3. Кореляційний та регресійний аналіз

3.1 Види зв'язків між явищами

Сучасна наука ґрунтується на допущенні про те, що всі явища у природі та суспільстві пов'язані між собою. Тому практично важливим є завдання кількісного вимірювання цих зв'язків та побудови їх математичних моделей, що дозволяють прогнозувати вплив одного фактору на інші. Неможливо керувати процесами, прогнозувати їх розвиток без вивчення характеру та щільності зв'язків між факторами, що впливають на них.

При вивченні різних природних та суспільних процесів ми виокремлюємо у них основні фактори, що впливають на їх розвиток. У статистиці фактори, що обумовлюють зміну інших, пов'язаних з ними факторів, називають *факторними ознаками*, фактори, що змінюються під впливом факторних ознак, називають *результативними ознаками*.

Розрізняють два типи зв'язків: функціональні та стохастичні. *Функціональним* називають такий зв'язок, при якому кожному значенню факторної ознаки або впорядкованому набору значень факторних ознак за деяким правилом ставлять у відповідність одне значення результативної ознаки. Функціональний зв'язок можна подати у математичній формі:

$$y = f\left(x\right) \tag{3.1}$$

або

$$y = f\left(x_1, x_2, ..., x_n\right). \tag{3.2}$$

Формула (3.1) – це модель однофакторної функціональної залежності, коли результативна ознака змінюється під впливом лише однієї факторної ознаки. Формула (3.2) – це модель багатофакторної функціональної залежності. Тут кожному впорядкованому набору *n* факторних ознак $\left(x_1, x_2, ..., x_n\right)$ відповідає єдине значення результативної ознаки *y*.

При функціональній залежності відомий повний набір факторів, що визначають значення результативної ознаки, та механізм їх впливу, записаний у вигляді конкретного рівняння. Модель багатофакторної функціональної залежності можна уточнювати, додаючи до неї нові факторні ознаки.

У моделях залежностей у вигляді функціональних зв'язків не враховують дію випадкових факторів. Крім того, у багатьох випадках нам невідомі всі фактори, що впливають на значення результативного фактору, а відомі фактори часто

вимірюють з похибками. У всіх цих випадках виникає невизначеність і для дослідження взаємозв'язків між явищами використовують статистичні моделі.

Статистичний зв'язок – це зв'язок між факторами, при якому зміні значення факторної ознаки відповідає зміна закону розподілу результативної ознаки. Значення результативної ознаки у цьому випадку не можна вказати точно, а лише з певною ймовірністю. Різні значення результативної ознаки – це реалізації деякої випадкової величини. У загальному випадку модель статистичного зв'язку можна записати у вигляді:

$$\tilde{y} = f\left(x_1, x_2, ..., x_n\right) + \Delta f\left(\Delta x_1, \Delta x_2, ..., \Delta x_n\right) + E\left(z_1, z_2, ..., z_m\right). \qquad (3.3)$$

У формулі (3.3) $x_1, x_2, ..., x_n$ – враховані факторні ознаки, $\Delta x_1, \Delta x_2, ..., \Delta x_n$ – похибки їх вимірювання, Δf – похибка моделі, обумовлена їх наявністю, $z_1, z_2, ..., z_m$ – невраховані фактори, E – похибка моделі, обумовлена наявністю неврахованих факторів, $\tilde{y}$ – розрахункове (теоретичне) значення результативної ознаки.

Окремим випадком статистичного зв'язку є *кореляційний зв'язок*, за наявності якого всі взаємопов'язані фактори є випадковими величинами. При кореляційному зв'язку у залежності від зміни факторної ознаки чи впорядкованого набору факторних ознак змінюється середнє значення результативної ознаки.

При дослідженні кореляційних залежностей потрібно вирішити наступні завдання:

1) попередній аналіз властивостей об'єкта моделювання;
2) визначення факту наявності кореляційного зв'язку, його форми та напряму;
3) вимірювання щільності зв'язку між факторами;
4) побудова математичної моделі зв'язку у вигляді рівняння, що пов'язує факторні та результативну ознаки;
5) оцінка адекватності отриманої моделі та її змістовна інтерпретація.

У залежності від напряму дії розрізняють прямі та обернені зв'язки. За наявності *прямого зв'язку* зі збільшенням факторної ознаки результативна ознака також збільшується, при зменшенні факторної ознаки – зменшується. Якщо збільшення факторної ознаки супроводжується зменшенням результативної ознаки, маємо *обернений зв'язок*.

У залежності від вигляду формули, що встановлює зв'язок між факторними та результативною ознаками розрізняють лінійні та нелінійні зв'язки. Лінійна залежність між факторною ознакою x та результативною ознакою y має вигляд $y = ax + b$, її графіком є пряма. Якщо факторних ознак кілька, то лінійна залежність має вигляд: $y = a_1x_1 + a_2x_2 + ... + a_nx_n$. Інші типи зв'язків є нелінійними.

