МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ТЕРНОПІЛЬСЬКИЙ НАЦІОНАЛЬНИЙ ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ

О. Я. Ковальчук

МАТЕМАТИЧНЕ МОДЕЛЮВАННЯ ТА ПРОГНОЗУВАННЯ В МІЖНАРОДНИХ ВІДНОСИНАХ

Підручник

Тернопіль
ТНЕУ
2019

УДК 519.86:330,4:327
К 56

Рецензенти:

Хіміч Олександр Миколайович, доктор фізико-математичних наук, професор, членкореспондент НАН України, заступник директора Інститут кібернетики НАН України імені В. М. Глушкова;

Максишко Наталія Костянтинівна, доктор економічних наук, професор, академік Академії економічних наук України, завідувачка кафедри економічної кібернетики Запорізького національного університету;

Грод Іван Миколайович, доктор фізико-математичних наук, доцент, професор кафедри математики та методики її викладання Тернопільського національного педагогічного університету імені Володимира Гнатюка.

Затверджено на засіданні вченої ради ТНЕУ
(протокол № 3 від 30.10.2019 р.)

Ковальчук О. Я.

К 56 Математичне моделювання тапрогнозування в міжнародних відносинах. Тернопіль : ТНЕУ, 2019. с. 412.

ISBN 978-966-654-570-4

У підручнику розглянуто питання застосування математичних методів та моделей у практиці дослідження міжнародних ситуацій і процесів для встановлення взаємозв'язків між суб'єктами міжнародних відносин, виявлення неочевидних ресурсів та можливостей взаємодії на міжнародній арені, прогнозування майбутніх станів та перевірки гіпотез про ймовірні сценарії розвитку ситуації й можливі сценарії дій.

Теоретичний матеріал розподілено на сім змістовних блоків, в яких викладено такі програмні питання: методологія застосування традиційних статистично-математичних методів для побудови моделей процесів та явищ міжнародних відносин, зокрема дисперсійного, кореляційно-регресійного, факторного, дискримінантного, кластерного і канонічного аналізів, прогнозування за моделлю ARIMA, аналізу виживання та аналізу відповідності. Висвітлено можливості застосування інструментарію data mining для дослідження міжнародних відносин, зокрема на основі аналізу асоціативних правил, нейронних мереж, споживчого скорингу та автоматичної класифікації текстів. Розглянуто можливості застосування інструментів моделювання для розробки оптимальних рішень у МВ, зокрема теорії ігор, теорії катастроф та інформаційних технологій big data.

Адресований студентам гуманітарних спеціальностей ВНЗ, які навчаються за напрямом «Міжнародні відносини» та вивчають дисципліну «Математичне моделювання та прогнозування в міжнародних відносинах». Буде корисним студентам та аспірантам, які працюють над курсовими, дипломними і науковими роботами, викладачам та фахівцям із міжнародних відносин.

УДК 519.86:330,4:327
К 56

ISBN 978-966-654-570-4

ЗМІСТ

ПЕРЕДМОВА

Ми живемо у світі, який швидко змінюється. Процеси прискорюються, а зв'язки стають складнішими. Стрімкий розвиток інформаційних технологій та збільшення обчислювальних потужностей сучасних комп'ютерів привели до різкого зростання ролі моделювання та прогнозування в міжнародних відносинах. Більш складні динамічні проблеми сучасної світової політики та економіки потребують більш складних і динамічних дій та рішень. Моделювання надає корисний інструмент для вирішення важливих проблем, з якими стикається світ сьогодні.

Моделі, побудовані з використанням математичних та статистичних методів, застосовують для вирішення завдань у різних сферах, включаючи соціальні науки та науку про державу, а також управління бізнесом, демографію, освіту та охорону здоров'я, енергетичну та екологічну політику, медичні науки, державну політику, економіку, міжнародні відносини (МВ) та багато інших дисциплін.

Незважаючи на безліч нових аналітичних інструментів та передових методів, які науковці застосовують для моделювання подій і процесів, що відбуваються у сфері міжнародних відносин, багато прийнятих рішень не вдається реалізувати на практиці, а окремі з них спричиняють несподівані побічні ефекти чи несприятливі наслідки.

Традиційні статистичні методи визначають залежну змінну як функцію кількох незалежних змінних і перевіряють взаємозв'язок між цими змінними на основі емпіричних даних. Ці моделі можуть виявляти статистичні кореляції та дають можливість робити короткострокові прогнози на основі встановлених кореляційних зв'язків.

За допомогою економетричних методів будують економіко-математичні моделі, які визначають кількісні та якісні взаємозв'язки між об'єктами (процесами чи явищами) міжнародних відносин на основі застосування математичних і статистичних методів. Враховуючи стрімку динаміку розвитку сучасного світу, науковці дедалі більше доповнюють традиційні стохастичні чи економетричні моделі політичних та економічних явищ і процесів методами динамічного моделювання, отриманими з багатопрофільних перспектив.

Методи багатовимірного моделювання використовують для пошуку значущих ознак процесу чи явища, що вивчається, агрегування початкових даних, виявлення спільних властивостей досліджуваних об'єктів та побудови функцій їх поділу на групи за схожими ознаками, виявлення взаємних залежностей між рівнями аналізованих факторів та наборами екзогенних та ендогенних чинників.

Методи інтелектуального аналізу даних (data mining) призначені для виявлення латентних взаємозв'язків між досліджуваними ознаками, кластеризації, класифікації, прогнозування, дослідження структур даних, аналізу неструктурованої інформації, виявлення асоціативних правил тощо.

Моделі теорії ігор та теорії катастроф застосовують для пошуку раціональних рішень з метою підвищення конкурентоспроможності й ефективності управлінн на регіональному і державному рівнях.

Технології великих даних (big data) використовують для аналізу різнорідної інформації великих обсягів і різноманітного складу, отриманої з різних джерел, та ефективного інтелектуального управління.

Науковий підхід, системне мислення та математичне моделювання надають інструменти для вироблення оптимальних рішень. Результати застосування математичних моделей можуть пояснити багато питань політології, економіки та міжнародних відносин, виявити неочевидні зв'язки, що спростять вирішення проблем у складних умовах, забезпечити людству виживання та процвітання в цьому нестабільному світі.

РОЗДІЛ 1

МОДЕЛЮВАННЯ ЯК МЕТОД ДОСЛІДЖЕННЯ МІЖНАРОДНИХ ВІДНОСИН

Моделювання у міжнародних відносинах – метод наукового дослідження, пов'язаний з абстрагованим відображенням реальних явищ, ситуацій та процесів МВ за допомогою ідеалізованих об'єктів і зв'язків між ними.

На сьогодні розроблено низку моделей міжнародних систем, соціально-економічних процесів, конфліктів, співпраці та співіснування, процесів прийняття зовнішньоекономічних рішень і т. ін. Сучасне моделювання є теоретичним методом аналізу міжнародних відносин, який припускає достатньо високий рівень абстракції та дає можливість симулювати конкретні явища і процеси МВ [51, с. 78].

Математичні підходи в аналізі міжнародних відносин використовують для вирішення тактичних (локальних) завдань і для аналізу стратегічних (глобальних) проблем. Математика є ефективним інструментом для побудови моделей міжнародних відносин різного ступеня складності та деталізованості. Переважно такі моделі використовують не для обчислення конкретних числових показників, а для виявлення закономірностей соціально-економічних і політичних ситуацій та процесів, зокрема політичної та економічної стабільності.

Засобами моделювання не можна вирішити всі завдання, що виникають при дослідженні МВ, оскільки на сучасному етапі розвитку не розроблено уніфікованої системи індикаторів, які б достовірно та повно відображали реальні процеси міжнародних відносин. Окрім того, при вивченні складних систем неможливо безпосередньо виміряти всі величини, які визначають їхні обсяги та характеристики (фактори). Нерідко є невідомими кількість та зміст цих факторів. Проте побудовані моделі дають можливість отримати нову інформацію про досліджуваний об'єкт.

1.1. Поняття про моделі та моделювання

Одним із основних методів отримання нової корисної інформації при дослідженні процесів та ситуацій МВ є моделювання. Це процес вивчення властивостей досліджуваного об'єкта шляхом побудови відповідної спрощеної моделі.

Об'єкт у загальнонауковому розумінні – це виділена частина навколишнього середовища, яку розглядають як єдине ціле. Об'єктами досліджень у МВ э міжнародні соціально-економічні та політичні ситуації, явища та процеси.

Модель – ідеальний чи матеріальний образ (замінник) досліджуваних об'єктів та зв'язків між ними.

Для опису та дослідження моделі використовують поняття та методи математики, у термінах якої поняття «модель» означає матеріальний або абстрактний об'єкт-замінник, що зберігає суттєві для конкретного дослідження, типові характеристики та властивості об'єкта-оригіналу.

Загальна структура моделі:

$$Y = F(X, P), \tag{1.1}$$

де X – множина вихідних (початкових) змінних, Y – множина результатних змінних системи (відгуків), P – множина параметрів, F – функція, функціонал, алгоритм або формальне представлення залежності результатних (залежних, ендогенних) змінних Y від факторних (незалежних, екзогенних) змінних X.

Метою моделювання у МВ є вивчення закономірностей та особливостей розвитку досліджуваного об'єкта, явища, системи, ситуації чи процесу для виявлення можливих ефективних дій щодо нього. Моделювання – один із способів розв'язання проблем реальності. Найчастіше цей метод застосовують у випадках, коли натурні експерименти надто витратні, небезпечні або неможливі. Моделювання передбачає нівелювання несуттєвих для дослідження ракурсів, властивостей і зв'язків об'єкта, подальше вивчення, аналіз та оптимізацію абстрактної моделі та відображення оптимального рішення в реальному світі [43, с. 11].

Якщо результати моделювання є достовірними, вважають, що побудована модель адекватно відображає процеси об'єктивної реальності. У такому випадку результати моделювання можуть стати основою для подальшого проєктування і прогнозування станів чи процесів об'єкта-оригіналу. Прийнятна адекватність моделі залежить від мети та ступеня деталізованості моделювання, використаних методів і прийнятих критеріїв.

Для побудови математичної моделі задачі (її формального опису у термінах математики у вигляді формул, рівнянь, нерівностей і т. ін.) необхідно виділити суттєві для конкретного дослідження характеристики об'єкта, який вивчають.

Модель має коректно відображати об'єкт реального світу, його властивості та зв'язки. Підтвердити адекватність обраної моделі можна лише за допомогою експериментального дослідження. Вибір математичної моделі залежить від типу, якості та повноти вихідних даних, допустимої для конкретного дослідження точності обчислень. Для підвищення якості результатів модель уточнюють, враховуючи нові особливості досліджуваного об'єкта.

При розробці математичної моделі необхідно виділити базові припущення, визначити вихідні дані та шукані результати і записати математичні співвідношення, які їх пов'язують.

Математичну модель будують на основі спрощень та ідеалізацій об'єкта. Вона лише наближено відображає реальні явища, а

вихідні дані – це неточні числа. Окрім того, метод розв'язування задачі є наближеним, а при обчисленнях використовують заокруглення. Ці причини зумовлюють похибки результатів. Важливим етапом моделювання є оцінювання точності розв'язку задачі (оцінювання похибки) та встановлення рівня значущості отриманих відгуків.

Моделювання застосовують у випадках, коли потрібно отримати уявлення про структуру об'єкта (ситуації, процесу); встановити, як ним управляти; отримати прогнозні оцінки прямих та опосередкованих наслідків реалізації визначених стратегій дій на об'єкт.

Моделювання – процес створення та використання моделей для розв'язування завдань. Міжнародні відносини, як і кожна наука, при описуванні фактів і явищ користується своєю мовою. При розробці моделей використовують мову, знаки та методи з галузі знань, в якій проводять формалізацію завдання за допомогою мови математики (мови цифр та математичних формул).

Класифікація моделей

Традиційно виділяють два основних способи побудови моделей:

► **фізичні** моделі (типу «сірого» ящика), при побудові яких на основі емпіричних досліджень виявляють закономірності функціонування системи, які потім відтворюють за допомогою моделі, досліджуючи на її основі поведінку системи; параметри моделі P в рівності (1.1) відповідають реальним процесам, які відбуваються в системі і мають фізичну інтерпретацію;

► **нефізичні** моделі (типу «чорного» ящика), при розробці яких без будь-якого фізичного обґрунтування припускають вид залежності F у рівності (1.1) та за даними емпіричних спостережень знаходять невідомі параметри P, непов'язані з фізикою реальних процесів, що відбуваються у системі, або ж їх взаємозв'язок є невідомим.

У науковій літературі моделі традиційно класифікують за різними ознаками (основами).

За способом подання всі моделі поділяють на **матеріальні** (предметні) та **інформаційні** (ідеальні) [43, с. 12].

Матеріальна модель відтворює геометричні та фізичні властивості об'єкта-оригіналу, має реальне втілення. Наприклад, глобус (модель земної кулі), макети ландшафтів та будівель, географічні карти і т. ін.

Інформаційна модель (у загальному розумінні) – інформація про характерні властивості та стани об'єкта, його взаємозв'язки із зовнішнім середовищем. Може бути представлена у словесно-мислительній (вербальній) чи знаковій формі.

Вербальна інформаційна модель (лат. «verbalis» – «усний») – модель, побудована на основі роздумів, логічних умовиводів. Може зберігатись у пам'яті людини в мислительній формі або передаватись за допомогою слів розмовної мови.

Знакова інформаційна модель – модель, в якій властивості та характеристики об'єкта описано за допомогою визначеної системи знаків: математичних та хімічних формул, азбуки Морзе, графів, діаграм, карт, схем і т. ін.

Знакове інформаційне моделювання – моделювання на основі побудови та дослідження знакових інформаційних моделей. Передбачає наявність правил, за якими можна використовувати знакові системи. Знакове інформаційне моделювання, зокрема математичне, є прогностичним, дає можливість виявляти нові корисні властивості досліджуваного об'єкта.

За способом реалізації знакові інформаційні моделі поділяють на **комп'ютерні** та **некомп'ютерні**.

Комп'ютерна модель реалізується засобами комп'ютерного середовища.

За фактором часу всі моделі поділяють на **статичні** та **динамічні**.

Матеріальні статичні моделі відображають просторові характеристики реального об'єкта та виражають особливості його функціонування. **Інформаційні моделі** описують структуру й характеристики об'єкта дослідження та характеризують процес його функціонування і динаміку змін, які визначаються набором його параметрів. До динамічних, зокрема, належать **алгоритмічні моделі**, що описують функціонування системи.

У більш вузькому розумінні **інформаційна модель** – опис властивостей і характеристик об'єкта однією з мов кодування (розмовною, науковою, графічною і т. ін.). Інформаційна модель містить не всю інформацію про об'єкт (процес) дослідження, а лише ту її частину, яка необхідна для вирішення виділених завдань.

Першою інформаційною моделлю реального об'єкта було слово. Інформаційні моделі описують розмовною та формальною мовами. Їх представляють у вигляді таблиць, ієрархічних та мережевих структур, діаграм.

Подання інформаційних моделей у вигляді таблиць реалізує відношення «один до одного». Представлення інформаційних моделей у вигляді різних схем з ієрархічною (підпорядкованою) структурою опису є відношенням «один до багатьох».

За результатною змінною виділяють **статичні** (*Y* не змінюється з часом) та **динамічні** (*Y* змінна у часі) **моделі**. Серед динамічних виділяють **неперервні** (*Y* – неперервна змінна) та **дискретні** (*Y* набуває дискретних значень).

З точки зору способу представлення залежності результатних змінних моделі від вихідних даних розрізняють **алгебраїчні, диференціальні, аналітичні, імітаційні моделі** і т. ін.

При дослідженні реальних процесів МВ використовують паралельно кілька різних методів моделювання та зіставляють отримані дані. Це збільшує достовірність результатів і прогнозних оцінок, зменшує ймовірність помилок.

Одним із найважливіших інструментів аналізу у МВ є математичне моделювання.

Математична модель – наближений (формальний) опис об'єкта за допомогою символів математики. Це функціональна залежність між характеристиками досліджуваного об'єкта (залежними та незалежними змінними), виражена у вигляді формули чи алгоритму.

За метою (цілями) моделювання виділяють **дескриптивні, оптимізаційні, багатокритеріальні, ігрові** та **імітаційні** моделі.

Дескриптивні моделі (лат. «descriptivus» – «опис») використовують для опису різних процесів, об'єктів. Наприклад, хімічні та фізичні процеси, які відбуваються всередині Землі.

Оптимізаційні моделі призначені для керування об'єктами та прийняття рішень. Наприклад, змінюючи ціни на товари, можна підібрати їх таким чином, щоб досягти максимального прибутку (оптимізувати процес).

Багатокритеріальні моделі будують для вирішення завдань з кількома цільовими функціями. Наприклад, знаючи ціни постачальників на сировину, її характеристики та обсяги, необхідні для виробництва, потрібно організувати постачання найякіснішої та найдешевшої сировини. Такі цілі не збігаються, тому при моделюванні буде декілька цільових функцій.

Ігрові моделі застосовують для опису та дослідження конфліктних ситуацій, в яких різні учасники мають інтереси, що не збігаються. Наприклад, конкуренція за ринок збуту.

Імітаційні моделі призначені для дослідження складних систем. Процес функціонування складної системи представляють за допомогою комп'ютерного алгоритму. Використовують у сучасній фізиці, зокрема для дослідження термоядерних процесів та прогнозування змін структури земної поверхні.

Властивості математичних моделей:

– адекватність (відповідність) досліджуваному об'єкту щодо виділених для дослідження властивостей (характеристик);

– не визначається однозначно досліджуваним об'єктом;

– лише наближено описує об'єкт досліджень;

– для одного і того самого об'єкта можна побудувати різні математичні моделі з різним ступенем наближення;

– універсальність (застосовність для досліджень у різних галузях людської діяльності).

Обчислювальний експеримент – методологія дослідження, заснована на вивченні математичної (інформаційної) моделі за допомогою логіко-математичних комп'ютерних алгоритмів.

При проведенні досліджень у сфері міжнародних відносин засобами моделювання вирішують два основних типи завдань:

► оцінні – виявлення специфіки об'єкта моделювання (наприклад, економічної ситуації, політичного процесу) та рівня інформаційної забезпеченості дослідження;

► операційні – виявлення характеру та форми моделі, які залежать від ступеня схожості з об'єктом моделювання (основні ознаки або узагальнюють, або максимально конкретизують).

Традиційно виділяють такі основні етапи моделювання:

– попередня орієнтація та аналіз системи, формулювання основних припущень та гіпотез, розроблення попередніх сценаріїв і нормативних установок;

– формалізація гіпотез (вербальна інтерпретація математичних виразів);

– відбір і формалізація необхідної інформації;

– дослідження моделі (перевірка на чутливість, адекватність, стійкість результатів);

– побудова альтернативних сценаріїв та модельні експерименти;

– якісний аналіз та інтерпретація результатів моделювання.

Останнім часом вчені-міжнародники дедалі частіше застосовують математичні методи аналізу даних при дослідженні різноманітних систем і процесів – природних, технічних, екологічних, економічних, соціальних та політичних. Математичне моделювання дає змогу розширити можливості традиційного аналізу, підвищити якість отриманих результатів та точність прогнозних оцінок. Використання математичних методів моделювання і прогнозування

виправдане у практиці досліджень міжнародних ситуацій та процесів, оскільки за їх допомогою можна виявити раніше не виділені взаємозв'язки між суб'єктами міжнародних відносин. Моделювання є суттєво важливим при виявленні прихованих ресурсів і можливостей взаємодії на міжнародній арені та необхідним для уточнення альтернатив ймовірних сценаріїв розвитку умов та способів дії. Проте часто доводиться враховувати багато суб'єктивних моментів та об'єктів, які не підлягають структуруванню. У деяких випадках складно провести формалізацію або інформація є неповною.

1.2. Завдання та методи моделювання

Залежно від цілі та завдання дослідження для однотипних об'єктів моделювання можна побудувати різні моделі.

Серед багатьох завдань моделювання виділяють такі основні задачі: моделювання, управління, ідентифікація, оптимізація, прогнозування [66, с. 11].

Задача моделювання (пряма задача) полягає у знаходженні значень результатних змінних Y при відомих значеннях вихідних змінних X, відомій моделі F та визначених параметрах P рівності (1.1).

Задача управління (зворотна задача) полягає у знаходженні таких значень вихідних змінних X, які забезпечують задані значення результатних змінних Y при відомій моделі F та фіксованих значеннях параметрів P.

У **задачі ідентифікації** відомими є множина вихідних змінних X, множина результатних змінних Y та множина моделей F. Потрібно визначити єдину модель f з множини запропонованих моделей F та її параметри P, які забезпечують при вихідних значення X відповідні результатні значення Y.

У формулюванні **задачі оптимізації** відомими є модель F, множина можливих вихідних значень X та критерій оптимізації K. Потрібно знайти значення вихідних змінних X, параметрів P та результатних змінних Y, які задовольняють заданий критерій K.

У **задачі прогнозування** відомими є вихідні та результатні значення моделі X_t, Y_t до моменту часу t. На заданому інтервалі прогнозування T потрібно визначити модель F та її параметри P, які забезпечують найкращий прогноз Y_{t+T}.

Серед великої кількості відомих методів моделювання виділяють аналітичне, математичне та імітаційне моделювання [66, с. 12].

Аналітичне моделювання передбачає представлення залежності F результатних змінних Y від її вихідних змінних X в аналітичному вигляді (опис за допомогою відомих аналітичних функцій). Перевагою цього методу моделювання є можливість отримати залежність $Y = f(X)$ в явному вигляді та застосувати до неї методи класичного математичного аналізу. Якщо можливо побудувати аналітичну модель системи, завжди надають перевагу цьому методу моделювання.

Знаходження залежності $Y = f(X)$ може виявитись настільки складним, що доводиться застосовувати спеціальне програмне забезпечення, а для окремих систем навіть відмовлятись від пошуку абстрактної залежності $Y = f(X)$ і задовольнятись наближеним розв'язком, який знаходять чисельними методами.

Окремі системи настільки складні, що, незважаючи на те, що їх функціонування можна описати анилітичними функціями, знаходження залежності $Y = f(X)$ в явному вигляді є неможливим. Наприклад, усі задачі математичного програмування мають достатньо простий аналітичний опис, але розв'язок може бути знайдений тільки в результаті виконання визначеної кількості кроків. Відомим є алгоритм відшукання точного розв'язку задачі, але розв'язок не може бути записаний в аналітичній формі. Такий метод моделювання називають **математичним моделюванням**. Алгоритм F знаходження точного розв'язку задачі може бути реалізований за допомогою спеціального програмного забезпечення або чисельних методів.

Підходи до аналітичного моделювання [68, с. 14]

Балансовий підхід припускає, що приріст змінної, яка описує процес, дорівнює різниці функції, що забезпечує збільшення кількості (функція приросту), і функції, що забезпечує зменшення кількості (функція смертності). На проміжку часу Δt

$$x(t + \Delta t) - x(t) = B(t, x(t), \Delta t) - D(t, x(t), \Delta t),$$

де $B(t, x(t), \Delta t)$ – збільшення змінної, $D(t, x(t), \Delta t)$ – зменшення змінної.

Одна з ефективних математичних моделей, що використовує цей підхід, – це модель Леонтьєва, яка базується на понятті «чиста галузь» [23].

Метод аналогій – застосування аналогій до раніше вивчених явищ.

Ієрархічний підхід – тільки у виняткових випадках вдається повністю побудувати модель (через багатофакторність процесу, складність залежностей і кількість зв'язків).

Підхід від простого до складного: будують ланцюг (ієрархію) моделей, що поступово ускладнюються, кожна з яких вміщає попередню як частковий випадок.

Лінійні та нелінійні моделі. Моделювання з використанням лінійних систем здійснюють з таких причин:

► при допустимих обмеженнях (малі часові проміжки) нелінійні процеси завжди можна апроксимувати лінійними (криву можна замінити відрізками прямих);

► точні розв'язки лінійних систем можна визначити в аналітичній формі;

► для лінійних систем існують ефективні методи оцінювання збурень (відхилень від розв'язку);

► розроблено високоякісні обчислювальні методи розв'язання лінійних систем.

Аналітичні методи розв'язання нелінійних систем є швидше винятком, ніж правилом. Навіть якщо є можливість отримати

розв'язок нелінійної системи в аналітичному вигляді, виникають складнощі при її якісному дослідженні.

Основною властивістю лінійних систем є принцип суперпозиції – принцип незалежного накладання дій або процесів, при якому результуючий ефект еквівалентний об'єднанню (сумі) ефектів, що зумовлені окремо кожним з них.

Нехай $f_1(t) = g_1(t), f_2(t) = g_2(t), \ldots, f_n(t) = g_n(t)$, тоді

$$f_1(t) + f_2(t) + \ldots + f_n(t) = g_1(t) + g_2(t) + \ldots + g_n(t); \; k f_i(t) = k g_i(t).$$

Оскільки на площині цей закон є рівнянням прямої лінії, такі системи називають **лінійними**.

Лінійні динамічні системи описують системами лінійних звичайних диференціальних рівнянь, рівнянь у частинних похідних і лінійними різницевими та інтегральними рівняннями. Швидкість зміни змінної пропорційна до її значення в поточний момент часу. Це основне припущення при побудові лінійних динамічних систем. Однією з перших моделей такого типу є модель зміни чисельності населення Мальтуса – найпростіша модель експоненціального зростання чисельність популяції за умови сталого приросту (необмежених ресурсів) [65, с. 24].

Основне твердження: швидкість зміни кількості населення пропорційна до чисельності, помноженої на коефіцієнт *k(t)*, що дорівнює різниці коефіцієнтів народжуваності $\alpha(t)$ та смертності $\beta(t)$: $\frac{dN(t)}{dt} = k(t)N(t)$, $k(t) = \alpha(t) - \beta(t)$.

Напрями модернізації лінійних моделей:

► припущення, що в околі стаціонарного (усталеного) процесу на динаміку системи діють невеликі збурювальні впливи – квазілінійні моделі $N(t) = A(t)N(t) + \varepsilon F(t, N(t))$;

► припущення, що коефіцієнт при змінній залежить і від $N(t)$ (немає обмеження області значень змінних, динаміка лінійної моделі наближена до реального процесу).

Найпростіша модель такого типу – система Лотки–Вольтерра (модель міжвидової конкуренції) [65, с. 68]:

$$\frac{dx}{dt} = (\alpha - \beta y)x\,,$$

$$\frac{dy}{dt} = (-\gamma + \delta x)y\ \ ,$$

де x – кількість жертв, y – кількість хижаків, t – час,
$\alpha, \beta, \gamma, \delta,$ – коефіцієнти, що відображають взаємодію між видами.

Існують системи, опис яких не підлягає представленню за допомогою аналітичних функцій, але процес їх функціонування може бути формалізований алгоритмом імітації.

Імітація – відтворення за допомогою комп'ютерної програми процесу функціонування складної системи в часі. У результаті багатократного використання імітаційної моделі отримують інформацію про властивості реальної системи. Такий метод моделювання називають **імітаційним моделюванням**.

Системний підхід до побудови моделей

Системний підхід до дослідження систем полягає у вивченні їх функціонування загалом, без врахування особливостей окремих складових. Цей підхід базується на використанні твердження, що навіть найкраще функціонування окремих підсистем та елементів системи не гарантує найкращого функціонування всієї системи, оскільки завжди існує взаємодія між її складовими.

Системний підхід передбачає формалізацію досліджуваної системи, що складається з опису таких складових:

– набору вихідних змінних системи та їх основних характеристик;

– набору результатних змінних системи;

– меж системи щодо її зовнішнього середовища;

– елементів системи та їх основних властивостей;

– зв'язків між елементами системи.

Чинники, які унеможливлюють покращення функціонування системи [70, с. 57]:

► змінюваність: жодна реальна система не є статичною впродовж тривалого проміжку часу (побудована модель має обмежений термін використання);

► наявність навколишнього середовища: в моделі має бути передбачений вплив зовнішнього середовища, який часто має випадковий характер;

► неінтуїтивна поведінка: наслідок може проявлятись пізніше, ніж причина;

► тенденція до погіршення характеристик функціонування: зношення окремих складових системи призводить до погіршення їх функціонування, що спричиняє непередбачувані наслідки;

► взаємозалежність: усі складові системи взаємопов'язані, погіршення характеристик функціонування однієї з них спричиняє зміну характеристик функціонування інших частин системи;

► організація: існує ієрархія підсистем, що підпорядковується цільовому призначенню системи.

Будь-яка система об'єктивна і водночас суб'єктивна з точки зору вибору меж системи та її елементів. Одному й тому самому об'єктивному процесу можна поставити у відповідність різні системи і тільки досвід дослідника, його інтуїція та здатність творчо мислити дають можливість здійснити вибір між багатьма варіантами і виконати дослідження системи оптимальним способом.

Моделі, побудовані із застосуванням системного підходу, називають **системними моделями**.

Опис системи разом із визначенням цілі та завдання дослідження становлять сутність **концептуальної** (лат. «conceptio» – «сприйняття») **моделі** системи.

Умовно виділяють такі етапи створення концептуальної моделі системи:

– визначення цілі дослідження системи (орієнтація);

– вибір рівня деталізації системи (стратифікація);

– визначення елементів системи (деталізація);

– визначення впливу зовнішнього середовища (локалізація);

– визначення зв'язків між елементами системи та із зовнішнім середовищем (структуризація).

Найбільш відомі методи, які застосовують для моделювання МВ

Узагальнення – сукупність дій зі зведення конкретних одиничних фактів в єдине ціле з метою виявлення загальних ознак і закономірностей, властивих досліджуваному явищу чи процесу [7, с. 20].

Рейтинговий метод – встановлення ступеня популярності особи (організації, угрупування), її діяльності, програм, планів, політики на конкретний момент часу; встановлення місця актора політичної арени серед собі подібних, визначене шляхом голосування, соціологічних опитувань, анкетування.

Вивчення документів – аналіз офіційних документів державних, політичних, релігійних та громадських організацій, що стосуються зовнішньо-політичної діяльності. Передбачає перевірку їх автентичності, встановлення достовірності відображених фактів, визначення основних понять та головної ідеї, отримання висновків про відображені у документі події.

Розпізнавання образів (об'єктів, сигналів, ситуацій, явищ, подій або процесів) – ідентифікація об'єкта чи визначення параметрів за його зображенням (оптичне розпізнавання) або аудіозаписом (акустичне розпізнавання) і т. ін. [60, с. 19].

Аналіз трендів – виділення трендів у наборах документів на визначений період. Тренд можна використати, наприклад, для виявлення змін інтересів споживачів.

Метод асоціацій та аналогій – ідентифікація асоціативних зв'язків між ключовими поняттями у множині документів.

Порівняння – зіставлення спільних та відмінних рис окремих явищ і процесів, що відбуваються на міжнародній арені. У міжнародних відносинах порівнюють показники розвитку за різні часові періоди, силовий потенціал держави (території, населення, озброєн-

ня), конкретні політичні, соціальні чи економічні факти, склад їхніх учасників (суспільні структури чи інституції, рухи, політичні лідери), суспільні настрої та інтереси націй (держав, спільнот) і т. ін.

Експеримент – верифікація та перевірка достовірності побудованих гіпотез. Експеримент у міжнародних відносинах є лише аналітичним представленням реальних фактів та дослідженням можливого розвитку подій. Оскільки багато гіпотез неможливо перевірити, дослідники зіставляють їх з обґрунтованими історичними фактами.

Експеримент у МВ – створення штучної ситуації з метою перевірки теоретичних гіпотез і висновків (імітаційні ігри).

Контент-аналіз – виявлення та аналіз специфічних характеристик інформаційних масивів з метою встановлення прихованого змісту чи сенсу, на якому наголошує автор.

Вперше контент-аналіз застосував американський учений Г. Лассуел для дослідження пропагандистської спрямованості політичних текстів [51, с. 53]. У загальному випадку контент-аналіз – систематизоване вивчення змісту письмового чи усного тексту за допомогою виділення у ньому найчастіше повторюваних лексем з метою виявити або оцінити відображені у документі соціальні факти та тенденції.

