Тема 5. Статистичне вивчення взаємозв'язків між явищами

5.1.Види взаємозв'язків

Усі природні та суспільні явища взаємопов'язані. Зв'язок між багатьма з них має причинно-наслідковий характер. Ознаки, що характеризують причини та умови зв'язку, називають *факторними*. Ознаки, що характеризують наслідки зв'язку, називають *результативними*. Факторні зв'язки поділяють на функціональні та стохастичні. При *функціональному зв'язку* кожному значенню факторної ознаки *x* відповідає єдине значення результативної ознаки у. Функціональні зв'язки вивчають у математиці та природничих науках. Наприклад, зв'язок між радіусом круга та його площею є функціональним.

На відміну від функціональних, стохастичні зв'язки є неоднозначними. Наприклад, залежність рівня знань студентів від забезпеченості сучасною навчальною літературою є стохастичною. При *стохастичному зв'язку* кожному значенню ознаки *x* відповідає певна множина значень *y*, які утворюють так званий умовний розподіл. При умовному розподілі значення *y* – це значення випадкової величини. При кожному значенні *x* можна вказати ймовірності отримання певних значень *y.* Якщо умовні розподіли замінюють одним параметром – середньою, то такий зв'язок називають *кореляційним*. Кореляційний зв'язок є різновидом стохастичного зв'язку і проявляється у змінні середніх значень x та y.

За напрямком розрізняють прямі та обернені зв'язки. *Прямий зв'язок* передбачає, що зі зростанням факторної ознаки x зростає і результативна ознака у. При *оберненому зв'язку* зростання факторної ознаки супроводжується спаданням результативної ознаки.

5.2. Основи кореляційно-регресійного аналізу

Головною характеристикою кореляційного зв'язку є лінія регресії. Лінія регресії *x* на *y* – функція, що пов'язує середні значення ознаки *y* з середніми значеннями ознаки *x*. У залежності від форми лінії регресії розрізняють лінійний та нелінійний зв'язки. Лінія регресії може бути представлена таблично, графічно, аналітично. Лінія регресії є неперервною і зображується у вигляді функції $\tilde{y} = f(x)$. Цю функцію називають рівнянням регресії, а $\tilde{y}$ називають теоретичними значеннями результуючої ознаки.

Сутність кореляційно-регресійного аналізу (КРА) розглянемо на наступному прикладі. Візьмемо сукупність людей середнього віку і визначимо

для них зріст та масу тіла. Відповідні пари показників – це точки у системі координат «зріст – маса». Ці точки на координатній площині утворюють поле кореляції. Нехай «зріст» – факторна ознака, «маса» –результативна ознака. Тут кожному значенню зросту може відповідати кілька значень маси. Маємо умовний розподіл показника маси. Можна його знайти середнє значення та стандартне відхилення, що відповідає кожному значенню зросту. Коли сукупність людей є досить великою, то їх розподіл за масою є близьким до нормального. У природі масових явищ нормальний розподіл досить поширений. Тут багато прикладів можна навести з біології, коли мова йде про норму, а не патологію. Нормально розвинені люди нормально розподілені за зростом, масою, артеріальним тиском тощо. Значно рідше нормальний розподіл зустрічається при дослідженні соціально-економічних явищ.

Побудувавши поле кореляції, ми бачимо, що між показниками «зріст» та «маса» існує стохастичний кореляційний прямий зв'язок: при зростанні зросту збільшується ймовірне середнє значення маси. У нашому прикладі кореляційне поле набуває певної форми і його можна моделювати певною функцією $\tilde{y} = f(x)$, де $\tilde{y}$ – теоретичне значення результативної ознаки.

При відсутності зв'язку між ознаками кореляційне поле являє множину хаотично розкиданих точок, що не групуються біля певної лінії – лінії регресії.

Кореляційно-регресійний аналіз складається з наступних етапів:

- Вибір форми лінії регресії;
- Визначення параметрів рівняння цієї лінії;
- Оцінка тісноти зв'язку;
- Перевірка істотності зв'язку.

При виборі функції, що визначає форму лінії регресії, використовують вигляд поля кореляції. Можливий перебір функцій, коли використовують рівняння регресії різних видів і з них вибирають найкраще.

Найбільш поширеною у статистичному аналізі є лінійна функція

$$\tilde{y} = a + bx. \qquad (5.1)$$

Тут параметр *b* називають *коефіцієнтом регресії*. Він показує, на скільки одиниць власного виміру у середньому змінюється значення ознаки *y* при збільшенні ознаки *x* на одиницю. Параметр а – це значення *y* при *x*=0.