За кількістю факторів, що впливають на результативну ознаку, розрізняють однофакторні та багатофакторні зв'язки. Однофакторні зв'язки називають також

парними. Якщо зв’язок багатофакторний, то всі фактори діють одночасно і у взаємозв’язку.

При дослідженні статистичних взаємозв’язків застосовують співставлення значень факторів та результативної ознаки, метод аналітичних групувань, кореляційний аналіз, регресійний аналіз та непараметричні методи. Найпростішим способом виявлення наявності зв’язку є *співставлення значень факторної ознаки та відповідних їм значень результативної ознаки*. Значення факторної ознаки записують у порядку зростання та аналізують характер зміни відповідних значень результативної ознаки. Недоліком такого підходу є неможливість знаходження кількісної міри зв’язку між даними факторами.

Статистичний зв’язок проявляється більш чітко, якщо для його вивчення застосувати *аналітичне групування* одиниць сукупності, що вивчається, на проміжки у відповідності зі зростанням факторної ознаки і для кожної групи знайти відповідне середнє значення результативної ознаки. Порівняння змін результативної та факторної ознак при аналітичному групуванні дає можливість встановити напрям та щільність зв’язку між ними, але не дозволяє визначити формулу, що встановлює зв’язок між факторами та результатом їх дії.

До задач *кореляційного аналізу* відносять кількісне вимірювання щільностей зв’язку між факторними та результативною ознаками, визначення невідомих зв’язків та оцінку факторів, що найбільше впливають на результативну ознаку.

Метою *регресійного аналізу* є знаходження аналітичного виразу (формули), що встановлює зв’язок між факторними та результативною ознакою. Визначається також ступінь впливу факторних ознак на результативну ознаку. Регресійні моделі можна використовувати для прогнозування значень результативної ознаки.

З допомогою непараметричних методів встановлюють зв’язок між якісними (атрибутивними) ознаками.

3.2 Однофакторний лінійний кореляційний та регресійний аналіз

Методика застосування однофакторної (парної) лінійної кореляції дозволяє визначити наявність впливу однієї факторної ознаки на результативну ознаку, для якого можна побудувати модель зв’язку між цими ознаками у вигляді лінійної функції. Крім того, у багатьох випадках нелінійні моделі можна звести до лінійної функції шляхом логарифмування або заміни змінної.

Перед побудовою лінійної моделі взаємозв’язку між двома факторами доцільно переконатися у наявності між ними лінійного зв’язку. Для цього визначають коефіцієнт кореляції між факторною та результативною ознаками. Розглянемо обчислення його статистичної оцінки.

Нехай ми маємо статистичний матеріал відносно спостережень за деякими двома явищами. За результатами цих спостережень потрібно побудувати лінійну однофакторну модель і кількісно оцінити щільність взаємозв’язку між явищами,

що вивчаються. Сукупності результатів спостережень за ними будемо розглядати як сукупності значень випадкових величин X (факторної ознаки) та Y (результативної ознаки). Нехай значення спостережень факторної ознаки X дорівнюють $x_1, x_2, ..., x_n$, а відповідні їм значення результативної ознаки – $y_1, y_2, ..., y_n$. Оцінка коефіцієнта кореляції між факторною та результативною ознаками здійснюється за формулою

$$r_{xy} = \frac{k_{xy}}{\sigma_x \sigma_y}, \tag{3.4}$$

де k_{xy} – *кореляційний* момент або *коваріація* між факторною та результативною ознаками,

$$k_{xy} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{n}. \tag{3.5}$$

$\overline{x}, \overline{y}$ – середні значення ознак X та Y,

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}, \tag{3.6}$$

σ_x, σ_y – стандартні відхилення цих ознак,

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n}}, \sigma_y = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \overline{y})^2}{n}}. \tag{3.7}$$

Формули (3.5) та (3.7) використовують, якщо кількість спостережень $n > 40$. Якщо $n \le 30$, то у цих формулах у знаменниках n замінюють на n–1.

Можна довести, що $|r_{xy}| \le 1$, причому чим ближчим є $|r_{xy}|$ до одиниці, тим тіснішим є лінійний зв'язок між факторами X та Y. Лінійний зв'язок вважається встановленим, якщо виконується нерівність

$$|r_{xy}| \ge 3\sigma(r_{xy}), \tag{3.8}$$

де $\sigma\left(r_{xy}\right)$ – стандартне відхилення коефіцієнта кореляції. Цей показник визначається за формулою:

$$\sigma\left(r_{xy}\right)=\frac{1-r_{xy}^2}{\sqrt{n}}. \tag{3.9}$$

Якщо кількість спостережень $n < 30$, то отриманий коефіцієнт кореляції перевіряють на істотність з допомогою *t*-критерію. Спочатку знаходять фактичне значення *t*-критерію за формулою

$$t_{факт.}=\frac{\left|r_{xy}\right|\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}. \tag{3.10}$$

Далі знаходять табличне значення *t*-критерію за вхідними даними $v=n-2$ та рівнем значущості α (звичайно вибирають $\alpha=0,05$). Якщо $t_{факт.}>t_{табл.}$, то розрахований коефіцієнт кореляції вважають істотним з ймовірністю 95%, тобто на його основі можна приймати рішення щодо наявності чи відсутності лінійного зв'язку між факторами.

