4. КОРЕЛЯЦІЙНИЙ АНАЛІЗ

Кореляцією (**кореляційним зв'язком**) між випадковими величинами (ознаками) називають наявність статистичного або ймовірнісного зв'язку між ними. При цьому закономірна зміна певних ознак призводить до закономірної зміни середніх значень інших, пов'язаних з ними ознак. **Кореляційним аналізом** називають сукупність методів виявлення кореляційного зв'язку. Тому його можна застосовувати для формалізованого подання моделей зв'язків між окремими компонентами системи або між окремими процесами, що відбуваються в ній. Наявність кореляційного зв'язку не означає існування причинно-наслідкового зв'язку між досліджуваними ознаками. Вона може бути зумовлена тим, що обидві ознаки мають причинно-наслідковий зв'язок з певним іншим фактором. Наприклад, існує кореляція між цінами на нафту й на золото. Проте вона пояснюється тим, що обидві ціни виражаються у доларах США й залежать від динаміки його індексу. Кореляція також може бути випадковою.

Сучасну класифікацію мір подібності запропонували австрійський та американський біостатистик та антрополог Роберт Сокал та британський таксономіст Пітер Сніс у 1963 р. Згідно з нею виокремлюють такі типи мір подібності [58]:

– міри асоціації, що відбивають різні співвідношення кількості ознак, що збігаються до загальної кількості ознак, а також близькі до них коефіцієнти спряженості (квантифіковані коефіцієнти зв'язку);

– вибіркові коефіцієнти зв'язку типу кореляції (нормовані косинусні міри);

– показники відстані у метричному просторі.

Перевірку зв'язку можна здійснювати лише для пов'язаних вибірок. Це означає, що між елементами обох досліджуваних вибірок існує взаємно однозначна відповідність, а кількість елементів у вибірках є однаковою.

Замість гіпотези про наявність кореляційного зв'язку часто розглядають протилежну гіпотезу про відсутність зв'язку між досліджуваними величинами. Нехай ознака A має r рівнів $A_1, A_2, \ldots A_r$, а ознака B – s рівнів $B_1, B_2, \ldots B_s$. Їх вважають **незалежними**, якщо події "ознака A набуває значення A_i" та "ознака B набуває значення B_j" є незалежними для всіх можливих пар i, j, тобто:

$$P\left(A_i, B_j\right) = P\left(A_i\right) P\left(B_j\right). \tag{4.1}$$

Це можна сформулювати в інший спосіб: ознаки є незалежними, якщо значення ознаки A не впливає на ймовірності реалізації можливих значень ознаки B:

$$P\left(B_j / A_i\right) = P\left(B_j\right), \quad \forall\left(A_i, B_j\right). \tag{4.2}$$

Кореляційний аналіз здійснюють на початковому етапі вирішення всіх основних проблем статистичного аналізу даних [4]. У проблемі статистичного аналізу залежностей і побудови регресійних моделей він дає змогу встановити сам факт існування зв'язку між змінними та оцінити ступінь його прояву. У проблемі класифікації даних за допомогою кореляційного аналізу отримують вихідну інформацію у вигляді коваріаційних і кореляційних матриць та інших характеристик парних порівнянь. Це дає змогу визначити подібні один до одного або до певних еталонів об'єкти, сформувати класи подібних об'єктів і здійснити класифікацію. У проблемі зменшення розмірності досліджуваного простору ознак також за допомогою коваріаційних і кореляційних матриць визначають ознаки, що можуть бути без втрати суттєвої інформації подані через інші наявні дані.

Загальна методика перевірки гіпотези про існування зв'язку між ознаками передбачає три основних етапи: визначення типу даних; перевірку гіпотези про відсутність зв'язку і, в разі її відхилення, оцінювання сили зв'язку. Тип вихідних даних суттєво впливає на вибір методів і критеріїв, які можна застосовувати на наступних етапах аналізу.

Для визначення сили зв'язку використовують різноманітні показники. Зазвичай їх прагнуть вибрати такими, щоб вони змінювалися від –1 до +1 або від 0 до 1. Значення, що є близькими за модулем до одиниці, свідчать про наявність сильного зв'язку. Близькі до нуля значення вказують або на відсутність будь-якого зв'язку, або на відсутність зв'язку того типу (найчастіше лінійного), для якого розроблено відповідний коефіцієнт. Знак коефіцієнта вказує на напрям зв'язку: прямий (для додатних значень) або зворотний (для від'ємних).

4.1. Кореляційний аналіз кількісних ознак

Методику кількісного оцінювання кореляції між ознаками вперше було запропоновано британським географом, антропологом та психологом Френсисом Гальтоном в 1888 р.

Універсальною характеристикою ступеня тісноти зв'язку між кількісними ознаками є коефіцієнт детермінації. **Вибірковий коефіцієнт детермінації** певної ознаки y за вектором незалежних ознак $\mathbf{X}=\left(x^{(1)}, x^{(2)}, ..., x^{(p)}\right)$ можна розрахувати як:

$$K_d\left(y;\mathbf{X}\right)=1-\frac{s_\varepsilon^2}{s_y^2}, \tag{4.3}$$

де

$$s_y^2=\frac{1}{n}\sum_{k=1}^{n}\left(y_k-\bar{y}\right)^2, \tag{4.4}$$

n – кількість спостережень, а вибіркове значення дисперсії нев'язок ε обчислюють за однією з таких формул:

$$s_\varepsilon^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(\mathbf{X}_i)\right)^2, \qquad (4.5)$$

де $\hat{f}(\mathbf{X}_i)$ є статистичною оцінкою невідомого значення функції регресії $f(\mathbf{X})$ у точці $\boldsymbol{X}_i$, або:

$$s_\varepsilon^2 = \frac{1}{m}\sum_{j=1}^{m}\frac{1}{\nu_j}\sum_{i=1}^{\nu_j}\left(y_{ij} - \bar{y}_{j*}\right)^2, \qquad (4.6)$$

де ν_j – кількість даних, що потрапили до j-го інтервалу групування;

y_{ji} – значення i-го спостереження досліджуваної ознаки, що потрапило до j-го інтервалу;

$\bar{y}_{j*} = \dfrac{\sum_{i=1}^{\nu_j} y_{ji}}{\nu_j}$ – її середнє значення за спостереженнями, які потрапили до j-го інтервалу;

m – кількість інтервалів.

Формулу (4.5) застосовують у випадку, коли за результатами попереднього аналізу встановлено, що умовна дисперсія $D(\varepsilon \mid \mathbf{X}) = \sigma_\varepsilon^2 = const$, тобто не залежить від **X**. Формулу (4.6) використовують, якщо ця умова не виконується, а також у всіх випадках, коли обчислення здійснюють за згрупованими даними. У цьому випадку необхідно попередньо здійснити групування даних. Для цього їх впорядковують за зростанням значень однієї з ознак (ознаки X). Потім задають кількість та межі інтервалів для цієї ознаки. Підраховують кількості точок, що потрапили до кожного інтервалу (ν_j), для змінної *Y* обчислюють загальне середнє $\bar{y}$ та середні за інтервалами $\bar{y}_{j*}$ й розраховують значення коефіцієнта детермінації за формулами (4.3, 4.4, 4.6).

Величина коефіцієнта детермінації може змінюватися в межах від нуля до одиниці й відображає частку загальної дисперсії досліджуваної ознаки, яка зумовлена зміною функції регресії $f(\mathbf{X})$. При цьому нульове значення коефіцієнта детермінації відповідає відсутності будь-якого зв'язку, а його рівність одиниці – наявності строго функціонального зв'язку. Оскільки цей коефіцієнт є універсальним показником зв'язку, він має відбивати й такі зв'язки, що є немонотонними функціями. Тому питання напряму зв'язку у цьому випадку не має сенсу.

Слід зазначити, що для обмеженого набору даних часто можна побудувати декілька різних адекватних регресійних моделей. Групування

даних також можна здійснювати різними способами. Тому існує певна невизначеність коефіцієнтів детермінації: при застосуванні різних регресійних моделей або різних способів групування ми будемо отримувати дещо різні значення коефіцієнта детермінації.

Інші поширені характеристики ступеня тісноти зв'язку між ознаками можна розглядати як окремі випадки коефіцієнта детермінації, отримані для конкретних математичних моделей зв'язку.

Розрізняють парні та частинні кореляційні характеристики. Парні характеристики розраховують за результатами вимірювань тільки досліджуваної пари ознак. Тому вони не враховують опосередкованого або спільного впливу інших ознак. Частинні характеристики є очищеними від впливу інших факторів, але для їх розрахунку необхідно мати вихідну інформацію не тільки про досліджувані ознаки, а й про всі інші, вплив яких необхідно усунути.

Для кількісних ознак найчастіше застосовують коефіцієнти кореляції Пірсона і Фехнера. **Коефіцієнт кореляції Пірсона (коефіцієнт кореляційного відношення Пірсона, парний коефіцієнт кореляції, вибірковий коефіцієнт кореляції, коефіцієнт Бравайса – Пірсона)** вимірює ступінь лінійного кореляційного зв'язку між кількісними скалярними ознаками. Він був запропонований К. Пірсоном у 1896 р. Часто, посилаючись на згадування К. Пірсона про ідеї математичного подання зв'язку, висловлені в 1846 р. відомим французьким фізиком та кристалографом Огюстом Браве, цей показник називають коефіцієнтом Бравайса – Пірсона (Бравайс – це викривлена транскрипція від французького Bravais, що закріпилася в літературі з кореляційного аналізу). Цей коефіцієнт розраховують за формулою:

$$r = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\sqrt{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}}. \qquad (4.7)$$

Коефіцієнт Пірсона можна виразити також через дисперсії σ_y і $\sigma_{\Delta y}$, друга з яких характеризує розкид емпіричних точок стосовно рівняння лінійної регресії $y = ax + b$, де a та b – коефіцієнти, визначені за методом найменших квадратів:

$$r = \frac{1}{\sqrt{1 + \left(\sigma_{\Delta y} / \sigma_y\right)^2}}. \qquad (4.8)$$

За умови достатньо великого обсягу спостережень $\left(N \geq 30\right)$ стандартне відхилення коефіцієнта кореляції Пірсона можна визначити за формулою:

$$\sigma_r = \frac{1 - r^2}{\sqrt{N}}. \qquad (4.9)$$

На рівні значущості 0,01 гіпотезу про наявність кореляційного зв'язку приймають, якщо $|r|/\sigma_r \geq 2,6$.

Застосування коефіцієнта Пірсона як міри зв'язку є обґрунтованим лише за умови, що спільний розподіл пари ознак є нормальним. Тому перед його розрахунком слід перевірити виконання цієї гіпотези. Якщо вона справедлива, то квадрат коефіцієнта кореляції Пірсона дорівнює коефіцієнту детермінації.

Значення коефіцієнта кореляції може змінюватися від −1 до +1. Значення −1 та +1 відповідають чіткій лінійній функціональній залежності, яка в першому випадку є спадною, а у другому – зростаючою. Для функціональної залежності $y = const$ коефіцієнт кореляції, як видно з наведеної формули, є невизначеним, оскільки в цьому випадку знаменник дорівнює нулю. Що ближчим є значення коефіцієнта кореляції до −1 або +1, то більш обґрунтованим є припущення про наявність лінійного зв'язку. Наближення його значення до нуля свідчить про відсутність лінійного зв'язку, але не є доказом відсутності статистичного зв'язку взагалі.

На рис. 4.1 показано дві серії точок, координати яких відповідають двом парам спряжених вибірок.

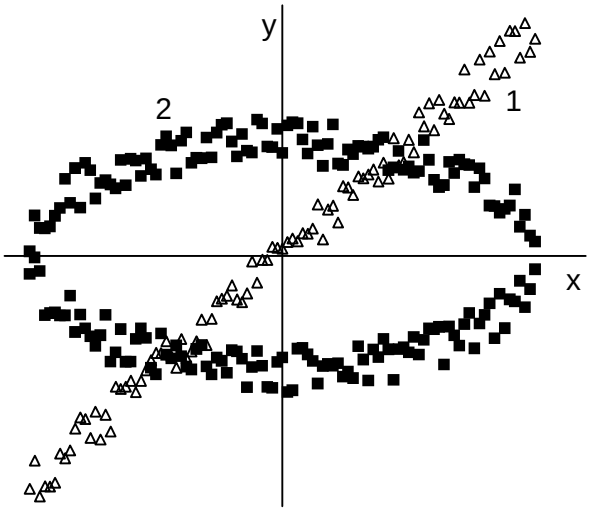


Рис. 4.1. Графічне зображення двох наборів тестових даних

Для обох пар вибірок є очевидним існування статистичного зв'язку між параметрами *x* та *y*. Але коефіцієнти кореляції для них дорівнюють, відповідно, $r_1 = 0,995$ і $r_2 = 0,006$. Близькість коефіцієнта кореляції до нуля для другої пари вибірок пов'язана не з відсутністю зв'язку, а з його нелінійністю. Для порівняння, коефіцієнти детермінації для тих самих пар вибірок дорівнюють 0,98 та 1,00.

