Розділ 12

ЕЛЕМЕНТИ КОРЕЛЯЦІЙНОГО ТА РЕГРЕСІЙНОГО АНАЛІЗУ

Кожній величині, яку отримують у результаті проведення експерименту, притаманний елемент випадковості, що виявляється більшою чи меншою мірою залежно від її природи.

У разі сумісної появи двох і більше величин у результаті проведення експерименту дослідник має підстави для встановлення певної залежності між ними, зв'язку.

Строгої функціональної залежності між змінними, у буквальному розумінні цього слова, у реальному світі не існує, бо вони перебувають під впливом випадкових факторів, наслідки якого передбачити практично неможливо. Тому між змінними існує особлива форма зв'язку, яку називають стохастичною і яка в математичній статистиці трансформується, не змінюючи своєї сутності, у статистичну залежність.

Наприклад, при дослідженні двох змінних X та Y зміна значень $X = x_i$ призводить до такої зміни значень Y, яку можна розбити на два компоненти: систематичну, що пов'язана із залежністю, котра існує між X та Y, і випадкову, яка зазнає впливу випадкових факторів.

Показником, що вимірює стохастичний зв'язок між змінними, є *коефіцієнт кореляції*, який свідчить, з певною мірою ймовірності, наскільки зв'язок між змінними близький.

За наявності кореляційного зв'язку між змінними необхідно виявити його форму функціональної залежності (лінійна чи нелінійна), а саме:

$$y = a_0 + a_1 x; \tag{12.1}$$

$$y = a_0 + a_1 x + a_2 x^2; \tag{12.2}$$

$$y = a_0 + \frac{a_1}{x}. \tag{12.3}$$

Наведені можливі залежності між змінними X і Y (12.1), (12.2), (12.3) називають *функціями регресії*. Форму зв'язку між змінними X і Y можна

встановити, застосовуючи кореляційні поля. Кожній точці з координатами x_i, y_i відповідає певне числове значення ознак *X* та *Y*.

Отже, на основі розміщення точок кореляційного поля дослідник має підстави для гіпотетичного припущення про лінійні чи нелінійні залежності між ознаками *X* і Y.

Припустимо, що нам відома функціональна залежність між випадковими величинами *Y* та *X* вигляду

$$Y = f(X; a_1; a_2; \dots; a_m) \tag{12.4}$$

з невідомими параметрами $a_1; a_2; \dots; a_m$.

Нехай внаслідок n незалежних випробувань одержані варіанти ознак Y та X, які оформлені у статистичній таблиці вигляду:

X	x_1	x_2	…	x_k	…	x_n
Y	y_1	y_2	…	y_k	…	y_n

Для знаходження оцінок параметрів функціональної залежності $a_1; a_2; \dots; a_m$ за даними вибірки застосуємо метод найменших квадратів. Цей метод ґрунтується на тому, що найімовірніші значення параметрів $a_1;\ a_2;\ \dots\ ;\ a_m$ повинні давати мінімум функції

$$S = \sum_{k=1}^{n} \left[y_k - f(x_k;\ a_1;\ a_2; \dots;\ a_m)\right]^2 \tag{12.5}$$

Якщо функція $f(x_k;\ a_1;\ a_2; \dots;\ a_m)$ має неперервні частинні похідні відносно невідомих параметрів $a_1;\ a_2; \dots;\ a_m$, то необхідною умовою існування мінімуму функції S буде система m рівнянь з m невідомими

$$\frac{\partial S}{\partial a_k} = 0, \quad k = 1, 2, \dots, m. \tag{12.6}$$

Знаходження функціональної залежності між випадковими величинами *X* та *Y* з використанням даних випробувань (або вибірки) називають вирівнюванням емпіричних даних вздовж кривої $y = f(x; a_1; a_2; \dots;\ a_m)$.

12.1. Рівняння лінійної парної регресії

Нехай між змінними *X* та *Y* теоретично існує певна лінійна залежність. Це твердження може ґрунтуватися на тій підставі, наприклад, що кореляційне поле для пар $(x_i;\ y_i)$ має такий вигляд (рис. 12.1).

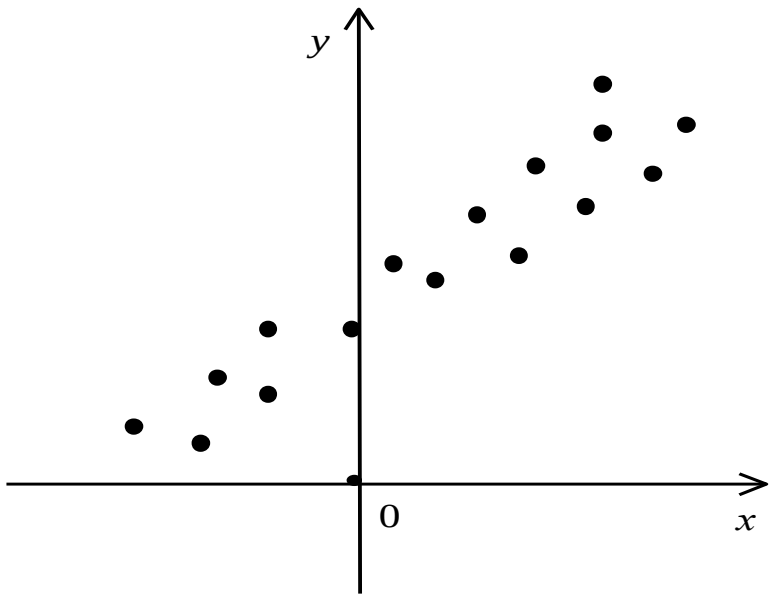

Рис. 12.1. Кореляційне поле для пар

Як бачимо, насправді між ознаками X та Y спостерігається не такий тісний зв'язок, як це передбачає функціональна залежність.

Окремі спостережувані значення y, як правило, відхилятимуться від передбаченої лінійної залежності під впливом випадкових збудників, які здебільшого є невідомими. Відхилення від передбаченої лінійної форми зв'язку можуть статися внаслідок неправильної специфікації рівняння, тобто ще з початку неправильно вибране рівняння, що описує залежність між X і Y.

