Змістовий модуль 3. Застосування кореляційно-регресійного аналізу при побудові складних систем

Тема 4. Побудова кореляційних та регресійних моделей

4.1 Сутність кореляційно-регресійного моделювання об'єктів та процесів

Адекватні моделі функціонування складних систем та їх окремих складових з врахуванням взаємозв'язків між ними та зовнішнім середовищем системи можна побудувати, використовуючи методи математичної статистики, зокрема, кореляційного та регресійного аналізу. Ці розділи математичної статистики створюють можливість дослідження за даними відповідних вибірок статистичної залежності між показниками діяльності системи. При наявності статистичної залежності величини, що є об'єктом дослідження, не пов'язані між собою функціонально, проте, як випадкові величини, вони мають сумісний розподіл ймовірностей.

Дослідження залежності випадкової величини від декількох випадкових та детермінованих величин є метою регресійного аналізу, що дозволяє отримати модель регресії між цими величинами. Регресійна модель описує статистичну залежність між величинами, проте не встановлює причинно-наслідковий зв'язок між величинами. Гіпотезу про наявність причинно-наслідкового зв'язку потрібно формулювати, виходячи зі змістовної моделі об'єкту дослідження. Прикладом кореляційного зв'язку може бути статистична взаємозалежність між окремим галузями економіки або параметрами різних частин людського тіла.

За наявності регресійного зв'язку одна випадкова величина (залежна змінна Y) залежить від кількох детермінованих факторів, що описуються незалежними змінними $x' = (x_1, x_2, ..., x_n)$, а також набору випадкових величин $Z' = (z_1, z_2, ..., z_l)$: $Y = f(x', Z')$. Прикладом регресійної залежності є залежність між врожайністю певної сільськогосподарської культури та природними і економічними факторами, що впливають на неї.

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 y. При *оберненому зв'язку* зростання факторної ознаки супроводжується спаданням результативної ознаки.

Далі розглянемо особливості застосування кореляційно-регресійного аналізу для дослідження статистичної залежності між вибірковими даними.

4.2 Багатовимірні статистичні сукупності

При статистичному моделюванні багатовимірних випадкових величин розрізняють генеральну сукупність та вибіркові сукупності. *Генеральною сукупністю* називають множину всіх можливих результатів спостережень, які можна отримати при даному комплексу умов. Об'єктом дослідження у багатовимірному статистичному аналізі є випадковий вектор (випадкова точка) *k*-вимірного евклідового простору, тобто *k*-вимірна випадкова величина $\bar{X} = (X_1, X_2, ..., X_k)$.

Функцією розподілу випадкового вектора X' називають детерміновану невипадкову величину, що визначається рівністю:

$$F(x') = P(X_1 < x_1, X_2 < x_2 ...., X_k < x_k) = P(X' < x'). \quad (4.1)$$

У рівності (4.1) $x' = (x_1, x_2, ..., x_k)^T$ – *k*-вимірний дійсний фіксований вектор.

Функція розподілу має наступні властивості.

1. $F(x') = 0$, якщо серед x_j є хоча б одна компонента, що дорівнює $-\infty$;
2. $F(x') = 1$, якщо всі компоненти вектора $\bar{x}$ дорівнюють $+\infty$.
3. $F(x')$ задовольняє формулу обчислення ймовірності потрапляння випадкової точки у *k*-вимірний паралелепіпед з гранями, паралельними координатним площинам.

Розрізняють *неперервні* k-вимірні випадкові величини, всі компоненти яких є неперервними одновимірними випадковими величинами, *дискретні* k-вимірні випадкові величини, всі компоненти яких є дискретними випадковими величинами, та *змішані* k-вимірні випадкові величини, серед компонент яких є неперервні та дискретні випадкові величини.

Функція розподілу $F(x')$ для неперервної k-вимірної випадкової величини за визначенням є неперервною.

Дискретна k-вимірна випадкова величина визначається заданням ймовірностей її потрапляння у довільну точку скінченної або зліченної множини допустимих точок.

Умовними розподілами випадкового вектора X' називають розподіли підсистеми $L\,(1 \le L \le k)$ його компонент за умови, що решта $k-L$ компонент є фіксованими. Ці компоненти від нефіксованих будемо відокремлювати похилою рискою. $P[x/y]$

Моментом l-го порядку $(l = l_1 + l_2 + ... + l_k)$ випадкового вектора X' відносно сталого вектору c' називають величину

$$M_{l_1,l_2,...,l_k} = M\left[\left(X_1 - c_1\right)^{l_1}\left(X_2 - c_2\right)^{l_2}...\left(X_k - c_k\right)^{l_k}\right]. \qquad (4.2)$$

При $c' = 0'$ моменти називають *початковими*, при $c' = (MX_1, MX_2, ..., MX_k)$ – *центральними моментами*.

При $l_1 = l_2 = ... = l_k = 1$ та $c' = 0'$ з формули (4.2) отримуємо *математичне сподівання* випадкового вектора X'.

На практиці достатньо обмежитися моментами до другого порядку включно.

Далі будемо використовувати наступні позначення:

$MX_j^l = M\left(X_j\right)^l$ – початковий момент l-го порядку, математичне сподівання l-го степеня j-ої компоненти вектора X';

$M\left(X_i X_j\right), i, j = 1,2,...,k$ – початковий мішаний момент другого порядку;

$\sigma_{jj} = \sigma_j^2 = MX_j^2 - \left(MX_j\right)^2 = M\left(X_j - MX_j\right)^2$ – дисперсія j-ої компоненти вектора X', $j = 1,2,...,k$;

$\sigma_{ij} = M\left[\left(X_i - MX_i\right)\left(X_j - MX_j\right)\right] = M\left(X_i X_j\right) - M\left(X_i\right)M\left(X_j\right)$ – центральний мішаний момент другого порядку, коефіцієнт коваріації i-ої та j-ої компонент вектора X', $i, j = 1,2,...,k$.

