Міністерство освіти і науки України

ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ СІЛЬСЬКОГО ГОСПОДАРСТВА ІМЕНІ ПЕТРА ВАСИЛЕНКА

Навчально-науковий інститут енергетики і комп'ютерних технологій

Кафедра біомедичної інженерії і теоретичної електротехніки

МЕТОД НАЙМЕНШИХ КВАДРАТІВ

Навчально-методичний посібник

Харків
2020

Міністерство освіти і науки України

ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
СІЛЬСЬКОГО ГОСПОДАРСТВА ІМЕНІ ПЕТРА ВАСИЛЕНКА

Навчально-науковий інститут енергетики і комп'ютерних технологій

Кафедра біомедичної інженерії і теоретичної електротехніки

МЕТОД НАЙМЕНШИХ КВАДРАТІВ

Навчально-методичний посібник

для студентів першого (бакалаврського) рівня вищої освіти денної та заочної форми навчання для інженерних спеціальностей

Затверджено
Рішенням Науково-методичної
ради ННІ ЕКТ ХНТУСГ
Протокол № 5
від 31 січня 2020 р.

Харків
2020

УДК 512.64(076.2)
Л-59

Схвалено
на засіданні кафедри біомедичної інженерії і теоретичної електротехніки
Протокол №6 від 27 січня 2020 р.

Метод найменших квадратів [Текст]: навч.-метод. посіб. для студентів першого (бакалаврського) рівня вищої освіти денної та заочної форми навчання інж спец. / Н. Г. Косуліна [та ін.]: Харків. нац. техн. ун-т сіл. госп-ва ім. П. Василенка. – Харків: ХНТУСГ, 2020. – 25 с.

Мета навчально-методичного посібника – вивчення методу найменших квадратів. Для набуття практичних навиків та закріплення теоретичних знань представлено значну кількість прикладів.

Видання призначене студентам першого (бакалаврського) рівня вищої освіти денної або заочної форми навчання інженерних спеціальностей.

Рецензенти:

Є. Л. Піротті, д-р фіз.-мат. наук, проф. кафедри комп'ютерної математики та математичного моделювання НТУ «ХПІ»;
М. П. Кунденко, д-р техн. наук, проф., зав. кафедри ІЕТП Харківського національного технічного університету сільського господарства імені Петра Василенка

Відповідальний за випуск (зав. каф.): **Н. Г. Косуліна,** д-р техн. наук, проф.

МЕТОД НАЙМЕНШИХ КВАДРАТІВ

1. Загальні поняття про метод найменших квадратів

Задача МНК розв'язується шляхом параметричної оцінки функції регресії, що описує залежність однієї величини Y, значення якої (y_i) спостерігають з випадковими похибками (θ_i), від групи невипадкових величин X_1, X_2, …, X.

Результати вимірювань деякої фізичної величини проводяться при незмінному її стані для усієї серії вимірювань. Але можливо, що сама вимірювана величина за час вимірювань змінюється внаслідок змінювання іншої величини, яка зв'язана з нею. В таких випадках буде також спостерігатися статистичне розсіяння, яке приводить до випадкових похибок.

Нехай в результаті вимірювань ми одержали ряд експериментальних точок з абсцисами $x_1, x_2, \dots, x_n$ і відповідними їм ординатами $y_1, y_2, \dots, y_n$.

Залежність y від x, яка зображується аналітичною залежністю $y=f(x)$, не може співпадати з експериментальними значеннями y_i всіх n точок. Одержимо ламану лінію, яка нічого загального не буде мати з шуканою аналітичною залежністю $y = f(x)$ тому, що форма ламаної не буде відтворюватись при повторних серіях вимірювань. Вимірювані значення y будуть в загальному випадку зміщені відносно шуканої кривої як в сторону більших, так і в сторону менших значень, внаслідок статистичного розсіяння.

Задача полягає в тому, щоб по даним експериментальним точкам провести криву, яка як можна ближче підходить до дійсної функціональної залежності $y = f(x)$. Це означає, що для всіх або деяких точок різниця

$$\Delta_i = y_i - f(x_i) \qquad (1.1)$$

буде відміна від нуля.

Параметри функції $y = f(x)$ підбирають таким чином, щоб сума квадратів різниць (1.1) була найменшою, тобто необхідно обернути в мінімум вираз:

$$z = \sum_{i=1}^{n} \Delta_i^2 = \sum_{i=1}^{n} [y_i - f(x_i)]^2 \quad . \tag{1.2}$$

Таким чином, при методі найменших квадратів наближення аналітичної функції $y = f(x)$ до експериментальної залежності вважається найкращим, якщо виконується умова мінімуму суми квадратів відхилень аналітичної функції від експериментальної залежності. Застосуванню методу найменших квадратів передують способи математичної статистики і теорії похибок.

Лінійна форма зв'язку між випадковими змінними займає особливе місце в теорії кореляції. Для такої форми зв'язку є лінійна функція y від x, тобто

$$y = a + bx \, , \tag{1.3}$$

де a і b – коефіцієнти рівняння регресії;

x – незалежна випадкова змінна.

При $x = 0$; $y = a$ – початок відліку; b – коефіцієнт регресії, який показує середнє зміщення залежної змінної при змінені незалежної змінної на одиницю. Коефіцієнт регресії завжди число іменоване. Якщо $b > 0$, то зв'язок прямий, якщо $b < 0$, то зв'язок обернений; якщо b=0, зв'язок відсутній. Лінійна залежність обумовлена двомірним нормальним законом розподілу пар випадкових величин (X, Y).