Показаний приклад свідчить, що в багатьох випадках для попереднього аналізу припущення про наявність і тип зв'язку між певними ознаками доцільно нанести наявні дані на графік.

Як видно, близькість коефіцієнта кореляції Пірсона до нуля в загальному випадку не є доказом незалежності ознак. Але можна довести, що у випадку, коли сумісний розподіл випадкових величин (x, y) є нормальним, рівність $r = 0$ свідчить про статистичну незалежність x і y.

Коефіцієнт кореляції Пірсона часто розглядають як універсальну міру кореляційного зв'язку. У багатьох пакетах загального призначення, зокрема в електронних таблицях MS Excel, не передбачено інших засобів його вимірювання. Але, як випливає з наведених вище даних, насправді сфера його обґрунтованого застосування є досить вузькою, оскільки лінійність залежності й нормальний розподіл даних навколо неї є скоріше винятком, ніж правилом.

При дослідженні багатовимірних сукупностей випадкових величин із коефіцієнтів кореляції, обчислених попарно між ними, можна побудувати квадратну симетричну кореляційну матрицю з одиницями на головній діагоналі. Вона є основним елементом при побудові багатьох алгоритмів багатовимірної статистики, наприклад у факторному аналізі. Довірчий інтервал вибіркової оцінки коефіцієнта кореляції для двовимірної нормальної генеральної сукупності:

$$r \in \left[\tanh\left(z(r) - \frac{N_{\frac{1+p}{2}}}{\sqrt{n-3}} \right); \tanh\left(z(r) + \frac{N_{\frac{1+p}{2}}}{\sqrt{n-3}} \right) \right], \qquad (4.10)$$

де n – обсяг вибірки;

$N_{\frac{1+p}{2}}$ – квантіль нормального розподілу;

p – значення довірчого рівня;

z (r) – z-перетворення (перетворення Фішера) вибіркового коефіцієнта кореляції r.

Коефіцієнт кореляції Пірсона можна застосовувати для перевірки гіпотези про значущість зв'язку. Для нормально розподілених вихідних даних величину вибіркового коефіцієнта кореляції вважають значимо відмінною від нуля, якщо виконується нерівність:

$$r^2 > \left[1 + (n-2)/t_\alpha^2\right]^{-1}, \qquad (4.11)$$

де t_α – критичне значення t-розподілу з $(n-2)$ степенями вільності.

Статистика $\sqrt{n-1}r$ має r-розподіл зі щільністю:

$$\varphi_{r(n)}(r) = \frac{1}{\sqrt{\pi}} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n-2}{2}\right)} \left(1-r^2\right)^{\frac{n-4}{2}} \quad (-1 < r < 1). \qquad (4.12)$$

Для великих за обсягом вибірок статистика $\sqrt{n-1}r$ наближається до стандартного нормального розподілу.

У випадку, коли між двома наборами ознак існує нелінійний зв'язок, для оцінювання ступеня його тісноти часто використовують **кореляційне відношення**, яке було запропоновано К. Пірсоном. Це можливо, якщо щільність розміщення емпіричних точок на координатній площині дає можливість їх групування за однією із змінних і підрахунку групових середніх значень другої змінної для кожного інтервалу. Тоді кореляційне відношення залежної змінної *y* за незалежною змінною *x* можна розрахувати за формулою:

$$\rho_{yx}^2 = s_{\bar{y}(x)}^2 / s_y^2, \qquad (4.13)$$

де

$$s_{\bar{y}(x)}^2 = \frac{1}{n}\sum_{j=1}^{s} \nu_j \left(\bar{y}_{j*} - \bar{y}\right)^2;$$

$$s_y^2 = \frac{1}{n}\sum_{j=1}^{s}\sum_{i=1}^{\nu_j} \left(y_{ji} - \bar{y}\right)^2;$$

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{s} \nu_j \bar{y}_{j*};$$

$$\bar{y}_{j*} = \left(\sum_{i=1}^{\nu_j} y_{ij}\right) / \nu_j,$$

n – обсяг вибірки; s – кількість інтервалів групування по вісі абсцис; v_j – кількість точок, що потрапили до j-го інтервалу. З погляду термінології, уведеної у попередньому розділі, кореляційне відношення є квадратним коренем з відношення факторної варіації ознаки до її загальної варіації.

Кореляційне відношення може змінюватися в інтервалі від нуля до одиниці. Із $\rho_{yx} = 1$ випливає наявність строгого функціонального зв'язку між досліджуваними ознаками, і навпаки, однозначний функціональний зв'язок між ними свідчить про те, що $\rho_{yx} = 1$. За відсутності зв'язку $\rho_{yx} = 0$, і навпаки, коли $\rho_{yx} = 0$, це означає, що для всіх інтервалів групування $\bar{y}_{j*} = \bar{y}$, тобто групові середні $\bar{y}_{j*}$ не залежать від x.

На відміну від коефіцієнта кореляції, кореляційне відношення не є симетричним: у загальному випадку $\rho_{yx} \neq \rho_{xy}$. Більше того, можливі ситуації, коли один із цих коефіцієнтів дорівнює нулю, другий – одиниці. Зок-

рема, це може спостерігатися для парних функцій за умови, що функція розподілу значень незалежної змінної є симетричною стосовно нуля. Для даних, що наведені на рис. 4.1, кореляційне відношення першої серії дорівнює приблизно 0,98 і в межах похибки обчислень збігається з коефіцієнтами детермінації і кореляції. Для другої серії $\rho_{yx} \approx 0,63$ і $\rho_{xy} \approx 0,72$.

Можна довести, що кореляційне відношення збігається з модулем коефіцієнта кореляції між тими самими змінними за наявності лінійного зв'язку, а також за відсутності зв'язку. В інших випадках воно перевищує модуль коефіцієнта кореляції. Це дає можливість використовувати їх різницю як характеристику ступеня відхилення зв'язку від лінійності. Для цього розраховують величину:

$$v^2 = \frac{(n-k)\left(\rho_{yx}^2 - r^2\right)}{(k-2)\left(1-\rho_{yx}^2\right)}, \qquad (4.14)$$

де n – кількість емпіричних точок;

k – кількість невідомих параметрів моделі. Ця величина приблизно підпорядковується F-розподілу з параметрами $s - 2$ та $n - s$. Якщо розраховане за формулою (4.13) значення перевищує точку v_α^2 розподілу $F(s-2, n-s)$, то гіпотезу про лінійний зв'язок відхиляють на рівні значущості α. Слід зазначити, що у зв'язку з можливістю різних способів групування даних значення кореляційного відношення, як і значення коефіцієнта детермінації, у загальному випадку є дещо невизначеним.

Коефіцієнт кореляції Фехнера розраховують за формулою:

$$r_F = \frac{C-H}{C+H} = \frac{2C-n}{n} = \frac{2C}{n} - 1, \qquad (4.15)$$

де C – кількість збігів знаків відхилень варіант від відповідних середніх;

H – кількість знаків, що не збігаються. Цей показник було запропоновано німецьким психологом Густавом Фехнером у 1860 р.

Значення коефіцієнта Фехнера можуть змінюватися в межах від −1 до +1. Як і коефіцієнт Пірсона, він показує наявність лінійного зв'язку: що ближчим до одиниці за модулем є значення коефіцієнта, то сильніший зв'язок. Малі значення абсолютної величини коефіцієнта свідчать про відсутність лінійного зв'язку, але цього недостатньо для твердження про відсутність будь-якого зв'язку взагалі. Зокрема, для наведених на рис. 4.1 наборів даних значення коефіцієнта Фехнера дорівнюють, відповідно, $r_{F1} = 0,941$ і $r_{F2} = -0,010$. Застосування для обчислення коефіцієнта лише кількості збігів або незбігів знаків відхилень від середніх значень можна розглядати як зведення первинної кількісної шкали до номінальної, що має призвести до втрати частини корисної інформації. Тому цей критерій

застосовують досить рідко, але у певних випадках, коли інформація про збіги й незбіги знаків відхилень потрібна й для інших цілей, він може виявитися зручнішим за критерій Пірсона.

Коваріацією називають змішаний момент другого порядку. Її розраховують за формулою:

$$Cov(X,Y)=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right). \qquad (4.16)$$

На відміну від інших показників, що характеризують наявність статистичного зв'язку, вона не є безрозмірною величиною. Також немає будь-яких обмежень на її значення. У загальному випадку за інших рівних умов вона збільшується (за модулем) із зростанням середніх значень досліджуваних показників. Це робить коваріацію незручною для застосування як показника сили зв'язку. Але у багатьох алгоритмах її використовують як проміжний показник, що застосовують у подальших розрахунках. У таких випадках важливою перевагою коваріації є необхідність виконання значно меншої кількості елементарних обчислень, ніж для аналогічних показників кореляції, таких як коефіцієнт Пірсона. Крім того вона має важливе теоретичне значення, що також у певних випадках приводить до доцільності її використання в аналізі даних.

Коваріація вибірки із самою собою є дисперсією. З наведеної формули можна отримати, що коефіцієнт кореляції Пірсона $r=\frac{Cov(X,Y)}{\sigma_X\sigma_Y}$, де σ_X^2, σ_Y^2 – дисперсії вибірок.

При аналізі багатовимірних вибірок часто застосовують **коваріаційні матриці**:

$$C=\begin{pmatrix} c_{11} & c_{12} & \dots & c_{1n} \\ c_{21} & c_{22} & \dots & c_{2n} \\ \dots & \dots & \dots & \dots \\ c_{n1} & c_{n2} & \dots & c_{nn} \end{pmatrix}, \qquad (4.17)$$

де $c_{ij}=\mathrm{cov}\left(x_i,x_j\right)$. Діагональні елементи матриці (4.17) є дисперсіями $c_{ii}=\sigma^2\left(x_i\right)$ відповідних рядів спостережень. Коваріаційна матриця є симетричною, тобто $c_{ij}=c_{ji}$.

4.2. Кореляційний аналіз порядкових ознак

Під **ранговою кореляцією** розуміють статистичний зв'язок між порядковими ознаками. Вихідні дані зазвичай подають у вигляді табл. 4.1, де елемент x_{ik} є рангом *i*-го об'єкта за *k*-ю властивістю.

Таблиця 4.1

Таблиця вихідних даних для рангового кореляційного аналізу

Порядковий номер об’єкта	Порядковий номер досліджуваної ознаки						
	0	1	2	…	k	…	p
1	x_{10}	x_{11}	x_{12}	…	x_{1k}	…	x_{1p}
2	x_{20}	x_{21}	x_{22}	…	x_{2k}	…	x_{2p}
…	…	…	…	…	…	…	…
i	x_{i0}	x_{i1}	x_{i2}	…	x_{ik}	…	x_{ip}
…	…	…	…	…	…	…	…
n	x_{n0}	x_{n1}	x_{n2}	…	x_{nk}	…	x_{np}

Завданнями аналізу в цьому випадку можуть бути: вивчення структури досліджуваних об’єктів; перевірка сукупної узгодженості ознак та умовне ранжирування об’єктів за ступенем тісноти зв’язку кожної з них з іншими ознаками; побудова єдиного групового впорядкування об’єктів (задача регресії на порядкових змінних).

У першому випадку кожну послідовність впорядкованих за k-ю ознакою n об’єктів подають як точку $\mathbf{X}^{(k)} = \left(x_1^{(k)},\ x_2^{(k)},\ ...,\ x_n^{(k)} \right),\ k = 0,\ 1,\ ...,\ p$ у n-вимірному просторі ознак. Найхарактернішими типами структури є такі.

1. Аналізовані точки рівномірно розкидані по всій області їх можливих значень. Це означає відсутність будь-якого зв’язку між досліджуваними ознаками.

2. Частина точок утворює ядро (кластер) із точок, що розташовані близько одна до одної, а інші випадково розкидані навколо цього ядра. Це відповідає існуванню підмножини узгоджених ознак.

3. Аналізовані точки утворюють декілька кластерів, розташованих відносно далеко один від одного. Це відповідає наявності декількох таких підмножин ознак, що існує істотний статистичний зв’язок між ознаками, які належать до однієї і тієї самої підмножини, і не існує значущого зв’язку між ознаками, які належать до різних підмножин.

Прикладом завдання другого типу є визначення узгодженості думок групи експертів з наступним впорядкуванням їх за рівнем компетентності. Для цього розраховують коефіцієнти конкордації для різних сукупностей досліджуваних змінних.

Вирішення завдань третього типу зводиться до побудови такого впорядкування, яке б у певному значенні було б найближчим до кожного з наданих впорядкувань досліджуваних ознак. Для цього часто застосовують середнє арифметичне або медіану наявних базових рангів. Це можна розглядати як задачу найкращого у певному розумінні відновлення невідомого ранжирування за наявними емпіричними даними, що зумовлює можливість її розгляду як задачі регресії.