Коваріаційною матрицею $[\Sigma]$ випадкового вектора X' називають математичне сподівання добутку центрованого випадкового вектора на цей же транспонований вектор:

$$\left[\Sigma\right] = M\left[\left(X' - MX'\right)\left(X' - MX'\right)^T\right]. \quad (4.3)$$

Коваріаційна матриця $\left[\Sigma\right]$ має вигляд:

$$\left[\Sigma\right] = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1k} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2k} \\ \dots & \dots & \dots & \dots \\ \sigma_{k1} & \sigma_{k2} & \dots & \sigma_{kk} \end{pmatrix}. \quad (4.4)$$

Коваріаційна матриця є симетричною та невід'ємно визначеною.

Коефіцієнт парної кореляції визначається за формулою:

$$\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}. \quad (4.5)$$

За абсолютною величиною цей показник не перевищує одиниці. Якщо коефіцієнт парної кореляції дорівнює нулю, то компоненти X_i та X_j є незалежними, при $\left|\rho_{ij}\right| = 1$ між цими компонентами спостерігається лінійна функціональна залежність.

Матрицю

$$\left[R\right] = \begin{pmatrix} 1 & \rho_{12} & \dots & \rho_{1k} \\ \rho_{21} & 1 & \dots & \rho_{2k} \\ \dots & \dots & \dots & \dots \\ \rho_{k1} & \rho_{k2} & \dots & 1 \end{pmatrix} \quad (4.6)$$

називають *кореляційною матрицею*. Вона симетрична та невід'ємно визначена.

Квадрат коефіцієнта кореляції називають *коефіцієнтом детермінації*.

Вибірку обсягом n з k-вимірної генеральної сукупності можна подати у вигляді наступної матриці даних:

$$\left[X'\right] = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}. \quad (4.7)$$

Її рядки будемо розглядати як n незалежних реалізацій k-вимірного випадкового вектора. Отже, елементи x_{ij} матриці $[X]$ можна розглядати або як одновимірні випадкові величини, незалежні по i, або як конкретні результати спостереження – координати n точок у k-вимірному евклідовому просторі.

Оцінку початкового моменту m-го порядку l-ої компоненти випадкового вектора X' обчислюють за формулою:

$$\bar{X}_l^m = \frac{1}{n}\sum_{i=1}^{n} x_{il}^m, l = 1,2,...,k. \quad (4.8)$$

При m=1 маємо середню арифметичну.

Оцінку коваріаційної матриці $[\Sigma]$ випадкового вектора X' (матриці вибіркових дисперсій та коефіцієнтів коваріації) визначають за формулою:

$$[S] = \begin{pmatrix} s_{11} & s_{12} & \dots & s_{1k} \\ s_{21} & s_{22} & \dots & s_{2k} \\ \dots & \dots & \dots & \dots \\ s_{k1} & s_{k2} & \dots & s_{kk} \end{pmatrix}. \quad (4.9)$$

Коефіцієнти цієї матриці мають вигляд:

$$s_{lj} = \frac{1}{n}\sum_{i=1}^{n}\left(x_{il} - \bar{x}_l\right)\left(x_{ij} - \bar{x}_j\right), l, j = 1,2,...,k. \quad (4.10)$$

Тут $s_{ll} = s_l^2$ – вибіркова дисперсія l-ої компоненти випадкового вектора X', s_{lj} – вибірковий коефіцієнт коваріації l-ої та j-ої компонент вектора X'. Замість матриці [S] використовують також незміщену оцінку матриці $[\Sigma]$:

$$\left[\hat{S}\right] = \begin{pmatrix} \hat{s}_{11} & \hat{s}_{12} & \dots & \hat{s}_{1k} \\ \hat{s}_{21} & \hat{s}_{22} & \dots & \hat{s}_{2k} \\ \dots & \dots & \dots & \dots \\ \hat{s}_{k1} & \hat{s}_{k2} & \dots & \hat{s}_{kk} \end{pmatrix} = \frac{n}{n-1}[S]. \quad (4.11)$$

Оцінку кореляційної матриці [R] можна знайти за формулою:

$$[R] = \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1k} \\ r_{21} & r_{22} & \dots & r_{2k} \\ \dots & \dots & \dots & \dots \\ r_{k1} & r_{k2} & \dots & r_{kk} \end{pmatrix}. \quad (4.12)$$

Тут $r_{lj} = \frac{s_{lj}}{s_l s_j}$ – оцінка парного коефіцієнта кореляції між l-ою та j-ою компонентами вектора X'.

Оцінку кореляційної матриці можна отримати також за формулою $R = \frac{1}{n} Z^T Z,$ де матриця Z має вигляд:

$$Z = \begin{pmatrix} z_{11} & z_{12} & \dots & z_{1k} \\ z_{21} & z_{22} & \dots & z_{2k} \\ \dots & \dots & \dots & \dots \\ z_{n1} & z_{n2} & \dots & z_{nk} \end{pmatrix}, \quad z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}. \quad (4.13)$$

Для отримання оцінок параметрів умовних розподілів потрібні певним чином організовані вибірки, тобто вибірки при фіксованих значеннях частини компонент генеральної сукупності. На практиці таку вибірку можна отримати за допомогою групування даних по закріпленим значенням частини ознак, що дорівнюють їх дискретним значенням або серединним значенням областей групування.

Розглянемо деякі оцінки параметрів умовних розподілів на прикладі двовимірної генеральної сукупності.

