Лекція №3.

ТЕМА 3 ОСНОВИ РЕГРЕСІЙНОГО ТА КОРЕЛЯЦІЙНОГО АНАЛІЗУ

Вчення про кореляцію (від латинського correlatio – співвідношення, взаємозв'язок) і регресію (від латинського regressio – рух назад) широко використовується при аналізі зв'язків різних явищ. Так, наприклад, економіці проводяться дослідження: залежності обсягів виробництва від цілого ряду факторів (розміру основних фондів, їхнього віку й т.д.); залежності продуктивності праці на підприємствах від рівня механізації й електрифікації виробництва, стажу й кваліфікації робітників; залежності попиту й споживання населення від рівня доходу, цін на товари тощо. Інтенсивно застосовується регресійний аналіз при вивченні залежності врожайності певної сільськогосподарської культури від природних і економічних факторів, що впливають на неї. Широко застосовуються методи кореляційного й регресійного аналізу в психології, соціології, педагогіці та в інших галузях науки і практики.

3.1 Функціональна, статистична й кореляційна залежності

У природничих науках здебільшого мають справу з *функціональними залежностями,* у яких кожному можливому значенню аргументу *X* (незалежної змінної) відповідає одне певне значення функції *Y* (залежної змінної) (наприклад, у математиці, при вивченні фізичних законів). Проте набагато частіше в навколишньому світі існує залежність, коли кожному фіксованому значенню однієї змінної відповідає не одне, а безліч значень іншої змінної. Це пояснюється тим, що залежна змінна піддана впливу ряду неконтрольованих або неврахованих факторів, а також численних неконтрольованих випадкових факторів. У цій ситуації залежна змінна *Y* є випадковою величиною. Змінна ж *X* може бути як детермінованою (тобто такою, що приймає цілком певні значення), так і випадковою величиною. Така залежність одержала назву *статистичної* (або *стохастичної, імовірнісної).*

Припустимо, що існує стохастична залежність випадкової змінної *Y* від *X*. При $X = x$ змінна *Y* у силу її стохастичної залежності від *X* може прийняти будь-яке значення з деякої нескінченності. Середнє цієї нескінченності називають *груповим генеральним середнім* змінної *Y* при $X = x$ або умовним математичним очікуванням випадкової величини *Y,* обчисленим за умови, що $X= x$. Ця величина позначається $M(Y|X=x)$. Зокрема, якщо стохастична залежність *Y* від *X* виявляється в тому, що при зміні *x* змінюється умовне математичне очікування $M(Y|X=x)$, тоді говорять, що має місце *кореляційна залежність* величини *Y* від *X*. Якщо ж $M(Y|X=x)$ залишаються незмінними, то говорять, що кореляційна залежність величини *Y* від *X* відсутня.

Якщо існує стохастична залежність випадкової змінної *X* від *Y,* то, аналогічно попередньому, можна ввести поняття умовного математичного очікування $M(Y|X=x)$ і кореляційної залежності величини *X* від *Y*.

Таким чином, кореляційна залежність може бути представлена в такому вигляді:

$$M(Y \mid X = x) = \varphi(x); \quad (3.1)$$

$$M(Y \mid Y = y) = \varphi(y), \quad (3.2)$$

де $\varphi(x) \neq const$, $\varphi(y) \neq const$.

Рівняння (3.1) і (3.2) називаються *модельними рівняннями регресії* відповідно *Y* по *X (Y* на *X)* і *X* по *Y (X* на В), функції $\varphi(x)$ й $\varphi(y)$ – *модельними функціями регресії,* а їхні графіки – *модельними лініями регресії.*

Слід зазначити, що введені поняття стохастичної і кореляційної залежностей належать до генеральної сукупності. Як оцінки умовних математичних очікувань приймають умовні середні, які визначають за даними спостережень (за вибіркою). На практиці дослідник, як правило, розташовує лише вибіркою пар значень (x_i;y_i) обмеженого обсягу.

Умовним середнім $\bar{y}_x(\bar{x}_y)$ називають середнє арифметичне спостережуваних значень *Y(X)*, що відповідають *X = x (Y = y).* Оскільки умовні математичні очікування *M(Y | X = x)* і *M(X | Y = y)* є відповідно функціями від *x* та y, то їхні оцінки, тобто умовні середні $\bar{y}_x$ й $\bar{x}_y$ також є функціями відповідно від *x* та y ,тобто

$$\bar{y}_x = \tilde{\varphi}(x, b_0, b_1, ..., b_k) \quad (3.3)$$

$$\bar{x}_y = \tilde{\psi}(y, c_0, c_1, ..., c_k) \quad (3.4)$$

де b_0,b_1, ..., b_k і c_0, c_1, ..., c_k – параметри.

Рівняння (16.3) і (16.4) називаються *вибірковими рівняннями регресії* відповідно *Y* по *X* і *X* по *Y,* функції $\tilde{\varphi}(x, b_0, b_1, ..., b_k)$ й $\tilde{\psi}(y, c_0, c_1, ..., c_k)$ – *вибірковими функціями регресії,* а їхні графіки – *вибірковими лініями регресії.*

Зв'язок або кореляція двох змінних називається *парною.* Якщо в рівняннях (16.3) і (16.4) зі збільшенням *x* та y, змінні $\bar{y}_x$ й $\bar{x}_y$ у середньому зростають (мають тенденцію до зменшення), то така парна кореляція буде *позитивною* (*негативною*). Нульова кореляція спостерігається при відсутності зв'язку між *X* і *Y.*

Діаграма, на якій зображується сукупність значень двох ознак, називається *кореляційним полем.* Кожна точка цієї діаграми має координати x_i, y_i, що відповідають розмірам ознак в *і-му* спостереженні. Три варіанти розподілу точок на кореляційному полі показані на рис. 3.1.

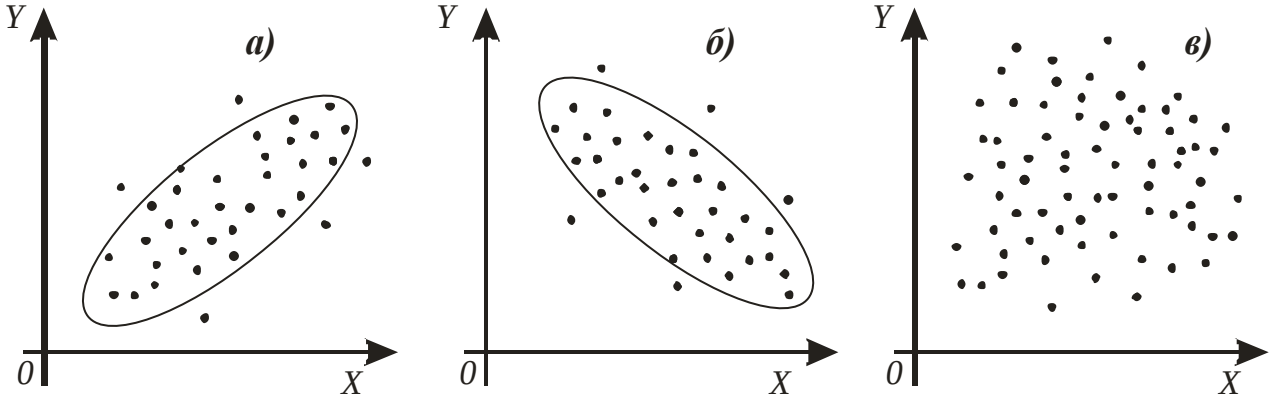

Рис. 3.1

На першому з них основна маса точок укладається в еліпс, більша вісь якого утворить позитивний кут з віссю *OX* (позитивна кореляція). Другий варіант відповідає негативній кореляції. Рівномірний розподіл точок у просторі *XY* свідчить про відсутність кореляційної залежності (рис. 3.1, в).