Будемо вважати, що специфікація рівняння вибрана правильно. Ураховуючи вплив на значення Y збурювальних випадкових факторів, лінійне рівняння зв'язку X і Y можна подати в такому вигляді:

$$y = a_0 + a_1 x + \varepsilon, \tag{12.7}$$

де a_0, a_1 є невідомі параметри регресії, ε є випадковою змінною, що характеризує відхилення y від гіпотетичної теоретичної регресії.

Отже, в рівнянні (12.7) значення «y» подається у вигляді суми двох частин: систематичної $a_0 + a_1 x$ і випадкової ε. Параметри a_0, a_1 є невідомими величинами, а ε є випадковою величиною, що має нормальний закон розподілу з числовими характеристиками: $M(\varepsilon) = 0; D(\varepsilon) = const$.

У результаті статистичних спостережень дослідник дістає характеристики для незалежної змінної x і відповідні значення залежної змінної y.

Отже, необхідно визначити параметри a_0, a_1. Але істинні значення цих параметрів дістати неможливо, оскільки ми користуємося інформацією, здобутою від вибірки обмеженого обсягу. Тому знайдені значення параметрів будуть лише статистичними оцінками істинних (невідомих нам) параметрів a_0, a_1.

На практиці найчастіше оцінки невідомих параметрів a_0, a_1 визначаються за методом найменших квадратів, розробка якого належить К. Гауссу і П. Лапласу. Цей метод почали широко застосовувати в економіко-статистичних обчисленнях, відколи була створена теорія регресії.

В основі методу найменших квадратів є принцип мінімізації суми квадратів залишків моделі.

Згідно з формулою (12.5) маємо

$$S = \sum_{k=1}^{n}\left[y_k - (a_0 + a_1 x_k)\right]^2 \tag{12.8}$$

Ця функція S неперервно диференційовна, тому згідно з необхідними умовами існування мінімуму S повинні виконуватись рівності $\frac{\partial S}{\partial a_0} = 0$ та $\frac{\partial S}{\partial a_1} = 0$.

У нашому випадку ці рівності мають вигляд:

$$\begin{cases} \sum_{k=1}^{n}\left[y_k - (a_0 + a_1 x_k)\right] = 0; \\ \sum_{k=1}^{n}\left[y_k - (a_0 + a_1 x_k)\right] \cdot x_k = 0; \end{cases}$$

Реалізація цього принципу дає можливість отримати систему нормальних рівнянь:

$$\begin{cases} n a_0 + a_1 \sum_{k=1}^{n} x_k = \sum_{k=1}^{n} y_k; \\ a_0 \sum_{k=1}^{n} x_k + a_1 \sum_{k=1}^{n} x_k^2 = \sum_{k=1}^{n} x_k y_k . \end{cases} \tag{12.9}$$

У цій системі n – кількість спостережень, $\sum_{k=1}^{n} x_k$, $\sum_{k=1}^{n} y_k$, $\sum_{k=1}^{n} x_k^2$, $\sum_{k=1}^{n} x_k y_k$ – величини, які можна розрахувати на основі вихідних спостережень над змінними Y і X.

Розв'язавши систему нормальних рівнянь, одержимо оцінки невідомих параметрів моделі $\hat{a}_0$ і $\hat{a}_1$:

$$\hat{a}_1 = \frac{\left(\sum_{k=1}^{n} x_k\right)\cdot\left(\sum_{k=1}^{n} y_k\right) - n\cdot\sum_{k=1}^{n} x_k y_k}{\left(\sum_{k=1}^{n} x_k\right)^2 - n\cdot\sum_{k=1}^{n} x_k^2}; \tag{12.10}$$

$$\hat{a}_0 = \frac{\left(\sum_{k=1}^{n} x_k\right)\cdot\left(\sum_{k=1}^{n} x_k y_k\right) - \left(\sum_{k=1}^{n} x_k^2\right)\cdot\left(\sum_{k=1}^{n} y_k\right)}{\left(\sum_{k=1}^{n} x_k\right)^2 - n\cdot\sum_{k=1}^{n} x_k^2}. \tag{12.11}$$

Тоді рівняння прямої регресії матиме вигляд:

$$\hat{Y} = \hat{a}_0 + \hat{a}_1 x. \tag{12.12}$$

Якщо кількість значень x_k та y_k велика, то обчислення параметрів $\hat{a}_0$ та $\hat{a}_1$ за формулами (12.10; 12.11) ускладнюється. Для спрощення обчислень поділимо два рівняння системи (12.9) на n. Отримаємо:

$$\begin{cases} a_0 + a_1 \dfrac{\sum_{k=1}^{n} x_k}{n} = \dfrac{\sum_{k=1}^{n} y_k}{n}; \\ a_0 \dfrac{\sum_{k=1}^{n} x_k}{n} + a_1 \dfrac{\sum_{k=1}^{n} x_k^2}{n} = \dfrac{\sum_{k=1}^{n} x_k y_k}{n}; \end{cases}$$

або

$$\begin{cases} a_0 + a_1 \bar{x} = \bar{y}; \\ a_0 \bar{x} + a_1 \overline{x^2} = \overline{xy}. \end{cases} \tag{12.13}$$

Розв’язавши систему рівнянь (12.13), отримали:

$$\hat{a}_1 = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2}; \tag{12.14}$$

$$\hat{a}_0 = \bar{y} - \hat{a}_1 \bar{x}. \tag{12.15}$$

Коваріацією вибірки називається величина:

$$K(X;Y) = \overline{xy} - \bar{x} \cdot \bar{y}. \tag{12.16}$$

Коефіцієнтом кореляції називають величину

$$r = \frac{K(X;Y)}{\sigma_x \cdot \sigma_y}, \tag{12.17}$$

де $\sigma_x = \overline{x^2} - (\bar{x})^2$ – середньоквадратичне відхилення змінної X, $\sigma_y = \overline{y^2} - (\bar{y})^2$ – середньоквадратичне відхилення змінної Y.