Нехай ми отримали вибірку з генеральної сукупності (X,Y) обсягом *n*:

$$(x_1,y_1),(x_2,y_2),(x_j,y_j),...,(x_n,y_n).$$

Таблиця 4.1 відображає загальний вигляд кореляційної таблиці. У рядку *x* у порядку зростання розташовані варіанти x_k, а у стовпчику *y* – варіанти y_l. На перетині стовпця x_k та рядка y_l знаходиться частота m_{kl}, що дорівнює кількості точок вибірки з координатами (x_k,y_l). У стовпчику m_y розташовані частоти $m_l=\sum_k m_{kl}$, а у рядку *x* – частоти $m_k=\sum_l m_{kl}$. Значення *n* дорівнює сумі частот будь-якого з одновимірних рядів, *x* чи *y*.

Таблиця 4.1. Загальний вигляд кореляційної таблиці

	…	x_k	…	m_y
…	…	…	…	…
y_l	…	m_{kl}	…	m_l
…	…	…	…	…
m_x	…	m_k	…	n

Якщо зафіксувати X на величині x_k, то отримаємо одновимірний згрупований ряд зі значеннями (варіантами) $y_1,...,y_l,...,y_s$ з відповідними частотами $m_{k1},...,m_{kl},...,m_{ks}$ і, відповідно з об'ємом m_k групи x_k.

Середня арифметична та початковий момент другого порядку

$$\overline{y}_k=\frac{1}{m_k}\sum_{l=1}^{s}y_l\cdot m_{kl},\ \overline{y}_k^2=\frac{1}{m_k}\sum_{l=1}^{s}y_l^2\cdot m_{kl},$$

а також вибіркова дисперсія $s_{yk}^2=\overline{y}_k^2-\left(\overline{y}_k\right)^2$ цього умовного згрупованого ряду є точковими оцінками для умовного математичного сподівання $MY/X=x_k$ та дисперсії $DY/X=x_k$ генеральної сукупності (X,Y).

Послідовність пар $(x_1,\overline{y}_1),(x_2,\overline{y}_2),...,(x_k,\overline{y}_k),...,(x_r,\overline{y}_r)$ є оцінкою регресії Y на X. Таку регресію називають *емпіричною*.

Величина

$$s_y^2 = \frac{\sum_{k=1}^{r} s_{yk}^2 m_k}{\sum_{k=1}^{r} m_k} \tag{4.14}$$

є оцінкою залишкової дисперсії Y, а величину

$$s_{y\,регр.}^2 = \frac{\sum_{k=1}^{s} (\overline{y}_k - \overline{y})^2 m_k}{\sum_{k=1}^{s} m_k} \tag{4.15}$$

називають дисперсією емпіричної регресії Y, яка є оцінкою генеральної дисперсії регресії $DY_{регр.} = M\left[MY / x - MY\right]^2$. Виконується рівність $s_y^2 = s_{y\,регр.}^2 + \overline{s}_y^2$.

4.3 Основні поняття кореляційного аналізу

Кореляційний аналіз є одним з методів статистичного аналізу взаємозалежності кількох ознак – компонент випадкового вектора X'.

Одним з основних показників взаємозалежності двох випадкових величин є парний коефіцієнт кореляції, що є мірою лінійної статистичної залежності між цими величинами. Це ж відноситься і до частинних та сукупних коефіцієнтів кореляції. Однією з вимог, за виконання якої застосовують кореляційний аналіз, є вимога лінійності статистичного зв'язку, тобто лінійності рівняння регресії. Загалом кореляційний аналіз застосовують , коли результати спостережень можна вважати випадковими та вибраними з генеральної сукупності, розподіленої за багатовимірним нормальним законом.

Основна задача кореляційного аналізу полягає у оцінці $\frac{k(k+3)}{2}$ параметрів, що визначають нормальний закон розподілу k-вимірного вектора X', зокрема, кореляційної матриці генеральної сукупності за вибіркою з неї. Крім того, при виконанні кореляційного аналізу здійснюється оцінка коефіцієнтів рівнянь регресії.

Для генеральної сукупності з двома ознаками, X та Y, її розподіл визначається 5 параметрами: $MX, MY, DX, DY, \rho = M\left[\frac{X-MX}{\sigma_X}\cdot\frac{Y-MY}{\sigma_Y}\right]$. Знаючи ці параметри, можна отримати рівняння ліній регресії, що описують зміну умовних

математичних сподівань у залежності від зміни відповідних значень випадкових аргументів:

$MY/X - MY = \beta_{YX}\left(X - MX\right)$ – пряма регресії Y на X;

$MX/Y - MX = \beta_{XY}\left(Y - MY\right)$ – пряма регресії X на Y;

$\beta_{YX} = \rho\dfrac{\sigma_Y}{\sigma_X}$ – коефіцієнт регресії Y на X;

$\beta_{XY} = \rho\dfrac{\sigma_X}{\sigma_Y}$ – коефіцієнт регресії X на Y.

Квадрат коефіцієнта кореляції (коефіцієнт детермінації) у кореляційній моделі характеризує частку дисперсії однієї випадкової величини, обумовленої зміною іншої.

Точкові оцінки параметрів генеральної сукупності через вибіркові параметри визначають з допомогою наведених раніше формул для обчислення вибіркових середніх та вибіркових дисперсій. Оцінка для коефіцієнта кореляції має вигляд:

$$r = \frac{\overline{xy} - \overline{x}\cdot\overline{y}}{s_x \cdot s_y}. \tag{4.16}$$

Оцінки коефіцієнтів регресії β_{YX}, β_{XY} отримаємо за формулами:

$$b_{yx} = r\cdot\frac{s_y}{s_x}, b_{xy} = r\cdot\frac{s_x}{s_y}. \tag{4.17}$$

Відповідно оцінки рівнянь регресії мають вигляд:

$$\overline{y/x} - \overline{y} = b_{yx}\left(x - \overline{x}\right), \overline{x/y} - \overline{x} = b_{xy}\left(y - \overline{y}\right). \tag{4.18}$$

У двовимірній моделі параметрами зв'язку є коефіцієнт кореляції ρ та коефіцієнти регресії β_{XY}, β_{YX}. У двовимірній моделі достатньо перевірити лише значущість коефіцієнта кореляції. Якщо він незначущий, то випадкові величини X та Y є незалежними.