Статистичні зв'язки між змінними можна вивчати методами кореляційного й регресійного аналізу. Основним завданням *регресійного аналізу* є встановлення форми залежності за достовірними даними, визначення функції регресії (процес *вирівнювання)* і вивчення залежності між змінними. Основним завданням *кореляційного аналізу* є виявлення зв'язку між випадковими змінними й оцінка її тісноти.

3.2 Лінійна парна регресія

Нехай у результаті незалежних спостережень над досліджуваною системою кількісних ознак *(X;Y)*, отримана *n* пара чисел $(x_1;y_1)$, $(x_2;y_2)$,..., $(x_n;y_n)$. За даними спостережень знайдемо вибіркове рівняння прямої лінії регресії, для визначеності *Y* по *X*

$$y = \alpha + \beta x. \qquad (3.5)$$

В формулі (3.5) $\bar{y}_x$ замінене на y, бо різні значення x ознаки X і відповідні їм значення y ознаки Y спостерігалися один раз й тому групувати дані немає необхідності. Якщо позначити через $\tilde{y}_i = \alpha + \beta x$ наближене значення y_i, обчислене з рівняння регресії (4.5), то величина $y_i - \tilde{y}_i$ є відхиленням наближеного значення $\tilde{y}_i$ від точного y_i (рис. 3.2).

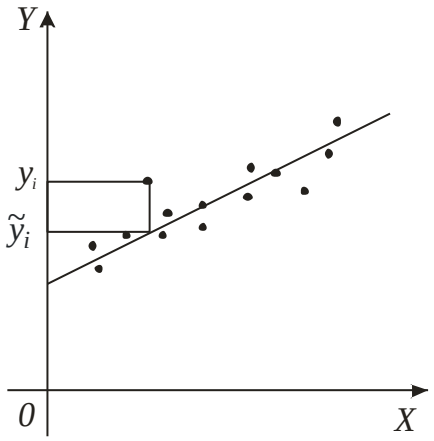

Рис. 3.2

Найбільш часте оцінювання параметрів α і β прямої регресії здійснюють на основі *методу найменших квадратів* (МНК), розробка якого належить К. Гауссу і П. Лапласу. Цей метод одержав широку область додатка в економіко-статистичних розрахунках після створення теорії регресії. Згідно з МНК параметри α і β прямої регресії вибирають так, щоб сума квадратів відхилень $y_i - \tilde{y}_i$ була мінімальної, тобто з умови мінімізації функції:

$$S(\alpha;\beta)=\sum_{i=1}^{n}(y_i-\tilde{y}_i)^2=\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)^2$$

Необхідною умовою існування мінімуму функції $S(\alpha;\beta)$ є рівність нулю часток похідних по невідомих параметрах α і β. Дорівнявши частки похідних S'_α й S'_β нулю, одержимо систему рівнянь для визначення α і β:

$$\begin{cases} S'_\alpha=\sum_{i=1}^{n}2(y_i-\alpha-\beta x_i)(-1)=0 \\ S'_\beta=\sum_{i=1}^{n}2(y_i-\alpha-\beta x_i)(-x_i)=0 \end{cases} \Rightarrow$$

$$\Rightarrow \begin{cases} \beta\sum_{i=1}^{n}x_i^2+\alpha\sum_{i=1}^{n}x_i=\sum_{i=1}^{n}x_iy_i \\ \beta\sum_{i=1}^{n}x_i+\alpha n=\sum_{i=1}^{n}y_i \end{cases} \qquad (3.6)$$

Розв’язавши цю систему, знайдемо шукані параметри:

$$\alpha=\frac{\left(\sum_{i=1}^{n}x_i^2\right)\left(\sum_{i=1}^{n}y_i\right)-\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}x_iy_i\right)}{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}$$

$$\beta = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \qquad (3.7)$$

Якщо потрібно за результатами спостережень одержати лінійне рівняння регресії X по Y, то в рівнянні регресії $y = \alpha + \beta x$ треба поміняти місцями змінні x та y. При цьому одержимо рівняння $x = \alpha' + \beta' y$, де α' і β' обчислюються за формулами:

$$\alpha' = \frac{\left(\sum_{i=1}^{n} y_i^2\right)\left(\sum_{i=1}^{n} x_i\right) - \left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i y_i\right)}{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}$$

$$\beta' = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2} \qquad (3.8)$$

Відзначимо, що регресійні прямі $y = \alpha + \beta x$ і $x = \alpha' + \beta' y$ різні. Перша пряма виходить у результаті розв'язання завдання мінімізації суми квадратів відхилень по вертикалі, а друга – при розв'язанні завдання зо мінімізації суми квадратів відхилень по горизонталі.

На практиці для знаходження рівнянь регресії складається наступна таблиця 3.1:

Таблиця 3.1

i	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	x_1	y_1	x_1^2	y_1^2	$x_1 y_1$
2	x_2	y_2	x_2^2	y_2^2	$x_2 y_2$
…	…	…	…	…	…
n	x_n	y_n	x_n^2	y_n^2	$x_n y_n$
$\sum$	$\sum_{i=1}^{n} x_i$	$\sum_{i=1}^{n} y_i$	$\sum_{i=1}^{n} x_i^2$	$\sum_{i=1}^{n} y_i^2$	$\sum_{i=1}^{n} x_i y_i$

В останньому рядку цієї таблиці суми й визначають коефіцієнти α і β або α' і β' у формулах (16.7) або (16.8) відповідно.

Приклад 1. За даними таблиці спостережень

x_i	1	1,5	2	2,5	3
y_i	2,1	2,2	2,7	2,8	2,85

скласти рівняння регресій *Y* по *X* і *X* по *Y*.

Складемо таблицю 3.2:

Таблиця 3.2

i	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	1	2,1	1	4,41	2,1
2	1,5	2,2	2,25	4,84	3,3
3	2	2,7	4	7,29	5,4
4	2,5	2,8	6,25	7,84	7
5	3	2,85	9	8,1225	8,55
$\sum$	10	12,65	22,5	32,5025	26,35

За формулами (3.7) при $n = 5$ одержуємо:

$$\alpha = \frac{22{,}5 \cdot 12{,}65 - 10 \cdot 26{,}35}{5 \cdot 22{,}5 - 100} \approx 1{,}69$$

$$\beta = \frac{5 \cdot 26{,}35 - 10 \cdot 12{,}65}{5 \cdot 22{,}5 - 100} \approx 0{,}42$$

Отже, рівняння регресії *Y* по *X* є
$y = \beta x + \alpha = 0{,}42 + 1{,}69$.

