МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Національний аерокосмічний університет ім. М. Є. Жуковського
«Харківський авіаційний інститут»

Д. С. Ревенко, В. О. Либа

МЕТОДИ ЕКОНОМІКО-СТАТИСТИЧНИХ ДОСЛІДЖЕНЬ

Навчальний посібник

Харків «ХАІ» 2014

УДК 330.45 (075.8)
ББК 65.05я73
Р32

Рецензенти: канд. екон. наук, доц. О. П. Мельникова,
канд. екон. наук, доц. О. П. Колісник

Ревенко, Д. С.

Р32 Методи економіко-статистичних досліджень [Текст] : навч. посіб. / Д. С. Ревенко, В. О. Либа. – Х.: Нац. аерокосм. ун-т ім. М. Є. Жуковського «Харк. авіац. ін.-т», 2014. – 64 с.

Розглянуто загальні методичні основи проведення економіко-статистичних досліджень економічних процесів і явищ. Описано методи аналізу динаміки, а також моделі й методи багатовимірного статистичного моделювання (кластерний, факторний і дискримінаційний аналіз).

Аналітичні можливості та межі застосування конкретних моделей і методів проілюстровано на прикладах.

Для студентів, спеціалістів, магістрантів, аспірантів економічних спеціальностей вищих навчальних закладів, що вивчають курс «Методи економіко-статистичних досліджень». Може бути корисним для системних аналітиків і фахівців у сферах маркетингових досліджень і фінансово-економічної діяльності.

Іл. 9. Табл. 12. Бібліогр.: 17 назв

УДК 330.45 (075.8)
ББК 65.05я73

ВСТУП

У сучасній економічній науці широко використовуються економіко-статистичні методи й моделі як для вирішення прикладних, практичних задач, так і для теоретичного моделювання соціально-економічних явищ і процесів. Економіко-статистичні методи стали складовою частиною методів економічної науки. Використання цих методів у поєднанні з обґрунтованим економічним аналізом надає нових можливостей для економічної науки й практики.

Економіко-статистичний аналіз даних стає невід'ємним атрибутом системи керування на всіх рівнях господарювання. Статистичні моделі використовують для діагностики стану об'єктів керування, при вивченні причиново-наслідкового механізму формування варіації й динаміки соціально-економічних явищ і процесів, у моніторингу економічної кон'юнктури, при прогнозуванні й прийнятті оптимальних управлінських рішень.

Сучасний економіст повинен добре знатися на економіко-статистичних методах досліджень, уміти їх використовувати на практиці для моделювання реальних економічних ситуацій, що сприятиме підвищенню рівня кваліфікації і загальній професійній культурі фахівця.

Вивчення курсу «Методи економіко-статистичних досліджень» базується на знаннях, отриманих з таких дисциплін, як вища математика, теорія ймовірностей, економетрика, статистика, економіко-математичне моделювання, економічний аналіз, мікро- і макроекономіка. Набуті знання також може бути використано при вивченні таких дисциплін, як наукове стажування магістра, економічна діагностика, маркетингові дослідження, при виконанні дипломних робіт спеціаліста і магістра за спеціальностями «Економіка підприємства» і «Маркетинг».

Метою курсу є вивчення теоретичних основ і можливостей практичного застосування методів аналізу часових рядів, багатовимірного статистичного аналізу для дослідження економічних систем різного рівня.

Предметом дисципліни є економіко-статистичне моделювання систем і процесів на базі методів аналізу часових рядів і багатовимірних кількісних методів визначення тенденцій, вивчення взаємозв'язків, побудови типології, класифікації і латентних структур у просторі факторів економічного середовища.

1. СТРУКТУРА ЧАСОВИХ РЯДІВ І ТРЕНДОВІ МОДЕЛІ

1.1. Часові ряди і їхні компоненти

Вивчення процесів змінення різних економічних явищ у часі – одна з найбільш важливих задач економіко-статистичних досліджень. Ця задача вирішується шляхом складання й аналізу рядів динаміки (іноді їх також називають часовими, або хронологічними, рядами).

Ряд динаміки є числовими значеннями певного статистичного показника в послідовні моменти або періоди часу (тобто їх розташовано в хронологічному порядку).

Числові значення того чи іншого статистичного показника, що становлять ряд динаміки, називають рівнями ряду й зазвичай позначають через y. Перший член ряду y_0 (або y_1) називають початковим рівнем, а останній y_n – кінцевим. Моменти або періоди часу, до яких належать рівні, позначають через t.

З практики дослідження динаміки явищ і прогнозування випливає, що значення часових рядів можуть містити такі компоненти (складові частини або структуротвірні елементи):

- тренд;
- сезонна компонента;
- циклічна компонента;
- випадкова компонента.

Під трендом розуміються зміни, що визначають загальний напрямок розвитку, основну тенденцію часового ряду. Це – систематична складова довгострокової дії.

Нарівні з довгостроковими тенденціями в часових рядах часто виникають коливання – періодичні складові рядів динаміки.

Якщо періодичні коливання не перевищують одного року, то їх називаю сезонними. Найчастіше причиною їх виникнення є причинні умови. При більшому періоді коливання вважається, що в часових рядах має місце циклічна складова. Прикладами можуть бути цикли ділової активності (які дослідив Н. Кондратьєв), демографічні, інвестиційні й ін.

В економічних часових рядах нечасто є можливість для виокремлення й подальшого аналізу циклічної компоненти, тому що ряди динаміки економічних показників найчастіше виявляються занадто «короткими» для проведення такого дослідження.

Якщо з часового ряду видалити тренд і періодичні складові, то залишиться нерегулярна компонента.

Економісти поділяють фактори, під дією яких формується нерегулярна компонента, на два види: різкої (раптової) дії і поточні.

Якщо часовий ряд подати у вигляді суми відповідних компонент, то отримана модель має назву адитивної:

$$y_t = u_t + s_t + v_t + \varepsilon_t, \qquad (1.1)$$

де y_t – рівні часового ряду;

u_t – трендова компонента;

s_t – сезонна компонента;

v_t – циклічна компонента;

ε_t – випадкова компонента,

а якщо часовий ряд подати у вигляді добутку, то модель називають мультиплікативною:

$$y_t = u_t s_t v_t \varepsilon_t . \quad (1.2)$$

Можна також виокремити ще один вид моделі змішаного типу:

$$y_t = u_t s_t v_t + \varepsilon_t . \quad (1.3)$$

На рис. 1.1 показано приклад часового ряду, у якому є наведені компоненти.

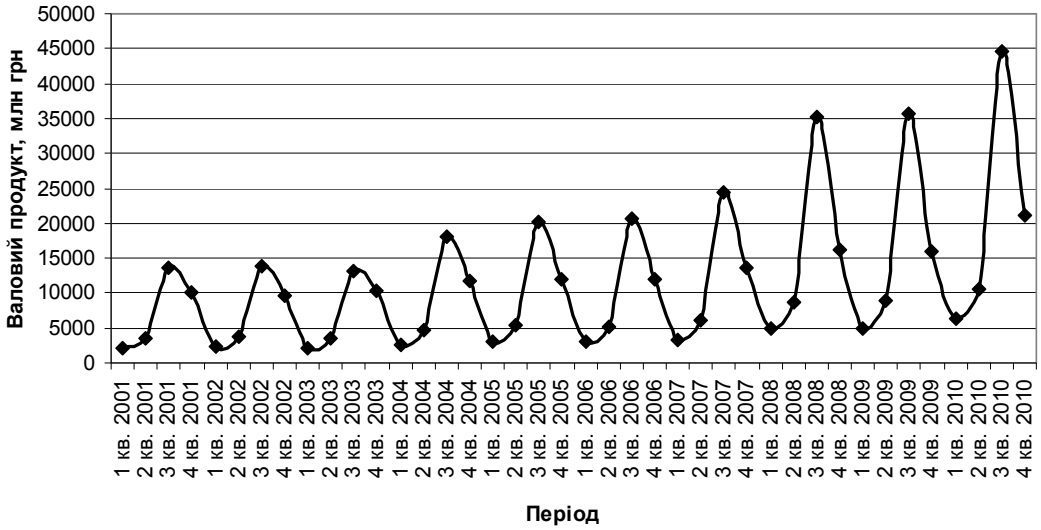


Рис. 1.1. Валовий продукт сільського господарства в Україні

Зазначимо, що в процесі формування значень рівнів кожного часового ряду не обов'язково мають брати участь одночасні компоненти. При зміненні значень одного показника може не бути трендової складової, іншого – періодичних складових, динаміка третього показника може описуватися лише випадковою складовою. Однак наявність випадкової, нерегулярної складової обов’язково є в усіх випадках.

1.2. Види моделей тренду

Для аналізу тенденцій на основі моделей тренду широко використовується залежність, що має назву рівняння тренду:

$$\overline{y}_t = f(t) + \varepsilon_t , \quad (1.4)$$

де $f(t)$ – детермінована невипадкова компонента процесу.

Сьогодні в літературі є опис кількох десятків моделей тренду, які умовно можна поділити на три класи залежно від типу динаміки розвитку процесів.

До І класу належать моделі, які використовуються для опису процесів з монотонним характером тенденції розвитку й відсутністю меж зростання. Ці умови справджуються для багатьох економічних показників, наприклад для більшості показників промислового виробництва в натуральному вираженні.

До ІІ класу належать моделі, з допомогою яких описують процес з межами зростання в досліджуваному періоді. З такими процесами часто стикаються в демографії, при вивченні потреб у товарах і послугах, дослідженні ефективності використання ресурсів і т. ін.

До ІІІ класу належать моделі, які називають кривими насичення, або S-подібними кривими, що мають точки перегину.

Найпростішим типом лінії тренду є пряма лінія, яка описується лінійним (тобто першого ступеня) рівнянням тренду:

$$\bar{y}_t = a + bt, \qquad (1.5)$$

де $\bar{y}_t$ – вирівняні, тобто позбавлені коливань, рівні тренду;

a – вільна складова рівняння, чисельно дорівнює середньому вирівняному рівню для моменту або періоду часу, взятого за початок відліку;

b – середня величина зміни рівня ряду за одиницю часу;

t – номер моментів або періодів часу, до яких належать рівні часового ряду (дата, місяць, квартал, рік).

Середнє змінення рівнів ряду за одиницю часу – головний параметр і константа прямолінійного тренду. Отже, цей тип тренду є придатним для кожного відображення тенденції приблизно одночасних змін: таких, що дорівнюють абсолютному приросту або абсолютному зниженню рівнів за однакові інтервали часу. З практики випливає, що такий характер динаміки трапляється досить часто.

Графічне зображення прямолінійного тренду – пряма лінія в системі прямокутних координат з лінійним (арифметичним) масштабом на обох осях (парабола 1-го порядку).

Під терміном «параболічний тренд» будемо розуміти тренд, виражений параболою 2-го порядку з рівнянням

$$\bar{y}_t = a + bt + ct^2. \qquad (1.6)$$

Параболи 3-го й більш високих порядків нечасто застосовують для відображення економічних тенденцій динаміки, вони є занадто складними для одержання надійних оцінок параметрів при обмеженій довжині часового ряду. Пряму лінію можна вважати одним із видів парабол – параболою 1-го порядку, яку вже було розглянуто раніше.

Значення параметрів параболи 2-го порядку такі: вільний член a – це середній рівень тренду на момент або період, взятий за початок відліку

часу, b – середній рівень тренду за весь періоду, приріст, який вже не є константою, а змінюється рівномірно із середнім прискоренням, що дорівнює $2c$, яке і є константою, головним параметром параболи 2-го порядку.

Тренд у вигляді параболи 2-го порядку застосовують для відображення таких тенденцій динаміки, яким властиве приблизно постійне прискорення абсолютних змін рівнів. Процеси такого роду трапляються на практиці набагато рідше, ніж процеси з однаковим зміненням, але, з іншого боку, будь-яке відхилення процесу від строго рівномірного зростання (або зниження) рівнів можна інтерпретувати як наявність прискорення. Більш того, існує чітке математичне правило: чим вищий порядок параболи, тим ближче лінія тренду до рівнів вихідного часового ряду. Якщо це правило довести до крайньої межі, то будь-який ряд з n рівнів можна точно відобразити параболою $(n-1)$-го порядку.

Рівняння тренду у вигляді параболи 2-го порядку застосовують до різних економічних процесів, які на деякому (зазвичай нетривалому) етапі розвитку мають приблизно постійне прискорення абсолютного приросту рівнів.

Тренд, виражений рівняннями

$$\overline{y}_t = ak^t, \tag{1.7}$$

$$\overline{y}_t = a + bk^t, \tag{1.8}$$

називають експонентним. Вільний член експоненти a дорівнює вирівняному рівню, тобто рівню тренду в момент або період часу, взятий за початок відліку. Основний параметр експонентного тренду k є постійним темпом змінення рівнів. Якщо $k > 1$, то маємо тренд з рівнями, що зростають, причому це зростання не просто прискорене, а зі зростальним прискоренням і зростальними похідними все більш високих порядків. Якщо $k < 1$, то має місце тренд, що є вираженням тенденції постійного зниження рівнів, які вповільнюються, причому дедалі сильніше.

Експонентний тренд є характерним для процесів, що відбуваються в середовищі, яке не створює ніяких обмежень для зростання рівня. Із цього випливає, що на практиці він може відбуватися тільки на обмеженому інтервалі часу, оскільки будь-яке середовище рано чи пізно створює обмеження, будь-які ресурси є вичерпними.

З усіх форм гіпербол розглянемо тільки найпростішу:

$$\overline{y}_t = a + \frac{b}{t}. \tag{1.9}$$

Якщо основний параметр гіперболи $b > 0$, то цей тренд є вираженням тенденції вповільнення зниження рівнів. Таким чином, вільний член гіперболи – це границя, до якої наближається рівень тренду.

Якщо параметр $b < 0$, то зі збільшенням t рівні тренду зростають і наближаються до величини a.

Якщо досліджуваний процес приводить до уповільнення збільшення якогось показника, але при цьому не наближається до якої-небудь границі, то гіперболічна форма тренду вже не є придатною. Тим більше не є придатною парабола з від'ємним прискоренням, по якій зростання, що вповільнюється, перетвориться згодом на зменшення рівнів. У цьому випадку тенденцію змінення найкраще можна відобразити логарифмічним трендом:

$$\bar{y}_t = a + b\ln(t). \qquad (1.10)$$

Логарифми збільшуються значно повільніше, ніж самі числа (номера періодів t), але це збільшення не є обмеженим. Добираючи початок відліку періодів (моментів) часу, можна знайти таку швидкість зменшення абсолютних змін, які найкраще відповідають фактичному ряду.

Якщо вплив обмежувального фактора починає виявлятися тільки після певного моменту (точки перегину), до якого процес розвивався за деяким експонентним законом, то для вирівнювання використовують S-подібні криві, найбільш відомі з яких – крива Гомперца й логістична крива (крива Перла – Ріда), яку можна описати формулою

$$\bar{y}_t = ka^{b^t}. \qquad (1.11)$$

Логістична форма тренду є придатною для опису такого процесу, при якому досліджуваний показник проходить повний цикл розвитку від нульового рівня спочатку повільно, але з прискоренням збільшується, потім прискорення стає нульовим у середині циклу, тобто збільшення уповільнюється за гіперболою в міру наближення до граничного значення показника.

1.3. Методи ідентифікації тренду у часовому ряді, його виду і параметрів

Процедура дослідження динаміки з використанням трендових моделей містить такі етапи:

– вибір однієї або кількох трендових моделей, форма яких відповідає характеру зміни часового ряду;

– ідентифікація трендової моделі;

– оцінювання параметрів вибраної трендової моделі;

– перевірка адекватності вибраної трендової моделі, оцінювання точності моделі й остаточний її вибір;

– розрахунок точкових або інтервальних теоретичних рівнів досліджуваного ряду, прогнозування.

Розв'язання будь-якої задачі з аналізу й прогнозування часових рядів починається з будування графіка досліджуваного процесу, тим більше що сучасні програмні засоби надають для цього користувачу великі можливості. При цьому не завжди чітко простежується наявність тренду в часовому ряді. У цих випадках перш ніж перейти до визначення тенденції

й виділення тренду, потрібно з'ясувати, чи існує взагалі тенденція в досліджуваному процесі.

Основні підходи до розв'язання цієї задачі базуються на статистичній перевірці гіпотез, критерії виявлення компонент ряду – на перевірці гіпотези про випадковість ряду, тобто по суті на статистичній перевірці гіпотези

$$H_0 : My(t) = a = const. \tag{1.12}$$

Розглянемо критерій серій, який часто використовується на практиці для перевірки наявності або відсутності тренду.

Застосування критерію серій, що базується на медіані вибірки, можна подати у вигляді послідовності кроків.

1. З вихідного ряду з рівнями $y_1, y_2, ..., y_n$ утворюють ранжований (варіаційний) ряд $y'_1, y'_2, ..., y'_n$, де y'_1 – найменше значення з рівнів вихідного ряду $y_1, y_2, ..., y_n$, де n – довжина часового ряду.

2. Визначають медіану (Me) цього варіаційного ряду. У випадку непарного значення довжини ряду n ($n = 2m+1$) $Me = y'_{m+1}$, у протилежному випадку ($n = 2m$) $Me = (y'_m + y'_{m+1})/2$.

3. Утворюють послідовність δ_l із плюсів і мінусів за таким правилом:

$$\delta_l = \begin{cases} +, \text{ якщо } y_t > Me, t = 1, 2, ..., n; \\ -, \text{ якщо } y_t < Me, t = 1, 2, ..., n. \end{cases} \tag{1.13}$$

Якщо значення рівня вихідного ряду y_t дорівнює медіані, то це значення пропускають. Очевидно, що загальна кількість знаків «+» і «–» заздалегідь є невідомою.

4. Підраховують $v(n)$ – кількість серій у сукупності δ_l, де під серією розуміється послідовність порядку плюсів і мінусів. Один плюс або один мінус теж буде вважатися серією.

Визначають $\tau_{max}(n)$ – довжину найдовшої серії.

5. Перевірка гіпотези ґрунтується на тому, що за умови випадковості ряду (за відсутності систематичної складової) довжина найдовшої серії не повинна бути занадто великою, а загальна кількість серій – занадто малою. Тому для того, щоб не відкидати гіпотезу про випадковість вихідного ряду (про відсутність систематичної складової), мають виконуватися такі нерівності:

$$v(n) > \left[\frac{1}{2}(n + 1 - 1{,}96\sqrt{n+1}\right];$$
$$\tau_{max}(n) < [1{,}431 \cdot ln(n+1)]. \tag{1.14}$$

Після підтвердження гіпотези про існування тренду в досліджуваному ряді наступною логічною задачею буде ідентифікація його типу, тобто знаходження кривої, з допомогою якої можна описати цей часовий ряд.

Одним із найбільш легких і популярних методів ідентифікації типу тренду є метод характеристик, що ґрунтується на розрахунку для заданого ряду спостережень $y_1, y_2, ..., y_n$ різних характеристик (моментів різних порядків, темпів зростання різниць, зворотних величин і т. ін.), що описують (ідентифікують) різні класи трендових моделей. Якщо за даними y_t буде визначено характеристику, що буде постійною (приблизно), то з її допомогою можна відновити клас кривих, які описують поводження тренду. Наприклад, якщо обчислення різниці першого порядку $u_t^{(1)} = y_{t+1} - y_t$ будуть постійними, то тренд можна описати лінійною функцією (1.5).

Відразу постає питання: що означає вираз «обчислені значення характеристик приблизно дорівнюють або є постійними»? Очікувати їхньої повної рівності за наявності значень рівнів y_t завад ε_t не доводиться. Тут треба вирішити два завдання:

– перед обчисленням характеристик спробувати зменшити вплив завад ε_t на значення рівнів;

– знайти критерій для визначення ступеня сталості характеристик.