Коефіцієнт рангової кореляції Спірмена (**показник кореляції рангів Спірмена, коефіцієнт кореляції рангів**) запропонований британським

психологом Чарльзом Едвардом Спірменом у 1904 р. Його використовують, якщо досліджується зв'язок між рядами даних, виміряними за порядковою шкалою. Його можна застосовувати також і для кількісних даних, але, як правило, це буває недоцільним. У найпростішому випадку досліджувані об'єкти класифікують за двома ознаками. Наприклад, ми можемо спочатку впорядкувати групу учнів за їх здібностями до математики, а потім – до іноземних мов. Місця, які *i*-й учень займе в обох списках, будуть його рангами r_i та s_i. Якщо досліджувані ознаки взаємопов'язані, то послідовність рангів $r_1, r_2, ..., r_n$ певною мірою корелює з послідовністю рангів $s_1, s_2, ..., s_n$.

Ступінь близькості двох послідовностей відображує величина:

$$S_\rho = \sum_{i=1}^{n} \left(r_i - s_i\right)^2 . \tag{4.18}$$

Якщо для нумерації об'єктів попередньо впорядкувати їх за однією з ознак, наприклад за зростанням рангів r_i, то формула (4.18) може бути записана так:

$$S_\rho = \sum_{k=1}^{n} \left(k - s_k\right)^2 . \tag{4.19}$$

Величина S_ρ набуде найменшого можливого значення $S_\rho = 0$ тоді й тільки тоді, коли послідовності повністю збігатимуться. Найбільше можливе значення $S_\rho = \frac{1}{3}\left(n^3 - n\right)$ відповідає випадку, коли послідовності є повністю протилежними, тобто для будь-яких i, j з нерівності $r_i > r_j$ випливає $s_i < s_j$, і послідовності рангів першої ознаки $r_i = \{1, 2, ..., n\}$ відповідає послідовність рангів другої $s_i = \{n, n-1, ..., 1\}$. Величину S_ρ незручно застосовувати як міру зв'язку, оскільки на її значення впливає кількість пар варіант досліджуваних рядів n.

З огляду на це, як міру зв'язку використовують коефіцієнт рангової кореляції Спірмена, значення якого розраховують за формулою:

$$\rho_s = 1 - \frac{6\left(S_\rho + B_x + B_y\right)}{n^3 - n}, \tag{4.20}$$

де B_x, B_y – поправки на об'єднання рангів у відповідних рядах, які обчислюють за формулою:

$$B_i = \frac{1}{12}\sum_{i=1}^{m} n_i \left(n_i^2 - 1\right), \tag{4.21}$$

де m – кількість груп об'єднаних рангів у вибірці;

n_i – кількість рангів у *i*-й групі.

Значення коефіцієнта можуть змінюватися в межах від −1 до +1, при цьому −1 відповідає повній протилежності послідовностей рангів, а +1 – їх повному збігу.

Коефіцієнт рангової кореляції Спірмена можна застосовувати як показник некорельованості вибірок. У цьому випадку розраховують величину:

$$t_p = \sqrt{n-2}\,\frac{\rho_s}{\sqrt{1-\rho_s^2}}. \qquad (4.22)$$

Для великих за обсягом вибірок ($n > 50$) статистика цього критерію наближається до розподілу Стьюдента з $(n-2)$ степенями вільності. Статистика $\sqrt{n-1}\rho_s$ для великих вибірок наближається до стандартного нормального розподілу.

Інший підхід використовує як міру подібності двох вибірок мінімальну кількість перестановок сусідніх об'єктів, потрібну для переведення послідовності рангів однієї вибірки до послідовності рангів іншої. Можна показати, що вона дорівнює кількості інверсій в однієї з цих послідовностей у випадку, коли інша послідовність впорядкована за зростанням. Нехай, наприклад, $n = 4$, послідовність r_i впорядкована за зростанням, а $s_i = \{4,\ 3,\ 1,\ 2\}$. Інверсіями є: $4 > 3;\ 4 > 1;\ 4 > 2;\ 3 > 1;\ 3 > 2$. Їх кількість $K = 5$. Найменше можливе значення кількості інверсій $K = 0$ відповідає повному збігу рангових послідовностей, а найбільше $K = \frac{n(n-1)}{2}$ – їх повній протилежності.

Як і в попередньому випадку, кількість інверсій залежить від обсягу вибірки і є незручною для застосування як показника кореляції. Для цього використовують **коефіцієнт рангової кореляції Кендалла** (**коефіцієнт кореляції рангів**, **ранговий коефіцієнт кореляції**). Він був запропонований британським статистиком Маурисом Кендаллом у 1938 р. Його розраховують за формулою:

$$\tau = 1 - \frac{2K}{\sqrt{\left(\frac{n(n-1)}{2} - B_x\right)\left(\frac{n(n-1)}{2} - B_y\right)}}, \qquad (4.23)$$

де r_j, s_i – масиви рангів аналізованих рядів;

n – кількість пар варіант у них. B_x, B_y – поправки на об'єднання рангів у відповідних рядах, які обчислюють за формулою:

$$B_i = \frac{1}{2}\sum_{i=1}^{m} n_i\,(n_i - 1), \qquad (4.24)$$

де m – кількість груп об'єднаних рангів у вибірці;

n_i – кількість рангів у i-й групі.

Для коефіцієнта рангової кореляції Кендалла у випадку великих вибірок статистика:

$$\tau\sqrt{\frac{9n(n-1)}{2(2n+5)}} \tag{4.25}$$

має розподіл, близький до стандартного нормального закону.

Коефіцієнт рангової кореляції Кендалла призначений для визначення сили кореляційного зв'язку між двома рядами даних за тих самих умов, що і коефіцієнт рангової кореляції Спірмена. Як і для коефіцієнта Спірмена, його значення можуть змінюватися в межах від −1 до +1, при цьому −1 відповідає повній протилежності послідовностей рангів, а +1 – їх повному збігу. Слід зазначити, що обчислення коефіцієнта Кендалла є більш трудомістким, але з іншого боку, він має ряд переваг порівняно із коефіцієнтом Спірмена. Основними з них є такі [23]:

– кращий рівень вивченості його статистичних властивостей, зокрема його вибіркового розподілу;
– можливість його застосування для визначення частинної кореляції;
– більша зручність перерахунку при додаванні нових даних.

4.3. Кореляційний аналіз номінальних ознак

Типовою ситуацією, коли необхідна перевірка зв'язку між номінальними ознаками, є обробка результатів соціологічних досліджень, що можуть містити такі комбінації ознак, як освіта, стать, професія, підтримка певної політичної партії, регіон проживання тощо.

При дослідженні зв'язків між **категоризованими** ознаками вихідні дані подають у вигляді таблиці спряженості (табл. 4.2). До категоризованих зараховують номінальні ознаки, а також порядкові ознаки, для яких є відомим скінченний набір можливих градацій.

Величини f_{ij} показують, скільки разів зустрічалася комбінація ознак, за якої рівень першої має значення i, а рівень другої – j; m_j є сумами стовпців, а n_i – сумами рядків. За даними табл. 4.2 можна оцінити значення ймовірностей, що входять до формули (4.1):

Таблиця 4.2

Таблиця спряженості категоризованих ознак

Рівні ознаки 1	Рівні ознаки 2				Разом
	1	2	…	r	
1	f_{11}	f_{12}	…	f_{1r}	n_1
2	f_{21}	f_{22}	…	f_{2r}	n_2
…	…	…	…	…	…
c	f_{c1}	f_{c2}	…	f_{cr}	n_c
Разом	m_1	m_2	…	m_r	S

$$p_{ij} = P\left(A_i B_j\right) = \frac{f_{ij}}{S};\ p_i = P\left(A_i\right) = \sum_{j=1}^{r} p_{ij} = \frac{n_i}{S};$$
$$p_j = P\left(B_j\right) = \sum_{i=1}^{c} p_{ij} = \frac{m_j}{S}. \tag{4.26}$$

Звідси для незалежних ознак маємо:

$$f_{ij} \approx n_i m_j / S. \tag{4.27}$$

Величини $\varphi_{ij} = n_i m_j / S$ є очікуваними частотами. Нульову гіпотезу про відсутність зв'язку відхиляють, якщо різницю між ними й частотами, що спостерігаються, не можна пояснити випадковими чинниками. Як критерій можна використовувати величину:

$$\chi^2 = \sum_{i=1}^{c}\sum_{j=1}^{r} \frac{\left(f_{ij} - \varphi_{ij}\right)^2}{\varphi_{ij}} = S\left[\sum_{i=1}^{c}\sum_{j=1}^{r} \frac{f_{ij}^2}{n_i m_j} - 1\right], \tag{4.28}$$

яка при достатньо великому обсязі вибірки наближається до розподілу χ^2 з кількістю степенів вільності $(r-1)(c-1)$. На практиці для можливості застосування критерію часто вважають достатнім, щоб усі значення f_{ij} були не меншими ніж п'ять. При збільшенні кількості степенів вільності мінімальні значення f_{ij} можуть бути дещо меншими.

На практиці частіше використовують **φ-коефіцієнт Пірсона**, або **середньоквадратичну спряженість** $\varphi^2 = \chi^2 / S$, яка може змінюватися від нуля до $\min\{r-1, c-1\}$.

Існує велика кількість показників ступеня тісноти статистичного зв'язку, призначених для категоризованих змінних, які не є універсальними, а відображають окремі властивості такого зв'язку.

Коефіцієнт спряженості Крамера був запропонований К.Х. Крамером у 1946 р. Його розраховують за формулою:

$$C = \left[\frac{\sum_{i=1}^{c}\sum_{j=1}^{r} \frac{f_{ij}^2}{n_i m_j} - 1}{\min\left(c-1, r-1\right)}\right]^{1/2} = \left[\frac{\varphi^2}{\min\left(c-1, r-1\right)}\right]^{1/2}. \tag{4.29}$$

Він змінюється в межах від нуля до одиниці. При цьому значення $C = 0$ свідчить про статистичну незалежність аналізованих ознак, а значення $C = 1$ – про можливість однозначного відтворення значень однієї

ознаки за відомими значеннями другої. Дисперсію оцінки коефіцієнта Крамера можна отримати з виразу:

$$\sigma_C^2 \approx \frac{1}{n \min\left(c-1, r-1\right)}. \tag{4.30}$$

Її довірчий інтервал:

$$\left[C - u_{1-\alpha}\sigma_C; C + u_{1-\alpha}\sigma_C\right], \tag{4.31}$$

де u_q – q-квантиль стандартного нормального розподілу.

Поліхоричний коефіцієнт спряженості Чупрова призначений для дослідження кореляції номінальних ознак у таблиці спряженості $r \times c$. Він був уведений російським статистиком О.О. Чупровим у 1926 р. Його значення розраховують за формулою:

$$T = \frac{J-1}{\sqrt{(r-1)(c-1)}};\ J = \sum_{i=1}^{c}\sum_{j=1}^{r}\frac{f_{ij}^2}{n_i m_j}. \tag{4.32}$$

Існує велика кількість коефіцієнтів, що характеризують кореляцію між ознаками у випадку, коли кожна з двох ознак може мати лише два рівні, які найчастіше відповідають наявності та відсутності ознаки. У цьому випадку таблиця спряженості має розмір 2×2 і її елементи позначають так: $a = f_{11}$, $b = f_{12}$, $c = f_{21}$, $d = f_{22}$.

Коефіцієнт (показник подібності) Жаккара, уведений в 1901 р. французьким геоботаніком Полем Жаккаром, обчислюють за формулою:

$$J = \frac{a}{a+b+c}. \tag{4.33}$$

Значення цього коефіцієнта можуть змінюватися в межах від нуля до одиниці.

Простий коефіцієнт зустрічальності (показник подібності Сокала й Міченера) запропонований Р. Сокалом та американським ентомологом Чарльзом Дунканом Міченером у 1958 р. Його розраховують за формулою:

$$J = \frac{a+d}{n} = \frac{a+d}{a+b+c+d}. \tag{4.34}$$

Як і в попередньому випадку, значення коефіцієнта можуть змінюватися в межах від нуля до одиниці.

Показник подібності Рассела і Рао запропонували в 1940 р. американський епідеміолог Поль Ф. Рассел та індийський і британський ентомолог Т. Рамакришна Рао. Його обчислюють як:

$$J = \frac{a}{n} = \frac{a}{a+b+c+d}. \tag{4.35}$$

Його значення також можуть змінюватися в межах від нуля до одиниці.

Коефіцієнт спряженості Бравайса – Пірсона (**показник подібності Чупрова**) був уведений О.О. Чупровим у 1923 р. Його розраховують за формулою:

$$C = \frac{ad - bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}. \qquad (4.36)$$

Значення цього коефіцієнта може змінюватися в межах від −1 до +1. Від'ємні значення коефіцієнта спряженості означають, що із збільшенням імовірності прояву одної ознаки, зменшується імовірність прояву іншою.