Аналогічно за формулами (3.8) знаходимо

$$\alpha' = \frac{32{,}5025 \cdot 10 - 12{,}65 \cdot 26{,}35}{5 \cdot 32{,}5025 - (12{,}65)^2} \approx -3{,}33$$

$$\beta' = \frac{5 \cdot 26{,}35 - 12{,}65 \cdot 10}{5 \cdot 32{,}5025 - (12{,}65)^2} \approx 2{,}11$$

Звідси рівняння *X* по *Y* є

$x = \beta' в + \alpha' = 2{,}11y - 3{,}33$

Якщо число вимірювань велике, то з метою спрощення розрахунків експериментальні дані потрібно групувати, тобто поєднувати в таблицю 3.3, названу *кореляційною.*

Таблиця 3.3

Y / X	y_1	y_2	…	y_k	n_{x_i}
x_1	n_{11}	n_{12}	…	n_{1k}	$n_{x_1} = \sum_{j=1}^{k} n_{1j}$

x_2	n_{21}	n_{22}	…	n_{1k}	$n_{x_2}=\sum_{j=1}^{k} n_{2j}$
…	…	…	…	…	…
x_s	n_{s1}	n_{s2}	…	n_{sk}	$n_{x_s}=\sum_{j=1}^{k} n_{sj}$
n_{y_j}	$n_{y_1}=\sum_{i=1}^{s} n_{i1}$	$n_{y_2}=\sum_{i=1}^{s} n_{i2}$	…	$n_{y_k}=\sum_{i=1}^{s} n_{ik}$	$n=\sum_{i=1}^{s}\sum_{j=1}^{k} n_{iy}$

Пояснимо, як заповнюється кореляційна таблиця. У першому стовпці (першому рядку) перераховуються у вибірці значення величини $X: x_i, i=\overline{1,s}$ $Y: y_j, j=\overline{1,k}$.

Якщо кількість різних значень величин *X і Y* велика або ці величини розподілені безкінечно, то робиться групування їхніх значень за інтервалами. У цьому випадку x_i і y_j. являють собою середини відповідних інтервалів.

На перетині рядків і стовпців указуються частоти n_{ij}, які дорівнюють числу появ у вибірці пари $(x_i;y_j)$ Якщо пари значенні ознак $(x_i;y_j)$ не спостерігалася, то у відповідному осередку таблиці ставиться риска.

В останньому рядку (останньому стовпці) вказуються числа n_{y_j}, $j=\overline{1,k}$ $(n_{x_i}, \quad i=\overline{1,s})$, що дорівнюють кількості появ у вибірці значень y_j (x_i) незалежно від того, у парі з яким зі значень величин *X(Y)* воно з'явилося.

Кореляційна таблиця містить всю інформацію, отриману в результаті вибіркових спостережень величин *X* і *Y*. Звідси з урахуванням частот появ змінних x_i і y_j, одержимо:

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{s} n_{x_i} x_i; \quad \sum_{i=1}^{n} y_i = \sum_{j=1}^{k} n_{y_j} y_j; \quad \sum_{i=1}^{n} x_1^2 = \sum_{i=1}^{s} n_{x_i} x_i^2;$$

$$\sum_{i=1}^{n} y_1^2 = \sum_{j=1}^{k} n_{y_j} y_j^2; \quad \sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{s}\sum_{j=1}^{k} n_{ij} x_i y_j$$

Підставивши ці суми у формули (16.7) і (16.8), одержимо:

$$\alpha = \frac{\left(\sum_{i=1}^{s} n_{x_i} x_i^2\right)\left(\sum_{j=1}^{k} n_{y_j} y_j\right) - \left(\sum_{i=1}^{s} n_{x_i} x_i\right)\left(\sum_{i=1}^{s}\sum_{j=1}^{k} n_{ij} x_i y_j\right)}{n\sum_{i=1}^{s} n_{x_i} x_i^2 - \left(\sum_{i=1}^{s} n_{x_i} x_i\right)^2}$$

(3.9)

$$\beta = \frac{n\sum_{i=1}^{s}\sum_{j=1}^{k} n_{ij} x_i y_j - \left(\sum_{i=1}^{s} n_{x_i} x_i\right)\left(\sum_{j=1}^{k} n_{y_j} y_j\right)}{n\sum_{i=1}^{s} n_{x_i} x_i^2 - \left(\sum_{i=1}^{s} n_{x_i} x_i\right)^2}$$

$$\alpha' = \frac{\left(\sum_{j=1}^{k} n_{y_j} y_j^2\right)\left(\sum_{i=1}^{s} n_{x_i} x_i\right) - \left(\sum_{j=1}^{k} n_{y_j} y_j\right)\left(\sum_{i=1}^{s}\sum_{j=1}^{k} n_{ij} x_i y_i\right)}{n\sum_{j=1}^{k} n_{y_j} y_j^2 - \left(\sum_{j=1}^{k} n_{y_j} y_j\right)^2} \qquad (3.10)$$

$$\beta' = \frac{n\sum_{i=1}^{s}\sum_{j=1}^{k} n_{ij} x_i y_j - \left(\sum_{j=1}^{k} n_{y_j} y_j\right)\left(\sum_{i=1}^{s} n_{x_i} x_i\right)}{n\sum_{j=1}^{k} n_{y_j} y_j^2 - \left(\sum_{j=1}^{k} n_{y_j} y_j\right)^2}$$

Як відомо, система рівнянь для визначення параметрів α і β рівняння прямої лінії регресії Y на X має вигляд (16.6). Передбачалося, що значення X *і* відповідні їм значення Y спостерігалися по одному разу. Тепер же допустимо, що отримано велику кількість даних, серед яких є повторювані, і вони згруповані у вигляді кореляційної таблиці. Запишемо систему (3.6) так, щоб вона відбивала дані кореляційної таблиці. Скористаємося тотожностями:

$$\sum_{i=1}^{n} x_i = n\bar{x}; \quad \sum_{i=1}^{n} y_i = n\bar{y}; \quad \sum_{i=1}^{n} x_1^2 = n\overline{x^2}; \quad \sum_{i=1}^{n} x_i y_i = n\overline{xy} \qquad (3.11)$$

де $\bar{x}, \bar{y}$ – вибіркові середні; $\overline{x^2}$ – вибіркове середнє квадрата; $\overline{xy}$ – вибіркове середнє добутку. Нагадаємо, що *вибіркова коваріація* s_{xy} визначається рівністю

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\cdot\bar{y} = \overline{xy} - \bar{x}\cdot\bar{y} \qquad (3.12)$$

Вибіркові дисперсії визначаються співвідношеннями:

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \overline{x^2} - \bar{x}^2$$

$$s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2 = \overline{y^2} - \bar{y}^2 \qquad (3.13)$$

За визначенням *вибірковий коефіцієнт кореляції*

$$\rho_{xy} = \frac{s_{xy}}{s_x s_y} \tag{3.14}$$

Підставивши праві частини (16.11) у систему (16.6), одержимо

$$\begin{cases} \beta n\overline{x^2} + \alpha n\bar{x} = n\overline{xy} \\ \beta n\bar{x} + \alpha n = n\bar{y} \end{cases} \Leftrightarrow \begin{cases} \overline{x^2}\beta + \bar{x}\alpha = \overline{xy} \\ \bar{x}\beta + \alpha = \bar{y} \end{cases} \tag{3.15}$$