Запишемо тепер вираз (1.2) з урахуванням (1.3) у вигляді

$$z(a, b) = \sum_{i=1}^{n} (y_i - f - bx)^2 \, . \tag{1.4}$$

Відомо, що мінімум функції можна знайти, якщо порівняти з нулем першу похідну.

Скорочуючи обидва вирази на -2 і виконуючи почленне підсумовування, одержимо систему нормальних рівнянь

$$\begin{cases} \sum y = an + b\sum x, \\ \sum yx = a\sum x + b\sum x^2. \end{cases} \tag{1.5}$$

Розв'язуючи систему (1.5), визначаємо параметри a і b. Параметри a і b можна визначити і за іншими робочими формулами, наприклад:

$$b = \frac{n\sum xy - \sum y \sum x}{n\sum x^2 - (\sum x)^2}; \quad a = \frac{\sum y \sum x^2 - \sum yx \sum x}{n\sum x^2 - (\sum x)^2}; \tag{1.6}$$

$$b = \frac{n\overline{yx} - \overline{x} \cdot \overline{y}}{n\overline{x}^2 - \left(\overline{x}\right)^2}, \qquad a = \overline{y} - b\overline{x}. \tag{1.7}$$

Розглянемо приклади щодо застосування метода найменших квадратів.

Приклад 1. При вимірюванні електричного опору R проволоки в залежності від температури t°C були одержані такі результати, які наведені в таблиці 1.

Таблиця 1 – Результати вимірювання електричного опору в залежності від температури

t,^{0}C	20.2	24.3	28.5	32.2	36.5	40.1	44.2	47.9
R,Ом	86.3	87.1	88.7	89.7	91.8	93.2	94.9	96.4

Необхідно знайти аналітичну залежність між змінними R і t^0.

Розв'язання.

Якщо на графік (рис. 1) нанести точки з таблиці 1, то можна висловити гіпотезу, що між змінними R і t^0 існує лінійна залежність $R=a+bt$. Невідомі параметри a і b знайдемо методом найменших квадратів.

Результати розрахунків, які допоможуть розв'язати систему (1.5), зводимо в таблицю 2.

Розв'яжемо систему (1.5), яка для наших змінних має вигляд:

$$\begin{cases} \sum R = an + b\sum t, \\ \sum Rt = a\sum t + b\sum t^2. \end{cases}$$

Підставимо в систему розрахункові значення з таблиці 2, одержимо

$$\begin{cases} 728{,}1 = 8a + 273{,}9b, \\ 25176{,}24 = 273{,}9a + 10035{,}93b. \end{cases}$$

Розв'язуючи систему, одержимо b=0,379; a=78,033.

Таким чином, рівняння регресії $R=a+bt$ з одержаними числовими значеннями параметрів має вигляд

$$\bar{R}_t = 78{,}033 + 0{,}379t.$$

Таблиця 2 – Результати розрахунків

№	Температура t^0	Опір R, Ом	Розрахункові величини			
			$R \cdot t^0$	R^2	t^2	Очікуване (розрахункове) значення опору, Ом
1	2	3	4	5	6	7
1	20,2	86,3	1743,26	7447,69	408,04	85,69
2	24,3	87,1	2116,53	7586,41	590,49	87,24
3	28,5	88,7	2527,95	7867,69	812,25	88,83
4	32,2	89,7	2888,34	8046,09	1036,84	90,24
5	36,5	91,8	3350,7	8427,24	1332,25	91,87
6	40,1	93,2	3737,32	8686,24	1608,01	93,23
7	44,2	94,9	4194,58	9006,01	1953,64	94,78
8	47,9	96,4	4617,56	9292,96	2294,41	96,19
Разом	273,9	728,1	25176,24	66360,33	10035,93	728,1

В середньому	34,24	91,01	3147,03	8285,04	1254,49	91,01

Підставляючи в дане рівняння значення температури з таблиці 2, одержимо розрахункові значення опору проводу від температури, які вносимо в 7 колонку таблиці 2. За цими значеннями будуємо теоретичну лінію регресії (рис. 1).

Правильність усіх розрахунків перевіряємо, порівнюючи суми фактичного та теоретичного значень електричного опору проводу в залежності від температури.

При криволінійній залежності система рівнянь складається так, як і для лінійної залежності. Наприклад, експериментальні точки розміщені на площині так, що можна висловити гіпотезу про параболічну залежність між змінними y і x.

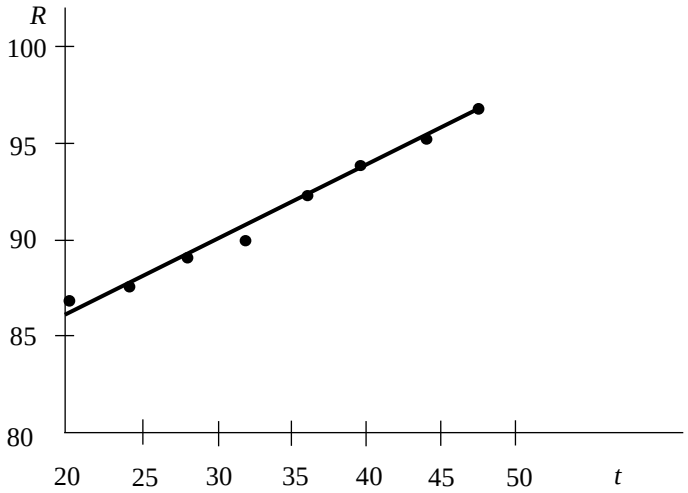

Рис. 1. Теоретична лінія регресії

Рівняння параболи має вигляд $y_x = a + bx + cx^2$. За допомогою метода найменших квадратів складаємо систему рівнянь для пошуку параметрів a, b, c:

$$\begin{cases} \sum y = an + b\sum x + c\sum x^2; \\ \sum yx = a\sum x + b\sum x^2 + c\sum x^3; \\ \sum yx^2 = a\sum x^2 + b\sum x^3 + c\sum x^4. \end{cases} \quad (1.8)$$

Якщо між змінними існує степенева, показникова або гіперболічна залежності, то їх можна привести за допомогою логарифмування функцій до лінійної залежності. Результати згладжування функцій наведені в таблиці 3.