Контент-аналіз вивчає документи в їх соціальному контексті та передбачає виокремлення первинних одиниць аналізу (лексем, змістових блоків) та одиниць виміру (слів, розділів, файлів). При дослідженні МВ одиницею аналізу є конкретна зовнішньоекономічна парадигма, подія міжнародного життя, ставлення до неї (позитивне чи негативне) акторів міжнародної арени.

Для обчислення результатів контент-аналізу використовують «коефіцієнт Яніса» – співвідношення між сприятливими та несприятливими (відносно конкретної позиції) оцінками, судженнями, аргументами [30, с. 68]:

$$C = \frac{t^2 - t \times f}{k \times n},\ t > f\ ,$$

$$C = \frac{t \times f - t^2}{k \times n}, \quad t < f,$$

де t – кількість сприятливих оцінок, f – кількість несприятливих оцінок, k – обсяг одиниць аналізу, n – загальний обсяг аналізованого тексту.

Івент-аналіз – аналіз публічної інформації з метою визначити та систематизувати дії у міжнародних відносинах за схемою:

– суб'єкт – ініціатор дії;

– сюжет дії;

– суб'єкт-мішень (об'єкт дії);

– дата дії.

При івент-аналізі міжнародні події систематизують, будують матричні таблиці, ранжують та досліджують засобами статистичних прикладних програм.

Суть методу аналізу подій полягає у дослідженні міжнародних економічних чи політичних процесів та визначенні основних тенденцій динаміки подій в окремих країнах і на міжнародній арені загалом [8, с. 112].

Когнітивне картування – вивчення особливостей мислення (понять та категорій) осіб, які приймають рішення у сфері зовнішньоекономічної політики.

Основним поняттям методу є поняття «схема» («карта») – графічне відображення позиції політичного діяча у вигляді блок-схеми або порівняльної таблиці, на основі чого можна зробити висновок про можливі взаємодії сторін [43, с. 129].

Когнітивне картування призначене для визначення основних понять, якими оперує політичний діяч, та знаходження причинно-наслідкових зв'язків між цими поняттями. Рішення приймають на основі індивідуального сприйняття подій особою, що ухвалює рішення. Цей метод не дає реального уявлення про механізми та закономірності процесів у сфері міжнародних відносин, тому є недосконалим. Окрім того, політичні діячі у публічних промовах не використовують лексем, що свідчать про їхні справжні наміри.

Етапи проведення когнітивного картування у МВ:

– виділення основних понять, якими оперує політичний діяч;

– відображення причинно-наслідкових зв'язків між поняттями;

– оцінювання значущості та щільності цих зв'язків.

На когнітивній карті зображають точки (основні теми) змісту та причинно-наслідкові зв'язки між ними. Через суб'єктивність людського мислення метод когнітивного картування можна ефективно використовувати лише у комбінації з іншими методиками.

Аналіз за допомогою індикаторів – виокремлення найважливіших факторів (індикаторів) поведінки учасників МВ. Наприклад, спосіб представництва делегації, економічні договори та угоди, міждержавні візити [51, с. 74].

Аналіз кореляцій – встановлення факту наявності або відсутності залежності між двома параметрами.

Факторний аналіз застосовують за необхідності обмеження кількості індикаторів (змінних). Індикатори, між якими виявлено сильну кореляцію (взаємозв'язок), вказують на одну і ту саму причину.

Метод індикаторів використовують для оцінювання характеру, структури та змісту систем міждержавних взаємодій. Він використовує статистичні прийоми збирання та опрацювання інформації для дослідження МВ.

Статистичний метод дослідження МВ – оцінювання кількісних параметрів масових суспільних явищ у причинно-наслідковому зв'язку з їхнім якісним змістом [50, с. 37]. Статистичне дослідження суспільних явищ (політична арифметика) – визначення кількісних параметрів явищ і процесів, їх аналіз та теоретичне узагальнення.

Системний метод дослідження міжнародних економічних відносин – розгляд об'єкта дослідження в його єдності та цілісності.

Прогнозні методи – наукові передбачення майбутніх явищ та процесів у міжнародних економічних відносинах.

Інтерполяція (метод статистичного прогнозування) – знаходження показника (приблизний розрахунок) у середині ряду на основі закономірностей розвитку явища за досліджуваний період.

Точність такого розрахунку залежить від стабільності показників динаміки – абсолютних приростів, темпів зростання.

Екстраполяція подій і явищ минулого на майбутній період – знаходження невідомих рівнів наприкінці чи на початку динамічного ряду (за межами наявних фактичних даних) відповідно до обраних індикаторів за визначеними часовими інтервалами (місяць, квартал рік і т. ін.). За середнім значенням індикатора будують хронологічний графік. Екстраполяцію проводять тільки для невеликих часових проміжків.

Екстраполяція на короткотривалий період можлива також на основі середнього абсолютного приросту, середнього темпу зростання та приросту.

Метод «мозкового штурму» – оперативна колективна генерація нестандартних ідей незалежних фахівців з різних предметних галузей, що полягає в інтенсивному аналізі проблемної ситуації, продукуванні думок, їх критиці та формулюванні контрідей, угрупуванні, відборі та оцінюванні висловлених ідей.

Регресійний аналіз – встановлення виду залежності (рівняння) між відгуком (залежною змінною) та факторами (незалежними змінними) [19, с. 146].

Аналіз тенденцій – виявлення закономірності розвитку явища в часі.

Спектральний аналіз – визначення основних коливань, їх частоти та фаз у складних динамічних структурах.

Основою методу є виділення структури коливного процесу (наприклад, популярність уряду, торгової марки) і побудова графіка коливань. Для цього збирають хронологічні дані, будують рівняння коливань і виділяють цикли, на основі яких отримують графіки [63, с. 20].

Прогальний аналіз – формування судження про відсутні факти, фактори, причини.

Побудова сценаріїв – визначення ймовірного розвитку подій МВ на основі аналізу конкретної ситуації. Це засіб прогнозування, за допомогою якого передбачають розвиток політичних та соціально-економічних подій [6, с. 88].

У сценарії (словесному описі прогнозованої ситуації) виокремлюють зв'язки між подіями та критичні точки, в яких фактори середовища можуть мати непропорційний вплив, визначають часовий інтервал, виділяють події, словесно описують їх сенс та кількісні характеристики за ранжованою шкалою. Сценарій у розгорнутій формі відображає можливі варіанти розвитку подій для їх подальшого дослідження та вибору найбільш реальних і оптимальних можливостей. Вироблення сценаріїв передбачає аналіз фактів і процесів МВ, які можна не враховувати під час абстрактних міркувань.

Метод експертного оцінювання – виявлення думок експертів у галузі МВ щодо конкретної проблеми. Проводять у формі опитування або ділової гри (групової діяльності експертів) за встановленими правилами у рамках розробленого сценарію. Це спосіб прогнозування та оцінювання майбутніх результатів дій на основі прогнозів фахівців [41, с. 417].

Застосування експертного оцінювання дає можливість отримати такі види інформації:

– інформація про окремі причинно-наслідкові зв'язки в конкретних умовах місця і часу;

– інформація про типові взаємозв'язки досліджуваних економічних явищ і процесів.

Основні завдання, які вирішують на основі отриманої від експертів інформації:

► ранжування (впорядкування, розміщення в порядку зростання чи спадання) факторів та їхніх показників за значущістю в динаміці досліджуваного явища чи процесу;

► ранжування міжнародних організацій, транснаціональних корпорацій чи їх структурних виробничих підрозділів за рейтингом на основі сукупності різних показників, що характеризують результати міжнародної економічної діяльності чи окремих її видів (фінансовий стан, рентабельність, платоспроможність, наявність міжнародних контрактів і т. ін.);

► попереднє оцінювання виконання плану за конкретним показником.

Дельфійський метод – систематичне та контрольоване обговорення конкретної проблеми кількома анонімними експертами у сфері МВ [51, с. 71].

Опитування експертів проводять у кілька турів, після кожного з яких їхні результати систематизують та узагальнюють. Це дає можливість визначити суттєві розходження результатів експертних оцінювань досліджуваної проблеми.

Аналіз причин розходжень в оцінках експертів дає можливість оцінити досліджувану проблему з різних позицій, виявити невідомі раніше аспекти та найбільш вагомі параметри, вибрати найвірогідніші сценарії розвитку подій (сценарії, які підтримала більшість експертів) чи найменш вірогідні (тому й становлять інтерес).

Теорія ігор – прикладна концепція, яка займається побудовою математичних моделей для дослідження конфліктних ситуацій [71].

Теорія катастроф спеціалізується на вивченні так званих хаотичних режимів функціонування складних систем та пов'язаних з ними перехідних явищ (криз та катастроф) [3].

Data mining (інтелектуальний аналіз даних) – міждисциплінарний розділ комп'ютерних наук, призначений для виявлення шаблонів у великих наборах даних із використанням методів штучного інтелекту, машинного навчання, статистики та баз даних [5].

Big data (великі дані) – аналіз даних у прогностичній аналітиці, обробці даних, в управлінні відносинами з клієнтами [79].

Моделювання у суспільних науках – метод дослідження, пов'язаний з абстрагованим відображенням реальних явищ та подій міжнародних економічних відносин за допомогою ідеалізованих об'єктів та зв'язків між ними.

На сьогодні розроблено низку моделей міжнародних систем, соціально-економічних процесів, конфліктів, співпраці та співіснування, процесів прийняття зовнішньоекономічних рішень і т. ін. Сучасне моделювання є теоретичним методом вивчення міжнародних відносин на достатньо високому рівні абстракції, який дає можливість симулювати конкретні явища і процеси МВ [51, с. 78].

Моделювання як спосіб пізнання використовують з давніх часів. Але з появою комп'ютера моделювання систем збагатилось появою принципово нових методів, зокрема це імітаційне та еволюційне моделювання, методи групового врахування аргументів. Моделі та методи моделювання використовують при створенні систем автоматизованого проєктування, систем прийняття рішень, систем автоматизованого керування, систем штучного інтелекту.

На сьогодні більшість рішень для проведення прикладних досліджень МВ реалізовано у пакетах прикладних програм. Зокрема, це статистичний пакет для проведення прикладних досліджень у соціальних науках «SPSS», система автоматизованого проєктування, орієнтована на підготовку інтерактивних документів з обчисленнями і візуальним супроводженням «Mathcad», одна з найбільш популярних статистичних програм для пошуку закономірностей, прогнозування, класифікації та візуалізації даних «Statistica», пакет для розв'язання математичних завдань та візуалізації графіків «Microsoft Mathematics», аналітична платформа «Deductor», програма для логічної обробки тексту «Text Miner» та багато інших.

ПИТАННЯ ДЛЯ САМОПІДГОТОВКИ ТА САМОКОНТРОЛЮ

Теоретичні запитання

1. Поняття «модель».
2. Загальна структура моделі.
3. Класифікація моделей.
4. Математичні моделі та їх види.
5. Властивості математичних моделей.
6. Загальне поняття про моделювання.
7. Завдання та методи моделювання.
8. Підходи до аналітичного моделювання.
9. Математичне моделюваня.
10. Лінійні та нелінійні моделі.
11. Поняття «імітаційне моделювання».
12. Системний підхід до побудови моделей.
13. Базові методики моделювання МВ.

РОЗДІЛ 2

СТАТИСТИЧНІ МЕТОДИ ДОСЛІДЖЕННЯ ЗВ'ЯЗКІВ МІЖ ФАКТОРАМИ

При моделюванні станів та процесів МВ для виявлення нових фактів чи підтвердження відомих закономірностей явища, що розглядається, у багатьох випадках використовують методи перевірки статистичних гіпотез на основі даних вибіркового спостереження.

Гіпотеза – наукове припущення, висунуте для пояснення конкретного явища, яке потрібно теоретично обґрунтувати та перевірити емпіричними даними. Гіпотези мають імовірнісний характер.

Статистична гіпотеза – припущення щодо виду або параметрів закону статистичного розподілу, якому підпорядковується генеральна сукупність. Її формулюють на заданому рівні статистичної значущості. Оцінки, отримані для емпіричних вибірок, поширюють на всю генеральну сукупність [53, с. 278].

Необхідність перевірки статистичних гіпотез часто виникає при моделюванні у різних сферах економіки, зокрема у МВ. Це завдання передбачає зіставлення й оцінювання властивостей випадкових явищ.

Перевірка статистичних гіпотез має велике практичне значення – за результатами вибіркового емпіричного спостереження роблять імовірнісні висновки щодо досліджуваної сукупності. Пе-

ревірку гіпотез проводять з використанням методів математичної статистики шляхом обчислення статистичних оцінок, на основі яких роблять з певною ймовірністю висновок про об'єкт дослідження [56, с. 313].

У загальному випадку завдання перевірки статистичних гіпотез зводиться до встановлення значущості обчислених оцінок для характеристики законів розподілу або окремих його параметрів.

2.1. Модельні розподіли

Міжнародні соціально-економічні та політичні процеси представляють у вигляді ряду послідовних значень досліджуваного показника, розташованих у хронологічному чи ранжованому порядку. Для встановлення основних властивостей та закономірностей МВ використовують ряди розподілу (табл. 2.1).

Таблиця 2.1

Країни ЄС, які прийняли найбільше біженців за 2017 р.

Країна	Німеччина	Франція	Швеція	Італія	Австрія	Англія
К-сть мігрантів	1256828	354880	317477	216027	165446	151621

Ряд розподілу – впорядкований розподіл одиниць досліджуваної сукупності на групи за однією варіативною ознакою.

Основні характеристики ряду розподілу:

– варіанта (окреме значення варіаційної ознаки);

– частота появи значень випадкової величини.

Ряди розподілу використовують при вивченні складу та структури сукупності, основних закономірностей розподілу одиниць за досліджуваною ознакою, при обчисленні середніх величин, показників варіації та взаємозв'язку і т. ін.

Описові статистики – це різні обчислювані показники, що характеризують розподіл значень змінної. Ці показники умовно можна розбити на декілька груп. Перша група – міри центральної тенденції, навколо яких «групують» дані: середнє значення, меді-

ана і мода. Друга група характеризує мінливість значень змінної відносно середнього: стандартне відхилення і дисперсія. Діапазон мінливості характеризується мінімумом, максимумом і розмахом. Асиметрія і ексцес представляють міру відхилення форми розподілу від нормального вигляду. Крім того, існують величини, що виражають похибки деяких статистик: стандартна помилка середнього, стандартна помилка асиметрії і стандартна помилка ексцесу.

Міри центральної тенденції

Існує три основні міри центральної тенденції розподілу: середнє значення, мода і медіана.

Середня величина – узагальнюючий показник, який характеризує типовий рівень варіативної ознаки в розрахунку на одиницю однорідної сукупності та може не збігатися з жодним з індивідуальних значень ознаки. Виражається в одиницях виміру ознаки. У середній узагальнюються типові риси одиниць сукупності та взаємно компенсуються індивідуальні відмінності елементів, зумовлені дією випадкових факторів [55, с. 217].

Усі види середніх поділяють на два класи:

► степеневі середні – середнє арифметичне, середнє гармонійне, середнє геометричне, середнє квадратичне;

► структурні (позиційні) середні – мода і медіана.

Середнє арифметичне застосовують при дослідженні закономірностей розподілу у випадку, коли обсяг ознаки для всієї сукупності є сумою індивідуальних значень її окремих елементів.

Середнє арифметичне просте для незгрупованих даних:

$$x_{сер} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n},$$

де x_i – індивідуальні значення ознаки, n – їх кількість.

Для розподілу [3 5 7 5 6 8 9] середнє значення дорівнює:

$$x_{сер} = \frac{3+5+7+5+6+8+9}{7} = 6{,}14.$$

Середнє арифметичне зважене розраховують для групованих даних за відомої частоти появи однакових значень (варіант) ознаки у сукупності:

$$x_{сер_{зв}} = \frac{x_1 f_1 + x_2 f_2 + \ldots + x_n f_n}{f_1 + f_2 + \ldots + f_n} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i},$$

де x_i – значення варіант сукупності, f_i – частота повторень значень відповідного варіанта.

Для розподілу [3 5 7 5 6 8 9] середнє значення дорівнює .

$$x_{сер} = \frac{3 + 5 \cdot 2 + 7 + 6 + 8 + 9}{7} = 6{,}14.$$

Частоту f називають **вагою** варіантів, множення x_i на f_i – **зважуванням** (відображення рівновагомості окремих варіантів).

Загальною середньою називають розраховану середню структурованої сукупності. Для її обчислення варантами обирають групові середні та їх ваги (групові частоти).

Середнє гармонійне використовують для сукупностей додатних чисел, які інтерпретують відповідно до їхнього темпу зростання, що передбачає обчислення добутку:

$$h = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}.$$

Для розподілу [3 5 7 5 6 8 9] середнє гармонійне дорівнює:

$$h = \frac{7}{\frac{1}{3} + \frac{1}{5} + \frac{1}{7} + \frac{1}{5} + \frac{1}{7} + \frac{1}{8} + \frac{1}{9}} = \frac{7}{0{,}33 + 0{,}2 + 0{,}14 + 0{,}2 + 0{,}17 + 0{,}13 + 0{,}1} = \frac{7}{1{,}3} = 5{,}5.$$

Середнє геометричне використовують для аналізу рядів динаміки:

$$g = \sqrt[n]{\prod_{i=1}^{n} x_i} = \prod_{i=1}^{n} \sqrt[n]{x_i}.$$

Для розподілу [3 5 7 5 6 8 9] середнє геометричне дорівнює $g = \sqrt[7]{3 \cdot 5 \cdot 7 \cdot 5 \cdot 6 \cdot 8 \cdot 9} = \sqrt[7]{226800} = 5{,}8$.

Структурні середні (моду та медіану) використовують в умовах неповноти вихідних даних (наприклад, захист комерційної таємниці) та для детальнішого вивчення властивостей розподілу і характеристик структури досліджуваної сукупності.

Мода (Мо) – значення ознаки, що найчастіше зустрічається в сукупності, тобто має найбільшу вагу (частоту, частку). Мода має практичне значення у соціологічних, маркетингових та інших дослідженнях, оскільки вказує, яке значення ознаки є найбільш імовірним, масовим. наприклад, найбільш поширена думка, рейтинг популярності, споживчий попит і т. ін.

Сукупності, в яких декілька варіантів мають однакову найбільшу частоту, називають **багатомодальними**. Такі сукупності є неоднорідними за досліджуваною ознакою.

У розподілі [3 5 7 5 6 8 9] мода дорівнює 5, оскільки число 5 зустрічається у ньому двічі.

Медіана (Ме) – значення варіативної ознаки, яке поділяє ранжований ряд даних на дві рівні частини: 50% одиниць досліджуваної сукупності мають значення ознаки менше, ніж медіана, а 50% – більше. Це серединне (центральне) значення сукупності.

Для розподілу [3 5 7 5 6 8 9] медіана становить 6, оскільки значення, що дорівнює 6, знаходиться в центрі послідовності [3 5 5 6 7 8 9].

Практичне використання медіани базується на її властивості: сума відхилень усіх варіантів від медіани є найменшою величиною. Цю величину використовують, наприклад, при плануванні розташування філій ТНК, у логістиці та маркетингу і т. ін.

Міри мінливості

Виділяють дві величини, що характеризують мінливість, або розкид, значень розподілу відносно середнього.

Дисперсія (показник середнього квадрата відхилень, міра розсіювання варіантів) дорівнює сумі квадратів відхилень кожного значення від середнього, поділеній на $n - 1$, де n – кількість значень у розподілі:

$$\sigma = \frac{\sum_{i-1}^{n} \left(x_i - x_{сер}\right)^2 n_i}{n}.$$

Для розподілу [3 5 7 5 6 8 9] дисперсія дорівнює

$$\sigma = \frac{\sum_{i=1}^{7}(3-6{,}14)^2 + (5-6{,}14)^2 \cdot 2 + (7-6{,}14)^2 + (6-6{,}14)^2 + (8-6{,}14)^2 + (9-6{,}14)^2}{7} = 3{,}55$$

Середнє квадратичне (стандартне) відхилення показує, на скільки в середньому відхиляються значення ознаки від середнього рівня; застосовують при розрахунках абсолютних і відносних показників варіації ознаки:

$$s = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - x_{сер}\right)^2 n_i}{n}}.$$

Для розподілу [3 5 7 5 6 8 9] середнє квадратичне дорівнює

$$s = \sqrt{\frac{\sum_{i=1}^{7}(3-6{,}14)^2 + (5-6{,}14)^2 \cdot 2 + (7-6{,}14)^2 + (6-6{,}14)^2 + (8-6{,}14)^2 + (9-6{,}14)^2}{7}} = 1{,}88$$

Характеристики діапазону розподілу

Додатковими мірами мінливості є 4 простих характеристики, що відображають межі розподілу та його розмах.

Мінімум дорівнює найменшому зі значень розподілу. Для розподілу [3 5 7 5 6 8 9] мінімум дорівнює 3.

Максимум дорівнює найбільшому зі значень розподілу. Для розподілу [3 5 7 5 6 8 9] максимум дорівнює 9.

Розмах становить різницю між максимумом і мінімумом розподілу. Для розподілу [3 5 7 5 6 8 9] розмах дорівнює 9 – 3 = 6.

Сума дорівнює сумі всіх значень розподілу. Для розподілу [3 5 7 5 6 8 9] сума дорівнює 3 + 5 + 7 + 5 + 6 + 8 + 9 = 43.

Показники варіації надають можливість оцінити залежність зміни ознаки від суттєвих факторів. Вони формують уявлення про інтенсивність структурних зрушень, щільність взаємозв'язків, точність результатів вибіркового спостереження.

Співвідношення варіант і частот формує закономірність розподілу: частоти змінюються відповідно до динаміки варіативної ознаки. Ці закономірності змін частот називають закономірностя-

ми розподілу. Вони яскраво проявляються тільки при масових спостереженнях [20, с. 236].

Динамічний ряд – сукупність ранжованих даних спостережень досліджуваного показника. Аналіз рядів динаміки є одним з найбільш ефективних засобів виявлення тенденцій та закономірностей розвитку соціально-економічних явищ.

За характером групувальної ознаки виділяють атрибутивні та варіаційні (кількісні) ряди розподілу.

Атрибутивні ряди розподілу формують за якісною ознакою. Варіанти розташовують у їх логічній послідовності.

Частковим випадком атрибутивних рядів розподілу є альтернативні ряди, елементи яких набувають лише двох взаємовиключних значень. Наприклад, розподіл країн за членством у СОТ, ЄС, участю в міждержавному рейтингу (так/ні) і т. ін.

Варіаційний ряд отримують у результаті групування та ранжування одиниць сукупності за кількісною ознакою. Виділяють дискретні та інтервальні (неперервні) варіаційні ряди розподілу.

У дискретних варіаційних рядах розподілу варіанти набувають значень лише цілих чисел (табл. 2.2).

Таблиця 2.2

Членство країн у ЄС за кількістю років

К-сть повних років	61	45	37	31	23	14	11	5
К-сть країн	5	3	1	2	4	10	2	1

Інтервальний варіаційний ряд складається з варіант, які є інтервалами. Наприклад, розмір надбавки за вислугу років (табл. 2.3).

Таблиця 2.3

Розмір щомісячної надбавки до посадового окладу

Стаж роботи	від 3 до 10 років	від 10 до 20 років	понад 20 років
Розмір надбавки, у %	10	20	30

Завдання дослідження закономірностей розподілу:

– визначення типового рівня ознаки (центру тяжіння розподілу);

– вимірювання варіації ознаки – ступеня групування індивідуальних значень ознаки навколо центра розподілу;

– оцінювання характеристик форми розподілу.

Центром розподілу називають варіанту, навколо якої групують інші значення ознаки. Основними характеристиками центра розподілу є середнє, мода та медіана, які для якісно однорідної сукупності незначно відрізняються між собою [11, с. 17].

Коефіцієнтами варіації називають розмах варіації, середнє арифметичне відхилення, квадратичне (стандартне) відхилення та дисперсію. Ці показники характеризують розподіл даних навколо середнього [62, с. 74].

Розмах варіації визначає максимальну амплітуду коливань значень ознаки у сукупності:

$$R = x_{max} - x_{min}.$$

Середнє арифметичне (лінійне) відхилення характеризує середню величину коливань значень ознаки навколо середнього рівня для незгрупованих та згрупованих даних відповідно:

$$d_{\text{лін}} = \frac{\sum_{i=1}^{n} \left| x_i - x_{cep} \right|}{n} = \frac{\sum_{i=1}^{k} \left| x_i - x_{cep} \right| f_i}{\sum_{i=1}^{k} f_i},$$

де n – кількість одиниць сукупності, k – кількість варіант.

Дисперсія – показник середнього квадрата відхилень значень ознаки від середнього рівня, міра розсіювання варіант:

$$\sigma^2 = \frac{\sum_{i=1}^{n} \left(x_i - x_{cep} \right)^2}{n} = \frac{\sum_{i=1}^{k} \left(x_i - x_{cep} \right)^2 f_i}{\sum_{i=1}^{k} f_i},$$

де n – кількість одиниць сукупності, k – кількість варіант.

Види дисперсій:

► загальна (σ^2) оцінює вплив постійних (систематичних) та випадкових факторів, що спричинили варіацію ознаки; характеризує варіацію ознаки навколо загальної середньої;

► групова (часткова, σ_i^2) характеризує вплив випадкових факторів (усіх, крім основи групування); розраховує відхилення варіант у групі від групової середньої; узагальнюючою мірою внутрішньогрупової варіації є середня з групових дисперсій ($\sigma_{i_{сер}}^2$);

► міжгрупова дисперсія (δ^2) оцінює вплив постійного фактора; характеризує відхилення групових середніх від загальної (систематичну варіацію).

Середнє квадратичне (стандартне) відхилення (σ) – квадратний корінь з дисперсії:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - x_{сер}\right)^2}{n}}.$$

Чим менше стандартне відхилення, тим більш однорідною є досліджувана сукупність.

Статистична обробка даних може бути проведена у середовищі будь-якого статистичного пакета. Наступний приклад демонструє аналіз даних за допомогою статистичного пакета «Statistica» (рис. 2.1–2.3) [36, с. 89].

Descriptive Statistics

Variable	Valid N	Mean	Geometric Mean	Harmonic Mean	Median	Mode	Frequency of Mode	Minimum	Maximum	Std.Dev.
GDP	182	21369,82	11681,66	5464,377	13652,17	Multiple	1	406,6671	124609,3	22826,00

Рис. 2.1. Вікно результатів обчислення описових статистик

У програмі «Statistica» буквою *p* позначають статистичну значущість (рівень значущості для перевірки H_0). При $p \le 0{,}05$ результати вважають значущими. Більш низький *p*-рівень відповідає вищому рівню надійності нульової гіпотези. При $p = 0{,}01$ надійність результатів зростає – статистична значущість (*p*-рівень) знаходиться у спадаючій залежності від надійності результату.

Valid N – кількість спостережень, придатних для аналізу.

Category	Frequency table: Var1 (Ex1.sta) K-S d=,14151, p> .20; Lilliefors p<,05 Count	Cumulative Count	Percent of Valid	Cumul % of Valid	% of all Cases	Cumulative % of All
10,00000<x<=20,00000	1	1	2,08333	2,0833	2,08333	2,0833
20,00000<x<=30,00000	0	1	0,00000	2,0833	0,00000	2,0833
30,00000<x<=40,00000	3	4	6,25000	8,3333	6,25000	8,3333
40,00000<x<=50,00000	1	5	2,08333	10,4167	2,08333	10,4167
50,00000<x<=60,00000	2	7	4,16667	14,5833	4,16667	14,5833
60,00000<x<=70,00000	7	14	14,58333	29,1667	14,58333	29,1667
70,00000<x<=80,00000	10	24	20,83333	50,0000	20,83333	50,0000
80,00000<x<=90,00000	9	33	18,75000	68,7500	18,75000	68,7500
90,00000<x<=100,0000	15	48	31,25000	100,0000	31,25000	100,0000
Missing	0	48	0,00000		0,00000	100,0000

Рис. 2.2. Таблиця частот розподілу

Щоб знайти квантиль розподілу, сформувати графік теоретичної функції розподілу або обчислити її значення у заданій точці, використовують імовірнісний калькулятор у програмі «Statistica».

Квантиль – значення ранжованої змінної, що відокремлює від варіаційного ряду визначену частку обсягу сукупності. Найбільш поширеними є **процентилі** (персентилі) та **квартилі**.

Процентилі ділять упорядковану сукупність на сто частин, тобто відокремлюють від сукупності по 0,01 частині (1%). **Квартилі** ділять сукупність на чотири рівні частини.

На коробчастій діаграмі маленький прямокутник відповідає значенню медіани, великий прямокутник – верхньому та нижньому квартилям, а вуса – найменшому та найбільшому значенням вибірки (рис. 2.3).

Статистичні ряди зображають у вигляді кривої, яка наближено представляє графік ряду при збільшенні обсягу сукупності та зменшенні довжини інтервалів групування.

Крива розподілу (графік щільності розподілу) – графічне зображення (неперервна лінія) змін частот варіаційного ряду, функціонально пов'язаних зі зміною варіант (співвідношення варіант і частот). Ця крива характеризує емпіричний або теоретичний розподіли.

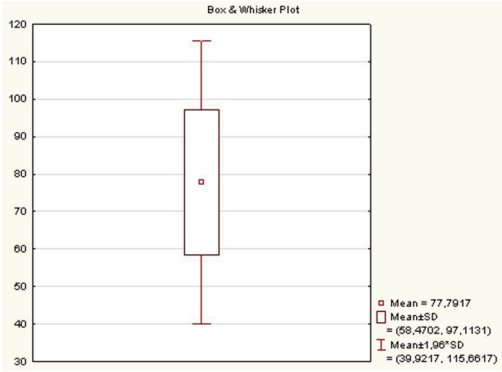


Рис. 2.3. Коробчаста діаграма

Емпіричною функцією розподілу (рис. 2.4) називають функцію $F_n(x)$, яка кожному значенню x ставить у відповідність частку подій $X \le x$ (закон зміни частоти події).

Теоретичною функцією розподілу випадкової величини x називають функцію дійсного аргументу $F(x) = P(X \le x)$ генеральної сукупності, яка відображає ймовірність появи значень у заданому діапазоні (рис. 2.4).

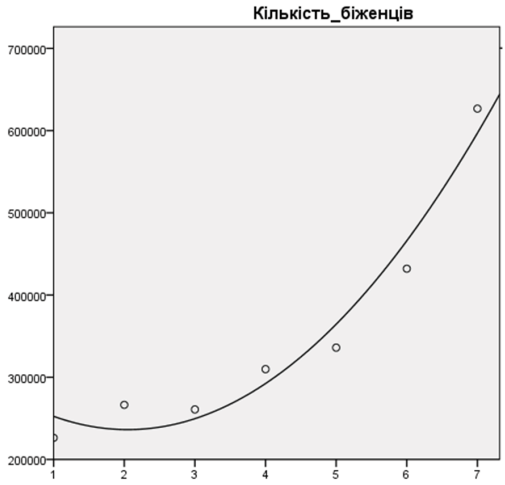


Рис. 2.4. Дані спостережень та графік теоретичного розподілу ряду динаміки

Емпірична функція визначає відносну частоту появи події, а теоретична – ймовірність її появи. Зі збільшенням обсягу вибірки емпірична функція розподілу наближається до теоретичної.

Аналіз варіаційних рядів полягає у зіставленні емпіричного та теоретичного розподілів і визначенні ступеня відхилень між ними.

Характеристики форми розподілу

За формою виділяють одно-, дво- та багатовершинні розподіли. Багатовершинність вказує на неоднорідність досліджуваної сукупності. В такому випадку проводять перегрупування даних для формування однорідних груп. Одновершинні розподіли характеризують якісно однорідні сукупності. Серед них виділяють гостро- і плосковершинні, симетричні та асиметричні (скошені).

