Метод найменших квадратів.

(1)
П5

5 Передбачається, що залежність ознаки Y від ознаки X має такий вигляд

$$y = f(x, \alpha_1, \alpha_2, \dots, \alpha_m)$$

де – значення ознаки X; y – значення ознаки Y; $\alpha_1, \alpha_2, \dots, \alpha_m$ – параметри, які належить визначити, та, що за результатами експерименту отримані такі емпіричні дані:

(*)

Значення ознаки X	x_1	x_2	...	x_i	...	x_n
Значення ознаки Y	y_1	y_2	...	y_i	...	y_n

Метод найменших квадратів стверджує, що найімовірніше значення параметрів $\alpha_1, \alpha_2, \dots, \alpha_m$ дає мінімум функції

$$S = \sum_{i=1}^{n} [y_i - f(x_i, \alpha_1, \alpha_2, \dots, \alpha_m)]^2.$$

Коли $f(x, \alpha_1, \alpha_2, \dots, \alpha_m)$ ~~непе~~ має неперервні частинні похідні за усіма своїми параметрами, то необхідна умова мінімуму функції S складає систему m рівнянь з m невідомими:

$$(*,*) \quad \frac{\partial S}{\partial \alpha_1} = 0, \quad \frac{\partial S}{\partial \alpha_2} = 0, \quad \dots, \quad \frac{\partial S}{\partial \alpha_m} = 0.$$

Визначення функціональної залежності між ознаками Y та X на основі експериментальних даних (*) називають <u>вирівнюванням</u> емпіричних даних вздовж кривої $y = f(x, \alpha_1, \alpha_2, \dots, \alpha_m)$.

Якщо $f(x, \alpha_1, \alpha_2, \dots, \alpha_m) = \alpha_1 x + \alpha_2$ то зазначеною кривою буде пряма лінія $y = \alpha_1 x + \alpha_2$. У цьому випадку система $(*, *)$ може бути перетворена у так звану нормальну систему методу найменших квадратів за умови вирівнювання за прямою:

$$\begin{cases} \alpha_1 \sum_{i=1}^{n} x_i^2 + \alpha_2 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i ; \\ \alpha_1 \sum_{i=1}^{n} x_i + \alpha_2 n = \sum_{i=1}^{n} y_i . \end{cases}$$

Система $(*, *)$ коли вирівнювання здійснюється за параболою $y = \alpha_1 x^2 + \alpha_2 x + \alpha_3$ може бути перетворена до такого вигляду:

$$\begin{cases} \alpha_1 \sum_{i=1}^{n} x_i^4 + \alpha_2 \sum_{i=1}^{n} x_i^3 + \alpha_3 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i^2 y_i ; \\ \alpha_1 \sum_{i=1}^{n} x_i^3 + \alpha_2 \sum_{i=1}^{n} x_i^2 + \alpha_3 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i ; \\ \alpha_1 \sum_{i=1}^{n} x_i^2 + \alpha_2 \sum_{i=1}^{n} x_i + \alpha_3 n = \sum_{i=1}^{n} y_i \end{cases}$$

Лінійна кореляційна залежність

Стверджується, що дві ознаки X та Y знаходяться в кореляційній залежності, якщо кожному значенню одного з них відповідає деякий розподіл іншого. Кореляційна залежність між ознаками X та Y задається за допомогою кореляційної таблиці:

Кореляційна таблиця

Y / X	y_1	y_2	$\ldots$	y_j	$\ldots$	y_m	m_x
x_1	m_{11}	m_{12}	$\ldots$	m_{1j}	$\ldots$	m_{1n}	m_{x1}
x_2	m_{21}	m_{22}	$\ldots$	m_{2j}	$\ldots$	m_{2n}	m_{x2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	m_{i1}	m_{i2}	$\ldots$	m_{ij}	$\ldots$	m_{in}	m_{xi}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	m_{k1}	m_{k2}	$\ldots$	m_{kj}	$\ldots$	m_{kn}	m_{xk}
m_y	m_{y1}	m_{y2}	$\ldots$	m_{yj}	$\ldots$	m_{yn}	N