Розв'язуючи систему (16.15) за формулами Крамера, одержимо:

$$\Delta = \begin{vmatrix} \overline{x^2} & \bar{x} \\ \bar{x} & 1 \end{vmatrix} = \overline{x^2} - \bar{x}^2 = s_x^2; \quad \Delta_\beta = \begin{vmatrix} \overline{xy} & \bar{x} \\ \bar{y} & 1 \end{vmatrix} = \overline{xy} - \bar{x}\cdot\bar{y} = s_{xy};$$

$$\Delta = \begin{vmatrix} \overline{x^2} & \overline{xy} \\ \bar{x} & \bar{y} \end{vmatrix} = \overline{x^2}\bar{y} - \bar{x}\,\overline{xy} = \overline{x^2}\bar{y} - \bar{x}^2\bar{y} + \bar{x}^2\bar{y} - \bar{x}\,\overline{xy} =$$

$$= \bar{y}\left(\overline{x^2} - \bar{x}^2\right) - \bar{x}\left(\overline{xy} - \bar{x}\cdot\bar{y}\right) = \bar{y}\cdot s_x^2 - \bar{x}\cdot s_{xy}$$

$$\alpha = \frac{\Delta_\alpha}{\Delta} = \frac{\bar{y}\cdot s_x^2 - \bar{x}\cdot s_{xy}}{s_x^2} = \bar{y} - \frac{s_{xy}}{s_x^2}\bar{x}; \quad \beta = \frac{\Delta_\beta}{\Delta} = \frac{s_{xy}}{s_x^2}$$

Підставивши коефіцієнти α і β у рівняння регресії $\bar{y} = \alpha + \beta x$, одержимо

$$\bar{y}_x = \bar{y} - \frac{s_{xy}}{s_x^2}\bar{x} + \frac{s_{xy}}{s_x^2} + \frac{s_{xy}}{s_x^2}x \Leftrightarrow \bar{y}_x - \bar{y} = \frac{s_{xy}}{s_x^2}\left(x - \bar{x}\right)$$

$$\frac{s_{xy}}{s_x^2} = \frac{s_{xy}}{s_x s_y}\cdot\frac{s_y}{s_x} = \left|(16.14)\right| = \rho_{xy}\frac{s_y}{s_x}$$

Таким чином, рівняння прямої регресії *Y* по *X* має вигляд

$$\bar{y}_x - \bar{y} = \rho_{xy}\frac{s_y}{s_x}\left(x - \bar{x}\right) \tag{3.16}$$

Аналогічно знайдемо, що рівняння прямої регресії *X* по *Y* запишеться у вигляді

$$\bar{x}_y - \bar{x} = \rho_{xy}\frac{s_x}{s_y}\left(y - \bar{y}\right) \tag{3.17}$$

З рівнянь (16.16) і (16.17) необхідно, що прямі регресії *Y* по *X* і *X* по *Y* проходять через точку $(\bar{x};\bar{y})$. Ці прямі збігаються, коли $\left|\rho_{xy}\right|^2 = 1$.

Величини $\rho_{xy}\dfrac{s_y}{s_x}$ й $\rho_{xy}\dfrac{s_x}{s_y}$ називаються *вибірковими коефіцієнтами лінійної регресії* й позначаються

$$\rho_{y|x}=\rho_{xy}\frac{s_y}{s_x};\quad \rho_{x|y}=\rho_{xy}\frac{s_x}{s_y} \tag{3.18}$$

Перемноживши праві й ліві частини рівностей (16.18), після добування кореня одержимо $\rho_{y|x}=\pm\sqrt{\rho_{y|x}\rho_{x|y}}$, тобто коефіцієнт кореляції є середнім геометричним коефіцієнтів лінійної регресії. Знак у правій частині ρ_{xy} збігається зі знаками $\rho_{y|x}$ й $\rho_{x|y}$.

Коефіцієнт регресії *Y* по *X* (*X* по *Y*) показує, на скільки одиниць у середньому змінюється змінна *Y(X)* при збільшенні змінної *X(Y)* на одну одиницю.

Якщо дані спостережень над ознаками *X* і *Y* задані у вигляді кореляційної таблиці з рівновіддаленими варіантами, то доцільно перейти до умовних варіант:

$$u_i=(x_i-C_1)/h_1;\quad v_j=(y_j-C_2)/h_2,$$

де C_1 і C_2 – хибні нулі варіант *X* і *Y* відповідно, h_1 і h_2 – кроки, тобто різниці між двома сусідніми варіантами *X* і *Y*. Тоді

$$\bar{x}=\frac{h_1}{n}\sum_{i=1}^{s}u_i n_{x_i}+C_1=h_1\bar{u}+C_1 \tag{3.19}$$

$$\bar{y}=\frac{h_2}{n}\sum_{j=1}^{k}v_j n_{y_j}+C_2=h_1\bar{v}+C_2 \tag{3.20}$$

$$s_x^2=\frac{h_1^2}{n}\sum_{i=1}^{s}u_i^2 n_{x_i}-(\bar{x}-C_1)^2=h_1^2\overline{u^2}-(\bar{x}-C_1)^2 \tag{3.21}$$

$$s_y^2=\frac{h_2^2}{n}\sum_{j=1}^{k}v_j^2 n_{y_j}-(\bar{y}-C_2)^2=h_2^2\overline{v^2}-(\bar{y}-C_2)^2 \tag{3.22}$$

Виразимо коваріацію через умовні варіанти. Маємо

$$s_{xy}=\frac{1}{n}\sum_{i=1}^{s}\sum_{j=1}^{k}x_i y_j n_{ij}-\frac{1}{n}\sum_{i=1}^{s}x_i n_{x_i}\frac{1}{n}=\frac{h_1h_2}{n}\sum_{i=1}^{s}\sum_{j=1}^{k}\frac{(x_i-C_1+C_1)}{h_1}\frac{y_j-C_2+C_2}{h_2}n_{ij}-$$

$$-\frac{h_1}{n}\sum_{i=1}^{s}\frac{(x_i-C_1+C_1)}{h_1}n_{x_i}\cdot\frac{h_2}{n}\sum_{j=1}^{k}\frac{y_j-C_2+C_2}{h_2}n_{y_j}=\frac{h_1h_2}{n}\sum_{i=1}^{s}\sum_{j=1}^{k}\left(u_i+\frac{C_1}{h_1}\right)\times$$

$$\times\left(v_j+\frac{C_2}{h_2}\right)n_{ij}-\frac{h_1h_2}{n^2}\sum_{i=1}^{s}\left(u_i+\frac{C_1}{h_1}\right)n_{x_i}\times\sum_{j=1}^{k}\left(v_j+\frac{C_2}{h_2}\right)n_{y_j}=\frac{h_1h_2}{n}\sum_{i=1}^{s}\sum_{j=1}^{k}\Big(u_iv_j+$$

$$+\frac{C_1}{h_1}v_j+\frac{C_2}{h_2}u_i+\frac{C_1C_2}{h_1h_2}\Big)n_{ij}-\frac{h_1h_2}{n^2}\left(\sum_{i=1}^{s}u_in_{x_i}+\frac{C_1}{h_1}n\right)\left(\sum_{j=1}^{k}v_jn_{y_j}+\frac{C_2}{h_2}n\right)=$$

$$=\frac{h_1h_2}{n}\sum_{i=1}^{s}\sum_{j=1}^{k}u_iv_jn_{ij}+\frac{C_1h_2}{n}\sum_{i=1}^{s}\sum_{j=1}^{k}v_jn_{ij}+\frac{C_2h_1}{n}\sum_{i=1}^{s}\sum_{j=1}^{k}u_in_{ij}+C_1C_2-\frac{h_1h_2}{n^2}\sum_{i=1}^{s}u_in_{x_i}\times$$

$$\times\sum_{j=1}^{k}v_jn_{y_j}-\frac{C_1h_2}{n}\sum_{i=1}^{s}\sum_{j=1}^{k}v_jn_{ij}-\frac{C_2h_1}{n}\sum_{i=1}^{s}\sum_{j=1}^{k}u_in_{ij}-C_1C_2=\frac{h_1h_2}{n}\sum_{i=1}^{s}\sum_{j=1}^{k}u_iv_jn_{ij}-$$

$$-\frac{h_1h_2}{n^2}\sum_{i=1}^{s}u_in_{x_i}\sum_{j=1}^{k}v_jn_{y_j}=h_1h_2\left(\overline{uv}-\bar{u}\cdot\bar{v}\right)$$