Перше завдання, зважаючи на адитивність моделі, можна вирішити, замінюючи вихідні значення рівнів y_t на усереднені, отримані з допомогою різних ковзних середніх.

Ступінь сталості характеристик визначається складніше. Тут можна запропонувати використовувати коефіцієнт варіації, який не менше за певний рівень, наприклад 20 … 30 %.

Розглянемо алгоритм методу характеристик, що складається з трьох основних етапів.

1. За заданим рядом рівнів $y_1, y_2, ..., y_n$ обчислимо усереднені їхні значення, використовуючи формули ковзного середнього. Зазвичай використовується проста ковзна середня з інтервалом $m = 2n + 1$. Після усереднення одержимо значення $y_{1-1/2m}, ..., y_{n-1/2m}$.

Операція усереднення виконується для зменшення впливу перешкод.

2. Для одержання усереднених значень обчислимо різні характеристики, з допомогою яких можна ідентифікувати вид тренду.

Нехай є набір усереднених значень $\bar{y}_1, \bar{y}_2, ..., \bar{y}_n$, для яких обчислимо такі характеристики:

а) перші різниці $u_t^{(1)} = \bar{y}_t - \bar{y}_{t-1}$ (характеризують пряму);

б) другі різниці $u_t^{(2)} = u_{t+1}^{(2)} - u_t^{(2)}$ (характеризують параболу);

в) темпи зростання $T_t = \bar{y}_{t+1} / \bar{y}_t$ (характеризують експоненту);

г) темпи зростання перших різниць $S_t = u_{t+1}^{(1)} / u_t^{(1)}$ (характеризують модифіковану експоненту);

д) зворотні значення рівнів $z_t = 1/y_t$;

е) прирости зворотних значень рівнянь $v_t = z_{t+1} - z_t$ (характеризують експоненту);

ж) темпи зростання перших різниць зворотних значень рівнів $w_t = v_{t+1}/v_t$ (характеризують криву Гомперца).

Результати розрахунків за методом характеристик рекомендується зводити в таблицю, в останньому рядку якої наводиться клас функцій, які можна використовувати для моделювання тренду при сталості відповідної характеристики.

3. На цьому етапі визначають, які з характеристик можна вважати приблизно постійними. Оскільки обчислені характеристики, як і рівні ряду, є випадковими величинами, то й перевірку їх сталості треба проводити з урахуванням можливого розкиду значень. Ураховуючи, що остаточне вирішення питання про адекватність моделі на цьому етапі не проводиться, на основі візуального аналізу значень характеристик можна вибрати найбільш придатне й запропонувати відповідні моделі для подальшого оброблення, якщо виникає спірне питання, то обчислюється коефіцієнт варіації для характеристик. Якщо для деяких характеристик цей коефіцієнт становить не більше 20 %, то їх можна взяти сталими й запропонувати відповідну модель тренду.

Після визначення типу трендової моделі необхідно знайти параметри цієї моделі.

Найпоширенішим методом знаходження параметрів трендової моделі при вирівнюванні рядів динаміки є метод найменших квадратів (МНК), у якому враховуються всі емпіричні рівні й має забезпечуватися мінімальна сума квадратів відхилень емпіричних значень рівнів y_t від теоретичних $\bar{y}_t$, тобто $\sum_{t=1}^{n}(y_t - \bar{y}_t)^2 \to min$.

Зокрема, при вирівнюванні за прямим видом функції $\bar{y}_t = a + bt$ параметри a і b визначаються шляхом розв'язання системи нормальних рівнянь, отриманої методом найменших квадратів,

$$\begin{cases} na + b\sum_{t=1}^{n} t = \sum_{t=1}^{n} y_t, \\ a\sum_{t=1}^{n} t + b\sum_{t=1}^{n} t^2 = \sum_{t=1}^{n} y_t t, \end{cases} \qquad (1.15)$$

де n – кількість рівнів ряду;

t – порядковий номер в умовному позначенні періоду або моменту часу;

y_t – рівні емпіричного ряду.

Розв'язання цієї системи має такий вигляд:

$$b = \frac{n\sum_{t=1}^{n} y_t t - \sum_{t=1}^{n} t \sum_{t=1}^{n} y_t}{n\sum_{t=1}^{n} t^2 - \left(\sum_{t=1}^{n} t\right)^2}; \quad (1.16)$$

$$a = \frac{\sum_{t=1}^{n} y_t \sum_{t=1}^{n} t^2 - \sum_{t=1}^{n} t \sum_{t=1}^{n} y_t t}{n\sum_{t=1}^{n} t^2 - \left(\sum_{t=1}^{n} t\right)^2}. \quad (1.17)$$

За таким самим принципом оцінюються й параметри параболічного й гіперболічного трендів.

У випадку розрахунків експонентного, логарифмічного й логістичного трендів застосовують логарифмування. При розв'язанні логарифмічного рівняння тренду логарифмують номери періодів (моментів) часу, а при розрахунку параметрів експонентного й логістичного трендів – самі рівні.

1.4. Оцінювання якості й адекватності трендових моделей

Питання про можливість застосування побудованих моделей з метою аналізу й прогнозування явища можна вирішити тільки після перевірки адекватності, тобто відповідності моделі досліджуваному процесу.

На практиці добір виду тренду, параметри якого визначаються методом найменших квадратів, проводиться в більшості випадків емпірично, шляхом будування ряду функцій і порівняння їх між собою за величиною середньоквадратичної похибки:

$$S = \sqrt{\frac{\sum_{t=1}^{n} (y_t - \bar{y}_t)^2}{n - p - 1}}, \quad (1.18)$$

де y_t – фактичні рівні часового ряду;

$\bar{y}_t$ – розрахункові значення рівнів часового ряду, отримані на основі трендової моделі;

n – кількість рівнів ряду;

p – кількість параметрів функції тренду.

Зазвичай перевірку відповідності вибраних трендових моделей реальному процесу будують на аналізі залишкової компоненти, яку отримують після виділення з досліджуваного ряду системних складових, якщо вони є в часовому ряді.

Ряд залишкових компонент буде отримано як відхилення фактичних рівнів ряду від теоретичних:

$$e_t = y_t - \bar{y}_t. \tag{1.19}$$

Уважається, що модель є адекватною описуваному процесу, якщо залишкова послідовність (ряд залишкових компонент) являє собою випадкову компоненту ряду. Тому при оцінюванні «якості» моделі перевіряють, чи задовольняє залишкова послідовність таким властивостям:

– випадковість коливань рівнів ряду;

– відповідність розподілу залишкової компоненти нормальному закону з нульовим математичним очікуванням;

– незалежність значень залишкових компонент ряду.

Відповідність нормальному закону розподілу залишкових компонент можна провести приблизно, наприклад, на основі підходу з урахуванням показників асиметрії A і ексцесу E, які обчислюються за формулами

$$A = \frac{1/n \sum_{t=1}^{n} e_t^3}{\sqrt{\left(1/n \sum_{t=1}^{n} e_t^2\right)^3}}; \tag{1.20}$$

$$E = \frac{1/n \sum_{t=1}^{n} e_t^4}{\sqrt{\left(1/n \sum_{t=1}^{n} e_t^2\right)^3}} - 3. \tag{1.21}$$

Як відомо, при нормальному законі розподілу показники асиметрії й ексцесу дорівнюють нулю.

Розглянемо останню властивість – незалежність значень рівнів ряду залишкових компонент. Якщо вид функції, що описує системну складову, вибрано невдало, то послідовні значення ряду залишків можуть не мати властивості незалежності, тому що можуть корелювати між собою. У цьому випадку кажуть, що існує автокореляція залишкових компонент.

Існує кілька способів виявлення автокореляції залишкових компонент. Найпоширенішим є тест, який спирається на критерій Дарбіна–Уотсона. Цей тест пов'язаний з перевіркою гіпотези про відсутність автокореляції першого порядку r_1, тобто автокореляції між сусідніми залишковими членами ряду. При цьому критична статистика визначається за формулою

$$d = \frac{\sum_{t=2}^{n} (e_t - e_{t-2})^2}{\sum_{t=1}^{n} e_t^2}. \tag{1.22}$$

Можна сказати, що $d \approx 2(1 - r_1)$. Якщо значення статистики d є близьким до нуля, то це означає наявність високої позитивної автокореляції (коефіцієнт r_1 близький до одиниці); близькість значення статистики d до чотирьох означає наявність високої негативної автокореляції (коефіцієнт r_1 близький до –1). У випадку відсутності автокореляції значення статистики d буде близьким до двох (коефіцієнт r_1 не сильно відрізняється від нуля).

У висновку зазначимо, що не може бути чисто формальних підходів до вибору й оцінювання трендових моделей. Успішне застосування трендових моделей на практиці можливе лише при поєднанні знань в області методів із глибоким знанням об'єкта дослідження зі змістовним аналізом досліджуваного явища.

Запитання для самоконтролю

1. З яких структурних компонентів складається часовий ряд?
2. З допомогою якої складової часового ряду можна описати його основну тенденцію?
3. Чим відрізняється сезонна компонента від циклічної?
4. Наявність якої компоненти є обов’язковою в усіх економічних часових рядах?
5. Яка з трендових моделей застосовується на практиці найчастіше?
6. Яким типом трендової моделі можна найбільш точно описати часовий ряд? Назвіть математичне правило.
7. Назвіть основні етапи дослідження часового ряду із застосуванням трендових моделей.
8. Для чого застосовується метод «критерій серій»? Назвіть його основні етапи.
9. Який метод застосовується для визначення типу трендової моделі?
10. Для чого в методі характеристик на другому етапі розраховуються різні статистичні характеристики часового ряду?
11. З допомогою якого статистичного показника визначається стабільність статистичних показників у методі характеристик?
12. Який метод найчастіше використовується при розрахунку параметрів трендових моделей? У чому його суть?
13. Для чого застосовують критерій Дарбіна–Уотсона?

Завдання для самостійної роботи

Завдання 1.1. Використовуючи дані про обсяг реалізованої продукції за 2000–2013 рр., визначити кількість серій, довжину найбільшої з них, перевірити гіпотезу про випадковість вихідного ряду, зробити відповідні висновки.

Таблиця 1.1

Вихідні дані про реалізацію продукції

Рік	2000	2001	2002	2003	2004	2005	2006
Обсяг реалізованої продукції, млн грн	146,9	176,5	155,2	159,2	165,2	169,3	162,2

Рік	2007	2008	2009	2010	2011	2012	2013
Обсяг реалізованої продукції, млн грн	173,3	175,7	183,4	180,4	186,4	192,4	197,5

Завдання 1.2. Провести добір видів функцій, які найбільш точно описують тенденцію змінення кількості внутрішніх туристів у країни за 2000–2013 рр.

Таблиця 1.2

Кількість внутрішніх туристів

Рік	2000	2001	2002	2003	2004	2005	2006
Внутрішні туристи, тис. чол.	1351	1488	1545	1922	1012	932	1039

Рік	2007	2008	2009	2010	2011	2012	2013
Внутрішні туристи, тис. чол.	2155	1387	1094	649	716	807	1465

Завдання 1.3. Визначити параметри лінійного тренду для обсягів капітальних інвестицій в основний капітал у фактичних цінах за 2007–2013 рр.

Таблиця 1.3

Інвестиції в основний капітал

Рік	2007	2008	2009	2010	2011	2012	2013
Інвестиції в основний капітал, млн грн	188486	194000	151777	163000	209130	208000	215212

Завдання 1.4. Провести вирівнювання даних виробництва насіння соняшнику сільськогосподарськими підприємствами за 12 років параболою другого порядку.

Таблиця 1.5

Виробництво насіння соняшнику

Рік	2002	2003	2004	2005	2006	2007
Виробництво насіння соняшнику, тис. т	3271	4254	3930	4706	5324	6500

Рік	2008	2009	2010	2011	2012	2013
Виробництво насіння соняшнику, тис. т	6526	6780	7570	8000	9600	11330

2. СТАТИСТИЧНИЙ АНАЛІЗ ПЕРІОДИЧНИХ КОЛИВАНЬ

2.1. Типи коливань і їхні характеристики

Дослідження періодичних коливань в економіці сягають корінням в минуле. Ще К. Жюгляр (1819–1905) при вивченні економічних часових рядів з метою виділення бізнес-циклів установив циклічність інвестицій (період циклу 7–11 років). Пізніше коливання досліджували С. Кітчин, С. Коваль, Н. Кондратьєв, які виявили цикли у відновленні оборотних коштів (період 3–5 років), цикли в будівництві (період 15–20 років), великі «хвилі Кондратьєва».

Потреби економічної практики стали потужним стимулом до вдосконалення статистичної методології в області виявлення, вимірювання, моделювання й прогнозування сезонних коливань.

Останніми роками одержали розвиток інтерактивні методи фільтрації компонент часових рядів (наприклад, метод Четверікова, Ферстера, Шискіна–Ейзенпреса й ін.). У багатьох із цих методів не враховується незмінність сезонної хвилі, що робить процес дослідження періодичних коливань більш гнучким. Однак із застосуванням процедур ковзних середніх втрачалася частина інформації на кінцях часових рядів. Цього вдалося уникнути завдяки сучасним підходам до сезонної декомпозиції (до коригування). Сьогодні у світовій практиці в економометричних пакетах наведено саме ці процедури.

Для моделювання періодичних коливань в економіці також стали застосовувати гармонійний аналіз, запозичений з природничих наук. Перші економічні роботи в цій галузі пов'язані з іменами Г. Мура й Беверіджа.

Зупинимося більш докладно на методологічних питаннях розрахунку сезонної складової.

Коливання рівнів часових рядів є предметом їх статистичного дослідження, оскільки:

– дає змогу висунути гіпотези про причини коливань і шляхи впливу на них;

– на основі параметрів коливання можна прогнозувати або враховувати фактор помилки прогнозу, тобто робити прогноз більш надійним і точним;

– на основі параметрів і прогнозів коливань можна якісно планувати діяльність будь-якої організації.

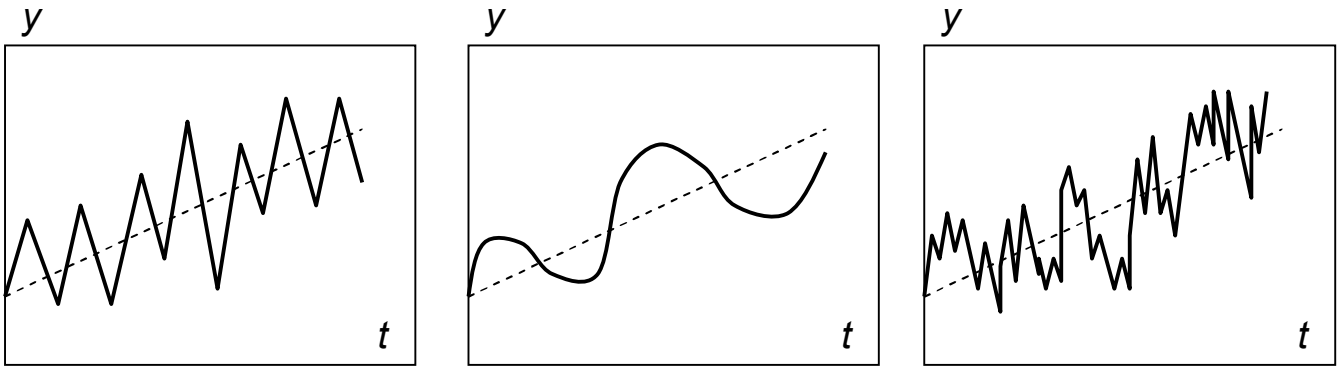
Коливання рівнів часового ряду можуть мати різну форму, різний розподіл у часі, різну амплітуду.

Усю різноманітність цих коливань можна подати як «суміш» у різних пропорціях трьох основних типів:

– пилкоподібні, або маятникові (рис. 2.1, *а*);

– довгоперіодичні (циклічні) (рис. 2.1, *б*);

– випадково розподілені в часі (рис. 2.1, *в*).



а б в

Рис. 2.1. Основні типи періодичних коливань часових рядів

Характерною рисою першого типу коливання є правильне, регулярне чергування відхилень від тренду вверх і вниз, тобто додатних і від'ємних за знаком через одне. Оскільки це схоже на коливання маятника годинника вправо-вліво, такі коливання називають також маятниковими. Назва ж «пилкоподібна» походить від виду графіка, схожого на зубці пилки.

Властивості пилкоподібного коливання такі: через часте змінення знака відхилення від тренду відбуваються акумуляції позитивних або негативних відхилень. Кількість додатних відхилень при досить великій довжині ряду дорівнює (точніше, наближається до рівності) кількості від'ємних коливань.

Розпізнати наявність пилкоподібних коливань як елемента в часовому ряді можна, по-перше, за видом графіка, по-друге, підрахунком кількості локальних екстремумів у ряді відхилень від тренду: чим ближче до кількості рівнів ряду, тим більше значення мають пилкоподібні коливання в їх загальному комплексі. Третій спосіб розпізнавання – за знаком і величиною коефіцієнта автокореляції відхилень від тренду зі зсувом (лагом) $t+1$ період.

Характерною рисою другого типу коливань (рис. 2.1, *б*) є наявність декількох (багатьох) підряд відхилень одного знака, що потім змінюється приблизно такою ж кількістю підряд відхилень протилежного знака. Потім весь цикл знову повторюється, причому зазвичай довжина всіх циклів є однаковою або хоча б приблизно однаковою. Якщо рівність окремих циклів істотно порушується, то кажуть про квазіциклічні коливання.

Розпізнати довгоперіодичні цикли коливань можна за видом графіка, підрахунком кількості екстремумів у ряді відхилень від тренду й за коефіцієнтом автокореляції відхилень 1-го порядку.

Характерною рисою третього типу коливань є хаотичність відхилень: після від'ємного відхилення від тренду може випливати знову від'ємне або навіть два-три від'ємних відхилень, а може й додатне. Іноді випадковий розподіл коливань називають «інтерференція коливань».

Для коливань третього виду характерними є такі властивості:

– знаки відхилень від тренду взаємопогашаються тільки на досить тривалому періоді через їх хаотичне чергування, а на коротких відрізках можуть акумулюватися;

– випадково розподілені в часі коливання неможливо прогнозувати.

Причиною випадкового розподілу коливань є наявність великого комплексу незалежних або слабозв'язаних факторів, що впливають на рівні досліджуваних явищ.

2.2. Статистичний аналіз сезонної нерівномірності на основі розрахунку індексів сезонності

Як вже було зазначено, у динамічних рядах часто спостерігаються сезонні коливання, під якими розуміють періодичне повторення із року в рік, підвищення й зниження рівнів в окремі місяці, квартали, тижні, дні. Сезонним коливанням піддаються внутрішньорічні рівні багатьох економічних показників. При графічному зображенні таких рядів сезонні коливання виявляються в підвищенні й зниженні рівнів у певні моменти.

При вивченні рядів динаміки, що містять «сезонну хвилю», її виділяють із загального коливання рівнів і вимірюють. Існує кілька методів для розв'язання цієї задачі. Усі вони базуються на порівнянні фактичних рівнів кожного моменту часового ряду із середнім рівнем, що дає змогу рівномірно розподілити річний показник по рівнях (або згладженими ковзними середніми, або вирівняними за трендом). При цьому для виміру «сезонної хвилі» розраховують або абсолютні різниці (відхилення) фактичних рівнів від середнього рівня (або від вирівняних), або відношення рівнів до середнього рівня за повний період (рік), так звані індекси сезонності:

$$I_s = \frac{y_t}{\bar{y}} 100\ \%. \qquad (2.1)$$

За наявності даних за кілька років розрахувати індекси сезонності можна по-різному.