Якщо *x* за своїм змістом не може набувати нульового значення, то *a* змістовно не інтерпретується, як вільний член рівняння регресії він має лише розрахункове значення.

У деяких випадках суть явища, що моделюється, приводить до необхідності використання нелінійних рівнянь регресії, наприклад, степеневої функції $\tilde{y} = ax^b$ або гіперболи $\tilde{y} = a + \frac{b}{x}$.

Визначення параметрів рівняння регресії проводиться методом найменших квадратів. Його основною умовою є мінімізація суми квадратів відхилень емпіричних значень результативної ознаки від теоретичних. Це дає можливості отримати найкращі оцінки параметрів регресії *a* та *b*. Маємо:

$$\sum\left(y-\tilde{y}\right)^2 \to \min. \tag{5.2}$$

Для їх обчислення у випадку використання лінійної функції (лінійної регресії) складають та розв'язують систему нормальних рівнянь:

$$\begin{cases} na + b\sum x = \sum y, \\ a\sum x + b\sum x^2 = \sum xy. \end{cases} \tag{5.3}$$

Розв'язуючи цю систему, отримують значення коефіцієнтів *a* та *b:*

$$a = \frac{\sum y \cdot \sum x^2 - \sum xy \cdot \sum x}{n\sum x^2 - \left(\sum x\right)^2}, b = \frac{n\sum xy - \sum x \cdot \sum y}{n\sum x^2 - \left(\sum x\right)^2}. \tag{5.4}$$

Визначення щільності зв'язку між x та y грунтується на визначенні дисперсій. Тут обчислюють факторну дисперсію

$$\sigma_{\tilde{y}}^2 = \frac{\sum\left(\tilde{y}_i - \overline{y}\right)^2}{n}. \tag{5.5}$$

Тут $\tilde{y}_i$ – теоретичні значення результативної ознаки, обчислені за рівнянням (5.1), $\overline{y}$ – середня емпіричних значень ознаки y_i, i=1, 2,…,n, n – кількість спостережень.

Факторна дисперсія (5.5) характеризує варіацію результативної ознаки, пов'язану з варіацією факторної ознаки.

Розраховують також залишкову випадкову дисперсію:

$$\sigma_{\varepsilon}^2 = \frac{\sum\left(\tilde{y}_i - y_i\right)^2}{n}. \tag{5.6}$$

Загальна дисперсія розраховується за формулою:

$$\sigma_y^2 = \sigma_{\tilde{y}}^2 + \sigma_{\varepsilon}^2 = \frac{\sum\left(y_i - \overline{y}\right)^2}{n}. \tag{5.7}$$

Вона характеризує варіацію результативної ознаки, не пов'язану з варіацією факторної ознаки.

Мірою щільності зв'язку у КРА є коефіцієнт детермінації:

$$R^2 = \frac{\sigma_y^2}{\sigma_{\tilde{y}}^2}. \tag{5.8}$$

Індекс кореляції $R = \sqrt{R^2}$ також характеризує щільність зв'язку. Він набуває значень від 0 (за відсутності лінійного зв'язку) до 1 (зв'язок між ознаками є функціональним).

При лінійному зв'язку між ознаками використовують також лінійний коефіцієнт кореляції

$$r = \frac{n\sum xy - \sum x \cdot \sum y}{\sqrt{\left(n\sum y^2 - \left(\sum y\right)^2\right) \cdot \left(n\sum x^2 - \left(\sum x\right)^2\right)}}. \tag{5.9}$$

Перевірку істотності зв'язку у КРА здійснюють за допомогою F-критерію Фішера:

$$F_R = \frac{R^2}{1 - R^2} \cdot \frac{n - m}{m - 1}, \tag{5.10}$$

де m – кількість параметрів регресії.

Залежність між факторною та результативною ознаками у багатьох випадках можна змоделювати рівнянням двочленної гіперболічної регресії виду $\tilde{y} = a + \frac{b}{x}$ (наприклад, залежність між собівартістю одиниці продукції та обсягом її виробництва). Вона відрізняється від лінійної лише тим, що замість величини x там присутня $1/x$. Тоді система нормальних рівнянь набуває вигляду:

$$\begin{cases} na + b\sum \frac{1}{x} = \sum y, \\ a\sum \frac{1}{x} + b\sum \frac{1}{x^2} = \sum \frac{y}{x}. \end{cases} \tag{5.11}$$