де $$n=\sum_{i=1}^{s}n_{x_i}=\sum_{j=1}^{k}n_{y_j}=\sum_{i=1}^{s}\sum_{j=1}^{k}n_{ij}$$

$$\bar{u}=\frac{1}{n}\sum_{i=1}^{s}u_in_{x_i};\quad \bar{v}=\frac{1}{n}\sum_{j=1}^{k}v_jn_{y_j};\quad \overline{uv}=\frac{1}{n}\sum_{i=1}^{s}\sum_{j=1}^{k}u_iv_jn_{ij}$$

Таким чином,

$$s_{xy}=\frac{h_1h_2}{n}\sum_{i=1}^{s}\sum_{j=1}^{k}u_iv_jn_{ij}-\frac{h_1h_2}{n^2}\sum_{i=1}^{s}u_in_{x_j}\sum_{j=1}^{k}v_jn_{y_j}\qquad(3.23)$$

Приклад 2. Знайти рівняння прямих регресії *Y* по *X* та *X* за *Y* за даними кореляційної таблиці 16.4 і побудувати їхні графіки.

Перейдемо до умовних варіант

$u_i=(x_i-C_1)/h_1=(x_i-20)/5;\qquad v_j=(y_i-C_2)/h_2=(y_i-50)/10$, де як хибні нулі C_1 і C_2 узяті варіанти відповідно x = 20 і y = 50, розташовані приблизно в середині варіаційних рядів, тобто C_1 = 20 і C_2 = 50, а h_1= 5 і h_2 = 10.

Таблиця 3.4

Y / X	30	40	50	60	70	n_{x_i}
5	2	—	—	—	—	2
10	6	5	—	—	—	11
15	—	3	7	4	—	14
20	—	—	40	9	4	53
25	—	—	2	6	7	15
30	—	—	—	—	5	5
n_{y_j}	8	8	49	19	16	n=100

Представимо кореляційну табл. 3.4 у вигляді табл. 3.5, де два передостанніх стовпці й два передостанні рядки містять наступні суми:

$$\sum_{i=1}^{6} u_i n_{x_i} = -17;\ \sum_{i=1}^{6} u_i^2 n_{x_i} = 111;\ \sum_{j=1}^{5} v_j n_{y_j} = 27;\ \sum_{j=1}^{5} v_j^2 n_{y_j} = 123;$$

Для зручності обчислення суми $\sum_{i=1}^{6}\sum_{j=1}^{5} u_i v_j n_{ij}$ спочатку розраховуємо $u_i v_j$ й проставляємо ці значення у верхньому правому кутку тих осередків, у яких $n_{ij} \neq 0$, а потім знаходимо добутки $u_i v_j n_{ij}$. Підсумовуючи їх по рядках і стовпцям, записуємо отримані результати відповідно в останньому стовпці й останньому рядку табл. 3.5. Підсумовуючи значення останніх стовпця й рядка, одержимо в правому нижньому вікні табл. 3.5 $\sum_{i=1}^{6}\sum_{j=1}^{5} u_i v_j n_{ij}$ =85; їхній збіг свідчить про правильність обчислень.

Використовуючи формули (3.19) – (16.23), одержимо:

$$\bar{x} = 5 \cdot (-17)/100 + 20 = 19{,}15; \qquad \bar{y} = 10 \cdot 27/100 + 50 = 52{,}7;$$

$$s_x^2 = 25 \cdot 111/100 - (19{,}15 - 20)^2 = 27{,}0275; \qquad s_x \approx 5{,}1988$$

Таблиця 3.5

X \ Y	u_i \ v_j	30 –2	40 –1	50 0	60 1	70 2	n_{x_i}	$u_i n_{x_i}$	$u_i^2 n_{x_i}$	$\sum_{j=1}^{5} u_i v_j n_{ij}$
5	–3	2 [6]	–	–	–	–	2	–6	18	12
10	–2	6 [4]	5 [2]	–	–	–	11	–22	44	34
15	–1	–	3 [1]	7 [0]	4 [–1]	–	14	–14	14	–1
20	0	–	–	40 [0]	9 [0]	4 [0]	53	0	0	0
25	1	–	–	2 [0]	6 [1]	7 [2]	15	15	15	20
30	2	–	–	–	–	5 [4]	5	10	20	20
n_{y_j}		8	8	49	19	16	100	–17	111	–
$v_j n_{y_j}$		–16	–8	0	19	32	27	–	–	–
$v_j^2 n_{y_j}$		32	8	0	19	64	123	–	–	–
$\sum_{i=1}^{6} u_i v_j n_{ij}$		36	13	0	2	34	–	–	–	85

За формулою (16.14) знайдемо вибірковий коефіцієнт кореляції $\rho_{xy} = 44{,}795/(5{,}1988 \cdot 10{,}7569) \approx 0{,}801$, а потім, використовуючи (3.16) і (3.17), одержимо шукані рівняння регресії

$$\bar{y}_x - 52{,}7 = 0{,}801 \cdot \frac{10{,}7569}{5{,}1988}(x - 19{,}15) \Rightarrow \bar{y}_x = 1{,}6574x + 20{,}9616$$

$$\bar{x}_y - 19{,}15 = 0{,}801 \cdot \frac{5{,}1988}{10{,}7569}(y - 52{,}7) \Rightarrow \bar{x}_y = 0{,}3871y - 1{,}2502$$

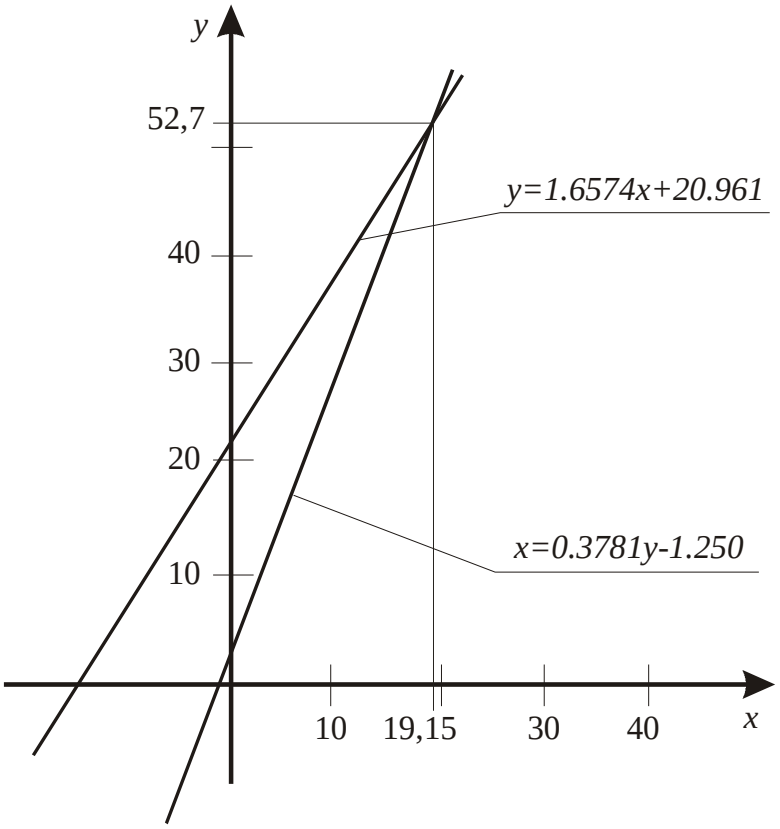

Рис. 3.3

На рис. 4.3 представлені графіки отриманих прямих регресії.