Асиметрія – ступінь відхилення графіка розподілу частот від симетричного вигляду відносно середнього значення. **Ексцес** – міра плосковершинності або гостровершинності графіка розподілу досліджуваної ознаки. **Коефіцієнт асиметрії** *As* та **коефіцієнт ексцесу** *Ex* визначають ступінь скошеності та рівень гостро- чи плосковершинності розподілу відповідно (рис. 2.5).

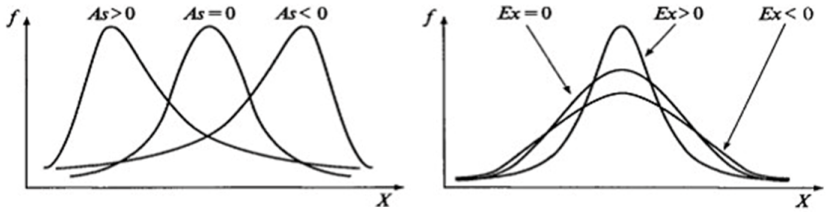


Рис. 2.5. Характеристики форми розподілу: асиметрія та ексцес

При симетричному розподілі частоти варіант, рівновіддалених у різні боки від центра тяжіння, рівні між собою. Серед асиметричних розподілів виділяють правосторонні ($x_{сер} > Me > Mo$) та лівосторонні ($x_{сер} < Me < Mo$) асиметрії, які є наслідками обмеженої варіації ознак в одному напрямку чи впливу домінантного фактора динаміки явища, що зумовлює зміщення центра розподілу.

Стандартна похибка є характеристикою точності, або стабільності, величини, для якої вона обчислюється.

Нормальний розподіл

Серед найбільш відомих теоретичних розподілів виділяють **нормальний розподіл (розподіл Лапласа–Гауса)**, який представляє нормальна крива (симетрична відносно максимальної ординати). Такий розподіл є результатом впливу значної кількості незалежних один від одного факторів на значення ознаки (рис. 2.6). Для нормального розподілу $As = Ex = 0$.

Нормальний розподіл повністю визначають два параметри – середнє арифметичне $x_{сер}$ та середнє квадратичне відхилення σ. Значення ознаки переважно розподілені навколо центра розподілу $x_{сер}$. Нормальний розподіл відображає значну кількість незалежних варіант спостережуваних ознак, які суттєво не відрізняються між собою.

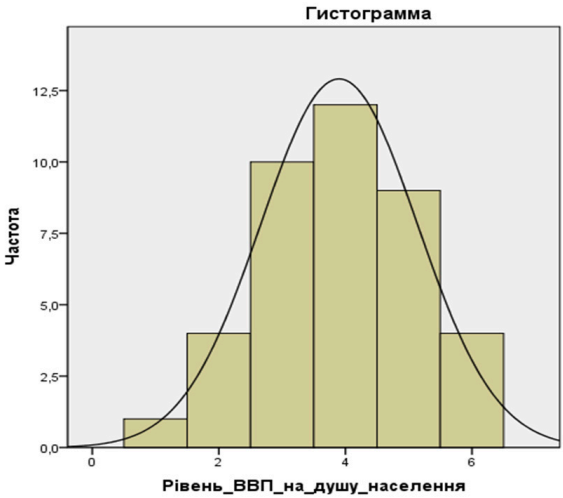


Рис. 2.6. Гістограма емпіричних даних та графік нормального розподілу

Поняття «нормальний розподіл» є основою багатьох методів емпіричних досліджень. Це своєрідний еталон для оцінювання характеристик інших розподілів.

Для перевірки ряду на відповідність нормальному розподілу будують *P-P* діаграму (рис. 2.7).

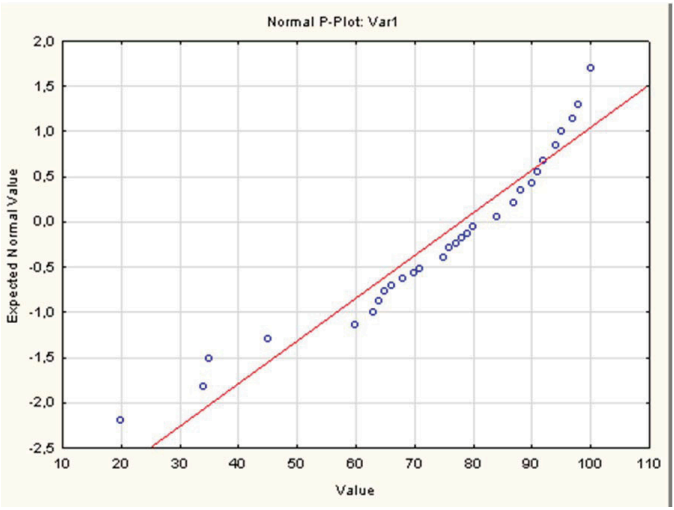


Рис. 2.7. *P-P* діаграма

Точки на діаграмі лежать поблизу суцільної червоної лінії – розподіл близький до нормального.

Для перевірки ряду розподілу на відповідність одному з відомих розподілів використовують *Q-Q* діаграму.

Приклади теоретичних функцій розподілу

Рівномірний розподіл – розподіл випадкової величини, який характеризує однакова ймовірність появи кожного із варіантів з визначеного інтервалу [*a, b*]. Рівномірному закону розподілу підпорядковуються, наприклад, похибки заокруглень, які виникають при розрахунках.

Біноміальний розподіл – дискретний імовірнісний розподіл, що з імовірністю *p* характеризує кількість успіхів (так/ні) у послідовності експериментів.

Розподіл Пуассона (закон розподілу рідкісних подій) стосується ймовірності появи сприятливої події у великій кількості експериментів.

Серед інших відомих законів щільності розподілу – геометричний, логарифмічний, бета- та гамма-розподіл і т. ін. [22, с. 109].

Для перевірки відповідності досліджуваного розподілу нормальному частоти розподілу даних спостережень порівнюють з теоретичними частотами, які характерні для нормального розподілу сукупності. Емпіричні дані підставляють у відповідну аналітичну формулу, за якою розраховують теоретичні частоти кривої нормального розподілу. Ступінь відповідності між фактичним і теоретичним розподілами оцінюють за допомогою показників (критеріїв) згоди.

Статистика володіє великою кількістю різноманітних методів перевірки статистичних гіпотез. При виборі методу для вирішення конкретного завдання враховують мету перевірки гіпотези, шкали вимірювання емпіричних даних, кількість вибірок, які потрібно порівняти, та їх залежність (незалежність) між собою.

Якщо потрібно порівняти більше, ніж 2 вибірки, застосовують методи дисперсійного аналізу (п. 2.3).

2.2. Перевірка статистичних гіпотез

Нульова гіпотеза H_0 – гіпотеза, яку перевіряють [8, с. 43].

Перевіряють переважно припущення про те, що обчислений за вибірковими емпіричними даними показник несуттєво відрізняється від його гіпотетичного значення (розрахованого теоретично) або визначеної величини за генеральною сукупністю. Для встановлення цього факту висувають наступну H_0: відмінність між емпіричними та теоретичними показниками дорівнює нулю. З огляду на це гіпотезу, яку перевіряють, називають нульовою (основною або робочою).

Приклади нульових гіпотез:

– H_0: середні значення двох вибірок істотно не відрізняються між собою;

– H_0: не існує залежності між досліджуваними змінними;

– H_0: розподіл вибірки відповідає нормальному закону з заданими параметрами.

У кожному випадку нульовій гіпотезі протиставляють конкуруючу гіпотезу, яка заперечує H_0.

Конкуруюча (альтернативна) гіпотеза H_1 – логічне заперечення H_0.

Приклади конкуруючих гіпотез для запропонованих вище нульових:

– H_1: середні значення двох вибірок істотно відрізняються між собою;

– H_1: між досліджуваними змінними існує залежність;

– H_1: розподіл вибірки не відповідає нормальному закону з заданими параметрами.

Для однієї H_0 переважно можна сформулювати кілька різних альтернативних гіпотез.

Статистичним гіпотезам властиві несуперечливість (відсутність логічних суперечностей у формулюванні) та повнота (сумарна ймовірність нульової та альтернативної гіпотез $p(H_0) + p(H_1) = 1$).

Гіпотези поділяють на **параметричні** (припущення про параметри генеральної сукупності) та **непараметричні** (припущення про розподіли). Якщо гіпотеза складається з одного твердження, її називають **простою**. **Складні** гіпотези містять декілька простих.

Якщо є можливість визначити напрям розбіжності досліджуваного показника порівнюваних вибірок, висувають **однобічні** гіпотези. Наприклад, H_0: дисперсія першої вибірки суттєво не перевищує дисперсії іншої. **Двобічні** гіпотези формулюють, якщо необхідно довести лише відмінності форми або значень показників розподілу ознак. Наприклад, H_0: середні значення двох вибірок істотно не відрізняються між собою [61, с. 163].

При перевірці статистичної гіпотези потрібно впевнитись в узгодженості емпіричних даних з висунутою H_0, встановити систематичність (чи випадковість) факторів, які спричиняють відмінності між гіпотезою і результатами спостереження. Результатом

перевірки гіпотези є прийняття рішення про вибір однієї з альтернатив (можливих взаємовиключних висновків). За підсумками перевірки гіпотези її приймають з певною ймовірністю або відхиляють [53, с. 27].

На практиці при перевірці нульової гіпотези використовують відомі модельні розподіли (статистичні критерії), які приблизно відповідають розподілу досліджуваного параметра (п. 2.1) [8, с. 43].

Критерій згоди – статистичний показник, на основі оцінювання величини якого роблять висновок при проведенні досліджень.

Значення критерію – величина, розрахована за досліджуваними вибірками.

Застосування критеріїв для прийняття (відхилення) статистичних гіпотез проводять на заданому рівні значущості (з довірчою ймовірністю).

Рівень статистичної значущості α – ймовірність помилкового відхилення правильної нульової гіпотези (випадкові відмінності між ознаками взято за істотні). Наприклад, якщо результати перевірки гіпотези (відмінності) достовірні на 5%-му рівні значущості, це означає, що з ймовірністю 0,05 вони все ж таки недостовірні (випадкові).

При дослідження МВ за низький прийнято 5%-й рівень статистичної значущості ($\alpha < 0,05$), достатній – 1%-й рівень ($\alpha < 0,01$), високий – 0,1%-й рівень ($\alpha < 0,001$). У таблицях критичних значень наводять значення критеріїв, які відповідають рівням статистичної значущості $\alpha < 0,05$, $\alpha < 0,01$ та $\alpha < 0,001$. У загальному випадку, якщо рівень статистичної значущості $\alpha \geq 0,05$, нульову гіпотезу про відсутність відмінностей відхиляють та приймають гіпотезу про статистичну достовірність відмінностей.

При перевірці двобічних гіпотез рівень значущості критерію (двобічного) у 2 рази більший, ніж для відповідних однобічних. Перед використанням однобічних критеріїв спочатку розраховують двобічні. Якщо гіпотезу про відсутність відмінності між вибірками прийнято, використання однобічного критерію є недоцільним.

Альтернативою рівню значущості є **довірчий рівень (інтервал)** $p = 1 - \alpha$. Після вибору рівня значущості для конкретного дослідження за емпіричними даними вибірки розраховують значення критерію та порівнюють його з обчисленим для заданого рівня значущості критичним (табличним) значенням. Якщо розраховане значення критерію більше за табличне, нульову гіпотезу відхиляють. В іншому разі H_0 приймають на заданому рівні значущості або проводять додаткові дослідження.

На практиці при проведенні емпіричних досліджень МВ використовують нормальний розподіл, χ^2-розподіл, розподіли Стьюдента і Фішера, критерій Колмогорова–Смирнова, Романовського, Ястремського і т. ін. Необхідною умовою використання цих критеріїв є достатня кількість спостережень – не менше 100 (20–30 для критерію Пірсона) [36, с. 71].

Одним з найбільш поширених є критерій Пірсона (*Xi*-квадрат):

$$\chi^2 = \sum \frac{\left(f_{теор} - f_{емп}\right)^2}{f_{теор}}.$$

Чим більшою є різниця між емпіричними $f_{емп}$ та теоретичними частотами $f_{теор}$, тим більше значення критерію Пірсона. Обчислене значення критерію $\chi^2_{ф}$ порівнюють з табличним (критичним) значенням $\chi^2_{табл}$.

Якщо $\chi^2_{ф} > \chi^2_{табл}$, розбіжність між емпіричними та теоретичними частотами істотна і її не можна пояснити випадковими коливаннями даних, а емпіричний розподіл є принципово відмінним від теоретичного.

Якщо $\chi^2_{ф} < \chi^2_{табл}$, відхилення фактичних частот від теоретичних вважають випадковим, неістотним. Емпіричний розподіл відповідає теоретичному.

Приклад 2.1. За умовними даними спостереження (табл. 2.4) розрахувати значення критерію Пірсона для перевірки H_0: форма державного устрою не впливає на членство у СОТ, тобто відмін-

ності, що спостерігаються серед країн-членів СОТ (31,8% в першій групі проти 53,3% у другій групі), є повністю випадковими і не пов'язані з формою державного устрою.

Таблиця 2.4

Умовні дані спостереження

		Членство у СОТ		Всього
		так	ні	
Форма державного устрою	унітарна	13	12	25
	федеративна	27	18	45
Всього		40	30	70

Для перевірки H_0 потрібно визначити, якими були б результати, якщо б форма державного устрою не впливала на членство у СОТ (розрахувати очікувані частоти появи значень для відповідних комірок таблиці спряженості).

У спостереженні не є членами СОТ всього 30 країн, що становить 42,9% від загальної кількості країн. Якщо форма державного устрою не впливає на членство у СОТ, в обох експериментальних групах (з унітарною та федеративною формами державного устрою) має спостерігатись однаковий відсоток країн, що не входять у СОТ – 42,9%.

42,9% від 25 (країни з унітарною формою державного устрою) становить 11 і від 45 (федеративна форма державного устрою) – 19. Це очікувані частоти – теоретично розрахована кількість країн, що не входять до складу СОТ в експериментальних групах (з різними формами державного устрою).

Аналогічно розрахуємо кількість країн, що входять у СОТ, – всього 40 країн (57,1% від загальної кількості). Очікувані частоти країн, що є членами СОТ, становлять 14 (57,1% від 25 країн з унітарною формою державного устрою) і 26 (57,1% від 45 країн з федеративною формою державного устрою).

Очікувані частоти не суттєво відрізняються від спостережуваних (табл. 2.5), тобто форма державного устрою країн не впливає на їх членство у СОТ.

Таблиця 2.5

Таблиця спряженості (взаємозв'язку) очікуваних частот

		Членство у СОТ		Всього
		так	ні	
Форма державного устрою	унітарна	14	11	25
	федеративна	26	19	45
Всього		40	30	70

Цей висновок можна кількісно виразити за допомогою критерію згоди Пірсона χ^2:

$$\chi^2 = \sum \frac{\left(f_{теор} - f_{емп}\right)^2}{f_{теор}},$$

де $f_{емп}$ та $f_{теор}$ – спостережені (емпіричні) та очікувані (теоретично розраховані) частоти відповідно. Сумування проводять за всіма комірками таблиці:

$$\chi^2 = \frac{(14-13)^2}{14} + \frac{(26-27)^2}{26} + \frac{(11-12)^2}{11} + \frac{(19-18)^2}{19} = 0{,}25.$$

Розраховане значення критерію χ^2 порівнюють з критичним (табличним значенням). Для цього обирають рівень значущості (наприклад, $\alpha = 0{,}05$), для якого довірчий рівень (інтервал) $p = 1 - \alpha = 0{,}95$, та обчислюють кількість ступенів вільності.

Ступені вільності (ступені свободи) – кількість незалежних змінних, які однозначно описують досліджуване явище.

Для χ^2 кількість ступенів вільності $df = (R - 1)(C - 1)$, де R і C – відповідно кількість рядків і стовпців таблиці спряженості. У прикладі 2.1 отримаємо:

$$df = (2 - 1)(2 - 1) = 1.$$

Порівняємо обчислене значення критерію $\chi^2_{ф} = 0{,}25$ з відповідним критичним значенням для $p = 0{,}95$ та $df = 1$ ($\chi^2_{табл} = 3{,}842$).

Обчислене $\chi^2_{ф} < \chi^2_{табл}$. Нульову гіпотезу про відсутність впливу форми державного устрою країни на її членство у СОТ приймаємо. При одному ступені вільності тільки в 5% випадків величина критерію $\chi^2_{ф}$ перевищує $\chi^2_{табл}$.

Критерій χ^2 передбачений і у програмному пакеті «Statistica» (рис. 2.8–2.9).

Критерій χ^2 для перевірки рядів розподілу на відповідність статистичним розподілам

Приклад 2.2. Перевірити гіпотезу про нормальний розподіл значень рівня ВВП для 40 країн-членів СОТ.

Висуваємо гіпотезу H_0: генеральна сукупність розподілена нормально.

Отримані оцінки застосування критерію χ^2 для перевірки результатів на відповідність нормальному розподілу представлені на рис. 2.8.

Variable: Рівень ВВП, Distribution: Normal (Ex6_1.sta)
Chi-Square = 2,16219, df = 3 (adjusted) , p = 0,53943

Upper Boundary	Observed Frequency	Cumulative Observed	Percent Observed	Cumul. % Observed	Expected Frequency	Cumulative Expected	Percent Expected	Cumul. % Expected	Observed-Expected
<= 100,00000	0	0	0,00000	0,0000	1,590092	1,59009	3,97523	3,9752	-1,59009
105,00000	5	5	12,50000	12,5000	2,107574	3,69767	5,26894	9,2442	2,89243
110,00000	5	10	12,50000	25,0000	3,684244	7,38191	9,21061	18,4548	1,31576
115,00000	5	15	12,50000	37,5000	5,378483	12,76039	13,44621	31,9010	-0,37848
120,00000	5	20	12,50000	50,0000	6,557305	19,31770	16,39326	48,2942	-1,55730
125,00000	5	25	12,50000	62,5000	6,676509	25,99421	16,69127	64,9855	-1,67651
130,00000	5	30	12,50000	75,0000	5,677176	31,67138	14,19294	79,1785	-0,67718
135,00000	5	35	12,50000	87,5000	4,031537	35,70292	10,07884	89,2573	0,96846
140,00000	5	40	12,50000	100,0000	2,390875	38,09380	5,97719	95,2345	2,60913
< Infinity	0	40	0,00000	100,0000	1,906204	40,00000	4,76551	100,0000	-1,90620

Рис. 2.8. Результати перевірки ряду на відповідність нормальному розподілу

Досягнутий рівень значущості дорівнює $p = 0{,}53943$. Гіпотезу про нормальність відхиляємо.

Аналогічно перевіряємо гіпотезу про рівномірний розподіл значень ряду. H_0: значення розподілені рівномірно. Досягнутий рівень значущості дорівнює $p = 0{,}99916$. Гіпотезу про рівномірну розподіленість відхиляємо.

Критерій χ^2 для гіпотези незалежності випадкових величин

Приклад 2.3. У таблиці наведено дані про 40 країн-членів СОТ, які було класифіковано щодо статі керівника держави та форми державного правління. Потрібно дослідити, чи можна на основі цих даних отримати висновок про наявність зв'язку між статтю керівника держави та формою державного правління.

H_0: політичний режим та членство у ЄС не пов'язані між собою.

Результати застосування критерію χ^2 для перевірки гіпотези про незалежність випадкових величин представлені на рис. 2.9.

Statistic	Statistics: Політичний режим(3) x Членство у ЄС(2)		
	Chi-square	df	p
Pearson Chi-square	10,90909	df=2	p=,00428
M-L Chi-square	14,02951	df=2	p=,00090

Рис. 2.9. Результати перевірки гіпотези незалежності випадкових величин

Досягнутий рівень значущості 0,004 – результати є статистично значущими. Кількість ступенів вільності $df = (3 - 1)(2 - 1) = 2$. $\chi^2 = 10{,}9$. $\chi^2(0{,}99;2) = 9{,}2$. Розраховане значення критерію більше за табличне. Отже, існує зв'язок між досліджуваними змінними.

Для оцінювання значущості отриманого значення χ^2 використовують критерій Романовського [36, с. 73]:

$$R = \frac{\chi^2 - k}{\sqrt{2k}},$$

де $k = s - 1 - r$ визначає кількість ступенів свободи, s – кількість груп (або часткових інтервалів), r – кількість зв'язків.

При $R \geq 3$ значення χ^2 вважають значущим, а порівнювані вибірки – істотно різними.

Критерії та тести для порівняння вибірок, які передбачають обчислення числового значення (параметра), називають **параметричними** [61, с. 176]. Такі критерії застосовують також при множинних порівняннях (порівняння двох груп вибірок – одну з іншою).

Параметричні критерії передбачають припущення, що розподіл ознаки в генеральній сукупності підпорядковується одному із відомих законів. Цю гіпотезу перевіряють до застосування параметричних тестів. Більшість таких критеріїв розроблено для нормально розподілених даних. Здебільшого параметричні тести більш потужні, ніж непараметричні.

Емпіричні дані досліджень МВ можуть бути представлені **незалежними** або **спряженими** (залежними) вибірками. Для незалежних вибірок критерії встановлюють статистичну значущість відмінностей між ними.

Приклади незалежних вибірок:

– мешканці двох різних держав (при демографічних дослідженнях);

– моніторинг громадської думки різними незалежними організаціями.

Для порівняння середніх значень двох незалежних нормальних вибірок з генеральних сукупностей, які мають відомі (рівні або нерівні) чи рівні невідомі дисперсії, застосовують *t*-критерій Стьюдента [36, с. 74]:

$$t = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}},$$

де σ_1^2, σ_2^2 – відомі внутрішньогрупові дисперсії, n_1 та n_2 – обсяги груп.

Для *m* груп рівних обсягів статистика має *t*-розподіл з кількістю ступенів вільності $m(n - 1)$, де *n* – загальна кількість одиниць досліджуваної сукупності.

Критерії, які застосовують до вибірок з попарно спряженими даними, називають **парними**.

Приклади спряжених вибірок:

– мешканці двох різних держав (при оцінюванні рівня бідності);

– моніторинг громадської думки до та після значущої події суспільного життя.

При аналізі **спряжених вибірок** їх порівняння проводять з метою встановлення впливу досліджуваного фактора. Обмеження на рівність дисперсій не накладають. Наприклад, H_0: немає різниці

між середніми значеннями вибірок. Значення критерію розраховують за формулою [36, с. 75]:

$$t = \frac{\sum_{i=1}^{n} \delta_i}{\sqrt{\frac{n\sum_{i=1}^{n} \delta_i^2 - \left(\sum_{i=1}^{n} \delta_i\right)^2}{n-1}}},$$

де n – кількість елементів у кожній із вибірок, $\delta_i = x_i - y_i$, x_i та y_i – відповідні значення елементів першої та другої вибірок.

Цей критерій ще називають **одновибірковим критерієм Стьюдента**. Відповідна статистика має розподіл Стьюдента з кількістю ступенів вільності $n - 1$.

Якщо дисперсії двох вибірок (або їх відношення) невідомі, а припущення про рівність дисперсій є необґрунтованим, застосовують **критерій Уєлча (Крамера–Уєлча)** [36, с. 75]:

$$d = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}},$$

де σ_1^2, σ_2^2 – розраховані за вибірками оцінки дисперсії.

***F*-критерій Фішера** використовують для порівняння дисперсій двох вибірок. Його значення розраховують за формулою [36, с. 76]:

$$F = \frac{\sigma_1^2}{\sigma_2^2},$$

де σ_1^2, σ_2^2 – значення оцінок більшої та меншої дисперсій відповідно. Кількості ступенів вільності для пошуку критичного значення обирають рівними $n_1 - 1$ та $n_2 - 1$ для кожної з вибірок відповідно. Гіпотезу про рівність дисперсій порівнюваних сукупностей відхиляють, якщо обчислене значення перевищує табличне на заданому рівні значущості.

Для отримання більш достовірних результатів при перевірці гіпотез прийнято розв’язувати завдання кількома методами та зіставляти отримані результати.

2.3. Дисперсійний аналіз

Дисперсійний аналіз, ANOVA (Analysis of Variations) – сукупність методів статистичної оцінки значущості залежності результатної ознаки від одного або кількох досліджуваних факторів (атрибутивних ознак) та встановлення ступеня їх впливу.

Фактори – контрольовані фактори, які впливають на значення результатної ознаки. **Рівень фактора** – значення, які характеризують конкретний його прояв. Ці значення зазвичай представляють у номінальній або порядковій шкалі вимірювань. Значення вимірюваної ознаки називають **відгуком**.

За допомогою методу дисперсійного аналізу перевіряють статистичні гіпотези про рівність середніх кількох нормально розподілених генеральних сукупностей. Наприклад, для оцінювання ефективності проведення соціально значущого заходу чи рекламної кампанії порівнюють середні показники до та після їх проведення (виявлення вірогідності відмінностей значень результатної ознаки) [8, с. 81].

При дисперсійному аналізі перевірку статистичної значущості відмінності між середніми декількох груп проводять на основі вибіркових дисперсій. Для цього загальну дисперсію (дисперсію варіації) розбивають на частини, одна з яких зумовлена випадковою помилкою (внутрішньогруповою мінливістю), інша – відмінністю середніх значень. Якщо ця відмінність значуща, нульову гіпотезу про рівність середніх значень відкидають на обраному рівні значущості.

У випадку, коли емпіричні значення факторів представлені у кількісній або порядковій шкалі, вихідні дані групують у ряди спостережень, що відповідають приблизно однаковим значенням фактора. Кількість груп не може бути надто великою, оскільки для отримання надійних результатів кожна з них має містити достатню кількість спостережень (не менше 3). Однак при виборі малої кількості груп неможливо врахувати суттєві особливості впливу

досліджуваного фактора на явище. Вибір способу групування даних залежить від їх обсягу та специфіки варіювання (мінливості) значень фактора.

Однофакторний дисперсійний аналіз

Дисперсійний однофакторний аналіз використовують при дослідженні зміни результатної ознаки під впливом зміни умов або градацій одного фактора. Метод полягає в перевірці гіпотези про рівність дисперсій за факторами та загальної дисперсії досліджуваної емпіричної сукупності.

Основною метою однофакторного дисперсійного аналізу є оцінювання величини впливу конкретного фактора на досліджуваний відгук або порівняння двох чи більше факторів між собою з метою встановлення різниці їх впливу на результатну ознаку **(контраст факторів)**. Метод передбачає попередню перевірку нульової гіпотези про відсутність будь-якого впливу досліджуваного фактора (факторів) – H_0: зміни значень ознаки в порівнюваних вибірках є випадковими і всі дані належать до однієї генеральної сукупності.

Обов'язковою умовою проведення однофакторного дисперсійного аналізу є наявність не менше двох градацій фактора, кожна з яких містить не менше трьох спостережень. При проведенні дисперсійного аналізу необхідно перевірити відповідність досліджуваної випадкової величини нормальному розподілу та відсутність відмінності дисперсій сукупностей [19, с. 423]. З цією метою використовують методи перевірки статистичних гіпотез.

Якщо нульову гіпотезу відкидають, наступним етапом є кількісне оцінювання впливу досліджуваного фактора і побудова довірчих інтервалів для отриманих характеристик. У випадку, коли нульова гіпотеза не може бути відкинутою, її приймають і роблять висновок про відсутність впливу. Якщо є підстави вважати, що такий вплив фактора на відгук має бути присутнім (наприклад, з теоретичних уявлень про досліджуваний об'єкт), необхідно пере-

вірити наявність інших факторів, що можуть приховувати (нівелювати) цей вплив.

Для проведення однофакторного дисперсійного аналізу розглядають не менше трьох незалежних вибірок з однієї генеральної сукупності, які отримані шляхом зміни значень одного з факторів, що не має кількісного вираження. Щодо аналізованих вибірок роблять припущення про відмінність їх вибіркових середніх та рівність вибіркових дисперсій. Результатом досліджень є підтвердження істотного впливу фактора на зміну вибіркових середніх відгуку чи встановлення факту випадковості такої зміни (розкиду), спричиненої незначними обсягами вибірок. Якщо вибірки належать до однієї генеральної сукупності, розкид даних між вибірками (групами) буде не більший, ніж розкид даних усередині цих вибірок (груп).

Дисперсійний аналіз полягає у дослідженні відмінності між середніми значеннями окремих груп спостережень, кожну з яких характеризує визначений рівень (або діапазон) фактора. При дослідженні статистичної значущості різниці між середніми значеннями груп порівнюють оцінки дисперсій.

Основою дисперсійного аналізу є представлення загальної дисперсії як суми внутрішньогрупової дисперсії та загальної дисперсії, розрахованої за всіма емпіричними даними. Якщо фактор є значущим (впливовим), дисперсія спостережень в окремій групі значно менша, ніж дисперсія всієї вибірки. Це пояснюється суттєвою різницею між груповими середніми.

При однофакторному дисперсійному аналізі вихідні дані представляють у вигляді таблиць, в яких кількість стовпців дорівнює кількості рівнів фактора, а кількість значень у кожному стовпці – кількості спостережень при відповідному рівні фактора (табл. 2.6). Для різних рівнів фактора кількість спостережень може бути різною.

Приклад 2.4. Дослідити вплив проведення передвиборчих агіткампаній на вибір електорату.

Таблиця 2.6

Умовні дані соціологічних опитувань до та після проведення передвиборчих агіткампаній

К-сть бажаючих віддати свій голос	До проведення агіткампанії	Після проведення агіткампанії
Спостереження 1	25	30
Спостереження 2	15	25
Спостереження 3	20	35
Групове середнє	20	30
Сума квадратів відхилень у групах	σ_1^2 =(25–20)² + (15–20)² + (20–20)² = 50	σ_2^2 = (30–30)² + (25–30)² + (35–30)² = 50
Загальне середнє	25	
Загальна сума квадратів відхилень	$\sigma_{заг}^2$ = (25–25)² + (15–25)² + (20–25)² + (30–25)² + (25–25)² + (35–25)² = 250	

Середні двох груп суттєво відрізняються (20 та 30 відповідно). Сума квадратів відхилень у кожній з груп дорівнює 50.

Залишкова (внутрішньогрупова) варіація є сумою квадратів відхилень спостережень від відповідних групових середніх

$$\sigma_{1,2}^2 = \sigma_1^2 + \sigma_2^2 = 50 + 50 = 100$$

і характеризує коливання значень досліджуваної ознаки, зумовлені випадковими відхиленнями від групових середніх (неврахованими або випадковими факторами).

Без врахування наявності окремих груп (впливу фактора на результат) загальна сума квадратів відхилень від загального середнього цих двох вибірок становить:

$$\sigma_{заг}^2 = 250.$$

Факторна (міжгрупова) варіація є зваженою сумою квадратів відхилень групових середніх від загального середнього:

$$\sigma^2_{\text{гр}} = (20-25)^2 + (30-25)^2 = 50$$

і характеризує коливання значень, зумовлені фактором, на основі якого здійснено групування даних.

Основна тотожність дисперсійного аналізу:

$$\sigma^2_{\text{заг}} = \sigma^2_{1,2} + \sigma^2_{\text{рсер}},$$

де $\sigma^2_{\text{рсер}}$ – сума квадратів, зумовлена різницею середніх значень між групами:

$$\sigma^2_{\text{рсер}} = \sigma^2_{\text{заг}} - \sigma^2_{1,2} = 250 - 100 = 150.$$

Перевірка значущості в дисперсійному аналізі заснована на порівнянні компоненти дисперсії, зумовленої міжгруповим розкидом, і компоненти дисперсії, зумовленої внутрішньогруповим розкидом. Отримані компоненти дисперсії порівнюють за допомогою *F*-критерію.