Коефіцієнт кореляції оцінює залежність між X та Y і має такі властивості:

1. $-1 \le r \le 1$ – залежність між X та Y тим сильніша, чим значення r ближче до 1. Якщо $r > 0$, то із зростанням X зростає і Y (кореляція додатна). Якщо $r < 0$, то при зростанні X величина Y у середньому спадає (кореляція від’ємна).

2. $r = \pm 1$ тоді і лише тоді, коли Y є лінійною функцією від X і навпаки.

3. Якщо статистичні змінні X та Y незалежні, то $r = 0$ ($K(X;Y) = 0$). Такі змінні X та Y, для яких $r = 0$, називають некорельованими, а для яких $r \neq 0$ – корельованими.

Коефіцієнт кореляції служить для оцінки тісноти лінійного зв'язку між випадковими змінними X та Y: чим $|r|$ ближче до 1, тим зв'язок сильніший, чим ближче $|r|$ до нуля – слабший. Від'ємний знак свідчить про обернений зв'язок, додатній – про прямий.

Коефіцієнтом детермінації називається величина R^2, яка визначається за формулою:

$$R^2 = r^2. \tag{12.18}$$

На основі коефіцієнта детермінації R^2 можна зробити висновок про ступінь значущості вимірюваного зв'язку на основі лінійної регресії. $R^2 \epsilon [0;1]$.

Оскільки коефіцієнт детермінації R^2 характеризує, якою мірою варіація залежної змінної визначається варіацією незалежної змінної, то що ближче R^2 до одиниці, то суттєвішим є зв'язок між цими змінними.

Приклад 12.1. 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. З таблиці бачимо, що зі збільшенням значень ознаки X залежна зміна Y має тенденцію до збільшення.

Тому припускаємо, що між ознаками X та Y існує лінійна функціональна залежність

$$\hat{Y} = \hat{a}_0 + \hat{a}_1 x.$$

Для визначення параметрів $\hat{a}_0$ та $\hat{a}_1$ скористаємося розрахунковою таблицею, що має такий вигляд:

№ з/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**

Скориставшись формулами (12.10; 12.11), де $n = 10$, отримаємо:

$$\hat{a}_1 = \frac{(\sum_{k=1}^{n} x_k)\cdot(\sum_{k=1}^{n} y_k) - n\cdot\sum_{k=1}^{n} x_k y_k}{(\sum_{k=1}^{n} x_k)^2 - n\cdot\sum_{k=1}^{n} x_k^2} = \frac{40\cdot 30 - 10\cdot 135{,}07}{40^2 - 10\cdot 172{,}56} = 1{,}12;$$

$$\hat{a}_0 = \frac{(\sum_{k=1}^{n} x_k)\cdot(\sum_{k=1}^{n} x_k y_k) - (\sum_{k=1}^{n} x_k^2)\cdot(\sum_{k=1}^{n} y_k)}{(\sum_{k=1}^{n} x_k)^2 - n\cdot\sum_{k=1}^{n} x_k^2} = \frac{40\cdot 135{,}07 - 172{,}56\cdot 30}{40^2 - 10\cdot 172{,}56} = -1{,}799.$$

Отже, рівняння регресії буде таким:

$$\hat{Y} = -1{,}799 + 1{,}12x.$$

Для обчислення r необхідно знайти $K(X;Y)$, σ_x, σ_y:

$$K(X;Y) = \overline{xy} - \bar{x}\cdot\bar{y} = 13{,}507 - 4\cdot 3 = 1{,}507;$$

$$\sigma_x = \sqrt{\overline{x^2} - (\bar{x})^2} = \sqrt{17{,}256 - 4^2} = \sqrt{1{,}256} = 1{,}12;$$

$$\sigma_y = \sqrt{\overline{y^2} - (\bar{y})^2} = \sqrt{10{,}824 - 3^2} = \sqrt{1{,}824} = 1{,}35;$$

$$r = \frac{K(X;Y)}{\sigma_x\cdot\sigma_y} = \frac{1{,}507}{1{,}12\cdot 1{,}35} = 0{,}996;$$

$$R^2 = r^2 = 0{,}996^2 = 0{,}992.$$

Як бачимо, коефіцієнт кореляції близький за своїм значенням до одиниці, що свідчить про те, що залежність між Xта Y є практично лінійною.

Коефіцієнт детермінації $R^2 = 0{,}992$. Це означає, що зміна обсягу прибутку підприємства на 99,2% визначається варіацією вартості основних фондів, і 0,8% – іншими випадковими факторами.

12.2. Парна нелінійна регресія

Якщо відображені на площині XOY групи точок $(x_i;\ y_i)$, розміщуються, нагадуючи деякі криві, то доцільно вважати, що між досліджуваними величинами існує нелінійна залежність. Тепер знову виникло завдання підібрати таку криву, яка б на основі методу найменших квадратів мала найменші відхилення від точок, здобутих при спостереженні, знайти її рівняння і визначити тісноту зв'язку.

Розглянемо деякі найпростіші види нелінійної кореляційної залежності.