Квадрат коефіцієнту кореляції називають *коефіцієнтом детермінації*. Його можна застосовувати для вимірювання щільності не лише лінійних, але й нелінійних залежностей. Його часто вимірюють у процентах.

Якщо, виходячи з розрахованого значення коефіцієнту кореляції, ми встановили наявність лінійної залежності між факторами *X* та *Y,* то далі визначаємо вигляд цієї залежності. Рівняння парної лінійної регресії має вигляд:

$$\tilde{y}_i=ax_i+b,\quad i=1,2,...,n. \tag{3.11}$$

Тут x_i – результати спостережень факторної ознаки *X*, $\tilde{y}_i$ – розраховані за регресійною моделлю (теоретичні) значення результативної ознаки, *a* та *b* – параметри регресії, які потрібно визначити. Параметр *a* у рівнянні регресії (3.11) показує, на скільки одиниць у середньому зміниться результативна ознака, якщо факторна ознака зміниться на одну одиницю її вимірювання. Цей параметр називають також *коефіцієнтом регресії*. При $a > 0$ у рівнянні (3.11) зв'язок між факторами є прямим, при $a < 0$ – оберненим. Параметри *a* та *b* можна знайти методом найменших квадратів за формулами, аналогічними до формул для визначення коефіцієнтів моделі лінійного тренду, розглянутих при вивченні часових рядів:

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}, \; b = \overline{y} - a\overline{x}\,. \qquad (3.12)$$

Приклад 3.1. У таблиці наведено дані про крадіжки вогнепальної зброї у деякому місті та про кількість злочинів, вчинених з її застосуванням, Побудувати лінійну регресійну модель взаємозв'язку між цими факторами.

Таблиця 3.1. Динаміка крадіжок вогнепальної зброї та злочинів, вчинених з її застосуванням у 2016-2021 рр.

Рік	2016	2017	2018	2019	2020	2021
Крадіжки вогнепальної зброї, *X*	773	1138	1396	1352	1336	1130
Злочини, вчинені з застосуванням вогнепальної зброї, *Y*	4481	8873	19154	18059	12160	9549

Нехай *X* – кількість крадіжок вогнепальної зброї, *Y* – кількість злочинів, вчинених з її застосуванням. Розташуємо дані таблиці 3.1 у порядку зростання фактору *X*. Отримаємо таблицю 3.2.

Таблиця 3.2. Крадіжки вогнепальної зброї *X* та злочини, вчинені з її застосуванням *Y* у 2016-2021 рр., розташовані у порядку зростання фактору *X*.

Крадіжки вогнепальної зброї, *X*	773	1130	1138	1336	1352	1396
Злочини, вчинені з застосуванням вогнепальної зброї, *Y*	4481	9549	8873	12160	18059	19154

З таблиці 3.2 видно, що зі зростанням факторної ознаки *X* результативна ознака *Y* також в основному зростає, за винятком 2017 р., коли при зростанні *X* фактор *Y* зменшився. З'ясуємо чи можна для моделювання взаємозв'язку між факторами *X* та *Y* використати модель лінійної регресії. Для цього перевіримо наявність кореляції між факторами, тобто розрахуємо коефіцієнт кореляції за формулами (3.4)-(3.7). Тут $n = 6$,

$$\sum_{i=1}^{6} x_i = 7125,\ \sum_{i=1}^{6} y_i = 72276,\ \overline{x} = \frac{7125}{6} = 1187{,}5,\quad \overline{y} = \frac{72276}{6} = 12046,$$

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{6}\left(x_i - \overline{x}\right)^2}{6-1}} = \sqrt{\frac{270151{,}5}{5}} = 232{,}4,$$

$$\sigma_y = \sqrt{\frac{\sum_{i=1}^{6}\left(y_i - \overline{y}\right)^2}{6-1}} = \sqrt{\frac{1{,}6022498\cdot 10^8}{5}} = 3660{,}8,$$

$$k_{xy} = \frac{\sum_{i=1}^{6}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{6-1} = \frac{5924419}{5} = 118483{,}8,\ r_{xy} = \frac{k_{xy}}{\sigma_x \sigma_y} = \frac{1184883{,}6}{232{,}4\cdot 5660{,}8} = 0{,}9.$$

Знайдений коефіцієнт кореляції є близьким до 1. Перевіримо його істотність, для цього розрахуємо значення *t*-критерію:

$$t_{факт.} = \frac{\left|r_{xy}\right|\sqrt{n-2}}{\sqrt{1-r_{xy}^2}} = \frac{0{,}9\cdot\sqrt{4}}{\sqrt{1-\left(0{,}9\right)^2}} = 4{,}13.$$

Знайдемо табличне значення *t*-критерію при $v = n - 2 = 4$ та рівню значущості $\alpha = 0{,}05$: $t_{табл.} = 2{,}78$, $t_{факт.} > t_{табл.}$. Коефіцієнт кореляції є істотним. Оскільки він дорівнює 0,9, тобто близький до одиниці, можна прийняти гіпотезу про наявність лінійної кореляції між факторами *X* та *Y*.