Статистика *r* пов'язана зі статистикою *t*, що має розподіл Стьюдента з *n*–2 ступенями вільності, формулою:

$$t = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}. \tag{4.19}$$

Знаючи межі для *t*, що відповідають рівню значущості α, можна отримати межі для r, використавши цю формулу. Побудовані таблиці таких меж. Для перевірки гіпотези $H_0 : \rho = 0$ за обраним рівнем значущості α (звичайно рівень значущості набуває значень 0,1; 0,05; 0,02; 0,01) а також величиною $\nu = n - 2$ знаходять $r_{табл.}$.Якщо абсолютна величина обчисленого значення r перевищує $r_{табл.}$,

то гіпотеза H_0 відхиляється з ймовірністю помилки α, у противному випадку гіпотезу H_0 приймають.

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.

4.4 Парна лінійна регресія

Найбільш поширеною у статистичному аналізі є лінійна функція

$$\tilde{y} = a + bx. \tag{4.20}$$

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 $\tilde{y} = a + \frac{b}{x}$.

У подальших формулах підсумовування здійснюється за всіма значеннями ознаки, що спостерігаються у вибірці і у значку суми індекси підсумовування опущені, як це зазвичай робиться у статистичних дослідженнях.

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:

$$\sum \left(y - \tilde{y}\right)^2 \to \min. \tag{4.21}$$

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

$$\begin{cases} na + b\sum x = \sum y, \\ a\sum x + b\sum x^2 = \sum xy. \end{cases} \tag{4.22}$$

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

$$a = \frac{\sum y \cdot \sum x^2 - \sum xy \cdot \sum x}{n\sum x^2 - \left(\sum x\right)^2}, b = \frac{n\sum xy - \sum x \cdot \sum y}{n\sum x^2 - \left(\sum x\right)^2}. \tag{4.23}$$

Коефіцієнти рівняння регресії розраховують, виходячи з даних вибірки, які можуть вказувати на лінійний характер зв'язку між показниками, хоча насправді у генеральній сукупності він не існує. Тому необхідно визначити ймовірність того, що лінійний зв'язок у вибірці свідчить про такий самий зв'язок у генеральній сукупності, тобто оцінити значущість коефіцієнтів регресії. Її оцінюють з

використанням t-критерію Стьюдента з $V = n - 2$ ступенями вільності і вибраним рівнем значущості α. Розрахункове значення критерію знаходять за формулою:

$$t_{розр.} = |b| \cdot \sqrt{\frac{\sum (x - \bar{x})^2}{\sum (y - \tilde{y})^2} \cdot (n - 2)}$$

Якщо $t_{розр.} > t_{крит.}$, то коефіцієнт регресії вважається значущим, тобто має значення і у генеральній сукупності.

Незміщена стандартна похибка коефіцієнта регресії дорівнює

$$\sigma_b = \sqrt{\frac{\sum (y - \tilde{y})^2}{\sum (x - \bar{x})^2} \cdot \frac{1}{n - 2}}.$$

Враховуючи цю рівність, знаходять граничну похибку коефіцієнта регресії $\Delta_b = t_{крит.} \cdot \sigma_b$. Межі довірчого інтервалу коефіцієнта регресії становлять $b \pm \Delta_b$.

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

$$\sigma_{\tilde{y}}^2 = \frac{\sum (\tilde{y}_i - \bar{y})^2}{n}. \qquad (4.24)$$

Тут $\tilde{y}_i$ – теоретичні значення результативної ознаки, обчислені за рівнянням (4.20), $\bar{y}$ – середня емпіричних значень ознаки y_i, i=1, 2,…,n, n – кількість спостережень.

Факторна дисперсія (4.24) характеризує варіацію результативної ознаки, пов'язану з варіацією факторної ознаки.

Розраховують також залишкову випадкову дисперсію:

$$\sigma_\varepsilon^2 = \frac{\sum (\tilde{y}_i - y_i)^2}{n}. \qquad (4.25)$$

Загальна дисперсія розраховується за формулою:

$$\sigma_y^2 = \sigma_{\tilde{y}}^2 + \sigma_\varepsilon^2 = \frac{\sum (y_i - \bar{y})^2}{n}. \qquad (4.26)$$

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

Мірою щільності зв'язку у кореляційно-регресійному аналізі (КРА) є коефіцієнт детермінації:

$$R^2 = \frac{\sigma_y^2}{\sigma_{\tilde{y}}^2}. \qquad (4.27)$$

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

$$r = \frac{n\sum xy - \sum x \cdot \sum y}{\sqrt{\left(n\sum y^2 - \left(\sum y\right)^2\right) \cdot \left(n\sum x^2 - \left(\sum x\right)^2\right)}}. \tag{4.28}$$

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

$$F_R = \frac{R^2}{1-R^2} \cdot \frac{n-m}{m-1}, \tag{4.29}$$

де *m* – кількість параметрів регресії.

Залежність між факторною та результативною ознаками у багатьох випадках можна змоделювати рівнянням двочленної гіперболічної регресії виду

$$\tilde{y} = a + \frac{b}{x} \tag{4.30}$$

(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:

$$\begin{cases} na + b\sum \frac{1}{x} = \sum y, \\ a\sum \frac{1}{x} + b\sum \frac{1}{x^2} = \sum \frac{y}{x}. \end{cases} \tag{4.31}$$

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

$$\begin{aligned} a &= \frac{\sum y \cdot \sum \frac{1}{x^2} - \sum \frac{y}{x} \cdot \sum \frac{1}{x}}{n\sum \frac{1}{x^2} - \left(\sum \frac{1}{x}\right)^2}, \\ b &= \frac{n\sum \frac{y}{x} - \sum \frac{1}{x} \cdot \sum y}{n\sum \frac{1}{x^2} - \left(\sum \frac{1}{x}\right)^2}. \end{aligned} \tag{4.32}$$

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

(4.20) звичайної лінійної регресії лише тим, що замість $\tilde{y}, x, a$ у рівнянні присутні їхні логарифми.