У цій таблиці $x_1, x_2, \ldots, x_i, \ldots, x_k$; $y_1, y_2, \ldots, y_j, \ldots, y_n$ – середини інтервалів або значення ознак X та Y, а $m_{x1}, m_{x2}, \ldots, m_{xi}, \ldots, m_{xk}$; $m_{y1}, m_{y2}, \ldots, m_{yj}, \ldots, m_{yn}$ – відповідні частоти; m_{ij} – частота, з якою зустрічається пара $(x_i; y_j)$; $N = \sum_{i=1}^{k} \sum_{j=1}^{n} m_{ij}$.

Кореляційна залежність між ознаками Y та X може бути замінена <u>функціональною</u>, якщо кожному значенню ознаки X поставити у відповідність умовне середнє ознаки Y, тобто для $X = x_i$ поставити у відповідність величину

$$\bar{Y}_{x_i} = \frac{\sum_{j=1}^{n} m_{ij} y_j}{m_{xi}}.$$

Якщо потім точки $(x_i; \bar{Y}_{x_i})$ вирівняти за методом найменших квадратів вздовж кривої $y = f(x, \alpha_1, \alpha_2, \ldots, \alpha_m)$, то остання називається лінією регресії y на x, а її рівняння – рівнянням регресії y на x. Аналогічно визначається лінія регресії x на y. Найпростішими і найважливішими випадками кривих регресії є прямі лінії.

Кутовий коефіцієнт прямої регресії y на x (x на y) називають коефіцієнтом регресії y на x (x на y) і позначають так $\rho_{y/x}$ ($\rho_{x/y}$). Коефіцієнти регресії можуть бути розраховані за формулами:

$$\rho_{y/x} = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\sigma_X^2}; \quad \rho_{x/y} = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\sigma_Y^2},$$

де $\overline{XY} = \dfrac{\sum_{i=1}^{k}\sum_{j=1}^{n} m_{ij} x_i y_j}{N}$ – середнє значення добутку ознак X та Y; $\overline{X}$ і $\overline{Y}$ – їх середні значення, а σ_X^2 та σ_Y^2 – їх дисперсії.

Рівняння прямих регресій мають ~~вигл~~ вигляд:

$$y - \overline{Y} = \rho_{y/x}(x - \overline{X});$$

$$x - \overline{X} = \rho_{x/y}(y - \overline{Y}).$$

Коефіцієнтом лінійної кореляції ознак X та Y називається величина

$$r = r(X, Y) = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\sigma_X \sigma_Y} = \pm\sqrt{\rho_{y/x}\,\rho_{x/y}}.$$

Коефіцієнт лінійної кореляції r характеризується такими властивостями:

1) $-1 \le r \le 1$;

2) $r[\alpha(X-x_0), \beta(Y-y_0)] = r(X,Y)$ $(\alpha>0, \beta>0)$;

3) якщо $r(X,Y)=\pm 1$, то між ознаками X і Y існує лінійна функціональна залежність (коли $r=1$ – пряма залежність, а коли $r=-1$ – зворотня залежність);

4) якщо $r(X,Y)=0$, то між ознаками X та Y відсутня лінійна кореляційна залежність;

5) $\rho_{y/x} = r\dfrac{\sigma_y}{\sigma_x}$ і $\rho_{x/y} = r\dfrac{\sigma_x}{\sigma_y}$.

Квадрат коефіцієнта лінійної кореляції дає коефіцієнт детермінації, який вимірює долю варіації Y, яка пояснюється впливом ознаки X, і навпаки.