Нехай зі зростанням однієї випадкової величини умовні середні другої зростають (спадають), досягають максимуму (мінімуму), а потім спадають (зростають). Тоді можна вважати, що між ними існує ***параболічна залежність*** виду:

$$y = a_0 + a_1 x + a_2 x^2 + \varepsilon. \tag{12.19}$$

Методом найменших квадратів на основі даних випробувань необхідно оцінити невідомі параметри a_0, a_1, a_2. Для цієї залежності формула (12.5) матиме вигляд:

$$S = \sum_{k=1}^{n} \left[y_k - (a_0 + a_1 x_k + a_2 x_k^2)\right]^2. \tag{12.20}$$

Необхідні умови існування мінімуму функції S є рівності нулю частинних похідних першого порядку: $\frac{\partial S}{\partial a_0} = 0; \frac{\partial S}{\partial a_1} = 0; \frac{\partial S}{\partial a_2} = 0.$

У нашому випадку:

$$\begin{cases} -2\sum_{k=1}^{n} \left[y_k - (a_0 + a_1 x_k + a_2 x_k^2)\right] x_k^2 = 0; \\ -2\sum_{k=1}^{n} \left[y_k - (a_0 + a_1 x_k + a_2 x_k^2)\right] x_k = 0; \\ -2\sum_{k=1}^{n} \left[y_k - (a_0 + a_1 x_k + a_2 x_k^2)\right] = 0. \end{cases}$$

Реалізація цього принципу дає можливість отримати систему нормальних рівнянь:

$$\begin{cases} a_2 \sum_{k=1}^{n} x_k^4 + a_1 \sum_{k=1}^{n} x_k^3 + a_0 \sum_{k=1}^{n} x_k^2 = \sum_{k=1}^{n} x_k^2 y_k ; \\ a_2 \sum_{k=1}^{n} x_k^3 + a_1 \sum_{k=1}^{n} x_k^2 + a_0 \sum_{k=1}^{n} x_k = \sum_{k=1}^{n} x_k y_k ; \\ a_2 \sum_{k=1}^{n} x_k^2 + a_1 \sum_{k=1}^{n} x_k + n \cdot a_0 = \sum_{k=1}^{n} y_k . \end{cases} \quad (12.21)$$

Система (12.21) є неоднорідною лінійною системою трьох рівнянь з невідомими a_0, a_1, a_2. Розв'язок цієї системи можна знайти різними методами (матричним, за правилом Крамера, методом Гаусса, а його вигляд буде громіздкий при доволі великій кількості випробувань *n*).

Розв'язавши систему нормальних рівнянь, одержимо оцінки невідомих параметрів моделі $\hat{a}_0$; $\hat{a}_1$; $\hat{a}_2$.

Тоді рівняння регресії матиме вигляд:

$$\hat{Y} = \hat{a}_0 + \hat{a}_1 x + \hat{a}_2 x^2. \quad (12.22)$$

Приклад 12.2. За даними (див. табл.) про господарську діяльність десяти підприємств побудувати парну кореляційно-регресійну модель залежності обсягів виробництва (*y* – показник) від основних виробничих фондів (*x* – фактор). Встановити форму зв'язку та математичне рівняння зв'язку шляхом побудови графіка кореляційної залежності. Знайти оцінки параметрів рівняння парної параболічної регресії $\hat{a}_0$; $\hat{a}_1$; $\hat{a}_2$.

x	2,8	4,2	5	5,4	5,8	7	8,2	9	10	10,2
y	7	9	9,8	11	10	13	12	11,4	8	6,6

Розв'язання. З таблиці бачимо, що зі збільшенням значень ознаки *X* залежна зміна *Y* має тенденцію до зростання, досягає максимуму, а потім спадає.

Тому припускаємо, що між ознаками *X* та *Y* існує параболічна функціональна залежність:

$$\hat{Y} = \hat{a}_0 + \hat{a}_1 x + \hat{a}_2 x^2.$$

Для визначення параметрів $\hat{a}_0$, $\hat{a}_1$ та $\hat{a}_2$ скористаємося розрахунковою таблицею, що має такий вигляд:

№ з/п	x_k	y_k	x_k^2	$x_k y_k$	y_k^2	x_k^3	x_k^4	$x_k^2 y_k$
1.	2,8	7	7,84	19,6	49	21,952	61,4656	54,88
2.	4,2	9	17,64	37,8	81	74,088	311,17	158,76
3.	5	9,8	25	49	96,04	125	625	245
4.	5,4	11	29,16	59,4	121	157,464	850,306	320,76
5.	5,8	10	33,64	58	100	195,112	1131,65	336,4
6.	7	13	49	91	169	343	2401	637
7.	8,2	12	67,24	98,4	144	551,368	4521,22	806,88
8.	9	11,4	81	102,6	129,96	729	6561	923,4
9.	10	8	100	80	64	1000	10000	800
10.	10,2	6,6	104,04	67,32	43,56	1061,21	10824,3	686,664
∑	**67,6**	**97,8**	**514,56**	**663,12**	**997,56**	**4258,19**	**37287,1**	**4969,74**

Підставивши у систему (12.21) дані з розрахункової таблиці, отримаємо:

$$\begin{cases} 37287{,}1a_2 + 4258{,}19a_1 + 514{,}56a_0 = 4969{,}74; \\ 4258{,}19a_2 + 514{,}56a_1 + 67{,}6a_0 = 663{,}12; \\ 514{,}56a_2 + 67{,}6a_1 + 10a_0 = 97{,}8. \end{cases}$$

Розв’язавши систему нормальних рівнянь, одержимо оцінки невідомих параметрів моделі $\hat{a}_0$; $\hat{a}_1$; $\hat{a}_2$.

$$\hat{a}_0 = -4{,}78;$$

$$\hat{a}_1 = 4{,}87;$$

$$\hat{a}_2 = -0{,}36.$$

Отже, рівняння регресії буде таким:

$$\hat{Y} = -4{,}78 + 4{,}87x - 0{,}36x^2.$$

Нехай зі зростанням однієї випадкової величини умовні середні другої спадають, але не на ту ж величину, як це буває в разі лінійної залежності, а розмір зміни ніби згасає. У такому разі можна вважати, що залежність ***гіперболічна***:

$$y = a_0 + \frac{a_1}{x} + \varepsilon. \tag{12.23}$$

Методом найменших квадратів на основі даних випробувань необхідно оцінити невідомі параметри a_0, a_1. Для цієї залежності формула (12.5) матиме вигляд:

$$S = \sum_{k=1}^{n} \left[y_k - \left(a_0 + \frac{a_1}{x_k} \right) \right]^2. \tag{12.24}$$

Необхідні умови існування мінімуму функції S є рівності нулю частинних похідних першого порядку: $\frac{\partial S}{\partial a_0} = 0; \frac{\partial S}{\partial a_1} = 0.$