Таблиця 3 – Результати згладжування функцій

Згладжувана функція	Приведена функція	Заміна змінних
$y = ax^b$	$n = \alpha\varepsilon + \beta$	$n = lg\gamma; \varepsilon = lgx;$ $\alpha = b; \beta = lga$
$y = ab^x$	$n = \alpha\varepsilon + \beta$	$n = lgy; \varepsilon = x$ $\alpha = lgb; \beta = lga$
$y = a + \frac{b}{x}$	$n = \alpha\varepsilon + \beta$	$n = y; \varepsilon = \frac{1}{x};$ $\alpha = b; \beta = a$
$y = a + \frac{b}{x}$	$n = \alpha\varepsilon + \beta$	$n = xy; \varepsilon = x$ $\alpha = a; \beta = a$
$y = \frac{1}{ax + b}$	$n = \alpha\varepsilon + \beta$	$n = \frac{1}{y}; \varepsilon = x$ $\alpha = a; \beta = b$
$y = \frac{x}{ax + b}$	$n = \alpha\varepsilon + \beta$	$n = \frac{x}{y}; \varepsilon = x$ $\alpha = a; \beta = b$

2. Система двох випадкових величин

В практичних задачах вимірювання приходиться розглядати не тільки одну, а дві і більше випадкових величин, які зв'язані з деяким дослідом.

Системою двох випадкових величин (X, Y) називається сукупність двох випадкових величин, які розглядаються сумісно. Кожну з величин X і Y називають складовими або компонентними.

Випадкові величини (X, Y) *називаються незалежними*, якщо закон розподілу кожної з них не залежить від того, яке значення прийняла друга випадкова величина. В іншому випадку *випадкові величини* (X, Y) *називаються залежними.*

Для незалежних неперервних випадкових величин щільність розподілу системи приймає вигляд

$$f(x, y) = F_1(x)F_2(y), \tag{2.1}$$

тобто щільність розподілу системи $f(x, y)$ дорівнює добутку щільностей розподілу окремих випадкових величин.

Аналогічно по відношенню до функції розподілу незалежних випадкових величин X і Y справедливо

$$F(x, y) = F_1(x)F_2(y). \tag{2.2}$$

Якщо умови (2.1) або (2.2) не виконуються, то випадкові величини залежні.

З більшості законів розподілу системи двох випадкових величин найбільше розповсюдження на практиці має нормальний закон розподілу.

Так, неперервна система випадкових величин (X, Y) розподілена за нормальним законом, якщо сумісна щільність має вигляд:

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-r_{xy}^2}} exp\left\{-\frac{1}{2\sqrt{1-r_{xy}^2}}\left[\frac{(x-m_x)^2}{\sigma_x^2} - \frac{2r_{xy}(x-m_x)(y-m_y)}{\sigma_x\sigma_y} + \frac{(y-m_y)^2}{\sigma_y^2}\right]\right\}, \tag{2.3}$$

де $-\exp(z) = e^z$ - показникова функція.

Параметри m_x, m_y, σ_x, σ_y, r_{xy}, які входять в формулу (2.3), є числовими характеристиками двовимірного нормального закону.

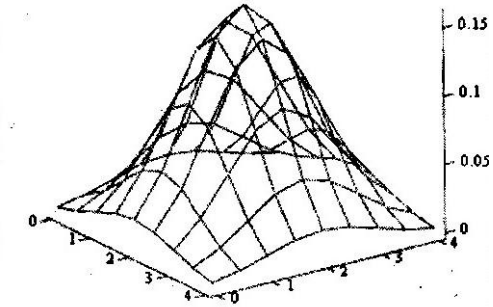

Рис. 2. Графік двомірного закону нормального розподілу

Кореляційний момент. Коефіцієнт кореляції

До числових характеристик однієї випадкової величини X відносяться: математичне сподівання $M[X]$, дисперсія $D[X]$, середнє квадратичне відхилення σ[X]. Аналогічні числові характеристики можна ввести і для системи випадкових величин.

Математичні сподівання для двох випадкових величин (X, Y), які входять в систему для неперервних випадкових величин, визначаються за формулами:

$$M[X] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xf(x,y)dxdy; \quad (2.3)$$

$$M[Y] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} yf(x,y)dxdy. \quad (2.4)$$

Ці два числа $M[X]$ *і* $M[Y]$ є координатами точки, яка називається центром розсіяння системи випадкових величин (X, Y).