У випадку справедливості гіпотези H_0 *F*-відношення

$$F = \frac{\dfrac{\sigma^2_{\text{рсер}}}{m-1}}{\dfrac{\sigma^2_{1,2}}{n-m}}$$

має розподіл Фішера з $m - 1$ та $n - m$ ступенями вільності (n – кількість одиниць сукупності, m – кількість груп).

Для прикладу 2.4 отримаємо:

$$F = \frac{\dfrac{150}{2-1}}{\dfrac{100}{6-2}} = \frac{150}{25} = 6.$$

Враховуючи кількість ступенів вільності $m - 1 = 1$ та $n - m = 4$ ($m = 2$ – кількість груп, $n = 6$ – кількість спостережень) та рівень значущості 0,05, за таблицею розподілу Фішера знайдемо критичне значення $F_{табл}(0{,}05;1;4) = 7{,}71$. Оскільки $F_{емп} < F_{табл}$, нульову гіпотезу приймаємо – групові середні відрізняються не суттєво.

Доходимо висновку, що проведення передвиборчих агіткампаній не впливає на думку респондентів.

Рангові методи не передбачають відповідності результатів спостережень нормальному розподілу та можуть бути застосованими як для оцінювання впливу кількісних даних з невідомим законом розподілу, так і для порядкових ознак [8, с. 90].

У таблиці рангового дисперсійного однофакторного аналізу замість результатів спостережень записують їх ранги r_{ij}, отримані шляхом впорядкування за зростанням усієї сукупності спостережень x_{ij} (табл. 2.7).

Для кожного рівня фактора (стовпця) розраховують суму рангів $R_j = \sum_{i=1}^{n_j} r_{ij}$ або відповідні середні ранги $\langle R_j \rangle = \frac{1}{n_j} \sum_{i=1}^{n_j} r_{ij}$.

Таблиця 2.7

Загальний вигляд таблиці даних рангового однофакторного аналізу

Спостереження	Вибірки			
	1	2	3	4
1	r_{11}	r_{12}	…	r_{1k}
2	r_{21}	r_{22}	…	r_{2k}
…	…	…	…	…
n_i	$n_{n_1 1}$	$n_{n_1 2}$	…	$n_{n_1 k}$

Якщо між стовпцями немає систематичної різниці, R_j будуть близькими до середнього рангу, розрахованого за всією сукупністю, який дорівнює:

$$\langle R \rangle = \frac{n+1}{2},$$

де n – загальна кількість одиниць сукупності.

Якщо нульова гіпотеза про рівність групових середніх є правильною, величини $\langle R_j \rangle - \frac{n+1}{2}$ мають бути відносно малими. Значення критерію обчислюють за формулою:

$$H = \frac{12}{n(n+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(n-1).$$

При $n_i \geq 5$ та $k \geq 4$ статистика критерію асимптотично наближається до χ^2 розподілу з кількістю ступенів вільності $k - 1$.

Якщо значення окремих спостережень збігаються, розглянуту схему аналізу можна застосовувати як наближену. Надійність її висновків знижується зі збільшенням кількості збігів. Для підвищення надійності використовують середні ранги (без заокруглення до цілого).

Якщо кількість збігів емпіричних даних велика, використовують модифіковану форму статистики Краскела–Уолліса [36, с. 83]:

$$H' = \frac{H}{1 - \left(\sum_{j=1}^{g} \frac{T_j}{n^3 - n} \right)},$$

де g – кількість груп спостережень, значення яких збігаються; $T_j = t_j^3 - t_j$, де t_j – кількість спостережень, значення яких збігаються у j-й групі.

Реалізація однофакторного дисперсійного аналізу у програмі «Statistica»

Приклад 2.5. За даними результатів досліджень рівня ВВП для 40 країн-членів СОТ (рис. 2.10) дослідити залежність величини загального експорту країни (за 2017 р.) від номінальної змінної ВВП на душу населення, значення якої розподілені за шістьма групами (факторами).

1 = "Вище $102400"
2 = "$51200-102400"
3 = "$25600-51200"
4 = "$6400-25600"
5 = "$1600-6400"
6 = "Нижче $1600"

1 Назва країни	2 ВВП на душу населення	3 Загальний експорт
Данія	3	65,463,453,453.00
Канада	3	4,654,456,654.00
Ісландія	3	65,654,564,654.00
Італія	3	65,456,654,656.00
Люксембург	2	36,545,564,654.00
Нідерланди	3	66,545,664,685.00
Норвегія	2	4,654,656,654.00
Португалія	4	3,654,654,654.00
США	2	6,545,645,665.00
Франція	3	3,465,468,545.00
Греція	4	265,564,654.00
Туреччина	5	3,654,564,654.00
Німеччина	3	29,878,941,322.00
Іспанія	3	3,654,654,654.00
Чехія	4	341,654,654.00

Рис. 2.10. Фрагмент файла даних для прикладу 2.5

Перевіряємо гіпотезу про рівність середніх у кожній з 6-ти груп, що відповідають шести рівням фактора.

Висуваємо H_0: впливу фактора немає.

Досягнутий рівень значущості $p = 0{,}000113$ (рис. 2.11), гіпотезу H_0 приймаємо. Фактор не впливає на досліджувану змінну.

Univariate Tests of Significance for Загальний експор
Sigma-restricted parameterization
Effective hypothesis decomposition

Effect	SS	Degr. of Freedom	MS	F	p
Intercept	234008,5	1	234008,5	3544,068	0,000000
ВВП на душу населення	2102,4	4	525,6	7,960	0,000113
Error	2311,0	35	66,0		

Рис. 2.11. Вікно результатів однофакторного дисперсійного аналізу

***G*-критерій Кокрена** (Кочрена) використовують для перевірки нульової гіпотези про рівність дисперсій k ($k \geq 2$) нормально розподілених генеральних сукупностей за незалежними вибірками рівного обсягу. Значення критерію обчислюють за формулою [36, с. 83]:

$$G = \frac{\max\limits_{1 \leq j \leq k} \sigma_j^2}{\sum\limits_{j=1}^{k} \sigma_j^2},$$

де σ_j^2 – дисперсія j-ї вибірки.

Розглянуті вище критерії дають змогу виявити різницю дисперсій сукупностей, але не дають можливості кількісно оцінити вплив фактора на досліджувану ознаку та встановити, для яких саме сукупностей дисперсії є різними.

Для обчислення кількісного впливу досліджуваного фактора застосовують адитивну модель, яка передбачає, що значення від-

гуку є сумою впливу фактора і незалежної від нього випадкової величини:

$$x_{ij} = a_j + \varepsilon_{ij} \ (j = 1, 2, ..., k; I = 1, 2, ..., n),$$

де a_j – невипадкові невідомі величини, що визначаються значеннями рівнів фактора; ε_{ij} – незалежні випадкові величини, які мають однаковий розподіл і відображають внутрішню мінливість, що не пов'язана із значеннями рівнів фактора.

Якщо при застосуванні однофакторного дисперсійного аналізу гіпотезу про рівність середніх відхиляють, наступним етапом є визначення вибірок, для яких ця різниця суттєва. З цією метою використовують метод лінійних контрастів. **Лінійним контрастом** у моделі адитивного впливу фактора на відгук називають лінійну функцію середніх значень k незалежних нормальних вибірок з невідомими рівними дисперсіями [8, с. 89]

$$L = \sum_{j=1}^{k} c_j m_j,$$

де c_i – відомі сталі, які задовольняють вимогу $\sum_{i=1}^{k} c_i = 0$, m_j – математичне сподівання для j-ї вибірки, яке для оцінювання лінійного контрасту заміняють відповідним середнім арифметичним.

Оцінку дисперсії лінійного контрасту розраховують за формулою:

$$S_L^2 = \sigma_1^2 \sum_{j=1}^{k} \frac{c_j^2}{n_j}.$$

Двофакторний дисперсійний аналіз

Двофакторний дисперсійний аналіз застосовують для взаємопов'язаних вибірок, які підпорядковуються нормальному розподілу. Емпіричні дані представляють у таблиці, заголовками стовпців якої є значення рівнів першого фактора, заголовками рядків – рівні другого фактора (табл. 2.8). Таблиця даних має розмірність $n \times k$, де n і k – кількість рівнів першого та другого факторів відповідно.

У таблиці двофакторного дисперсійного аналізу можлива неоднорідність даних у стовпцях, якщо вплив другого фактора є суттєвим.

Для опису даних використовують адитивну модель, яка передбачає, що значення відгуку – це сума внесків окремо кожного із факторів b_i та t_j і незалежної від факторів випадкової компоненти ε_{ij}:

$$x_{ij} = b_i + t_j + \varepsilon_{ij}.$$

Таблиця 2.8

Загальний вигляд таблиці даних двофакторного аналізу

Рівні фактора 2	Рівні фактора 1			
	1	2	…	*k*
1	x_{11}	x_{12}	…	x_{1k}
2	x_{21}	x_{22}	…	x_{2k}
…	…	…	…	…
n	x_{n1}	x_{n2}	…	x_{nk}

Якщо випадкова компонента ε_{ij} підпорядковується нормальному розподілу з нульовим середнім і рівними для всіх *i, j* дисперсіями σ^2, застосовують двофакторний дисперсійний аналіз (дисперсійний аналіз за двома ознаками).

Нульова гіпотеза може полягати в рівності ефектів стовпців між собою (відсутність впливу першого фактора)

$$H_{01}: \tau_1 = \tau_2 = \ldots = \tau_k = 0$$

або рівності ефектів рядків між собою (відсутність впливу другого фактора)

$$H_{02}: \beta_1 = \beta_2 = \ldots = \beta_n = 0.$$

Для перевірки нульової гіпотези, як і у разі однофакторного дисперсійного аналізу, розраховують дві оцінки дисперсії.

Дисперсія для H_{01} становить:

$$\sigma_1^2 = \frac{1}{(n-1)(k-1)} \sum_{i=1}^{n} \sum_{j=1}^{k} \left(x_{ij} - \langle x_i \rangle - \langle x_j \rangle + \langle x \rangle \right)^2,$$

де $\langle x \rangle = \frac{1}{nk} \sum_{i=1}^{n} \sum_{j=1}^{k} x_{ij}$ – загальне середнє за всіма спостереженнями;

$\langle x_i \rangle = \frac{1}{k} \sum_{j=1}^{k} x_j$ – середнє за *i*-м рядком;

$\langle x_j \rangle = \frac{1}{k}\sum_{i=1}^{n} x_j$ – середнє за j-м стовпцем.

Оцінка дисперсії σ_1^2 є незміщеною (дорівнює її математичному сподіванню) і не залежить від справедливості нульової гіпотези.

Дисперсія для H_{02} становить:

$$\sigma_2^2 = \frac{n}{k-1}\sum_{j=1}^{k}\left(\langle x_j \rangle + \langle x \rangle\right)^2.$$

Вона є незміщеною лише за умови справедливості нульової гіпотези. Чим більша різниця між результатами дії першого фактора, тим більшим є значення σ_2^2.

Для перевірки достовірності гіпотези H_{01} розраховують відношення дисперсій:

$$F = \frac{\sigma_2^2}{\sigma_1^2} = \frac{n(n-1)(k-1)\sum_{j=1}^{k}\left(\langle x_j \rangle - \langle x \rangle\right)^2}{(k-1)\sum_{i=1}^{n}\sum_{j=1}^{k}\left(x_j - \langle x_i \rangle - \langle x_j \rangle + \langle x \rangle\right)^2},$$

яке має F-розподіл Фішера з кількостями ступенів вільності $(k-1)$ і $(n-1)(k-1)$. Нульову гіпотезу приймають на рівні значущості α, якщо $F < F_{1-\alpha}$.

Для перевірки H_{02} використовують формулу:

$$F = \frac{\sigma_2^2}{\sigma_1^2} = \frac{k(k-1)(n-1)\sum_{i=1}^{n}\left(\langle x_i \rangle - \langle x \rangle\right)^2}{(n-1)\sum_{j=1}^{k}\sum_{i=1}^{n}\left(x_j - \langle x_i \rangle - \langle x_j \rangle + \langle x \rangle\right)^2}.$$

Якщо припущення, необхідні для застосування двофакторного дисперсійного аналізу, не виконуються, застосовують непараметричний ранговий критерій Фрідмана (Фрідмана, Кендалла та Сміта), який не накладає обмежень на тип розподілу емпіричних даних. Однак роблять припущення, що величини ε_{ij} є незалежними, однаково розподіленими та неперервними.

При перевірці нульової гіпотези H_{01}: $\tau_1 = \tau_2 = \ldots = \tau_k = 0$ (відсутність впливу першого фактора) вихідні дані представляють у

вигляді прямокутної матриці, в якій n рядків відповідають рівням першого фактора, а k стовпців – рівням другого фактора. У комірках таблиці записують результати вимірювань параметрів одиниць спостережень (груп спостережень) при фіксованих значеннях рівнів обох факторів. Дані представляють у вигляді середніх значень вимірюваного параметра за всіма вимірюваннями або одиницями досліджуваної сукупності.

Для застосування критерію у таблиці вихідних даних абсолютні величини результатів вимірювань заміняють відповідними рангами. Ранжування проводять за кожним рядком окремо – величини x_{ij} впорядковують для кожного фіксованого значення i, отримуючи при цьому k значень відповідних рангів r_{ij}. Це дає можливість усунути вплив другого фактора, значення якого для кожного рядка є однаковими.

Критерій обчислюють за формулою:

$$S = \left[\frac{12}{nk(k+1)} \sum_{j=1}^{k} \left(\sum_{i=1}^{n} r_{ij} \right)^2 \right] - 3n(k+1) .$$

За необхідності перевірити нульову гіпотезу H_{02}: $\beta_1 = \beta_2 = \ldots = \beta_n = 0$ (відсутність впливу другого фактора) вихідні дані ранжують за стовпцями і повторюють описану вище процедуру із заміною n на k і навпаки. При справедливості нульової гіпотези і $n \to \infty$ S-статистика Фрідмана асимптотично наближається до статистики χ^2 з $k-1$ ступенем вільності, тому нульову гіпотезу можна прийняти на рівні значущості α, якщо $S < \chi^2_{1-\alpha}(k-1)$.

Реалізація двофакторного дисперсійного аналізу у програмі «Statistica»

Приклад 2.6. За даними результатів досліджень рівня ВВП для 40 країн-членів СОТ (рис. 2.12) дослідити залежність величини загального експорту країни (за 2017 р.) від ***ВВП на душу населення*** та **Форми державного устрою**. Значення номінальної змінної ВВП на душу населення розподілені за шістьма групами

(як у прикладі 2.5). Фактор **Форма державного устрою** має три градації:

1 = "унітарна"
2 = "федеративна"
3 = "конфедеративна"

3 Загальний експорт	2 ВВП на душу населення	3 Загальний експорт	4 Форма державного устрою
Данія	3	65,463,453,453.00	1
Канада	3	4,654,456,654.00	1
Ісландія	3	65,654,564,654.00	1
Італія	3	65,456,654,656.00	1
Люксембург	2	36,545,564,654.00	1
Нідерланди	3	66,545,664,685.00	2
Норвегія	2	4,654,656,654.00	1
Португалія	4	3,654,654,654.00	1
США	2	6,545,645,665.00	1
Франція	3	3,465,468,545.00	1
Греція	4	265,564,654.00	1
Туреччина	5	3,654,564,654.00	1
Німеччина	3	29,878,941,322.00	1
Іспанія	3	3,654,654,654.00	2

Рис. 2.12. Файл даних для прикладу 2.6

Висуваємо гіпотезу H_0: впливу факторів немає.

У результаті проведення багатофакторного дисперсійного аналізу (рис. 2.13) досягнуто рівень значущості $p_A = 0{,}000468$ – вплив фактора *A* присутній. Для фактора *B* досягнутий рівень значущості $p_B = 0{,}579382 > 0{,}01$, гіпотезу про відсутність впливу фактора *B* на відгук приймаємо – загальний обсяг експорту не залежить від форми державного устрою.

Univariate Tests of Significance for Загальний експор
Sigma-restricted parameterization
Effective hypothesis decomposition

Effect	SS	Degr. of Freedom	MS	F	p
Intercept		0			
ВВП на душу населення	1659,721	3	553,2404	7,899894	0,000468
Форма державного устрою		0			
ВВП на душу населення*Форма державного устрою	139,895	3	46,6317	0,665869	0,579382
Error	2170,972	31	70,0314		

Рис. 2.13. Вікно результатів багатофакторного дисперсійного аналізу

Критерій Пейджа (L-критерій) призначений для перевірки нульової гіпотези H_{01}: $\tau_1 = \tau_2 = \ldots = \tau_k = 0$ або H_{02}: $\beta_1 = \beta_2 = \ldots = \beta_n = 0$, на противагу альтернативі $\tau_1 \le \tau_2 \le \ldots \le \tau_k$ або $\beta_1 \le \beta_2 \le \ldots \le \beta_n$ відповідно, де принаймні одна із нерівностей є строгою. Для впорядкованих альтернативи він є потужнішим за критерій Фрідмана. Значення критерію обчислюють за формулами [36, с. 87]:

$$L_1 = \sum_{j=1}^{k} j r_j \,,\; L_2 = \sum_{i=1}^{n} j r_i \,,$$

де

$$r_j = \sum_{i=1}^{n} r_{ij} \,,\; r_i = \sum_{j=1}^{k} r_{ij} \,.$$

Для великих вибірок застосовують апроксимацію статистики Пейджа:

$$L_1^* = \frac{L - nk(k+1)^2/4}{\sqrt{n\left(k^3 - k\right)^2/144(k-1)}},\; L_2^* = \frac{L - nk(k+1)^2/4}{\sqrt{k\left(n^3 - n\right)^2/144(n-1)}},$$

які за умови справедливості відповідних нульових гіпотез підпорядковуються нормальному розподілу. У разі, коли в рядках вихідної таблиці є однакові значення, необхідно використовувати середні ранги. При цьому точність висновків буде незначною відповідно до того, чим більшою є кількість таких збігів.

***Q*-критерій Кокрена** використовують у випадках, коли групи однорідних об'єктів піддаються впливу більше, ніж двох факторів, а відгук може набувати одного з двох альтернативних варіантів. Нульова гіпотеза полягає в рівності ефектів впливу всіх факторів. Значення критерію розраховують за формулою [36, с. 88]:

$$Q = \frac{(c-1)\left(c\sum_{j=1}^{c} T_j^2 - \left(\sum_{j=1}^{c} T_j\right)^2\right)}{c\sum_{i=1}^{r} T_i - \sum_{i=1}^{r} T_i^2},$$

де $T_j = \sum_{i=1}^{r} x_{ij} (j = 1, 2, \ldots, c)$ – суми стовпців; $T_j = \sum_{i=1}^{r} x_{ij}$ (i = 1, 2, …, r) – суми рядків; c – кількість стовпців (вибірок); r – кількість рядків (факторів).

Довірчий рівень визначається функцією розподілу χ^2 з кількістю ступенів вільності $c - 1$.

Двофакторний дисперсійний аналіз дає можливість визначити існування ефектів впливу, проте не дає змоги встановити, для яких саме стовпців існує цей ефект.

При вирішенні цієї проблеми застосовують метод множинних порівнянь Шеффе для залежних вибірок. Значення критерію розраховують за формулою:

$$t = \frac{\left(\sum_{i=1}^{r} c_i \overline{x_i}\right)^2}{\frac{(r-1)S}{c} \sum_{i=1}^{r} c_i^2},$$

де c_i (i = 1, 2, …, r) – константи, $S = \sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij}^2 - \frac{T^2}{rc}$ – залишковий середній квадрат, $S = \sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij}^2 - \frac{T^2}{rc}$ – загальна сума, c – кількість стовпців (вибірок), r – кількість рядків (факторів).

Довірчий рівень визначається функцією розподілу Фішера з параметрами (r – 1) та (r – 1)(c – 1) при дослідженні ефекту рядків і (c − 1) та (r – 1)(c – 1) при дослідженні ефекту стовпців.

ПИТАННЯ ТА ЗАВДАННЯ ДЛЯ САМОПІДГОТОВКИ ТА САМОКОНТРОЛЮ

Теоретичні запитання

1. Визначення поняття «гіпотеза». Статистична гіпотеза.
2. Нульова та альтернативна гіпотези.
3. Ряди розподілу.
4. Основні характеристики ряду розподілу.
5. Середня величина як узагальнюючий показник ряду розподілу.
6. Середнє арифметичне (просте та зважене).
7. Середнє гармонійне.
8. Середнє геометричне.
9. Середнє квадратичне.
10. Структурні середні: мода та медіана.
11. Закономірність розподілу.
12. Атрибутивні ряди розподілу.
13. Варіаційні ряди.
14. Коефіцієнти варіації: розмах, середнє арифметичне відхилення, квадратичне (стандартне) відхилення, дисперсія.
15. Види дисперсій.
16. Крива розподілу.
17. Емпірична та теоретична функції розподілу.
18. Характеристики форми розподілу: асиметрія та ексцес.
19. Нормальний розподіл.
20. Рівномірний розподіл.
21. Однобічні та двобічні гіпотези.
22. Параметричні та непараметричні тести.
23. Критерії згоди.
24. Рівень статистичної значущості.
25. Критерій Пірсона.
26. *T*-критерій Стьюдента.
27. *F*-критерій Фішера.
28. Однофакторний дисперсійний аналіз.
29. Контраст факторів.

30. Основна тотожність дисперсійного аналізу.
31. Залишкова (внутрішньогрупова) варіація.
32. Факторна (міжгрупова) варіація.
33. Однофакторний дисперсійний аналіз.
34. Двофакторний дисперсійний аналіз.

Практичні завдання

Завдання 1. Обчислити основні описові статистики для значень *Індексу людського розвитку* (*ІЛР*) країн ЄС за 2018 р.

Завдання 2**.** Знайти квантиль розподілу для змінної *ВВП* країн ЄС за 2018 р. Сформувати графік щільності та функції розподілу при $p = 0{,}01$.

Завдання 2. Отримати таблицю частот для значень змінної *Політичний режим.*

Завдання 4. Побудувати діаграму та коробчасту діаграму для вибірки – значень змінної *Місце у світовому рейтингу*.

Завдання 5. Побудувати *P–P* діаграму для значень змінної *Рівень ВВП* та зробити висновок про відповідність нормальному розподілу.

Завдання 6. Побудувати *Q–Q* діаграму для значень змінної *ВНД на душу населення* та зробити висновок про відповідність обраного розподілу досліджуваним даним.

Завдання 7. Перевірити гіпотезу про нормальний розподіл значень змінної *ІЛР 2018 р.* – застосувати критерій згоди χ^2.

Завдання 8. Побудувати таблицю спряженості для зміних *Політичний режим* та *Форма державного устрою.*

Завдання 9. Застосувати критерій незалежності випадкових величин χ^2 для встановлення залежності змінних *Очікувана*

тривалість життя у 2018 р. та *Очікувана тривалість навчання дітей шкільного віку у 2018 р.* для країн ЄС.

Завдання 10. Дослідити залежність змінної *Індекс людського розвитку у 2018 р.* від номінальної змінної *Форма державного правління.*

Завдання 11. Дослідити залежність змінної *Індекс людського розвитку у 2018 р.* від змінних *Членство в ЄС, Членство у НАТО* та *Членство у СОТ*. Пояснити отримані результати.

РОЗДІЛ 3

ПРОГНОЗНИЙ ІНСТРУМЕНТАРІЙ МОДЕЛЮВАННЯ МІЖНАРОДНИХ СОЦІАЛЬНО-ЕКОНОМІЧНИХ ПРОЦЕСІВ

3.1. Парадигма прогнозування

Міжнародні відносини як система міждержавних взаємодій охоплюють соціально-економічні, військово-політичні, дипломатичні, національні, культурні, гуманітарні та інші взаємозв'язки між державами, соціальними угрупуваннями, суспільними організаціями та окремими політичними діячами.

Передбачення наслідків та вибір оптимальної стратегії дій акторів міжнародної арени у кожній конкретній ситуації є головним завданням прогнозування процесів міжнародних відносин. Дослідженню закономірностей прогнозування науковці приділяють належну увагу лише впродовж останніх десятиліть. Це пов'язано зі стрімким розвитком міжнародних соціально-економічних і політичних процесів та ускладненням управління у сфері МВ.

Прогнозування – процес передбачення майбутнього стану предмета чи явища на основі систематичного аналізу інформації про минуле і сучасне, про якісні та кількісні характеристики його

розвитку. Прогноз має випадковий, імовірнісний характер. Його будують на підставі аргументованих наукових уявлень про стан і розвиток об'єкта дослідження. Результатом прогнозування є **прогноз** – науково обґрунтоване судження щодо можливих станів об'єкта в майбутньому, ймовірний розвиток сучасних тенденцій.

Прогнозні методи у МВ – наукові передбачення майбутніх явищ та процесів у міжнародних відносинах.

Метод прогнозування – сукупність прийомів і способів мислення, за допомогою яких на основі ретроспективних даних спостережень та зовнішніх і внутрішніх зв'язків об'єкта прогнозування роблять висновки щодо майбутнього стану його розвитку.

На сьогодні відомо понад 150 різних методів прогнозування, з яких на практиці використовують 15–20.

Методика прогнозування – сукупність методів і правил розробки прогнозів конкретних об'єктів.

Етапи прогнозування [40, с. 11]:

– передпрогнозна орієнтація;
– розробка завдання на прогноз;
– прогнозна ретроспекція;
– прогнозний діагноз;
– прогнозна проспекція;
– верифікація прогнозу;
– коректування прогнозу;
– синтез прогнозів.

Основні функції прогнозування МВ:

► науковий аналіз економічних, соціальних, політичних, гуманітарних, науково-технічних процесів і тенденцій;

► дослідження об'єктивних зв'язків міжнародних соціально-економічних та політичних явищ економіко-політичного розвитку в конкретних умовах;

► оцінювання досягнутого рівня розвитку, дійсної ситуації і виявлення тенденцій, які можуть бути у майбутньому, передбачення нових ситуацій та їх оцінювання;

► виявлення можливих альтернатив розвитку світової економіки та міжнародної політики у перспективі, нагромадження наукового матеріалу для обґрунтованого вибору оптимальних рішень.

Принципи наукового прогнозування:

– варіантність – розробка декількох варіантів прогнозу з врахуванням особливостей висунутої гіпотези, постановки мети (у нормативному прогнозуванні) та можливих варіантів прогнозного середовища;

– верифікативність – перевірка достовірності, точності й обґрунтованості прогнозів;

– неперервність – коректування прогнозів за отримання нових даних про об'єкт прогнозування;

– рентабельність – перевищення економічного ефекту від використання прогнозу над витратами на його розробку;

– системність – узгодженість всіх прогнозованих показників та параметрів прогнозів;

– узгодженість нормативних і пошукових прогнозів різної природи та різних періодів випередження.

За функціональною ознакою (спрямуванням прогнозування) виділяють пошуковий і нормативний прогнози.

Пошуковий прогноз ґрунтується на умовному перенесенні на майбутнє закономірностей розвитку об'єкта дослідження в минулому і сучасному за умов збереження наявних тенденцій.

Нормативний прогноз розробляють на основі заздалегідь визначених цілей. Його завдання – визначити шляхи і терміни досягнення бажаних станів об'єкта прогнозування в майбутньому.

Основні джерела прогнозної інформації [40, с. 23]:

► набутий досвід, що ґрунтується на знанні закономірностей перебігу та розвитку досліджуваних подій, процесів, явищ;

► екстраполяція наявних тенденцій, закони розвитку яких у минулому й сучасному відомі;

► побудова моделей об'єктів прогнозування щодо очікуваних або планованих умов.

Способи прогнозування:

– **експертний** – прогнозування та оцінювання майбутніх результатів дій на основі використання професійного досвіду й інтуїції висококваліфікованих експертів у галузі МВ; застосовують в умовах неповної інформації або за відсутності надійних методів оцінювання подій минулого;

– **інтерполяція** (метод статистичного прогнозування) – знаходження показника (приблизний розрахунок) у середині ряду на основі закономірностей розвитку явища за досліджуваний період;

– **екстраполяція** – вивчення тенденцій минулого і сучасного розвитку об'єкта та перенесення виявлених закономірностей на майбутнє;

– **моделювання** – дослідження пошукових і нормативних моделей прогнозованого об'єкта щодо очікуваних або спланованих змін його стану.

У процесі економіко-політичного прогнозування використовують як загальні наукові, так і специфічні методи та підходи до дослідження [40, с. 24]:

► історичний метод – розгляд окремого явища у взаємозв'язку його історичних форм;

► комплексний метод полягає в розгляді явищ у їх взаємозалежності на основі методів дослідження різних наук, які вивчають ці явища;

► системний метод передбачає дослідження кількісних і якісних закономірностей перебігу ймовірнісних процесів у складних системах;

► структурний метод призначений для виявлення складових та причин досліджуваного явища;

► системно-структурний – метод дослідження системи у розвитку як динамічного цілого, вивчення її структурних елементів і взаємозв'язків між ними.

Питання про істинність прогнозування тісно пов'язане з проблемою критеріїв істинності, які поділяють на дві групи: практичні критерії (практика як критерій істини на всіх стадіях прогнозуван-

ня) і логічні або непрямі критерії (можливість перевірки прогнозів, їх адекватність, логічна несуперечність).

Основні параметри прогнозів:

– достовірність прогнозу;
– джерело помилки прогнозу;
– обґрунтованість прогнозу;
– помилка прогнозу;
– період підстави для прогнозу;
– період випередження;
– прогнозний горизонт;
– точність прогнозу.

Сьогодні при прогнозуванні розвитку міжнародної політики необхідно враховувати не тільки всі нюанси політичних, економічних, національних й інших зв'язків між народами окремих країн, а й діяльність партій, рухів, соціальних груп і навіть окремих осіб, визначати їх вплив на державне життя [57, с. 117].

Планування міжнародних політичних процесів має жорсткий характер. Усі прийняті рішення впливають на окремі елементи міжнародної політики. Від якості прогнозів залежить майбутнє міжнародної політики.

Дослідження сучасної міжнародної економіко-політичної ситуації для прийняття оптимальних рішень щодо майбутніх дій здійснюється у двох взаємопов'язаних напрямах:

► розробка сучасного наукового інструментарію прогнозування конфліктних ситуацій і проведення дієвих заходів щодо їх уникнення чи нейтралізації;

► моделювання нових форм і методів міжнародного співробітництва для забезпечення стабільного мирного співіснування у світі.

Основні групи факторів, що визначають становище держави на світовій арені [69, с. 62]:

– територія, природні умови;
– демографічні характеристики;
– загальні показники економічного розвитку;

– державний і приватний сектори економіки;
– структура промислового виробництва;
– галузеві господарства;
– ступінь розвитку матеріальної інфраструктури;
– внутрішній ринок товарів;
– територіальні аспекти соціально-економічного розвитку;
– виробництво послуг;
– фінанси, кредит, грошовий обіг, цінова політика;
– зовнішньоекономічні відносини;
– ринок праці;
– розвиток науки та техніки;
– проблеми управління приватним капіталом;
– соціальні потреби населення;
– індикатори безпеки і якості життя;
– показники екології;
– політична організація суспільства;
– характеристики політичних лідерів;
– особливості керівного політичного органу;
– внутрішня політика держави;
– реакція суспільства на політику уряду;
– стабільність режиму;
– збройні сили;
– військово-промисловий потенціал;
– зовнішня політика;
– регіональні проблеми;
– міжнародні взаємодії та конфлікти;
– глобальні проблеми сучасності.

На сучасному етапі розвитку суспільства надважливого значення набуло прийняття своєчасних ефективних управлінських рішень у сфері МВ. Суттєво зросла роль відповідно математичних матодів у формуванні достовірних, науково обґрунтованих висновків та прогнозів.

3.2. Базові методики прогнозування міжнародних відносин

Серед інструментів економіко-політичної прогностики важливу роль відіграють економіко-математичні методи, методи економіко-математичного моделювання, статистичної екстраполяції, елімінування, експертних оцінок і т. ін.