У нашому випадку:

$$\begin{cases} \sum\limits_{k=1}^{n} \left[y_k - \left(a_0 + \dfrac{a_1}{x_k} \right) \right] = 0; \\ \sum\limits_{k=1}^{n} \left[y_k - (a_0 + a_1 \dfrac{1}{x_k}) \right] \cdot \dfrac{1}{x_k} = 0. \end{cases}$$

Реалізація цього принципу дає можливість отримати систему нормальних рівнянь:

$$\begin{cases} n a_0 + a_1 \sum_{k=1}^{n} \frac{1}{x_k} = \sum_{k=1}^{n} y_k; \\ a_0 \sum_{k=1}^{n} \frac{1}{x_k} + a_1 \sum_{k=1}^{n} \frac{1}{x_k^2} = \sum_{k=1}^{n} \frac{y_k}{x_k}. \end{cases} \tag{12.25}$$

Система (12.25) є неоднорідною лінійною системою двох рівнянь з невідомими $a_0,\ a_1$.

Розв’язавши систему нормальних рівнянь, одержимо оцінки невідомих параметрів моделі $\hat{a}_0; \hat{a}_1$.

Тоді рівняння регресії матиме вигляд:

$$\hat{Y} = \hat{a}_0 + \frac{\hat{a}_1}{x}. \tag{12.26}$$

Приклад 12.3. Підприємства легкої промисловості регіону отримали інформацію, що характеризує залежність обсягу випуску продукції (Y, млн грн) від обсягу капіталовкладень (X, млн грн). Установити форму залежності між X і Y, знайти рівняння регресії і оцінити тісноту зв'язку.

x_k	3	7	7	10	12	14	17	20	21	22
y_k	13	15	19	22	21	20	26	30	26	27

Розв'язання. Припускаємо, що між ознаками X та Y існує гіперболічна функціональна залежність

$$\hat{Y} = \hat{a}_0 + \frac{\hat{a}_1}{x}.$$

Для визначення параметрів $\hat{a}_0$ та $\hat{a}_1$ скористаємося розрахунковою таблицею, що має такий вигляд:

№ з/п	x_k	y_k	$\frac{1}{x_k}$	$\frac{1}{x_k^2}$	$\frac{y_k}{x_k}$	y_k^2
1.	3	13	0,333333	0,111111	4,333333	169
2.	7	15	0,142857	0,020408	2,142857	225
3.	7	19	0,142857	0,020408	2,714286	361
4.	10	22	0,1	0,01	2,2	484
5.	12	21	0,083333	0,006944	1,75	441
6.	14	20	0,071429	0,005102	1,428571	400
7.	17	26	0,058824	0,00346	1,529412	676
8.	20	30	0,05	0,0025	1,5	900
9.	21	26	0,047619	0,002268	1,238095	676
10.	22	27	0,045455	0,002066	1,227273	729
∑	133	219	1,075707	0,184268	20,06383	5061

Скориставшись формулами (12.25), отримаємо систему нормальних рівнянь:

$$\begin{cases} 10a_0 + 1{,}075707a_1 = 219; \\ 1{,}075707a_0 + 0{,}184268a_1 = 20{,}06383. \end{cases}$$

Розв’язавши цю систему нормальних рівнянь, одержимо оцінки невідомих параметрів моделі $\hat{a}_0$; $\hat{a}_1$.

$$\hat{a}_0 = 27{,}38;$$

$$\hat{a}_1 = -50{,}97.$$

Отже, рівняння регресії буде таким:

$$\hat{Y} = 27{,}38 - \frac{50{,}97}{x.}.$$

Для обчислення r необхідно знайти $K(X;Y)$, σ_x, σ_y:

$$K(X;Y) = \overline{\left(\frac{y}{x}\right)} - \frac{1}{\bar{x}} \cdot \bar{y} = 2{,}006383 - 0{,}1075707 \cdot 21{,}9 = -0{,}3494;$$

$$\sigma_x = \sqrt{\overline{\left(\frac{1}{x^2}\right)} - \left(\overline{\frac{1}{x}}\right)^2} = \sqrt{0{,}184268 - 0{,}1075707^2} = \sqrt{0{,}0068553} = 0{,}082797;$$

$$\sigma_y = \sqrt{\overline{y^2} - (\bar{y})^2} = \sqrt{506{,}1 - 21{,}9^2} = \sqrt{26{,}49} = 5{,}146844;$$

$$r = \frac{K(X;Y)}{\sigma_x \cdot \sigma_y} = \frac{-0{,}3494}{0{,}082797 \cdot 5{,}146844} \approx -0{,}82.$$

Як бачимо, коефіцієнт кореляції $|r|$ близький до одиниці, що свідчить про те, що зв’язок між Xта Y є тісним.

Нехай зі зростанням однієї випадкової величини умовні середні другої зростають (спадають), але не на ту ж величину, як це буває в разі лінійної залежності, а розмір зміни ніби прискорюється (згасає). У такому разі можна вважати, що залежність ***показникова***:

$$y = a_0 \cdot {a_1}^x + \varepsilon. \qquad (12.27)$$

Для оцінки невідомих параметрів a_0, a_1 рівняння показникової регресії, насамперед зведемо її до лінійного вигляду прологарифмувавши ліву і праву частину рівняння (12.27).

$$\ln y = \ln(a_0 \cdot {a_1}^x) \quad \Rightarrow \quad \ln y = \ln a_0 + x \ln a_1.$$

Введемо нові змінні $y' = \ln y, \quad b_0 = \ln a_0; \;\; b_1 = \ln a_1$. Тоді рівняння регресії матиме вигляд:

$$y' = b_0 + b_1 \cdot x. \tag{12.28}$$

Методом найменших квадратів оцінимо невідомі параметри b_0, b_1, розв'язавши систему нормальних рівнянь:

$$\begin{cases} nb_0 + b_1 \sum_{k=1}^{n} x_k = \sum_{k=1}^{n} y'_k; \\ b_0 \sum_{k=1}^{n} x_k \; + b_1 \sum_{k=1}^{n} x_k^2 = \sum_{k=1}^{n} x_k y'_k \,. \end{cases} \tag{12.29}$$

Розв'язавши систему нормальних рівнянь (12.29), одержимо оцінки невідомих параметрів моделі $\hat{b}_0$; $\hat{b}_1$.