Легко показати, що цей показник є окремим випадком φ-коефіцієнта Пірсона для таблиць 2×2.

Коефіцієнт асоціації Юла був уведений відомим британським статистиком Джорджем Удні Юлом у 1900 р. Його визначають із співвідношення:

$$Q = \frac{ad - bc}{ad + bc}. \qquad (4.37)$$

Коефіцієнт колігації Юла, що також був запропонований Дж.У. Юлом в 1912 р., обчислюють як:

$$Y = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}. \qquad (4.38)$$

Він не має переваг порівняно з коефіцієнтом асоціації. Значення обох коефіцієнтів змінюються в межах від −1 до +1.

Хеммінгова відстань (**метрика Хеммінга**) $H = a + d$ також може застосовуватися для визначення кореляції. Проте, як і коваріація, вона не є безрозмірною величиною і може набувати будь-яких невід'ємних значень (верхньою межею є загальна кількість спостережень n). Цей показник був уведений відомим американським математиком Ричардом Веслі Хеммінгом у 1950 р.

4.4. Кореляційний аналіз змішаних ознак

Коефіцієнт Гауера був запропонований британським статистиком Джоном Кліффордом Гауером у 1971 р. Його застосовують у тому випадку, коли досліджувані ознаки виміряні в різних шкалах. Обчислення елементів матриці подібності здійснюють за формулою:

$$s_{ij} = \frac{\sum_{k=1}^{p} S_{ijk}}{\sum_{k=1}^{p} W_{ijk}}, \ \left(i = 1, ..., n;\ j = 1, ..., n\right), \qquad (4.39)$$

де S_{ijk} $\left(i, j = 1, ..., n;\ k = 1, ..., p\right)$ – внесок ознаки у подібність об'єктів;

W_{ijk} – вагова змінна ознаки;

p – кількість ознак, що характеризують об'єкт;

n – кількість об'єктів.

Для дихотомічних ознак алгоритм підрахунку внеску ознаки і визначення вагових коефіцієнтів збігається з коефіцієнтом Жаккара. Для порядкових ознак алгоритм підрахунку внеску ознаки збігається з хеммінговою відстанню, узагальненою на порядкові змінні, а вагові коефіцієнти беруть рівними одиниці для всіх ознак. Для кількісних ознак:

$$S_{ijk} = 1 - \frac{\left| x_{ik} - x_{jk} \right|}{R_k}, \tag{4.40}$$

де x_{ik}, x_{jk} – значення k-ї змінної для об'єктів i та j;

R_k – розкид k-ї ознаки, обчислений за всіма об'єктами.

Бісеріальний коефіцієнт кореляції запропоновано К. Пірсоном. Його призначено для дослідження кореляції в таблицях розміром $2 \times n$, які є дихотоміями за певною номінальною ознакою і класифікаціями за номінальною або порядковою ознакою, що класифікується за q класами і може бути впорядкованою або невпорядкованою. Вихідний розподіл має бути двовимірним нормальним.

При класифікації за порядковою ознакою бісеріальний коефіцієнт:

$$r_b = \frac{\left(\overline{x_1} - \overline{x} \right) n_1}{n s_x z_k}, \tag{4.41}$$

де $\overline{x_1}$ – середнє за першим рядком;

$\overline{x}$ – загальне середнє за всією таблицею;

s_x – вибіркове середнє квадратичне відхилення;

n_1 – чисельність першого рядка; n – загальна чисельність усіх вибірок;

z_k – ордината щільності нормального розподілу в точці k, де k – розв'язок рівняння:

$$1 - F(k) = n_1 / n. \tag{4.42}$$

Значення бісеріального коефіцієнта кореляції можуть змінюватися від −1 до +1. Його похибку можна визначити за формулою:

$$m_{r_b} = \frac{1 - r_b}{\sqrt{n}}. \tag{4.43}$$

Вона має t-розподіл з кількістю степенів вільності $(n-2)$.

Бісеріальний коефіцієнт кореляції за таблицею Келлі – Вуда запропонований американським психологом Луїсом Л. Терстоуном (Louis L. Thurstone) в 1928 р. Його розраховують за формулою:

$$r_b = \frac{\left|\overline{x_1} - \overline{x_2}\right| pq}{s_x \varsigma}, \tag{4.44}$$

де $p = n_1 / n$ – частка частот у рядку, що визначається умовою $p > q$;

q – частка частот в іншому рядку;

ς – ордината в точці межі класів частот першого та другого рядків, яка визначається за таблицею Келлі – Вуда. Похибку коефіцієнта визначають за формулою:

$$m_r = \frac{\frac{\sqrt{pq}}{\varsigma} - r_b^2}{\sqrt{n}}. \tag{4.45}$$

У випадку класифікації за номінальною ознакою обчислення бісеріального коефіцієнта кореляції можна здійснить за формулою:

$$r_\eta = \sqrt{\frac{\frac{1}{n}\sum_{i=1}^{q} n_i \left(\frac{\mu_i}{\sigma_i}\right)^2 - \left(\frac{\mu_y}{\sigma_y}\right)^2}{1 + \frac{1}{n}\sum_{i=1}^{q} n_i \left(\frac{\mu_i}{\sigma_i}\right)^2}}, \tag{4.46}$$

де n – загальний обсяг даних;

n_i – кількість даних в i-му перетині;

m_i/s_i – оцінка в перетині i, одержувана за таблицею нормального інтеграла від відносної частоти першої з двох якісних ознак;

m_y/s_y – оцінка, одержувана за таблицею нормального інтеграла від відносної частоти першої якісної ознаки за всією таблицею.

У випадку, коли одна із змінних дихотомізована, а інша – виміряна в кількісній шкалі, обчислюють **точково-бісеріальний коефіцієнт кореляції**, який визначається за формулою:

$$r_{pb} = \frac{\left|\overline{x_p} - \overline{x}\right|}{s_x} \sqrt{\frac{n_p}{n_q}}, \tag{4.47}$$

де $\overline{x_p}$ – середнє варіант кількісної вибірки, які відповідають подіям верхнього (першого) рівня дихотомічної вибірки;

$\overline{x}$ – середнє кількісної вибірки;

s_x – середнє квадратичне кількісної вибірки;

n_p – кількість подій у верхній (з рівнем 1) групі;

n_q – кількість подій у нижній (з рівнем 2) групі.

При цьому передбачається, що дихотомічна змінна може набувати лише два значення: 1 (верхній рівень) та 0 (нижній рівень). З погляду теорії точково-бісеріальну кореляцію можна розглядати як окремий випадок коефіцієнта кореляції Пірсона.

Величину точково-бісеріального коефіцієнта кореляції вважають відмінною від нуля на рівні значущості α, якщо виконується нерівність:

$$r_{pb}\sqrt{\frac{n-2}{1-r_{pb}^2}} \geq t_\alpha, \tag{4.48}$$

де t_α – критичне значення t-розподілу з $(n-2)$ степенями вільності.

4.5. Множинна кореляція

Про множинну кореляцію мова йде в тому випадку, коли певна ознака може бути пов’язана не з однією, а із сукупністю декількох інших ознак.

У реальних дослідженнях можлива ситуація, коли на певну ознаку може впливати не одна, а декілька інших. В таких випадках парні показники кореляції будуть давати неправильну інформацію щодо наявності зв’язку між відповідними показниками, оскільки ці їх значення будуть викривлятися невраховуваними ознаками.

Для уникнення помилок використовують частинні показники кореляції, що усувають такий вплив. Ідея введення таких показників вперше була висунута Г.У. Юлом у 1896 р., а пізніше розвинена ним та К. Пірсоном.

Якщо досліджувані ознаки задовольняють багатовимірний нормальний розподіл, **частинний коефіцієнт кореляції** між двома ознаками i та j при фіксованих значеннях інших ознак розраховують за формулою:

$$r_{ijX^{(i,\,j)}} = -\frac{R_{ij}}{\sqrt{R_{ii}R_{jj}}}, \tag{4.49}$$

де R_{kl} – алгебраїчне доповнення для елемента r_{kl} у кореляційній матриці. Цей показник запропоновано К. Пірсоном у 1897 р.

Для тривимірної ознаки звідси можна отримати:

$$r_{01(2)} = \frac{r_{01} - r_{02}r_{12}}{\sqrt{\left(1-r_{02}^2\right)\left(1-r_{12}^2\right)}}. \tag{4.50}$$

Частинні коефіцієнти кореляції порядку k, тобто такі, що не враховують опосередкований вплив k інших змінних, можна розрахувати за коефіцієнтами порядку $k-1$, використовуючи рекурентну формулу:

$$r_{01(2,\,3,\,\ldots,\,k+1)} = \frac{r_{01(2,\,\ldots,\,k)} - r_{0\,k+1(2,\,\ldots,\,k)} r_{1\,k+1(2,\,\ldots,\,k)}}{\sqrt{\left(1-r^2_{0\,k+1(2,\,\ldots,\,k)}\right)\left(1-r^2_{1\,k+1(2,\,\ldots,\,k)}\right)}}. \qquad (4.51)$$

Частинні коефіцієнти кореляції мають всі властивості парних коефіцієнтів кореляції. Вони є показниками наявності лінійного зв'язку між двома незалежними ознаками, який не залежить від впливу інших ознак.

Тісноту зв'язку між декількома змінними у випадку множинної регресії можна оцінити за допомогою **коефіцієнта множинної кореляції**:

$$R = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}}, \qquad (4.52)$$

де Y_i – значення змінної, взяті з кореляційної таблиці;

y_i – відповідні значення, розраховані за рівнянням регресії.

Крім того, застосовують множинний коефіцієнт кореляції, якій є мірою лінійної кореляції між певною змінною y та сукупністю величин $X_1, X_2, \ldots, X_n$ і визначається як звичайний парний коефіцієнт кореляції між y та множинною лінійною регресією за $X_1, \ldots, X_n$. При цьому припускають, що досліджувана сукупність підпорядковується багатовимірному нормальному закону.

Множинний коефіцієнт кореляції, запропонований у 1935 р. Г. Хотелінгом, є окремим випадком коефіцієнтів канонічної кореляції. Його можна розрахувати за формулою:

$$R^2_{yX} = 1 - \frac{|R|}{R_{00}}, \qquad (4.53)$$

де $|R|$ – визначник кореляційної матриці.

Його також можна визначити за частинними коефіцієнтами кореляції:

$$R^2_{yX} = 1 - \left(1-r^2_{01}\right)\left(1-r^2_{02(1)}\right)\left(1-r^2_{03(12)}\right)\ldots\left(1-r^2_{0n(1,\,2,\,\ldots,\,n-1)}\right). \qquad (4.54)$$

Наприклад множинний коефіцієнт кореляції між певною ознакою z та двома іншими ознаками (x, y) дорівнює:

$$R_z = R_{z/xy} = \sqrt{\frac{r^2_{zx} + r^2_{zy} - 2r_{xy}r_{zx}r_{zy}}{1-r^2_{xy}}}.$$

Множинний коефіцієнт кореляції мажорує будь-який парний або частинний коефіцієнт кореляції, що характеризує статистичні зв'язки досліджуваної ознаки. Як видно з формули (4.52), додавання нових ознак не може зменшувати коефіцієнт множинної кореляції.

Для багатовимірних нормальних сукупностей виконується рівність:

$$K_d\left(y, X\right) = R_{yX}^2 . \tag{4.55}$$

Подання математичних об'єктів називають **канонічним**, якщо кожному об'єкту однієї множини відповідає один і тільки один об'єкт іншої множини й ця відповідність є взаємно однозначною. **Канонічний кореляційний аналіз** здійснюють між двома сукупностями (групами) вибірок. Він призначений для визначення лінійної функції від перших p компонент і лінійної функції від q компонент, що залишилися, таких, щоб коефіцієнт кореляції між цими лінійними функціями набув найбільшого можливого значення. Чисельності груп (кількість вибірок у першій та другій групах, p та q) можуть різнитися, але необхідною умовою є рівна кількість варіант у всіх вибірках, що становлять обидві групи. Матриця взаємної кореляції двох груп вибірок має вигляд:

$$R = \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix}, \tag{4.56}$$

де R_{11} – матриця взаємної кореляції p змінних першої групи розміром $p \times p$; R_{22} – матриця взаємної кореляції q змінних другої групи розміром $q \times q$; R_{12} – матриця взаємної кореляції змінних першої та другої груп розміром $p \times q$.

Розв'язання зводиться до узагальненої проблеми власних значень:

$$R_{12}^T R_{11}^{-1} R_{12} v = \lambda R_{22} v, \tag{4.57}$$

де λ – вектор власних значень розміром q.

Квадратні корені з власних значень називають **канонічними кореляціями**.