Інтерполяція – обчислення значення показника (приблизний розрахунок) у середині ряду на основі закономірності розвитку досліджуваного явища за визначений період. Розрахунок невідомих рівнів ряду проводять на основі середнього абсолютного приросту або середнього темпу зростання. На точність обчислень впливає стабільність показників динаміки – абсолютних приростів та темпів зростання [36, с. 112].

Приклад 3.1. За даними значень загального імпорту України за 2004–2014 рр. визначити невідомий рівень ряду динаміки (табл. 3.1).

Таблиця 3.1

Загальний імпорт України за 2004–2014 рр.

Рівні ряду	Роки										
	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Імпорт, млрд. дол.	36	44	–	72	99	56	73	99	105	101	113
Абсол. приріст, млрд. дол.	–	8	–	–	27	-43	17	26	6	-4	12
Темп зростання, %	100	122	–	–	137	56	130	135	106	96	111

Середньорічний абсолютний приріст за досліджуваний період:

$$\Delta_{сер} = (113 - 36) / 10 = 7{,}7 \text{ млн. дол.},$$

$$y_{2006} = y_{2005} + \Delta_{сер} = 44 + 7{,}7 = 51{,}7 \text{ млн. дол.}$$

Середньорічний темп зростання:

$$\text{Тз}_{\text{сер}} = \sqrt[10]{\frac{113}{36}} = 1{,}12 = 112\%,$$

$$y_{2006} = y_{2006} \times \text{Тз}_{\text{сер}} = 44 \times 1{,}12 = 49{,}28$$

При прогнозуванні рівнів динамічного ряду застосовують один із методів статистичного прогнозування **екстраполяцію тренда**. Цей метод полягає у знаходженні значення показника за межами відомого ряду динаміки. Цей метод передбачає поширення тенденції ряду на минуле (ретроспективна екстраполяція) або майбутнє (прогнозна екстраполяція).

Екстраполяцію на короткотривалий період проводять на основі середнього абсолютного приросту. Для прикладу 3.1 отримаємо:

$$y_{2003} = y_{2004} - \Delta_{сер} = 36 - 7{,}7 = 28{,}3 \text{ млн. дол.},$$

$$y_{2015} = y_{2014} + \Delta_{сер} = 113 + 7{,}7 = 120{,}7 \text{ млн. дол.}$$

або середнього темпу зростання:

$$y_{2003} = y_{2004} \div \text{Тз}_{\text{сер}} = 36 \div 1{,}12 = 32{,}1$$

$$y_{2015} = y_{2014} \times \text{Тз}_{\text{сер}} = 113 \times 1{,}12 = 126{,}56$$

Екстраполяцію на довготривалий період проводять за допомогою рівняння тренда, оскільки тенденції в майбутньому можуть змінитися під впливом нових умов і факторів.

Елімінування

Вибір технічних прийомів прогнозування залежить від характеру зв'язків між явищами. Якщо взаємозв'язок між результатним показником та факторами, що його визначають, має функціональний характер (кожному значенню факторної ознаки відповідає одне значення результатної), в аналітичній практиці використовують прийоми елімінування.

Елімінування – усунення, виключення впливу всіх, окрім одного, факторів на величину результатного показника. Цей прийом базується на припущенні, що всі фактори змінюються незалежно один від одного: спочатку змінюється один, а всі інші залишаються незмінними; потім змінюється другий; потім третій і т. д. за незмінних інших. Це дає змогу визначити вплив кожного фактора на величину досліджуваного показника окремо.

В аналітичній практиці широко застосовують такі прийоми елімінування [7, с. 117]:

– спосіб ланцюгових підстановок;

– спосіб абсолютних різниць;

– спосіб відносних різниць;

– індексний спосіб.

Спосіб ланцюгових підстановок є найбільш універсальним. Його використовують для розрахунку впливу факторів у моделях всіх типів. Цей спосіб полягає у визначенні впливу окремих факторів на зміну величини результатного показника за допомогою поступової заміни базисної величини кожного факторного показника у факторній моделі на фактичну величину у звітному періоді. З цією метою визначають умовні величини результатного показника, які враховують зміну одного, потім двох, трьох факторів, припускаючи, що інші фактори є незмінними.

Порівняння величини результатного показника до та після заміни рівня конкретного показника нейтралізує (елімінує) вплив усіх факторів, окрім одного, дає можливість визначити його індивідуальний вплив на приріст результатного показника. При цьому спочатку підлягають заміні кількісні параметри, далі – структурні, після них – якісні. Якщо у модель входить багато кількісних, структурних або якісних показників, основні показники заміняють перед похідними, первинні – перед вторинними.

Розглянемо залежність результатного показника від трьох факторів: $Y = X_1X_2X_3$, причому Y^0 та Y^1 – його базисний і звітний рівні відповідно:

$$Y^0 = X_1^0 X_2^0 X_3^0,$$

$$Y^1 = X_1^1 X_2^1 X_3^1.$$

Послідовно замінюємо базисні параметри, що входять у модель, на звітні та розраховуємо умовні результатні показники:

– 1-й умовний показник (перша заміна):

$$Y_{\text{ум1}} = X_1^1 X_2^0 X_3^0;$$

– 2-й умовний показник (друга заміна):

$$Y_{\text{ум2}} = X_1^1 X_2^1 X_3^0;$$

– 3-й показник (третя заміна):

$$Y_{\text{ум3}} = X_1^1 X_2^1 X_3^1.$$

Вплив факторів на відхилення значення показника Y^1 від значення показника Y^0:

► вплив фактора X^1 на зміну Y:

$$\Delta Y_{X_1} = X_1^1 X_2^0 X_3^0 - X_1^0 X_2^0 X_3^0;$$

► вплив фактора X_2 на зміну Y:

$$\Delta Y_{X_2} = X_1^1 X_2^1 X_3^0 - X_1^1 X_2^0 X_3^0;$$

► вплив фактора X_3 на зміну Y:

$$\Delta Y_{X_3} = X_1^1 X_2^1 X_3^1 - X_1^1 X_2^1 X_3^0;$$

$$\Delta Y = \Delta Y_{X_1} + \Delta Y_{X_2} + \Delta Y_{X_3}.$$

Використання способу ланцюгових підстановок передбачає знання взаємозв'язків факторів, уміння правильно їх класифікувати та систематизувати.

Приклад 3.2. Способом ланцюгових підстановок розрахувати вплив факторів на величину результатного показника $Y = X_1X_2X_3$, якщо для базисного рівня цього показника Y^0: $X_1^0 = 2,\ X_2^0 = 0{,}5,\ X_3^0 = 4$; для звітного рівня Y^1: $X_1^1 = 4,\ X_2^1 = 0{,}25,\ X_3^1 = 6$.

$$\Delta Y_{X_1} = X_1^1 X_2^0 X_3^0 - X_1^0 X_2^0 X_3^0 = 4 \cdot 0{,}5 \cdot 4 - 2 \cdot 0{,}5 \cdot 4 = 8 - 4 = 4;$$

$$\Delta Y_{X_2} = X_1^1 X_2^1 X_3^0 - X_1^1 X_2^0 X_3^0 = 4 \cdot 0{,}25 \cdot 4 - 4 \cdot 0{,}5 \cdot 4 = 4 - 8 = -4;$$

$$\Delta Y_{X_3} = X_1^1 X_2^1 X_3^1 - X_1^1 X_2^1 X_3^0 = 4 \cdot 0{,}25 \cdot 6 - 4 \cdot 0{,}25 \cdot 4 = 6 - 4 = 2;$$

$$\Delta Y = \Delta Y_{X_1} + \Delta Y_{X_2} + \Delta Y_{X_3} = 2.$$

Спосіб абсолютних різниць є спрощеним варіантом способу ланцюгових підстановок. Його використовують для розрахунку впливу факторів на приріст результатного показника у моделях, в яких результатний показник представлений у вигляді добутку факторів $Y = X_1X_2$, або в змішаних моделях типу $Y = (X_1 - X_2)X_3$.

Якщо результатний показник розраховують як частку від ділення факторів або він представлений залежністю факторів, використовують тільки спосіб ланцюгових підстановок.

Особливо ефективним спосіб абсолютних різниць є в тому разі, коли вихідні дані містять абсолютні відхилення за факторними показниками. Алгоритм розрахунків за допомогою способу ланцюгових підстановок:

$$\Delta Y_{X_1} = (X_1^1 - X_1^0)X_2^0 X_3^0;$$
$$\Delta Y_{X_2} = X_1^1(X_2^1 - X_2^0)X_3^0;$$
$$\Delta Y_{X_3} = X_1^1 X_2^1(X_3^1 - X_3^0);$$
$$\Delta Y = \Delta Y_{X_1} + \Delta Y_{X_2} + \Delta Y_{X_3}.$$

Приклад 3.3. Використовуючи спосіб абсолютних різниць, розрахувати вплив факторів на величину результатного показника $Y = X_1X_2X_3$, якщо для базисного рівня цього показника Y^0: $X_1^0 = 2,\ X_2^0 = 0{,}5,\ X_3^0 = 4$; для звітного рівня Y^1: $X_1^1 = 4,\ X_2^1 = 0{,}25,\ X_3^1 = 6$.

$$\Delta Y_{X_1} = (X_1^1 - X_1^0)X_2^0 X_3^0 = (4-2)\cdot 0{,}5\cdot 4 = 4;$$
$$\Delta Y_{X_2} = X_1^1(X_2^1 - X_2^0)X_3^0 = 4\cdot(0{,}25-0{,}5)\cdot 4 = -4;$$
$$\Delta Y_{X_3} = X_1^1 X_2^1(X_3^1 - X_3^0) = 4\cdot 0{,}25\cdot(6-4) = 2;$$
$$\Delta Y = \Delta Y_{X_1} + \Delta Y_{X_2} + \Delta Y_{X_3} = 4 - 4 + 2 = 2.$$

Спосіб відносних різниць використовують для вимірювання впливу факторів на результатний показник тільки в моделях, де результатний показник представлений у вигляді добутку факторів $Y = X_1X_2X_3$ та в комбінованих моделях $Y = (X_1 - X_2)X_3$.

Розрахунки дії факторів на досліджуваний показник проводять на основі відносних показників їх зміни, виражених у відсотках або коефіцієнтах.

Для визначення впливу факторів за допомогою способу відносних різниць для моделі типу $Y = X_1X_2X_3$ спочатку необхідно розрахувати відносні відхилення факторних показників за формулою:

$$\Delta X\% = \frac{X^1 - X^0}{X^0}100\%.$$

Для визначення впливу першого фактора на результатний показник необхідно базисне значення результатного показника помножити на відносне відхилення першого фактора, виражене у відсотках, і поділити на 100. Для спрощення розрахунків відносне відхилення факторів доцільно розраховувати в коефіцієнтах:

$$\Delta Y_{X_1} = \frac{Y^0 \cdot \Delta X_1 \%}{100}.$$

Вплив другого фактора визначають множенням базисного значення результатного показника, скоригованого на дію першого фактора, на відносне відхилення другого фактора, виражене у відсотках. Отриманий результат ділять на 100:

$$\Delta Y_{X_2} = \frac{(Y^0 + \Delta Y_{X_1})\Delta X_2 \%}{100}.$$

Вплив третього фактора (і всіх наступних) визначають аналогічно: базисне значення результатного показника коригують на результат дії першого та другого факторів. Отриманий результат множать на відносне відхилення третього фактора:

$$\Delta Y_{X_3} = \frac{(Y^0 + \Delta Y_{X_1} + \Delta Y_{X_2})\Delta X_3 \%}{100}.$$

Приклад 3.4. Використовуючи спосіб відносних різниць, розрахувати вплив факторів на величину результатного показника $Y = X_1X_2X_3$, якщо для базисного рівня цього показника Y^0: $X_1^0 = 2,\ X_2^0 = 0{,}5,\ X_3^0 = 4$; для звітного рівня Y^1: $X_1^1 = 4,\ X_2^1 = 0{,}25,\ X_3^1 = 6$.

$$Y^0 = X_1^0 X_2^0 X_3^0 = 2 \cdot 0{,}5 \cdot 4 = 4;$$

$$\Delta X_1 \% = \frac{X_1^1 - X_1^0}{X_1^0} 100\% = \frac{4-2}{2} 100\% = 100\%;$$

$$\Delta X_2 \% = \frac{X_2^1 - X_2^0}{X_2^0} 100\% = \frac{0{,}25-0{,}5}{0{,}5} 100\% = -50\%;$$

$$\Delta X_3 \% = \frac{X_3^1 - X_3^0}{X_3^0} 100\% = \frac{6-4}{4} 100\% = 50\%;$$

$$\Delta Y_{X_1} = \frac{Y^0 \cdot \Delta X_1 \%}{100} = \frac{4 \cdot 100\%}{100} = 4\%;$$

$$\Delta Y_{X_2} = \frac{(Y^0 + \Delta Y_{X_1})\Delta X_2 \%}{100} = \frac{(4+4)\cdot(-50\%)}{100} = -4\%;$$

$$Y_{X_3} = \frac{(Y^0 + \Delta Y_{X_1} + \Delta Y_{X_2})\Delta X_3 \%}{100} = \frac{(4+4-4)50\%}{100} = 2\%;$$

$$\Delta Y = \Delta Y_{X_1} + \Delta Y_{X_2} + \Delta Y_{X_3} = 4 - 4 + 2 = 2\%.$$

Спосіб відносних різниць використовують у разі, коли потрібно розрахувати вплив великої кількості факторів (5–10). На відміну від попередніх способів, у цьому разі значно скорочується кількість розрахунків.

Результати обчислень, проведених за всіма розглянутими способами елімінування, є однаковими. Незначні відмінності можуть бути зумовлені неточністю розрахунків відносних відхилень, виражених у коефіцієнтах або відсотках. З огляду на це рекомендується обчислювати їх з точністю до чотирьох знаків для відхилень у коефіцієнтах або двох знаків для відхилень у відсотках.

Методи експертних оцінок

Прогнозування є важливим етапом аналізу результатів дослідження та початковим етапом планування. Воно охоплює попередній і кінцевий (формальний) прогнози, для яких розробляють один або декілька можливих сценаріїв дій.

У випадках, коли статистичні дані відсутні чи є неповними для кількісного оцінювання явищ, для яких не існує традиційних способів вимірювання, для прогнозування майбутніх подій використовують **методи експертних оцінок** [15, с. 140].

Експерти, базуючись на власному досвіді та аналізі виділених факторів, формують судження про вплив факторів на подію, оцінюючи ймовірності їх дії на результатну ознаку об’єкта дослідження. При вирішенні проблем в умовах невизначеності думка групи експертів дає більш надійні результати, ніж думка одного експерта.

Отримані експертні оцінки аналізують та оцінюють їх достовірність. Оцінку результатів експертного оцінювання проводять за коефіцієнтом конкордації, який показує ступінь згоди думок експертів. Найбільш достовірні оцінки отримують за умов узгодженості думок експертів.

Коефіцієнт конкордації W розраховують за формулою [36, с. 119]:

$$W = \frac{12}{m^2\left(n^3 - n\right)} \sum_{j=1}^{n} \left(\sum_{i=1}^{m} \left(R_{ij} - \frac{n+1}{2} \right) \right)^2,$$

де n – кількість об'єктів оцінювання, m – кількість експертів, R_{ij} – ранг j-го об'єкта, представленого i-м експертом.

Якщо об'єкти оцінювання мають однакові ранги, коефіцієнт конкордації розраховують за формулою:

$$W = \frac{12}{\frac{1}{12} m^2\left(n^3 - n\right) - m\sum_{j=1}^{m} T_j} \sum_{j=1}^{n} \left(\sum_{i=1}^{m} \left(R_{ij} - \frac{n+1}{2} \right) \right)^2,$$

де T_j обчислюють за формулою:

$$T_j = \frac{\sum_{i=1}^{L_i} \left(n_i^2 - n_i \right)}{12},$$

де L_i – кількість груп однакових рангів, n_i – кількість елементів i-ї групи для j-го експерта.

Статистичну значущість коефіцієнта конкордації перевіряють порівнянням величини $n(m-1)\cdot W$ з табличним значенням розподілу χ^2 при рівні значущості $\alpha = 0{,}001$ та $n - 1$ ступенях свободи.

Якщо коефіцієнт конкордації є незначущим, використовують методику виведення з дослідження експерта, думка якого не узгоджується з думкою інших експертів. Для цього будують матрицю коефіцієнтів кореляції (взаємозв'язків) Пірсона $r(k, i)$ або рангових коефіцієнтів кореляції Спірмена $r_S(k, i)$ (п. 3.3.1) та виявляють експерта, оцінка якого задовольняє умову:

$$r_j(k,i) = \min_{i=1,\ldots,m} \left\{ r(k,i) \right\}.$$

Це означає, що думка цього експерта найменше узгоджується з думками інших експертів. Його оцінки у подальших розрахунках не враховують.

Алгоритм повторюють, поки коефіцієнт конкордації не буде значущим.

Приклад 3.5. Група з трьох експертів оцінила можливі виграші кожної з чотирьох запропонованих стратегій (табл. 3.2). Перевірити ступінь узгодженості думок експертів.

Таблиця 3.2

Умовні дані для прикладу 3.5

Експерти	Виграші стратегій			
	Стратегія 1	Стратегія 2	Стратегія 3	Стратегія 4
I	7	6	3	5
II	5	9	4	12
III	6	9	4	7

Для перевірки за коефіцієнтом конкордації знайдемо ранги стратегій за оцінками кожного з експертів (табл. 3.3).

Таблиця 3.3

Ранги стратегій за оцінками кожного з експертів

Експерти	Ранги стратегій			
	Стратегія 1	Стратегія 2	Стратегія 3	Стратегія 4
I	4	3	1	2
II	2	3	1	4
III	2	4	1	3

У групах рангів оцінок окремих експертів немає однакових, тому коефіцієнт конкордації розрахуємо за першою формулою.

Для прикладу 3.5 кількість стратегій $n = 4$, кількість експертів $m = 3$.

$$\frac{n+1}{2} = \frac{4+1}{2} = 2{,}5\,,\ \frac{12}{m^2\left(n^3-n\right)} = \frac{12}{3^2\left(4^3-4\right)} = \frac{12}{540} \approx 0{,}022\,.$$

Результати подальших обчислень представлено в табл. 3.4.

Таблиця 3.4

Результати розрахунків

Розрахункові формули	Отримані результати			
	Стратегія 1	Стратегія 2	Стратегія 3	Стратегія 4
$R_j - \frac{n+1}{2}$	1,5	0,5	–1,5	–0,5
	–0,5	0,5	–1,5	1,5
	–0,5	1,5	–1,5	0,5
$\sum_{i=1}^{m}\left(R_{ij} - \frac{n+1}{2}\right)$	0,5	2,5	–4,5	1,5
$\left(\sum_{i=1}^{m}\left(R_{ij} - \frac{n+1}{2}\right)\right)^2$	0,25	6,25	20,25	2,25
$\sum_{j=1}^{n}\left(\sum_{i=1}^{m}\left(R_{ij} - \frac{n+1}{2}\right)\right)^2$	29			

Перевіримо значущість коефіцієнта конкордації:

$$\frac{n+1}{2} = \frac{4+1}{2} = 2{,}5\,,\ \frac{12}{m^2\left(n^3-n\right)} = \frac{12}{3^2\left(4^3-4\right)} = \frac{12}{540} \approx 0{,}022\,.$$

$$n\times(m-1)\times W = 4(3-1)\times 0{,}638 = 5{,}104.$$

Критичне значення χ^2 (0,001; 4 − 1) = 16,27. Оскільки величина $n\times(m-1))\times W$ менша за табличне значення χ^2, коефіцієнт конкордації не є значущим, а думки експертів неузгоджені.

Виокремимо експерта, оцінки якого є найбільш неузгодженими. Для цього побудуємо матрицю парних коефіцієнтів кореляції Пірсона (табл. 3.5).

Таблиця 3.5

Матриця парних коефіцієнтів кореляції Пірсона

Експерти	I	II	III
I	1		
II	0,23035	1	
III	0,657143	0,797366	1

Найменшим є значення коефіцієнта кореляції, який показує узгодженість думок першого та другого експертів, тому одного з

них необхідно виключити з експертизи. Доцільно вивести першого експерта, тому що його оцінки є менш узгодженими з оцінками третього експерта.

Розрахуємо коефіцієнт конкордації без врахування оцінок першого експерта (табл. 3.6):

$$\frac{n+1}{2} = \frac{4+1}{2} = 2{,}5\,,\ \frac{12}{m^2\left(n^3-n\right)} = \frac{12}{2^2\left(4^3-4\right)} = \frac{12}{240} = 0{,}05\,.$$

Коефіцієнт конкордації:

$$W = \frac{12}{m^2\left(n^3-n\right)}\sum_{j=1}^{n}\left(\sum_{i=1}^{m}\left(R_{ij}-\frac{n+1}{2}\right)\right)^2 = 0{,}05\cdot 18 = 0{,}9\,.$$

Таблиця 3.6

Результати розрахунків для двох експертів

Розрахункові формули	Отримані результати			
	Стратегія 1	Стратегія 2	Стратегія 3	Стратегія 4
$R_j - \frac{n+1}{2}$	–0,5	0,5	–1,5	1,5
	–0,5	1,5	–1,5	0,5
$\sum_{i=1}^{m}\left(R_{ij}-\frac{n+1}{2}\right)$	–1	2	-3	2
$\left(\sum_{i=1}^{m}\left(R_{ij}-\frac{n+1}{2}\right)\right)^2$	1	4	9	4
$\sum_{j=1}^{n}\left(\sum_{i=1}^{m}\left(R_{ij}-\frac{n+1}{2}\right)\right)^2$	18			

Значення коефіцієнта конкордації без врахування думки першого експерта значно збільшилося, однак теж не є значущим. Це пов'язано не з тим, що думки експертів не узгоджуються, а з тим, що кількість експертів надто мала. Виграш стратегій розраховують як середнє арифметичне експертних оцінок.

Коефіцієнт компетенції

Використання коефіцієнта конкордації ґрунтується на припущенні, що чим більш узгоджені думки експертів, тим достовірнішими є їх оцінки. Однак практика показує, що це не завжди правильно, і експерт, який не згоден з думкою більшості, може дати найточніші оцінки.

Щоб врахувати думки всіх експертів, обробку результатів експертного оцінювання потрібно виконувати за допомогою коефіцієнта компетентності експерта.

Цей метод базується на використанні попередньої оцінки компетентності експертів, які беруть участь у дослідженні. Оцінювання експертів проводять за критеріями компетентності, серед яких можуть бути рівень освіти, загальний стаж роботи, стаж роботи за проблемою дослідження, посада і т. ін.

Крім того, важливим критерієм є оцінка надійності експерта, яку розраховують як відношення кількості його правильних оцінок до кількості всіх проведених експертиз. Правильними вважаються ті оцінки, які з часом підтвердила практика.

При розрахунку коефіцієнтів компетентності експертів необхідно використовувати єдину для всіх критеріїв шкалу оцінювання.

Коефіцієнт компетентності розраховують за формулою [36, с. 124]:

$$KK_i = \frac{\sum_{j=1}^{m} k_{ij}}{\sum_{i=1}^{n}\sum_{i=1}^{m} k_{ij}},$$

де n – кількість експертів, m – кількість критеріїв оцінювання експертів, k_{ij} – бал, отриманий i-м експертом за j-м критерієм.

Приклад 3.6. За вихідними даними прикладу 3.5 знайти виграш стратегій з урахуванням коефіцієнта компетентності експертів. Оцінювання проведено за трьохбальною шкалою (табл. 3.7).

Таблиця 3.7

Коефіцієнти компетентності експертів

Експерти	Виграші стратегій			Суми балів кожного з експертів
	Критерій 1	Критерій 2	Критерій 3	
I	1	2	2	5
II	2	3	3	8
III	2	1	1	4
Загальна сума балів				17

Для I експерта $KK_1 = \frac{\sum_{j=1}^{m} k_{ij}}{\sum_{i=1}^{n}\sum_{i=1}^{m} k_{ij}} = \frac{5}{17} \approx 0{,}2941$, для другого і третього відповідно $KK_2 = \frac{8}{17} \approx 0{,}4706$ та $KK_3 = \frac{4}{17} \approx 0{,}2353$.

Розрахуємо виграші стратегій з урахуванням коефіцієнта компетентності експертів:

$$v_j = \boldsymbol{K}_{\ i} \cdot v_j \ , \ j = \overline{1,m} \ ,$$

де v_i – виграш і-ї стратегії, v_{ij} – оцінка виграшу *i*-ї стратегії *j*-м експертом (табл. 3.8).

Таблиця 3.8

Результати аналізу узгодженості експертних висновків з урахуванням коефіцієнта компетентності експертів

Експерти	KK_i	Отримані результати			
		Стратегія 1	Стратегія 2	Стратегія 3	Стратегія 4
I	0,2941	7	6	3	5
II	0,4706	5	9	4	12
III	0,2353	6	9	4	7
Виграш з урахуванням KK_i		5,8235	8,1177	3,7059	8,7648

Оптимальною є стратегія 4.

3.3. Методи дослідження зв'язків між факторами у МВ

Емпіричні дані, отримані при аналізі явищ та процесів міжнародних відносин, є кількісними характеристиками, сформованими під впливом множини факторів, більшість з яких можна корегувати. Неконтрольовані фактори можуть набувати випадкових значень та зумовлювати випадковість отриманих результатів. З огляду на це вивчення об'єктивних зв'язків між факторами та результатом є одним з найважливіших завдань емпіричних досліджень МВ.

Фактор – причина та умови, за яких досліджують закономірності зв'язку. **Факторні (незалежні)** ознаки характеризують фактори (зумовлюють зміни інших, пов'язаних із ними ознак). **Результатні ознаки (відгуки)** характеризують наслідки (змінюються під дією факторних ознак).

Залежність між ознаками може проявлятися у функціональній або стохастичній формі [72, с. 57].

Функціональний вид зв'язку характеризує однозначна відповідність між зміною факторної ознаки та зміною величини результату: кожному можливому значенню факторної ознаки x відповідає єдине однозначно визначене значення результатної ознаки y. Функціональну залежність між відгуком та факторними ознаками можна представити в аналітичному вигляді $y_{сер} = f(x)$.

Такі залежності характерні переважно для систем, які вивчають природничі науки. Для дослідження суспільних явищ їх застосовують рідко. Зокрема, при дослідженні зв'язків між елементами розрахункових формул економічних показників використовують адитивну ($a + b + c$) та мультиплікативну ($a = bc$, $c = a/b$) моделі.

Стохастичний (статистичний) вид зв'язку передбачає, що між зміною факторної та результатної ознак немає однозначної відповідності: одному й тому самому значенню фактора відповідає множина значень результатної ознаки. Це спричинено впливом

на значення ознаки значної кількості різнорідних факторів. Вплив окремих із них проявляється лише загалом при масовому спостереженні даних експериментальних досліджень.

Залежності між явищами суспільного життя формуються під впливом багатьох різнорідних і взаємозалежних факторів, тому є найбільш складними.

Різновидом стохастичного зв'язку є **кореляційний зв'язок**, при якому зі зміною факторної ознаки змінюється середнє значення відгуку. Термін «кореляція» означає співвідношення, відповідність. Цей вид зв'язку найчастіше спостерігається при вивченні суспільних явищ. Для таких досліджень властиво, що, крім впливу істотних факторів, які формують рівень результатної ознаки, проявляється дія багатьох інших, неврахованих і випадкових факторів.

На відміну від функціональної залежності, кореляційний зв'язок є неповним (визначають лише співвідношення між фактором і відгуком). Це пояснюється тим, що $y_{сер}$ залежить не лише від аргументу x, а й піддається впливу інших факторів.

Кореляційні зв'язки проявляються тільки у масових явищах. З їх допомогою встановлюють тенденцію змін результатної ознаки при зміні факторної величини. Кореляційну залежність можна встановлювати для двох показників (парна кореляція) або для декількох показників (множинна кореляція).

Методи, які використовують для виявлення кореляційного зв'язку:

– порівняння паралельних рядів даних;

– аналітичне групування (побудова групових та кореляційних таблиць);

– графічне зображення кореляційного поля;

– дисперсійний і кореляційно-регресійний аналіз.

Порівняння паралельних рядів є найпростішим прийомом і передбачає зіставлення ряду значень факторної ознаки та ряду відповідних значень результатної ознаки. Значення фактора ранжують і визначають співвідношення та напрямок зміни величини результатної ознаки.

Приклад 3.7. За умовними даними (табл. 3.9) про витрати ТНК на рекламу (факторна ознака) та кількість продажів (результатна ознака) встановити наявність кореляційного зв'язку.

Таблиця 3.9

Умовні дані продажів

№ ТНК	1	2	3	4	5	6	7	8	9	10
Рекламні витрати (ум. грош. од.)	13	13	13	14	14	14	14	15	15	15
К-сть продажів	100	120	116	120	130	143	117	150	140	160

Загалом для всієї сукупності зростання витрат ТНК на рекламу приводить до збільшення кількості продажів, хоча в окремих випадках така закономірність не спостерігається (наприклад, при зіставленні даних ТНК із номерами 2 та 3 або 8 та 9). Цей факт можна пояснити наявністю інших факторів впливу.

Прямим кореляційним зв'язком називають закономірність розвитку явища, за якої зростання величини факторної ознаки приводить до збільшення величини результатної ознаки. Наприклад, зі зниженням курсу гривні спостерігається зменшення купівельної спроможності українців.

Зворотний кореляційний зв'язок між ознаками припускають у випадках, коли зі збільшенням факторної ознаки величина відгуку має тенденцію до зменшення. Наприклад, зі збільшенням курсу долара спостерігається тенденція до зниження продажів.

Наявність великої кількості значень результатної ознаки, які відповідають одному й тому самому значенню ознаки-фактора, ускладнює сприйняття паралельних рядів, особливо при великій кількості одиниць досліджуваної сукупності. У таких випадках для встановлення факту наявності кореляційного зв'язку між ознаками використовують статистичні таблиці.

Аналітичне групування є одним із найважливіших методів дослідження взаємозв'язків. Усі спостереження поділяють на групи за величиною факторної ознаки і для кожної з груп обчислюють середні значення результатної ознаки – будують групові статистичні

таблиці та порівнюють зміни середніх, виявляють характер зв'язку. У прикладі 3.7 факторна ознака представлена трьома варіантами повторюваних значень. Результатом групування є табл. 3.10.

Таблиця 3.10

Групова статистична таблиця для прикладу 3.7

Групи ТНК (за рекламними витратами)	Кількість ТНК у групі	Середня кількість продажів ТНК у групі
13	3	112
14	4	127,5
15	3	150

Порівнюючи середні значення результатної ознаки за групами, можна припустити наявність прямого кореляційного зв'язку між досліджуваними ознаками: зі збільшенням витрат ТНК на рекламу можна очікувати зростання продажів.

Ще одним із найбільш поширених методів виявлення зв'язку між ознаками є використання **кореляційних таблиць** (таблиць співзалежності, спряженості двох ознак). У кореляційних таблицях значення факторної ознаки розташовують у рядках, а результатної – у стовпцях. На перетині рядків і стовпців такої таблиці записують частоту повторення відповідної комбінації значень *x* та *y* (рис. 3.1). Для кожного варіанта розподілу фактора розраховують середнє значення результатної ознаки. Середні порівнюють та роблять висновок про залежність змінних.

Таблица сопряженности Форма_державного_устрою * Членство_у_СОТ

Частота

		Членство_у_СОТ		Итого
		так	ні	
Форма_державного_устрою	унітарна	7	15	22
	федеративна	8	7	15
Итого		15	22	37

Рис. 3.1. Приклад кореляційної таблиці

Якщо частоти, розташовані на діагоналі матриці спряженості, впорядковані за зростанням з лівого верхнього до правого нижнього кута (більшим значенням фактора відповідають більші значення результату), роблять припущення про наявність прямого кореляційного зв'язку. Якщо величина частот збільшується з правого верхнього до лівого нижнього кута матриці (більшим значенням фактора відповідають менші значення результату), припускають наявність зворотного зв'язку між ознаками.