Оцінки $\hat{a}_0$ та $\hat{a}_1$ невідомих параметрів здійснимо за формулами:

$$\hat{a}_0 = e^{\hat{b}_0}; \quad \hat{a}_1 = e^{\hat{b}_1}.$$

Тоді рівняння регресії матиме вигляд:

$$\hat{Y} = \hat{a}_0 \cdot \hat{a}_1{}^x. \tag{12.30}$$

Приклад 12.4. Дослідити виробничий процес у регіоні (табл.) за допомогою класичної виробничої моделі, що описує залежність між обсягом валової продукції (y) та обсягом основного капіталу (x). Установити форму залежності між X і Y, знайти рівняння регресії і оцінити тісноту зв'язку.

$\boldsymbol{x_k}$	60	64	72	76	78	82	84	92	94	98
$\boldsymbol{y_k}$	6,4	6,8	7,8	10,4	15,4	25,4	30,4	40,8	50,4	56,2

Розв'язання. Припускаємо, що між ознаками X та Y існує показникові функціональна залежність

$$\hat{Y} = \hat{a}_0 \cdot \hat{a}_1{}^x.$$

Прологарифмуємо ліву і праву частину рівняння та введемо додаткові позначення.

$$\ln y = \ln(a_0 \cdot a_1{}^x) \quad \Rightarrow \qquad \ln y = \ln a_0 + x \ln a_1$$

$y' = \ln y, \;\; b_0 = \ln a_0; \;\; b_1 = \ln a_1.$

Тоді рівняння регресії матиме вигляд:

$$y' = b_0 + b_1 \cdot x.$$

Для визначення параметрів b_0 та b_1 скористаємося розрахунковою таблицею, що має такий вигляд:

№ з/п	x_k	y_k	$y'_k = \ln y_k$	x_k^2	$y'_k \cdot x_k$	$(y'_k)^2$
1.	60	6,4	1,856298	3600	111,3779	3,445842
2.	64	6,8	1,916923	4096	122,683	3,674592
3.	72	7,8	2,054124	5184	147,8969	4,219424
4.	76	10,4	2,341806	5776	177,9772	5,484054
5.	78	15,4	2,734368	6084	213,2807	7,476766
6.	82	25,4	3,234749	6724	265,2494	10,4636
7.	84	30,4	3,414443	7056	286,8132	11,65842
8.	92	40,8	3,708682	8464	341,1988	13,75432
9.	94	50,4	3,919991	8836	368,4792	15,36633
10.	98	56,2	4,028917	9604	394,8338	16,23217
∑	800	250	29,2103	65424	2429,79	91,77552

$$b_1 = \frac{(\sum_{k=1}^{n} x_k)\cdot(\sum_{k=1}^{n} y'_k) - n\cdot\sum_{k=1}^{n} x_k y'_k}{(\sum_{k=1}^{n} x_k)^2 - n\cdot\sum_{k=1}^{n} x_k^2} = \frac{800\cdot 29{,}2103 - 10\cdot 2429{,}79}{800^2 - 10\cdot 65424} = 0{,}065285;$$

$$b_0 = \frac{(\sum_{k=1}^{n} x_k)\cdot(\sum_{k=1}^{n} x_k y'_k) - (\sum_{k=1}^{n} x_k^2)\cdot(\sum_{k=1}^{n} y'_k)}{(\sum_{k=1}^{n} x_k)^2 - n\cdot\sum_{k=1}^{n} x_k^2} = \frac{800\cdot 2429{,}79 - 65424\cdot 29{,}2103}{800^2 - 10\cdot 65424} = -2{,}301779.$$

Знайдемо оцінки $\hat{a}_0$ та $\hat{a}_1$ невідомих параметрів, використовуючи такі формули:

$$\hat{a}_0 = e^{b_0} = e^{-2{,}301779} = 0{,}1001;$$

$$\hat{a}_1 = e^{b_1} = e^{0{,}065285} = 1{,}0675.$$

Тоді рівняння регресії матиме вигляд:

$$\hat{Y} = 0{,}1001 \cdot 1{,}0675^{x}.$$

Для оцінки тісноти зв'язку використовуємо коефіцієнт кореляції:

$$r = \frac{K(X;Y)}{\sigma_x \cdot \sigma_{y'}} = \frac{9{,}2966}{11{,}93314711 \cdot 0{,}803203423} = 0{,}97;$$

$$K(X;Y') = \overline{xy'} - \bar{x} \cdot \overline{y'} = 242{,}979 - 80 \cdot 2{,}92103 = 9{,}2966;$$

$$\sigma_x = \sqrt{\overline{x^2} - (\bar{x})^2} = \sqrt{6542{,}4 - 80^2} = 11{,}93314711;$$

$$\sigma_y = \sqrt{\overline{(y')^2} - (\overline{y'})^2} = \sqrt{9{,}177552 - 2{,}92103^2} = 0{,}803203423.$$

Як бачимо, коефіцієнт кореляції близький за своїм значенням до одиниці, що свідчить про те, що зв'язок між Xта Y є тісним.

12.3. Множинна лінійна регресія

На практиці здебільшого залежна змінна Y пов'язана з впливом не одного, а кількох аргументів. У цьому разі регресію називають множинною. Водночас якщо аргументи в функції регресії в першій степені, то множинна регресія називається *лінійною*, в іншому разі – *множинною нелінійною регресією*.

Лінійна множинна регресія

Визначення статистичних точкових оцінок

Розглянемо лінійну залежність Y від m аргументів $(X_1, X_2, \dots, X_m)$.