Побудуємо лінійну регресійну модель $y = ax + b$. Використавши дані таблиці 3.2, за формулами (3.12) знаходимо коефіцієнти *a* та *b*:

$$\sum_{i=1}^{6} x_i = 7125,\ \sum_{i=1}^{6} y_i = 72276,\ \sum_{i=1}^{6} x_i \sum_{i=1}^{6} y_i = 5{,}149665\cdot 10^8,$$

$$\sum_{i=1}^{6} x_i y_i = 91752169,\ n\sum_{i=1}^{6} x_i y_i = 5{,}5051301\cdot 10^8,\ \sum_{i=1}^{6} x_i^2 = 8731089,$$

$$n\sum_{i=1}^{6} x_i^2 = 52386534,\ \left(\sum_{i=1}^{6} x_i\right)^2 = 50765625.$$

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{5{,}5051301\cdot 10^8 - 5{,}149665\cdot 10^8}{52386534 - 50765625} = 21{,}9,$$

$$b = \overline{y} - a\overline{x} = 12046 - 21{,}9\cdot 1187{,}5 = -13960{,}3.$$

Таким чином, рівняння лінійної регресії має вигляд:

$$\tilde{y}_i = 21{,}9x_i - 13960{,}3.$$

Використовуючи це рівняння та дані таблиці 3.2, знаходимо теоретичні значення $\tilde{y}_i$ результативної ознаки *Y*, наприклад для $x_1 = 773$ маємо:

$$\tilde{y}_1 = 21{,}9 \cdot 773 - 13960{,}3 \approx 2968.$$

Найпростішу оцінку якості отриманої лінійної моделі можна отримати, розрахувавши середню відносну похибку апроксимації. Дані для розрахунку наведені у таблиці 3.3.

Таблиця 3.3. Дані для розрахунку середньої відносної похибки апроксимації

X	Y	$\tilde{Y}$	$\lvert Y - \tilde{Y} \rvert$
773	4481	2968	1513
1130	9549	10787	1238
1138	8873	10962	2089
1336	12160	15298	3138
1352	18059	15649	2410
1396	19154	16612	2542

Середня відносна похибка апроксимації $A = \dfrac{\sum_{i=1}^{n} \lvert y_i - \tilde{y}_i \rvert}{\sum_{i=1}^{n} y_i} = \dfrac{12930}{72276} \approx 0{,}179 = 17{,}9\%$.

Відхилення фактичних значень результативної ознаки від її середнього значення можна подати у вигляді:

$$\sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} \left(y_i - \tilde{y}_i \right)^2 + \sum_{i=1}^{n} (\tilde{y}_i - \overline{y})^2, \qquad (3.13)$$

де $\sum_{i=1}^{n} (y_i - \overline{y})^2$ – *загальна варіація* результативної ознаки, $\sum_{i=1}^{n} (y_i - \tilde{y}_i)^2$ – *залишкова варіація*, $\sum_{i=1}^{n} (\tilde{y}_i - \overline{y})^2$ – *варіація результативної ознаки, що пояснюється рівнянням регресії.*

Рівність (3.13) є основою оцінки якості отриманого рівняння регресії: чим більша частина загальної варіації результативної ознаки пояснюється регресією, тим кращою є лінійна математична модель залежності між факторною та результативною ознаками і факторна ознака для побудови моделі вибрана вірно. Співвідношення варіації результативної ознаки, обумовленої регресією, та її загальної варіації називають *коефіцієнтом детермінації*. Його визначають за формулою:

$$R^2 = \frac{\sum_{i=1}^{n}\left(\tilde{y}_i - \overline{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}. \tag{3.14}$$

Величину

$$\eta = \sqrt{R^2} = \sqrt{\frac{\sum_{i=1}^{n}\left(\tilde{y}_i - \overline{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}} \tag{3.15}$$

називають *теоретичним кореляційним відношенням*.

Його застосовують для оцінки щільності зв'язку при лінійному та нелінійному зв'язках між факторною та результативною ознаками, тобто воно є більш загальною мірою щільності зв'язку між факторами у порівнянні з коефіцієнтом кореляції r_{xy} . Для нелінійних залежностей теоретичне кореляційне відношення називають *індексом кореляції*.