Графічно взаємозв'язок двох ознак зображають за допомогою поля кореляції. В прямокутній системі координат на осі абсцис відкладають значення факторної ознаки, а на осі ординат – результатної й отримують точковий графік, який називають **«поле кореляції»** (рис. 3.2).

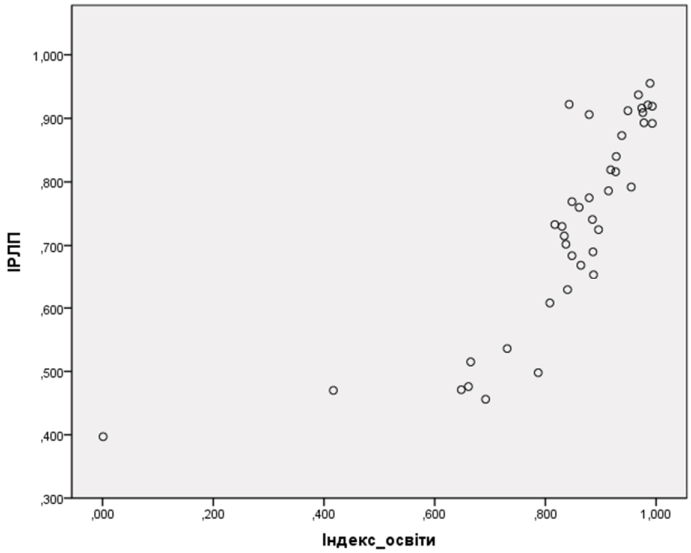


Рис. 3.2. Приклад поля кореляції

За розміщенням точок на графіку роблять висновок про напрям і силу зв'язку:

► якщо точки розташовані хаотично, припускають, що щільний зв'язок між ознаками відсутній;

► якщо точки сконцентровані навколо діагоналі від нижнього лівого кута координат до верхнього правого, йдеться про щільний прямий зв'язок;

► якщо точки розташовані навколо діагоналі від верхнього лівого кута координат до правого нижнього, припускають зворотний зв'язок між досліджуваними ознаками.

Якщо на поле кореляції нанести середні значення результатної ознаки і з'єднати їх відрізками, отримаємо емпіричну лінію зв'язку, яка відображає форму взаємозалежності між ознаками. У разі, коли отримана лінія є прямою, роблять висновок про рівномірну зміну залежних ознак. Якщо емпірична лінія зв'язку є кривою, спостерігається нерівномірна (криволінійна) залежність.

Розглянуті прийоми характеризують лише загальні риси зв'язку, його тенденцію. Для встановлення впливів окремо кожного з факторів та визначення щільності зв'язків між ознаками застосовують методи кореляційно-регресійного та дисперсійного аналізу (п. 2.3).

Прогнозний інструментарій моделювання традиційно використовує методи кореляційно-регресійного аналізу. Макроекономічні моделі дають можливість прогнозувати всі найважливіші показники розвитку МВ. Зовнішньоекономічний прогноз визначає результати експорту та імпорту, надає інформацію про можливі політичні відносини з іншими країнами.

Кореляційний аналіз – метод дослідження стохастичної взаємозалежності між випадковими величинами. Найпростішим випадком є аналіз двох вибірок з генеральної сукупності.

Завдання **кореляційного** аналізу – кількісна оцінка тісноти (сили, щільності) статистичного зв'язку між двома ознаками. Він не встановлює причин залежності між досліджуваними ознаками, а лише виявляє наявність залежності, її силу та напрям [56, с. 327].

Регресійний аналіз полягає у визначенні аналітичного виразу кореляційного зв'язку – описі вигляду та параметрів функції зв'язку (регресійної моделі). Термін «регресія» означає повернення до середньої.

За кількістю факторних ознак, які входять у регресійну модель, виділяють одно- та багатофакторні моделі.

Застосування кореляційного аналізу щільно пов'язане з регресійним аналізом, тому його часто називають кореляційно-регресійним.

Вимоги до застосування кореляційно-регресійного аналізу:

– однорідність одиниць досліджуваної сукупності та їх підпорядкованість нормальному закону розподілу;

– достатня кількість спостережень;

– незалежність між собою обраних для дослідження факторів.

3.3.1. Моделі лінійної регресії

Кореляційно-регресійний аналіз базується на припущенні про те, що залежність між значенням факторної ознаки x і середнім значенням результатної $y_{сер}$ можна представити у вигляді функції (лінії регресії):

$$y_{сер} = f(x),$$

яку називають рівнянням простої парної регресії – **однофакторною регресійною моделлю**.

Якщо значення залежних ознак змінюються приблизно рівномірно (емпірична лінія зв'язку, або лінія групових середніх наближається до прямої), зв'язок між факторами можна описати за допомогою лінійної функції:

$$y_{сер} = a + bx,$$

де a і b – параметри лінійного регресійного рівняння.

Параметр b **(коефіцієнт регресії)** визначає ефект впливу x на $y_{сер}$. Він показує, на скільки одиниць у середньому змінюється результатна ознака $y_{сер}$ зі зміною факторної ознаки x на одну одиницю виміру. При прямому зв'язку між залежними ознаками b – величина додатна, при оберненому – від'ємна.

Параметр a **(вільний член рівняння регресії)** – це значення $y_{сер}$ при $x = 0$. Якщо межі варіації не містять нуля, цей параметр має лише розрахункове значення – показує усереднений вплив на результатну ознаку неврахованих або невиділених для дослідження факторів.

Рівняння регресії відображає закономірності зв'язку між x та $y_{сер}$ для сукупності загалом, а не для окремих її елементів. Вплив інших факторів нівелюється за принципом «за інших однакових умов» [72, с. 67].

На практиці дослідження МВ проводять за великою кількістю спостережень. Вихідні дані представляють у зведеній кореляційній таблиці.

Рівняння регресії застосовують для прогнозування показників, тому важливе значення має оцінювання значущості параметрів регресійного рівняння. Значущість коефіцієнта регресії b оцінюють за допомогою t-критерію Стьюдента – фактичні дані підставляють у формулу критерію і обчислюють його розрахункове значення (п. 2.2). Обчислене за емпіричними даними значення критерію порівнюють з критичним (табличним) значенням. Якщо $t_{емп} > t_{табл}$, коефіцієнт регресії b вважають значущим із заданою ймовірністю.

Для встановлення надійності прогнозування середніх значень результатної ознаки за емпіричними значеннями факторної ознаки оцінюють адекватність регресійної моделі. Для оцінювання надійності застосовують F-критерій Фішера і метод перевірки гіпотез (п. 2.2). Якщо фактичне значення критерію перевищує критичне, то зв'язок між ознаками не випадковий. Якщо $F_{емп} > F_{табл}$, гіпотезу про значущість рівняння приймають.

Для розрахунку теоретичних прогнозних значень результатної ознаки в отримане рівняння регресії підставляють конкретні значення факторної ознаки. Прогноз показника $y_{сер}$ розраховують за умов збереження загальної тенденції розвитку явища на майбутнє за емпіричними даними минулого періоду. Правильність розрахунку перевіряють рівністю сумарних теоретичних та емпіричних значень результатної ознаки при їх зіставленні.

Для прогнозних моделей, побудованих на основі рівнянь регресії, характерними є слабкі екстраполяційні властивості (поширення кількісних характеристик і висновків на іншу сукупність, інший час, за межі досліджуваної сукупності). Вони не відобра-

жають тенденцій розвитку суспільних явищ і процесів та можуть бути використані лише для короткочасних прогнозів. Інтерпретація моделей регресії дає можливість виявляти лише резерви розвитку досліджуваних явищ.

Для кількісного оцінювання **щільності (сили) зв'язку** (узгодженості варіацій взаємопов'язаних ознак) використовують коефіцієнти із такими спільними властивостями:

► за відсутності зв'язку значення коефіцієнта наближається до нуля;

► при функціональному зв'язку значення коефіцієнта наближається до одиниці;

► за наявності кореляційного зв'язку коефіцієнт виражається дробом (переважно десятковим), який збільшується за абсолютною величиною зі зростанням тісноти зв'язку.

Найпоширенішим є лінійний **коефіцієнт кореляції Пірсона** (r), який характеризує тісноту і напрям зв'язку між двома корелюючими ознаками за умови наявності між ними лінійної залежності [36, с. 134]:

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

Коефіцієнт змінюється в межах від – 1 до 1. При прямому зв'язку r – величина додатна, при зворотному – від'ємна. Знаки коефіцієнтів кореляції та регресії збігаються.

Для одержання висновків про практичне застосування побудованої регресійної моделі значенням коефіцієнта надають якісної оцінки, яка визначається за шкалою Чеддока (табл. 3.11).

Таблиця 3.11

Значення коефіцієнта Чеддока

Значення коефіцієнта *r*	**|±0,1|–|±0,3|**	**|±0,3|–|±0,5|**	**|±0,5|–|±0,7|**	**|±0,7|–|±0,9|**	**|±0,9|–|±0,99|**
Вид зв'язку	практично відсутній	слабкий	помірний	щільний	дуже щільний

При оцінюванні значущості коефіцієнта регресії *b* за допомогою *t*-критерію Стьюдента висувають та перевіряють гіпотезу про рівність лінійного коефіцієнта кореляції нулю (H_0: $r = 0$). Якщо фактичне значення критерію перевищує критичне, то зв'язок між ознаками вважають невипадковим.

Більш загальним завданням регресійного аналізу є встановлення залежності між кількома незалежними факторами ($x_1, x_2, \ldots, x_m$) та відгуком (результатною ознакою $y_{сер}$). Наприклад, залежність динаміки курсу долара від ВВП, темпів інфляції, рівня емісії, облікової ставки та інших факторів.

Дослідження лінійного зв'язку між результатною ознакою та кількома факторами називають **множинною (багатофакторною) регресією** [41, с. 467].

Побудова простих лінійних регресійних моделей у середовищі «Statistica» [35, с. 19; 52, с. 69]

Приклад 3.8. Побудувати діаграму розсіювання даних, графік регресійної прямої залежності змінної значення Індексу глобального миру держав світу за 2018 р. (змінна *Global Peace Index*) від кількості поточних конфліктів, зафіксованих у цих країнах (незалежна змінна *Number of ongoing crisis and disasters*). Записати регресійну модель.

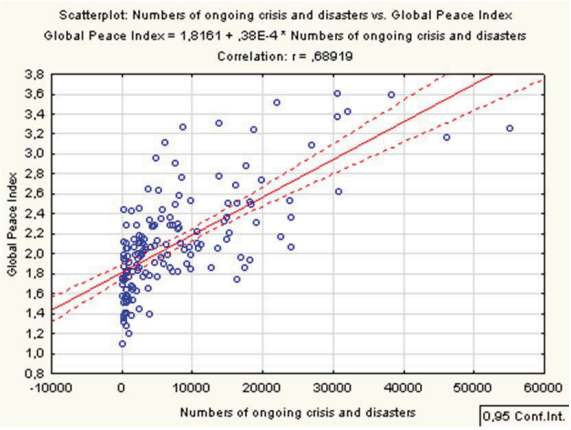


Рис. 3.3. Графік діаграми розсіювання і регресійної прямої

Отримано таку однофакторну регресійну модель:

$$Y = 1{,}81 + 0{,}000038\,X.$$

Відносно невелике значення вільного члена рівняння (1,81) свідчить про незначну частку неврахованих та випадкових факторів, не включених у рівняння регресії. Коефіцієнт при факторній змінній вказує, що при збільшенні значення незалежної змінної на одну одиницю виміру значення залежної змінної збільшиться на 0,000038.

Коефіцієнт кореляції $r = 0{,}67$ свідчить про існування помірного зв'язку між аналізованими змінними.

Модель множинної лінійної регресії:

$$y_i = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \ldots + \alpha_k x_{ik} + \varepsilon_k, \quad i = 1, n, \qquad (3.1)$$

де n – кількість спостережень, k – кількість факторів.

Коефіцієнт регресії α_i показує, на яку величину в середньому зміниться результатна ознака y_i, якщо змінну x_i збільшити на одну одиницю виміру при фіксованих значеннях інших змінних, які входять у модель. α_0 є нормованим коефіцієнтом і може бути як додатним, так і від'ємним.

Матрична форма запису рівняння (3.1):

$$Y = Xa + \varepsilon, \qquad (3.2)$$

де Y — вектор залежної змінної розмірності $n \times 1$, який представляє n спостережень значень y_i; X – матриця n спостережень незалежних

змінних $X_1, X_2, ..., X_k$, розмірність матриці $X - n \times (k + 1)$, k – кількість факторів, введених у модель. Додатковий фактор X_0 складається з одиниць. Його вводять для обчислення вільного члена. Як початкові дані можуть бути використані часові ряди (п. 4.1) або просторова вибірка; a – вектор невідомих параметрів розмірності $(k + 1) \times 1$, який потрібно оцінити; ε – вектор випадкових відхилень (збурень) розмірності $n \times 1$; ε відображає той факт, що зміна y_i неточно описується зміною пояснюючих змінних X, оскільки існують інші фактори, невраховані в моделі.

Рівняння (3.2) у матричній формі:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ ... \\ y_n \end{pmatrix}, \; X = \begin{pmatrix} 1 & x_{11} & ... & x_{1k} \\ 1 & x_{21} & ... & x_{2k} \\ ... & ... & ... & ... \\ 1 & x_{n1} & ... & x_{nk} \end{pmatrix}, \; a = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ ... \\ \alpha_n \end{pmatrix}, \; \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ ... \\ \varepsilon_n \end{pmatrix}. \qquad (3.3)$$

Рівняння (3.3) містить значення невідомих параметрів α_1, α_2, ..., α_n. Ці величини оцінюють на основі вибіркових спостережень, тому отримані розрахункові показники не є істинними, а лише їх статистичними оцінками. Модель лінійної регресії, в якій замість істинних значень параметрів підставлені їх оцінки (саме такі регресії застосовуються на практиці), має такий вигляд:

$$Y = XA + \varepsilon = \hat{Y} + \varepsilon,$$

де A – вектор оцінок параметрів; $\varepsilon = Y - XA$ – вектор «оцінюваних» відхилень регресії, залишки регресії; $\hat{Y}$ – оцінка значень Y, що дорівнює XA.

Для побудови рівняння регресії використовують **метод найменших квадратів** (МНК), суть якого полягає в мінімізації суми квадратів відхилень фактичних значень результатної ознаки від його розрахункових значень (п. 4.1):

$$\sum_{i=1}^{n} \left(y_i - \overline{y}_i \right)^2 \to \min .$$

Параметри регресійного рівняння обчислюють за методом найменших квадратів:

$$A = \left(X^T X\right)^{-1} X^T Y .$$

Щоб отримати якнайкращий результат регресійного аналізу, заснованого на звичайному МНК, потрібно забезпечити виконання умов Гаусса–Маркова [36, с. 138]:

1. Математичне сподівання випадкової складової у будь-якому спостереженні має дорівнювати нулю:

$$M(\varepsilon_i) = 0 \text{ для всіх } i = \overline{1,n}.$$

Випадкова складова може бути додатною або від'ємною, але не може мати систематичного зміщення в жодному з двох можливих напрямків.

Якщо рівняння регресії містить вільний член, ця умова виконується автоматично, оскільки роль константи полягає у визначенні будь-якої систематичної складової Y, яку не враховують пояснюючі змінні, введені в рівняння регресії.

2. Дисперсія випадкової складової має бути постійною для всіх спостережень. Не має бути апріорної причини для того, щоб в одних спостереженнях дисперсія спричинювала більшу похибку, ніж в інших:

$$D(\varepsilon_i) = D(\varepsilon_j) = \sigma_\varepsilon^2 \text{ для будь-яких спостережень } i \text{ та } j.$$

3. Умова незалежності припускає відсутність систематичного зв'язку між значеннями випадкової складової у будь-яких двох спостереженнях. Наприклад, якщо в одному спостереженні випадкова складова є великою та додатною, це не має зумовлювати систематичну тенденцію до того, що вона буде такою самою в наступному спостереженні.

Випадкові складові мають бути незалежними одна від одної:

$$M\left(\varepsilon_i, \varepsilon_j\right) = 0 \ \left(i \neq j\right).$$

Ця умова означає, що відхилення регресії (і сама залежна змінна) не корелюють. Для часового ряду y_i це означає відсутність автокореляції ряду (п. 4.1).

4. У моделі (3.1) збурення ε_i (або залежна змінна y_i) є випадковою величиною, а пояснююча змінна – невипадковою.

Якщо ця умова виконана, то теоретична коваріація (числова характеристика залежності випадкових величин, яка виникає внаслідок невизначеності результату перемножування двох сукупностей чисел) між незалежною змінною і випадковою складовою дорівнює нулю.

Разом з умовами Гаусса–Маркова передбачається нормальність розподілу випадкової складової.

У тих випадках, коли виконуються передумови, оцінки, отримані методом найменших квадратів, будуть мати властивості незміщеності, спроможності й ефективності.

Незміщена оцінка – точкова оцінка (число, обчислене на основі вибірки, ймовірно близьке до оцінюваного параметра вибірки), математичне сподівання якої дорівнює параметру, що оцінюється.

Спроможність оцінки – при постійному збільшенні обсягу вибірки вона наближається до значення параметра, який оцінює.

Якість моделі регресії позв'язують з адекватністю (відповідністю) моделі спостережуваним (емпіричним) даним. Перевірку адекватності моделі регресії спостережуваним даним проводять на основі аналізу залишків $\varepsilon_i = y_i - \hat{y}_i$.

Аналіз залишків дає можливість отримати представлення, наскільки добре підібрана модель і наскільки правильно вибрано метод оцінювання коефіцієнтів. Згідно із загальними припущеннями регресійного аналізу залишки мають бути незалежними (майже незалежними) однаково розподіленими випадковими величинами.

При аналізі якості моделі регресії використовують **коефіцієнт детермінації** [53, с. 360]:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}\varepsilon_i^{\,2}}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

де $\hat{y}_i$ – передбачене (розраховане за рівнянням регресії) значення залежної змінної; $\bar{y}$ – середнє значення залежної змінної.

Коефіцієнт детермінації визначає частку варіації результатної ознаки у, враховану в моделі та зумовлену впливом на неї факторів, введених у модель.

Чим ближче R^2 до одиниці, тим вища якість моделі.

Для оцінювання якості регресійних моделей доцільно також використовувати **коефіцієнт множинної кореляції** (індекс кореляції) [36, с. 139]:

$$R = \sqrt{1 - \frac{\sum_{i=1}^{n} \varepsilon_i^{\,2}}{\sum_{i=1}^{n} (y_i - \bar{y})^2}} = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}}.$$

Цей коефіцієнт універсальний, оскільки відображає щільність зв'язку та точність моделі, а також може бути використаний при будь-якій формі зв'язку між змінними.

Перевірка значущості побудованого рівняння загалом і окремих його параметрів

Оцінити значущість рівняння регресії означає встановити, чи відповідає математична модель, що виражає залежність між Y та X, фактичним даним і чи достатньо введених у рівняння пояснюючих змінних X для опису залежної змінної Y.

Оцінювання значущості рівняння регресії проводять для того, щоб перевірити, чи придатне рівняння регресії для практичного використання (зокрема, для прогнозування).

Для перевірки значущості моделі регресії використовують F-критерій Фішера:

$$F = \frac{R^2 / k}{(1 - R^2)(n - k - 1)}.$$

Якщо розраховане значення з k та $n - k - 1$ ступенями вільності, де k – кількість факторів, введених у модель, більше від табличного на заданому рівні значущості, модель вважають значущою.

Як міру точності застосовують незміщену оцінку дисперсії залишкової компоненти, яка є відношенням суми квадратів рівнів

залишкової компоненти до величини $n - k - 1$. Квадратний корінь з цієї величини (σ_ε) називають **стандартною помилкою** [36, с. 139]:

$$\sigma_\varepsilon = \sqrt{\frac{1}{n-k-1}\sum_{i=1}^{n}\varepsilon_i^2}\ .$$

Значущість окремих коефіцієнтів регресії перевіряють за допомогою *t*-статистики шляхом перевірки гіпотези про рівність нулю *j*-го параметра рівняння (окрім вільного члена):

$$t_{a_j} = \frac{a_j}{\sigma_{a_j}},$$

де σ_{a_j} – стандартне (середньоквадратичне) відхилення коефіцієнта рівняння регресії a_j.

Величина σ_{a_j} є квадратним коренем з добутку незміщеної оцінки дисперсії σ_ε^2 та *j*-го діагонального елементу матриці, оберненої до матриці X^TX:

$$\sigma_{a_j} = \sigma_\varepsilon\sqrt{b_{ij}},$$

де b_{ij} – діагональний елемент матриці $(X^TX)^{-1}$.

Якщо розрахункове значення *t*-критерію з $n - k - 1$ ступенями вільності перевершує його табличне значення на заданому рівні значущості, коефіцієнт регресії вважають значущим. В іншому разі фактор, що відповідає цьому коефіцієнту, виключають з моделі (при цьому її якість не погіршується).

Рівняння регресії використовують для розрахунку значень показника в заданому діапазоні зміни параметрів. Воно обмежено придатне для розрахунку за межами цього діапазону, тобто його можна застосовувати для вирішення завдань інтерполяції та в обмеженій мірі – для екстраполяції (п. 4.2).

Прогноз, отриманий підстановкою в рівняння регресії очікуваного значення параметра, є точковим. Вірогідність реалізації такого прогнозу дуже мала. Доцільно визначити довірчий інтервал прогнозу.

Щоб визначити область можливих значень результатного показника при розрахованих значеннях факторів, потрібно враховувати такі можливі причини помилок:

– розсіювання спостережень відносно лінії регресії;

– помилки, зумовлені математичним апаратом побудови лінії регресії.

Помилки першого типу вимірюють за допомогою характеристик точності, зокрема величиною σ_{ε}. Помилки другого типу зумовлені фіксацією числового значення коефіцієнтів регресії, тоді як насправді є випадковими, нормально розподіленими.

Для лінійної моделі регресії при прогнозуванні індивідуальних значень межі довірчого інтервалу розраховують як $Y \pm U\left(X_{\text{прогноз}}\right)$. Величину відхилення від лінії регресії $U\left(X_{\text{прогноз}}\right)$ обчислюють за формулою:

$$U\left(X_{\text{прогноз}}\right) = \sigma_{\acute{o}} t_{\alpha} \sqrt{1 + X^{T}_{\text{прогноз}} \left(X^{T} X\right)^{-1} X_{\text{прогноз}}},$$

де $X^{T}_{\text{прогноз}} = \left(1,\ X_{1\text{прогноз}},\ X_{2\text{прогноз}},\ \ldots,\ X_{k\ddot{i}\,\text{прогноз}}\right)$, t_{α} – табличне значення t-статистики Стьюдента на заданому рівні значущості α.

Однією з умов адекватного застосування регресійної моделі є припущення про лінійну незалежність пояснюючих (факторних) змінних: розв'язання задачі можливе лише тоді, коли стовпці та рядки матриці початкових даних лінійно незалежні (жоден з рядків/стовпців матриці не можна представити лінійною комбінацією інших рядків/стовпців). Для економічних показників ця умова виконується не завжди.

До лінійної залежності нормальних рівнянь приводить **мультиколінеарність** (сильна взаємна корельованість) пояснюючих змінних. Мультиколінеарність може спричиняти, наприклад, наявність у декількох незалежних змінних однакового часового тренда з незначними коливаннями.

Способи визначення наявності мультиколінеарності:

► аналіз матриці коефіцієнтів парної кореляції: явище мультиколінеарності в початкових даних вважають встановленим, якщо коефіцієнт парної кореляції між двома змінними більший 0,8;

► дослідження матриці X^TX: якщо визначник матриці X^TX близький до нуля, це свідчить про наявність мультиколінеарності.

Для усунення або зменшення мультиколінеарності використовують такі методи:

– виключення одного з двох сильно пов'язаних факторів: із двох пояснюючих змінних, які мають високий коефіцієнт кореляції, одну змінну виключають з розгляду. Вибір змінної для виключення проводять на підставі економічних міркувань. Якщо з економічної точки зору жодній із змінних не можна надати перевагу, залишають ту з них, яка має більший коефіцієнт кореляції із залежною змінною. Це найпростіший, але не завжди найефективніший метод;

– перехід від первинних факторів до їх головних компонент, кількість яких може бути меншою, і подальше повернення до первинних факторів;

– використання стратегії покрокового відбору (алгоритми покрокової регресії).

Найбільш поширені схеми побудови рівняння множинної регресії:

► метод включення факторів: ознаку включають у рівняння, якщо її введення істотно збільшує значення коефіцієнта множинної кореляції. Таким способом послідовно відбирають фактори, які чинять істотний вплив на результатну ознаку навіть за умовах мультиколінеарності системи ознак, відібраних як аргументів зі змістовних міркувань. При цьому першим у рівняння включають фактор, який найбільш щільно корелює з Y, другим – фактор, який в парі з першим з відібраних дає максимальне значення коефіцієнта множинної кореляції, і т. д. На кожному кроці обчислюють нове значення коефіцієнта множинної кореляції (більше, ніж на попередньому кроці). Так визначають внесок кожного відібраного фактора в пояснену дисперсію Y;

► метод виключення – відсіювання факторів з повного їх набору: після побудови рівняння регресії та оцінювання значущості всіх його коефіцієнтів з моделі виключають той фактор, коефіцієнт при якому незначущий і має найменше значення *t*-статистики за абсолютною величиною. Після цього отримують нове рівняння множинної регресії і знову проводять оцінювання значущості всіх коефіцієнтів регресії, що залишилися. Якщо і серед них виявляться незначущі, знову виключають фактор із найменшим значенням *t*-критерію. Процес виключення факторів припиняють, коли всі регресійні коефіцієнти є значущими.

Жодна з цих процедур не гарантує отримання оптимального набору змінних. Проте при практичному застосуванні вони дають можливість формувати достатньо якісні набори факторів, які істотно впливають на відгук.

Традиційно кількість факторів, включених у модель, має бути у 6–7 разів меншою за обсяг сукупності, за якою будують регресію. При порушенні цього співвідношення кількість ступенів вільності залишкової дисперсії буде дуже малою, що спричинить статистичну незначущість параметрів рівняння регресії, розрахункове значення *t*-критерію буде меншим за його табличне значення.

Особливим випадком мультиколінеарності при аналізі часових вибірок є наявність у складі змінних лінійних або нелінійних трендів. У цьому разі рекомендують спочатку виділити і виключити тренди, а потім визначити параметри регресії за залишками. Ігнорування наявності трендів у залежній і незалежній змінних призводить до **помилкової кореляції** (завищення міри впливу незалежних змінних на результатну ознаку).

Вибір кількості спостережень залежить від вимог до точності та надійності оцінок параметрів. Досягнення бажаної точності забезпечують як обсяг вибірки, так і розташування прогнозних значень факторів. Чим більше вони розкидані від середніх вибіркових значень, тим менша точність прогнозу. Мінімальний необхідний обсяг вибірки має дорівнювати 6–8 спостереженням на кожну змінну при прогнозуванні близько середніх значень факторів. При

віддаленні прогнозних значень факторів від середніх мінімальний обсяг вибірки зростає пропорційно до квадрата відхилень від них.

Суттєвою перешкодою до застосування регресії є обмеженість початкової інформації. При цьому цінність інформації може знижуватися не лише через мультиколінеарність, залежність залишків, невеликий обсяг вибірки і т. ін., а й через її «засміченість» (вплив нових, раніше неврахованих обставин).

Спостереження, які значно відхиляються від загалу, можуть бути результатом або дії значної кількості порівняно малих випадкових факторів, які в окремих випадках приводять до великих відхилень, або ж це викиди, які можна виключити з дослідження як аномальні. Якщо на кілька десятків спостережень припадає не менше трьох аномальних відхилень, припускають наявність одного або декількох неврахованих факторів, які проявляються тільки у вигляді аномальних спостережень.

Передумовою застосування множинного лінійного регресійного аналізу є відсутність функціонального зв'язку (залежності) між факторами.

Багатофакторний кореляційно-регресійний аналіз дає можливість оцінити вплив на досліджуваний результатний показник кожного із врахованих у регресійній моделі факторів при фіксованих середніх рівнях інших факторів. Модель можна застосовувати для прикладного аналізу лише після встановлення її адекватності (всі змінні статистично значущі).

Розглянуті вище методи вимірювання взаємозв'язків між ознаками називають **параметричними**, оскільки вони передбачають використання основних параметрів розподілу – середніх величин і дисперсій. Параметричні методи можна застосовувати лише для моделей з кількісними ознаками-факторами та нормально розподіленою результатною ознакою (як кількісною, так і якісною).

У випадках, коли фактори не можна виміряти кількісно або не підтверджено гіпотезу про нормальний розподіл результатної ознаки (кількісної чи якісної) для сукупностей незначного обсягу,

застосовують **непараметричні** методи дослідження взаємозв'язків з такими характеристиками [53, с. 380]:

– не потребують числового вираження значень ознак;

– не передбачають обчислення параметрів розподілу;

– не накладають обмежень на вид розподілу ознак у сукупності.

Непараметричні методи забезпечують лише оцінювання щільності зв'язку та перевірку його істотності, але не дають можливості представити зв'язок аналітично. Основою обчислення щільності зв'язку між атрибутивними ознаками є побудова таблиць співзалежності (взаємного спряження), в яких представлені комбінаційні розподіли сукупностей за факторною ознакою (за рядками) та результатною (за стовпцями). Найбільш поширеними є таблиці 2 × 2 (табл. 3.12) [72, с. 89].

Таблиця 3.12

Таблиця спряженості 2 × 2 у загальному вигляді

Ознака	A	Не A	åB
B	*a*	*b*	*a* + *b*
Не B	*c*	*d*	*c* + *d*
∑A	*a* + *c*	*b* + *d*	*a* + *b* + *c* + *d*

Для вимірювання щільності зв'язку між двома альтернативними ознаками використовують **коефіцієнт асоціації** K_A та **коефіцієнт контингенції** K_K:

$$K_A = \frac{d - b}{d + b},$$

$$K_K = \frac{d - b}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}.$$

Зв'язок вважають підтвердженим, якщо $K_A \geq 0{,}5$ чи $K_K \geq 0{,}3$.

Приклад 3.9. Проаналізувати Індекс рівня життя (ІРЖ) країн, виділивши дві групи: країни-члени ЄС та країни, що не входять у ЄС.

Таблиця 3.13

Умовні дані для прикладу 3.9

Ознака	Входять у ЄС	Не входять у ЄС
Високий ІРЖ	20	30
Низький ІРЖ	8	40

$$K_A = \frac{800 - 240}{800 + 240} = \frac{560}{1040} = 0{,}54,$$

$$K_K = \frac{800 - 240}{\sqrt{50 \cdot 48 \cdot 28 \cdot 70}} = \frac{560}{2168{,}87} = 0{,}26.$$

$K_A > 0{,}5$, тому у досліджуваній вибірці може йтися про існування зв'язку індексу рівня життя з членством у ЄС.

У разі, коли кожна з якісних ознак складається більше, ніж із двох груп (неквадратні таблиці), для визначення щільності зв'язку застосовують **коефіцієнт взаємної спряженості Пірсона–Чупрова**, який набуває значення від 0 до 1.

Коефіцієнт взаємної зв'язаності Чупрова *K* використовують у разі неоднакової кількості рядків і стовпців таблиці спряженості ($k_1 \times k_2$) [61, с. 218]:

$$K = \sqrt{\frac{\varphi^2}{(k_1 - 1)(k_2 - 1)}},$$

де k_1, k_2 – кількість груп першої та другої ознак відповідно (ознаки x та y). **Коефіцієнт взаємної зв'язаності Пірсона** C застосовують у разі, коли кількість рядків і кількість стовпців у таблиці спряженості збігається ($k_1 = k_2$):

$$C = \sqrt{\frac{\varphi^2}{1 + \varphi^2}},$$

де

$$\varphi^2 = \sum_{y=1}^{k_1} \left(\frac{\sum_{x=1}^{k_2} \left(\frac{n_{xy}^2}{n_x} \right)}{n_y} \right) - 1.$$

Значення коефіцієнтів Чупрова K та Пірсона C змінюються від 0 до 1.