Лінійна модель у цьому разі набуває такого вигляду:

$$y_i = a_0 + a_1 x_{i1} + a_2 x_{i2} + \cdots + a_m x_{im}. \qquad (12.31)$$

Для вибірки обсягу n матимемо систему лінійних рівнянь

$$y_1 = a_0 + a_1x_{11} + a_2x_{12} + \cdots + a_mx_{1m} + \varepsilon_1;$$

$$y_2 = a_0 + a_1x_{21} + a_2x_{22} + \cdots + a_mx_{2m} + \varepsilon_2; \qquad (12.32)$$

$$y_3 = a_0 + a_1x_{31} + a_2x_{32} + \cdots + a_mx_{3m} + \varepsilon_3;$$

$$y_n = a_0 + a_1x_{n1} + a_2x_{n2} + \cdots + a_mx_{nm} + \varepsilon_n.$$

Параметри рівняння (12.31) є величинами сталими, але невідомими. Ці параметри оцінюють статистичними точковими оцінками $\hat{a}_0$; $\hat{a}_1$; $\hat{a}_2$; …;$\hat{a}_m$, які отримують шляхом обробки результатів вибірки, і є величинами випадковими.

Для визначення компонентів $\hat{a}_0$; $\hat{a}_1$; $\hat{a}_2$; …;$\hat{a}_m$ (статистичних точкових оцінок) застосовується метод найменших квадратів. Необхідно, щоб сума квадратів відхилень фактичних даних від теоретичних була мінімальною. Цю вимогу можна представити так:

$$S = \sum_{i=1}^{n}\left[y_i - (a_0 + a_1x_{i1} + a_2x_{i2} + \cdots + a_mx_{im})\right]^2. \qquad (12.33)$$

Ця функція S неперервно диференційовна, тому згідно з необхідними умовами існування мінімуму S повинні виконуватись рівності частинних похідних нулю.

У нашому випадку ці рівності мають вигляд:

$$\begin{cases} \sum_{i=1}^{n}\left[y_i - (a_0 + a_1x_{i1} + a_2x_{i2} + \cdots + a_mx_{im})\right] = 0; \\ \sum_{i=1}^{n}\left[y_i - (a_0 + a_1x_{i1} + a_2x_{i2} + \cdots + a_mx_{im})\right] \cdot x_{i1} = 0; \\ \sum_{i=1}^{n}\left[y_i - (a_0 + a_1x_{i1} + a_2x_{i2} + \cdots + a_mx_{im})\right] \cdot x_{i2} = 0; \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \sum_{i=1}^{n}\left[y_i - (a_0 + a_1x_{i1} + a_2x_{i2} + \cdots + a_mx_{im})\right] \cdot x_{im} = 0. \end{cases}$$

Реалізація цього принципу дає можливість отримати систему нормальних рівнянь:

$$\begin{cases} na_0 + a_1 \sum_{i=1}^{n} x_{i1} + a_2 \sum_{i=1}^{n} x_{i2} + \cdots + a_m \sum_{i=1}^{n} x_{im} = \sum_{i=1}^{n} y_i; \\ a_0 \sum_{i=1}^{n} x_{i1} + a_1 \sum_{i=1}^{n} x_{i1}^2 + a_2 \sum_{i=1}^{n} x_{i1}x_{i2} + \cdots + a_m \sum_{i=1}^{n} x_{i1}x_{im} = \sum_{i=1}^{n} x_{i1}y_i; \\ a_0 \sum_{i=1}^{n} x_{i2} + a_1 \sum_{i=1}^{n} x_{i1}x_{i2} + a_2 \sum_{i=1}^{n} x_{i2}^2 + \cdots + a_m \sum_{i=1}^{n} x_{i2}x_{im} = \sum_{i=1}^{n} x_{i2}y_i; \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_0 \sum_{i=1}^{n} x_{im} + a_1 \sum_{i=1}^{n} x_{i1}x_{im} + a_2 \sum_{i=1}^{n} x_{i2}x_{im} + \cdots + a_m \sum_{i=1}^{n} x_{im}^2 = \sum_{i=1}^{n} x_{im}y_i. \end{cases}$$

Розв’язавши систему нормальних рівнянь, одержимо оцінки невідомих параметрів моделі $\hat{a}_0$ $\hat{a}_1$; $\hat{a}_2$; …;$\hat{a}_m$.

Вираження для a_0 , a_1 , a_2…, a_m стає доволі складним, тому доцільно це зробити з допомогою матричної алгебри.

У векторно-матричній формі система (12.32) набуває такого вигляду:

$$Y = XA + \varepsilon, \qquad (12.34)$$

де

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{pmatrix}; \quad A = \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ \cdots \\ a_m \end{pmatrix}; \; X = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1m} \\ 1 & x_{21} & \ldots & x_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n1} & \ldots & x_{nm} \end{pmatrix}; \; \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \\ \varepsilon_n \end{pmatrix}.$$

Матрицю X розміром $n \times (m + 1)$ називають регресійною, а елементи x_{ij} цієї матриці – регресорами. Статистична оцінка буде визначатись вектором

$$\hat{A} = \begin{pmatrix} \hat{a}_0 \\ \hat{a}_1 \\ \hat{a}_2 \\ \cdots \\ \hat{a}_m \end{pmatrix}.$$

Рівняння (24) представимо у вигляді:

$$S = (Y - XA)' \cdot (Y - XA) = Y'Y - 2A'X'Y + A'X'XA, \qquad (12.35)$$

де символ штрих (′) означає операцію транспонування.

Знайдемо частинну похідну окресленого виразу за компонентами вектора A і прирівняємо до нуля:

$$\frac{\partial S}{\partial A} = -2X'Y + 2X'XA = 0.$$

Звідси, отримаємо систему рівнянь у матричній формі, якій повинен задовольняти вектор A при дотриманні вимоги:

$$X'XA = X'Y. \qquad (12.36)$$

Якщо до матриці $X'X$ існує обернена матриця $(X'X)^{-1}$, то отримаємо розв'язком системи нормальних рівнянь вектор-стовпець шуканих оцінок параметрів регресії:

$$A = (X'X)^{-1} \cdot (X'Y). \qquad (12.37)$$

На відміну від простої моделі регресії алгоритм визначення параметрів багатофакторної моделі є більш складним та трудомістким.