Знайдемо коефіцієнт детермінації у прикладі 3.1:

$$R^2 = \frac{\sum_{i=1}^{n}\left(\tilde{y}_i - \overline{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2} = \frac{1,2957568 \cdot 10^8}{1,6022498 \cdot 10^8} \approx 0,81.$$

Відповідне теоретичне кореляційне відношення $\eta = \sqrt{R^2} = \sqrt{0,81} = 0,9$. Встановлено, що коли $\left|R^2 - r_{xy}^2\right| \le 0,1$, то гіпотеза про наявність лінійного зв'язку між факторами підтверджується. У прикладі 3.1 $\left|R^2 - r_{xy}^2\right| = \left|0,8087 - 0,81\right| = 0,0013 \le 0,1$, тому між факторами *x* та *y* є лінійна кореляційна залежність.

Якщо значення R^2 та r_{xy}^2 суттєво відрізняються між собою, то зв'язок між факторами є нелінійним.

Значення теоретичного кореляційного відношення $0 \le \eta \le 1$. Чим ближче воно до одиниці, тим щільнішим є взаємозв'язок між ознаками.

Оскільки кількість значень факторів у прикладі 3.1 $n < 30$, то потрібно перевірити значущість (суттєвість) отриманих коефіцієнтів *a* та *b* лінійної регресійної моделі $y = ax + b$. Для цього застосовують *t*-критерій.

Перевіримо значущість коефіцієнта *b.* Для цього обчислюємо розрахункове значення *t*-критерію за формулою:

$$t_{b,p.} = |b| \cdot \frac{\sqrt{n-2}}{\sigma_{зал.}}. \tag{3.16}$$

У формулі (3.16) залишкова дисперсія $\sigma_{зал.}$ обчислюється за формулою:

$$\sigma_{зал.} = \sqrt{\frac{\sum_{i=1}^{n} \left(y_i - \tilde{y}_i\right)^2}{n}}. \tag{3.17}$$

Для прикладу 3.1 маємо:

$$\sigma_{зал.} = \sqrt{\frac{\sum_{i=1}^{n} \left(y_i - \tilde{y}_i\right)^2}{n}} = \sqrt{\frac{30302642}{6}} \approx 2247{,}3$$

Отже, $t_{b,p.} = |b| \cdot \frac{\sqrt{n-2}}{\sigma_{зал.}} = \frac{13960{,}3 \cdot \sqrt{6-2}}{2247{,}3} \approx 12{,}42$. Далі знайдене значення $t_{b,p.}$ порівнюють з табличним значенням *t*-критерію. Це значення знаходять у статистичній таблиці *t*-розподілу (розподілу Стьюдента) при вхідних параметрах $v = n - 2$ та вибраному рівню значущості α. Найчастіше вибирають $\alpha = 0{,}05$. У нашому прикладі $v = 6 - 2 = 4$. При таких вхідних параметрах табличне значення *t*-критерію $t_{табл.} = 2{,}78$. Якщо $t_{b,p.} > t_{табл.}$, то параметр *b* вважається значущим. У нашому випадку 12,42 > 2,78, тому параметр *b* є значущим.

Визначимо значущість коефіцієнта *a* лінійної регресійної моделі. Для цього розрахуємо значення *t*-критерію:

$$t_{a,p.} = |a| \cdot \frac{\sqrt{n-2}}{\sigma_{зал.}} \cdot \sigma_x. \tag{3.17}$$

У прикладі 3.1 $\sigma_{зал.} = 2247{,}3$, $\sigma_x = 232{,}4$, $a = 21{,}9$. Отже, отримаємо:

$$t_{a,p.} = 21{,}9 \cdot \frac{\sqrt{6-2} \cdot 232{,}4}{2247{,}3} \approx 4{,}53.$$

Для нашого прикладу $t_{a,p.} > t_{табл.}$, тому коефіцієнт a є значущим.

Для перевірки значущості моделі у цілому застосовують критерій Фішера. Розрахункове значення цього критерію для випадку парної кореляції визначають за формулою:

$$F_{p.} = \frac{r_{xy}^2}{1 - r_{xy}^2}(n-2). \tag{3.19}$$

Знайдене $F_{p.}$ порівнюють з табличним значенням критерія Фішера $F_{табл.}$, яке визначають за заданим рівнем значущості α та параметрами $v_1 = 1$, $v_2 = n - 2$ ($v_1 = 1$ для регресійної моделі з 2 коефіцієнтами, зокрема, парної лінійної регресії). При $v_1 = 1$, $v_2 = 6 - 2 = 4$ та вибраним рівнем значущості $\alpha = 0{,}05$ у статистичній таблиці розподілу Фішера знаходимо $F_{табл.} = 7{,}71$. Якщо $F_{p.} > F_{табл.}$, то рівняння регресії у цілому є значущим. Для прикладу 3.1 маємо:

$$F_{p.} = \frac{0{,}81 \cdot 4}{1 - 0{,}81} \approx 17{,}05.$$

Отже, $F_{p.} > F_{табл.}$, рівняння регресії у цілому для прикладу 3.1 є значущим.