3.3 Коефіцієнт кореляції

Для кількісної характеристики тісноти лінійного кореляційного зв'язку між двома величинами X і Y генеральної сукупності раніше було уведене поняття *коефіцієнта лінійної кореляції,* обумовленого співвідношенням

$$r_{xy} = \frac{M(XY) - m_x m_y}{\sigma_x \sigma_y} \tag{3.24}$$

де $m_x = M(X),\ m_y = M(Y),\quad \sigma_x = \sqrt{D(X)},\quad \sigma_y = \sqrt{D(Y)}$

Відомо, що якщо величини X і Y незалежні, то $r_{xy} = 0$; якщо $r_{xy} = \pm 1$, то X і Y зв'язані лінійною функціональною залежністю, причому, дорівнює $r_{xy} = 1$ у випадку зростаючої залежності й $r_{xy} = -1$ у випадку спадної; $|r_{xy}| \le 1$. При цьому $r_{xy} > 0$, якщо при зростанні однієї величини (наприклад, X) M(Y) збільшується, і негативний у протилежному випадку.

На практиці для оцінки тісноти лінійного кореляційного зв'язку між величинами X і Y за результатами вибіркових спостережень використовується вибірковий коефіцієнт лінійної кореляції, обумовлений формулою (3.25), тобто

$$\rho_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{s_x s_y} \tag{3.25}$$

Формулу (3.25) звичайно застосовують у трохи модифікованому вигляді. Так, підставивши в неї розгорнуті вираження для s_x, s_y і s_{xy}, легко одержати наступну формулу:

$$\rho_{xy} = \frac{n\sum_{i=1}^{s}\sum_{j=1}^{k} x_i y_j n_{ij} - \left(\sum_{i=1}^{s} x_i n_{x_i}\right)\left(\sum_{j=1}^{k} y_j n_{y_j}\right)}{\sqrt{n\sum_{i=1}^{s} x_i^2 n_{x_i} - \left(\sum_{i=1}^{s} x_i n_{x_i}\right)^2} \cdot \sqrt{n\sum_{j=1}^{k} y_j^2 n_{y_j} - \left(\sum_{j=1}^{k} y_j n_{y_j}\right)^2}} \tag{3.26}$$

Якщо дані не згруповані у вигляді кореляційної таблиці й представляють n пару чисел $(x_i; y_j)$, то для обчислення коефіцієнта кореляції у формулі (3.26) варто взяти $n_{ij} = n_{x_i} = n_{y_j} = 1,\ i = j$, а $\sum_{i=1}^{s}\sum_{j=1}^{k}$ замінити на $\sum_{i=1}^{n}$.

Коефіцієнт кореляції ρ_{xy} є безрозмірною величиною (тому що розмірністю чисельника й знаменника є розмірності добутку XY); його величина не залежить від вибору одиниць виміру обох змінних; величина ρ_{xy} приймає значення на відрізку [–1;1]. Близька до нуля величина коефіцієнта кореляції свідчить про відсутність лінійного зв'язку змінних, але не заперечує можливість існування іншої форми залежності між ними.

Хоча вибірковий коефіцієнт кореляції ρ_{xy} представляє значиму оцінку для r_{xy}, однак більш надійну оцінку близькості ρ_{xy} до r_{xy} за даними вибірки можна дати лише в тому випадку, коли розподіл величин X і Y досить близький до нормальної форми. Наприклад, для оцінки r_{xy} нормально розподіленої генеральної сукупності, у випадку більших вибірок $(n \geq 50)$, можна скористатися формулою

$$\rho_{xy} - u_\gamma \frac{1-\rho_{xy}^2}{\sqrt{n}} \leq r_{xy} \leq \rho_{xy} + u_\gamma \frac{1-\rho_{xy}^2}{\sqrt{n}} \tag{3.27}$$

де u_γ – корінь рівняння $2\Phi(\varepsilon) = \gamma$, що визначається з таблиці за заданою довірчою імовірністю γ, а величина

$$\delta = u_\gamma \frac{1-\rho_{xy}^2}{\sqrt{n}} \tag{3.28}$$

називається точністю оцінки.

Наведемо схему знаходження точності δ і довірчих границь, що відповідають надійності γ:

$$\left.\begin{array}{l} n \\ \rho_{xy} \\ \gamma \end{array}\right\} \to 2\Phi(\varepsilon)=\gamma \xrightarrow{\text{П4}} \varepsilon = u_\gamma \to \delta = u_\gamma \frac{1-\rho_{xy}^2}{\sqrt{n}} \longrightarrow \begin{array}{l} \rho_{xy}-\delta \\ \rho_{xy}+\delta \end{array} \qquad (3.29)$$

Приклад 3. Знайти довірчий інтервал для оцінки коефіцієнта кореляції r_{xy} з надійністю $\gamma = 0{,}95$, якщо $\rho_{xy} = 0{,}23$, $n = 320$.

За схемою (17.6),

$$\left.\begin{array}{l} n = 320 \\ \rho_{xy} = 0{,}23 \\ \gamma = 0{,}95 \end{array}\right\} \to 2\Phi(\varepsilon) = 0{,}95 \to \varepsilon = 1{,}96 \to$$

$$\to \delta = 1{,}96\frac{1-(0{,}23)^2}{\sqrt{230}} \approx 0{,}104 \longrightarrow \begin{array}{l} 0{,}23-0{,}104 \\ 0{,}23+0{,}104 \end{array}$$

Звідси, $0{,}126 < r_{xy} < 0{,}334$.

Застосовуючи коефіцієнт кореляції як міру зв'язку, потрібно мати на увазі, що він отриманий на основі даних вибірки й, отже, підданий впливу випадковості.

Якщо обсяг вибірки невеликий, то знайти вибіркову помилку цієї величини досить складно, тому на практиці зазвичай замість визначення помилки коефіцієнта кореляції перевіряють гіпотезу про його значимість (істотність), тобто чи суттєво ρ_{xy} відрізняється від нуля чи цю відмінність можна приписати впливу випадковості, пов'язаної з вибіркою. Інакше кажучи, виникає необхідність при заданому рівні значимості a перевірити нульову гіпотезу $H_0 : r_{xy} = 0$ при конкуруючій гіпотезі $H_0 : r_{xy} \neq 0$.

Якщо нульова гіпотеза відкидається, то це означає, що ρ_{xy} істотно відрізняється від нуля, а X та Y корельовані, тобто пов'язані лінійною залежністю. Якщо нульова гіпотеза буде прийнята, то не ρ_{xy} значимо, а X та Y некорельовані, тобто не пов'язані лінійною залежністю.

Як критерій перевірки нульової гіпотези приймемо випадкову величину

$$T = \frac{\rho_{xy}\sqrt{n-2}}{\sqrt{1-\rho_{xy}^2}}, \qquad (3.30)$$

яка має розподіл Стьюдента з $k = n–2$ ступенями волі. Число ступенів волі менше числа спостережень на 2, оскільки у формулу вибіркового коефіцієнта кореляції входять середні вибіркові значення X і Y, для розрахунку яких

використаються дві лінійні формули їхньої залежності від спостережень випадкових величин.

