001	– 0.002	– 0.003	– 0.005	– 0.007
8	(4.0, 8.8, 5.4, 7.7)	0.9	0.000	– 0.005	– 0.007	– 0.010	– 0.014
9	(3.1, 9.7, 5.4, 7.7)	0.9	0.000	– 0.001	– 0.003	– 0.004	– 0.001
10	(3.1, 8.8, 6.3, 7.7)	0.9	0.000	0.000	0.000	0.000	– 0.006
11	(3.1, 8.8, 5.4, 8.6)	0.9	0.000	0.004	0.004	0.004	– 0.001
12	(4.0, 7.9, 6.3, 7.7)	0.9	0.000	0.000	– 0.002	– 0.004	– 0.011
13	(4.0, 7.9, 5.4, 8.6)	0.9	0.004	0.022	0.020	0.019	0.015
14	(4.0, 8.8, 4.5, 8.6)	0.9	0.000	– 0.002	– 0.005	– 0.006	– 0.007
15	(5.9, 8.7, 4.7, 5.4)	1.0	0.000	– 0.001	– 0.003	– 0.004	– 0.010
16	(4.9, 9.7, 4.7, 5.4)	1.0	0.000	0.010	0.011	0.011	0.013
17	(4.9, 8.7, 5.7, 5.4)	1.0	0.000	0.000	0.011	0.003	0.004
18	(4.9, 8.7, 4.7, 6.4)	1.0	0.000	0.001	0.001	0.001	– 0.002
19	(5.9, 7.7, 5.7, 5.4)	1.0	0.000	– 0.007	– 0.010	– 0.010	– 0.010
20	(5.9, 7.7, 4.7, 6.4)	1.0	0.005	0.017	0.013	0.012	0.012
21	(5.9, 8.7, 3.7, 6.4)	1.0	0.000	– 0.002	– 0.005	– 0.007	– 0.012
22	(2.0, 9.0, 2.2, 7.0)	0.9	0.003	0.010	0.013	0.014	0.015
23	(1.1, 9.9, 2.2, 7.0)	0.9	0.000	– 0.007	– 0.011	– 0.014	– 0.014
24	(1.1, 9.0, 3.1, 7.0)	0.9	0.005	0.021	0.023	0.022	0.017
25	(1.1, 9.0, 2.2, 7.9)	0.9	0.002	0.012	0.013	0.012	0.010
26	(2.0, 8.1, 3.1, 7.9)	0.9	– 0.074	– 0.030	– 0.022	– 0.018	– 0.015
27	(2.0, 8.1, 3.1, 7.9)	0.9	– 0.074	– 0.030	– 0.022	– 0.018	– 0.015
28	(2.0, 9.0, 1.3, 7.9)	0.9	0.000	0.000	0.001	0.004	0.010

Аналізуючи табл. 4, можна зробити декілька висновків.

По-перше, якщо при $p=1$ не враховувати спостереження з номерами 26 і 27, то максимальне відхилення серед інших 26-ти значень становить всього 0.005, тоді як при $p=2$ воно становить 0.022, причому наявні також інші доволі вагомі відхилення: -0.018 (номери 26 і 27), 0.019 (номер 13), 0.014 (номери 22–23). Загалом серед всіх відхилень при $p=1$ є лише декілька відмінних від нуля (а саме 10), тоді як при $p=2$ таких відхилень 26. Такі висновки вказують на те, що для даного прикладу МНМ $(p=1)$ демонструє суттєво кращі результати, ніж МНК $(p=2)$, відкидаючи аномальні спостереження, тоді як МНК враховує їх, зміщуючи розв'язок.

По-друге, по мірі того як значення параметра p прямує від 1 до 2, відхилення зростають. Однак при $p=5$ вони або несуттєво зменшуються, або залишаються незмінними. Очевидно, залежність значення параметра p від відхилень є нелінійною, тому необхідно провести додаткові дослідження цієї залежності для коректного підбору параметра p.

Однак остаточний вибір значення параметра p визначається конкретним набором спостережень. Наприклад, наявність викидів під номерами 26 та 27 може бути спричинена похибкою зчитування даних або помилкою експерта, який визначив значення f_i; в такому випадку повторне зчитування та перегляд значень експертом можуть покращити ситуацію. З іншого боку, якщо випадкова величина, яка характеризує похибку результатів спостережень, має нормальний розподіл ймовірностей, доцільним буде використання МНК. Якщо ж вона має розподіл Пуассона, краще використовувати МНМ. Можливість використання параметра p дозволяє спростити розв'язання задач, описаних в цьому підрозділі, адже замість трьох методів (МНМ, МНК та мінімаксного методу) можна використовувати лише один, а також гнучко підбирати значення параметра p для отримання «найкращих» розв'язків.