3.3 Нелінійна регресія

Зміну результативної ознаки у багатьох випадках можна відобразити на координатній площині графіком, відмінним від прямої, наприклад, кривою, що за виглядом нагадує параболу, гіперболу або інші криві. При знаходженні рівняння зв'язку між факторною ознакою *x* та результативною ознакою *y* крива регресії вибирається у вигляді кривої, найбільш близької до такого графіка.

Якщо при рівномірному зростанні *x* значення *y* зростають прискорено, тобто коли для рівних значень $\Delta x = x_i - x_{i-1}$ приблизно однаковими є другі різниці рівнів часового ряду: $\Delta^2 y_i = \Delta y_i - \Delta y_{i-1} \approx \text{const}$, де $\Delta y_i = y_i - y_{i-1}$, то для моделювання залежності між факторною та результативною ознаками використовують рівняння параболи тобто квадратичної функції $\tilde{y} = a_0 + a_1 x + a_2 x^2$. Невідомі параметри

a_0, a_1, a_2 цього рівняння знаходимо згідно з методом найменших квадратів шляхом розв'язання нормальної системи рівнянь:

$$\begin{cases} na_0 + a_1\sum_{i=1}^{n} x_i + a_2\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i, \\ a_0\sum_{i=1}^{n} x_i + a_1\sum_{i=1}^{n} x_i^2 + a_2\sum_{i=1}^{n} x_i^3 = \sum_{i=1}^{n} x_i y_i, \\ a_0\sum_{i=1}^{n} x_i^2 + a_1\sum_{i=1}^{n} x_i^3 + a_2\sum_{i=1}^{n} x_i^4 = \sum_{i=1}^{n} x_i^2 y_i. \end{cases} \quad (3.20)$$

Для розв'язання системи (3.20) можна використати системи комп'ютерної алгебри або отримати коефіцієнти параболи за допомогою пакету статистичного аналізу даних STATISTICA.

Приклад. Побудувати модель залежності між урожайністю пшениці та кількістю внесених добрив за даними спостереження на 5 ділянках, наведеними у таблиці 3.4

Таблиця 3.4. Розрахункова таблиця для визначення квадратичної залежності

Внесено мінеральних добрив, ц/га, x	Урожайність, ц/га, y	x^2	x^3	x^4	xy	x^2y	$\tilde{y}$
1	16	1	1	1	16	16	16,2
2	19	4	8	16	38	76	18,5
3	20	9	27	81	60	180	20,4
4	22	16	64	256	88	352	21,9
5	23	25	125	625	115	375	23,0
n=5, Σ =	100	55	225	979	317	1199	100

За даними таблиці 3.4 складемо систему нормальних рівнянь:

$$\begin{cases} 5a_0 + 15a_1 + 55a_2 = 100, \\ 15a_0 + 55a_1 + 225a_2 = 317, \\ 55a_0 + 225a_1 + 979a_2 = 1199. \end{cases}$$

Розв'язавши цю систему, отримаємо:

$$a_0 = 13,41;\ a_1 = 2,98;\ a_2 = -0,214.$$

Отже, квадратична залежність між факторами *x* та *y* має вигляд:

$$\tilde{y} = 13,41 + 2,98x - 0,214x^2 .$$

Підставляючи сюди послідовно значення *x*, отримуємо теоретичні значення результативного фактору $\tilde{y}$, наведені у останньому стовпчику таблиці 3.4.

Якщо між факторною та результативною ознаками спостерігається обернена залежність, то рівняння для неї шукають або у вигляді рівняння прямої, або у вигляді рівняння гіперболи $\tilde{y} = a_0 + \frac{a_1}{x}$. Рівняння гіперболи доцільно використовувати, коли нульове значення результативної ознаки не має сенсу.

Згідно з методом найменших квадратів для знаходження параметрів a_0, a_1 у рівнянні гіперболи використовують систему нормальних рівнянь:

$$\begin{cases} na_0 + a_1 \sum\limits_{i=1}^{n} \frac{1}{x_i} = \sum\limits_{i=1}^{n} y_i, \\ a_0 \sum\limits_{i=1}^{n} \frac{1}{x_i} + a_1 \sum\limits_{i=1}^{n} \frac{1}{x_i^2} = \sum\limits_{i=1}^{n} \frac{y_i}{x_i}. \end{cases}$$

Розв’язавши цю систему, отримуємо значення коефіцієнтів a_0 та a_1 рівняння гіперболічної залежності $\tilde{y} = a_0 + \frac{a_1}{x}$.

3.4. Проблема автокореляції. Критерій Дарбіна-Ватсона

У багатьох випадках для знаходження рівняння парної регресії використовують паралельні часові ряди, тобто розглядають пару послідовностей значень факторної та результативної ознак, взятих за ряд періодів часу. У цьому випадку може спостерігатися певна залежність наступного значення фактору від його попереднього значення. Таку залежність називають *автокореляцією*. Наявність автокореляції може впливати на точність коефіцієнта кореляції та теоретичного кореляційного відношення.