За даними за кілька років розраховують середнє значення рівня для кожного моменту $\hat{y}_t$, а також середній рівень за весь період $\bar{y}$. Потім визначають індекси сезонності як процентне відношення середніх рівнів для кожного рівня до загального рівня ряду (за всі роки)

$$I_{si} = \frac{\hat{y}_t}{\bar{y}} 100\ \%. \qquad (2.2)$$

Цей метод використовують в основному в тих випадках, коли рівні однакових моментів у різні роки різняться незначно.

Якщо ж спостерігається тенденція до підвищення або зниження рівнів з року в рік, то ефективніше розраховувати індекси сезонності за такою схемою: для кожного року окремо розраховують індекси сезонності за

формулою (2.1), а потім з індексів однакових моментів знаходять середнє арифметичне.

Одним із ефективних інструментів візуалізації й розпізнавання образів сезонних компонентів є будування радіальних (павутинчастих) діаграм у відносних або абсолютних величинах (рис. 2.2).

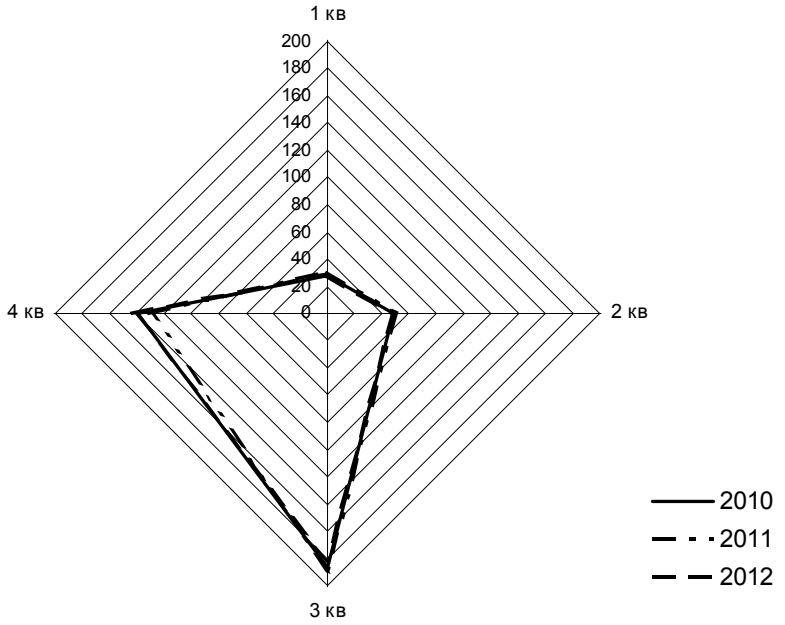


Рис. 2.2. Радіальна діаграма індексів сезонності виробництва валового продукту сільського господарства в Україні

2.3. Дослідження періодичних коливань методами спектрального аналізу

При вивченні структури часових рядів і аналізі періодичних складових важливим є спектральний аналіз, що базується на фундаментальній математичній теорії, основою якої є узагальнення розвинення за методом Фур‘є. Подальшого розвитку цей підхід набув у роботах А. Колмогорова, Н. Вінера, Дж. Т’юки, А. Хінчина та ін..

Можна виокремити три основні напрямки практичного використання спектрального аналізу:

– одержання корисних описових статистик;

– важливий засіб діагностики часового ряду;

– перевірка постульованих теоретичних моделей.

Однак особливо ефективним є спектральний метод при узагальненні підходу для двох або більше рядів. Метод дає змогу оцінити характеристики, що відображають тісноту зв'язку між періодичними компонентами рядів, визначити взаємні змінення (випередження й запізнювання) для цих компонентів, а також простежити в часі за зміненням цих характеристик (якщо такі змінення є значними).

Порівняно з традиційними методами вивчення сезонності спектральний аналіз має великі переваги. Спектральний аналіз дає змогу

одночасно визначити період (частоту) різних періодичних коливань та їх інтенсивність (амплітуду). Велика кількість традиційних методів базується на припущенні, що характерні параметри коливання (наприклад, період коливання) вже є відомими.

Крім того, при застосуванні традиційних методів, на відміну від спектрального аналізу, зазвичай припускається незмінність сезонної хвилі.

Вихідним для перетворення Фур'є найкраще взяти не первинний ряд за кілька років, а усереднений ряд місячних рівнів, у якому елімінована (виключено) тренд і (або) в основному загашено випадкові коливання.

Вирівнювання здійснюють з допомогою ряду Фур'є, у якому рівні можна виразити як функцію часу.

Моделювання рекомендується проводити в трьох випадках, коли в емпіричному ряді спостерігається періодичність змінення рівнів. У цьому випадку періодичність коливань рівнів динамічного ряду можна подати у вигляді синусоїдальних коливань. Оскільки ці коливання являють собою гармонійні коливання, синусоїди, отримані при вирівнюванні за рядом Фур'є, називають гармоніками різних порядків. Показник k у рівнянні визначає кількість гармонік. Зазвичай при використанні методу Фур'є розраховують кілька гармонік (частіше не більше чотирьох) і потім вже визначають, з якою кількістю гармонік ряд Фур'є якнайкраще відображає змінення рівнів ряду.

При вирівнюванні за рядом Фур'є періодичні коливання рівнів динамічного ряду наведено у вигляді суми декількох синусоїд (гармонік), накладених одна на одну.

Так, наприклад, при $k = 1$ ряд Фур'є буде мати вигляд

$$\tilde{y}_t = a_0 + a_1 \cos t + b_1 \sin t, \qquad (2.4)$$

а при $k = 2$ відповідно

$$\tilde{y}_t = a_0 + a_1 \cos t + b_1 \sin t + a_2 \cos 2t + b_2 \sin 2t \qquad (2.5)$$

і т. д.

Параметри рівняння теоретичних рівнів, що визначається рядом Фур'є, знаходять, як і в інших випадках, методом найменших квадратів, причому без відображення формули, яку використовують для розрахунку параметрів ряду Фур'є:

$$a_0 = \frac{\sum_{t=1}^{n} y_t}{n}, \qquad (2.6)$$

$$a_k = \frac{2\sum_{t=1}^{n} y_t \cos kt}{n}, \qquad (2.7)$$

$$b_k = \frac{2\sum_{t=1}^{n} y_t \sin kt}{n}. \qquad (2.8)$$

Як бачимо, параметри рівняння залежать від значень y_t і пов'язаних з ними послідовних значень $\cos kt$ і $\sin kt$.

Останнє значення t відрізняється від нуля на $2\pi/n$, де n – кількість рівнів емпіричного ряду. Наприклад, для вивчення сезонних коливань протягом року необхідно взяти $n=12$ (за кількістю місяців). Тоді, подаючи періоди як частоти довжини кола, ряд динаміки можна записати в такому вигляді:

$$0; \frac{\pi}{6}; \frac{\pi}{3}; \frac{\pi}{2}; \frac{2\pi}{3}; \frac{5\pi}{6}; \pi; \frac{7\pi}{6}; \frac{4\pi}{3}; \frac{3\pi}{2}; \frac{5\pi}{3}; \frac{11\pi}{6}.$$

При обчисленнях слід зважувати на те, що в чотирьох квадратах від нуля до другого косинуси й синуси чотири рази набувають тих самих абсолютних значень, а саме: 0; 0,5; 0,866; 1, узятих зі знаками «плюс» і «мінус». Для обчислення синусів і косинусів гармонік найкраще користуватися табл. 2.1.

Таблиця 2.1

Значення синусів і косинусів гармонік при $n=12$

t	$\cos t$	$\cos 2t$	$\cos 3t$	$\cos 4t$	$\sin t$	$\sin 2t$	$\sin 3t$	$\sin 4t$
0	1	1	1	1	0	0	0	0
$\pi/6$	0,866	0,5	0	-0,5	0,5	0,866	1	0,866
$\pi/3$	0,5	-0,5	-1	-0,5	0,866	0,866	0	-0,866
$\pi/2$	0	-1	0	1	1	0	-1	0
$2\pi/3$	-0,5	-0,5	1	-0,5	0,866	-0,866	0	0,866
$5\pi/6$	-0,866	0,5	0	-0,5	0,5	0,5	-0,866	-0,866
π	-1	1	-1	1	0	0	0	0
$7\pi/6$	-0,866	0,5	0	-0,5	-0,5	0,866	-1	0,866
$4\pi/3$	-0,5	-0,5	1	-0,5	-0,866	0,866	0	-0,866
$3\pi/2$	0	-1	0	1	-1	0	1	0
$5\pi/3$	0,5	-0,5	-1	-0,5	-0,866	0,866	0	0,866
$11\pi/6$	0,866	0,5	0	-0,5	-0,5	-0,866	-1	-0,866

Оскільки t в річній динаміці означає номер конкретного місяця, відповідно $t=0$ відповідає січню, $t=\pi/6$ відповідає лютому й т. д.

Після одержання рівняння моделі й розрахунку теоретичних значень ряду оцінюють якість отриманої моделі. Стандартним методом оцінювання якості моделі є розрахунок середньоквадратичної помилки (1.18).

Запитання для самоконтролю

1. На які основні типи поділяють форми коливань?
2. Для чого застосовують радіальні діаграми?

3. Які існують методи розрахунку індексів сезонності, чим вони різняться?

4. У яких основних напрямках економіко-статистичних досліджень використовують спектральний аналіз?

5. На якій фундаментальній математичній теорії базується метод спектрального аналізу?

6. Які переваги має спектральний аналіз на відміну від традиційних методів визначення циклічної компоненти?

7. Скільки гармонік розраховують при застосуванні рядів Фур'є на практиці?

8. За яким показником оцінюють якість побудованої моделі на основі рядів Фур'є?

9. Який тип коливань називають інтерференцією коливань?

10. Які коливання називають квазіциклічними?

Завдання для самостійної роботи

Завдання 2.1. Провести вирівнювання даних з продажу зимового одягу за методом Фур'є. Вихідні дані наведено в табл. 2.2.

Таблиця 2.2

Динаміка продажу одягу

Місяць	1	2	3	4	5	6	7	8	9	10	11	12
Обсяг продажу зимового одягу, тис. грн	37	40	44	52	46	70	60	48	46	38	36	35

Завдання 2.2. Використовуючи дані, отримані у попередній задачі, розрахувати глибину сезонних коливань продажів одягу, визначити змінення цих коливань за кварталами.

3. МЕТОДИ І МОДЕЛІ КОРЕЛЯЦІЙНО-РЕГРЕСІЙНОГО АНАЛІЗУ

3.1. Види зв'язку між змінними, класифікація функцій регресії

Між більшістю явищ і процесів в економіці є постійний взаємний і об'єктивний всеохоплюючий зв'язок. Дослідження взаємозв'язків між об'єктивно існуючими явищами й процесами має велике значення для економіки, оскільки дає можливість глибше зрозуміти складний механізм причинно-наслідкових відносин між явищами. Для дослідження

інтенсивності, виду й форми залежностей широко застосовується кореляційно-регресійний аналіз, що є методичним інструментом при вирішенні завдань прогнозування, планування й аналізу господарської діяльності підприємств, а також маркетингу.

Розрізняють функціональну і стохастичну залежності між економічними явищами й процесами.

У випадку функціональної залежності є однозначне відображення множини A на множину B. Множину A називають областю визначення функцій, а множину B – множиною значень функції.

Функціональна залежність трапляється нечасто. У більшості випадків функція Y або аргумент X – випадкові величини. X і Y піддано дії різних випадкових факторів, серед яких можуть бути фактори, що є загальними для двох випадкових величин.

Стохастичними називають процеси, на проходження яких значно впливають фактори випадкового характеру і обставини, що важко передбачити.

Стохастичну залежність виражають з допомогою функцій, які називають регресією.

Існують кілька видів регресії.

1. Регресія відносно кількості змінних.

Проста регресія – регресія між двома змінними. Множинна регресія – це регресія між залежною змінною Y і декількома пояснювальними змінними $x_1, x_2, ..., x_n$. Множинна лінійна регресія має такий вигляд:

$$Y = a_0 + a_1 x_1 + a_2 x_2 + ... + a_n x_n, \qquad (3.1)$$

де Y – функція регресії;

$x_1, x_2, ..., x_n$ – незалежні змінні;

$a_1, a_2, ..., a_n$ – коефіцієнти регресії;

a_0 – вільний член рівняння;

n – кількість факторів моделі.

2. Регресія відносно форми залежності:

– лінійна регресія, яку виражено лінійною функцією;

– нелінійна регресія, яку виражено нелінійною функцією.

3. Залежно від характеру регресії розрізняють додатну й від'ємну регресії.

4. За типом об'єднання явищ розрізняють:

– безпосередню регресію, коли залежна й пояснювальна змінні пов'язані безпосередньо одна з одною;

– непряму регресію, коли пояснювальна змінна діє на залежну через кілька інших змінних;

– помилкову регресію, яка виникає при формальному підході до досліджуваних явищ без з'ясування того, чим спричинено цей зв'язок.

3.2. Методика кореляційного аналізу

Будь-який причинний вплив може виражатися або функціональним, або кореляційним зв'язком, але не кожна функція або кореляція відповідає причинній залежності між явищами. Тому необхідним є обов'язкове дослідження причинно-наслідкових зв'язків.

Дослідження кореляційних зв'язків називають кореляційним аналізом, а дослідження однобічних стохастичних залежностей – регресійним аналізом.

Задачами кореляційного аналізу є такі:

– вимір ступеня щільності (зв'язаності, сили), форми й напрямку взаємозв'язку двох явищ;

– добір факторів, що найбільш істотно впливають на результативну ознаку, на основі виміру щільності зв'язку між явищами;

– виявлення невідомих причинних зв'язків.

Кореляція безпосередньо є засобом виявлення причинних зв'язків між явищами, але з її допомогою встановлюють ступінь необхідності цих зв'язків і достовірність суджень про їх наявність. Причинний характер зв'язків з'ясовується з допомогою логіко-професійних суджень, що розкривають механізм зв'язків.

Отже, основною задачею кореляційного аналізу є визначення кореляційної залежності між ознаками.

Кореляційна залежність – це залежність випадкових величин (ознак), при якій змінення середнього значення однієї випадкової величини відповідає зміненню середнього значення іншої.

Для виявлення наявності й характеру такого зв'язку в статистиці використовують кілька методів: застосування паралельних даних, графічний метод, метод аналітичних груп і кореляційних таблиць, розрахунок коефіцієнтів кореляції.

Одним із простих методів визначення системного зв'язку між змінними є розрахунок коефіцієнта Фехнера (коефіцієнта кореляції знаків), що базується на порівнянні поводження відхилень індивідуальних значень кожної ознаки (x і y) від своєї середньої величини. При цьому до уваги беруть не величини відхилень $(x_i - \bar{x})$ і $(y_i - \bar{y})$, а їхні знаки («+» і «–»). Визначивши знаки відхилення від середньої величини в кожному ряді, підраховують кількість збігів і розбіжностей усіх пар знаків. Якщо збіг знаків позначити символом C, а розбіжностей – H, то коефіцієнт Фехнера можна записати як відношення різниці кількості пар збігів і розбіжностей знаків до їх суми, тобто до загальної кількості спостережуваних одиниць:

$$K_{ф} = \frac{\sum_{i=1}^{n} C - \sum_{j=1}^{m} H}{\sum_{i=1}^{n} C + \sum_{j=1}^{m} H}. \qquad (3.2)$$

Очевидно, якщо знаки всіх відхилень за кожною ознакою збігаються, то $\sum_{j=1}^{m} H = 0$ і $K_ф = 1$. Це характеризує наявність прямого зв'язку. Якщо всі знаки збігаються, то $\sum_{i=1}^{n} C = 0$ і $K_ф = -1$ (зворотний зв'язок). Коефіцієнт Фехнера може набувати значень від -1 до 0 і від 0 до +1. При цьому чим ближче до одиниці, тим більше (сильніше) залежність між x і y.

Коефіцієнт Фехнера характеризує не тільки щільність зв'язку, але і його наявність і напрямок, оскільки залежить тільки від знаків і при цьому не враховується величина самих відхилень x і y від їхніх середніх величин.

Зв'язок між кількісними ознаками вимірюють через їх варіацію. Виміряти залежність (зв'язок) між двома величинами, що корелюються, означає визначити, наскільки варіація результативної ознаки обумовлена варіацією факторної ознаки.

За показники щільності зв'язку між кількісними ознаками крім коефіцієнта Фехнера найбільш часто використовують лінійний коефіцієнт кореляції, який розраховують лише у випадку лінійної залежності між ознаками. Якщо форму зв'язку між x і y не визначено, його розраховують з метою визначення, чи є залежність лінійною.

Як і коефіцієнт Фехнера, лінійний коефіцієнт кореляції будується на основі відхилень індивідуальних значень x і y від відповідної середньої величини. Однак на відміну від $K_ф$ у лінійному коефіцієнті кореляції враховують не тільки знаки, але й значення відхилень $(x - \bar{x})$ і $(y - \bar{y})$, виражені для порівняння в одиницях середнього квадратичного відхилення кожної ознаки, де $\bar{x}$ і $\bar{y}$ – середні величини.

Лінійний коефіцієнт кореляції має такий вигляд:

$$r_{xy} = \frac{\sum_{i=1}^{n}\left((x_i - \bar{x})(y_i - \bar{y})\right)}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}. \qquad (3.3)$$

Значення, яких набуває r_{xy}, знаходяться в межах від –1 до +1. При додатному значенні r_{xy} має місце додатна кореляція, тобто зі збільшенням (зменшенням) змінної x змінна y відповідно збільшується (зменшується). При від'ємному значенні r_{xy} має місце від'ємна кореляція, тобто зі збільшенням (зменшенням) значень x значення y відповідно зменшуються (збільшуються).

Іноді показники щільності зв'язку можна якісно оцінити за шкалою Чеддока (табл. 3.1).

Таблиця 3.1

Шкала Чеддока

Кількісна міра щільності зв'язку $\left\|r_{xy}\right\|$	Якісна характеристика сили зв'язку
0,1 … 0,3	Слабка
0,3 … 0,5	Помірна
0,5 … 0,7	Помітна
0,7 … 0,9	Висока
0,9 … 0,99	Досить висока

Ще одним ефективним інструментом дослідження щільності зв'язків між факторами є будування корелограм.

Корелограма відображає кореляційне поле, утворене факторами x і y. У корелограмі коефіцієнт кореляції відображає зашумленість лінійної залежності, але за ним неможливо визначити нахил (рис. 3.1).

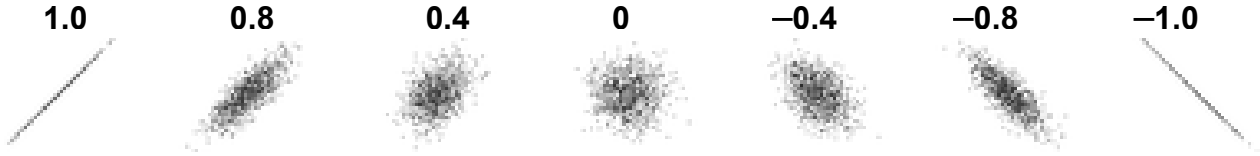


Рис. 3.1. Корелограми й оцінювання зв'язку між факторами

Інтерпретуючи значення коефіцієнта кореляції, слід мати на увазі, що його розраховано за x і y для обмеженої кількості спостережень і він змінюється випадково, як і самі значення x і y.

Іншими словами, як будь-який вибірковий показник, коефіцієнт кореляції містить випадкову похибку й не завжди однозначно відображає дійсно реальний зв'язок між досліджуваними показниками.

Для того щоб оцінити істотність (значущість) самого r_{xy} і відповідно реальність вимірюваного зв'язку між x і y, необхідно розрахувати середньоквадратичну похибку коефіцієнта кореляції σ_r.