Дисперсії випадкових величин $X\ i\ Y$, які належать системі (X, Y) , для неперервних випадкових величин визначають за формулами:

$$D[X] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x - m_x)^2 f(x, y)dxdy; \quad (2.5)$$

$$D[Y] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(y - m_y)^2 f(x, y)dxdy. \quad (2.6)$$

Середні квадратичні відхилення випадкових величин $X\ i\ Y$ шукають за формулами:

$$\sigma_x = \sqrt{D[X]}, \sigma_y = \sqrt{D[Y]}. \quad (2.7)$$

Впровадимо нову числову характеристику для системи випадкових величин – коваріацію K_{xy} (кореляційний момент). Коваріація характеризує ступінь зв'язку між випадковими величинами $X\ i\ Y$.

Коваріацією K_{xy} двох залежних випадкових величин називається математичне сподівання добутку відхилень величин $X\ i\ Y$ від їх математичних сподівань

$$K_{xy} = M\big[(X - m_x)(Y - m_y)\big]. \quad (2.8)$$

За означенням, коваріація не змінюється при заміні місцями індексів:

$$K_{xy} = K_{yx}\ .$$

Коваріація для неперервних випадкових величин обчислюється за формулою:

$$K_{xy} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x - m_x)\big(y - m_y\big)f(x, y)dxdy\ \ . \quad (2.9)$$

Існує інша формула, яка більш зручніша для обчислення коваріації. Коваріація двох випадкових величин дорівнює математичному сподіванню їх добутку, мінус добуток математичних сподівань цих величин

$$K_{xy} = M[X \cdot Y] - M[X] \cdot M[Y], \quad (2.10)$$

де математичне сподівання добутку випадкових величин знаходиться за формулою:

$$M[X \cdot Y] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf(x,y)dxdy. \qquad (2.11)$$

Для незалежних випадкових величин коваріація дорівнює нулю. Покажемо це на прикладі. Користуючись формулою (2.9) і вираховуючи (2.1), одержимо:

$$K_{xy} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - m_x)\big(y - m_y\big) f_1(x) f_2(y) dxdy \, .$$

Даний інтеграл можна переписати у вигляді добутку двох інтегралів :

$$K_{xy} = \int_{-\infty}^{\infty} (x - m_x) f_1(x) dx \int_{-\infty}^{\infty} (y - m_y) f_2(y) \, dy = 0. \qquad (2.12)$$

В формулі (2.12) інтеграли дорівнюють нулю тому, що математичне сподівання відхилення випадкової величини від її математичного сподівання дорівнює нулю.

Розмірність коваріації дорівнює добутку розмірностей випадкових величин. Для характеристики зв'язку між випадковими величинами зручніше працювати з безрозмірною величиною r_{xy}, яку *називають коефіцієнтом кореляції*

$$r_{xy} = \frac{K_{xy}}{\sigma_x \sigma_y}, \qquad (2.13)$$

де σ_x, σ_y – середні квадратичні відхилення випадкових величин X і Y .

Коефіцієнтом кореляції r_{xy} *двох випадкових величин* X і Y *називають* відношення кореляційного моменту до добутку середніх квадратичних відхилень цих величин.

Якщо випадкові величини X і Y незалежні між собою, тоді $r_{xy} = 0 \, \big(K_{xy} = 0\big)$. Обернене твердження невірне, тобто коефіцієнт кореляції, може дорівнювати нулю $\big(r_{xy} = 0\big)$, але величини X і Y

можуть і не бути незалежними. Випадкові величини, для яких $(r_{xy} = 0)$, називаються *некорельованими* .

Розглянемо випадок, коли випадкові величини $X\ i\ Y$ зв'язані лінійною залежністю

$$Y = bX + a, \tag{2.14}$$

де a і b – невипадкові величини.

Знайдемо коефіцієнт кореляції r_{xy} випадкових величин $X\ i\ Y$. За означенням

$$r_{xy} = \frac{K_{xy}}{\sigma_x \sigma_y}.$$

Коваріацію шукаємо за формулою $K_{xy} = M\big[(X - m_x)(Y - m_y)\big]$. З врахуванням (2.14), одержимо:

$$K_{xy} = M[(X - m_x)(bX + a - M)(bX + a)] = bM[(X - m_x)(bX + a - bm_x - a)] = M[(X - m_x)(bX - bm_x)] = bM\big[(X - m_y)^2\big] = bD_x. \tag{2.15}$$

Дисперсія випадкової величини Y дорівнює:

$D_y = D[bX + a] = D[bX] = b^2 D_x$. Відповідно $\sigma_y = |b|\sigma_x$.

Підставимо одержані значення в формулу для коефіцієнта кореляції r_{xy} :

$$r_{xy} = \frac{bD_x}{\sigma_x \sigma_y} = \frac{b\sigma_x^2}{\sigma_x |b| \sigma_x} = \frac{b}{|b|},$$

тобто

$$r_{xy} = \begin{cases} -1 \text{ коли } b < 0; \\ 0 \text{ коли } b = 0; \\ 1 \text{ коли } b > 0. \end{cases} \tag{2.16}$$

В загальному випадку значення коефіцієнта кореляції r_{xy} задовольняє умові

$$-1 \le r_{xy} \le 1.$$

Якщо коефіцієнт кореляції додатний, то зв'язок між змінними додатний. Це значить, що при зростанні x збільшується y. Якщо коефіцієнт кореляції від'ємний, то зв'язок між змінними від'ємний. Це значить, що при зростанні x зменшується y.

Коефіцієнт кореляції не залежить від вибору початку відліку і одиниці вимірювання. Таким чином, змінні $x\ i\ y$ можна зменшувати або збільшувати в k разів, а також додавати до змінних $x\ i\ y$ або віднімати від них одне і те ж число b. В результаті величина коефіцієнта корекції не зміниться.