Нехай рівняння лінійної регресії побудовано та має вигляд:

$$y_i = ax_i + b + u_i, i = 1,2,...,n.$$

Тут u_i – похибка моделі лінійної регресії за *i*-й рік, $u_i = y_i - \tilde{y}_i$. Явище автокореляції полягає у тому, що у будь-який і-й рік залишок u_i не є випадковою величиною, а залежить від величини залишку u_{i-1} у попередній період часу.

Внаслідок цього при використанні рівняння регресії для прогнозування можуть виникати значні помилки.

Для визначення наявності чи відсутності автокореляції застосовують критерій Дарбіна-Ватсона (коефіцієнт автокореляції), який розраховують за формулою:

$$d = \frac{\sum_{i=2}^{n} (u_i - u_{i-1})^2}{\sum_{i=1}^{n} u_i^2} \qquad (3.20)$$

Можливі значення цього критерію знаходяться у інтервалі від 0 до 4. Якщо автокореляція залишків відсутня, то $d \approx 2$. На практиці застосування критерію Дарбіна-Ватсона ґрунтується на порівнянні величини *d* з критичними значеннями d_L та d_U, які знаходять з статистичної таблиці критичних значень d_L та d_U для коефіцієнтів автокореляції за критерієм Дарбіна-Ватсона. Входами цієї таблиці є кількість спостережень *n*, число незалежних змінних моделі *k* (для парної лінійної регресії $k = 1$) та рівень значущості α. На практиці найчастіше вибирають $\alpha = 0,05$.

При порівнянні величини *d* з критичними значеннями d_L та d_U можливі три випадки.

1. Якщо $d < d_L$, то гіпотеза про незалежність відхилень відхиляється, наявна додатна автокореляція.
2. Якщо $d > d_U$, то гіпотеза про незалежність відхилень приймається, автокореляція відсутня.
3. Якщо $d_L < d < d_U$, то для прийняття рішення про наявність автокореляції немає достатніх підстав.

При від'ємній автокореляції значення *d* знаходяться у інтервалі від 2 до 4, тому для перевірки автокореляції потрібно визначити величину $d' = 4 - d$ та порівнювати її з критичними значеннями коефіцієнта автокореляції d_L та d_U.

3.5 Багатофакторний лінійний кореляційний та регресійний аналіз

Здебільшого природні та суспільні явища, що є об'єктами дослідження, залежать не від одного, а від кількох факторів. Кореляційну залежність результативної ознаки від кількох факторних ознак називають *множинною регресією*. Головними завданнями, що виникають при побудові моделі множинної регресії (рівняння, що встановлює зв'язок між факторними та результативною ознаками) є наступні:

1) визначення факторних ознак, що найбільше впливають на результативну ознаку;

2) вірний вибір моделі регресії.

Далі будемо розглядати лінійну множинну регресію, що описується рівнянням:

$$\tilde{y}_i = a_1 x_{1i} + a_2 x_{2i} + ... + a_m x_{mi} + b\,, \tag{3.21}$$

Коефіцієнти $a_1, a_2, ..., a_n, b$ цього рівняння знаходять методом найменших квадратів (МНК). Після визначення коефіцієнтів моделі її потрібно перевірити на адекватність, використовуючи кореляційний аналіз.

Розглянемо знаходження коефіцієнтів множинної лінійної регресії за допомогою МНК. Основна ідея цього методу полягає у тому, що підібрати коефіцієнти моделі $a_1, a_2, ..., a_n, b$ так, щоб сума квадратів відхилень фактичних значень результативної ознаки від її теоретичних значень, обчислених за рівнянням моделі, була мінімальною. Отже, потрібно визначити параметри $a_1, a_2, ..., a_n, b$ так, щоб вираз

$$F = \sum_{i=1}\left(y_i - \tilde{y}_i\right)^2 = \sum_{i=1}^{n}\left(y_i - a_1 x_{1i} - a_2 x_{2i} - ... - a_m x_{mi} - b\right)^2 \to \min\,. \tag{3.22}$$

Використавши необхідні умови екстремуму функції, що залежить від m+1 аргументів $a_1, a_2, ..., a_n, b$, тобто прирівнявши до нуля її частинні похідні за цими аргументами, отримуємо систему (3.23) лінійних алгебраїчних рівнянь відносно невідомих коефіцієнтів множинної лінійної регресії:

$$\begin{cases} a_1\sum\limits_{i=1}^{n} x_{1i}^2 + a_2\sum\limits_{i=1}^{n} x_{1i}x_{2i} + ... + a_m\sum\limits_{i=1}^{n} x_{1i}x_{mi} + b\sum\limits_{i=1}^{n} x_{1i} = \sum\limits_{i=1}^{n} x_{1i}y_i, \\ a_1\sum\limits_{i=1}^{n} x_{2i}x_{1i} + a_2\sum\limits_{i=1}^{n} x_{2i}^2 + ... + a_m\sum\limits_{i=1}^{n} x_{2i}x_{mi} + b\sum\limits_{i=1}^{n} x_{2i} = \sum\limits_{i=1}^{n} x_{2i}y_i, \\ ..... \\ a_1\sum\limits_{i=1}^{n} x_{mi}x_{1i} + a_2\sum\limits_{i=1}^{n} x_{mi}x_{2i} + ... + a_m\sum\limits_{i=1}^{n} x_{mi}^2 + b\sum\limits_{i=1}^{n} x_{mi} = \sum\limits_{i=1}^{n} x_{mi}y_i, \\ a_1\sum\limits_{i=1}^{n} x_{1i} + a_2\sum\limits_{i=1}^{n} x_{2i} + ... + a_m\sum\limits_{i=1}^{n} x_{mi} + b\cdot n = \sum\limits_{i=1}^{n} y_i. \end{cases} \tag{3.23}$$