Приклад 4.1. 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 4.2.

Розв'язання. За даними таблиці обчислюємо параметри рівняння регресії:

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

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

Таблиця 4.2. 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

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:

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

Тоді факторну дисперсію визначаємо з формули (4.7):

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

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.

Для моделювання періодичних коливань результативної ознаки під дією певних факторів використовують криві регресії у вигляді тригонометричного многочлена:

$$\tilde{y} = a_0 + \sum_{k=1}^{n} \left(a_k \cos kx + b_k \sin kx \right)$$

Такі криві використовуються, зокрема, при моделюванні сезонних явищ. Якщо є результати щомісячного спостереження за сезонним явищем протягом року, то аргумент *x* як фактор часу може набувати значень від $2\pi/12$ до 2π. Застосувавши метод найменших квадратів, можна знайти невідомі параметри рівняння регресії періодичного вигляду:

$$a_0 = \frac{1}{n}\sum y,\ a_n = \frac{2}{n}\sum y \cos kx,\ b_n = \frac{2}{n}\sum y \sin kx.$$

Оцінку значущості коефіцієнтів нелінійної регресії можна проводити аналогічно до оцінки значущості коефіцієнтів лінійного рівняння, користуючись *t*-критерієм, попередньо звівши модель до лінійної. Наприклад, для оцінки коефіцієнта a_1 періодичного рівняння регресії $\tilde{y} = a_0 + a_1 \cos x + a_2 \sin x$ проводимо заміну $\cos x = u$ і для знаходження розрахункового значення критерія використовуємо формулу

$$t_{розр.} = |a_1| \cdot \sqrt{\frac{\sum (u - \bar{u})^2}{\sum (y - \tilde{y})^2} \cdot (n-2)}.$$

4.5 Багатофакторний кореляційно-регресійний аналіз

Значення багатьох показників, що описують функціонування складної системи, формується під впливом не однієї, а кількох факторних ознак. При цьому кожна з факторних ознак окремо не чинить вирішального впливу на значення результативної ознаки, але спільний вплив факторних ознак є визначальним. Для кількісного аналізу характеру впливу декількох факторних ознак на значення результативної ознаки використовують моделі множинної кореляції та множинної регресії.

Розглянемо сутність багатофакторного кореляційно-регресійного аналізу. На його першому етапі необхідно вирішити, які саме фактори потрібно включити у модель. Необхідною умовою такого включення є наявність статистичного зв'язку між даним фактором та результативною ознакою. Помилкою є спроба врахувати при побудові моделі всі фактори, що можуть впливати на результативну ознаку. Проте, якщо між факторами існує функціональний або дуже щільний статистичний зв'язок (явище мультиколінеарності), то не потрібно включати їх у модель разом, оскільки один з них виражається через інший. Включення у модель взаємопов'язаних факторів призводить до розширення інтервалів довіри, незначущості t-статистики для оцінки параметрів моделі, труднощів обчислювального характеру, пов'язаних із розв'язанням системи нормальних рівнянь для визначення коефіцієнтів моделі. Рівняння множинної регресії об'єктивно відображає явище лише тоді, коли фактори є кореляційно незалежними, тобто мультиколінеарність відсутня.

Другим кроком кореляційно-регресійного аналізу є статистичний аналіз факторів з метою перевірки основних припущень класичної регресійної моделі.

Кількісною мірою щільності зв'язку між результативною ознакою та факторами є множинний (сукупний) коефіцієнт кореляції, який характеризує міру спільного впливу незалежних факторів $x_1, x_2, ..., x_n$ на величину результативної ознаки y.

Коефіцієнт множинної кореляції розраховують за формулою:

$$R = \frac{\sum (y - \bar{y})(y - \tilde{y})}{\sqrt{\sum (y - \bar{y})^2 (y - \tilde{y})^2}}. \qquad (4.33)$$

Значення множинного коефіцієнта кореляції може набувати значень у проміжку від 0 до 1. Чим ближче це значення до 1, тим щільнішим є множинний

зв'язок. Величину R^2 називають *множинним коефіцієнтом детермінації*. Він характеризує частку ознаки варіації результативної ознаки, що зумовлена впливом факторів, відображених у регресійній моделі. Коефіцієнт детермінації множинної регресії можна звести до вигляду:

$$R^2 = \frac{\sum\left(\tilde{y} - \overline{y}\right)^2}{\sum\left(y - \overline{y}\right)^2}. \tag{4.34}$$

Рівняння лінійної багатофакторної регресії має вигляд:

$$\tilde{y} = b_0 + b_1 x_1 + b_2 x_2 + ... + b_m x_m, \tag{4.35}$$

де m – кількість факторів $x_i, i = 1,2,...,m$.

Згідно з методом найменших квадратів можемо записати:

$$S = \sum_{i=1}^{n}\left(y_i - \left(b_0 + b_1 x_{1i} + ... + b_m x_{mi}\right)\right)^2 \to \min$$

Прирівнявши частинні похідні функції S за змінними $b_0, b_1, ..., b_m$ до нуля, отримаємо систему m+1 нормальних рівнянь з m+1 невідомими:

$$\begin{cases} mb_0 + b_1\sum x_1 + b_2\sum x_2 + ... + b_m x_m = \sum y, \\ b_0\sum x_1 + b_1\sum x_1^2 + b_2\sum x_1 x_2 + ... + b_m\sum x_1 x_m = \sum y x_1, \\ b_0\sum x_2 + b_1\sum x_1 x_2 + b_2\sum x_2^2 + ... + b_m\sum x_2 x_m = \sum y x_2, \\ ....., \\ b_0\sum x_m + b_1\sum x_1 x_m + b_2\sum x_2 x_m + ... + b_m\sum x_m^2 = \sum y x_m. \end{cases} \tag{4.36}$$

З системи (4.36) знаходимо коефіцієнти рівняння (4.35) лінійної множинної регресії.

