

Тема 6

Кореляційно-регресійний
аналіз при вивченні
причинно-наслідкових
зв'язків

План

1. Сутність кореляційно-регресійного аналізу
2. Знаходження рівняння регресії
3. Визначення щільності зв'язку
4. Непараметричні методи вивчення взаємозв'язків

1. Сутність кореляційно- регресійного аналізу

Суспільні явища та процеси знаходяться в постійному взаємному всеохоплюючому об'єктивному зв'язку. Для того, щоб зрозуміти сутність будь-якого явища чи процесу слід дослідити та розкрити його зв'язок з іншими явищами та процесами.

За змістом виділяють зв'язки, що виражаються у дії ознак-причин на ознаку-наслідків (причинно-наслідковий зв'язок) і у дії ознак-умов на ознаку-наслідок. Причинно-наслідкові зв'язки дозволяють розкрити сутність подій, які є головними, на які повинна бути звернута особлива увага дослідника. Важливий вплив на результат мають й ознаки-умови. Ознаки-причини і ознаки-умови, як правило, виступають як єдині факторні ознаки, а ознаки-наслідки — як результативні ознаки

За кількістю взаємодіючих факторів виділяють зв'язки однофакторні і багатфакторні.

При однофакторних зв'язках результативна ознака пов'язана з однією факторною, **при багатфакторних** — з двома або більшою кількістю факторних ознак

За загальним напрямом зв'язки можуть бути прямі і зворотні.

При прямих зв'язках із збільшенням ознаки х збільшується й ознака у, при зворотних — при збільшенні ознаки х ознака у зменшується

Зв'язки бувають прямолінійні і криволінійні. Під прямолінійним зв'язком розуміють зв'язок, який можна описати рівнянням прямої; криволінійним називають зв'язок, який можна описати рівнянням будь-якої кривої (окрім прямої).

За характером залежності явищ розрізняють функціональні і кореляційні (стохастичні) зв'язки.

Функціональні — це такі зв'язки, коли зміні факторної ознаки на одиницю відповідає зміна результативної ознаки на суворо визначену величину.

Наприклад, при збільшенні радіуса кола на 1 см довжина кола завжди збільшується на 6,28 см, адже вона визначається за формулою: $l = 2\pi r$

У випадку функціональної залежності наявне однозначне відображення множини A в множині B . Множина A називається областю визначення функції, а B – множиною значень функції. Якщо y_i – відображення x_i , причому y_i – елемент множини B , а x_i – елемент множини A , то це записується у вигляді рівності $y = f(x)$, y_i називається значенням функції в точці x_i . Наведене рівняння відображає правило відповідності незалежної змінної x залежній змінній y .

Таким чином, функціональний зв'язок можна представити наступним рівнянням: $y_i = f(x_i)$, (7.1) де y_i — результативна ознака; x_i — факторна ознака; $f(x_i)$ — невідомий функціональний зв'язок цих ознак.

Кореляційні зв'язки – це такі зв'язки, коли при одному і тому ж значенні факторної ознаки зустрічаються різні значення результативної ознаки; при цьому, між ними існує таке співвідношення, що певній зміні факторної ознаки відповідають середні зміни результативної ознаки. Таким чином, якщо випадкові змінні причинно обумовлені і можна у імовірнісному розумінні говорити про їх зв'язок, то існує кореляційний (стохастичний) зв'язок чи кореляція.

Кореляційний зв'язок можна виразити наступним рівнянням:

$$y_i = \varphi(x_i) + \varepsilon_i, \quad (7.2)$$

де $\varphi(x_i)$ – частина результативної ознаки, яка сформулювалася під дією врахованих відомих факторних ознак (однієї або багатьох), що знаходяться у кореляційному зв'язку (φ) з ознакою y_i ; ε_i – частина результативної ознаки, яка виникає внаслідок дії другорядних і випадкових факторів.

2. Знаходження рівнянь регресії

Рівняння регресії дозволяє визначити, яким в середньому буде значення результативної ознаки (y) при тому чи іншому значенні факторної ознаки (x), якщо решту факторів, що впливають на y і не пов'язані з x , розглядати незмінними, тобто абстрагуватися від них.

Коли із зростанням (або зменшенням) значень x значення y збільшуються (або зменшуються) більш-менш рівномірно, утворюючи на графіку пряму лінію, дослідник робить припущення про наявність прямолінійної форми зв'язку.

Коли із зростанням (або зменшенням) значень x значення y збільшуються (або зменшуються) більш-менш рівномірно, утворюючи на графіку пряму лінію, дослідник робить припущення про наявність прямолінійної форми зв'язку. У цьому випадку рівняння регресії записується наступним чином:

$$\bar{y}_x = a + bx .$$

Криволінійна форма зв'язку може виражатися різними кривими, найпростіші з яких:

- 1) парабола другого порядку $\bar{y}_x = a + bx + cx^2$;
- 2) гіпербола $\bar{y}_x = a + b/x$;
- 3) показова функція $\bar{y}_x = ab^x$ (останню можна привести до лінійного вигляду $\lg \bar{y}_x = \lg a + x \lg b$);
- 4) степенева функція $\bar{y}_x = ax^b$ тощо.