Систему (3.23) називають *системою нормальних рівнянь*.

Розв'язуючи цю систему з використанням стандартних математичних пакетів прикладних програм, отримуємо невідомі коефіцієнти лінійної множинної регресії.

Оскільки на результативну ознаку впливають *m* факторних ознак, то потрібно визначити щільність зв'язку між результативною та кожною з факторних ознак, також необхідно визначити щільність зв'язку між результативною ознакою та всією сукупністю факторних ознак, включених до моделі множинної лінійної регресії. При побудові лінійної однофакторної моделі ми визначаємо один парний коефіцієнт кореляції, для множинної лінійної регресії кількість парних коефіцієнтів кореляції, які необхідно обчислити, дорівнює $C_{m+1}^2 = \frac{m(m+1)}{2}$.

Коефіцієнти парної кореляції між факторними ознаками x_k та x_s, де $k = 1,2,...,m, \quad s = 1,2,...,m, \quad k \neq s$, визначаються за формулами

$$r_{ks} = \frac{\sum_{i=1}^{n}(x_{ki} - \overline{x}_k)(x_{si} - \overline{x}_s)}{n\sigma_k\sigma_s}. \tag{3.24}$$

Коефіцієнти парної кореляції між факторними ознаками x_k, *k*=1,2,..,*m*, та результативною ознакою y обчислюються за формулами:

$$r_{x_k y} = \frac{\sum_{i=1}^{n}(x_{ki} - \overline{x}_k)(y_i - \overline{y})}{n\sigma_k\sigma_y}. \tag{3.25}$$

Стандартні відхилення $\sigma_k = \sigma_{x_k}, k = 1,2,...,m,$ обчислюються за формулою:

$$\sigma_k = \sqrt{\frac{\sum_{i=1}^{n}(x_{ki} - \overline{x}_k)^2}{n}}, \quad k = 1,2,...,n, \sigma_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n}}, \quad \overline{x}_k = \frac{\sum_{i=1}^{n} x_{ki}}{n}, \overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}. \tag{3.26}$$

У формулах (3.24), (3.25), (3.26) $\sigma_k, \sigma_s, \sigma_y$ – стандартні (середні квадратичні) відхилення факторних ознак x_k, x_s та *y*, $\overline{x}_k, \overline{x}_s, \overline{y}$ – їх середні арифметичні. Якщо кількість спостережень $n < 30$, то у знаменниках цих формул замість *n* використовують *n*–1.

Близькість коефіцієнтів парної кореляції до нуля свідчить про відсутність лінійного зв'язку між відповідними змінними.

За отриманими коефіцієнтами парної кореляції можна робити висновки про щільність зв'язку факторних ознак з результативною ознакою та між собою. Тому ці коефіцієнти доцільно використовувати з метою попереднього відбору факторів для включення їх у якості змінних до складу регресійної моделі. Не рекомендується включати до рівняння регресії фактори, що слабко пов'язані з результативною ознакою, але щільно пов'язані між собою. Якщо, наприклад, $r_{yx_1} = 0,85, r_{yx_2} = 0,61, r_{x_1x_2} = 0,92$, то до рівняння регресії доцільно включати фактор x_1, а фактор x_2 не включати, оскільки він тісно пов'язаний з x_1, а його кореляція з у менша, ніж кореляція фактора x_1. Не можна включати до регресійної моделі фактори, функціонально пов'язані між собою, коефіцієнти кореляції між якими близькі до одиниці.

На основі коефіцієнтів парної кореляції обчислюють загальний показник щільності зв'язку всіх факторів, що входять до складу рівняння регресії, з результативною ознакою. – множинний коефіцієнт детермінації R^2. Для випадку лінійної двофакторної множинної регресії

$$y_i = a_1x_{1i} + a_2x_{2i} + b, \quad i = 1,2,...,n,$$

коефіцієнт множинної детермінації обчислюють за формулою:

$$R^2 = \frac{r_{x_1y}^2 + r_{x_2y}^2 - 2r_{yx_1}r_{yx_2}r_{x_1x_2}}{1 - r_{x_1x_2}^2}.$$

Коефіцієнт множинної детермінації визначає частку зміни результативної ознаки, викликаної зміною факторних ознак, що входять до складу рівняння множинної лінійної регресії.