Для випадкової вибірки обсягу n з $\left(p+1\right)$-вимірного нормального розподілу коефіцієнт множинної кореляції вважають відмінним від нуля на рівні значущості α, якщо виконується нерівність:

$$R \geq \sqrt{\frac{pF}{\nu + pF}}, \tag{4.58}$$

де F – значення оберненої функції F-розподілу для довірчого рівня $\left(1-\alpha\right)$ та кількості степенів вільності p та $\left(n-p-1\right)$.

Коефіцієнт конкордації призначений для дослідження зв'язків між порядковими ознаками, кількість яких є більшою, ніж два. Як міру узго-

дженості беруть суму квадратів відхилень сум рангів спостережень (об'єктів) від їх спільного середнього рангу:

$$S_W = \sum_{i=1}^{n}\left(S_i - \overline{S}\right) = \sum_{i=1}^{n} S_i^2 - \frac{\left(\sum_{i=1}^{n} S_i\right)^2}{n};$$
$$\overline{S} = \frac{\sum_{i=1}^{n} S_i}{n};\ S_i = \sum_{j=1}^{k} R_{ij}, \tag{4.59}$$

де R_{ij} – ранг i-го спостереження за j-ю ознакою.

Коефіцієнт конкордації Кендалла (***W*-коефіцієнт Кендалла**) обчислюють за формулою:

$$W = \frac{12S_W}{k^2\left(n^3 - n\right)}. \tag{4.60}$$

Цей показник було запропоновано М. Кендаллом у 1939 р. Його значення може змінюватися в межах від нуля до одиниці, при цьому він дорівнює одиниці лише за умови, що всі досліджувані ранжирування збігаються. Коефіцієнт конкордації дорівнює нулю, якщо $k \geq 3$ і всі ранжирування є випадковими впорядкуваннями вихідної вибірки.

Середнє за всіма можливими парами ранжирувань значення коефіцієнта Спірмена за відсутності об'єднаних рангів пов'язано з коефіцієнтом конкордації співвідношенням:

$$\overline{\rho}_s = \frac{kW - 1}{k - 1}. \tag{4.61}$$

Величина $(k-1)\dfrac{W}{1-W}$ має F-розподіл з кількостями степенів вільності $(n-1)$ та $\left((n-1)(k-1)-2\right)$. Високі значення функції F-розподілу свідчать про високий рівень узгодженості між ранжируваннями.

При $n > 7$ величина $k(n-1)W$ за відсутності зв'язку між ознаками має розподіл, близький до χ^2 з $(n-1)$ степенем вільності. Якщо:

$$k(n-1)W > \chi_\alpha^2(n-1), \tag{4.62}$$

то гіпотезу про відсутність рангової кореляції можна відкинути при рівні значущості α.

4.6. Приклади здійснення кореляційного аналізу

Перевірка гіпотези про наявність зв'язку між кількісними ознаками

Побудуємо дві послідовності. Перша з них є арифметичною прогресією з першим членом –2 й різницею 0,05. Елементи другої розраховані за формулою: $y_i = 2x_i + \varepsilon_i$, де ε_i – елементи згенерованої за допомогою електронних таблиць MS Excel рівномірної випадкової послідовності, заданої на відрізку $[-0,5; 0,5]$.

Для перевірки нульової гіпотези про наявність зв'язку скористаємося відповідною процедурою пакета SPSS (Analyze/Correlate/Bivariate). У відповідному вікні задаємо: вибірки, зв'язок між якими необхідно перевірити; значення коефіцієнтів кореляції, які треба розрахувати; вказуємо характер гіпотези – однобічна чи двобічна; а також, за необхідністю, додаткові опції. На рис. 4.2 наведено результати кореляційного аналізу, отримані у пакеті SPSS, а на рис. 4.3 – графік, з якого видно наявність близького до лінійного зв'язку між ознаками.

Correlations

		VAR00001	VAR00006
VAR00001	Pearson Correlation	1	,991**
	Sig. (2-tailed)		,000
	Sum of Squares and Cross-products	110,700	225,011
	Covariance	1,384	2,813
	N	81	81
VAR00006	Pearson Correlation	,991**	1
	Sig. (2-tailed)	,000	
	Sum of Squares and Cross-products	225,011	465,288
	Covariance	2,813	5,816
	N	81	81

**. Correlation is significant at the 0.01 level (2-tailed).

Correlations

			VAR00001	VAR00006
Kendall's tau_b	VAR00001	Correlation Coefficient	1,000	,927**
		Sig. (2-tailed)	.	,000
		N	81	81
	VAR00006	Correlation Coefficient	,927**	1,000
		Sig. (2-tailed)	,000	.
		N	81	81
Spearman's rho	VAR00001	Correlation Coefficient	1,000	,991**
		Sig. (2-tailed)	.	,000
		N	81	81
	VAR00006	Correlation Coefficient	,991**	1,000
		Sig. (2-tailed)	,000	.
		N	81	81

**. Correlation is significant at the 0.01 level (2-tailed).

Рис. 4.2. Результати кореляційного аналізу, отримані за допомогою пакета SPSS, у випадку лінійного зв'язку

Бачимо, що у цьому випадку, як коефіцієнт кореляції Пірсона, так і рангові коефіцієнти кореляції достатньо точно визначають наявність лінійного зв'язку між ознаками.

В електронних таблицях MS Excel є вбудовані засоби для розрахунку коефіцієнта кореляції Пірсона. Це функція КОРРЕЛ () та процедура "Кореляція" пакета аналізу, який викликають з пункту меню "Сервіс/Аналіз даних". Якщо пакет аналізу не встановлено, то це можна зробити, користуючись пунктом меню "Сервіс/Надбудови".

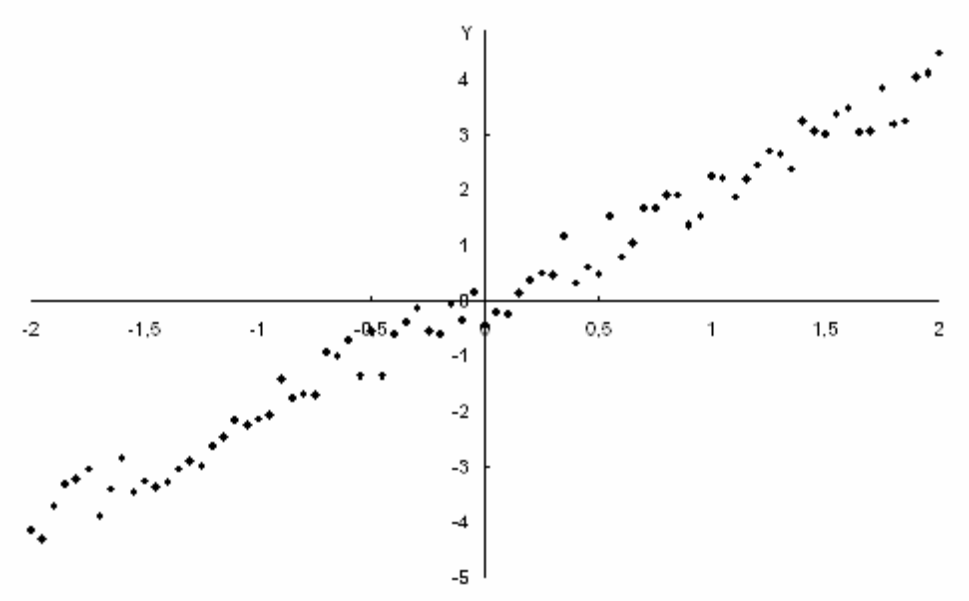


Рис. 4.3. Графік зв'язку між досліджуваними ознаками у випадку лінійного зв'язку

Різниця між цими засобами полягає в тому, що за допомогою пакета аналізу ми отримуємо кореляційну матрицю для даних, що розташовані у декількох сусідніх стовпчиках або рядках робочого аркушу.

При спробі розрахувати коефіцієнти кореляції для даних, між якими є розриви, буде отримано повідомлення про помилку. Якщо ж ми використовуємо функцію КОРРЕЛ (), то вимога щодо відсутності розривів між даними не висувається, але ця функція дає змогу розраховувати лише значення коефіцієнта кореляції між двома множинами даних.

Для прикладу, що розглядається, в обох випадках одержуємо одне й те саме значення коефіцієнта кореляції 0,991.

Розглянемо інший приклад. Елементи другої послідовності у цьому випадку розрахуємо за формулою: $y_i = 2\left(x_i^2 + \varepsilon_i\right)$, де ε_i – елементи згенерованої за допомогою електронних таблиць MS Excel рівномірної випадкової послідовності, заданої на відрізку $[-0,5;\ 0,5]$.

На рис. 4.4 наведено графік, що демонструє зв'язок між ознаками, а результати, одержані за допомогою пакета SPSS, – на рис. 4.5.

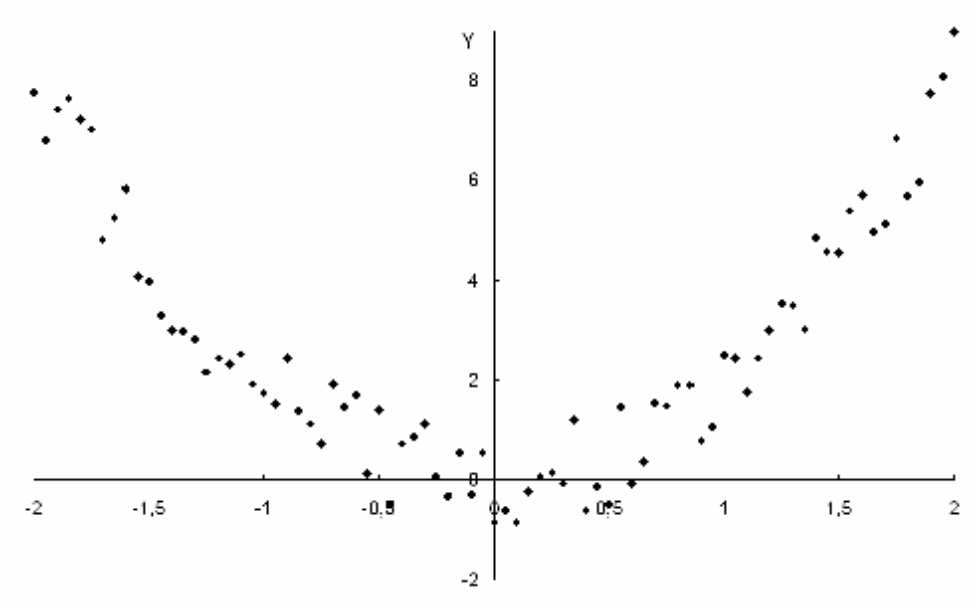


Рис. 4.4. Графік зв’язку між досліджуваними ознаками у випадку нелінійного зв’язку

Correlations

		VAR00001	VAR00003
VAR00001	Pearson Correlation	1	,030
	Sig. (2-tailed)		,791
	Sum of Squares and Cross-products	110,700	7,222
	Covariance	1,384	,090
	N	81	81
VAR00003	Pearson Correlation	,030	1
	Sig. (2-tailed)	,791	
	Sum of Squares and Cross-products	7,222	525,735
	Covariance	,090	6,572
	N	81	81

Correlations

			VAR00001	VAR00003
Kendall's tau_b	VAR00001	Correlation Coefficient	1,000	-,009
		Sig. (2-tailed)	.	,903
		N	81	81
	VAR00003	Correlation Coefficient	-,009	1,000
		Sig. (2-tailed)	,903	.
		N	81	81
Spearman's rho	VAR00001	Correlation Coefficient	1,000	,019
		Sig. (2-tailed)	.	,864
		N	81	81
	VAR00003	Correlation Coefficient	,019	1,000
		Sig. (2-tailed)	,864	.
		N	81	81

Рис. 4.5. Результати кореляційного аналізу, отримані за допомогою пакета SPSS, у випадку нелінійного зв’язку

Бачимо, що у цьому випадку, коефіцієнти кореляції, що розраховуються у пакеті SPSS, не виявляють наявного нелінійного зв'язку. Всі розраховані коефіцієнти є близькими до нуля.