Оцінювання істотності (значущості) лінійного коефіцієнта кореляції базується на порівнянності значення r_{xy} з його середньоквадратичною похибкою:

$$\left|r_{xy}\right| \Big/ \sigma_r \ . \tag{3.4}$$

Середньоквадратичну похибку можна розрахувати за такою формулою (при $n > 50$):

$$\sigma_r = 1 - r_{xy}^2 \Big/ \sqrt{n} \ . \tag{3.5}$$

В інших випадках розраховують довірчий інтервал і перевіряють значущість на основі t-критерію Стьюдента.

Для аналізу зв'язків між ознаками, які виміряно за порядковими шкалами, застосовують так звані рангові коефіцієнти кореляції. Основою

рангових коефіцієнтів є ранг показника – номер спостереження в упорядкованій сукупності. Для цього оцінювання використовують коефіцієнти кореляції Кендела й Спірмена.

3.3. Методика регресійного аналізу

Методика проведення регресійного аналізу складається з кількох етапів:

1. Встановлення форми залежності (лінійна або нелінійна; додатна або від'ємна і т. д.).

2. Визначення функцій регресії і встановлення впливу факторів на незалежну змінну. Важливо не тільки визначити форму регресії, навести загальну тенденцію змінення залежної змінної, але й з'ясувати, яка була б дія на залежну змінну головних факторів, якби інші не змінювалися і було виключено випадкові елементи. Для цього визначають функцію регресії у вигляді математичного рівняння.

3. Оцінювання незалежних значень змінної, тобто вирішення завдань екстраполяції й інтерполяції. Під час екстраполяції поширюються тенденції, визначені в минулому, на майбутній період. Екстраполяція широко використовується в прогнозуванні. Під час інтерполяції знаходять невідомі значення, що відповідають моментам часу між відомими моментами, тобто обчислюють значення залежної змінної всередині інтервалу заданих значень факторів.

Розроблення кореляційно-регресійної моделі й дослідження економічних процесів мають виконуватися за етапами показаними на рис. 3.2.

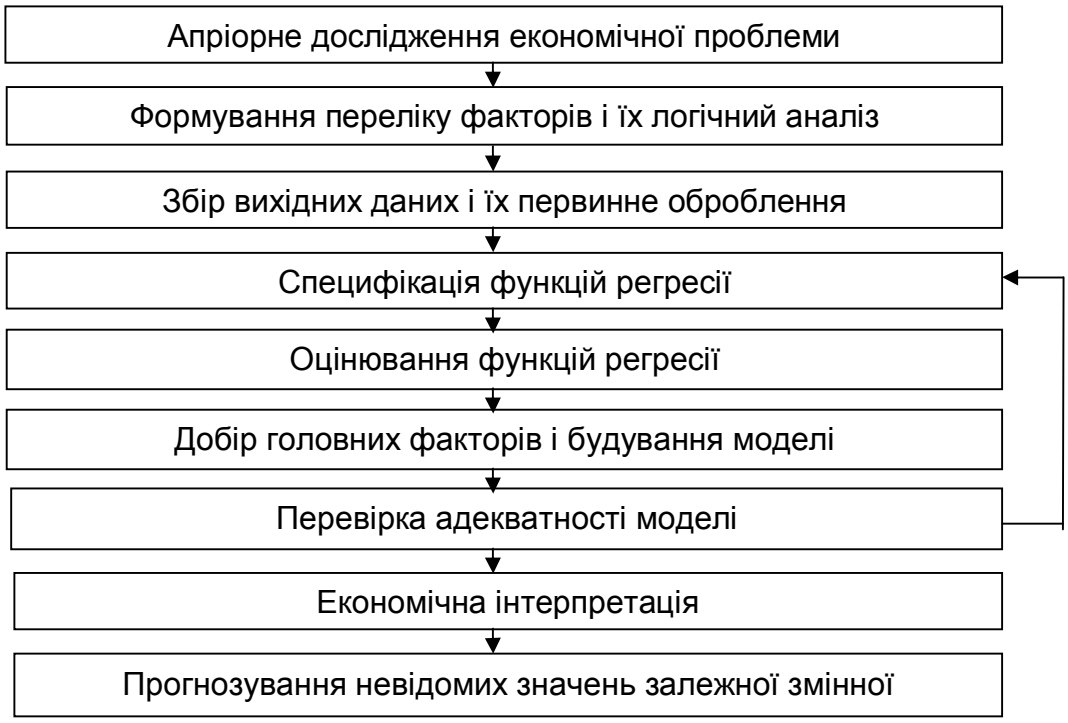


Рис. 3.2. Етапи будування регресійної моделі

Розглянемо докладніше зміст етапів.

Апріорне дослідження економічної проблеми. Відповідно до мети роботи на основі знань макро- і мікроекономіки конкретизуються явища, процеси, а залежність між ними підлягає оцінюванню. При цьому мається на увазі насамперед чітке визначення економічних явищ, установлення об'єкта й періодів дослідження.

На цьому етапі дослідження має бути сформульовано економічно осмислені й прийнятні гіпотези про залежність економічних явищ.

Формування переліку факторів і їх логічний аналіз. Для визначення найбільш адекватної кількості змінних у регресійній моделі насамперед орієнтуються на логічний аналіз економічного явища. Виходячи з фізичного змісту явища, роблять класифікацію змінних на залежну й пояснювальну.

Збір вихідних даних і їх первинне оброблення. При будуванні моделі вихідну інформацію збирають у трьох видах:

– динамічні (часові) ряди;

– просторова інформація;

– інформація про роботу декількох об'єктів в одному розмірі часу.

Обсяг вибірки залежить від кількості факторів моделі із урахуванням вільного члена. Для одержання статистично значущої моделі потрібен мінімальний обсяг вибірки, що визначається так:

$$n_{min} = 5(m + n), \tag{3.6}$$

де m – кількість факторів моделі;

n – кількість вільних членів у рівнянні.

Специфікація функції регресії. На цьому етапі дослідження дається конкретне формулювання гіпотези про форму зв'язку (лінійна або нелінійна, проста або множинна й т. д.). При цьому використовуються різні критерії перевірки спроможності гіпотетичного виду залежності, та перевіряються передумови кореляційно-регресійного аналізу.

Оцінювання функцій регресії. Тут визначаються числові значення параметрів регресії й обчислення кількох показників, що характеризують точність регресійного аналізу.

Добір головних факторів. Вибір факторів – основа для будування багатофакторної кореляційно-регресійної моделі. На цьому етапі формування переліку факторів і їх логічного аналізу збирають всі можливі фактори й факти. Після цього необхідно вибирати більш раціональний перелік факторів. При цьому проводять аналіз факторів на мультиколінеарність – попарну кореляційну залежність між факторами.

Мультиколінеарна залежність є, якщо коефіцієнт парної кореляції $r_{ij} \to 0{,}70 + 0{,}80$.

Негативний вплив мультиколінеарності полягає в такому:

– ускладнюється процедура вибору головних факторів;

– спотворюється суть коефіцієнта множинної кореляції (припускається незалежність факторів);

– ускладнюються обчислення при будуванні самої моделі;

– зменшується точність оцінювання параметрів регресії, спотворюється оцінка дисперсії.

Наслідком зменшення точності є надійність коефіцієнтів регресії і частково неприйнятність їх використання для інтерпретації як заходу впливу відповідної пояснювальної змінної на залежну змінну.

Оцінки коефіцієнта стають дуже чутливими до вибіркових спостережень. Невелике збільшення обсягу вибірки може призвести до дуже сильних змінень оцінок. Крім того, стандартні похибки оцінок входять до формули критерію значущості, тому застосування самих критеріїв стає також ненадійним. З наведеного ясно, що дослідник має встановити стохастичну мультиколінеарність і за можливості усунути її.

Для вимірювання мультиколінеарності можна використовувати коефіцієнт множинної детермінації

$$D = R^2, \tag{3.7}$$

де R – коефіцієнт множинної кореляції, що розраховується за формулою

$$R = \sqrt{1 - \frac{\sigma^2_{ост}}{\sigma^2_{общ}}}, \tag{3.8}$$

де $\sigma^2_{ост}$ – залишкова дисперсія залежної змінної;

$\sigma^2_{общ}$ – загальна дисперсія залежної змінної.

Якщо фактори не є мультиколінеарними, то

$$D = \sum_{j=1}^{m} d_{yj}. \tag{3.9}$$

Тут d_{yj} – коефіцієнт парної детермінації, $d_{yj} = r^2_{yi}$, де r_{yi} – коефіцієнт парної кореляції між j-м фактором і залежною змінною y.

За наявності мультиколінеарності рівність (3.9) не виконується. Тому як міра мультиколінеарності використовується така різниця:

$$M = D - \sum_{j=1}^{m} d_{yj}. \tag{3.10}$$

Чим меншою є ця різниця, тим меншою буде мультиколінеарність. Для усунення мультиколінеарності використовують метод виключення змінних.

Цей метод полягає в тому, що пояснювальні змінні (фактори), які високо корелюються, усувають з регресії, і її заново оцінюють. Добір змінних, які підлягають виключенню, проводять з допомогою коефіцієнтів парної кореляції. З досвіду випливає, що якщо $|r_{yj}| \geq 0{,}70$, то одну зі змінних можна виключити з аналізу, але яку саме, вирішують, виходячи з керованості факторів на рівні підприємства.

Зазвичай в моделі залишають той фактор, що забезпечує її поліпшення в планованому році. Можлива ситуація, коли обидва мультиколінеарних фактори керуються на рівні підприємства. Вирішити питання про виключення того чи іншого фактора можна тільки відповідно до процедури добору головних факторів.

Добір факторів не є самостійним процесом, його доповнює будування моделі. Прийняття рішення про виключення факторів проводять на основі аналізу значень спеціальних статистичних характеристик і з урахуванням керованості факторів на рівні підприємства.

Процедура добору головних факторів. Ця процедура обов'язково містить такі етапи.

1. Аналіз факторів на мультиколінеарність і її виключення. Тут проводять аналіз значень коефіцієнтів парної кореляції r_{ij} між факторами x_i і x_j.

2. Аналіз щільності взаємозв'язку факторів x із залежною змінною y.

Для аналізу щільності взаємозв'язку x і y використовують коефіцієнт парної кореляції $r_{x_i y}$, який визначають і подають у вигляді кореляційної матриці (табл. 3.2).

Таблиця 3.2

Кореляційна матриця

Номер змінної	x_1	x_2	x_3	…	x_m	y
x_1	1	$r_{x_1x_2}$	$r_{x_1x_3}$	…	$r_{x_1x_m}$	r_{x_1y}
x_2	$r_{x_1x_1}$	1	$r_{x_2x_3}$	…	$r_{x_2x_m}$	r_{x_4y}
x_3	$r_{x_3x_1}$	$r_{x_3x_2}$	1	…	r_{x3x_m}	r_{x_3y}
…	…	…	…	…	…	…
x_m	$r_{x_mx_1}$	$r_{x_mx_2}$	$r_{x_mx_3}$	…	1	r_{x_my}
y	r_{yx_1}	r_{yx_2}	r_{yx_3}	…	r_{yx_m}	1

Фактори, для яких $r_{x_i y} \to 0$, тобто не пов'язані з y, підлягають виключенню в першу чергу. Фактори, для яких $r_{x_i y}$ має найменше

значення, можна виключити з моделі. Питання про їх остаточне виключення вирішують під час аналізу інших статистичних показників.

3. При аналізі коефіцієнтів β_k ураховують вплив факторів на змінну y і розходження у рівні їх коливань. Коефіцієнт β_k є показником того, як змінюється функція зі зміненням середнього квадратичного відхилення на одну одиницю при фіксованому значенні інших аргументів:

$$\beta_k = a_k \frac{\sigma_{x_k}}{\sigma_y}, \tag{3.11}$$

де σ_{x_k} – середнє квадратичне відхилення k-го фактора;

σ_y – середнє квадратичне відхилення функції;

a_k – коефіцієнт регресії при k-му факторі.

Із двох факторів x_i і x_j можна виключити той фактор, для якого β_k має менше значення.

Припустимо, виключенню підлягає один із мультиколінеарних факторів x_i або x_j. Обидва фактори керуються на рівні підприємства, коефіцієнти регресії a_i і a_j є статистично значущими. Фактор x_i більш тісно пов'язаний з y, тобто $r_{x_i y} > r_{x_j y}$, але при цьому $\beta_{x_i} > \beta_{x_j}$. У цьому випадку зазвичай виключенню підлягає фактор x_j.

4. Перевірка коефіцієнтів регресії на статистичну значущість.

Перший спосіб. Перевірку статистичної значущості a_k за критерієм Стьюдента проводять за формулою

$$t_k = \frac{a_k}{S_{a_k}}, \tag{3.12}$$

де a_k – коефіцієнт регресії при k-му факторі;

S_{a_k} – стандартне відхилення оцінки параметра a_k.

Кількість ступенів свободи f статистики t_k дорівнює $n-m-1$, де m – кількість факторів моделі.

Значення t_k порівнюють із критичним значенням $t_{f,a}$, при заданих рівні значущості α і кількості ступенів свободи f (двостороння критична область).

Якщо $t_k \geq t_{f,a}$, то a_k істотно більше 0, а фактор x_k впливає на y. При цьому фактор x_k залишаємо в моделі. Якщо $t_k < t_{f,a}$, то фактор виключаємо з моделі.

Другий спосіб. Перевірка статистичної значущості a_k за критерієм Фішера:

$$F_k = \left(\frac{a_k}{S_{a_k}} \right)^2 = t^2, \qquad (3.13)$$

де t^2 – багатовимірний аналог критерію Стьюдента.

Кількість ступенів свободи статистики F_k є такою: $f_1 = 1,\ f_2 = n - m - 1$. Значення F_k порівнюють із критичним значенням $F_{f_1 f_2}$, при заданих рівні значущості α і кількості ступенів свободи f_1, f_2.

Якщо $F_k \geq F_{f_1 f_2 \alpha}$, то α_k істотно більше нуля, а фактор x_k впливає на y. При цьому фактор x_k залишаємо в моделі. Якщо $F_k < F_{f_1 f_2 \alpha}$, то фактор виключаємо з моделі.

5. Аналіз факторів на керованість. Під час логічного аналізу на основі економічних значень дослідник має зробити висновок щодо можливості розроблення організаційно-технічних заходів, спрямованих на поліпшення (змінення) вибраних факторів на рівні заходів. Якщо це можливо, то ці фактори є керованими. Некеровані фактори на рівні підприємства можна виключити з моделі.

6. Будування нової регресійної моделі без виключених факторів. Для цієї моделі визначають коефіцієнт множинної детермінації D.

7. Дослідження доцільності виключення факторів з моделі з допомогою коефіцієнта детермінації.

Перш ніж прийняти рішення про виключення змінних з аналізу через їх незначущий вплив на залежну змінну, проводять дослідження з допомогою коефіцієнта детермінації.

У першій регресії міститься m пояснювальних змінних, у другій – тільки частина з них – m_1. При цьому $m = m_1 + m_2$, тобто до другої регресії не включено m_2 пояснювальних змінних. Тепер слід перевірити, чи пояснюють змінні m_2 варіацію змінної y. Для цього використовують статистику, яка має F-розподіл з $f_1 = m - m_1 = m_2$ і $f_2 = n - m - 1$ ступенями свободи:

$$F = \frac{(D_m - D_{m_1})(n - m - 1)}{(m - m_1)(1 - D_m)}, \qquad (3.14)$$

де D_m – коефіцієнт детермінації регресії з m пояснювальними змінними;

D_{m_1} – коефіцієнт детермінації регресії з факторами m_1.

Різниця $(D_m - D_{m_1})$ у чисельнику формули є мірою додаткового пояснення варіації змінної y шляхом включення m_2 змінних.

Критичне значення $F_{f_1 f_2}$ знаходять за таблицею F-розподілу за заданим рівнем значущості α і ступенями свободи f_1 і f_2. Якщо $F_k \geq F_{f_1 f_{2\alpha}}$, то включення додатково пояснювальних змінних не впливає значно на змінну y. Якщо $F_k < F_{f_1 f_{2\alpha}}$, то пояснювальні змінні m_2 впливають на варіацію змінної y і, отже, у цьому випадку всі змінні m_2 не можна виключати з моделі.

При реалізації першої ситуації ($F_k \geq F_{f_1 f_{2\alpha}}$) фактори остаточно виключаються з моделі.

Перевірка адекватності моделі

Цей етап аналізу містить кілька процедур.

Оцінювання значущості коефіцієнта детермінації. Це оцінювання необхідне для вирішення питання, чи впливають вибрані фактори на залежну змінну. Оцінювання значущості D слід проводити тому, що може скластися така ситуація, коли величина коефіцієнта детермінації буде цілком обумовлена випадковими коливаннями у вибірці, на основі якої його обчислюють. Це пояснюється тим, що величина D істотно залежить від обсягу вибірки.

Для оцінювання значущості коефіцієнта множинної детермінації використовують таку статистику:

$$F = \frac{D(n-m-1)}{m(1-D)}, \tag{3.15}$$

яка має F-розподіл з $f_1 = m$ і $f_2 = n-m-1$ ступенями свободи. Тут $D = R^2$; m – кількість пояснювальних змінних (факторів).

Значення статистики F, обчислене за емпіричними даними, порівнюють з табличними значеннями $F_{f_1 f_{2\alpha}}$. Критичне значення визначають за заданим α і ступенями свободи f_1 і f_2. Якщо $F_k > F_{f_1 f_{2\alpha}}$, то обчислений коефіцієнт детермінації значно відрізняється від нуля і, отже, включені до регресії змінні добре пояснюють залежну змінну, що дає змогу казати про значущість самої регресії (моделі).

Перевірка якості добору теоретичного рівняння. Цей етап проводять з використанням середньої похибки апроксимації регресії, яку визначають за формулою

$$E = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - y_{iT}}{y_{iT}}\right| 100\%. \tag{3.16}$$

Економічна інтерпретація. Результати регресійного аналізу порівнюють з гіпотезами, сформованими на першому етапі дослідження, і оцінюють їх істинність в економічних умовах.

Прогнозування невідомих значень залежної змінної. Отримане рівняння регресії застосовують при прогностичному аналізі. Прогноз одержують шляхом підставлення в моделі регресії кількісно оцінених параметрів значень факторів. Слід підкреслити, що прогнозування результатів за регресією краще піддається змістовній інтерпретації, ніж проста екстраполяція тенденцій, тому що повніше враховується природа досліджуваного явища.

Запитання для самоконтролю

1. Які існують типи зв’язків між економічними явищами і процесами?
2. Що називають функцією регресії?
3. За якими ознаками класифікують функції регресії?
4. Що називають кореляційною залежністю?
5. Сформулюйте основну задачу кореляційного аналізу.
6. Які недоліки має коефіцієнт Фехнера?
7. Що характеризує лінійний коефіцієнт кореляції?
8. Що вимірюють за шкалою Чеддока?
9. Чим відрізняється зв'язок між факторами, якщо коефіцієнт кореляції становить +1 і -1?
10. Для чого використовують корелограму?
11. Назвіть основні завдання регресійного аналізу.
12. З яких етапів складається алгоритм будування регресійної моделі?
13. Що таке коефіцієнт детермінації?
14. Що таке мультиколінеарність?
15. Як в регресійних моделях визначають мультиколінеарність? Як її позбутися?
16. Для чого будують кореляційну матрицю?
17. Якими методами перевіряють адекватність регресійної моделі?

Завдання для самостійної роботи

Завдання 3.1. За даними, які наведено в табл. 3.3, розрахувати параметри рівняння регресії. Зробити висновки.

Таблиця 3.3

Вихідні дані для проведення аналізу

Готова продукція, тис. грн	954	10200	814	426	101	4439
Чистий дохід від реалізації продукції, тис. грн	9548	33176	41700	8990	2504	20200

Закінчення табл. 3.3.