Якщо одна із взаємопов'язаних ознак – кількісна, а друга – альтернативна, показником щільності є бісеріальний коефіцієнт кореляції. Наприклад, залежність рівня доходів (кількісна ознака) від рівня освіти (атрибутивна).

Оцінювання всіх непараметричних показників проводять за допомогою *t*-критерію Стьюдента.

У випадках, коли досліджувані ознаки виміряні у номінальній чи порядковій шкалі (наприклад, стать, політичний режим, членство у СОТ), застосовують **методи рангової кореляції**. Переважно це дані соціологічних досліджень (анкет), міждержавних рейтингів, експертних оцінок і т. ін.

Непараметричні методи рангової кореляції базуються на принципі ранжування значень статистичного ряду. Кожній одиниці сукупності має відповідати порядковий номер за величиною значення окремої ознаки – ранг (натуральне число). Ранжування проводять за принципом надання переваг окремо за кожною ознакою. При ранжуванні значень факторної та результатної ознак використовують один принцип – за зростанням або спаданням значень ознаки. Кількість рангів дорівнює обсягу сукупності. Зі збільшенням обсягу ступінь «розпізнаваності» елементів зменшується, тому рангові оцінки щільності зв'язку доцільно використовувати для сукупностей невеликих обсягів.

Однією з рангових оцінок щільності є коефіцієнт кореляції рангів **Спірмена *p***, при розрахунку якого використовують різниці рангів *d* факторної та результатної ознак для кожної одиниці сукупності [72, с. 7]:

$$p = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2 - 1\right)}.$$

Цей коефіцієнт має ті самі властивості, що й лінійний коефіцієнт кореляції, – змінюється в межах від 0 до 1, характеризує щільність зв'язку та вказує його напрям. Зв'язок між ознаками вважають статистично значущим, якщо значення коефіцієнта рангової кореляції Спірмена $p > 0{,}5$.

Приклад 3.10. За даними табл. 3.14 оцінити зв'язок між Індексом розвитку людського потенціалу (ІРЛП) та Індексом освіти для 10-ти вибраних країн.

Таблиця 3.14

Значення ІРЛП та індексу освіти за 2017 р. для окремих країн

№ з/п	Країна	Індекс освіти y	ІРЛП x	Ранги		$d = R_x - R_y$	d^2
				R_x	R_y		
1	Мексика	0,879	0,775	5	4	1	1
2	М'янма	0,787	0,498	8	9	–1	1
3	Намібія	0,808	0,608	7	7	0	0
4	Нігерія	0,648	0,471	10	10	0	0
5	Нідерланди	0,985	0,921	3	2	1	1
6	Нова Зеландія	0,993	0,919	1	3	–2	4
7	Норвегія	0,989	0,955	2	1	1	1
8	Пакистан	0,665	0,515	9	8	1	1
9	Парагвай	0,864	0,669	6	6	0	0
10	Перу	0,885	0,741	4	5	-1	1
	Разом	–	–	–	–	–	10

Обчислене $p = 0{,}94 > 0{,}5$, що свідчить про щільний прямий зв'язок між Індексом освіти та Індексом розвитку людського потенціалу.

Лінійне моделювання в середовищі «Statistica»: множинна регресія [35, с. 21; 52, с. 72]

Приклад 3.11. Побудувати лінійну регресійну модель залежності ІЛР країни I від показників тривалості життя T (у роках), величини валового національного доходу (ВНД) на душу населення Vd (у доларах), середньої тривалості отриманої освіти дорослого населення Tod (25 років і більше) та очікуваної тривалості навчання дітей шкільного віку Tnd (рис. 3.4).

Індекс людського розвитку (ІЛР) є узагальненим вимірником оцінки людського розвитку за трьома основними складовими: здоров'я і довголіття, доступність освіти та гідний рівень життя.

Загальний показник має діапазон від 0,001 до 1 і розраховується для кожної країни. За величиною ІЛР визначають місце країни у світовому рейтингу.

1 Країна	2 I	3 T	4 Vd	5 Tod	6 Tnd
Норвегія	0,955	81,3	48688	12,6	17,5
США	0,937	78,7	43480	13,3	16,8
ФРН	0,920	80,6	35431	12,2	16,4
Японія	0,912	83,6	32545	11,6	15,3
Корея	0,909	80,7	28231	11,6	17,2
Франція	0,893	81,7	30277	10,6	16,1
Італія	0,881	82,0	26158	10,1	16,2
Великобританія	0,875	80,3	32538	9,4	16,4
Угорщина	0,831	74,6	16088	11,7	15,3
Польща	0,821	76,3	17776	10,0	15,2
Білорусь	0,793	70,6	13385	11,5	14,7
Росія	0,788	69,1	14461	11,7	14,3
Румунія	0,786	74,2	11011	10,4	14,5
Казахстан	0,754	67,4	10451	10,4	15,3
Україна	0,740	68,8	6428	11,3	14,8
Туреччина	0,722	74,2	13710	6,5	12,9
Китай	0,699	73,7	7945	7,5	11,7
Нігер	0,304	55,1	701	1,4	4,9

Рис. 3.4. Фрагмент файла даних для прикладу 3.11

Специфікація моделі (аналітична форма економетричної моделі):

$$I = b_0 + b_1 T + b_2 Vd + b_3 Tod + b_4 Tnd + \varepsilon,$$

де I – залежна змінна (відгук), всі інші – незалежні (фактори).

Отримано оцінки регресійної моделі (рис. 3.5–3.6):

```
 Multiple Regression Results

Dependent: I                Multiple R =  ,99726347     F = 636,8730
                                 R?=  ,99453444    df =   4,14
No. of cases: 19            adjusted R?=  ,99297285     p =  ,000000
            Standard error of estimate:  ,012059384
Intercept:  -,257763767  Std.Error: ,0483793  t(   14) = -5,328  p =  ,0001

          T b*=,408              Vd b*=,044              Tod b*=,310
        Tnd b*=,327
```

Рис. 3.5. Оцінки регресійної моделі

► **Standard error of estimate** – стандартна помилка оцінки – міра розсіювання спостережуваних значень щодо регресійної прямої;

► **Intercept** – оцінка вільного члена b_0 регресії, якщо обрано регресію, що містить вільний член;

► **Std.Error** – стандартна помилка оцінки вільного члена b_0;

► ***t, p*** – значення *t*-критерію та рівня значущості *p* для перевірки гіпотези про рівність нулю вільного члена b_0;

► ***F, df*** – значення *F*-критерію, кількість ступенів вільності та рівень значущості *p* для перевірки гіпотези про залежність факторів і показників.

З отриманих результатів аналізу (рис. 3.5) випливає, що залежність між показником і факторами сильна (R^2~1). Побудована лінійна регресія адекватно описує взаємозв'язок між показником і факторами, вільний член статистично значущий ($p < 0{,}05$).

Друга частина інформаційного вікна (рис. 3.5) містить інформацію про значущі (при *T*, *Tod*, *Tnd*) та незначущі (при ***Vd***) оцінки регресійних коефіцієнтів. Пояснення **Significant *b** are highlighted in red** означає, що значущі коефіцієнти виділено червоним**;** *b** – стандартизовані коефіцієнти b_1, b_2, b_3, b_4 (коефіцієнти при незалежних змінних).

Результати дослідження залежності відгуку – Індексу людського розвитку країни від регресорів (показників тривалості життя, величини ВНД на душу населення, середньої тривалості освіти дорослого населення та очікуваної тривалості навчання дітей шкільного віку) зображено на рис. 3.6.

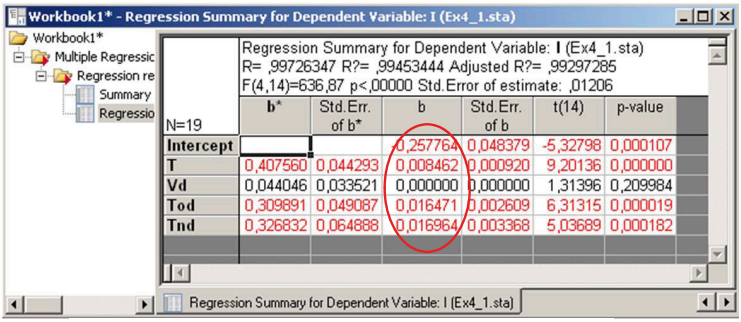


Рис. 3.6. Результати виконання множинної лінійної регресії

У першому стовпці таблиці розташовані значення коефіцієнтів *b** (стандартизовані коефіцієнти регресійного рівняння), у другому – стандартні помилки цих коефіцієнтів, у третьому (стовпець *b*) – точкові оцінки параметрів моделі:

– вільний член $b_0 = –0,257764$;

– коефіцієнт b_1 (при незалежній змінній *T*) = 0,4;

– коефіцієнт b_2 (при незалежній змінній *Vd*) = 0,04;

– коефіцієнт b_3 (при незалежній змінній *Tod*) = 0,31;

– коефіцієнт b_4 (при незалежній змінній *Tnd*) = 0,33.

Коефіцієнти *b** оцінюють за стандартизованими даними, що мають вибіркове середнє, що дорівнює 0, і стандартне відхилення, яке дорівнює 1. Величини *b** дають змогу оцінити внески кожного фактора в прогнозування показника.

У наступних стовпцях таблиці результатів (рис. 3.4) містяться стандартні помилки для b_0, b_1, b_2, b_3, b_4, значення статистик *t*-критерію та рівня значущості *p*. За підсумковою таблицею регресії можна побудувати таку модель:

$$I = –0,26 + 0,4 \times T + 0,04 \times Vd + 0,31 \times Tod + 0,33 \times Tnd + \varepsilon.$$

Для оцінювання значущості отриманих коефіцієнтів регресійного рівняння використовують *t*-критерій Стьюдента – стовпець *t*(14). У пакеті «Statistica» значення *t*-критерію (t_p) визначаються як відношення взятого за модулем коефіцієта регресії (стовпець *b*) до його стандартної помилки (стовпець **St Err. of *b***) (рис. 3.11). Табличне значення *t*-критерію з рівнем значущості $\alpha = 0,01$ і кількістю ступенів вільності $df = n – m – 1 = 19 – 4 – 1 = 14$ (*n* – кількість спостережень, *m* – кількість незалежних змінних) $t_{табл} = 4,140$. Порівняємо значення $t_{емп}$ і $t_{табл}$ для кожного з отриманих параметрів:

– $t_{емп} = 5,32 > t_{табл}$ – для вільного члена b_0;

– $t_{емп} = 9,2 > t_{табл}$ – для коефіцієнта b_1;

– $t_{емп} = 1,31 < t_{табл}$ – для коефіцієнта b_2;

– $t_{емп} = 6,31 > t_{табл}$ – для коефіцієнта b_3;

– $t_{емп} = 5,04 > t_{табл}$ – для коефіцієнта b_4.

Статистично значущими є коефіцієнти b_0, b_1, b_3, b_4. Коефіцієнт b_2 сформований під впливом випадкових факторів. Відповідно фактор *Vd* можна виключити з моделі як неінформативний. Аналогічний висновок можна зробити, порівнюючи значення рівня значущості (стовпець ***p*-level**, рис. 3.6) з прийнятим рівнем $\alpha = 0{,}01$. Для b_0, b_1, b_3, b_4 показник імовірності випадкових значень параметрів регресії менший за 1% (0,01×100%). Можна зробити висновок про те, що отримані коефіцієнти статистично значущі й надійні. Для b_2 робимо висновок про випадковість його значення, оскільки $\alpha = 0{,}2 \times 100\% = 20\% > 1\%$. Це дає можливість розглядати b_2 як неінформативний фактор. Його можна вилучити з рівняння для покращення моделі.

Вільний член b_0 оцінює агрегований вплив інших (неврахованих у моделі) факторів на результат *I*. Коефіцієнти b_1, b_2, b_3, b_4 вказують на те, що зі збільшенням на одиницю значень *T*, *Vd*, *Tod* та *Tnd* залежна змінна *I* зростає на 0,4, 0,04 0,31 та 0,33 відповідно. Порівнювати ці значення не можна, оскільки вони залежать від одиниць вимірювання кожної ознаки і є незрівнянними між собою. Для порівняння використовують відносні показники – β-коефіцієнти (стовпець *b**, рис. 3.7).

Визначення оптимальної кількості пояснюючих змінних, що адекватно описують зміни відгуку

При підготовці даних прогнозист має знайти набір регресорів, здатних суттєво вплинути на процес прогнозування значень залежної змінної, та визначити ті з них, що необхідні для прогнозування майбутніх значень ***I*.** Важливим показником оцінювання моделі лінійної регресії є коефіцієнт детермінації R^2.

При визначенні кількості регресорів (змінних, що можуть вплинути на прогнозування майбутніх значень) намагаються позбутися тих, в яких 95%-і довірчі інтервали (відповідні коефіцієнти b_i можуть бути нулями).

З отриманих результатів (див. рис. 3.6) можна зробити висновок, що незначущою є змінна *Vd*. Коефіцієнт детермінації $R^2 = 0{,}99453$. Претендентами на пояснюючі змінні є *T*, *Tod*, *Tnd*.

Далі будемо включати або виключати з моделі змінні за допомогою «регресії вперед» та «регресії назад». Для цього виконаємо процедуру поступового включення змінних: спочатку, перебираючи всі регресори, будують модель простої регресії та перевіряють, яке значення *F*-критерію Фішера при перевірці гіпотези про незначущість змінної є найбільшим. Таку змінну залучають до множини регресорів (рис. 3.7–3.11).

Regression Summary for Dependent Variable: I (Ex4_1.sta)
R= ,96760538 R?= ,93626017 Adjusted R?= ,93251077
F(1,17)=249,71 p<,00000 Std.Error of estimate: ,03737

N=19	b*	Std.Err. of b*	b	Std.Err. of b	t(17)	p-value
Intercept			0,067751	0,047649	1,42188	0,173149
Tnd	0,967605	0,061232	0,050224	0,003178	15,80219	0,000000

Рис. 3.7. Результати виконання множинної лінійної регресії на I етапі включення змінних у модель

Regression Summary for Dependent Variable: I (Ex4_1.sta)
R= ,98858175 R?= ,97729388 Adjusted R?= ,97445562
F(2,16)=344,33 p<,00000 Std.Error of estimate: ,02299

N=19	b*	Std.Err. of b*	b	Std.Err. of b	t(16)	p-value
Intercept			-0,259072	0,067479	-3,83929	0,001448
Tnd	0,684290	0,064770	0,035518	0,003362	10,56493	0,000000
T	0,348283	0,064770	0,007231	0,001345	5,37723	0,000062

Рис. 3.8. Результати виконання множинної лінійної регресії на II етапі включення змінних у модель

Regression Summary for Dependent Variable: I (Ex4_1.sta)
R= ,99692548 R?= ,99386042 Adjusted R?= ,99263250
F(3,15)=809,39 p<,00000 Std.Error of estimate: ,01235

N=19	b*	Std.Err. of b*	b	Std.Err. of b	t(15)	p-value
Intercept			-0,300298	0,036814	-8,15709	0,000001
Tnd	0,324364	0,066413	0,016836	0,003447	4,88408	0,000198
T	0,440002	0,037654	0,009136	0,000782	11,68548	0,000000
Tod	0,317518	0,049909	0,016876	0,002653	6,36198	0,000013

Рис. 3.9. Результати виконання множинної лінійної регресії на III етапі включення змінних у модель

N=19	Regression Summary for Dependent Variable: I (Ex4_1.sta) R= ,99726347 R?= ,99453444 Adjusted R?= ,99297285 F(4,14)=636,87 p<,00000 Std.Error of estimate: ,01206					
	b*	Std.Err. of b*	b	Std.Err. of b	t(14)	p-value
Intercept			-0,257764	0,048379	-5,32798	0,000107
Tnd	0,326832	0,064888	0,016964	0,003368	5,03689	0,000182
T	0,407560	0,044293	0,008462	0,000920	9,20136	0,000000
Tod	0,309891	0,049087	0,016471	0,002609	6,31315	0,000019
Vd	0,044046	0,033521	0,000000	0,000000	1,31396	0,209984

Рис. 3.10. Результати виконання множинної лінійної регресії на IV етапі включення змінних у модель

Процес залучення змінних у модель продовжуємо, поки не отримано результат, де є незначуща змінна (*Vd*).

У заданому прикладі із 4 регресорів обрано всі 4. Будемо поступово включати з цього набору покроково ту змінну, яка має найменше значення *F*-критерію Фішера.

Формуємо нову регресійну модель, регресорами якої будуть всі змінні початкової моделі *T, Vd, Tod* та *Tnd*. Далі запускаємо процедуру покрокового виключення незалежних змінних. При цьому на кожному етапі коефіцієнт детермінації має бути достатньо великим ($R^2 > 0{,}75$). Оцінки оптимального варіанта досліджуваної моделі:

N=19	Regression Summary for Dependent Variable: I (Ex4_1.sta) R= ,99692548 R?= ,99386042 Adjusted R?= ,99263250 F(3,15)=809,39 p<,00000 Std.Error of estimate: ,01235					
	b*	Std.Err. of b*	b	Std.Err. of b	t(15)	p-value
Intercept			-0,300298	0,036814	-8,15709	0,000001
T	0,440002	0,037654	0,009136	0,000782	11,68548	0,000000
Tod	0,317518	0,049909	0,016876	0,002653	6,36198	0,000013
Tnd	0,324364	0,066413	0,016836	0,003447	4,88408	0,000198

Рис. 3.11. Результати побудови оптимальної моделі множинної лінійної регресії

Оптимальна регресійна модель має вигляд:

$$I = -0{,}3 + 0{,}44 \times T + 0{,}32 \times Tod + 0{,}33 \times Tnd + \varepsilon.$$

Оцінки коефіцієнтів мають додатні знаки – Індекс людського розвитку країни тим вищий, чим більший показник тривалості життя, більша середня тривалість отриманої освіти дорослого населення і більша очікувана тривалість навчання дітей шкільного віку.

Оцінювання щільності парних залежностей

Причинна залежність – зв'язок між процесами, коли зміна одного з них є наслідком зміни іншого.

Оцінити щільність парних залежностей включених у модель факторів можна за допомогою матриці парних коефіцієнтів кореляції (рис. 3.12).

Variable	Correlations (Ex4_1.sta)				
	T	Vd	Tod	Tnd	I
T	1,000000	0,800621	0,633247	0,813462	0,904927
Vd	0,800621	1,000000	0,589225	0,698720	0,781307
Tod	0,633247	0,589225	1,000000	0,898583	0,887616
Tnd	0,813462	0,698720	0,898583	1,000000	0,967605
I	0,904927	0,781307	0,887616	0,967605	1,000000

Рис. 3.12. Матриця парних коефіцієнтів кореляції

Отримані значення парних коефіцієнтів кореляції свідчать про щільний зв'язок Індексу людського розвитку країни *I* як з показником тривалості життя *T* – 0,9049, так і з середньою тривалістю отриманої освіти дорослого населення *Tod* – 0,8876 та з очікуваною тривалістю навчання дітей шкільного віку *Tnd* – 0,9676. При цьому потрібно враховувати щільний міжфакторний зв'язок *Tod* з *Tnd* (0,8985), який приблизно дорівнює зв'язку *I* з *Tod*. Для покращення моделі фактор *Vd* можна вивести як недостатньо статистично надійний.

Оцінити щільність зв'язку значень двох змінних без впливу всіх інших змінних, представлених у рівнянні множинної регресії, можна за допомогою матриці лінійних коефіцієнтів часткової кореляції (рис. 3.13).

Variable	Variables currently in the Equation; DV: I (Ex4_1.sta)						
	b* in	Partial Cor.	Semipart Cor.	Tolerance	R-square	t(14)	p-value
T	0,407560	0,926340	0,181805	0,198988	0,801012	9,201361	0,000000
Vd	0,044046	0,331334	0,025962	0,347425	0,652575	1,313958	0,209984
Tod	0,309891	0,860260	0,124738	0,162025	0,837975	6,313148	0,000019
Tnd	0,326832	0,802746	0,099521	0,092722	0,907278	5,036888	0,000182

Рис. 3.13. Матриця лінійних коефіцієнтів часткової кореляції

Матриця лінійних коефіцієнтів часткової кореляції містить коефіцієнти *b**, часткові коефіцієнти кореляції (**Psrtical Cor**), напівчасткові коефіцієнти кореляції (**Semipart Cor**), толерантності (**Tolerance**), коефіцієнти детермінації (***R*-square**), значення *t*-критерію та рівні значущості *p* – імовірності відхилення гіпотези про значущість часткового коефіцієнта кореляції.

Коефіцієнти часткової кореляції дають точнішу характеристику щільності залежності двох ознак, ніж коефіцієнти парної кореляції, тому що «очищають» парну залежність від взаємодії даної пари ознак з іншими представленими в моделі ознаками. Найбільше Індекс людського розвитку країни *I* пов'язаний з показником тривалості життя *T* – 0,9263, з середньою тривалістю отриманої освіти дорослим населенням *Tod* – 0,8603 та очікуваною тривалістю навчання дітей шкільного віку *Tnd* – 0,8027 порівняно зі зв'язком *I* з ВНД *Vd* – 0,3313 (рис. 3.13). Цей факт свідчить про доцільність виключення фактора *Vd* з моделі.

Напівчасткова кореляція – кореляція фактора та показника в припущенні, що контролюється вплив інших факторів на заданий фактор, але не контролюється вплив факторів на показник. Якщо напівчасткова кореляція мала, тоді як часткова кореляція відносно велика, то відповідний фактор може мати самостійну «частку» у поясненні мінливості залежної змінної, тобто «частку», що не пояснюється іншими факторами. З таблиці (рис. 3.13) видно, що фактори *T, Tod* та *Tnd* мають самостійну частку в поясненні мінливості відгуку.

***R*-square** (коефіцієнт детермінації) – квадрат коефіцієнта множинної кореляції між за даною змінною та всіма іншими змінними, що входять у рівняння регресії.

Toleranse (толерантність) = 1 – *R*-square.

$t(14)$ – значення критерію Стьюдента для перевірки гіпотези про значущість часткового коефіцієнта кореляції із зазначеною (у дужках) кількістю ступенів вільності.

***p*-value** – ймовірність відхилення гіпотези про значущість часткових коефіцієнтів кореляції. Часткові коефіцієнти кореляції змінних *T*, *Tod* та *Tnd* значущі при рівні значущості $p \le 0{,}00018$, а частковий коефіцієнт кореляції змінної *Vd* не є значущим ($p = 0{,}2$).

Значення коефіцієнтів множинної детермінації дають можливість зробити висновок про високу (більше за 80%) детермінованість результатної ознаки *I* в моделі з факторними ознаками *T*, *Tod* та *Tnd*.

Оцінювання адекватності моделі на основі аналізу залишків

Залишки – це різниці між спостережуваними значеннями (емпіричними) та модельними (аналітичними), тобто значеннями, підрахованими за моделлю з оціненими параметрами. Модель можна вважати задовільною, якщо залишки некорельовані й розподілені (приблизно) за нормальним законом.

Аналіз залишків моделі на нормальність

Щоб перевірити відповідність залишків нормальному розподілу, потрібно побудувати таблицю залишків (рис. 3.14). Перевіряємо, чи виходять залишки за межі інтервалу $(-3s, 3s)$, де s – емпіричне середньоквадратичне відхилення залишків (на графіку залишок позначений *). З рис. 3.14 видно, що залишки не виходять за межі інтервалу $(-3s, 3s)$. Значення залишків розташовані у третьому стовпці таблиці. Середнє залишків 0, медіана – 0,0019.

Графіки залежності регресійних залишків від експериментальних значень вихідних змінних дають можливість перевірити припущення про однорідність і незалежність помилок, що є передумовами застосування методу найменших квадратів, і локалізувати викиди. Якщо такі умови виконуються, графіки будуть мати симетричний, випадковий і рівномірний розподіли точок. Графіки емпіричної функції розподілу залишків на нормальному ймовірнісному розподілі **(Probability plots)** і гістограми дають змогу перевірити справедливість припущення про нормальний розподіл залишків.

Raw Residual (Ex4_1.sta)
Dependent variable: I

Case	Raw Residuals −3s . . . 0 . . +3s	Observed Value	Predicted Value	Residual	Standard Pred. v.	Standard Residual	Std.Err. Pred.Val	Mahalanobis Distance	Deleted Residual	Cook's Distance
1		0,955000	0,949720	0,005280	0,98523	0,42764	0,004066	1,00434	0,005923	0,006236
2		0,937000	0,925994	0,011006	0,81980	0,89129	0,004695	1,65524	0,012866	0,039243
3		0,920000	0,918054	0,001946	0,76444	0,15759	0,004082	1,01971	0,002185	0,000855
4		0,912000	0,916816	-0,004816	0,75581	-0,39003	0,007439	5,58588	-0,007560	0,034014
5		0,909000	0,922311	-0,013311	0,79412	-1,07799	0,004365	1,30163	-0,015212	0,047404
6		0,893000	0,896051	-0,003051	0,61102	-0,24706	0,004340	1,27664	-0,003481	0,002455
7		0,881000	0,892037	-0,011037	0,58303	-0,89384	0,005049	2,06271	-0,013253	0,048162
8		0,875000	0,868060	0,006940	0,41585	0,56203	0,006690	4,33674	0,009824	0,046453
9		0,840000	0,833628	0,006372	0,17576	0,51606	0,004718	1,68004	0,007461	0,013325
10		0,831000	0,836281	-0,005281	0,19426	-0,42770	0,003704	0,67196	-0,005803	0,004968
11		0,821000	0,821439	-0,000439	0,09077	-0,03555	0,003366	0,39018	-0,000474	0,000027
12		0,793000	0,786261	0,006739	-0,15452	0,54577	0,004744	1,70901	0,007906	0,015124
13		0,788000	0,769198	0,018802	-0,27349	1,52269	0,005724	2,92035	0,023948	0,202053
14		0,786000	0,797219	-0,011219	-0,07811	-0,90855	0,002960	0,08669	-0,011903	0,013344
15		0,754000	0,748564	0,005436	-0,41737	0,44023	0,007940	6,49543	0,009268	0,058237
16		0,740000	0,768125	-0,028125	-0,28097	-2,27768	0,005514	2,64181	-0,035130	0,403476
17		0,722000	0,704464	0,017536	-0,72486	1,42018	0,006316	3,76141	0,023749	0,241922
18		0,699000	0,696569	0,002432	-0,77992	0,19692	0,005577	2,72442	0,003055	0,003121
19		0,304000	0,309211	-0,005211	-3,48085	-0,42200	0,010742	12,67581	-0,021430	0,569907
Minimum		0,304000	0,309211	-0,028125	-3,48085	-2,27768	0,002960	0,08669	-0,035130	0,000027
Maximum		0,955000	0,949720	0,018802	0,98523	1,52269	0,010742	12,67581	0,023948	0,569907
Mean		0,808421	0,808421	-0,000000	0,00000	-0,00000	0,005370	2,84211	-0,000424	0,092122
Median		0,831000	0,833628	0,001946	0,17576	0,15759	0,004744	1,70901	0,002185	0,034014

Рис. 3.14. Таблиця залишків множинної регресійної моделі

Для візуального аналізу розподілу залишків можна також використати нормальні ймовірнісні графіки. Побудуємо *P–P* діаграму порівняння залишків моделі з нормальним розподілом (рис. 3.15). З отриманого графіка можна вірогідно припустити, що залишки розподілені нормально – розкид не дуже великий.

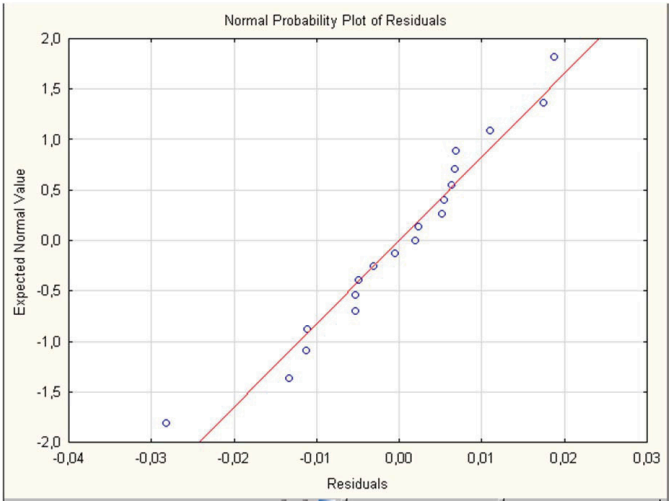


Рис. 3.15. Графік нормального розподілу залишків

Виконаємо графічне порівняння прогнозних значень показника і спостережених значень (рис. 3.16).

Перевірка розподілу залишків на відповідність нормальному закону

Однією з умов коректного застосування регресійного аналізу є відповідність розподілу залишків нормальному закону. З побудованого графіка (рис. 3.17) видно, що розподіл залишків загалом відповідає нормальному закону.

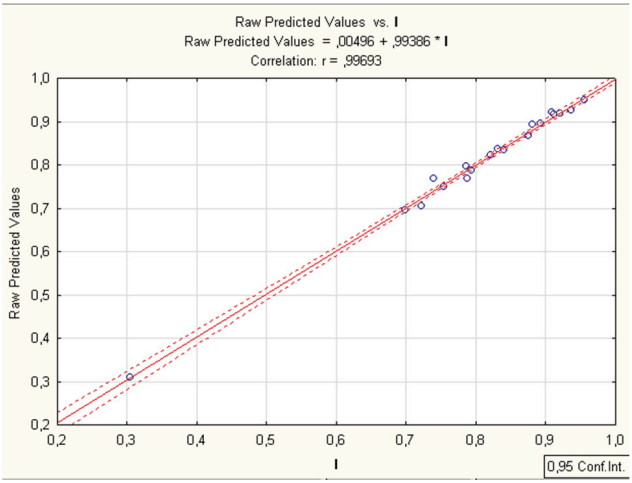


Рис. 3. 16. Графік порівняння прогнозних і спостережених значень

Якщо залишки не розподілені за нормальним законом, для стабілізації дисперсії даних застосовують перетворення залежних та незалежної змінних, наприклад, логарифмічне перетворення залежних змінних або знаходження квадратного кореня.

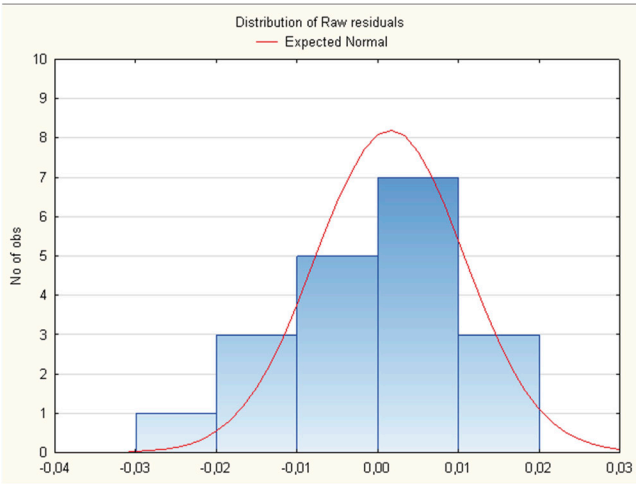


Рис. 3.17. Графік розподілу залишків

Аналіз викидів

16-те спостереження ідентифікується як викид (рис. 3.18).