На практиці про розподіл ознак X та Y в генеральній сукупності визначають за даними вибірки. За цими даними можливо знайти вибірковий коефіцієнт лінійної кореляції r_B, який є випадковою величиною. За умови достатньо великого об'єму вибірки $r_B \approx r$. Якщо розподіл ознак X та Y досить близький до нормального, то можливо наближено вважати r_B також нормальною випадковою величиною, середнє квадратичне відхилення якої дорівнює $\dfrac{1-r^2}{\sqrt{N}}$, де N – об'єм вибірки.

Приклад розв'язання задачі:

Задача 1.

Компанія-перевізник провела статистичне дослідження на основних маршрутах і одержала залежність між вартістю перевезення ξ (в умовних одиницях за 1 км) і довжиною маршруту η (тис. км). Результати дослідження наведені в таблиці:

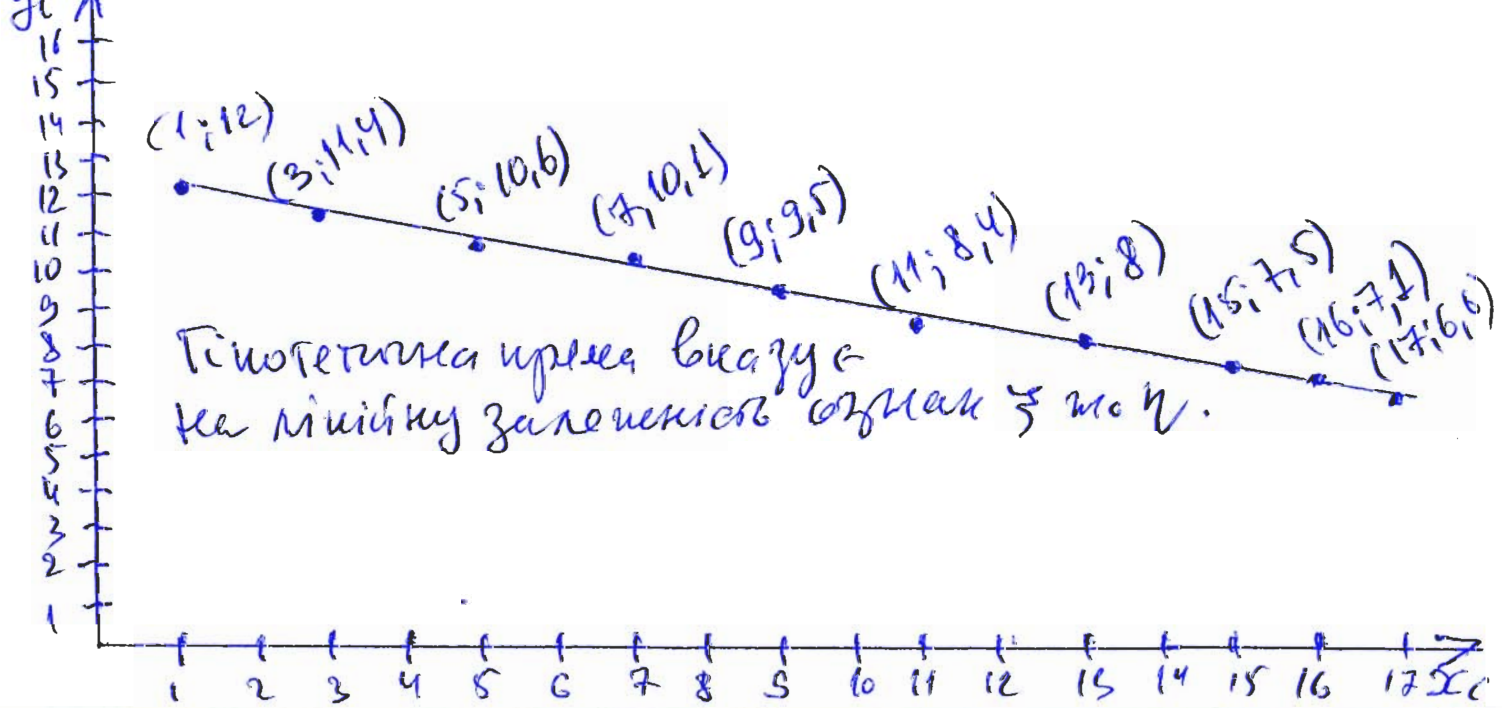
Значення вартості x_i	1	3	5	7	9	11	13	15	16	17
Довжина маршруту y_i	12	11,4	10,6	10,1	9,5	8,4	8	7,5	7,1	6,6