Розв’язавши цю систему, отримаємо наступні вирази для параметрів *a* та *b:*

$$a = \frac{\sum y \cdot \sum \frac{1}{x^2} - \sum \frac{y}{x} \cdot \sum \frac{1}{x}}{n\sum \frac{1}{x^2} - \left(\sum \frac{1}{x}\right)^2},$$
$$b = \frac{n\sum \frac{y}{x} - \sum \frac{1}{x} \cdot \sum y}{n\sum \frac{1}{x^2} - \left(\sum \frac{1}{x}\right)^2}. \qquad (5.12)$$

Для розрахунку параметрів рівняння регресії, що має форму степеневої функції $\tilde{y} = ax^b$ необхідно привести цю функцію до лінійного вигляду шляхом логарифмування: $\lg \tilde{y} = \lg a + b \lg x$. Отримане рівняння відрізняється від рівняння (5.1) звичайної лінійної регресії лише тим, що замість $\tilde{y}, x, a$ у рівнянні присутні їхні логарифми.

Приклад 5.1. За допомогою КРА визначити наявність та характер статистичного зв’язку між ознаками «вік устаткування» та «витрати на ремонт».Вихідні дані та проміжні розрахунки наведено у таблиці 5.1.

Таблиця 5.1. Вік устаткування та витрати на ремонт для групи підприємств

№	Вік устаткування, років, x	Витрати на ремонт, тис. г.о., y	x^2	xy	$\tilde{y}$	$(y-\tilde{y})^2$	$(y-\bar{y})^2$
1	4	1,5	16	6,0	0,868	0,399	1,44
2	5	2,0	25	10,0	1,479	0,271	0,49
3	5	1,4	25	7,0	1,479	0,006	1,69
4	6	2,3	36	13,8	2,090	0,044	0,16
5	8	2,7	64	21,6	3,312	0,374	0,0
6	10	4,0	100	40,0	4,534	0,285	1,69
7	8	2,3	64	18,4	3,312	1,024	0,16
8	7	2,5	49	17,5	2,700	0,040	0,04
9	11	6,6	121	72,6	5,145	2,117	15,21
10	6	1,7	36	10,2	2,090	0,152	1,0
Разом	70	27	536	217,1	27,010	4,712	21,92

За даними таблиці обчислюємо параметри рівняння регресії:

$$a = \frac{27 \cdot 536 - 217{,}1 \cdot 70}{10 \cdot 536 - 70 \cdot 70} \approx -1{,}576;$$
$$b = \frac{10 \cdot 217{,}1 - 70 \cdot 27}{10 \cdot 536 - 70 \cdot 70} \approx 0{,}611.$$

Отже, маємо прямий зв'язок між віком устаткування та витратами на його ремонт. Лінійне рівняння регресії має вигляд:

$$\tilde{y} = -1{,}576 + 0{,}611x\,.$$

Підставляючи у це рівняння значення x, отримуємо теоретичні значення $\tilde{y}$. Залишкова дисперсія дорівнює:

$$\sigma_{\varepsilon}^2 = \frac{\sum(\tilde{y}_i - y_i)^2}{n} = \frac{4{,}712}{10} = 0{,}4712\,.$$

Загальна дисперсія дорівнює:

$$\sigma_y^2 = \frac{\sum(y_i - \overline{y})^2}{n} = \frac{21{,}92}{10} = 2{,}192\,.$$

Тоді факторну дисперсію визначаємо з формули 95.7):

$$\sigma_{\tilde{y}}^2 = 2{,}192 - 0{,}4712 = 1{,}7208\,.$$

Коефіцієнт детермінації дорівнює:

$$R^2 = \frac{1{,}7208}{2{,}192} \approx 0{,}785\,.$$

Отже, 78,5% загальної варіації витрат на ремонт устаткування залежить від варіації його віку.

Індекс кореляції $R = \sqrt{R^2} = \sqrt{0{,}785} \approx 0{,}886$ є близьким до 1, що свідчить про досить тісний прямий зв'язок між віком устаткування та витратами на його ремонт.

Для перевірки істотності індексу кореляції застосовують таблицю критичних значень F-критерію Фішера. Спочатку розрахуємо значення цього критерію:

$$F_R = \frac{R^2}{1-R^2} \cdot \frac{n-m}{m-1} = \frac{0,785}{1-0,785} \cdot \frac{10-2}{2-1} \approx 54,6 .$$

При рівні значущості $\alpha = 0,01$, $n-m=8$, $m-1=1$ табличне критичне значення F-критерію становить 11,26, що менше, ніж фактичне значення цього критерію, що становить 54,6. Таким чином, обчислений нами індекс кореляції є істотним та адекватно відображає щільність взаємозв'язку між віком устаткування та витратами на його ремонт.