Для оцінки значущості знайдених коефіцієнтів користуємось t-критерієм, причому

$$t_{i,розр.} = \frac{b_i}{\sigma_{b_i}}, \tag{4.37}$$

де σ_{b_i} – середня квадратична похибка i-го коефіцієнта регресії [24, 29]. На практиці для розрахунку коефіцієнтів лінійної множинної регресії та їх середніх квадратичних похибок використовують стандартну функцію ЛИНЕЙН у електронних таблицях Excel. Крім того, ця стандартна функція дає змогу розрахувати коефіцієнт детермінації, F-статистику та стандартну похибку σ_y.

Вибравши рівень значущості α, за таблицями розподілу Стьюдента для $V = n - m - 1$ ступенів вільності знаходимо $t_{крит.}$ Якщо $t_{i,розр.} > t_{крит.}$, то кофіцієнт рівняння b_i слід вважати значущим і його можна використовувати для дослідження впливу i-ої факторної ознаки на результативну ознаку. У такий спосіб відкидаємо фактори, неістотні з точки зору впливу на результативну ознаку. У результаті за

набором значущих факторів визначаємо остаточний вигляд рівняння лінійної множинної регресії, а також коефіцієнт множинної кореляції.

На підставі розрахованих значень σ_{b_i} та вибраного рівня значущості α можна знайти довірчі межі коефіцієнтів регресії у генеральній сукупності:

$$b_i - t_{\frac{\alpha}{2}}\sigma_{b_i} \le \beta_i \le b_i + t_{\frac{\alpha}{2}}\sigma_{b_i}. \quad (4.38)$$

Оцінку значущості коефіцієнта множинної кореляції можна здійснити з використанням *F*-критерію із $V_1 = m, V_2 = n - m - 1$ ступенями вільності:

$$F_{розр.} = \frac{R^2}{1 - R^2} \cdot \frac{n - m - 1}{m}. \quad (4.39)$$

Цей же критерій використовують і для оцінки адекватності регресійної моделі, тобто оцінки значущості рівняння регресії.

Гранична похибка оцінки за рівнянням множинної регресії визначається за формулою (4.40):

$$\Delta_y = t_{\frac{\alpha}{2}} \sqrt{\frac{\sum (y - \tilde{y})^2}{n - m - 1}}. \quad (4.40)$$

Крім моделей лінійної множинної регресії, при дослідженні складних систем використовують і нелінійні моделі. Наприклад, для аналізу залежності факторів виробництва від факторів, що його забезпечують, використовують статистичну залежність (мультиплікативну модель):

$$y = b_0 x_1^{b_1} x_2^{b_2} \ldots x_m^{b_m}. \quad (4.41)$$

Логарифмуючи цю функцію, отримуємо:

$$\ln y = \ln b_0 + b_1 \ln x_1 + b_2 \ln x_2 + \ldots + b_m \ln x_m.$$

Систему нормальних рівнянь та коефіцієнти даної моделі отримуємо аналогічно до випадку лінійної залежності між факторною та результативними ознаками. Прикладами мультиплікативних моделей є мультиплікативні виробничі функції.

Приклад 4.2. Згідно з даними про обсяг виробленої продукції *P*, витрати праці *L* на її виробництво, а також вартість виробничих фондів *K*, наведених у таблиці 4.3, Побудувати мультиплікативну виробничу функцію, що встановлює залежність між обсягом виробництва (результативний показник) та факторами – вартістю виробничих фондів та витратами праці. Здійснити оцінку адекватності отриманої моделі.

Розв'язання. Мультиплікативну виробничу функцію можна записати у вигляді функції

$$P = b_0 \cdot L^{b_1} \cdot K^{b_2}.$$

Таблиця 4.3. Дані про обсяги виробництва, витрати праці та вартість виробничих фондів на підприємствах галузі за рік

№	Обсяг виробництва (*P*), тис. у.г.о.	Витрати праці (L), тис. у.г.о.	Вартість виробничих фондів (К), тис. у.г.о.
1	16,207	1,426	7,905
2	16,250	1,539	7,956
3	16,091	1,002	7,704
4	16,105	1,078	7,792
5	16,211	1,499	7,911
6	16,117	1,156	7,840
7	16,301	1,697	8,165
8	16,278	1,625	8,098
9	16,144	1,215	7,853
10	16,263	1,936	7,993
11	16,186	1,355	7,896

Після логарифмування отримуємо:

$$\ln P = \ln b_0 + b_1 \ln L + b_2 \ln K .$$

Позначимо $Y = \ln P, X_1 = \ln L, X_2 = \ln K, B_0 = \ln b_0$. Отримаємо множинну лінійну регресійну модель у вигляді: $Y = B_0 + b_1 X_1 + b_2 X_2$.

Значення Y, X_1, X_2 наведені у таблиці 4.4.

Для знаходження невідомих параметрів рівняння множинної регресії скористаємось стандартною функцією ЛИНЕЙН електронних таблиць Excel. Отримаємо рівняння:

$$Y = 2,505 + 0,112 X_1 + 0,134 X_2 .$$

За допомогою цієї функції отримуємо також стандартні похибки оцінки коефіцієнтів рівняння $\sigma_{B_0} = 0,074, \sigma_{B_1} = 0,003, \sigma_{B_2} = 0,036$, множинний коефіцієнт детермінації $R^2 = 0,964$, значення *F*-статистики $F = 109,897$.