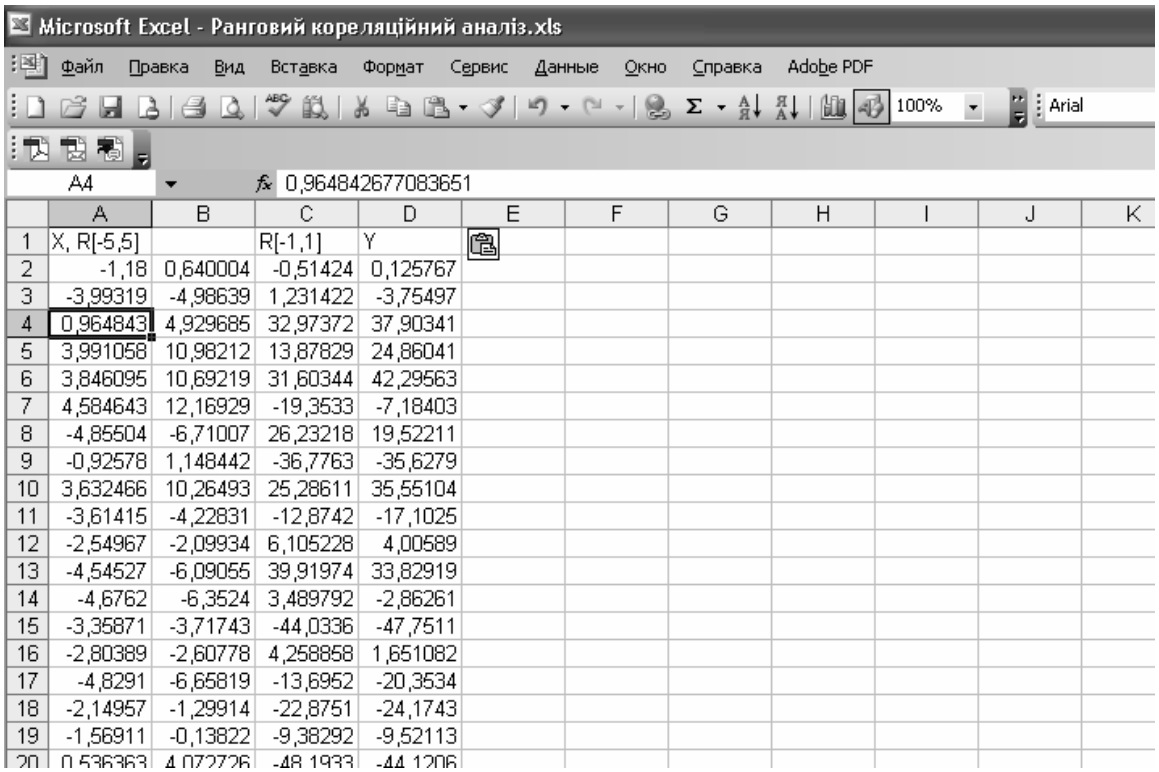
На жаль, у пакеті SPSS та електронних таблицях MS Excel не передбачено можливості розрахунку коефіцієнта детермінації. Тому розрахуємо його за допомогою електронних таблиць MS Excel, використовуючи можливість створення власних розрахункових формул. Вибірковий коефіцієнт детермінації певної ознаки y за вектором незалежних ознак $\mathbf{X} = \left(x^{(1)}, x^{(2)}, ..., x^{(p)} \right)$ можна розрахувати за формулами (4.3; 4.4; 4.6).

У досліджуваному випадку розраховане значення коефіцієнта детермінації дорівнює 0,946, що свідчить про наявність сильного зв'язку між ознаками.

Перевірка гіпотези про наявність зв'язку між порядковими ознаками

Електронні таблиці MS Excel не містять вбудованих доданків, які давали б змогу розраховувати коефіцієнти рангової кореляції. Проте вони дають змогу зробити для цього розрахункову форму.

Проілюструємо це за допомогою такого прикладу. На робочому аркуші електронних таблиць MS Excel сформуємо два стовпчики, що містять елементи досліджуваних вибірок (рис. 4.6).



	A	B	C	D	E	F	G	H	I	J	K
1	X, R[-5,5]		R[-1,1]	Y							
2	-1,18	0,640004	-0,51424	0,125767							
3	-3,99319	-4,98639	1,231422	-3,75497							
4	0,964843	4,929685	32,97372	37,90341							
5	3,991058	10,98212	13,87829	24,86041							
6	3,846095	10,69219	31,60344	42,29563							
7	4,584643	12,16929	-19,3533	-7,18403							
8	-4,85504	-6,71007	26,23218	19,52211							
9	-0,92578	1,148442	-36,7763	-35,6279							
10	3,632466	10,26493	25,28611	35,55104							
11	-3,61415	-4,22831	-12,8742	-17,1025							
12	-2,54967	-2,09934	6,105228	4,00589							
13	-4,54527	-6,09055	39,91974	33,82919							
14	-4,6762	-6,3524	3,489792	-2,86261							
15	-3,35871	-3,71743	-44,0336	-47,7511							
16	-2,80389	-2,60778	4,258858	1,651082							
17	-4,8291	-6,65819	-13,6952	-20,3534							
18	-2,14957	-1,29914	-22,8751	-24,1743							
19	-1,56911	-0,13822	-9,38292	-9,52113							
20	0,536363	4,072726	-48,1933	-44,1206							

Рис. 4.6. Фрагмент робочого аркуша з даними

Перший стовпчик сформуємо як рівномірну випадкову послідовність обсягом 100 елементів, задану на відрізку [–5; 5]. Обираємо в головному меню пункт “Сервіс” й підпункти “Аналіз даних”, “Генерація випадкових чисел”. При цьому з’являється діалогове вікно генератора випадкових чисел (рис. 4.7).

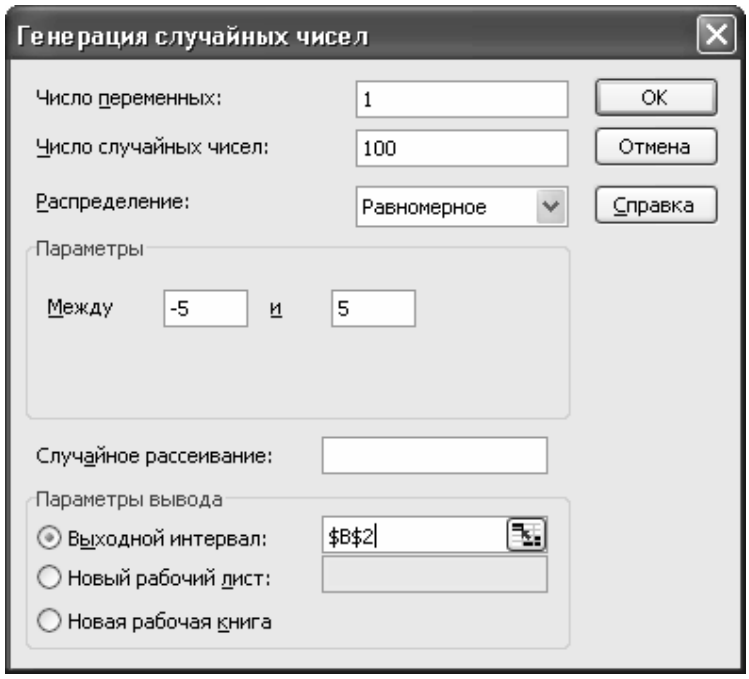


Рис. 4.7. Діалогове вікно генератора випадкових чисел

У цьому вікні задаємо: кількість змінних (кількість вибірок, які необхідно сформувати); кількість випадкових чисел (обсяг кожної вибірки); тип розподілу; параметри розподілу, а також посилання на верхню ліву комірку діапазону, в якому мають бути розташовані згенеровані дані.

Потім згенеруємо елементи другої вибірки, використовуючи формулу:

$$=2*A2+3+C3,$$

де A2 – посилання на комірку, де міститься значення відповідного елемента першої вибірки, а C3 – на комірку з елементом рівномірної випадкової послідовності, заданої на відрізку $[-b; b]$.

Після цього сформуємо стовпчики, що містять значення рангів елементів досліджуваних вибірок. Для цього скористаємося формулою:

$$=РАНГ (A2;\$A\$2:\$A\$101;1),$$

де A2 – посилання на комірку, яка містить елемент, ранг якого необхідно визначити; $A$2:$A$101 – посилання на діапазон комірок, де містяться усі елементи досліджуваних вибірок; 1 – вказівка на те, що ранжирування елементів необхідно здійснювати за зростанням.

Далі формуємо стовпчик квадратів різниць рангів, та в окремій комірці записуємо суму квадратів цих різниць. Потім розраховуємо коефіцієнт кореляції Спірмена, використовуючи формулу:

=1–6*H102/(K1^3-K1),

де H102 – посилання на комірку, у якої записано суму квадратів різниць рангів, K1 – посилання на комірку, де записано кількість елементів у кожній вибірці. На рис. 4.8, 4.9 наведено результати розрахунку рангового коефіцієнта кореляції Спірмена для різних значень параметра b, а також відповідні кореляційні поля досліджуваних ознак, а на рис. 4.10 – залежність коефіцієнта кореляції Спірмена від значення параметра b.

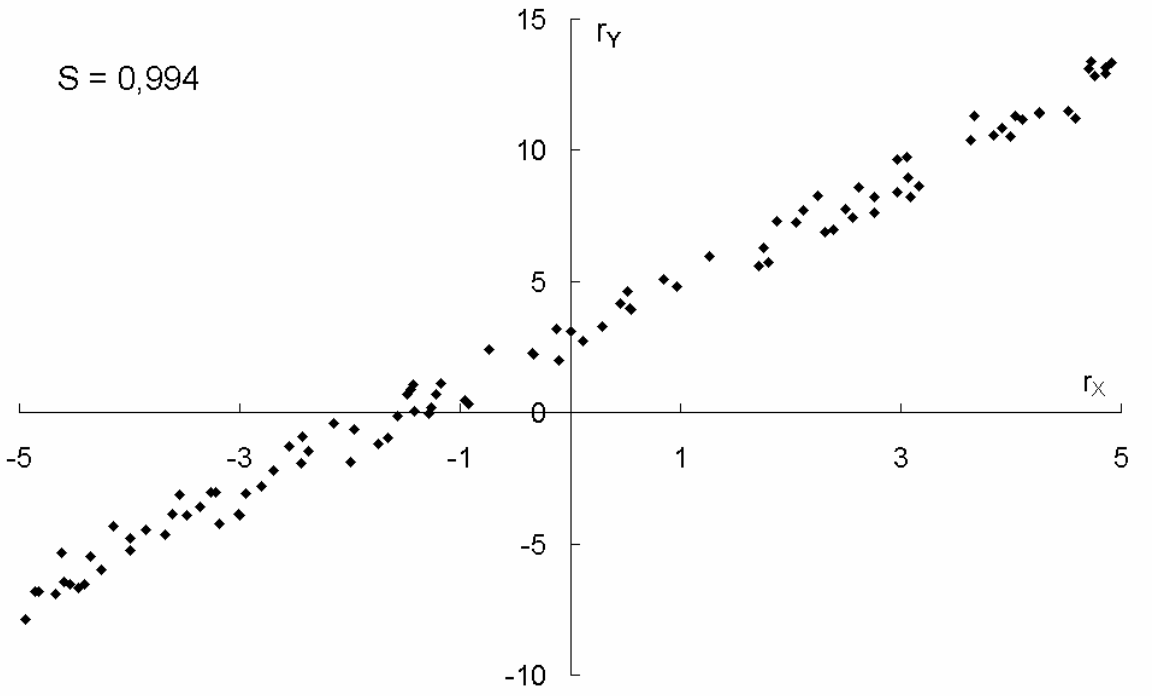


Рис. 4.8. Кореляційне поле досліджуваних ознак при $b = 1$

Більш зручним для дослідження рангової кореляції є застосування спеціалізованих статистичних пакетів, зокрема пакету SPSS. Розглянемо його використання на тому самому прикладі, що і у попередньому випадку.

До вікна даних заносимо стовпчики із значеннями досліджуваних вибірок (рис. 4.11). У пункті меню Analyze обираємо Correlate/ Bivariate Correlations. Після цього з’являється вікно вибору параметрів цієї процедури (рис. 4.12). У ньому треба позначити, між якими змінними шукатимемо кореляцію, які саме коефіцієнти кореляції необхідно розрахувати, а також який тип гіпотези (однобічну чи двобічну) ми розглядаємо.