У випадку наявності прямолінійної форми зв'язку між факторною ознакою x та результативною ознакою y , лінія регресії визначається функцією вигляду $y = a + bx$, параметри якої a і b знаходяться шляхом розв'язку системи нормальних рівнянь, складеної з врахуванням вимог методу найменших квадратів:

$$\begin{cases} na + b \sum x = \sum y \\ a \sum x + b \sum x^2 = \sum xy, \end{cases}$$

де n – кількість пар взаємопов'язаних величин у сукупності; $\sum x$ – сума значень факторної ознаки; $\sum x^2$ – сума квадратів значень факторної ознаки; $\sum y$ – сума значень результативної ознаки; $\sum xy$ – сума добутків значень факторної ознаки на значення результативної ознаки.

В табл. 7.1 показано розрахунок лінійного рівняння регресії між вартістю основних засобів і валовою продукцією за даними десяти підприємств (дані умовні).

Таблиця 7.1. Розрахунок лінійного рівняння регресії

Основні виробничі засоби, тис. грн. (x)	Валова продукція, тис. грн. (y)	x^2	xy	$\bar{y}_x = 1,024 + 2,12x$
1,2	2,8	1,44	3,36	1,5
1,6	4,0	2,56	6,40	2,4
2,5	3,8	6,25	9,50	4,3
3,8	6,5	14,44	24,70	7,0
4,3	8,0	18,49	34,40	8,1
5,5	10,1	30,25	55,55	10,6
6,0	9,5	36,00	57,00	11,7
8,0	12,5	64,00	100,00	15,9
9,1	18,3	82,81	166,53	18,3
10,0	24,5	100,00	245,00	20,2
$\Sigma x = 52,0$ $n = 10$	$\Sigma y = 100,0$	$\Sigma x^2 = 356,24$	$\Sigma xy = 702,44$	$\Sigma \bar{y}_x = 100,0$

За визначеними сумами будуємо систему рівнянь:

$$\begin{cases} 10a + 52b = 100 \\ 52a + 356,24b = 702,44. \end{cases}$$

Розв'язуючи систему рівнянь, знаходимо $a = -1,024$; $b = 2,12$. Таким чином, рівняння регресії матиме наступний вигляд:

$$\bar{y}_x = -1,024 + 2,12x.$$

3. Визначення щільності зв'язку

Кореляційне відношення застосовується у всіх випадках кореляційної залежності незалежно від форми зв'язку (прямолінійної чи криволінійної). Другий показник, як свідчить його назва, є мірою тісноти зв'язку лише для лінійної залежності між x і y .

Теоретичне *кореляційне відношення* – це відносна величина, яку отримують в ході порівняння міжгрупової дисперсії із загальною дисперсією. Кореляційне відношення розраховується за наступною формулою:

$$\eta^2 = \frac{\delta_{\text{міжгр.}}^2}{\sigma_{\text{заг.}}^2} \quad (7.3)$$

де, η^2 – кореляційне відношення;

$\delta_{\text{міжгр.}}^2$ – міжгрупова дисперсія;

$\sigma_{\text{заг.}}^2$ - загальна дисперсія.

Міжгрупова дисперсія вимірює варіацію результативної ознаки y за рахунок впливу тільки групувальної ознаки x .

Загальна дисперсія вимірює варіацію результативної ознаки y , зумовлену впливом всіх можливих факторів.

Міжгрупова дисперсія розраховується за наступною формулою:

$$\delta_{\text{міжгр.}}^2 = \frac{\sum_{i=1}^m (\bar{y}_i - \bar{y})^2 \times f_i}{\sum_{i=1}^m f_i}, \quad (7.4)$$

де, $\bar{y}_i$ – середнє значення результативної ознаки в окремій групі;

$\bar{y}$ – загальна середня результативної величини;

f_i – частота.

Загальна дисперсія розраховується за наступною формулою:

$$\sigma_{\text{заг.}}^2 = \bar{y}^2 - \bar{y}^2 \quad (7.5)$$

В основі обчислення кореляційного відношення лежить правило складання дисперсій, згідно якого загальна дисперсія дорівнює сумі міжгрупової дисперсії і середньої з групових дисперсій, тобто $\sigma_{\text{заг.}}^2 = \delta^2 + \bar{\sigma}_i^2$.

Середня з групових дисперсій $\bar{\sigma}_i^2$ розраховується за наступною формулою:

$$\bar{\sigma}_i^2 = \frac{\sum_{i=1}^m \sigma_{\text{гpi.}}^2 f_i}{\sum_{i=1}^m f_i}, \quad (7.6)$$

де $\bar{\sigma}_i^2$ - середня з групувальних дисперсій;

$\sigma_{\text{гpi.}}^2$ - дисперсія в окремій групі;

f_i – частота.