Висновки. Для розв'язання задачі визначення параметрів лінійної регресії можна використовувати як класичні методи – МНК та МНМ при нормальному та пуасонівському розподілах випадкової величини похибки спостережень відповідно, так і методи мінімізації негладких функцій, якщо змінити постановку задачі. В такому випадку за допомогою значення параметра p можна плавно регулювати розв'язок для відкидання аномальних спостережень або їх урахування у випадку надійності. Модифіковані формули для обчислення значення цільової функції та її субградієнту дозволяють використовувати довільні значення параметра $p \geq 1$, у тому числі доволі великі. У випадку апроксимації спостережень квадратичною функцією схема зведення дозволяє розв'язувати цю задачу як задачу визначення параметрів лінійної регресії.

Список літератури

1. Демиденко Е.З. Линейная и нелинейная регрессии. М.: Финансы и статистика, 1981. 304 с.
2. Shor N.Z., Stetsyuk P.I. Constructing Utility Functions by Methods of Nondifferentiable Optimization. In: *Constructing and Appling Objective Functions, Lecture Notes in Economics and Mathematical Systems*. V. 510. Berlin: Springer-Verlag, 2002. P. 215–232. https://doi.org/10.1007/978-3-642-56038-5_10
3. Стецюк П.И., Колесник Ю.С., Лейбович М.М. О робастности метода наименьших модулей. *Компьютерная математика*. 2002. С. 114–123.
4. Стецюк П.И., Колесник Ю.С. К вопросу выбора метода аппроксимации результатов измерения. *Интеллектуальные информационно-аналитические системы и комплексы*. 2000. С. 62–67.
5. Стецюк П.И., Стовба В.А., Мартынюк И.С. Алгоритм метода эллипсоидов для нахождения *Lp*-решения системы линейных уравнений. *Теорія оптимальних рішень*. 2017. С. 139–146. http://dspace.nbuv.gov.ua/handle/123456789/131449
6. Шор Н.З. Метод отсечения с растяжением пространства для решения задач выпуклого программирования. *Кибернетика*. 1977. № 1. С. 94–95.
7. Стецюк П.И., Стовба В.А., Жмуд А.А. Метод эллипсоидов для нахождения решения переопределенной СЛАУ. *Теорія оптимальних рішень*. 2018. № 17. С. 115–123. http://dspace.nbuv.gov.ua/handle/123456789/144980
8. Стецюк П.И., Била Г.Д., Стовба В.А. Метод эллипсоидов для нахождения *Lp*-решения системы линейных уравнений. *Інформатика та системні науки* (*ІСН-2017*): матеріали VIII Всеукраїнської науково-практичної конференції за міжнародною участю. Полтава, 16–18 березня 2017 р.
9. Кларк Ф. Оптимизация и негладкий анализ. М.: Наука, 1988. 280 с.
10. Gruber J. Opening Remarks: A Retrospection over 35 Years of Work. *Constructing and Appling Objective Functions, Lecture Notes in Economics and Mathematical Systems*. V. 510. Berlin: Springer-Verlag, 2002. P. 3–13. https://doi.org/10.1007/978-3-642-56038-5_1

Одержано 11.10.2020

Стовба Віктор Олександрович,
аспірант Інституту кібернетики імені В.М. Глушкова НАН України, Київ.
vik.stovba@gmail.com

УДК 519.85

В.А. Стовба

Метод эллипсоидов для нахождения параметров линейной регрессии

Институт кибернетики имени В.М. Глушкова НАН Украины, Киев
Переписка: vik.stovba@gmail.com

Введение. Задача определения параметров линейной регрессии может быть сформулирована как задача минимизации негладкой функции, которая представлена как L_p -норма вектора-невязки системы линейных уравнений. Для ее решения можно использовать методы минимизации негладких функций, например, субградиентные методы. В работе [7] рассмотрено приложение метода эллипсоидов для нахождения L_p -решения переопределенной системы линейных уравнений при $1 \le p \le 2$.