Якщо коефіцієнт кореляції дорівнює одиниці, то зв'язок між змінними функціональний. Випадкові величини, для яких $r_{xy} = 0$, називаються некорельованими. Некорельованість не слід змішувати з незалежністю тому, що незалежні випадкові величини завжди корельовані. Отже обернене стверджування невірне тому, що некорельовані величини можуть бути залежними ще і функціонально.

Таким чином, коефіцієнт кореляції характеризує ступінь наближення залежності зв'язку між випадковими величинами до лінійної функціональної залежності зв'язку між випадковими величинами. Значення коефіцієнта кореляції визначає наскільки залежність між випадковими змінними близька до лінійної функціональної. Якщо $|r_{xy}| = 0{,}8 \div 0{,}9$, то можна впевнено вважати, що незалежно від форми зв'язку між змінними $x\ i\ y$, вона достатньо щільна для того, щоб дослідити її форму.

Приклад. Визначити щільність зв'язку між ознаками, що досліджуються - температурою t°C та електричним опором проводу R, Ом. Необхідні дані для розрахунку коефіцієнта кореляції наведені в таблицях 1 і 2.

Розв'язання. Для лінійної залежності між двома випадковими величинами коефіцієнт кореляції визначається за допомогою формули:

$$r = \frac{xy - x \cdot y}{\sigma_x \sigma_y}, \tag{2.17}$$

$xy = \frac{\sum xy}{n}$; $x = \frac{\sum x}{n}$; $y = \frac{\sum y}{n}$; $\sigma_x = \sqrt{\frac{\sum x^2}{n} - x^2}$; $\sigma_y = \sqrt{\frac{\sum y^2}{n} - y^2}$.

Для розглянутого прикладу y змінюємо на R, а x на t .

$Rt = \frac{\sum Rt}{n} = \frac{25176,1}{8} = 3147,03$ $t = \frac{\sum t}{n} = 34,24$; К $= \frac{\sum R}{n} = \frac{728,1}{8} =$ 91,01;

$$\sigma_t = \sqrt{\frac{\sum t^2}{n} - (t)^2} = \sqrt{\frac{10035,93}{8} - (34,24)^2 = \sqrt{82,02}} = 9,06\ ;$$

$$\sigma_R = \sqrt{\frac{\sum R^2}{n} - (R)^2} = \sqrt{\frac{66360,33}{8} - (91,01)^2 = \sqrt{12,22}} = 3,496;$$

$$r_{Rt} = \frac{R_t - R \cdot t}{\sigma_t \sigma_R} = \frac{3147,03 - 91,01 \cdot 34,24}{9,06 \cdot 3,496} = \frac{30,848}{31,674} = 0,974.$$

Коефіцієнт кореляції показує, що між електричним опором і температурою існує щільний зв'язок.

Для оцінки значущості фактору, який вивчається, щодо загальної варіації признаку визначають коефіцієнт детермінації, який дорівнює квадрату коефіцієнта кореляції, тобто r_{xy}^2. Якщо коефіцієнт детермінації виражаємо в відсотках, тоді він показує, що варіація залежної змінної (результативного признаку) на стільки-то відсотків обумовлена варіацією фактору

$$d = r_{xy}^2 \cdot 100\%. \qquad (2.18)$$

В розглянутому прикладі 2 коефіцієнт детермінації дорівнює $r_{Rt}^2 = (0,974)^2 \cdot 100\% = 94,87\%$, тобто 94,87% загальної варіації електричного опору проводу обумовлено температурою, а остання частина (5,13%) іншими факторами, які в даній задачі не враховувались.

При дослідженні кореляційної залежності між випадковими величинами експериментальним шляхом одержують сукупність пар значень випадкових величин X і Y. Цю сукупність пар чисел можна розглядати як випадкову вибірку з генеральної сукупності усіх можливих значень двомірної випадкової величини $(X\ i\ Y)$. Тому рівняння ліній регресії, одержаних на основі експериментальних даних, називають вибірковими.

При великому числі дослідів одна і та ж пара чисел (x, y) буде спостерігатись декілька раз. В такому випадку результати спостережень зручніше записувати у вигляді таблиці, в якій вказується, скільки разів спостерігалась кожна пара чисел. Така таблиця називається кореляційною таблицею (табл. 4).

Таблиця 4 – Кореляційна таблиця

	x_1	x_2	x_3	…	x_k	n_x
y_1	n_{11}	n_{21}	n_{31}	…	n_{k1}	n_1
y_2	n_{12}	n_{22}	n_{32}	…	n_{k2}	n_2
y_3	n_{13}	n_{23}	n_{33}	…	n_{k3}	n_3
…	…	…	…	…	…	…
y_i	n_{1i}	n_{2i}	n_{3i}	…	n_{ki}	n_i
n_y	m_1	m_2	m_3	…	m_k	n

В кореляційній таблиці часто вказують не спостережені в експерименті значення випадкової величини, а інтервали, в які ці значення попадають. Можна замість інтегралів вказувати середини інтегралів.

Рівняння парної лінії регресії x на y мають вигляд відповідно

$$y - \overline{y}_b = r_b \frac{\sigma_y}{\sigma_x}\left(y - \overline{y}_b\right) \text{ і } x - \overline{x}_b = r_b \frac{\sigma_x}{\sigma_y}\left(x - \overline{x}_b\right).$$

Коефіцієнт кореляції обчислюють за формулою

$$r_b = \frac{\overline{xy}_b - \overline{x}_b \cdot \overline{y}_b}{\sigma_x \sigma_y},$$

де r_b – вибіркове значення коефіцієнта кореляції.

