Лекція 20. Елементи регресивного аналізу

За результатами експерименту можна отримати наближене вираження (оцінку) функції регресії (тобто вибіркові рівняння регресії) вигляду:

$$\overline{y}_x = f(x) \text{ та } \overline{x}_y = \varphi(y),$$

де $\overline{y}_x$ - умовне середнє змінної Y при фіксованому значенні X, $\overline{x}_y$ - умовне середнє змінної X при фіксованому значенні Y.

Кореляційна залежність між величинами X і Y може бути подана як функціональна залежність між $\overline{y}_x$ і X або $\overline{x}_y$ і Y.

Методи знаходження таких залежносте і оцінки їхніх статистичних властивосте складають зміст *регресивного аналізу*.

20.1.Метод найменших квадратів. Парна лінійна регресія.

Завдання: за заданими значеннями $(x_1, y_1),\ldots,(x_n, y_n)$ і відповідними частотами цих значень n_{ij}, а також за умовними середніми $\overline{x}_{y_j}$ та $\overline{y}_{x_i}$ знайти функції регресії $f(x)$ і $\varphi(y)$.

Найпростішою формою кореляційно-регресивного зв’язку є лінійний зв'язок між двома величинами – парна лінійна регресія. Її рівняння:

$\overline{y}_x = a + bx$ або $\overline{x}_y = c + dy$.

У кожному з цих рівнянь по дві невідомі величини – коефіцієнти лінійної регресії. Так у рівнянні $\overline{y}_x = a + bx$ це параметри a і b. Однозначно визначити їх за вибіркою неможливо – через вплив випадкових чинників.

Вплив цих випадкових відхилень (похибок, помилок, збурень, шуму) на спостережувані значення подають у наступному вигляді:

$$\overline{y}_{x_i} = a + bx_i + \varepsilon_i, \ \left(i = \overline{1,n}\right),$$

де ε_i - випадкова змінна, вона характеризує відхилення значень вибірки від теоретичної регресії.

Потрібно знайти значення невідомих параметрів a і b, щоб випадкові відхилення ε_i у сукупності були близькі до нуля.

Суть методу найменших квадратів: знати таку значення параметрів a і b, щоб була як можна меншою зважена сума квадратів відхилень:

$$\sum_{i=1}^{n} \varepsilon_i^2 \cdot m_{x_i} \to \min .$$

Тоді буде знайдено рівняння прямої, найменш віддаленої від усіх точок $\left(x_i, \overline{y}_{x_i}\right)$ з урахуванням їхніх вагів.

Записавши необхідну умову існування її мінімуму (рівність нулю двох частинних похідних), отримують лінійну систему двох рівнянь із двома змінними a і b.

$$\begin{cases} na + b \cdot \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ a \cdot \sum_{i=1}^{n} x_i + b \cdot \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i \cdot y_i \end{cases} \qquad (20.1)$$

Розв'язавши систему (20.1) відносно параметрів a і b, знайдемо:

$$a = \overline{y} - b \cdot \overline{x};$$

$$b = \frac{\frac{\sum x_i y_i}{n} - \overline{x} \cdot \overline{y}}{\frac{\sum x_i^2}{n} - \left(\overline{x}\right)^2} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} \cdot \left(\overline{x}\right)^2} = \frac{K_{xy}^*}{\sigma_x^2}. \qquad (20.2)$$

де $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\overline{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$, $\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$, $\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i \cdot y_i$, n - число вимірювань.

Помноживши ліву і праву частини (20.2) на $\frac{\sigma_x}{\sigma_y}$, дістанемо:

$$\frac{\sigma_x}{\sigma_y} \cdot b = \frac{K_{xy}^*}{\sigma_x^2} \cdot \frac{\sigma_x}{\sigma_y} = \frac{K_{xy}^*}{\sigma_x \cdot \sigma_y} = r_{xy} \quad \Rightarrow \quad b = r_{xy} \cdot \frac{\sigma_y}{\sigma_x}, \qquad (20.3)$$

де r_{xy} - парний коефіцієнт кореляції між ознаками X і Y.

Тоді

$$a = \bar{y} - b \cdot \bar{x} = \bar{y} - r_{xy} \cdot \frac{\sigma_y}{\sigma_x} \cdot \bar{x}. \qquad (20.4)$$

З урахуванням (20.3), (20.4) рівняння лінійної парної регресії набере такого вигляду:

$$y_x = r_{xy} \cdot \frac{\sigma_x}{\sigma_y} \cdot (x - \bar{x}) + \bar{y} \qquad (20.5)$$

Для визначення оцінки параметрів лінійної залежності (регресії) Y на X $y_x = a + b \cdot x_i$, можна також застосувати формулу через незміщену статистичну оцінку (виправлену дисперсію) ознаки X: $b = \frac{K^*_{xy}}{S_x^2}$.