Оскільки альтернативна гіпотеза має вигляд $H_0 : r_{xy} \neq 0$, то критична область є двосторонньою. Перевірку нульової гіпотези будемо здійснювати за наступною схемою:

$$\left.\begin{array}{l} n; k = n-2 \\ \rho_{xy} \\ \alpha \end{array}\right\} \rightarrow H_0 : r_{xy} = 0 \rightarrow T_{\text{набл}} = \frac{\rho_{xy}\sqrt{n-2}}{\sqrt{1-\rho_{xy}^2}} \rightarrow$$

$$\rightarrow H_1 : r_{xy} \neq 0 \rightarrow t_{\text{кр}}(\alpha; k) \xrightarrow{Д\,7} t_{\text{кр}}^{\text{дв}} \begin{cases} |T_{\text{набл}}| < t_{\text{кр}}^{\text{дв}} \rightarrow \textit{прийнята}, \\ |T_{\text{набл}}| > t_{\text{кр}}^{\text{дв}} \rightarrow \textit{відхилена} \end{cases} \tag{3.31}$$

Приклад 4. За вибіркою обсягу n = 62, взятою з нормальної двовимірної генеральної сукупності (X; Y), знайдений вибірковий коефіцієнт кореляції ρ_{xy} = 0,3. Потрібно при рівні значимості α = 0,01 перевірити нульову гіпотезу $H_0 : r_{xy} = 0$ при альтернативній гіпотезі $H_1 : r_{xy} \neq 0$.

Використовуючи схему (17.8), одержимо

$$\left.\begin{array}{l} n = 62; k = 60 \\ \rho_{xy} = 0{,}3 \\ \alpha = 0{,}01 \end{array}\right\} \rightarrow H_0 : r_{xy} = 0 \rightarrow T_{\text{набл}} = \frac{0{,}3 \cdot \sqrt{60}}{\sqrt{1-0/09}} \approx 2{,}436 \rightarrow$$

$$\rightarrow H_1 : r_{xy} \neq 0 \rightarrow t_{\text{кр}}(0{,}01; 60) \xrightarrow{\text{П}\,6} t_{\text{кр}}^{\text{дв}} = 2{,}66 \rightarrow |T_{\text{набл}}| < t_{\text{кр}}^{\text{дв}} \rightarrow \textit{прийнята}$$

Таким чином, немає підстав відкинути нульову гіпотезу. Інакше кажучи, ρ_{xy} незначно відрізняється від нуля, тобто X і Y некорельовані, тобто не пов'язані лінійною залежністю.

3.4 Поняття про множинну кореляцію

У тих випадках, коли досліджується кореляційний зв'язок між величинами, число яких більше двох, уводять поняття *множинної* кореляції.

Так, при дослідженні кореляційного зв'язку між трьома кількісними ознаками X, Y і Z можна ввести рівняння регресії

$$\bar{z}_{xy} = f(x; y)$$

де $\bar{z}_{xy}$ – середнє значення величини *Z*, що відповідає певним значенням *x* та *y*. Геометричною інтерпретацією цього рівняння є деяка поверхня в прямокутній системі координат тривимірного простору.

У найбільш простому випадку лінійної кореляційної залежності ознак *X*, *Y* і *Z* вибіркове рівняння регресії має вигляд

$$\bar{z}_{xy} = ax + by + c$$

Нехай у результаті незалежних спостережень над досліджуваною системою кількісних ознак (*X*; *Y*; *Z*), отримані *n* сукупностей чисел (x_1,y_1,z_1), (x_2, y_2, z_2), …, (x_n, y_n, z_n)... Припустимо, що функція регресії лінійна, тобто

$$z = ax + by + c \qquad (3.32)$$

В (17.9) $\bar{z}_{xy}$ замінене на z , тому що різні значеннях *x* і *y* ознак *X* і *Y*, і відповідні їм значення *z* ознаки *Z* спостерігалися по одному разу. Якщо позначити $\bar{z}_i = ax_i + by_i + c$ наближене значення z_i, обчислене з рівняння регресії (17.9), то величина $z_i - \bar{z}_i$, є відхиленням наближеного значення $\bar{z}_i$, від точного z_i. Коефіцієнти *a*, *b* і *з* рівняння регресії (17.9) знайдемо з вимоги методу найменших квадратів:

$$S(a;b;c) = \sum_{i=1}^{n}\left(z_i - \tilde{z}_i\right)^2 = \sum_{i=1}^{n}\left(z_i - ax_i - by_i - c\right)^2 \to \min$$

Необхідні умови мінімуму функції *S* утворять систему

$$\begin{cases} \dfrac{\partial S}{\partial a} = 2\sum\limits_{i=1}^{n}\left(z_i - ax_i - by_i - c\right)\left(-x_i\right) = 0, \\ \dfrac{\partial S}{\partial b} = 2\sum\limits_{i=1}^{n}\left(z_i - ax_i - by_i - c\right)\left(-y_i\right) = 0 \\ \dfrac{\partial S}{\partial c} = 2\sum\limits_{i=1}^{n}\left(z_i - ax_i - by_i - c\right)\left(-1\right) = 0, \end{cases}$$

яка в результаті тотожних перетворень набуває такого вигляду

$$\begin{cases} a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i y_i + c\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i z_i, \\ a\sum_{i=1}^{n} x_i y_i + b\sum_{i=1}^{n} y_i^2 + c\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} y_i z_i \\ a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} y_i + cn = \sum_{i=1}^{n} z_i. \end{cases} \quad (3.33)$$

Розв'язуючи систему (3.10) щодо невідомих параметрів *a*, *b* і *c*, а потім, підставляючи їхні значення в (3.9), одержимо шукане рівняння регресії.

Зручніше, однак, знайти готові формули для розрахунку параметрів, а не розв'язувати систему щоразу. Коефіцієнти регресії *a*, *b* і параметр *c* у цьому випадку обчислюються за наступними формулами:

$$a = \frac{\rho_{xz} - \rho_{yz}\rho_{xy}}{1-\rho_{xy}^2}\frac{s_z}{s_x}; \quad b = \frac{\rho_{yz} - \rho_{xz}\rho_{xy}}{1-\rho_{xy}^2}\frac{s_z}{s_y}; \quad c = \bar{z} - a\bar{x} - b\bar{x} \quad (3.34)$$

Тут $\rho_{yz}, \rho_{xz}, \rho_{xy}$ – вибіркові коефіцієнти лінійної кореляції між відповідними ознаками; s_x, s_y, s_z – середні квадратичне відхилення; $\bar{x}, \bar{y}, \bar{z}$ – середні значення.

Тіснота зв'язку ознаки *Z* з ознаками *X* і *Y*, оцінюється *вибірковим сукупним коефіцієнтом кореляції*

$$R = \sqrt{\left(\rho_{xz}^2 - 2\rho_{xy}\rho_{xz}\rho_{yz} + \rho_{yz}^2\right)/\left(1-\rho_{xy}^2\right)} \quad (3.35)$$

причому $0 \le R \le 1$.