Готова продукція, тис. грн	14200	83	12100	24200	2998	2007
Чистий дохід від реалізації продукції, тис. грн	40787	4745	27982	45868	15962	3695

Завдання 3.2. Використовуючи кореляційно-регресійний аналіз, провести статистичне оброблення даних машинобудівних підприємств з табл. 3.4.

Таблиця 3.4

Вихідні дані для проведення аналізу

Номер підприємства	Дохід від реалізації продукції y, тис. грн	Основні засоби (залишкова вартість) x_1, тис. грн	Матеріальні витрати x_2, тис. грн
1	11436	25336	7722
2	39301	10560	12617
3	102262	11862	32035
4	50856	27296	31654
5	1189	1663	878
6	10459	5298	5894
7	2703	4772	1450
8	139524	85058	130342
9	44710	8023	21349
10	5123	2302	2325
11	33078	42370	14131
12	53124	102736	25708

Завдання 3.3. Використовуючи дані з попереднього прикладу, розрахувати щільність зв’язку між наведеними показниками y, x_1 і x_2. Зробити висновки.

4. КЛАСТЕРНИЙ АНАЛІЗ

4.1. Суть кластерного аналізу

Класифікація об'єктів за осмисленими групами (кластерам) є важливою процедурою в області статистичних досліджень. Її широко застосовують економісти й статистики.

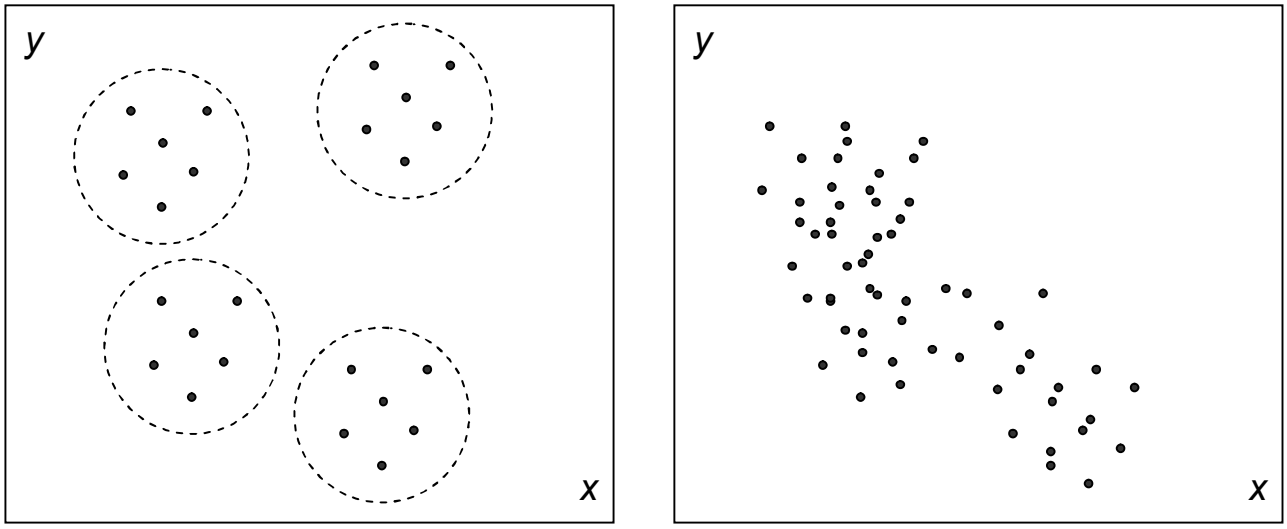
Метою кластерного аналізу є класифікація об'єктів на відносно гомогенні (однорідні) групи, виходячи з розглядуваного набору змінних.

Об'єкти в групі майже схожі відносно змінних й відрізняються від об'єктів в інших групах.

Кластерний аналіз складається з методів, з допомогою яких класифікують об'єкти або події за однорідними групами, що мають назву кластерів. Кластерний аналіз також називають класифікаційним аналізом, або числовою таксономією (систематикою).

Кластерний аналіз є одним із методів розвідувального аналізу даних, створених для виявлення яких-небудь можливих груп у всій сукупності даних. Основним критерієм для об'єднання даних є відстань: об'єкти, розташовані «близько» один від одного, мають потрапляти в той самий кластер, тоді як «досить далекі» об'єкти мають бути в різних кластерних групах.

На рис. 4.1, *а* показано ідеальну ситуацію кластеризації, коли кластери чітко відділено один від одного на основі двох змінних.



а б

Рис. 4.1. Дві протилежні ситуації кластеризації

На рис. 4.1, *б* показано реальну ситуацію, коли границі деяких кластерів обкреслено нечітко, і віднесення деяких змінних до конкретного кластера є неочевидним, оскільки багато з них не можна згрупувати в той чи інший кластер. Ця ситуація найчастіше трапляється на практиці.

Кластерний аналіз можна звести до чотирьох основних задач:

– розроблення типології й класифікації;

– дослідження корисних концептуальних схем групування об'єктів;

– породження гіпотез на основі дослідження даних;

– перевірка гіпотез або дослідження для визначення, чи дійсно виділені тим або іншим чином типи (групи) є серед існуючих даних.

На рис. 4.2 показано загальну схему проведення кластерного аналізу.

На першому етапі формулюють проблему кластерізації шляхом визначення змінних, на базі яких її буде проведено. Потім вибирають відповідний спосіб виміру відстані. Міра відстані є показником того, наскільки об'єкти, які кластеризують, схожі або не схожі між собою.

Розроблено кілька методів кластерізації, і досліднику необхідно вибрати найбільш прийнятний для вирішення цієї проблеми. Рішення про кількість кластерів приймає також дослідник. Сформовані кластери потрібно аналізувати з урахуванням змінних, використаних для їх одержання. Для профілювання кластерів можна використовувати додаткові явно виражені змінні. І нарешті, дослідник має оцінити достатність (якість) процесу кластеризації.

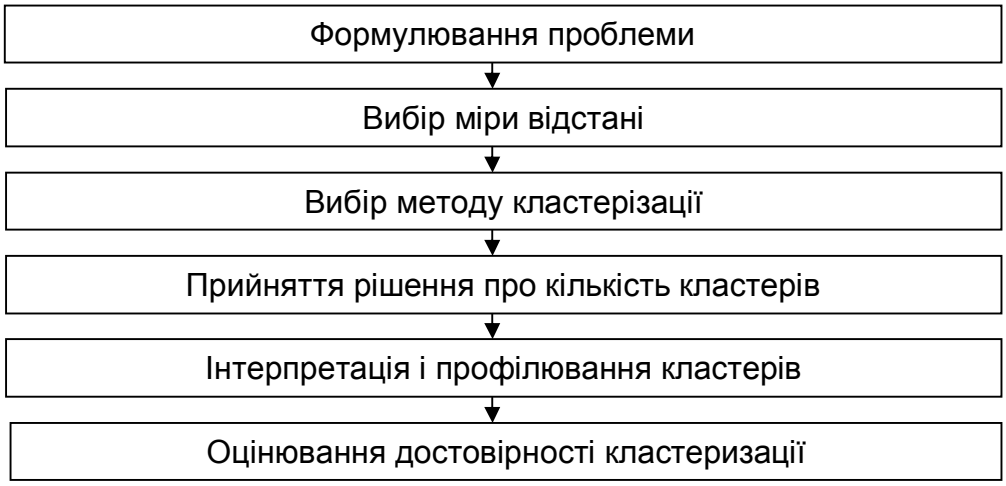


Рис. 4.2. Схема виконання кластерного аналізу

4.2. Методика кластерного аналізу

Можливо, найважливіша частина формулювання проблеми кластеризації – це вибір змінних, на основі яких проводять кластеризацію. Навіть одна стороння змінна (яка не належить до групи) може спотворити результати кластеризації. Завдання полягає в тому, щоб з допомогою вибраного набору змінних можна було описати подібність між об'єктами з урахуванням ознак, що стосуються проблеми дослідження. Змінні слід вибирати, виходячи з досвіду минулих досліджень, теорії або гіпотези, що тестується. Дослідник повинен мати інтуїцію й уміти робити висновки.

Мета кластеризації – групування схожих об'єктів. Тому для того щоб оцінити, наскільки вони схожі або не схожі, необхідно використовувати якусь одиницю виміру, наприклад відстань між двома об'єктами. Об'єкти з найменшими відстанями більше схожі між собою, ніж об'єкти з більшими відстанями. Існує кілька способів обчислення відстані між двома об'єктами:

– евклідовий простір або його квадрат (міра подібності, яка найчастіше використовується); евклідова відстань (геометрична відстань у багатомірному просторі) дорівнює квадратному кореню із суми різниць значень для кожної змінної:

$$d_{ij} = \sqrt{\sum_{k=1}^{m} \left(x_{ik} - x_{jk}\right)^2}\ , \qquad (4.1)$$

де x_{ij} – значення j-ї ознаки i-го об'єкта;

m – кількість ознак;

– відстань міських кварталів (манхетенська відстань) між двома об'єктами – сума абсолютних різниць у значеннях для кожної змінної;

– відстань Чебишева між двома об'єктами – максимальна абсолютна різниця в значеннях для будь-якої змінної;

– Хемінгова відстань

$$d_{ij} = \sum_{k=1}^{m} \left| x_{ik} - x_{jk} \right|^2 . \qquad (4.2)$$

Доцільно використовувати різні міри подібності й потім порівнювати результати, оскільки використання різних способів виміру відстані дає різні результати кластеризації. Вибравши міру подібності, можна вибрати метод кластеризації.

Методи кластеризації можуть бути ієрархічними й неієрархічними. Ієрархічна кластеризація характеризується будуванням ієрархічної або деревоподібної структури.

Ієрархічні методи можуть бути агломеративними (об'єднувальними) і дивізіонними. Агломеративна кластеризація починається з кожного об'єкта в окремому кластері. Кластери поєднують, групуючи об'єкти щоразу в усе більші й більші кластери. Цей процес триває доти, доки всі об'єкти не стануть членами одного єдиного кластера.

Дивізійна кластеризація починається з усіх об'єктів, згрупованих в єдиному кластері. Кластери розділяють (розщеплюють) доти, доки кожен об'єкт не опиниться в окремому кластері.

Зазвичай в економічних дослідженнях використовують агломеративні методи, наприклад методи зв'язку, дисперсійні й центроїдні методи. Методи зв'язку містять методи одиночного, повного й середнього зв'язку.

Основою методу одиночного зв'язку є розрахунок мінімальної відстані, або правило найближчого сусіда:

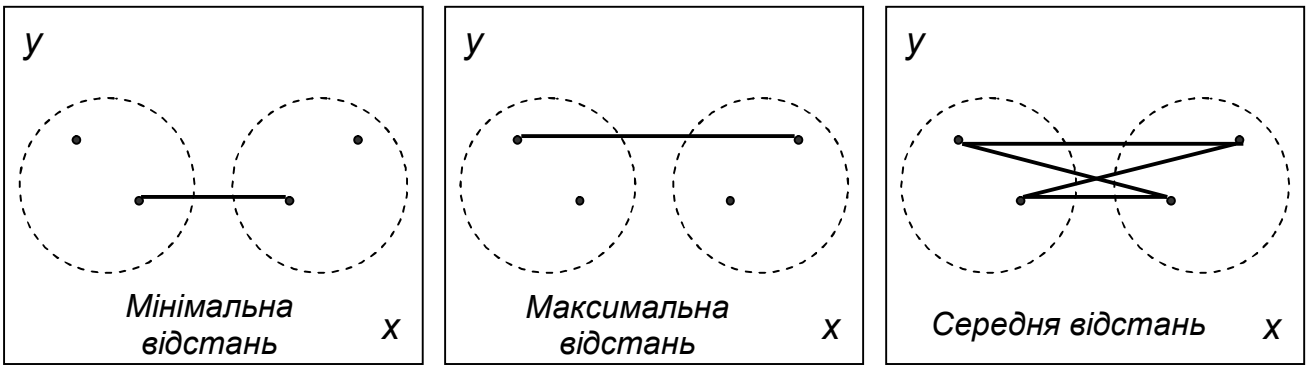
$$\begin{aligned} d(G_l, G_k) &= \min d(x_i, x_j), \\ x_i &\in G_l, \\ x_j &\in G_k . \end{aligned} \qquad (4.3)$$

При формуванні кластера першими поєднують два об'єкти, відстань між якими є мінімальною. Далі визначають наступну за величиною найменшу відстань, і в кластер з першими двома об'єктами вводять третій об'єкт. На кожній стадії відстань між двома кластерами являє собою відстань між їхніми найближчими точками (рис. 4.3, *а*).

На будь-якій стадії два кластери з'єднують по єдиній найкоротшій відстані між ними. Цей процес продовжують доти, доки всі об'єкти не буде об'єднано в кластер. Якщо кластери погано визначено, то метод

одиночного зв'язку працює недостатньо добре. Метод повного зв'язку аналогічний методу одиночного зв'язку, за винятком того, що його основою є розрахунок максимальної відстані між об'єктами, або правило далекого сусіда (рис. 4.3. *б*). У методі повного зв'язку відстань між двома кластерами обчислюють як відстань між двома їхніми окремими точками:

$$d(G_l, G_k) = max\, d(x_i, x_j),$$
$$x_i \in G_l, \qquad (4.4)$$
$$x_j \in G_k.$$



а б в

Рис. 4.3. Методи зв'язку для процедури кластеризації:
а – одиночний зв'язок; б – повний зв'язок; в – середній зв'язок

Метод середнього зв'язку діє аналогічно. Однак за цим методом відстань між двома кластерами визначають як середнє значення всіх відстаней, що вимірюються між об'єктами двох кластерів, при цьому кожну пару становлять об'єкти з різних кластерів (рис. 4.3, *в*):

$$d(G_l, G_k) = d(\bar{x}(l), \bar{x}(k)). \qquad (4.5)$$

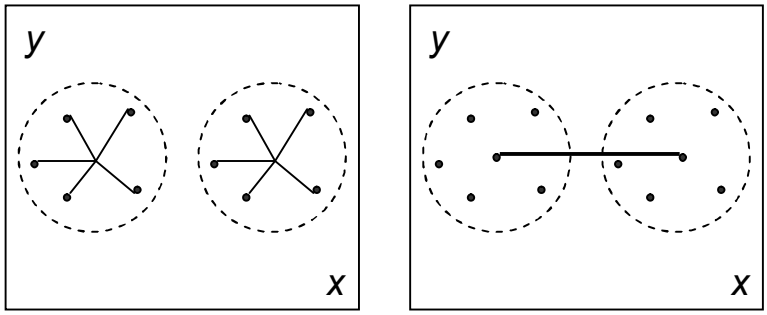
На рис. 4.3 видно, що в методі середнього зв'язку використовується інформація про всі відстані між парами, а не тільки про мінімальну або максимальну. Із цієї причини віддають перевагу зазвичай методу середнього зв'язку, а саме методу одиночного (або повного) зв'язку.

Дисперсійні методи формують кластери таким чином, щоб мінімізувати внутрішньокластерну дисперсію.

Широко відомим дисперсійним методом, який використовують для цієї мети, є метод Варда. Для кожного кластера обчислюють середні всіх змінних. Потім для кожного об'єкта обчислюють квадрати евклідових відстаней до кластерних середніх (рис. 4.4, *а*).

Ці квадрати відстаней сумуються для всіх об'єктів. На кожній стадії поєднують два кластери з найменшим приростом у повній внутрішньокластерній дисперсії. У центроїдних методах відстань між кластерами являє собою відстань між їх центроїдами (середніми для всіх змінних), як показано на рис. 4.4, *б*.

Ці методи містять послідовний граничний метод, паралельний граничний метод і оптимізаційний розподіл. При послідовному граничному методі вибирають центр кластера й всі об'єкти, що знаходяться в границях заданого центра порогового значення, групують разом. Потім вибирають новий кластерний центр і процес повторюють для незгрупованих точок. Після того як об'єкт віднесено до кластера із цим новим центром, його вже не розглядають як об'єкт для подальшої кластеризації.



а б

Рис. 4.4. Інші агломеративні методи кластеризації:
а – метод Варда; б – центроїдний метод

Аналогічно працює паралельний граничний метод, за винятком того, що одночасно вибирають декілька кластерних центрів, і об'єкти в границях граничного рівня групують із найближчим центром.

Метод оптимізаційного розподілу відрізняється від двох викладених вище граничних методів тим, що об'єкти можна згодом поставити у відповідність до інших кластерів (перерозподілити), щоб оптимізувати сумарний критерій, такий, як середня внутрішньокластерна відстань для певної кількості кластерів.

Вибір методу кластеризації залежить від міри відстаней і навпаки. Наприклад, квадрати евклідових відстаней використовують нарівні з методом Варда й центроїдним методом.

Ще один корисний графічний засіб відображення результатів кластеризації – це деревоподібна діаграма (дендрограма) (рис. 4.5).

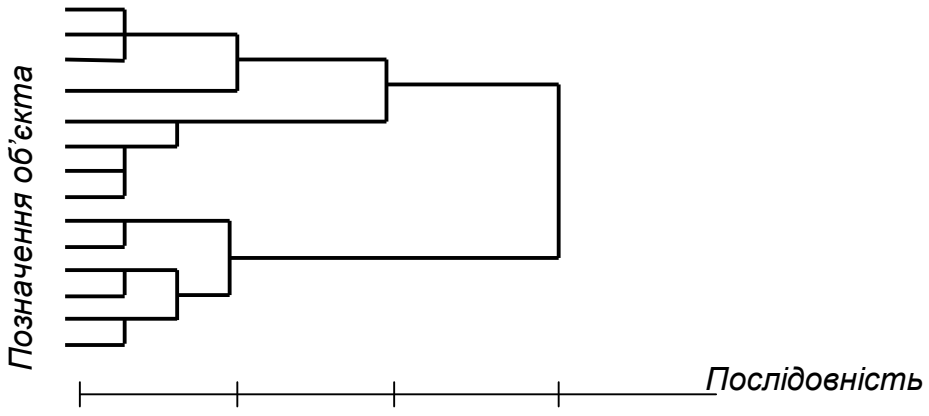


Рис. 4.5. Приклад деревоподібної діаграми

Деревоподібну діаграму читають зліва направо. Вертикальні лінії відображають кластери, об'єднані разом. Положення лінії відносно шкали відображає відстань, на якій кластери об'єднано. Оскільки багато відстаней на перших стадіях об'єднання приблизно однакової величини, важко описати послідовність, у якій об'єднано перші кластери. Однак зрозуміло, що на останніх двох стадіях відстані, на яких кластери мають об'єднатися, є досить великими. Ця інформація має сенс при прийнятті рішення про кількість кластерів.

Прийняття рішення про кількість кластерів – головне питання кластерного аналізу. Тут немає твердих правил, що дають змогу швидко прийняти рішення, але керуються таким.

1. При визначенні кількості кластерів можна брати до уваги теоретичні й практичні судження.

2. В ієрархічній кластеризації як критерій можна використовувати відстань, на якій поєднують кластери.

3. У неієрархічній кластеризації можна побудувати графік залежності сумарної внутрішньогрупової дисперсії і міжгрупової дисперсії від кількості кластерів. Точка, у якій спостерігається вигин або різкий поворот, є показником прийнятної кількості кластерів. Збільшення кількості кластерів зазвичай є недоцільним.

4. Відносні розміри кластерів мають бути досить виразними.

Запитання для самоконтролю

1. Назвіть основну мету кластерного аналізу.
2. Назвіть основні етапи кластерного аналізу.
3. Що таке евклідова відстань?
4. Які існують методи визначення відстані між об’єктами в кластерному аналізі?
5. Які основні методи кластерізації використовують на практиці? Класифікуйте їх.
6. Чим відрізняється метод Варда від центроїдного методу?
7. Що таке дендрограма?
8. Чим керуються при вирішенні питання про кількість кластерів?