Існують інші формули для обчисленні вибіркового коефіцієнта кореляції, які приводяться в підручниках по кореляційному аналізу.

3. Приклади використання методу найменших квадратів

Приклад. Данні про обсяг випуску продукції (Y) та коштам основних промислово-виробничих фондів (X) по 60 підприємствам згруповані і навелені в таблиці 5.

Таблиця 5 – Дані для розрахунку

Обсяг продукції	Вартість фондів (X)	0-2	2-4	4-6	6-8	8-10	m_j
	Центри інтервалів	1	3	5	7	9	
0-0,2	0,1	2	2				4
0,2-0,4	0,3	2	7	10			19
0,4-0,6	0,5		2	17	7		26
0,6-0,8	0,7			4	3	2	9
0,8-1,0	0,9					2	2
n_i		4	11	31	10	4	n=60

Скласти рівняння регресії та обчислити коефіцієнт кореляції.

Розв'язання. Знайдемо умовно середні $\overline{y}_x$ відповідних значенням $X = 1,3,5,7,9$. Так, при значенні $x = 1$ значення $y = 0,1$ спостерігалось 2 рази, а при $y = 0,3$ теж 2 рази (дивись таблицю 4). Тепер умовне середнє $\overline{y}_x$ для $x = 1$ буде дорівнювати

$$\overline{y}_1 = \frac{2 \cdot 0,1 + 2 \cdot 0,3}{4} = 0,2.$$

Аналогічно знаходяться умовні середні $\overline{y}_3$, $\overline{y}_5$, $\overline{y}_7$, $\overline{y}_9$. Складаємо таблицю, в якій вказані спостережені значення X і відповідні їм умовні середні величини $\overline{y}_x$.

Таблиця 6 – Залежність умовних середніх величин $\overline{Y}_x$ від X

X	1	3	5	7	9
$\overline{Y}_x$	0,2	0,3	0,46	0,56	0,8

З таблиці 6 видно, що зі зростанням значень величини *x* зростають умовні середні $\overline{y}_x$. Якщо нанести точки (x_i, $\overline{y}_{x_i}$) на координатну сітку (рис. 3), то побачимо, що вони наближено належать однієї прямій. Це дає нам можливість висловити гіпотезу про лінійну залежність між змінними (x_i, $\overline{y}_{x_i}$).

Використаємо метод найменших квадратів. Представимо таблицю 5 в зручному для обчислень вигляді.

X,Y	1	3	5	7	9	n_y	yn_y	y^2n_y	Σyxn_{xy}
0,1	2	2				4	0,4	0,04	0,8
0,3	2	7	10			19	5,7	1,71	21,9
0,5		2	17	7		26	13	6,5	70
0,7			4	3	2	9	6,3	4,41	41,3
0,9					2	2	1,8	1,62	16,2
n_x	4	11	31	10	4	n=60	27,2	14,28	150,2
xn_x	4	33	155	70	36	298			
x^2n_x	4	99	775	490	324	1692			

$$\bar{x} = \frac{\sum xn_x}{n} = \frac{298}{60} = 4{,}97, \quad \bar{y} = \frac{\sum yn_y}{n} = \frac{27{,}2}{60} = 0{,}45;$$

$$\overline{x^2} = \frac{\sum x^2n_x}{n} = \frac{1692}{60} = 28{,}2;$$

$$\overline{y^2} = \frac{\sum y^2n_y}{n} = \frac{14{,}28}{60} = 0{,}238;$$

$$\overline{xy} = \frac{\sum yxn_{xy}}{n} = \frac{150{,}2}{60} = 5{,}5;$$

$\sigma_x = \sqrt{x^2 - (x)^2} = \sqrt{28{,}2 - (4{,}97)^2} = \sqrt{3{,}499} = 1{,}87;$
$\sigma_y = \sqrt{y^2 - (y)^2} = \sqrt{0{,}238 - (0{,}45)^2} = \sqrt{0{,}0355} = 0{,}188.$

Обчислимо вибірковий коефіцієнт кореляції:
$\overline{r_{xy}} = \frac{\overline{x\bar{y}} - \bar{x}\cdot\bar{y}}{\sigma_x \cdot \sigma_y} = \frac{2{,}503 - 4{,}97 \cdot 0{,}45}{1{,}87 \cdot 0{,}188} = 0{,}76.$
Складаємо вибіркове лінійне рівняння регресії
$y_x = y + \bar{r}_{xy}\frac{\sigma_y}{\sigma_x}(x - \bar{x})$;
$\overline{y_x} = 0{,}45 + 0{,}76\frac{0{,}187}{1{,}87}(x - 4{,}97) = 0{,}072 + 0{,}076x.$
Таким чином,

$$\overline{y_x} = 0{,}072 + 0{,}076x.$$

За одержаним рівнянням регресії обчислюємо значення $\overline{y_x}$ при заданих значеннях *x* і побудуємо пряму лінію, графік якої наносимо на координатну сітку (рис. 3).