В задачі потрібно

1) встановити форму залежності між ξ і η;
2) знайти рівняння лінійної регресії η на ξ та ξ на η;
3) обчислити коефіцієнт кореляції вибірки та оцінити силу лінійного зв'язку між ξ та η.

Розв'язок.

1) Для обґрунтування залежності ознак ξ та η графічно зобразимо точки (x_i, y_i).



2) Зв'язок між ознаками ξ та η, які вивчаються, може бути виражений рівнянням прямої лінії регресії η на ξ: $\bar{y}_x = ax + b$.
Для обчислення параметрів a і b та коефіцієнта кореляції складаємо розрахункову таблицю:

(*)

											сума в рядку
x_i	1	3	5	7	9	11	13	15	16	17	97
y_i	12	11,4	10,6	10,1	9,5	8,4	8	7,5	7,1	6,6	91,2
x_i^2	1	9	25	49	81	121	169	225	256	289	1225
y_i^2	144	129,96	112,36	102,01	90,25	70,56	64	37,5	50,41	43,56	844,3
x_iy_i	12	34,2	53	70,7	85,5	92,4	104	112,5	113,6	105,6	783,5

Знайдемо рівняння лінійної регресії η на ξ.

$$\bar{y}_x = ax + b$$

Застосуємо метод найменших квадратів і складаємо систему рівнянь для визначення параметрів a і b.

$$\begin{cases} a\sum\limits_{i=1}^{n} x_i^2 + b\sum\limits_{i=1}^{n} x_i = \sum\limits_{i=1}^{n} x_i y_i \\ a\sum\limits_{i=1}^{n} x_i + bn = \sum\limits_{i=1}^{n} y_i \end{cases};$$

$n = 10$

Підставляємо значення із таблиці (*) і отримуємо систему рівнянь:

$$\begin{cases} 1225 \cdot a + 97 \cdot b = 783,5 \\ 97 \cdot a + 10 \cdot b = 91,2 \end{cases}$$

$$a = \frac{783,5 - 97b}{1225}, \quad 97\,\frac{783,5 - 97b}{1225} + 10b = 91,2 \Rightarrow$$

$$\Rightarrow 97(0,64 - 0,08b) + 10b = 91,2 \Rightarrow b = \frac{29,12}{2,24} = 13;$$

$$a = \frac{783,5 - 97 \cdot 13}{1225} = \frac{783,5 - 1261}{1225} = \frac{-477,5}{1225} = -0,4;$$

$$\underline{\bar{y}_x = -0,4x + 13.}$$

Знайдемо рівняння лінійної регресії ξ на η

$$\overline{x}_y = ay + b.$$

Для визначення параметрів a та b за методом найменших квадратів запишемо відповідну систему рівнянь.

$$\begin{cases} a\sum\limits_{i=1}^{n} y_i^2 + b\sum\limits_{i=1}^{n} y_i = \sum\limits_{i=1}^{n} x_i y_i \\ a\sum\limits_{i=1}^{n} y_i + bn = \sum\limits_{i=1}^{n} x_i \end{cases}$$

Із таблиці (*) підставляємо розраховані дані. Отримуємо таку систему рівнянь:

$$\begin{cases} 844{,}3\cdot a + 91{,}2b = 783{,}5 \\ 91{,}2a + 10b = 97 \end{cases}$$

Розв'яжемо систему рівнянь.