Розрахуємо *t*-статистики коефіцієнтів рівняння за формулою (4.37):

$$t_{0,розр.} = \frac{2,505}{0,074} = 33,916; t_{1,розр.} = \frac{0,012}{0,03} = 4, t_{2,розр.} = \frac{0,134}{0,036} = 3,701.$$

З таблиці *t*-розподілу при $\alpha = 0,05$ та $V = 11 - 2 - 1 = 8$ ступенів вільності знаходимо $t_{крит.} = 2,306$, що менше від розрахункових значень. Отже, всі коефіцієнти моделі є статистично значущими з ймовірністю 0,95.

Таблиця 4.4. Значення змінних Y, X_1, X_2.

№	1	2	3	4	5	6	7	8	9	10	11
Y	2,785	2,788	2,778	2,779	2,786	2,780	2,791	2,790	2,782	2,789	2,784
X_1	0,355	0,431	0,002	0,075	0,405	0,145	0,529	0,486	0,195	0,661	0,304
X_2	2,067	2,074	2,042	2,053	2,068	2,059	2,100	2,092	2,061	2,079	2,066

Обчислений коефіцієнт детермінації $R^2 = 0,964$ близький до 1, що свідчить про дуже щільний зв'язок між обсягами виробництва продукції та факторами, що їх визначають.

З таблиці *F*-розподілу знаходимо для $\alpha = 0,05$ та значень $V_1 = 2, V_2 = 8$ ступенів вільності значення $F_{крит.} = 4,46$. Оскільки $F_{розр.} = 109,897$ значно перевищує критичне значення, то можна стверджувати, що побудована модель є статистично значущою.

Враховуючи, що $b_0 = e^{2,505} = 12,238$, отримуємо виробничу функцію:

$$P = 12,238 \cdot L^{0,012} \cdot K^{0,134}.$$

Розглянуті вище регресійні моделі є вибірковими моделями, побудованими на основі даних певної вибірки. Модель, яка відображає взаємозв'язок між факторами для всієї генеральної сукупності, називають узагальненою регресійною моделлю. Вона має вигляд:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_m x_m + \varepsilon,$$

де $\beta_0, \beta_1, ..., \beta_m$ – параметри моделі, які потрібно оцінити, ε – неспостережувана випадкова величина.

У теорії математичної статистики доводиться, що математичні сподівання параметрів вибіркової лінійної регресії, розраховані методом найменших квадратів, дорівнюють параметрам узагальненої моделі для генеральної сукупності при наступних класичних припущеннях:

- випадкова величина ε є нормально розподіленою з математичним сподіванням, що дорівнює нулю, та одиничною дисперсією;
- випадкові величини (фактори) незалежні між собою;
- значення випадкової величини ε не залежить від значень факторів;

– регресійну модель визначено вірно (у модель включені лише істотні для вимірювання зв'язку незалежні змінні і вірно підібрана аналітична форма зв'язку).

Якщо яке-небудь з цих припущень не виконується, то результати дослідження можуть бути помилковими. Підкреслимо, що модель є адекватною, коли всі змінні є статистично значущими.

Розглянемо вибір оптимальної регресійної моделі, для якої виконується остання вимога. Для цього розроблено декілька підходів: метод усіх можливих регресій, метод виключення та крековий регресійний аналіз.

Перший підхід передбачає побудову кожного з усіх можливих регресійних рівнянь. Їх кількість дорівнює 2^m, де m – кількість факторів. Отримані статистично значущі моделі ранжують за значеннями коефіцієнта детермінації та стандартного відхилення залишків. Остаточний вибір моделі здійснюється на основі якісного аналізу, у зв'язку з чим вибір оптимальної моделі певною мірою є суб'єктивним. Використання методу всіх можливих регресій пов'язане зі значними витратами часу, тому його доцільно використовувати при невеликій кількості факторів, що враховуються у моделі.

В основі методу виключень та крокового аналізу знаходиться використання часткового F-критерію. Нехай S_1 – сума квадратів відхилень, обумовлена регресією, що враховує k факторів, σ_ε^2 – оцінка дисперсії залишку цієї моделі, S_2 – сума квадратів залишків, обумовлена регресією за умови включення у модель лише перших $k-1$ факторів. Величину $S = S_1 - S_2$ називають додатковою сумою квадратів. Вона порівнюється з дисперсією σ_ε^2 за допомогою F-критерію. Розглянутий варіант критерію Фішера називають *частковим F-критерієм*. Його застосовують для перевірки гіпотези, що $\beta_k = 0$. При цьому кількості ступенів вільності $V_1 = 1, V_2 = n - k - 1$.

Згідно з алгоритмом методу виключень спочатку будують регресійну модель, що містить всі фактори, далі для кожного фактору обчислюють величину часткового F-критерію, серед обчислених часткових F-критеріїв знаходять $F_{\min}$. Якщо $F_{\min} < F_{крит.}$, то фактор, якому відповідає $F_{\min}$, виключається з числа факторів моделі.

При застосуванні крокового аналізу рухаються у зворотному напрямі: модель формують послідовним включенням додаткових факторів. На першому етапі вибирають фактор, що має найбільший коефіцієнт кореляції з результативним показником, далі будують модель парної регресії, яку перевіряють на адекватність. Якщо побудоване рівняння парної регресії є незначущим, то процедура побудови моделі припиняється. У протилежному випадку знаходять наступну змінну, яка має

найбільший коефіцієнт кореляції з результативною ознакою і будують нове рівняння регресії, де враховуються вже 2 фактори. Після цього розраховують частковий *F*-критерій. За результатами тестування змінна або залишається у моделі, або замість неї вводиться нова змінна.Цей прцес продовжують до повного перегляду факторних ознак.