Для лінійної регресії X на Y отримують аналогічні формули ($x_y = c + dy$):

$d = \frac{K^*_{xy}}{\sigma_y^2}$ або через виправлену дисперсію $d = \frac{K^*_{xy}}{S_y^2}$, тоді $c = \bar{x} - d \cdot \bar{y}$:

$$x_y = r_{xy} \cdot \frac{\sigma_x}{\sigma_y} \cdot (y - \bar{y}) + \bar{x}. \qquad (20.6)$$

Прямі регресії Y на X і X на Y співпадають коли $|r_{xy}| = 1$, тобто у разі функціональної лінійної залежності між величинами X і Y.

Співвідношення $\sqrt{b \cdot d} = |r_{xy}|$ використовують для контролю обчислень.

Коефіцієнт детермінації $R^2 = r_{xy}^2$, дорівнює квадрату коефіцієнта множинної кореляції. Коефіцієнт детермінації R^2 показує, в якій мірі варіації значень величини Y залежить від значень фактору X.

Приклад 20.5. Залежність обсягу отриманого прибутку деяким умовним підприємством регіону від вартості основних виробничих фондів наведено парним статистичним розподілом вибірки:

Основні фонди, млн грн, x_k	2,5	2,8	3	3,2	3,5	4,2	4,5	5	5,3	6
Прибуток, млн грн, y_k	1,2	1,5	1,7	2,2	2,6	3,1	3,4	4,2	4,7	5,4

Методом найменших квадратів визначити оцінки невідомих параметрів лінійної парної регресії. Обчислити коефіцієнт кореляції та детермінації, зробити висновки.

Розв'язання: Нехай між ознаками X та Y існує лінійна функціональна залежність $y_x = a + b \cdot x_i$.

Для визначення параметрів a та b скористаємося методом найменших квадратів, що має такий вигляд:

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

№ п/п	x_k	y_k	x_k^2	$x_k y_k$	y_k^2
1	2,5	1,2	6,25	3,0	1,44
2	2,8	1,5	7,84	4,2	2,25
3	3	1,7	9	5,1	2,89
4	3,2	2,2	10,24	7,0	4,84
5	3,5	2,6	12,25	9,1	6,76
6	4,2	3,1	17,64	13,0	9,61
7	4,5	3,4	20,25	15,3	11,56
8	5	4,2	25	21,0	17,64
9	5,3	4,7	28,09	24,9	22,09
10	6	5,4	36	32,4	29,16
$\sum$	**40**	**30**	**172,56**	**135,07**	**108,24**

$$\begin{cases} 10 \cdot a + 40 \cdot b = 30 \\ 40 \cdot a + 172{,}56 \cdot b = 135{,}07 \end{cases} \Rightarrow \begin{cases} a = -1{,}799 \\ b = 1{,}12 \end{cases}$$

Отже, рівняння регресії буде $y_x = -1{,}799 + 1{,}12 \cdot x_i$.

Для обчислення коефіцієнта кореляції r_{xy}, визначимо кореляційний момент $K^*_{xy} = \overline{xy} - \bar{x} \cdot \bar{y} = 13,507 - 4 \cdot 3 = 1,507$.

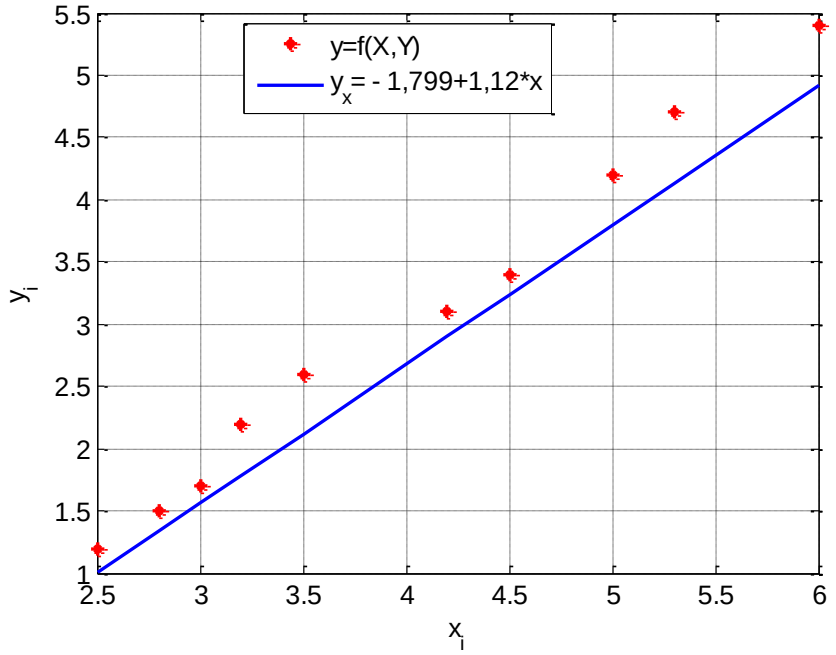
$$S_x^2 = \frac{n}{n-1} D_x = \frac{10}{9} \cdot 1,2544 = 1,394; \quad S_x = 1,18;$$

$$S_y^2 = \frac{n}{n-1} D_y = \frac{10}{9} \cdot 1,8225 = 2,025; \quad S_y = 1,42.$$