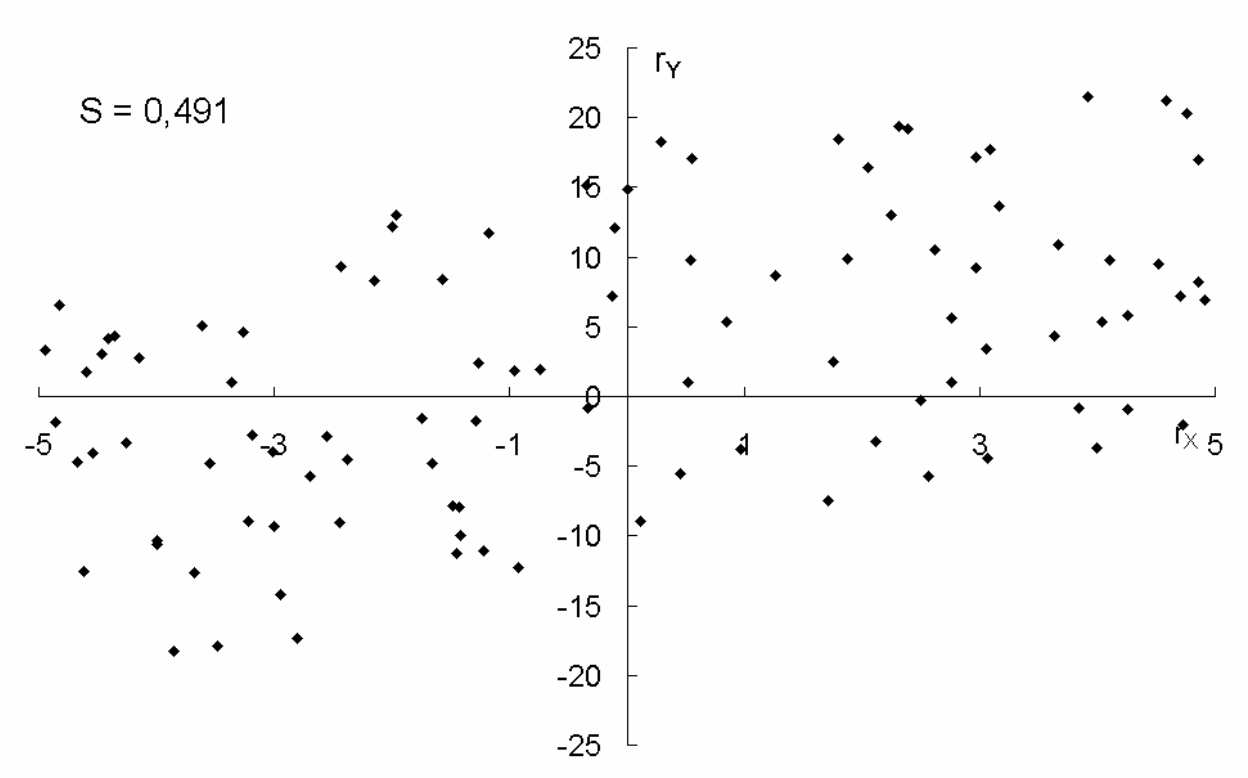


Рис. 4.9. Кореляційне поле досліджуваних ознак при b = 15

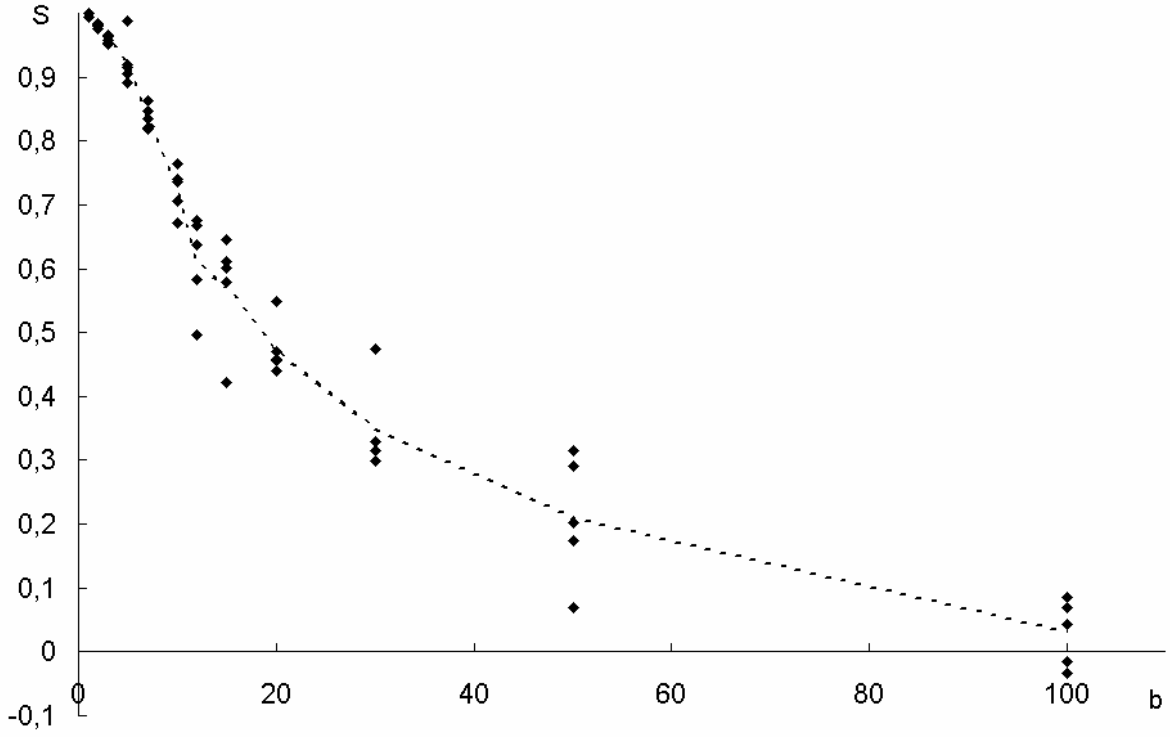


Рис. 4.10. Залежність коефіцієнта Спірмена від параметра b

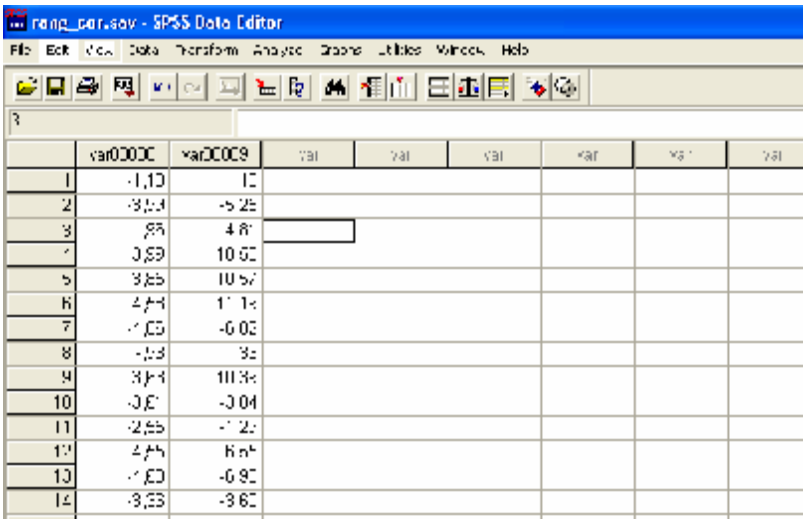


Рис. 4.11. Вікно даних пакету SPSS

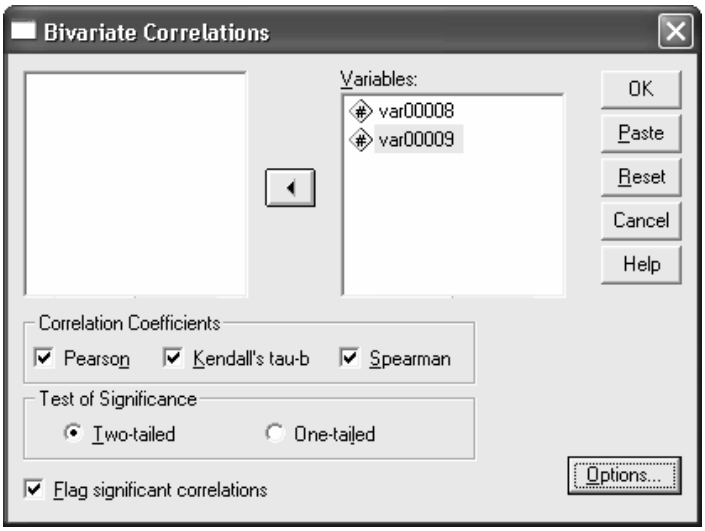


Рис. 4.12. Вікно вибору параметрів кореляційного аналізу

У вікні Options (рис. 4.13) зазначаємо, які додаткові статистичні параметри необхідно розрахувати, а також спосіб обробки пропущених даних.

Результати аналізу наведено на рис. 4.14. Бачимо, що всі коефіцієнти, що розраховувалися є достатньо близькими між собою. Це пов'язано насамперед з тим, що зв'язок між досліджуваними змінними є близьким до лінійного.

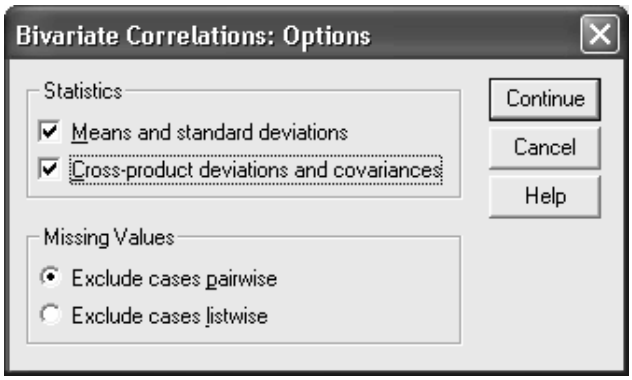


Рис. 4.13. Вікно задання додаткових параметрів кореляційного аналізу

Descriptive Statistics

	Mean	Std. Deviation	N
VAR00008	-,1454	3,09236	100
VAR00009	2,6854	6,23077	100

Correlations

		VAR00008	VAR00009
VAR00008	Pearson Correlation	1	,996**
	Sig. (2-tailed)	,	,000
	Sum of Squares and Cross-products	946,704	1899,604
	Covariance	9,563	19,188
	N	100	100
VAR00009	Pearson Correlation	,996**	1
	Sig. (2-tailed)	,000	,
	Sum of Squares and Cross-products	1899,604	3843,427
	Covariance	19,188	38,822
	N	100	100

**. Correlation is significant at the 0.01 level (2-tailed).

Correlations

			VAR00008	VAR00009
Kendall's tau_b	VAR00008	Correlation Coefficient	1,000	,939**
		Sig. (2-tailed)	,	,000
		N	100	100
	VAR00009	Correlation Coefficient	,939**	1,000
		Sig. (2-tailed)	,000	,
		N	100	100
Spearman's rho	VAR00008	Correlation Coefficient	1,000	,994**
		Sig. (2-tailed)	,	,000
		N	100	100
	VAR00009	Correlation Coefficient	,994**	1,000
		Sig. (2-tailed)	,000	,
		N	100	100

**. Correlation is significant at the .01 level (2-tailed).

Рис. 4.14. Результати кореляційного аналізу

Розглянемо далі приклад параболічної моделі. Першу вибірку згенеруємо так само, як і у попередньому випадку. Потім згенеруємо елементи другої вибірки, використовуючи формулу:

$$=2*A2*A2+3+C3,$$

де A2 – посилання на комірку, де міститься значення відповідного елемента першої вибірки, а C3 – на комірку з елементом рівномірної випадкової послідовності, заданої на відрізку $[-b; b]$.

На рис. 4.15, 4.16 наведено результати розрахунку рангового коефіцієнта кореляції Спірмена для різних значень параметра b, а також відповідні кореляційні поля досліджуваних ознак. Але для коефіцієнта Спірме-

на у цьому випадку розраховували три значення: S – відповідає усієї сукупності даних; S_- – значенням $x < 0$, S_+ – значенням $x > 0$.

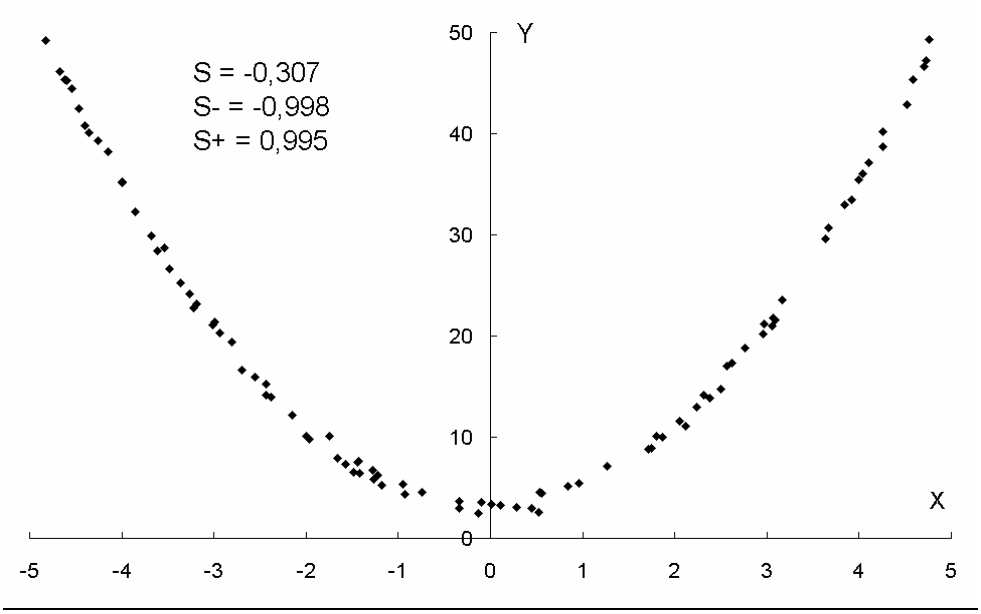


Рис. 4.15. Кореляційне поле досліджуваних ознак при $b = 1$

Безпосереднє застосування процедур кореляційного аналізу пакету SPSS не надає значних переваг при проведенні кореляційного аналізу досліджуваних ознак. Але цей пакет дає змогу визначити коефіцієнт детермінації, що відповідає окремим типам моделей зв'язку між досліджуваними ознаками. Застосування відповідних процедур можливо лише для кількісних даних.