Приклад 12.5. На основі наведених у таблиці даних залежність обсягу отриманого прибутку підприємствами регіону від розміру основних виробничих фондів та затрат праці побудувати множинну лінійну регресійну модель.

Таблиця

Номер підприємства	Прибуток, млн грн, y	Основні фонди, млн грн, x_1	Затрати праці, млн днів, x_2
1	1,2	2,5	4,0
2	1,5	2,8	4,2
3	1,9	3,0	3,6
4	2,2	3,6	4,6
5	2,8	3,9	4,3
6	3,1	4,2	5,1
7	3,4	4,5	5,3
8	4,5	5,0	4,8
9	4,8	5,6	5,4
10	5,4	6,0	5,8

Розв'язування. Попередній аналіз вхідної інформації дає можливість зробити висновок про наявність лінійної форми зв'язку між вибраними економічними показниками:

$$\hat{y} = a_0 + a_1 x_1 + a_2 x_2,$$

де y – прибуток, млн грн; x_1 – вартість основних виробничих фондів, млн грн; x_2 – затрати праці, млн днів.

Для знаходження оцінок параметрів моделі використаємо математичний апарат матричної алгебри.

Введемо позначення:

$$A = \begin{pmatrix} \hat{a}_0 \\ \hat{a}_1 \\ \hat{a}_2 \end{pmatrix}; \qquad Y = \begin{pmatrix} 1{,}2 \\ 1{,}5 \\ 1{,}9 \\ 2{,}2 \\ 2{,}8 \\ 3{,}1 \\ 3{,}4 \\ 4{,}5 \\ 4{,}8 \\ 5{,}4 \end{pmatrix}; \qquad X = \begin{pmatrix} 1 & 2{,}5 & 4{,}0 \\ 1 & 2{,}8 & 4{,}2 \\ 1 & 3{,}0 & 3{,}6 \\ 1 & 3{,}6 & 4{,}6 \\ 1 & 3{,}9 & 4{,}3 \\ 1 & 4{,}2 & 5{,}1 \\ 1 & 4{,}5 & 5{,}3 \\ 1 & 5{,}0 & 4{,}8 \\ 1 & 5{,}6 & 5{,}4 \\ 1 & 6{,}0 & 5{,}8 \end{pmatrix}.$$

Вектор-стовпець шуканих оцінок параметрів регресії знайдемо, користуючись формулою:

$$A = (X'X)^{-1} \cdot (X'Y),$$

де X' – матриця, транспонована до матриці X.

1. Знаходимо добуток двох матриць:

$$X'X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2{,}5 & 2{,}8 & 3{,}0 & 3{,}6 & 3{,}9 & 4{,}2 & 4{,}5 & 5{,}0 & 5{,}6 & 6{,}0 \\ 4{,}0 & 4{,}2 & 3{,}6 & 4{,}6 & 4{,}3 & 5{,}1 & 5{,}3 & 4{,}8 & 5{,}4 & 5{,}8 \end{pmatrix} \cdot$$

$$\cdot \begin{pmatrix} 1 & 2{,}5 & 4{,}0 \\ 1 & 2{,}8 & 4{,}2 \\ 1 & 3{,}0 & 3{,}6 \\ 1 & 3{,}6 & 4{,}6 \\ 1 & 3{,}9 & 4{,}3 \\ 1 & 4{,}2 & 5{,}1 \\ 1 & 4{,}5 & 5{,}3 \\ 1 & 5{,}0 & 4{,}8 \\ 1 & 5{,}6 & 5{,}4 \\ 1 & 6{,}0 & 5{,}8 \end{pmatrix} = \begin{pmatrix} 10 & 14{,}1 & 47{,}1 \\ 41{,}1 & 181{,}5 & 200{,}2 \\ 47{,}1 & 200{,}2 & 226{,}2 \end{pmatrix}.$$

2. Знаходимо обернену матрицю до матриці $(X'X)$:

$$(X'X)^{-1} = \begin{pmatrix} 8{,}92 & 1{,}22 & -2{,}93 \\ 1{,}22 & 0{,}4 & -0{,}61 \\ -2{,}93 & -0{,}61 & 1{,}51 \end{pmatrix}.$$

3. Знаходимо добуток матриць X' та Y:

$$X'Y = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2{,}5 & 2{,}8 & 3{,}0 & 3{,}6 & 3{,}9 & 4{,}2 & 4{,}5 & 5{,}0 & 5{,}6 & 6{,}0 \\ 4{,}0 & 4{,}2 & 3{,}6 & 4{,}6 & 4{,}3 & 5{,}1 & 5{,}3 & 4{,}8 & 5{,}4 & 5{,}8 \end{pmatrix} \cdot \begin{pmatrix} 1{,}2 \\ 1{,}5 \\ 1{,}9 \\ 2{,}2 \\ 2{,}8 \\ 3{,}1 \\ 3{,}4 \\ 4{,}5 \\ 4{,}8 \\ 5{,}4 \end{pmatrix} =$$

$$= \begin{pmatrix} 30{,}8 \\ 141{,}84 \\ 152{,}77 \end{pmatrix}.$$

4. Знаходимо вектор-стовпець шуканих оцінок параметрів регресії:

$$A = (X'X)^{-1} \cdot (X'Y) = \begin{pmatrix} 8{,}92 & 1{,}22 & -2{,}93 \\ 1{,}22 & 0{,}4 & -0{,}61 \\ -2{,}93 & -0{,}61 & 1{,}51 \end{pmatrix} \cdot \begin{pmatrix} 30{,}8 \\ 141{,}84 \\ 152{,}77 \end{pmatrix} =$$

$$= \begin{pmatrix} -0{,}97 \\ 1{,}4 \\ -0{,}37 \end{pmatrix}.$$

Отже, нами отримано таку множинну лінійну регресійну модель:

$$\hat{y} = -0{,}97 + 1{,}4x_1 - 0{,}37x_2.$$