Інтервали кореляційних відношень:

- 1) від 0 до 0,16 – зв'язок відсутній;
- 2) від 0,16 до 0,25 – зв'язок поганий;
- 3) від 0,25 до 0,40 – зв'язок слабкий;
- 4) від 0,40 до 0,60 – зв'язок середній;
- 5) від 0,60 до 0,80 – зв'язок сильний;
- 6) від 0,80 до 0,99 – зв'язок дуже сильний;
- 7) 1 – повний взаємозв'язок.

4. Непараметричні методи вивчення взаємозв'язків

Коефіцієнт Спірмена

Так, для вимірювання тісноти залежності між величинами y та x можна використовувати досить простий показник, що називається *коефіцієнтом кореляції рангів* або *коефіцієнтом Спірмена* (ρ):

$$\rho = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}. \quad (7.14)$$

Для розрахунку цього показника значення випадкових величин x та y нумеруються (кожна окремо) в порядку зростання (або зменшення), тобто їм присвоюється відповідний ранг (N_x і N_y) – порядковий номер у ряді. Якщо зустрічається декілька однакових значень x (або y), то кожному значенню присвоюється ранг, рівний частці від ділення суми рангів, що припадає на ці значення, на число цих рівних значень.

Потім ранги окремих значень факторної ознаки зіставляють із рангами значень результативної ознаки. Різницю рангів позначимо d . Коли ранги результативної ознаки повністю збігаються із рангом факторної ознаки, то кожне значення $d = N_x - N_y = 0$ і $\sum d^2 = 0$. Відповідно у формулі Спірмена $\rho = 1$, тобто можна говорити майже про повну пряму залежність.

Якщо ранги йдуть суворо у протилежному напрямі, тобто першому рангу фактора x відповідає n -й (останній) ранг результативної ознаки y , другому рангу x відповідає $n-1$ ранг y і так далі, то у цьому випадку максимальна величина $\sum d^2$ буде дорівнювати $\frac{n(n^2-1)}{3}$ і, отже, $\frac{6\sum d^2}{n(n^2-1)}$ може мати максимальне значення 2. І тоді за формулою

Спірмена $\rho = -1$, що свідчить майже про повну зворотну залежність між x і y .

Якщо ж зв'язок між зміними x і y відсутній ($\rho = 0$), то, імовірно, в цьому випадку спостерігатиметься рівність $\sum d^2 = \frac{n(n^2-1)}{6}$.

Визначивши в кожному випадку конкретну $\sum d^2$, отримаємо ρ у межах від 0 до ± 1 .

Коефіцієнт Фехнера

Другим найпростішим непараметричним показником, що використовується для вимірювання тісноти залежності між двома корельованими показниками, є *коефіцієнт Фехнера*. Він будується на порівнянні поведінки відхилень окремих варіант від їх середньої величини за кожною ознакою. При цьому береться до уваги не величина самих відхилень, а їх знаки. Визначивши відхилення від середньої в кожному ряді, порівнюють знаки і підраховують кількість збігів та розбіжностей знаків. Якщо збіги знаків позначити символом Σ , а розбіжності – P , то тоді коефіцієнт Фехнера можна записати як:

$$K_{\Phi} = \frac{\Sigma \Sigma - \Sigma P}{\Sigma \Sigma + \Sigma P}. \quad (7.15)$$

Значення цього коефіцієнту коливається у межах від -1 до $+1$. Чим ближче до 1 , тим сильніший зв'язок. Знак “+” або “–” вказує напрям зв'язку. Якщо $\Sigma \Sigma = \Sigma P$, то $K_{\Phi} = 0$ і зв'язку немає.

Коефіцієнт асоціації

Крім рангового коефіцієнту кореляції, щільність зв'язку між атрибутивними ознаками можна вимірювати за допомогою спеціального *коефіцієнту асоціації*. В принципі, це вимірювання базується на спостереженні частоти спільної появи атрибутивних ознак; зв'язок між ними вважається тим сильніше, чим вище відносна частота такої появи.

У найпростішому випадку – наявність двох атрибутивних ознак з двома варіантами – дані можуть бути зображені в такій чотирьохклітинній комбінаційній таблиці (табл. 7.5).

Таблиця 7.5. Комбінаційна таблиця співвідношення двох атрибутивних ознак

Значення другої ознаки	Значення першої ознаки		
	1-ше	2-ге	Разом
1-ше	a	b	$a + b$
2-ге	c	d	$c + d$
Разом	$a + c$	$b + d$	n

Тут a, b, c, d – частоти ознак, що співставляються, n – загальна сума частот.

Тіснота зв'язку двох атрибутивних ознак, які мають по два варіанти, визначається за допомогою *коефіцієнту асоціації* (r_{ac}), формула якого є наступною:

$$r_{ac} = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}. \quad (7.16)$$

Цей коефіцієнт за своїм змістом і числовим значенням, який він може приймати, подібний до лінійного і рангового коефіцієнтів.

Дякую за
увагу!
Гарного
дня!