Тіснота зв'язку між *Z* і *X* (при постійному *Y*), між *Z* і *Y*, (при постійному *X*) оцінюється відповідно *частковими вибірковими коефіцієнтами кореляції:*

$$\rho_{xz(y)} = \frac{\rho_{xz} - \rho_{xy}\rho_{yz}}{\sqrt{\left(1-\rho_{xy}^2\right)\left(1-\rho_{yz}^2\right)}}; \quad \rho_{yz(x)} = \frac{\rho_{yz} - \rho_{xy}\rho_{xz}}{\sqrt{\left(1-\rho_{xy}^2\right)\left(1-\rho_{xz}^2\right)}} \quad (3.36)$$

Ці коефіцієнти показують ступінь лінійної залежності між спостережуваними значеннями *Z* і *X*, а також *Z* і *Y*, якщо вплив третьої ознаки усунуто.

Приклад 5. Нехай змінна *z* перебуває в лінійній залежності від змінних *x* та *y*. Для оцінок параметрів рівняння регресії зібрані наступні дані:

Таблиця 3.6

i	1	2	3	4	5	6	7	8	9	10
z_i	10	12	17	13	15	10	14	12	16	18
x_i	2	2	8	2	6	3	5	3	9	10

y_i	1	2	10	4	8	4	7	3	10	11

Знайти рівняння лінійної регресії, виходячи із системи (3.10) і формул (3.11), а також сукупний і частковий вибіркові коефіцієнти кореляції.

Складемо розрахункову табл. 3.7

Таблиця 3.7

i	x_i	x_i^2	y_i	y_i^2	z_i	z_i^2	$x_i y_i$	$x_i z_i$	$y_i z_i$
1	2	4	1	1	10	100	2	20	10
2	2	4	2	4	12	144	4	24	24
3	8	64	10	100	17	289	80	136	170
4	2	4	4	16	13	169	8	26	52
5	6	36	8	64	15	225	48	90	120
6	3	9	4	16	10	100	12	30	40
7	5	25	7	49	14	196	35	70	98
8	3	9	3	9	12	144	9	36	36
9	9	81	10	100	16	256	90	144	160
10	10	100	11	121	18	324	ПО	180	198
$\sum$	50	336	60	480	137	1947	398	756	908

Побудуємо систему рівнянь (3.33), використовуючи розрахунки табл. 3.7

$$\begin{cases} 336a + 398b + 50c = 756, \\ 398a + 480b + 60c = 908, \\ 50a + 60b + 10c = 137 \end{cases}$$

Розв’язок системи щодо невідомих параметрів дає шукані оцінки: a = 0,1285; b = 0,6117; c = 9,3872.

Рівняння регресії має вигляд

$$z = 0{,}1285a + 0{,}6117b + 9{,}3872$$

Для розв’язку цього прикладу за допомогою виразів (3.11) знайдемо:

$$\bar{x} = 5;\ \bar{y} = 6;\ \bar{z} = 13{,}7;\ \ \overline{x^2} = 33{,}6;\ \ \overline{y^2} = 48;\ \overline{z^2} = 194{,}7;$$

$$\overline{xy} = 39{,}8;\ \ \overline{xz} = 75{,}6;\ \ \overline{yz} = 90{,}8;$$

$$s_x = \sqrt{\overline{x^2} - \bar{x}^2} = \sqrt{33{,}6 - 25} \approx 2{,}9325757;$$

$$s_y = \sqrt{\overline{y^2} - \bar{y}^2} = \sqrt{48 - 36} \approx 3{,}4641016;$$

$$s_z = \sqrt{\overline{z^2} - \bar{z}^2} = \sqrt{194{,}7 - 187{,}69} \approx 2{,}6476405$$

$$s_{xy} = \overline{xy} - \bar{x} \cdot \bar{y} = 39{,}8 - 30 = 9{,}8;$$
$$s_{xz} = \overline{xz} - \bar{x} \cdot \bar{y} = 75{,}6 - 68{,}5 = 7{,}1;$$
$$s_{yz} = \overline{yz} - \bar{y} \cdot \bar{z} = 90{,}8 - 82{,}2 = 8{,}6$$

$$\rho_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{9{,}8}{2{,}9325757 \cdot 3{,}4641016} \approx 0{,}9646865;$$
$$\rho_{xz} = \frac{s_{xz}}{s_x s_z} = \frac{7{,}1}{2{,}9325757 \cdot 2{,}6476405} \approx 0{,}9144292;$$
$$\rho_{yz} = \frac{s_{yz}}{s_y s_z} = \frac{8{,}6}{3{,}4641016 \cdot 2{,}6476405} \approx 0{,}9376674$$

$$a = \frac{0{,}9144292 - 0{,}9376674 \cdot 0{,}9646865}{1 - (0{,}9646865)^2} \cdot \frac{2{,}6476405}{2{,}9325757} \approx 0{,}1285;$$
$$b = \frac{0{,}9376674 - 0{,}9144292 \cdot 0{,}9646865}{1 - (0{,}9646865)^2} \cdot \frac{2{,}6476405}{3{,}4641016} \approx 0{,}6117;$$
$$c = 13{,}7 - 0{,}1285 \cdot 5 - 0{,}6117 \cdot 6 \approx 9{,}3872$$

Тіснота зв'язку ознаки Z з ознаками X і Y, X (при постійному Y), Y (при постійному X) визначається за формулами (3.12) і (3.13):

$$\rho_{xz(y)} = \frac{0{,}9144292 - 0{,}904555}{\sqrt{0{,}069380 \cdot 0{,}120780}} \approx 0{,}1079;$$
$$\rho_{yz(x)} = \frac{0{,}9376674 - 0{,}8821375}{\sqrt{0{,}069380 \cdot 0{,}163819}} \approx 0{,}5209$$

3.5 Питання для самоперевірки теми 3

1. Дайте визначення функціональної залежності між ознаками X і Y.
2. Яку залежність називають статистичною або стохастичною?
3. Що називається умовним середнім $\bar{y}_x(\bar{x}_y)$ спостережуваних значень $Y(X)$?
4. Які рівняння називають модельними рівняннями регресії?
5. Що називається кореляційним полем?
6. Сформулюйте основні завдання регресійного аналізу.
7. Запишіть модель лінійної парної регресії Y по X і X по Y.
8. Запишіть формули для знаходження параметрів α і β $(\alpha'$ й $\beta')$ прямій регресії.

9. Чим відрізняються регресійні прямі $y = \alpha + \beta x$ й $x = \alpha' + \beta' y$?
10. Поясніть, як заповнюється кореляційна таблиця.
11. Чому дорівнює вибірковий коефіцієнт кореляції?
12. Чому рівні вибіркові коефіцієнти лінійної регресії?
13. Дайте визначення кореляційної залежностей між ознаками X і Y.
14. Встановіть зв'язок r_{xy} з $r_{y|x}$ і $r_{x|y}$.
15. Наведіть виведення $\bar{x}$, $\bar{y}$, s_x^2 і s_y^2 через умовні варіанти.
16. Виразіть коваріацію через умовні варіанти.
17. Для чого використовується вибірковий коефіцієнт лінійної кореляції? Наведіть формулу.
18. Наведіть схему знаходження довірчих границь для оцінки r_{xy}.
19. Наведіть схему перевірки нульової гіпотези про рівність нулю r_{xy}.
20. Запишіть модель лінійної регресійної залежності ознак *X, Y* і *Z.*
21. Запишіть формули для знаходження параметрів *a, b* і *з* лінійного рівняння регресії.
22. Для чого використовуються вибіркові сукупний і частковий коефіцієнти кореляції? Наведіть формули.