$$a = \frac{783{,}5 - 91{,}2b}{844{,}3} \Rightarrow 91{,}2(0{,}93 - 0{,}1b) + 10b = 97;$$

$$\Rightarrow 84{,}8 - 9{,}12b + 10b = 97; \Rightarrow 0{,}88b = 12{,}2 \Rightarrow b = \frac{12{,}2}{0{,}88} =$$

$$= 13{,}9 \approx 14$$

$$a = \frac{783{,}5 - 91{,}2\cdot 13{,}9}{844{,}3} = \frac{783{,}5 - 1267{,}68}{844{,}3} = \frac{-484{,}18}{844{,}3} =$$

$$= -0{,}57; \Rightarrow a = -0{,}57,\ b = 14 \quad \underline{\overline{x}_y = -0{,}57x + 14.}$$

3) Обчислимо коефіцієнт кореляції r вибірки. Коефіцієнт кореляції визначається за формулою

$$r(\xi,\eta) = \frac{\overline{\xi\eta} - \overline{\xi}\cdot\overline{\eta}}{\sigma_\xi\cdot\sigma_\eta},$$

Необхідно знайти вибіркові середнє добутку ознак ξ та η: $\overline{\xi\eta} = \frac{1}{n}\sum\limits_{i=1}^{n} x_i y_i$; середні вибіркові

$$\overline{\xi} = \frac{1}{n}\sum_{i=1}^{n} x_i\,; \quad \overline{\eta} = \frac{1}{n}\sum_{i=1}^{n} y_i\,;$$

та вибіркові середньо квадратичні відхилення

$$\sigma_\xi = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}_B\right)^2}\,; \quad \sigma_\eta = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \overline{y}_B\right)^2}$$

Застосуємо обчислені значення із таблиці (*) і знайдемо вибіркові середні:

$$\overline{\xi\eta} = \frac{1}{n}\sum_{i=1}^{n} x_i \cdot y_i = \frac{1}{10}\cdot 783{,}5 = 78{,}35\,;$$

$$\overline{\xi} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{10}\cdot 97 = 9{,}7\,; \quad \overline{\eta} = \sum_{i=1}^{n} y_i = \frac{91{,}2}{10} = 9{,}12\,;$$

Для визначення σ_ξ і σ_η складемо таблицю для $(x_i - \overline{x}_B)^2$ та $(y_i - \overline{y}_B)^2$

(*,*)

$(x_i - \overline{x}_B)^2$	75,69	44,89	22,09	7,29	0,49	1,69	10,89	28,09	39,69	53,29
$(y_i - \overline{y}_B)^2$	8,3	5,2	2,2	0,77	0,14	0,52	1,25	2,62	4,08	6,35

Застосовуючи значення із таблиці (*, *) знайдемо σ_ξ та σ_η

$$\sigma_\xi = \left[\frac{1}{9}\left(75{,}69 + 44{,}89 + 22{,}09 + 7{,}29 + 0{,}49 + 1{,}69 + 10{,}89 + 28{,}09 + 39{,}69 + 53{,}29\right)\right]^{1/2} =$$

$$= \sqrt{\frac{284{,}1}{9}} = 5{,}61\,; \quad \sigma_\xi = 5{,}61$$

$$\sigma_\eta = \left[\frac{1}{9}\left(8{,}3 + 5{,}2 + 2{,}2 + 0{,}77 + 0{,}14 + 0{,}52 + 1{,}25 + 2{,}62 + 4{,}08 + 6{,}35\right)\right]^{1/2} = \sqrt{\frac{31{,}43}{9}} = 1{,}87$$

$$\sigma_\eta = 1{,}87$$

$$r(\xi, \eta) = \frac{78,35 - 9,7 \cdot 9,12}{5,61 \cdot 1,87} = \frac{78,35 - 88,46}{10,49} =$$

$$= \frac{-10,11}{10,49} = -0,96$$

Ознаки ξ та η від'ємно корельовані. Це визначається тим, що коли ознака ξ зростає — ознака η спадає. Значення r вказує на те, що кореляція ознак ξ та η висока.