Цель работы. Расширить алгоритм на базе метода эллипсоидов для решения задачи определения параметров линейной регрессии для произвольных значений параметра $p \ge 2$ так, чтобы при больших значениях p решение задачи совпадало с решением, полученным минимаксным методом, который соответствует значению $p = \infty$. Описать формулировку задачи аппроксимации наблюдений квадратичной функцией как задачи определения параметров линейной регрессии. Проанализировать результаты работы алгоритма для большого количества наблюдений и выбросов. Сравнить результаты работы минимаксного метода и метода эллипсоидов для задачи определения параметров линейной регрессии при больших значениях параметра p.

Результаты. Разработан способ вычисления значения целевой функции и ее субградиента при больших значениях параметра p, проверенного на примере аппроксимации линейной функцией наблюдений, содержащих аномалии. Алгоритм на основе метода эллипсоидов демонстрирует монотонность при изменении параметров линейной регрессии с помощью регулировании параметра p, позволяя, таким образом, отбрасывать или учитывать те или иные наблюдения. В работе [3] показано, что преимущество следует отдавать методу наименьших модулей (МНМ), поскольку он игнорирует аномальные наблюдения и точно восстанавливает линейную функцию. Результаты экспериментов с большим количеством наблюдений и выбросов при $p = 1$ подтвердили этот вывод: МНМ игнорирует группы аномалий и адекватно аппроксимирует наблюдения линейной функцией. Метод наименьших квадратов отклоняется от оптимальной линейной функции, если имеется группа аномалий в той или иной области. При использовании больших значений параметра p решение задачи приближается к решению, получаемому минимаксным методом.

Выводы. Алгоритм на основе метода эллипсоидов позволяет находить параметры линейной регрессии при произвольном значении параметра $p \ge 1$. Это дает возможность использовать три известных метода: метод наименьших модулей, метод наименьших квадратов и минимаксный метод – как его частные случаи. При этом, устремляя p к единице, можно регулировать интенсивность игнорирования аномальных наблюдений, что дает возможность использовать информацию из внешних источников (экспертные оценки, показания измерительных приборов, статистические прогнозы и т. п.) для более адекватного восстановления аппроксимирующей функции.

Ключевые слова: метод эллипсоидов, линейная регрессия, аномальные наблюдения.

UDC 519.85

V. Stovba

Ellipsoid Method for Linear Regression Parameters Determination

V.M. Glushkov Institute of Cybernetics of the NAS of Ukraine, Kyiv
Correspondence: vik.stovba@gmail.com

Introduction. Linear regression parameters determination can be formulated as a non-smooth function minimization problem, which is L_p -norm of residual of the linear equations system. To solve it non-smooth function minimization methods can be used, e.g. subgradient methods. The article [7] considers ellipsoid method application for finding L_p -solution of redefined linear equations system with $1 \le p \le 2$.

The purpose of the paper is to extend the algorithm based on the ellipsoid method for a linear regression parameters determination problem with an arbitrary value of parameter $p \geq 2$ so that under big values of p the solution of the problem equals minimax method solution, which corresponds to $p = \infty$ case. To describe the formulation of observation approximation problem with quadratic function as linear regression parameters determination problem. To analyze algorithm work results for great number of observations and outliers. To compare the minimax method and the ellipsoid method algorithm work results for linear regression parameters determination problem with big values of parameter p.

Results. The way of calculation of objective function and its subgradient values with large values of parameter p was developed and verified on example of observation approximation containing outliers with linear function. Algorithm based on ellipsoid method changes linear function parameters monotonically using parameter p adjusting, thereby permits to reject or consider these or those observations. It is shown in [3] that Least Absolute Deviations method (LAD) is advised to be used as far as it ignores outliers and reconstructs linear function accurately. Experiment results with big number of observations and outliers using $p = 1$ confirmed that conclusion: LAD ignores outlier groups and approximates observations with linear function adequately. Least Square Method (LSM) deviates from optimal linear function if a group of outliers is present in particular area. In case of using big values of parameter p problem solution converges to minimax method solution.

Conclusions. Algorithm based on ellipsoid method permits to determine linear regression parameters with arbitrary value of parameter $p \geq 1$. So, three known methods can be used – LAD, LSM and minimax method – as its special cases. Moreover, directing p to 1, intensity of outliers ignoring can be regulated, that gives a possibility to use external sources of information (expert opinions, measuring devices readings, statistical forecasts, etc.) for more correct and adequate approximation function reconstruction.

Keywords: ellipsoid method, linear regression, outliers.