4.6 Непараметричні методи дослідження зв'язків між показниками діяльності складної системи

Розглянуті вище методи вимірювання взаємозв'язків між ознаками називають параметричними, оскільки вони ґрунтуються на використанні середніх величин та дисперсій, які є основними параметрами розподілу. Параметричні методи не можна застосовувати, якщо ознаки неможливо виміряти кількісно або не виконується припущення про нормальний розподіл результативної ознаки для малих сукупностей. В таких випадках використовують непараметричні методи оцінки зв'язку, які не вимагають використання числових значень ознак та обчислення параметрів розподілу. Застосування непараметричних методів надає менші можливості для дослідження взаємозв'язків між ознаками, оскільки дозволяє лише оцінити щільність зв'язку та перевірити його істотність, але не дає можливості побудови регресійної моделі.

В основі обчислення щільності зв'язку між атрибутивними (якісними) ознаками знаходиться побудова таблиці взаємного спряження (взаємозалежності) (таблиця 4.5), у якій наводяться комбінаційні розподіли сукупностей за факторною та результативною ознаками.

Таблиця 4.5. Загальний вигляд таблиці взаємного спряження

Групи за ознакою x	Групи за ознакою y						
	Група 1	Група 2	…	Група j	…	Група m_2	Разом
Група 1	f_{11}	f_{12}	…	f_{1j}	…	f_{1m_2}	f_{10}
Група 2	f_{21}	f_{22}	…	f_{2j}	…	f_{2m_2}	f_{20}
…	…	…	…	…	…	…	…
Група i	f_{i1}	f_{i2}		f_{ij}		f_{im_2}	f_{i0}
…	…	…	…	…	…	…	…
Група m_1	$f_{m_1 1}$	$f_{m_1 2}$	…	$f_{m_1 j}$	…	$f_{m_2 m_1}$	$f_{m_1 0}$
Разом	f_{01}	f_{02}	…	f_{0j}	…	$f_{0m_2 1}$	n

Величина f_{ij} – це число спостережень на перетині *i*-го рядка та *j*-го стовпця, тобто частота групи *i* у групі *j,* а f_{i0} та f_{0j} – відповідно підсумкові частоти ха ознакою *x* та ознакою *y*. У випадку відсутності стохастичної залежності між ознаками частки умовних розподілів збігаються і дорівнюють часткам безумовного розподілу (часткам розподілу по підсумковому рядку). Розбіжність між фактичною кількістю спостережень у клітинках таблиці 4.4 і теоретично можливою за повної відсутності зв'язку оцінюють за допомогою показника χ^2, який розраховують за формулою:

$$\chi^2 = n\left[\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\frac{f_{ij}^2}{f_{i0}f_{j0}} - 1\right]. \tag{4.42}$$

За відсутності зв'язку між ознаками $\chi^2 = 0$.

Для вимірювання щільності зв'язку між ознаками використовують кілька коефіцієнтів спряження. Найчастіше використовують *коефіцієнт Чупрова*. Він обчислюється за формулою:

$$K_Ч = \sqrt{\frac{\chi^2}{n\sqrt{(m_1-1)(m_2-1)}}}. \tag{4.43}$$

Тут *n* – кількість спостережень.

Якщо кількості виділених груп за кожною ознакою рівні, тобто $m_1 = m_2$ і між ознаками існує функціональний зв'язок, то коефіцієнт Чупрова дорівнює 1. Проте, якщо $m_1 \neq m_2$, то значення коефіцієнта Чупрова відмінне від 1 навіть за наявності функціонального зв'язку між ознаками.

Модифікацією коефіцієнта Чупрова є коефіцієнт Крамера:

$$K_K = \sqrt{\frac{\chi^2}{n(m-1)}}. \tag{4.44}$$

Тут $m = \min(m_1, m_2)$.

Оцінити щільність зв'язку між якісними ознаками можна також за допомогою коефіцієнта Пірсона:

$$K_П = \sqrt{\frac{\chi^2}{n+\chi^2}}. \tag{4.45}$$

Значення коефіцієнтів Чупрова, Крамера та Пірсона коливаються у межах від 0 до 1. Коефіцієнт Чупрова враховує кількість виділених груп за кожною ознакою і дає найбільш обережну оцінку щільності зв'язку. Якщо значення цього коефіцієнта $K_Ч \geq 0{,}3$, то можна говорити про помірний або щільний зв'язок між

ознаками. Перевірка істотності зв'язку здійснюється на основі χ^2-критерію з $V=(m_1-1)(m_2-1)$ ступенями вільності.

Приклад 4.3. На основі даних, наведених у таблиці 4.6, дослідити щільність зв'язку між категоріями працівників підприємства та задоволеністю рівня оплати праці.

Розв'язання. Обчислимо значення χ^2. За формулою (4.42) маємо:

$$\chi^2 = 131\left(\frac{625}{71\cdot 40}+\frac{225}{60\cdot 40}+\frac{1600}{71\cdot 80}+\frac{1600}{60\cdot 80}+\frac{36}{71\cdot 11}+\frac{25}{60\cdot 11}-1\right)=$$

$$=1,69.$$

Обчислимо значення коефіцієнта Чупрова:

$$K_Ч=\sqrt{\frac{1,69}{131\sqrt{2\cdot 1}}}=0,1.$$

Таблиця 4.6. Дані щодо задоволеності рівнем оплати праці різних категорій працівників підприємства

Група працівників	Кількість працівників		
	Задоволений оплатою праці	Незадоволений оплатою праці	Разом
Управлінський та інженерно-технічний персонал	25	15	40
Робітники	40	40	80
Допоміжний персонал	6	5	11
Разом	71	60	131

Оскільки значення цього коефіцієнта менше 0,3, то можна говорити, про дуже слабкий зв'язок між ознаками, що розглядаються. Аналогічний висновок отримуємо, використавши χ^2-критерій. Для рівня значимості $\alpha=0,05$ та $V=(3-1)(2-1)=2$ ступенів вільності з таблиць розподілу χ^2 отримуємо $\chi^2_{табл.}=5,99>1,69$, тому слід прийняти гіпотезу про відсутність зв'язку між ознаками, що розглядаються.