Завдання для самостійної роботи

Завдання 4.1. Провести класифікацію п’яти точок, які характеризують обсяги продажів телефонів x_1 і планшетів x_2 за день у п’яти магазинах. Зробити відповідні висновки.

Таблиця 4.1

Вихідні дані

Магазин	1	2	3	4	5
x_1	4	7	2	2	0
x_2	3	4	0	1	4

5. МОДЕЛІ Й МЕТОДИ ФАКТОРНОГО АНАЛІЗУ

5.1. Застосування факторного аналізу

Факторний аналіз надає користувачеві адекватний інструмент дослідження системи ознак, що, зі свого боку, у деяких випадках дає змогу виявити логічну структуру складних явищ, відокремити взаємозалежні й взаємозамінні ознаки від незалежних, істотні від несуттєвих, обґрунтувати вибір тієї чи іншої системи ознак, оцінити її інформативність, перевірити або висунути гіпотези про взаємозв'язки. З його допомогою можна не тільки констатувати наявність якої-небудь проблеми, але й знайти шлях до її усунення.

Інакше кажучи, факторний аналіз часто використовують для зниження розмірності даних, щоб знайти невелику кількість факторів, які пояснюють більшу частину дисперсії, спостереженої для значної кількості явних змінних. Факторний аналіз може також використовуватися з метою формування гіпотез про механізми причинних зв'язків або перевірки змінних перед подальшим аналізом.

Часто в літературі також зазначається, що головними цілями факторного аналізу є зменшення кількості змінних, тобто редукція даних і визначення структури взаємозв'язків між змінними. Тому факторний аналіз використовується або як метод зменшення даних, або як метод класифікації змінних.

Виникнення факторного аналізу пов'язують з виходом 1901 року статті англійського вченого К. Пірсона «Перехід по лініях і площинах до візуальних систем точок у просторі», у якій було викладено ідею побудови головних осей. Визначальними для формування факторного аналізу в самостійний великий розділ статистичної науки стали роботи британського психолога Ч. Спірмена. 1904 року було опубліковано фундаментальну статтю «Загальні відомості про об'єктивні рішення й простір».

На основі теоретичних передумов Ч. Спірмена в 40-ві роки появилися глибокі розробки американських статистиків і математиків: Л. Гутмана, С. Хотелінга, Л. Терстоуна, К. Хользінгера, С. Рао, С. Барта, Д. Лоулі, А. Максвелла й ін.

Факторний аналіз використовують в таких ситуаціях:

– для визначення основних факторів, які пояснюють зв'язок у наборі змінних.

– для визначення нового, меншого за розміром, набору некорельованих змінних, що заміняють вихідний набір корельованих змінних, на основі якого далі виконують багатовимірний аналіз (регресійний або дискримінантний);

– для перетворення набору, більшого за розміром, на менший набір ясно виражених змінних для використання їх при наступному багатовимірному аналізі.

Факторний аналіз широко використовується в маркетингу при сегментації ринку для визначення латентних змінних з метою групування споживачів; при розробленні товарної і рекламної стратегій, стратегії ціноутворення.

Етапи виконання факторного аналізу показано на рис. 5.1.

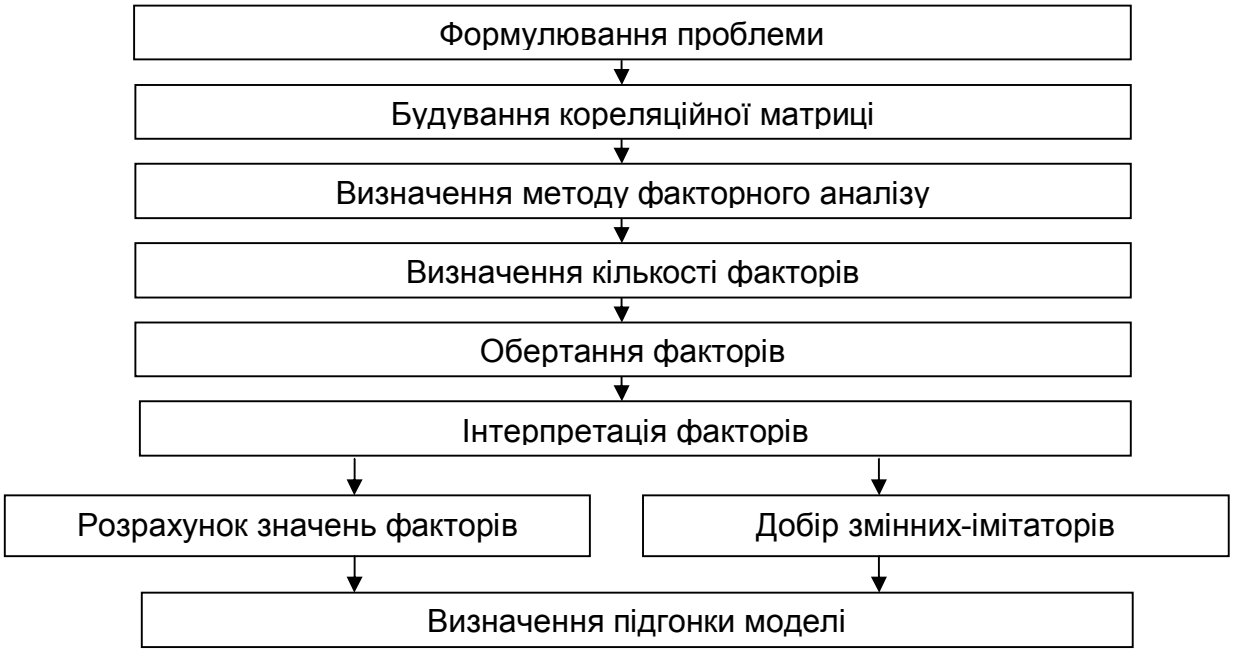


Рис. 5.1. Загальна схема виконання факторного аналізу

Суть першого етапу полягає у формулюванні проблеми факторного аналізу й визначенні змінних, що піддаються факторному аналізу. Потім будують кореляційну матрицю змінних, вибирають метод факторного аналізу, кількість факторів, які варто виділити, і метод обертання факторів. Далі вибрані фактори інтерпретують. Залежно від цілей обчислюють значення факторів або добирають змінні-замінники для подання факторів у подальшому багатовимірному аналізі. І нарешті, перевіряють, наскільки добре підігнано факторну модель.

5.2. Загальний алгоритм факторного аналізу

Модель факторного аналізу має вигляд

$$Z = AF, \tag{5.1}$$

де $Z = (Z_1, Z_2, ..., Z_m)^T$ – вектор нормалізованих ознак;

$F = (F_1, F_2, ..., F_k)^T$ – вектор-стовпець усіх загальних факторів;

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mk} \end{pmatrix}$$ – матриця навантажень на загальні фактори.

Модель факторного аналізу являє собою систему факторних рівнянь. Для того щоб визначити загальні фактори для всіх ознак, необхідно обчислити матрицю A, яку розраховують таким чином, щоб фактори ранжувалися за ступенем впливу на ознаку. Для цього спочатку знаходять власні числа кореляційної матриці $\hat{R}$ із рівняння

$$det(\hat{R} - \Lambda E) = 0, \tag{5.2}$$

де $\hat{R} = \begin{pmatrix} r(z_1, z_1) & r(z_1, z_2) & \cdots & r(z_1, z_m) \\ r(z_2, z_1) & r(z_2, z2) & \cdots & r(z_2, z_m) \\ \cdots & \cdots & \cdots & \cdots \\ r(z_m, z_1) & r(z_m, z_2) & \cdots & r(z_m, z_m) \end{pmatrix}$ – оцінка кореляційної матриці;

Λ – діагональна матриця власних чисел;

E – одинична матриця;

$r(z_i, z_j)$ – коефіцієнт кореляції між ознаками i та j.

Коефіцієнт кореляції між ознаками розраховують за вихідними даними з матриці «об'єкт-властивість»:

$$Z = \begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1m} \\ z_{21} & z_{22} & \cdots & z_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ z_{n1} & z_{n2} & \cdots & z_{nm} \end{pmatrix}. \tag{5.3}$$

З огляду на те, що Z_j – нормована випадкова величина,

$$r(z_i, z_j) = \sum_{k=1}^{m} Z_{ik} Z_{jk}. \tag{5.4}$$

Потім для одержання нормованого вектора переходу від вихідних ознак до головних компонентів (факторів) вирішують систему рівнянь

$$(\hat{R} - \lambda E) V^T = 0, \tag{5.5}$$

де λ – відповідне власне число.

За знайденою ортогональною матрицею V знаходимо матрицю вагових коефіцієнтів (навантажень):

$$A = V\Lambda^{1/2}. \tag{5.6}$$

Коефіцієнти матриці A є коефіцієнтами кореляції між центровано-нормованими вихідними ознаками й ненормованими головними компонентами і є показником наявності, міри й спрямованості лінійного зв'язку між відповідними вихідними ознаками й головними компонентами.

Початок алгоритму факторного аналізу, а саме перехід від матриці вихідних даних X до матриці стандартизованих значень ознак Z і далі до матриці кореляції R і коваріації S, не становить будь-яких ускладнень. Перехід від матриці X до матриці Z здійснюється після обчислення всіх елементів x_{ij} за формулою $z_{ij} = \frac{x_{ij} - \overline{x_j}}{\sigma_j}$.

Як відомо, для стандартизованих значень z_{ij} математичне сподівання дорівнює нулю ($E(z) = 0$) і дисперсія $D(z) = 1$.

На наступному кроці простим перемноженням скаляра $1/n$ і матриць Z' і Z отримаємо матрицю парних кореляцій: $R = 1/n\, Z'Z$. Перехід від матриці X до матриці Z можна пропустити, і тоді факторне рішення знайдемо не за матрицею R, а за матрицею коваріацій $S = 1/n\, X'X$ (тут бажано, щоб аналізовані ознаки X мали одні й ті самі одиниці виміру).

Виконання четвертого кроку алгоритму обумовлюється вирішенням першої проблеми – будуванням зредукованої матриці кореляцій (коваріацій). Проблема є актуальною саме для методів факторного аналізу, оскільки в методі головних компонент береться, що всю варіацію вихідних ознак повністю пояснюють латентні фактори, при цьому матриця парних кореляцій R розмірністю $m \times m$ залишається як редукована R_h, у якій всі множині $\sum h_j^2 = 1$:

$$R = \begin{pmatrix} 1 & r_{12} & r_{13} & \cdots r_{1m} \\ r_{21} & 1 & r_{23} & \cdots r_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ r_{m1} & r_{m2} & r_{m3} & \cdots\ 1 \end{pmatrix}. \qquad (5.7)$$

У факторному аналізі матриця кореляції R перетворюється на матрицю R_h з $h_1^2 < 1$, тобто варіацію ознак ($x_j, j = \overline{1,m}$) можна пояснити на 100 % і дещо менше, з урахуванням існування їхньої прихованої характерності:

$$R_h = \begin{pmatrix} h_1^2 & r_{12} & r_{13} & \cdots r_{1m} \\ r_{21} & h_2^2 & r_{23} & \cdots r_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ r_{m1} & r_{m2} & r_{m3} & \cdots h_m^2 \end{pmatrix}, \qquad (5.8)$$

Існують досить прості методи пошуку множин h_i^2.

Метод найбільшої кореляції. На головній діагоналі з додатним знаком записують найбільший за величиною коефіцієнт кореляції.

Метод Варта. У кожному стовпці матриці R спочатку знаходять середнє значення коефіцієнтів кореляції $\overline{r_j}$, потім, якщо $\overline{r_j}$ є порівняно великим, за множину беруть значення, яке є трохи більшим від найбільшого в стовпці коефіцієнта кореляції, і, якщо $\overline{r_j}$ є порівняно малим значенням, то множина буде дещо меншою за найбільший в стовпці коефіцієнт кореляції.

Метод тріад. Множини для кожного j-го стовпця матриці R обчислюють за формулою $n_j^2 = \frac{r_{ik} r_{il}}{r_{kl}}$, де r_{ik} і r_{il} – коефіцієнти кореляції, що є найбільшими в стовпці.

Метод малого центроїда. Для кожної змінної j будують кореляційну матрицю розмірністю 4×4. Додаючи саму змінну до цієї матриці, записують оцінки кореляції трьох інших змінних, особливо тісно пов'язаних з першою. За даними малої матриці кореляції розраховують множину:

$$h_j^2 = \frac{\left(\sum\limits_i r_{il} \right)^2}{\sum\limits_{i,j} r_{i,j}}, \qquad (5.9)$$

де $\sum\limits_i r_{il}$ – сума елементів першого стовпця;

$\sum\limits_{i,j} r_{i,j}$ – сума всіх елементів матриці 4×4.

Інша проблема виникає на етапі будування матриці відображення A і полягає у виборі оптимального методу для пошуку вагових коефіцієнтів a_{ij} елементів матриці A. Найкращі рішення зазвичай знаходять з допомогою сучасних методів факторного аналізу: головних факторів, максимальної правдоподібності та ін. У загальному випадку виділені фактори не обов'язково є ортогональними, тоді вектори (стовпці) матриці A будуть лінійно залежними.

Будування матриці факторного відображення і вирішення обертання простору загальних факторів не є обов'язковим. Потреба в цьому виникає, коли просторове розташування загальних факторів F_r є нелогічним або важко піддається інтерпретації.

Можливість виникнення алогічних перших результатів аналізу пояснюється положенням факторних осей в просторі, що нечітко визначається, або, інакше кажучи, відсутністю спочатку будь-якого просторового прив'язування для осей F_r.

На рис. 5.2 показано два різних положення в просторі факторних осей (F_1 і F_2). Легко помітити, що змінення положення F_1 і F_2 одночасно призводить до змінення координат початкових ознак x_j. Мета повороту – перетворення координат (факторних навантажень) таким чином, щоб факторотвірні ознаки мали найбільші навантаження, близькі до одиниці ($|a_{ir}| \to 1$), а інші – мінімальні значення, близькі до нуля, тобто добиваються економічного опису даних.

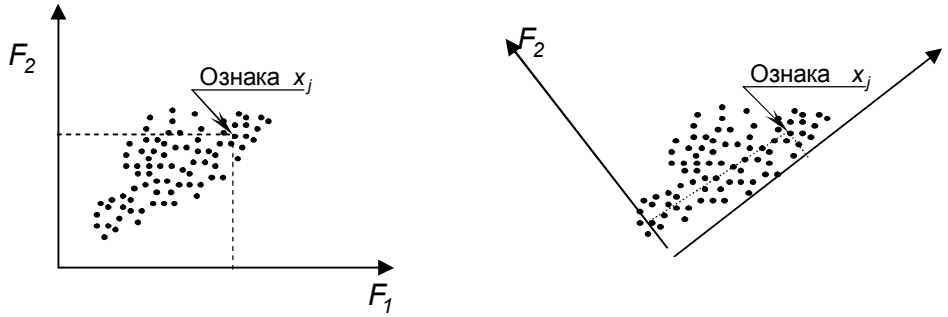
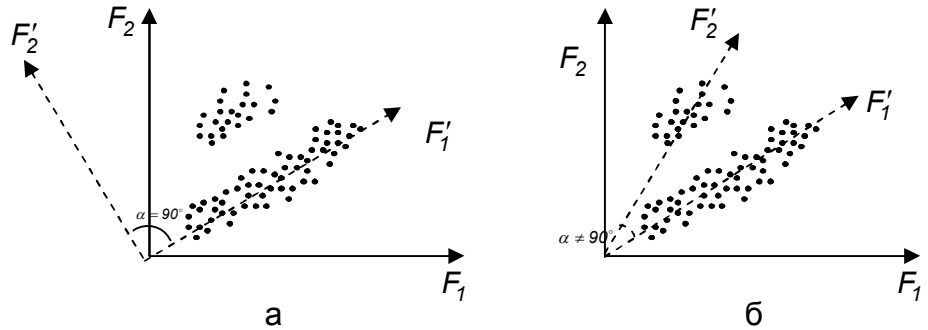


Рис. 5.2. Вихідні ознаки x_j у просторі загальних факторів F_1 і F_2

Повороти осей можуть бути ортогональними і косокутними (рис. 5.3).

При косокутному повертанні, хоча його і більш важко здійснити й інтерпретувати, значно підвищується можливість оптимального відображення згущень ознак у просторі R^F (рис. 5.3, *б*).

На рис. 5.3, *а* показано вісь F', яка після повороту осі F займає більш раціональне положення, але через жорсткість осьової конкуренції положення F_2 віддаляється від оптимального. На рис. 5.3, *б* показано оптимізацію положення відразу обох осей F_1 і F_2 після косокутного повертання ($\alpha \neq 90^\circ$). Повертання простору загальних факторів F_r не змінює величин множини h_j^2.



а б

Рис. 5.3. Гіпотетичні результати ортогонального (а) і косокутного (б) повертання простору загальних факторів

На заключному етапі алгоритму розраховують матрицю значень факторів F. Її елементи – це факторні значення f_{ir} для кожної одиниці спостереження. Тим самим визначаємо положення n об'єктів у просторі R^F з кількістю факторних осей r (рис. 5.4).

Алгоритми факторного аналізу різняться, як бачимо, трудомісткістю, їх повне виконання можливе за умови використання технічних засобів.

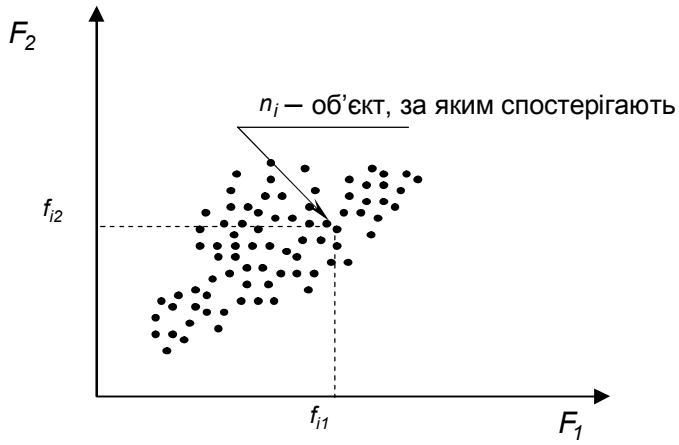


Рис. 5.4. Відображення об'єкта n_i, за яким спостерігають у просторі двох загальних факторів F_1 і F_2

5.3. Загальна математична модель методу головних компонент

З методів, що дають змогу узагальнювати значення елементарних ознак, метод головних компонент вирізняється простою логічною конструкцією. Метод головних компонент дає можливість за кількістю вихідних ознак виділити m головних компонент або узагальнених ознак. Простір головних компонент є ортогональним.

Математична модель головних компонент базується на логічному припущенні, що значення множини взаємозалежних ознак продовжують деякий загальний результат. Припустивши лінійну форму зв'язку ознак x_j, запишемо в матричній формі рівняння залежності результату F від X: $F = XB$, де B – вектор параметричних значень лінійного рівняння зв'язку. Умовою виконання такої рівності є відповідність дисперсій, тобто $D(X) = D(XB)$. Оскільки X – багатовимірна випадкова величина, її дисперсійна оцінка – це коваріаційна матриця S. Постійну величину B виносимо за знак дисперсій і підносимо до квадрата, після чого отримуємо

$$D(F) = B'SB. \tag{5.10}$$

Пошук головних компонент зводиться до задачі послідовного визначення першої головної компоненти F_1, що має максимальну дисперсію, другої головної компоненти, що має другу за величиною дисперсію і т. д. Така задача має розв'язок за умови введення обмежень

$$B'B = b_1^2 + b_2^2 + ... + b_m^2 = 1. \tag{5.11}$$

При $B'B = 1$ максимізуємо $B'SB$, використовуючи метод множинників Лагранжа. Функція Лагранжа для дисперсії має вигляд $r = B'SB - \lambda(B'B - 1)$, а необхідна умова екстремуму $\frac{\partial r}{\partial B} = 2SB - 2\lambda B = 0$. Звідси $SB - \lambda B = 0$.