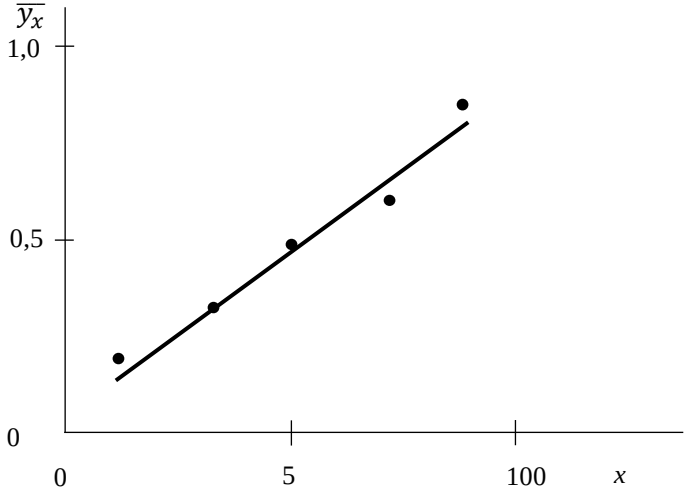

Рис. 3. Графік залежності $\overline{y_x}$ від *x* відповідно до рівняння регресії

Коефіцієнт детермінації дорівнює

$d = r_в^2 \cdot 100\% = 0{,}76^2 \cdot 100\% = 58\%.$

Таким чином, на обсяг випуску продукції (Y) кошти основних промислово-виборчих фондів (X) впливають на 58%, а останні фактори на 42%.

Оцінка значущості вибіркового коефіцієнта кореляції

Після обчислення вибіркового коефіцієнта кореляції необхідно перевірити його значущість. Нехай двомірна генеральна сукупність (X, Y) розподілена за нормальним знаком. З цієї сукупності добута вибірка обсягу n, по якій знайдено вибірковий коефіцієнт кореляції $\overline{r_в} \neq 0$. Висловимо нульову гіпотезу $H_0: r_2 = 0$ тобто гіпотезу про рівність нулю генерального коефіцієнта кореляції.

Якщо нульова гіпотеза приймається, то випадкові величини X і Y некорельовані, а у протилежному випадку корельовані, тобто буде справедливою конкуруюча гіпотеза $H_1: r_2 \neq 0$.

Обчислимо спостерігаєме значення критерію

$$T_{\text{спостер}} = r_в \frac{\sqrt{n-2}}{\sqrt{1-r_в^2}} \qquad (2.20)$$

і по таблиці критичних точок розподілу Стьюдента, при заданому рівні значущості α і числу степеней вільності $k = n - 2$ можна знайти критичну точку $t_{kp}(\alpha; k)$ двосторонньої критичної області. Якщо $\left|\text{T}_{\text{спостер}}\right| < t_{kp}$, то немає підстави відкинути нульову гіпотезу. Якщо $\left|T_{\text{спостер}}\right| > t_{kp}$ – нульову гіпотезу відкидаємо і приймаємо конкуруючу гіпотезу, тобто про значущість вибіркового коефіцієнта кореляції.

Приклад. Перевірити значущість вибіркового коефіцієнта кореляції, обчисленого в прикладі (див. попередній). Обсяг вибірки n=60, вибірковий коефіцієнт кореляції $r_в = 0{,}76$. При заданому рівні значущості $\alpha = 0{,}05$ перевірити нульову гіпотезу про рівність нулю генерального коефіцієнта кореляції при конкуруючій гіпотезі $H_1: r_2 \neq 0$.

Розв'язання. Знайдемо спостерігаєме (емпіричне) значення критерію

$\mathrm{T}_{\text{спостер}} = r_{\text{в}} \frac{\sqrt{n-2}}{\sqrt{1-r_{\text{в}}^2}} = 0{,}76 \cdot \frac{\sqrt{60-2}}{\sqrt{1-0{,}76^2}} = 8{,}91$.

Критична область двостороння. За таблицею критичних точок розподілу Стьюдента для рівня значущості $\alpha = 0{,}05$ і числа степеней вільності $k = 60 - 2 = 58$ знайдемо критичну точку двосторонньої критичної області $t_{kp}(0{,}05; 58) = 2{,}008$. Так як $\mathrm{T}_{\text{спостер}} > t_{kp}$, то нульову гіпотезу відкидаємо. Це значить, що коефіцієнт кореляції значимо відрізняється від нуля, і ми робимо висновок, що випадкові величини $X\ i\ Y$ коррельовані.

Встановимо інтервал довіри для коефіцієнта кореляції генеральної сукупності з нерівності

$r_{\text{в}} - \delta < r_2 < r_{\text{в}} + \delta$,

де $\delta = \frac{t \cdot (1-r_{\text{в}}^2)}{\sqrt{n}}$; n- обсяг вибірки ; t- аргумент функції Лапласа $\Phi(t)$.

Для рівня значущості $\alpha = 0{,}05$ надійність дорівнює

$y = 1 - 0{,}05 = 0{,}95;$

$\Phi(t) = 0{,}95 = 1{,}96. \delta = \frac{t \cdot (1-r_{\text{в}}^2)}{\sqrt{n}} = \frac{1{,}96 \cdot (1-0{,}76^2)}{\sqrt{60}} = 0{,}106.$

Тепер інтервал довіри має вигляд:

$r_{\text{в}} - \frac{t \cdot (1-r_{\text{в}}^2)}{\sqrt{n}} < r_2 < r_{\text{в}} + \frac{t \cdot (1-r_{\text{в}}^2)}{\sqrt{n}}$;

$0{,}654 < r_2 < 0{,}866$.

Перевіримо значущість лінійного рівняння регресії $\overline{y_x} = a + bx$ за критерієм Фішера. Спочатку висловимо робочу гіпотезу $H_0: \overline{y_x}$ – лінійне рівняння значуще.