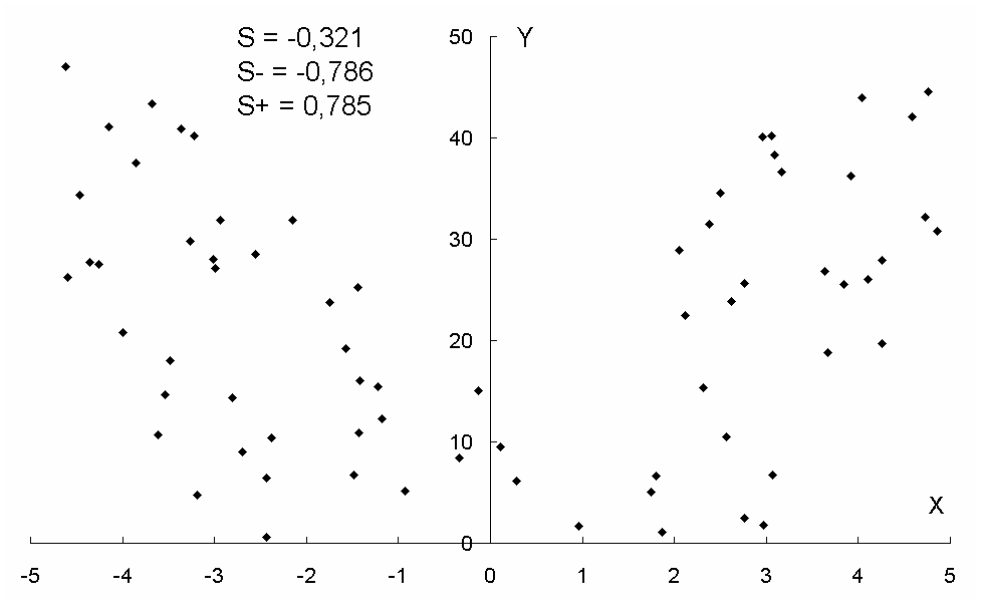


Рис. 4.16. Кореляційне поле досліджуваних ознак при $b = 20$

Розглянемо більш докладно методику визначення коефіцієнта детермінації та інших характеристик моделі зв’язку в пакеті SPSS. У пункті Analyze головного меню обираємо Regression/Curve Estimation. Після цього з’являється діалогове вікно підбору моделі зв’язку (рис. 4.17).

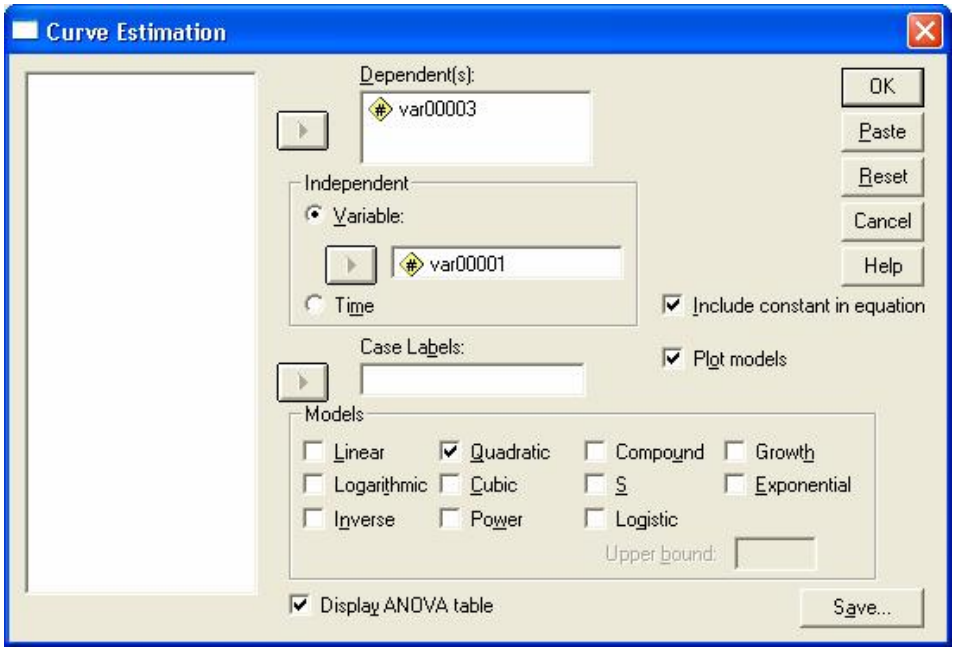


Рис. 4.17. Діалогове вікно підбору моделі зв’язку

У цьому вікні зазначаємо залежну й незалежну змінні, тип моделі, необхідність виведення таблиці ANOVA, графіка, а також збереження результатів на сторінці даних. Результати для розглянутих вище значень параметра b наведено на рис. 4.18–4.21.

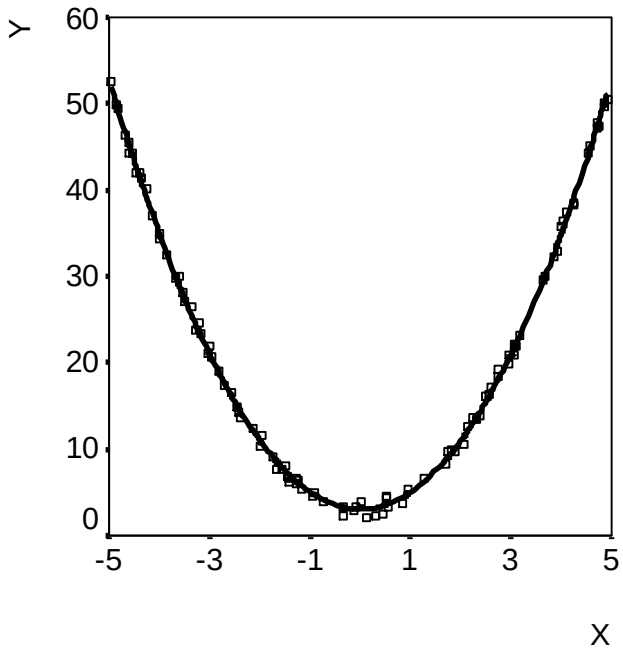


Рис. 4.18. Параболічна модель для $b = 1$

```
MODEL:  MOD_1.

Dependent variable.. VAR00003          Method.. QUADRATI

Listwise Deletion of Missing Data

Multiple R           ,99933
R Square             ,99867
Adjusted R Square    ,99864
Standard Error       ,57182

            Analysis of Variance:

               DF   Sum of Squares      Mean Square

Regression      2        23761,560      11880,77995
Residuals      97           31,717           ,32698

F =   36334,75101       Signif F =  ,0000

-------------------- Variables in the Equation -------------------

Variable                   B        SE B       Beta        T  Sig T

VAR00001     4,24885409E-05     ,018585  8,475E-06      ,002 ,9982
VAR00001**2       1,995714     ,007403    ,999333   269,570 ,0000
(Constant)        3,016919     ,090626               33,290 ,0000
```

Рис. 4.19. Характеристики параболічної моделі для $b = 1$

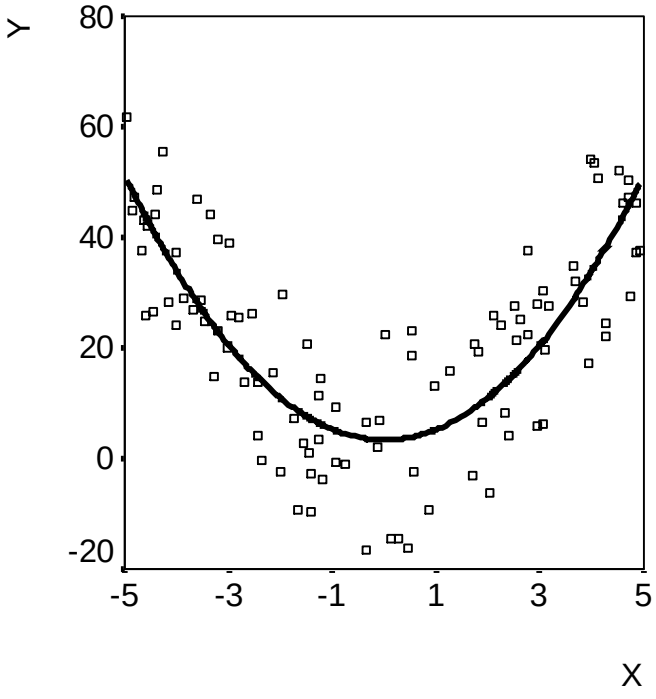


Рис. 4.20. Параболічна модель для $b = 20$

```
MODEL:  MOD_2.

Dependent variable.. VAR00009          Method.. QUADRATI

Listwise Deletion of Missing Data

Multiple R           ,79548
R Square             ,63278
Adjusted R Square    ,62521
Standard Error     11,43645

            Analysis of Variance:

               DF   Sum of Squares      Mean Square

Regression      2        21861,932        10930,966
Residuals      97        12686,870          130,792

F =      83,57488       Signif F =  ,0000

-------------------- Variables in the Equation -------------------

Variable                   B        SE B       Beta        T  Sig T

VAR00008             ,000850     ,371697    ,000141     ,002  ,9982
VAR00008**2         1,914277     ,148066    ,795477    12,928 ,0000
(Constant)          3,338370    1,812522                1,842  ,0686
```

Рис. 4.21. Характеристики параболічної моделі для $b = 20$

Наведені дані свідчать про те, що значення коефіцієнту рангової кореляції, розраховані окремо за спадною та висхідною гілками параболи, є близькими до квадратного кореня з коефіцієнта детермінації. Це підтверджує те, що у даному випадку коефіцієнт рангової кореляції Спірмена є адекватною мірою нелінійного зв'язку між ознаками, якщо він розраховується за інтервалами, що відповідають ділянкам монотонного спадання або згасання функції регресії.

Контрольні питання

1. Що називають кореляцією двох випадкових величин?
2. Які ознаки вважають статистично незалежними?
3. Які проблеми аналізу даних потребують попередньої перевірки наявності статистичного зв'язку між досліджуваними ознаками?
4. Якою є загальна методика перевірки гіпотези про наявність статистичного зв'язку?
5. З якою метою на початковому етапі кореляційного аналізу перевіряють тип даних?
6. Що є універсальною характеристикою статистичного зв'язку між кількісними ознаками?

7. Якими є переваги й недоліки застосування коефіцієнта детермінації?

8. Для заданого набору даних визначити коефіцієнт детермінації і зробити висновок про наявність кореляційного зв'язку.

9. У чому полягає різниця між парними та частинними кореляційними характеристиками?

10. Що вимірює парний коефіцієнт кореляції Пірсона?

11. Для заданого набору даних розрахувати значення парного коефіцієнта кореляції Пірсона і зробити висновок про наявність кореляційного зв'язку.

12. Що називають кореляційним відношенням двох випадкових величин? Які властивості зв'язку характеризує цей показник?

13. Що вимірює коефіцієнт кореляції Фехнера?

14. Для заданого набору даних розрахувати значення коефіцієнта кореляції Фехнера і зробити висновок про наявність кореляційного зв'язку.

15. Що називають коваріацією випадкових величин? Як цей показник пов'язаний з парним коефіцієнтом кореляції Пірсона?

16. Яких значень можуть набувати показники корельованості кількісних ознак? Які висновки можна зробити на основі значень цих показників?

17. Що називають ранговою кореляцією?

18. На якій властивості корельованих ознак ґрунтується коефіцієнт рангової кореляції Спірмена?

19. На якій властивості корельованих ознак ґрунтується коефіцієнт рангової кореляції Кендалла?

20. Яких значень можуть набувати показники рангової кореляції і про що свідчать їх значення?

21. Для заданого набору даних розрахувати значення коефіцієнтів рангової кореляції Спірмена й Кендалла і зробити висновок про наявність кореляційного зв'язку.

22. Які показники використовують для перевірки корельованості номінальних ознак? У чому полягають особливості застосування окремих показників?

23. Для заданого набору даних перевірити корельованість даних за допомогою критерію χ^2, а також коефіцієнта Крамера й поліхоричного коефіцієнта спряженості Чупрова.

24. Які показники використовують для перевірки корельованості ознак, що виміряні у шкалах різного типу? Якими є особливості їх застосування?

25. Які показники використовують для дослідження корельованості декількох ознак? Якими є особливості їх застосування?