Вибірковий коефіцієнт кореляції

$$r_{xy} = \frac{K^*_{xy}}{\sqrt{S_x^2 \cdot S_y^2}} = \frac{1,507}{1,18 \cdot 1,42} \approx 0,9.$$

Побудуємо кореляційне поле та регресивну функцію



Прямі регресії на кореляційному полі

Визначимо коефіцієнт детермінації $R^2 = r_{xy}^2 = 0,9^2 = 0,81$.

Коефіцієнт детермінації $R^2 = 0,81$. Це означає, що зміна обсягу прибутку підприємства на 81% визначається варіацією вартості основних фондів і 19% – іншими випадковими факторами.

Приклад 20.2. Дано вибірку

x_i	3,72	3,09	3,47	3,25	3,34	3,11	3,51	3,34	3,68	3,40
y_i	1,49	2,73	3,32	3,32	3,69	3,67	3,30	2,55	3,11	3,60

Обчислити коефіцієнт кореляції, визначити і побудувати прямі регресії Y на X та X на Y.

i	x_i	$x_i-\overline{x}_B$	y_i	$y_i-\overline{y}_B$	$(x_i-\overline{x}_B)(y_i-\overline{y}_B)$	$(x_i-\overline{x}_B)^2$	$(y_i-\overline{y}_B)^2$
1	3,72	0,33	1,49	– 1,59	– 0,525	0,109	2,528
2	3,09	– 0,3	2,73	– 0,35	0,105	0,09	0,123
3	3,47	0,08	3,32	0,24	0,019	0,006	0,058
4	3,25	– 0,14	3,32	0,24	– 0,034	0,020	0,058
5	3,34	– 0,05	3,69	0,61	– 0,03	0,003	0,372
6	3,11	– 0,28	3,67	0,59	– 0,165	0,078	0,348
7	3,51	0,12	3,30	0,22	0,026	0,014	0,048
8	3,34	– 0,05	2,55	– 0,53	0,027	0,003	0,281
9	3,68	0,29	3,11	0,03	0,009	0,084	0,001
10	3,40	0,01	3,60	0,52	0,005	0,000	0,270
∑	33,91		30,78		– 0,56	0,41	4,09

Визначимо оцінки параметрів лінійної залежності (регресії) Y на X $y_x = a + b \cdot x_i$, та лінійної регресії X на Y: $x_y = c + dy$.

Для обчислення коефіцієнта кореляції r_{xy}, визначимо кореляційний момент $K^*_{xy} = \dfrac{1}{n-1} \cdot \sum_{j=1}^{n} \sum_{i=1}^{k} \left(x_j - \overline{x}\right) \cdot \left(y_i - \overline{y}\right) = \dfrac{-0,56}{9} = -0,06$.

$$S_x^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2 = \frac{1}{9} \cdot 0,41 = 0,05; \quad S_x = 0,22;$$

$$S_y^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2 = \frac{1}{9} \cdot 4,09 = 0,45; \quad S_y = 0,67.$$

Вибірковий коефіцієнт кореляції

$$r_{xy} = \frac{K^*_{xy}}{S_x \cdot S_y} = -\frac{0,06}{0,22 \cdot 0,67} \approx -0,41.$$

Оцінимо параметри:

$b = \dfrac{K_{xy}^*}{S_x^2} = -\dfrac{0,06}{0,05} = -1,2$, тоді $a = \overline{y} - b \cdot \overline{x} = 3,08 + 1,2 \cdot 3,39 = 7,15$.

Оцінимо параметри:

$d = \dfrac{K_{xy}^*}{S_y^2} = -\dfrac{0,06}{0,45} = -0,13$, тоді $c = \overline{x} - d \cdot \overline{y} = 3,39 + 0,13 \cdot 3,08 = 3,8$.

Таким чином, прямі регресії мають наступні рівняння:

$$y_x = a + b \cdot x_i \quad \Rightarrow \quad y_x = 7,15 - 1,2 \cdot x_i$$
$$x_y = c + d \cdot y_i \quad \Rightarrow \quad x_y = 3,8 - 0,13 \cdot y_i$$

Для перевірки застосуємо співвідношення:

$\sqrt{b \cdot d} = |r_{xy}| \quad \Rightarrow \quad \sqrt{(-1,2) \cdot (-0,13)} = 0,39$.

З урахуванням округлення помилок можна вважати, що $0,39 \approx |-0,41|$.