Отже, отримаємо $|S - \lambda E|B = 0$, і характеристичне рівняння для пошуку власного числа λ_j буде мати вигляд $|S - \lambda E| = 0$. Із кількох значень характеристичних власних чисел λ_j вибираємо найбільше λ_1. Знаходимо вектор B_1 значень для першої головної компоненти F_1, для другого за величиною характеристичного власного числа λ_2 – вектор значень другої компоненти B_2 і так далі до λ_m і B_m для F_m (де m – кількість аналізованих ознак). Тут B – вектор величин, які є координатами головних компонент F_r у просторі ознак R^x і характеризують сили зв'язку r-ї головної компоненти і j-ї ознаки x_j.

Якщо вихідну матрицю даних X попередньо стандартизувати, то матриця коваріацій S перетвориться на матрицю парних кореляцій R і вектор B стане власним вектором стандартизованих даних U. Вирішальне рівняння в матричній формі набуває вигляду $(R - \lambda E)U = 0$.

Результатом застосування методу головних компонент є дані матриці відображення A. Остаточний запис залежності значень вихідних ознак від значень головних компонентів має вигляд

$$Z = AF' \text{ або } z_{ij} = a_{j1}f_{1i} + a_{j2}f_{12} + ... + a_{jr}f_{ri}, \qquad (5.12)$$

а залежність значень головних компонент від значень елементарних ознак

$$F = A^{-1}Z', \text{ або } f_{ri} = \frac{1}{\lambda_r}(a_{1r}z_{i1} + a_{2r}z_{i2} + ... + a_{mr}z_{im}), \qquad (5.13)$$

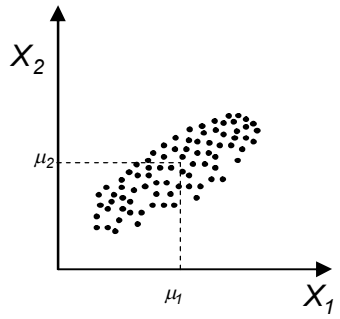
де z_{ij} – значення j стандартизованої змінної i-го об'єкта спостереження;

f_{ri} – r-та головна компонента F_r i-го об'єкта спостереження;

a_{jr} – ваговий коефіцієнт r-ї головної компоненти j-ї змінної;

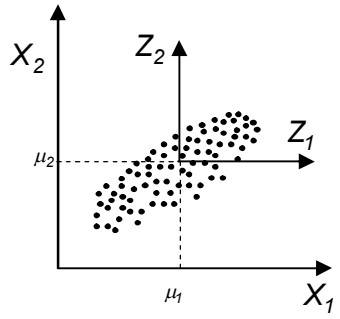
a_{mr} – ваговий коефіцієнт (характеристика сили зв'язку) m-ї елементарної ознаки r-ї головної компоненти.

Процедуру виокремлення головних компонент можна показати геометрично, якщо випадкова величина є двовимірною. На рис. 5.5 видно, що завдання виділення головних компонент зводиться до поетапного розв'язання класичних задач аналітичної геометрії: змінення масштабу простору, повороту координатної системи, координатного відображення векторів у старій і новій (після повороту) системах координат. На рис. 5.5, *в* можна бачити відображення Z на F і навпаки.



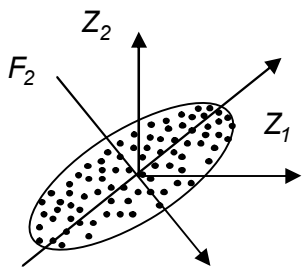
а

Спочатку є деякий емпіричний розподіл даних у двовимірному просторі ознак з центром (μ_1, μ_2).



б

Якщо початковий простір ознак піддається центруванню і стандартизації даних, то систему координат переносять у центр розподілу даних.



в

Шляхом розв'язання матричного рівняння $(R - \lambda E)V = 0$ знаходять параметри еліпса, який описує об'єкти, емпірично розподілені у нормованому просторі ознак R^z, потім знаходять положення головних компонент (осей), які узагальнюють варіацію ознак Z_1 і Z_2.

Рис. 5.5. Геометрична інтерпретація алгоритму методу головних компонент

5.4. Обчислювальна процедура методу головних компонент

Розв'язання задачі методом головних компонент зводиться до поетапного перетворення матриці вихідних даних розмірністю $n \times m$ (де n – кількість об'єктів спостереження, m – кількість елементарних аналітичних ознак):

$$X \longrightarrow Z \longrightarrow R(S) \begin{array}{c} \nearrow \Lambda \searrow \\ U \longrightarrow V \end{array} \searrow A \longrightarrow F$$

тут Z – матриця стандартизованих значень ознак (елементи матриці обчислюють за формулою $z_{ij} = \dfrac{x_{ij} - \overline{x_j}}{\sigma_j}$);

R – матриця парних кореляцій, $R = \dfrac{1}{n} Z'Z$.

Якщо попередня стандартизація даних не проводилася, то на цьому кроці отримують матрицю $S = \frac{1}{n} X'X$, елементи матриці X для розрахунку S будуть центрованими величинами: $x_{ij} = x_{ij} - \overline{x_j}$;

Позначимо через Λ діагональну матрицю власних (характеристичних) чисел:

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ 0 & 0 & \lambda_3 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & \lambda_m \end{pmatrix}. \qquad (5.14)$$

Множину значень λ_j знаходимо шляхом розв'язання характеристичного рівняння $|R - \lambda E| = 0$. Власні числа λ_j – це характеристики варіації, або показники дисперсії кожної головної компоненти. Сумарне значення $\sum \lambda_j$ дорівнює сумі дисперсій елементарних ознак X_j за умови стандартизації вихідних даних, коли $D(z_{ij}) = 1$, $\sum \lambda_j$ дорівнює кількості елементарних ознак m.

Позначимо через V матрицю нормованих власних (характеристичних) векторів. Кількість векторів V_j спочатку дорівнює m, тобто $j = \overline{1,m}$. Отримаємо V_j перетворенням ненормованих власних векторів U:

$$V_j = \frac{U_j}{|U_j|}, \qquad (5.15)$$

де $|U_j|$ – норма вектора U, тобто $|U_j| = (u_{1j}^2 + u_{2j}^2 + ... + u_{mj}^2)^{\frac{1}{2}}$.

Зі свого боку, власні вектори U_j знаходять з матричного рівняння $|R - \lambda E|U = 0$ для кожного λ_j при $j = \overline{1,m}$, що еквівалентно розв'язанню m систем лінійних рівнянь:

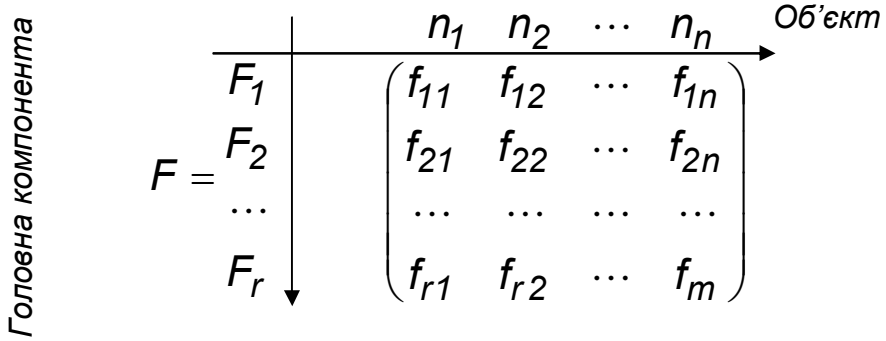
$$\begin{aligned} &(1 - \lambda_j)u_{1j} + r_{12}u_{2j} + r_{13}u_{3j} + ... + r_{1m}u_{mj} = 0, \\ &r_{21}u_{1j} + (1 - \lambda_j)u_{2j} + r_{23}u_{3j} + ... + r_{2m}u_{mj} = 0, \\ &r_{31}u_{1j} + r_{32}u2_{3j} + (1 - \lambda_j)u_{3j} + ... + r_{3m}u_{mj} = 0, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &r_{m1}u_{1j} + r_{m2}u_{2j} + r_{m3}u_{3j} + ... + (1 - \lambda_j)u_{mj} = 0. \end{aligned} \qquad (5.16)$$

Значення власних векторів отримують, задаючи довільну величину однієї компоненти кожного вектора і для спрощення розрахунків прирівнюючи її до одиниці.

Позначимо через A матрицю факторного відображення. Її елементи a_{rj} – вагові коефіцієнти. Спочатку A має розмірність $m \times m$ – за кількістю елементарних ознак x_j, потім в аналізі залишається r найбільш значущих компонент, $r \le m$. Обчислюємо матрицю A за відомими даними матриці власних чисел Λ і нормованих власних векторів V за формулою $A = V\Lambda^{1/2}$.

Матрицю F значень головних компонент розмірністю $r \times n$, знаходимо за формулами $F = A^{-1}Z'$, $F = \Lambda^{-1}A'Z'$ або $F = \Lambda^{-1/2}V'Z'$.

Матрицю F у загальному вигляді запишемо так:



5.5. Оцінювання рівня інформативності та пошук назв для головних компонент

Алгебраїчними перетвореннями матриці вихідних даних X виокремлюють головні компоненти F і встановлюють їх просторове розташування. Потім завдання розпізнавання головних компонент, визначення для них назв вирішують суб'єктивно на основі коефіцієнтів a_{jr} з матриці відображення A.

Для кожної головної компоненти F множину значень a_{jr} умовно розбивають на чотири підмножини з нечіткими границями:

W_1 – підмножина незначущих вагових коефіцієнтів;

W_2 – підмножина значущих вагових коефіцієнтів;

W_3 – підмножина значущих вагових коефіцієнтів, які не беруть участі у формуванні назви головної компоненти;

$W_2 - W_3$ – підмножина значущих вагових коефіцієнтів, які беруть участь у формуванні назви.

Додаткове виділення підмножини W_3 пояснюється прагненням до більш простої структури головної компоненти, яка завжди легше піддається інтерпретації. На своїх границях підмножина W_3 має критичні значення: $a_{кр_1}$ – максимальна кількість ознак, що пояснюють головну

компоненту, $a_{кр_2}$ – мінімальна кількість пояснювальних ознак. Загальний склад множини вагових коефіцієнтів показано на рис. 5.6.

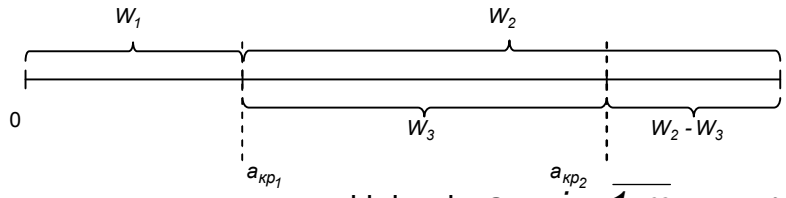


Рис. 5.6. Склад множин вагових коефіцієнтів a_{jr}, $\overline{j = 1,m}$, для r-ї головної компоненти

Підтвердження значущості ознак (x_j або z_j), що беруть участь у формуванні назви головної компоненти, отримують розрахунковим шляхом при визначенні коефіцієнта інформативності:

$$K_i = \frac{\sum_{j=1}^{m} a_{jr}^2 \{W_2 - W_3\}}{\sum_{j=1}^{m} a_{jr}^2}. \qquad (5.18)$$

Набір пояснювальних ознак вважається задовільним, якщо $K_i = 0{,}75...0{,}95$.

Запитання для самоконтролю

1. У яких ситуаціях використовують факторний аналіз?
2. З яких етапів складається загальна схема факторного аналізу?
3. Коли в маркетингу використовують факторний аналіз?
4. З яких основних кроків складається алгоритм факторного аналізу?
5. Які методи використовують у факторному аналізі для пошуку множин?
6. Що дає процес обертання у факторному аналізі?
7. Сформулюйте основну задачу методу головних компонент.
8. З яких кроків складається метод головних компонент?
9. Поясніть суть обчислювальної процедури методу головних компонент.
10. Як оцінюють рівень інформативності у факторному аналізі?

Завдання для самостійної роботи

Завдання 5.1. Знайти власні числа і власні вектори матриці парних кореляцій R, зробити відповідні висновки.

$$R = \begin{pmatrix} 1 & 0{,}641 & 0{,}104 \\ 0{,}641 & 1 & 0{,}352 \\ 0{,}104 & 0{,}352 & 1 \end{pmatrix}.$$

6. МОДЕЛІ Й МЕТОДИ ДИСКРИМІНАНТНОГО АНАЛІЗУ

6.1. Основні задачі дикримінантного аналізу

Дискримінантний аналіз, як і кластерний, найбільш яскраво відображає риси багатовимірного аналізу й класифікації, факторний аналіз використовують при дослідженні зв'язку.

Дискримінантний аналіз як розділ багатовимірного статистичного аналізу містить статистичні методи класифікації багатовимірних спостережень у ситуації, коли дослідник має так звані навчальні вибірки (класифікація з навчанням).

Загальна задача дискримінантного аналізу формулюється таким чином: задано k кластерів спостережень, описаних у m-вимірному просторі, і необхідно визначити, до якого кластера належить нове спостереження. Дискримінантний аналіз є наступним етапом багатовимірної класифікації, коли виокремлені класи описуються з допомогою певного правила, тобто спеціальних визначальних функцій, кожна з яких відповідає певному кластеру й набуває максимального значення тільки для спостережень цього кластера. Факт належності спостереження цьому кластеру підтверджується тим, що визначальна функція має максимальне значення.

Також дискримінантний аналіз використовують для аналізу даних у тому випадку, коли залежна змінна є категоріальною, а предикати (незалежні змінні) – інтервальними.

Цілі дискримінантного аналізу:

– визначення дискримінантних функцій або лінійних комбінацій незалежних змінних, з допомогою яких можна якнайкраще розділити (дискримінувати) категорії (групи) залежної змінної;

– перевірка існування між групами значущих розходжень з урахуванням незалежних змінних;

– визначення предикатів, що роблять найбільший внесок у міжгрупове розходження;

– віднесення випадків до однієї із груп (класифікація), виходячи зі значень предикатів;

– оцінювання точності класифікації даних по групах.

Найбільш відомим методом використання дискримінантного аналізу в економіці підприємства є визначення ймовірності банкрутства підприємства за системою економічних показників. Уперше вирішення цього завдання запропонував американський учений Е. Альтман 1968 року, розробивши коефіцієнт імовірності банкрутства.

Вибір методу дискримінантного аналізу залежить від кількості категорій залежних змінних. Якщо аналізуються дві категорії залежної змінної, то аналіз називають дискримінантним, якщо три або більше – множинним дискримінантним.

Модель дискримінантного аналізу має такий вигляд:

$$D = b_0 + b_1 X_1 + b_2 X_2 + ... + b_k X_k , \qquad (6.1)$$

де D – дискримінантний показник (дискримінант);

b_k – вага;

X – предиктор, або незалежна змінна.

Коефіцієнти, або ваги b_k, визначають таким чином, щоб групи максимально різнилися значеннями дискримінантної функції. Це відбувається тоді, коли відношення міжгрупової суми квадратів до внутрішньогрупової суми квадратів для дискримінантних показників є максимальним.

Будь-яка інша лінійна комбінація предикатів приводить до меншого значення цього відношення.

Виконання дискримінантного аналізу містить такі стадії: формулювання проблеми, обчислення коефіцієнтів дискримінантної функції, визначення значущості, інтерпретація й перевірка ймовірності.

Етапи виконання дискримінантного аналізу показано на рис. 6.1.

Рис. 6.1. Схема виконання дискримінантного аналізу

Перший крок дискримінантного аналізу – формулювання проблеми шляхом визначення цілей, залежної змінної і незалежних змінних. Залежна змінна має складатися з двох або більше взаємовиключних і взаємовичерпних категорій. Якщо залежну змінну виміряно з допомогою інтервальної або відносної шкали, то її треба насамперед звести до статусу категоріальної.

Наступний крок – поділ вибірки на дві частини, одну з них (аналізовану вибірку) використовують для обчислення дискримінантної функції, іншу (перевірну вибірку) – для перевірки дискримінантної функції. Обидві частини аналізують по черзі, міняючи їх місцями. Такий аналіз має назву подвійної перехресної перевірки.

І нарешті, перевірку ймовірності дискримінантної функції пропонують проводити неодноразово. Щоразу вибірку слід поділяти на дві частини: для аналізу й перевірки.

6.2. Класична модель дикримінантного аналізу

Нагадаємо, що метою дискримінантного аналізу є добір лінійних комбінацій таких показників, які найкращім чином поділяють сукупності. Ці сукупності можна поділити з допомогою прямої (дикримінантної функції), середня точка між сукупностями в такому випадку буде пороговим значенням. Цю функцію і порогові значення можна використати для класифікації майбутніх спостережень.

Нехай результатом спостережень над об'єктом є реалізація функції $X^{(0)} = (x_1^{(0)}, x_2^{(0)}, ..., x_m^{(0)})$ m-вимірного випадкового вектора $X = (x_1, x_2, ..., x_m)$. Відомо, що цей об'єкт належить до однієї із l генеральних сукупностей, до одного із класів K_j $(j = 1, 2, ..., l)$, відносно яких передбачається, що кожний клас K_j подано вибіркою обсягом n_j. Ці вибірки називають навчальними.

Потрібно побудувати правило дискримінації – правило розпізнавання класу, до якого належить об'єкт, що не ввійшов до вибірки $X^{(0)}$.

Розглянемо випадок для одновимірної випадкової величини $X (m = 1)$ і двох класів $l = 2$, взявши, що питома вага об'єктів кожного класу в загальній сукупності є однаковою. На рис. 6.2 зображено графіки функцій щільності $f_1(x)$ і $f_2(x)$ нормальних випадкових величин, що різняться тільки математичними сподівання a_1 і a_2.