Відповідно критерію Фішера шукаємо його розрахункове значення через відношення двох дисперсій:

$$F_p = \frac{S_{ad}^2}{s_y^2} \text{ (якщо } S_{ad}^2 > s_y^2) \text{ або}$$

$$F_p = \frac{S_y^2}{S_{ad}^2} \text{ (якщо } S_y^2 > S_{ad}^2).$$

Дисперсія досліду обчислюються за формулою

$$S_y^2 = \frac{1}{n-1} \sum (y_i - \bar{y})^2, \qquad (2.21)$$

де y_{ix} - поточна точка досліду, коли $x = x_i$;

$\overline{y_{ix}}$ - середнє значення признаку;

m - число значущих факторів.

Табличне значення критерію Фішера $F_t(\alpha, k_1, k_2)$ шукається за залежністю від рівня значущості α і ступенів вільності k_1 і k_2. Значення k_1 належить більшій дисперсії $k_1 = n - m$. Значення k_2 залежить від обсягу вибірки $k_2 = n - 1$.

Якщо $F_p < F_1$ - висунута гіпотеза приймається, а якщо $F_p > F_1$, гіпотеза відхиляється.

Приклад. Перевірити за критерієм Фішера адекватність лінійного рівняння регресії, одержаного в задачі (див. попередній приклад) $\overline{y_x} = 0{,}072 + 0{,}076_x$. Необхідні дані для розрахунків наведені в таблиці 6.

Розв'язання. Дисперсії досліду і адекватності обчислимо за формулою (2.21). Результати розрахунків зводимо в таблицю 7.

Таблиця 7 – Результати розрахунків

№ п/п	n=60; $\bar{y} = 0{,}45$; $\bar{y}_x = 0{,}45$						
	x_i	y_i	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$\bar{y}_{x_i}$	$\bar{y}_{x_i} - \bar{y}_x$	$(\bar{y}_{x_i} - \bar{y}_x)^2$
1	2	3	4	5	6	7	8
1	1	0,2	-0,25	0,0625	0,148	-0,302	0,0912
2	3	0,3	-0,15	0,0225	0,3	-0,15	0,025
3	5	0,43	0,01	0,0001	0,452	0,002	0,000004
4	7	0,56	0,11	0,0121	0,604	0,154	0,0237
5	9	0,8	0,35	0,1225	0,756	0,306	0,0936
$\sum$				0,2197	2,26		0,1137
В середньому					0,45		

Дисперсія досліду $S_y^2 = \frac{1}{n-1}\sum(y_i - \bar{y})^2 = \frac{0{,}2197}{59} = 0{,}0037$.

Дисперсія адекватності $S_{ad}^2 = \frac{1}{n-m}\sum(Y_{ix} - \overline{y_{ix}})^2 = \frac{0{,}1137}{59} = 0{,}001972.$

Очевидно, що X є значущим фактором, тому $m = 1.$

Порівнюючи значення дисперсій S_y^2 і S_{ad}^2, відмітимо, що $S_y^2 > S_{ad}^2$, тому використовуємо для розрахункового значення критерію адекватності формулу

$F_p = \frac{S_y^2}{S_{ad}^2} = \frac{0{,}0037}{0{,}001972} = 1{,}876.$

Критичне значення точок розподілу Фішера $F_{kp}(k_1, k_2)$ для $\alpha = 0{,}05$ шукаємо з відповідних таблиць. $F_{kp} = 2{,}69$. Порівнюючи значення критеріїв бачимо, що $F_p < F_{ad}$, тому лінійна модель адекватна.

Висновки

Метод найменших квадратів (МНК), завдяки широкій сфері застосування, посідає виняткове місце серед методів математичної статистики. Задачею МНК є оцінка закономірностей, які спостерігаються на тлі випадкових коливань, та її використання для подальших розрахунків, зокрема, для прогнозів. Особливу роль відіграють МНК у техніці, визначаючи концепцію та методологію розв‘язання широкого кола задач метрологічної оцінки результатів експерименту.

Список використаної літератури

1. ДСТУ 4116-2002 Метрологія. Державна повірочна схема для засобів вимірювань електричної потужності і коефіцієнта потужності у діапазоні частот від 40 до 20000 Гц.
2. О.А.-Б. Ахмадов, С.О. Ахмадов, С.Р. Карпенко, С.П. Токар, В.С. Писчиков. Вторинний еталон електричної потужності для розширеного діапазону частот. Вісник інженерної академії України. – Випуск №1, 2011. – С. 243 – 247.
3. О.А.-Б. Ахмадов, С.О. Ахмадов, С.Р. Карпенко. Розрахунок невизначеності Вторинного еталону електричної потужності для розширеного діапазону частот. Вісник інженерної академії України. – Випуск №3-4, 2011. – С. 137 – 142.

4. ДСТУ 3231:2007 Метрологія. Еталони одиниць вимірювань державні, первинні та вторинні. Основні положення, порядок розроблення, затвердження, реєстрації, зберігання та застосування.
5. Макаров Е.Г. «Самоучитель MathCad 14»: 2008.

Навчальне видання

МЕТОД НАЙМЕНШИХ КВАДРАТІВ

Навчально-методичний посібник

КОСУЛІНА Наталія Геннадіївна,
ЛЯШЕНКО Геннадій Анатолійович,
ЗОТОВА Ольга Сергіївна,
ПОЛЯНОВА Надія Володимирівна

Формат 60x84/16. Гарнітура Times New Roman
Папір для цифрового друку. Друк ризографічний.
Ум. друк. арк. __.
Наклад __пр.
Харківський національний технічний університет сільського господарства імені Петра Василенка
61002, м. Харків, вул. Алчевських, 44