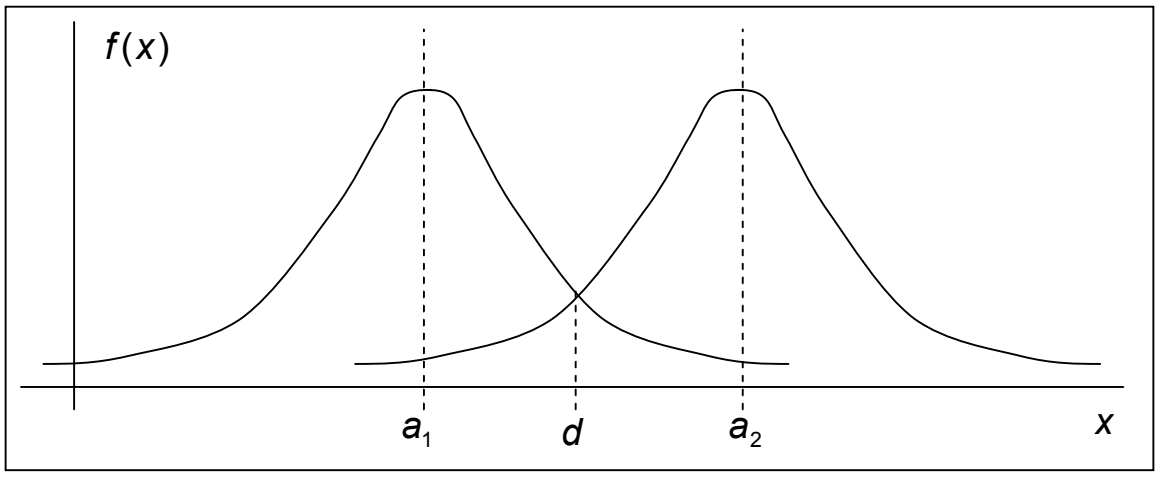


Рис. 6.2. Функції щільностей імовірності

Нехай d – деяка точка на осі x і правило дискримінації таке, що об'єкт, який класифікується, належить до першого класу тоді й тільки тоді, коли $x \le d$ (тут x – значення випадкової величини X певного об'єкта), і до

другого класу в інших випадках. Тут можна припуститися помилок двох видів:

– об'єкт, що належить до першого класу, буде віднесено до другого, а ймовірність цієї помилки

$$p_1 = P\{X > d\} = \int_d^{\infty} f_1(x)dx = 1 - \int_{-\infty}^{d} f_1(x)dx; \qquad (6.2)$$

– об'єкт, що належить до другого класу, буде віднесено до першого, а ймовірність цієї помилки

$$p_2 = P\{X \le d\} = \int_{-\infty}^{d} f_2(x)dx. \qquad (6.3)$$

Точку d знайдемо як розв'язок такої екстремальної задачі: за умови ймовірності $p_1 = p_2$, що рівносильно умові

$$\int_{-\infty}^{d} (f_1(x) + f_2(x))dx = 1, \qquad (6.4)$$

потрібно мінімізувати ймовірність p_2, тобто

$$\int_{-\infty}^{d} f_2(x)dx \to min. \qquad (6.5)$$

Використавши для розв'язання задачі метод множників Лагранжа, одержимо

$$\begin{cases} f_1(d) = \dfrac{1-\lambda}{\lambda}; \\ \int_{-\infty}^{d} (f_1(x) + f_2(x))dx = 1. \end{cases} \qquad (6.6)$$

Таким чином, задачі (6.4) і (6.5) рівносильні системі (6.6), що містить вимогу рівності ймовірності помилок і рівняння

$$\frac{f_1(d)}{f_2(d)} = const, \qquad (6.7)$$

де константа не залежить від d.

З урахуванням нормальності розподілу це рівняння рівнозначно рівнянню

$$\left(x - \frac{1}{2}(a_1 + a_2)\right)(a_1 + a_2)\frac{1}{\delta^2} = C, \qquad (6.8)$$

де C – деяка стала величина, що залежить від d;

δ – середньоквадратичне відхилення.

З розглянутого тривіального випадку (6.7) безпосередньо випливає, що

$$d = \frac{(a_1 + a_2)}{2}. \qquad (6.9)$$

Підставивши $x = d$ в (6.5), одержимо $C = 0$. Таким чином, сформульоване вище класифікаційне правило є еквівалентним такому: об'єкт належить до першого класу тоді й тільки тоді, коли

$$\left(x - \frac{1}{2}(a_1 + a_2) \right)(a_1 + a_2)\frac{1}{\delta^2} \geq 0, \qquad (6.10)$$

а в усіх інших випадках – до другого.

Класифікаційне правило для m-вимірного випадкового вектора X: об'єкт із координатами $X^{(0)} = (x_1^{(0)}, x_2^{(0)}, ..., x_m^{(0)})$ належить до першого класу тоді й тільки тоді, коли

$$\left(X^{(0)} - \frac{1}{2}(\hat{a}_1 + \hat{a}_2) \right)\hat{\Sigma}^{-1}(\hat{a}_1 + \hat{a}_2)^r \geq 0, \qquad (6.11)$$

і до другого – в усіх інших випадках.

У співвідношенні (6.11) $\hat{a}_1 = (\hat{a}_1^{(1)}, \hat{a}_2^{(1)}, ..., \hat{a}_m^{(1)})$ – вектор середніх значень випадкової величини $X_1, X_2, ..., X_m$ у вибірці обсягом n_1 із першого класу; $\hat{a}_2 = (\hat{a}_1^{(2)}, \hat{a}_2^{(2)}, ..., \hat{a}_m^{(2)})$ – вектор середніх значень цих величин у вибірці обсягом n_2 з другого класу; $\hat{\Sigma}$ – розрахована за навчальними вибірками оцінка коваріаційної матриці вектора X, що є загальною для двох класів.

Статистичні оцінки характеристик навчальних вибірок розраховують за формулами

$$\hat{a}_j^{(1)} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{ij}, \qquad (6.12)$$

$$\hat{a}_j^{(2)} = \frac{1}{n_2}\sum_{i=1}^{n_2} x_{ij}, \qquad (6.13)$$

$$cov(X_j, X_k) = \frac{cov^{(1)}(X_j, X_k)n_1 + cov^{(2)}(X_j, X_k)n_2}{n_1 + n_2 - 2}, \qquad (6.14)$$

$$cov^{(1)}(X_j, X_k) = \frac{1}{n_1}\sum_{i=1}^{n_1}(x_{ij} - \hat{a}_j^{(1)})(x_{ik} - \hat{a}_k^{(1)}), \qquad (6.15)$$

$$cov^{(2)}(X_j, X_k) = \frac{1}{n_2}\sum_{i=1}^{n_2}(x_{ij} - \hat{a}_j^{(2)})(x_{ik} - \hat{a}_k^{(2)}), \qquad (6.16)$$

де x_{ij} – значення випадкової величини X_j для простору i-го об'єкта;

n_1 – обсяг вибірки класу K_1;

n_2 – обсяг вибірки класу K_2.

За співвідношенням (6.11) знаходять вид дискримінаційної функції для двох сукупностей, розподілених за нормальним законом.

Якщо $l \geq 2$, а частини об'єктів у загальній генеральній сукупності дорівнюють $П_1, П_2, ..., П_l$, то узагальнене класифікаційне правило формулюється таким чином: об'єкт з координатами $X^{(0)} = \{x_1^{(0)}, x_2^{(0)}, ..., x_m^{(0)}\}$ належить до класу з номером j_0 тоді й тільки тоді, коли

$$\left(X^{(0)} - \frac{1}{2}(\hat{a}_{j0} + \hat{a}_j) \right) \hat{\Sigma}^{-1} (\hat{a}_{j0} + \hat{a}_j)^r \geq ln(\frac{\pi_j}{\pi_{j0}}) \qquad (6.17)$$

для всіх $j = 1, 2, ..., l$. При виконанні цього правила втрати від неправильної класифікації є постійними й мінімальними.

6.3. Загальний алгоритм дискримінантного аналізу

При дискримінантному аналізі застосовують статистичні методи класифікації багатовимірних об'єктів у ситуації, коли дослідник має так звані навчальні вибірки (класифікація з навчанням).

У загальному випадку задача дискримінації формулюється таким чином: нехай результатом спостереження над об'єктом є реалізація k-вимірного випадкового вектора $X = (x_1, x_2, ..., x_k)^T$; потрібно визначити правило, відповідно до якого за спостережуваним значенням вектора X об'єкт належить до однієї з можливих сукупностей $\varphi_i, i = 1, 2, ..., l$.

Для будування правила дискримінації весь вибірковий простір R значень вектора X розбивають на області R_i, $i = 1, 2, ..., l$ так, що при вилученні X в R_i об'єкт належатиме сукупності φ_i.

Правило дискримінації вибирається відповідно до певного принципу оптимальності на основі апріорної інформації, яку можна подати у вигляді як деяких відомостей про функцію k-вимірного розподілу ознак у кожній сукупності, так і вибірок із цих сукупностей. Апріорні ймовірності можуть бути заданими і незаданими.

Найчастіше вихідну інформацію про розподіл подають вибірками. У цьому випадку задача дискримінації ставиться таким чином.

Нехай $x_1^1, ..., x_j^i, ..., x_n^l$ – вибірка із сукупності φ_i, $i = 1, 2, ..., l$, $j = 1, 2, ..., n$, причому j-й об'єкт вибірки подано k-вимірним вектором параметрів $x_j^i = (x_{j1}^i, ..., x_{jq}^i, ..., x_{jk}^i)^T$. Проведено додаткове спостереження $x = (x_1, ..., x_k)^T$ над об'єктом, що належить до сукупності φ_i. Потрібно побудувати правило віднесення спостереження x до однієї із сукупностей.

Зазвичай в задачі дискримінації переходять від вектора ознак до лінійної функції та дискримінантної функції (гіперплощини), що якнайкраще розподіляють сукупність вибіркових точок. Ці точки використовують для оцінювання параметрів статистичних функцій розподілу. Зазвичай для будування функції використовують нормальний розподіл.

Нехай є дві генеральні сукупності X і Y, що мають тривимірний закон розподілу з невідомими, але рівними коваріаційними матрицями. З них взято початкові вибірки з обсягами n_1 в X і n_2 в Y:

$$X = \begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \end{pmatrix}, \; Y = \begin{pmatrix} y_{11} & y_{12} & y_{13} \\ y_{21} & y_{22} & y_{23} \\ y_{31} & y_{32} & y_{33} \end{pmatrix}. \qquad (6.18)$$

Метою дискримінантного аналізу є віднесення нового спостереження (рядка матриці Z) або до X, або до Y:

$$Z = \begin{pmatrix} z_{11} & z_{12} & z_{13} \\ z_{21} & z_{22} & z_{23} \\ z_{31} & z_{32} & z_{33} \end{pmatrix}. \qquad (6.19)$$

Для розв'язання задачі за початковими вибірками визначено вектори середніх:

$$\overline{X} = \begin{pmatrix} \overline{x_1} \\ \overline{x_2} \\ \overline{x_3} \end{pmatrix}; \; \overline{Y} = \begin{pmatrix} \overline{y_1} \\ \overline{y_2} \\ \overline{y_3} \end{pmatrix}. \qquad (6.20)$$

Алгоритм містить такі кроки.

1. Визначення оцінки коваріаційних матриць:

$$S_x = (S_{ki})_x \text{ і } S_y = (S_{ki})_y; \; \overline{x_j} = \frac{1}{n_1} \sum_{i=1}^{n_1} x_{ij} \; . \qquad (6.21)$$

Знайдемо елемент матриці S_x:

$$S_{ki}(x) = \frac{1}{n_1} \sum_{i=1}^{n_1} (x_{ij} - \overline{x_j})(x_{jk} - \overline{x_k}) = \overline{x_j x_k} - \overline{x_j}\,\overline{x_k}; \; j,k = 1,2,3, \qquad (6.22)$$

де $\overline{x_j}$ і $\overline{x_k}$ – середні значення.

2. Визначення незміщеної оцінки сумарної коваріаційної матриці:

$$\hat{S} = \frac{1}{n_1 + n_2 - 2}(n_1 S_x + n_2 S_y). \qquad (6.23)$$

3. Визначення матриці $\hat{S}^{-1}$, оберненої до матриці $\hat{S}$.

4. Обчислення вектора оцінок коефіцієнтів дискримінантної функції: $a = \hat{S}^{-1}\left(\overline{x} - \overline{y}\right)$.

5. Обчислення оцінок векторів значень дискримінантної функції для матриць вихідних даних: $\hat{U}_x = Xa, \; \hat{U}_y = Ya$.

6. Обчислення середніх значень оцінок дискримінантної функції:

$$\overline{\hat{U}_x} = \frac{1}{n_1}\sum_{i=1}^{n_1}\hat{U}_{xi},\ \overline{\hat{U}_y} = \frac{1}{n_2}\sum_{i=1}^{n_2}\hat{U}_{yi}\ . \qquad (6.24)$$

7. Визначення константи: $\hat{C} = \frac{1}{2}\left(\overline{\hat{U}_x} + \overline{\hat{U}_y}\right)$.

8. Одержання дискримінантної функції для v-го спостереження, що підлягає дискримінації, з рівняння

$$\hat{U}_v = z_{v1}a_1 + z_{v2}a_2 + z_{v3}a_3\ . \qquad (6.25)$$

Якщо $\hat{U}_v \geq \hat{C}$, то v-те спостереження треба віднести до X, якщо ж $\hat{U}_v < \hat{C}$, то до сукупності Y.

Запитання для самоконтролю

1. Чим відрізняються кластерний і факторний види аналізу від дискримінантного?
2. Сформулюйте загальну задачу дискримінантного аналізу.
3. Якими є цілі дискримінантного аналізу?
4. Який вигляд має модель дискримінантного аналізу?
5. Якими є основні етапи в алгоритмі дискримінантного аналізу?
6. У чому полягає суть класичної моделі дискримінантного аналізу?
7. Які основні етапи містить алгоритм дискримінантного аналізу?

Завдання для самостійної роботи

Завдання 6.1. З допомогою методів дискримінантного аналізу за наведеними даними (табл. 6.1) провести дискримінацію підприємств, які характеризуються продуктивністю праці X_1, коефіцієнтом змінності обладнання X_2, питомою вагою втрат від браку X_3, і віднести їх до відповідного класу. Дати економічну інтерпретацію і зробити висновки.

Таблиця 6.1

Вихідні данні

Клас	Номер підприємства	X_1	X_2	X_3
А	1	9,26	1,37	0,23
	2	9,38	1,49	0,39
	3	12,11	1,44	0,43
	4	10,81	1,42	0,18
Б	5	5,49	1,10	0,05
	6	6,61	1,23	0,48
	7	4,32	1,39	0,41
	8	7,37	1,38	0,62
В’	9	6,7	0,79	0,39
	10	9,42	0,7	0,72

БІБЛІОГРАФІЧНИЙ СПИСОК

1. Орлов, А. И. Прикладная статистика [Текст] : учебник для вузов / А. И. Орлов. – М. : Экзамен, 2004. – 656 с.
2. Булашев, С. В. Статистика для трейдеров [Текст] / С. В. Булашев. – М. : Спутник+, 2003. – 245 с.
3. Сондерс, М. Методы проведения экономических исследований [Текст] : пер. с англ. / М. Сондерс, Ф. Льюис, Э. Торнхилл. – М. : Эксмо, 2006. – 640 с.
4. Дубров, А. М. Многомерные статистические методы [Текст] : учебник / А. М. Дубров, В. С. Мхитарян, Л. И. Трошин. – М. : Финансы и статистика, 2003. – 352 с.
5. Афанасьев, В. Н. Анализ временных рядов и прогнозирование [Текст] : учебник / В. Н. Афанасьев, М. М. Юзбашев. – М. : Финансы и статистика, 2001. – 228 с.
6. Антохонова, И. В. Методы прогнозирования социально-экономических процессов [Текст] : учеб. пособие / И. В. Антохонова. – Улан-Удэ : Изд-во ВСГТУ, 2004. – 212 с.
7. Карп, Д. Б. Эконометрика: основные формулы с комментариями [Текст] : учеб.-метод. пособие / Д. Б. Карп. – Владивосток : Изд-во ДВГАЭУ, 2004. – 50 с.
8. Вуколов, Э. А. Основы статистического анализа [Текст] : учеб. пособие / Э. А. Вуколов. – М. : Форум, 2007. – 460 с.
9. Солодухін, С. В. Методи економіко-статистичних досліджень [Текст] : навч. посіб. / С. В. Солодухін, О. М. Ісаєнко, В. В. Хорошун. – Запоріжжя : ЗДІА, 2012. – 132 с.
10. Грицюк, С. Н. Математические методы и модели в экономике [Текст] : учебник / С. Н. Грицюк, Е. В. Мирзоева, В. В. Лысенко. – Ростов н/Д. : Феникс, 2007. – 348 с.
11. Добронец, Б. С. Интервальная математика [Текст] : учеб. пособие / Б. С. Добронец. – Красноярск : Краснояр. гос. ун-т, 2004. – 216 с.
12. Алтунин, А. Е. Модели и алгоритмы принятия решений в нечетких условиях [Текст] : монография / А. Е. Алтунин, М. В. Семухин. – Тюмень : Изд-во Тюмен. гос. ун-та, 2000. – 352 с.
13. Теория статистики [Текст] : учебник / под ред. Г. Л. Громыко. – М. : Инфра-М, 2002. – 414 с.
14. Дубнов, П. Ю. Обработка статистической информации с помощью SPSS [Текст] / П. Ю. Дубнов. – М. : ООО «Изд-во АСТ» ; Изд-во «НТ Пресс», 2004. – 221 с.
15. Халафян, А. А. STATISTICA 6. Статистический анализ данных [Текст] : учебник / А. А. Халафян. – М. : ООО «Бином-Пресс», 2007. – 512 с.
16. Ефимова, М. Р. Общая теория статистики [Текст] : учебник / М. Р. Ефимова, Е. В. Петрова, В. Н. Румянцев. – М. : Инфра-М, 2002. – 416 с.
17. Наследов, А. Д. SPSS 15: профессиональный статистический анализ данных [Текст] / А. Д. Наследов. – СПб. : Питер, 2008. – 416 с.

ЗМІСТ

ВСТУП…… 3
1. СТРУКТУРА ЧАСОВИХ РЯДІВ І ТРЕНДОВІ МОДЕЛІ……..... 4
1.1. Часові ряди і їхні компоненти…… 4
1.2. Види трендових моделей……... 5
1.3. Методи ідентифікації тренду у часовому ряді, його виду і параметрів…… 8
1.4. Оцінювання якості й адекватності трендових моделей……. 12
2. СТАТИСТИЧНИЙ АНАЛІЗ ПЕРІОДИЧНИХ КОЛИВАНЬ…… 16
2.1. Типи коливань і їхні характеристики…… 16
2.2. Статистичний аналіз сезонної нерівномірності на основі розрахунку індексів сезонності…….. 18
2.3. Дослідження періодичних коливань методами спектрального аналізу…… 19
3. МЕТОДИ І МОДЕЛІ КОРЕЛЯЦІЙНО-РЕГРЕСІЙНОГО АНАЛІЗУ……. 22
3.1. Види зв’язку між змінними, класифікація функцій регресії……. 22
3.2. Методика кореляційного аналізу…….. 24
3.3. Методика регресійного аналізу……. 27
4. КЛАСТЕРНИЙ АНАЛІЗ…….. 35
4.1. Суть кластерного аналізу……... 35
4.2. Методика кластерного аналізу…….. 37
5. МОДЕЛІ Й МЕТОДИ ФАКТОРНОГО АНАЛІЗУ…….. 42
5.1. Застосування факторного аналізу…… 42
5.2. Загальний алгоритм факторного аналізу…… 43
5.3. Загальна математична модель методу головних компонент…… 49
5.4. Обчислювальна процедура методу головних компонент……. 50
5.5. Оцінювання рівня інформативності та пошук назв для головних компонент…….. 52
6. МОДЕЛІ Й МЕТОДИ ДИСКРИМІНАНТНОГО АНАЛІЗУ……..…… 54
6.1. Основні задачі дикримінантного аналізу…… 54
6.2. Класична модель дискримінантного аналізу…… 56
6.3. Загальний алгоритм дискримінантного аналізу…… 59
Бібліографічний список…… 62

Навчальне видання

Ревенко Даніїл Сергійович
Либа Василь Олексійович

МЕТОДИ ЕКОНОМІКО-СТАТИСТИЧНИХ ДОСЛІДЖЕНЬ

Редактор О.Ф. Серьожкіна

Зв. план, 2014
Підписано до видання 03.10.2014
Ум. друк. арк 3,5. Обл.-вид. арк. 4. Електронний ресурс

Видавець і виготовлювач
Національний аерокосмічний університет ім. М. Є. Жуковського
«Харківський авіаційний інститут»
61070, Харків-70, вул. Чкалова, 17
http://www.khai.edu
Видавничий центр «ХАІ»
61070, Харків-70, вул. Чкалова, 17
izdat@khai.edu

Свідоцтво про внесення суб'єкта видавничої справи
до Державного реєстру видавців, виготовлювачів і розповсюджувачів
видавничої продукції сер. ДК № 391 від 30.03.2011