Standard Residual: **I** (Ex4_1.sta)
Outliers

Case	Standard Residuals −5.	−4.	−3.	±2.	3.	4.	5.	**Observed Value**	Predicted Value	Residual	Standard Pred. v.	Standard Residual	Std.Err. Pred.Val
16	.	.	.	*|	.	.	.	0,740000	0,768125	-0,028125	-0,280975	-2,27768	0,005514
Minimum	.	.	.	*|	.	.	.	0,740000	0,768125	-0,028125	-0,280975	-2,27768	0,005514
Maximum	.	.	.	*|	.	.	.	0,740000	0,768125	-0,028125	-0,280975	-2,27768	0,005514
Mean	.	.	.	*|	.	.	.	0,740000	0,768125	-0,028125	-0,280975	-2,27768	0,005514
Median	.	.	.	*|	.	.	.	0,740000	0,768125	-0,028125	-0,280975	-2,27768	0,005514

Рис. 3.18. Аналіз викидів

Якщо виключити його з аналізу, регресія теж зміниться. Чи потрібно враховувати спостереження-викиди в моделі, однозначно стверджувати неможливо. Це вирішує замовник дослідження.

Обчислимо прогнозні значення залежної величини для заданих значень регресорів та знайдемо інтервал надійності (рис. 3.19).

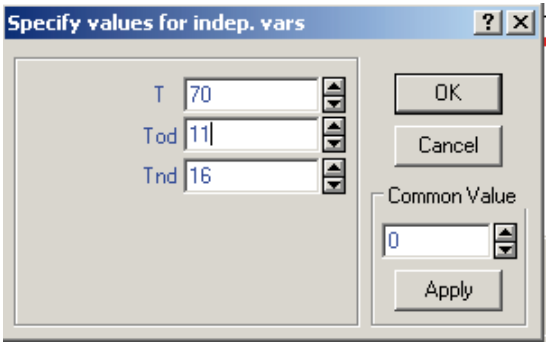


Рис. 3.19. Значення регресорів прогнозної моделі

Передбачуване значення для $I = 0{,}8016$. З імовірністю 0,95 дійсне значення I має потрапити в інтервал (0,7886;0,8147) (рис. 3.20).

Predicting Values for (Ex4_1.sta)
variable: I
Exclude cases: 16

Variable	b-Weight	Value	b-Weight * Value
T	0,008399	70,00000	0,587935
Tod	0,017261	11,00000	0,189867
Tnd	0,018011	16,00000	0,288168
Intercept			-0,264327
Predicted			0,801643
-95,0%CL			0,788627
+95,0%CL			0,814659

Рис. 3.20. Прогнозні значення Індексу людського розвитку для заданих значень регресорів

Побудова лінійних моделей множинної регресії у середовищі SPSS

Приклад 3.12. Побудувати регресійну модель для прогнозування обсягів реалізації продукції ТНК за умовними даними (табл. 3.15).

Таблиця 3.15

Умовні дані для прикладу 3.12

Обсяг реалізації (млн. дол.), Y	Період (міс.), X_1	Витрати на рекламу (тис. дол.), X_2	Ціна (дол.), X_3	Середня ціна у конкурентів (дол.), X_4	Індекс споживчих витрат (%), X_5
126	1	4	15	17	100
137	2	4,8	14,8	17,3	98,4
148	3	3,8	15,2	16,8	101,2
191	4	8,7	15,5	16,2	103,5
274	5	8,2	15,5	16	104,1
370	6	9,7	16	18	107
432	7	14,7	18,1	20,2	107,4
445	8	18,7	13	15,8	108,5
367	9	19,8	15,8	18,2	108,3
367	10	10,6	16,9	16,8	109,2
321	11	8,6	16,3	17	110,1
307	12	6,5	16,1	18,3	110,7
331	13	12,6	15,4	16,4	110,3
345	14	6,5	15,7	16,2	111,8
364	15	5,8	16	17,7	112,3
384	16	5,7	15,1	16,2	112,9

Відбір факторних ознак для побудови регресійної моделі на основі аналізу матриці коефіцієнтів кореляції

Побудуємо матрицю парної кореляції всіх змінних (рис. 3.21).

Коефіцієнт кореляції Пірсона r між двома змінними визначає силу зв'язку між кожною з пар змінних та набуває значень від –1 до +1: чим ближче абсолютне значення до одиниці, тим сильніший зв'язок.

Аналіз матриці коефіцієнтів парної кореляції (див. рис. 3.21) свідчить, що залежна змінна *обсяг реалізації* має щільний зв'язок з *Індексом споживчих витрат* (r_{Y,X_5} = 0,816), з *витратами на рекламу* (r_{Y,X_2} = 0,646) і з часовим періодом (r_{Y,X_1} = 0,678). Окрім того, фактори X_2 та X_5 щільно пов'язані між собою (r_{X_1,X_5} = 0,96), що свідчить про наявність мультиколінеарності. З цих двох змінних залишимо в моделі X_5 (*Індекс споживчих витрат*).

Correlations

		Обсяг реалізації (млн. дол.)	Період (міс.)	Витрати на рекламу (тис. дол.)	Ціна (дол.)	Середня ціна у конкурентів (дол.)	Індекс споживчих витрат (%)
Обсяг реалізації (млн. дол.)	Pearson Correlation	1	,678**	,646**	,233	,226	,816**
	Sig. (2-tailed)		,004	,007	,385	,399	,000
	N	16	16	16	16	16	16
Період (міс.)	Pearson Correlation	,678**	1	,106	,174	-,051	,960**
	Sig. (2-tailed)	,004		,695	,520	,851	,000
	N	16	16	16	16	16	16
Витрати на рекламу (тис. дол.)	Pearson Correlation	,646**	,106	1	-,003	,204	,209
	Sig. (2-tailed)	,007	,695		,990	,448	,421
	N	16	16	17	16	16	17
Ціна (дол.)	Pearson Correlation	,233	,174	-,003	1	,698**	,235
	Sig. (2-tailed)	,385	,520	,990		,003	,380
	N	16	16	16	16	16	16
Середня ціна у конкурентів (дол.)	Pearson Correlation	,226	-,051	,204	,698**	1	,031
	Sig. (2-tailed)	,399	,851	,448	,003		,910
	N	16	16	16	16	16	16
Індекс споживчих витрат (%)	Pearson Correlation	,816**	,960**	,209	,235	,031	1
	Sig. (2-tailed)	,000	,000	,421	,380	,910	
	N	16	16	17	16	16	17

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

Рис. 3.21. Матриця коефіцієнтів парної кореляції

На основі аналізу матриці коефіцієнтів парної кореляції робимо висновок про доцільність побудови двофакторної регресійної моделі $Y = f(X_2, X_5)$.

Побудова лінійного рівняння регресії зі значущими факторами

Оцінювання параметрів регресії проведемо методом найменших квадратів. Як незалежні оберемо змінні X_1–X_5. Використаємо метод виключення.

Отримані результати регресійного аналізу

1. Таблиця послідовного виключення змінних з регресійної моделі (рис. 3.22).

2. Зведення моделі представляє значення коефіцієнта детермінації, коефіцієнта множинної кореляції, стандартної помилки, коефіцієнта Дурбіна–Уотсона послідовно для всіх моделей, у тому числі оптимальної (рис. 3.25).

3. Результати дисперсійного аналізу та значення *t*-критерію, отримані на кожному кроці (**ст.св.** – кількість ступенів вільності, **Знч.** – значущість) (рис. 3.24).

4. Таблиця коефіцієнтів рівняння регресії, стандартних похибок коефіцієнтів рівняння регресії, стандартизованих коефіцієнтів та статистики, яку використовують для перевірки значущості коефіцієнтів регресійної моделі (рис. 3.25).

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	Індекс споживчих витрат (%), Середня ціна у конкурентів (дол.), Витрати на рекламу (тис. дол.), Ціна (дол.), Період (міс.)	.	Enter
2	.	Ціна (дол.)	Backward (criterion: Probability of F-to-remove >= ,100).
3	.	Середня ціна у конкурентів (дол.)	Backward (criterion: Probability of F-to-remove >= ,100).
4	.	Період (міс.)	Backward (criterion: Probability of F-to-remove >= ,100).

a. All requested variables entered.
b. Dependent Variable: Обсяг реалізації (млн. дол.)

Рис. 3.22. Таблиця включення/виключення змінних на кожному етапі

Model Summary[e]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,944[a]	,891	,836	41,649	
2	,943[b]	,889	,848	40,044	
3	,939[c]	,882	,852	39,537	
4	,927[d]	,859	,837	41,473	1,357

Рис. 3.23. Зведення регресійної моделі

Оптимальна модель лінійної регресії містить два значущих фактори (*затрати на рекламу* та *Індекс споживчих витрат*) і має такий вигляд:

$$Y = -1471{,}314 + 9{,}568X_1 + 15{,}754X_2.$$

Коефіцієнти рівняння регресії показують, що при збільшенні витрат на рекламу на 1 тис. дол. обсяги реалізації підвищаться на 9,568 млн. дол., а при збільшенні Індексу споживчих витрат на 1% зростуть на 15,754 млн. дол.

ANOVA[e]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	141371,778	5	28274,356	16,300	,000[a]
	Residual	17346,660	10	1734,666		
	Total	158718,438	15			
2	Regression	141079,525	4	35269,881	21,995	,000[b]
	Residual	17638,912	11	1603,537		
	Total	158718,438	15			
3	Regression	139960,250	3	46653,417	29,845	,000[c]
	Residual	18758,188	12	1563,182		
	Total	158718,438	15			
4	Regression	136358,334	2	68179,167	39,639	,000[d]
	Residual	22360,104	13	1720,008		
	Total	158718,438	15			

Рис. 3.24. Таблиця дисперсійного аналізу

Оцінювання якості моделі

Коефіцієнт детерміації $R^2 = 0{,}859$ (рис. 3.23) відображає частку варіації результатної ознаки, зумовлену впливом досліджуваних факторів. Для прикладу 3.12 близько 86% варіації залежної змінної враховані в моделі та зумовлені впливом включених факторів.

Коефіцієнт множинної кореляції $R = 0{,}927$ свідчить про щільність зв'язку залежної змінної Y з усіма пояснюючими змінними, включеними в модель.

Перевірка значущості рівняння регресії

Про значущість рівняння регресії свідчить **Знч** < 0,001.

Розраховане значення критерію Фішера $F_{емп} = 39{,}639$. Кількість ступенів свободи для критерію Фішера k та $n - k - 1$, де n – кількість спостережень, k – кількість факторів.

На заданому рівні значущості 0,95 $F_{табл}(2;13) = 19{,}4$. $F_{емп} > F_{табл}$, тому можна зробити висновок, що отримане рівняння регресії

$$Y = -1471{,}314 + 9{,}568X_1 + 15{,}754X_2$$

є значущим та адекватним.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-3017,396	1094,485		-2,757	,020
	Період (міс.)	-13,419	10,378	-,621	-1,293	,225
	Витрати на рекламу (тис. дол.)	6,672	3,009	,319	2,218	,051
	Ціна (дол.)	-6,477	15,779	-,067	-,410	,690
	Середня ціна у конкурентів (дол.)	12,238	14,410	,136	,849	,416
	Індекс споживчих витрат (%)	30,476	11,525	1,337	2,644	,025
2	(Constant)	-2914,326	1024,234		-2,845	,016
	Період (міс.)	-12,569	9,778	-,582	-1,285	,225
	Витрати на рекламу (тис. дол.)	7,125	2,691	,340	2,648	,023
	Середня ціна у конкурентів (дол.)	7,930	9,492	,088	,835	,421
	Індекс споживчих витрат (%)	29,151	10,637	1,279	2,740	,019
3	(Constant)	-2957,613	1009,969		-2,928	,013
	Період (міс.)	-14,316	9,431	-,663	-1,518	,155
	Витрати на рекламу (тис. дол.)	7,229	2,654	,345	2,724	,018
	Індекс споживчих витрат (%)	30,951	10,285	1,358	3,009	,011
4	(Constant)	-1471,314	259,766		-5,664	,000
	Витрати на рекламу (тис. дол.)	9,568	2,266	,457	4,223	,001
	Індекс споживчих витрат (%)	15,753	2,467	,691	6,386	,000

a. Dependent Variable: Обсяг реалізації (млн. дол.)

Рис. 3.25. Таблиця коефіцієнтів та оцінок регресійної моделі

Знаходження прогнозних значень результатної ознаки за регресійною моделлю

Для отримання прогнозних значень обсягів реалізації за відомими значеннями значущих факторів $X_2 = 5{,}75$ та $X_5 = 112{,}468$ потрібно ввести їх значення у початковий файл даних (рис. 3.26).

Для побудови прогнозу необхідно вибрати значущі змінні оптимальної моделі та передбачити збереження інтервалів прогнозу для окремих значень з вірогідністю 90%.

Y	X1	X2	X3	X4	X5
126	1	4,0	15,0	17,0	100,0
137	2	4,8	14,8	17,3	98,4
148	3	3,8	15,2	16,8	101,2
191	4	8,7	15,5	16,2	103,5
274	5	8,2	15,5	16,0	104,1
370	6	9,7	16,0	18,0	107,0
432	7	14,7	18,1	20,2	107,4
445	8	18,7	13,0	15,8	108,5
367	9	19,8	15,8	18,2	108,3
367	10	10,6	16,9	16,8	109,2
321	11	8,6	16,3	17,0	110,1
307	12	6,5	16,1	18,3	110,7
331	13	12,6	15,4	16,4	110,3
345	14	6,5	15,7	16,2	111,8
364	15	5,8	16,0	17,7	112,3
384	16	5,7	15,1	16,2	112,9
.	.	5,8	.	.	112,5

Рис. 3.26. Фрагмент файла даних з відомими значеннями факторів, за якими буде отримано прогноз

Отримано результати прогнозування за моделлю регресії: точковий прогноз, верхня та нижня межі (рис. 3.27).

PRE_1	RES_1	LICI_1	UICI_1
142,24674	-16,24674	59,70119	224,79229
124,69687	12,30313	40,05231	209,34143
159,23651	-11,23651	78,03246	240,44055
242,35334	-51,35334	165,01260	319,69408
247,02086	26,97914	170,17685	323,86487
307,05682	62,94318	231,32030	382,79333
361,20003	70,79997	282,49880	439,90127
416,80185	28,19815	332,70365	500,90006
424,17653	-57,17653	338,01935	510,33371
350,32471	16,67529	274,11289	426,53653
345,36547	-24,36547	268,45428	422,27666
334,72353	-27,72353	256,11282	413,33424
386,78970	-55,78970	309,39496	464,18444
352,05169	-7,05169	272,19034	431,91303
353,23023	10,76977	272,06331	434,39716
361,72512	22,27488	279,57913	443,87110
355,39830	.	273,94332	436,85327

Рис. 3.27. Результати прогнозування

З вірогідністю 90% обсяг реалізації в прогнозованому місяці становитиме від 273,94 до 436,85 млн. дол.

3.3.2. Нелінійні регресійні моделі

У багатьох практичних застосуваннях моделювання економічних залежностей лінійними рівняннями дає задовільний результат. Більшість лінійних взаємозв'язків випливає із відомих закономірностей. Однак при нелінійному співвідношенні економічних показників використання методів лінійного регресійного аналізу приводить до помилок результатів та спрощених або навіть неправильних висновків, отриманих на основі аналітичного рівняння.

У випадках, коли між економічними явищами існують нелінійні співвідношення, їх представляють за допомогою відповідних нелінійних функцій.

Класи нелінійних регресій

1. Нелінійні регресії – за пояснюючими змінними, включеними в рівняння, але лінійні за оцінюваними параметрами. Цей клас нелінійних регресій містить рівняння, в яких залежна змінна лінійно пов'язана з параметрами [54, с. 42].

Наприклад:

– поліноми різних степенів $y_i = a_0 + a_1 x_i + a_2 x_i^{\,2} + a_3 x_i^{\,3} + \ldots + a_k x_i^{\,k} + \varepsilon_i$,

– рівностороння гіпербола $y_i = a + \frac{b}{x_i} + \varepsilon_i$.

2. Регресії, нелінійні за оцінюваними параметрами. До цього класу регресій відносять рівняння, в яких залежна змінна нелінійно пов'язана з параметрами.

Наприклад:

► степенева функція $y_i = a x_i^b$;

► показникова функція $y_i = a b^{x_i}$.

Якщо нелінійна модель внутрішньо нелінійна, вона не може бути зведена до лінійної функції і для оцінювання її параметрів використовують ітеративні процедури, успішність яких залежить від вигляду рівнянь і особливостей використаного ітераційного підходу (приклад 3.15).

При оцінюванні параметрів регресій, нелінійних за пояснюючими змінними, використовують **підхід лінеаризації** (приклад 3.13),

який називають **«заміна змінних»**. Нелінійні пояснюючі змінні замінюють новими «лінійними» змінними та зводять нелінійну регресію до лінійної. До нової, «перетвореної» регресії можна застосовувати звичайний метод найменших квадратів (приклад 3.14).

Можливість переходу до лінійної моделі потрібно використовувати завжди, оскільки в цьому разі параметри регресії обчислюють безпосередньо, а не визначають за допомогою ітерацій.

Побудова нелінійних регресійних моделей у SPSS

Приклад 3.13. За умовними даними (табл. 3.16) побудувати модель залежності обсягів випуску продукції від температури (при використанні відповідного технологічного процесу) за допомогою полінома другого степеня $y_i = a_0 + a_1 x_i + a_2 x_i^2 + \varepsilon_i$. Початкові дані та результати розрахунків відобразити на графіку [31, с. 28; 54, с. 57].

Таблиця 3.16

Умовні дані для прикладу 3.13

Обсяг випуску продукції, *Y*	Температура, *X*
127	600
139	625
147	650
147	675
155	700
154	725
153	750
148	115
146	800
136	825
129	850

Спосіб 1. Використання заміни змінних

Введемо нову змінну Z $(z_i = x_i^2)$ (рис. 3.28).

Обсяг_продукції_Y	Температура_X	Z
127	600	360000
139	625	390625
147	650	422500
147	675	455625
155	700	490000
154	725	525625
153	750	562500
148	775	600625
146	800	640000
136	825	680625
129	850	722500

Рис. 3.28. Фрагмент файла даних для лінеаризованої регресійної моделі

Отримано двофакторне рівняння лінійної регресії $y_i = a_0 + a_1 x_i + a_2 z_i + \varepsilon_i$, для оцінювання параметрів якого застосуємо звичайний метод найменших квадратів. Використовуючи лінійну регресію, отримаємо модель $Y = -712{,}105 + 2{,}391X_1 - 0{,}002X_2$ (рис. 3.29).

Коэффициенты[a]

Модель		Нестандартизованные коэффициенты		Стандартизованные коэффициенты	t	Знч.	95,0%% доверительный интервал для B	
		B	Стд. Ошибка	Бета			Нижняя граница	Верхняя граница
1	(Константа)	-712,105	54,372		-13,097	,000	-837,486	-586,724
	Температура_X	2,391	,151	20,426	15,810	,000	2,042	2,740
	Z	-,002	,000	-20,465	-15,840	,000	-,002	-,001

a. Зависимая переменная: Обсяг_продукції_Y

Рис. 3.29. Параметри та оцінки лінійної регресійної моделі

Спосіб 2. Використання нелінійної регресії

Побудуємо нелінійну модель за даними табл. 3.16 (рис. 3.30):

Имя	Тип	Ширина	Десятич...	Метка
Y	Числовая	10	0	Обсяг_продукції
X	Числовая	10	0	Температура

Рис. 3.30. Структура файла даних для нелінійного регресійного аналізу

Побудуємо нелінійну регресійну модель такого вигляду: $a_0 + a_1 x + a_2 x^2$.

Представлення результатів нелінійної регресії

Для досягнення заданого рівня точності виконано вісім ітераційних кроків (рис. 3.31).

История итераций[b]

Номер итерации[a]	Сумма квадратов остатков	Параметр		
		a0	a1	a2
1.0	228175,000	,000	,000	,000
1.1	94282,367	17,719	,024	3,138E-5
2.0	94282,367	17,719	,024	3,138E-5
2.1	2156,789	71,989	,066	4,077E-5
3.0	2156,789	71,989	,066	4,077E-5
3.1	805,464	80,257	,182	,000
4.0	805,464	80,257	,182	,000
4.1	620,192	-19,944	,463	,000
5.0	620,192	-19,944	,463	,000
5.1	325,646	-221,846	1,026	-,001
6.0	325,646	-221,846	1,026	-,001
6.1	38,325	-625,657	2,150	-,001
7.0	38,325	-625,657	2,150	-,001
7.1	29,105	-712,105	2,391	-,002
8.0	29,105	-712,105	2,391	-,002
8.1	29,105	-712,105	2,391	-,002

Производные вычисляются числовым способом.

Рис. 3.31. Таблиця ітерацій

Обчислені кінцеві значення всіх трьох параметрів моделі разом з відповідною стандартною помилкою та довірчим інтервалом (рис. 3.32):

Оценки параметра

Параметр	Оценка	Стд. Ошибка	Доверительный интервал 95 %	
			Нижняя граница	Верхняя граница
a0	-712,105	54,372	-837,486	-586,724
a1	2,391	,151	2,042	2,740
a2	-,002	,000	-,002	-,001

Рис. 3.32. Оцінки параметрів нелінійної регресійної моделі

У результаті ітераційних обчислень отримаємо таку саму модель, що й при обчисленнях першим способом: $Y = –712{,}105 + 2{,}391\ X_1 – 0{,}002X_2$.

Розрахункові значення моделі (рис. 3.33).

Y	X	PRED_	RESID
127	600	128,48	-1,48
139	625	137,72	1,28
147	650	144,90	2,10
147	675	150,01	-3,01
155	700	153,06	1,94
154	725	154,04	-,04
153	750	152,97	,03
148	775	149,83	-1,83
146	800	144,62	1,38
136	825	137,36	-1,36
129	850	128,03	,97

Рис. 3.33. Прогнозні значення, отримані за нелінійною регресійною моделлю

Графік початкових даних та отриманої моделі (рис. 3.34).

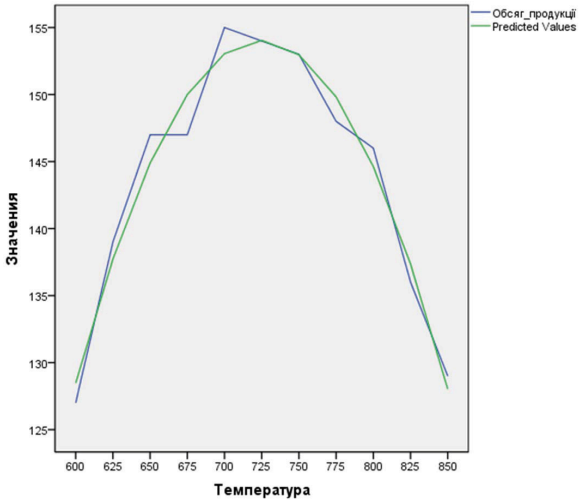


Рис. 3.34. Графік початкових даних та нелінійної регресійної моделі

Приклад 3.14. За умовними даними (табл. 3.17) побудувати виробничу функцію, яка описує залежність обсягів випуску хімічної промисловості від факторів «праця» та «капітал»:

$$Y = a_0 K^{a_1} L^{a_2},$$

де Y – обсяг промислового випуску хімічної промисловості у грошовому еквіваленті;

K – основні фонди в грошовому вираженні;

L – трудовий фактор – чисельність промислово-виробничого персоналу (ПВП) у хімічній промисловості.

Таблиця 3.17

Умовні дані для прикладу 3.14

Роки	Обсяг промисловості (млн. дол.), Y	Основний капітал, (млрд. дол.), K	Чисельність ПВП (чол.), L
2006	94,7	0,2	5636
2007	112,4	0,3	5520
2008	224500	1160,6	4502
2009	302000	1160,6	4515
2010	486900	1159,3	5454
2011	283	2,4	4109
2012	392,2	2,4	3897
2013	493,9	2,5	3958
2014	696,4	1,6	4086

Спосіб 1. Використання заміни змінних

Щоб провести лінеаризацію моделі за параметрами, потрібно прологарифмувати обидві частини рівняння $Y = a_0 K^{a_1} L^{a_2}$:

$$\ln Y = \ln a_0 + \ln a_1 K + a_2 \ln L.$$

На етапі моделювання вільний член $\ln a_0$ виключають як параметр, який погіршує статистичні властивості моделі. У вихідній мультиплікативній моделі a_0 беруть відповідно за одиницю. Цей параметр в економічній літературі інтерпретують як коефіцієнт нейтрального технічного прогресу [36, с. 203].

Результати побудови лінійної регресійної моделі (рис. 3.35).

Коэффициенты[a]

Модель		Нестандартизованные коэффициенты		Стандартизованные коэффициенты	t	Знч.
		B	Стд. Ошибка	Бета		
1	(Константа)	-14,107	7,996		-1,764	,128
	lnK	,985	,039	,990	24,978	,000
	lnL	2,339	,948	,098	2,467	,049

a. Зависимая переменная: lnY

Рис. 3.35. Обчислені коефіцієнти лінеаризованого рівняння регресії

Отримано рівняння $\ln \hat{Y} = 0{,}985 \ln K + 2{,}339 \ln L$.

Перейдемо до початкових змінних від прогнозних (змінна *PRE_1*) за допомогою функції *EXP*(*PRE_1*) і отримаємо виробничу функцію:

$$Y = K^{0{,}985} + L^{2{,}339}.$$

Спосіб 2. Використання нелінійної регресії

Задамо модель виразом $Y = K^{a_1} L^{a_2}$.

Представлення результатів нелінійно регресії

Для досягнення заданого рівня точності виконано 19 ітераційних кроків. Отримано оцінки параметрів моделі, відповідну стандартну похибку та довірчий інтервал (рис. 3.36).

Оценки параметра

Параметр	Оценка	Стд. Ошибка	Доверительный интервал 95 %	
			Нижняя граница	Верхняя граница
a1	,474	,256	-,131	1,079
a2	1,109	,212	,607	1,610

Рис. 3.36. Оцінки параметрів моделі

Отримаємо виробничу функцію з коефіцієнтами a_1 та a_2, які відрізняються від отриманих способом заміни змінних: $Y = K^{0{,}474} + L^{1{,}109}$.

Порівняємо графік початкових даних та графіки значень, отриманих за двома побудованими регресійними моделями (рис. 3.37).

Друга модель краще апроксимує початкові дані (мінімізується відхилення моделі від початкових даних, а не від їх логарифмів).

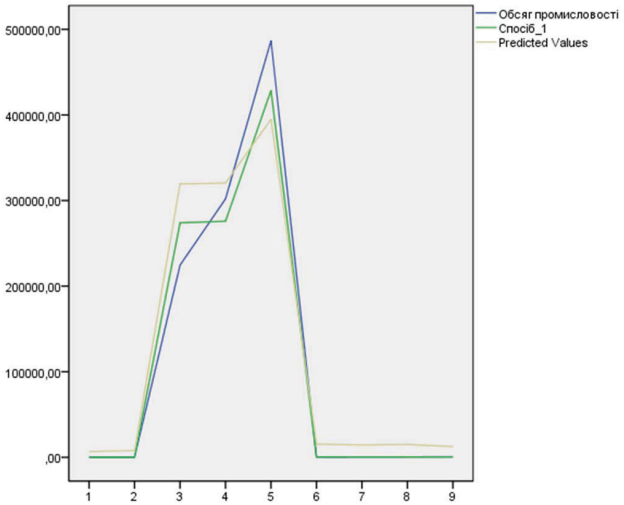


Рис. 3.37. Графіки початкових даних і розрахованих значень способами 1 та 2

Приклад 3.15. За умовними даними про норму безробіття та темпи інфляції (табл. 3.20) побудувати діаграму розсіювання та модель регресії залежності темпів інфляції від норми безробіття. Візуалізувати результати розрахунків.

Таблиця 3.18

Умовні дані для прикладу 3.15

Темпи інфляції, ***Y***	1,0	1,1	1,2	1,3	1,7	2,9	2,9	4,2	5,4
Норма безробіття, ***X***	6,5	5,4	5,5	5,0	4,4	3,7	3,7	3,5	3,4

За даними табл. 3.18 отримаємо лінійну регресійну модель $\hat{y}_i = 7{,}955 - 1{,}214x_i$ (рис. 3.38).

Коэффициенты[a]

Модель		Нестандартизованные коэффициенты		Стандартизованные коэффициенты	t	Знч.
		B	Стд. Ошибка	Бета		
1	(Константа)	7,955	1,345		5,915	,001
	Норма безробіття	-1,214	,287	-,848	-4,226	,004

a. Зависимая переменная: Темпи інфляції

Рис. 3.38. Обчислені коефіцієнти рівняння лінійної регресії

Побудуємо діаграму розсіювання для змінних X (відсоток безробітних у загальній чисельності робочої сили) та Y (темп інфляції) на основі даних табл. 3.18 (рис. 3.39).

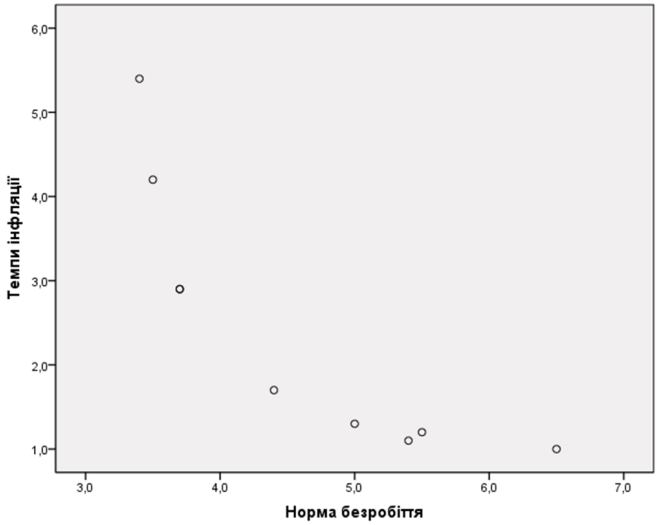


Рис. 3.39. Діаграма розсіювання

Характер розташування точок на діаграмі розсіювання свідчить про наявність нелінійного зв'язку між заданими змінними *темпи інфляції* та *норми безробіття*. Очевидно, що пряма $\hat{y}_i = 7{,}955 - 1{,}214x_i$, підібрана методом найменших квадратів, не відповідає характеру статистичних даних, хоча значення коефіцієнта детермінації $R^2 = 0{,}718$ є достатньо високим (рис. 3.40).

Сводка для модели

Модель	R	R-квадрат	Скорректированный R-квадрат	Стд. ошибка оценки
1	,848[a]	,718	,678	,8864

а. Предикторы: (конст) Норма безробіття

Рис. 3.40. Зведення лінійної регресійної моделі

При підборі моделей до реальних статистичних даних потрібно аналізувати не лише коефіцієнт детермінації, а й відповідність підібраної моделі характеру вихідної інформації. На практиці підбір найкращої моделі регресії виконують за допомогою спеціальних програмних продуктів, зокрема «Statistica», «SPSS», «СТАТЭКСПЕРТ» і т. ін.

Побудова нелінійної моделі темпів інфляції

Побудовано модель з великим коефіцієнтом детермінації R^2 = 0,998 (рис. 3.41):

$$\hat{y}_i = \frac{x_i}{-6{,}321 + 2{,}045 x_i}.$$

Оценки параметра

	Параметр	Оценка	Стд. Ошибка	Доверительный интервал 95 %	
				Нижняя граница	Верхняя граница
Асимптотический	a1	-6,321	,149	-6,673	-5,969
	a2	2,045	,043	1,943	2,147

Рис. 3.41. Оцінки параметрів регресійної моделі

Отримано прогнозні значення за нелінійною регресійною моделлю (рис. 3.42).

Y	X	PRED_	RESID
1,0	6,5	,93	,07
1,1	5,4	1,14	-,04
1,2	5,5	1,12	,08
1,3	5,0	1,28	,02
1,7	4,4	1,64	,06
2,9	3,7	2,97	-,07
2,9	3,7	2,97	-,07
4,2	3,5	4,18	,02
5,4	3,4	5,38	,02

Рис. 3.42. Обчислені прогнозні значення та залишки

Графік моделі свідчить про відповідність підібраної моделі характеру статистичних даних (рис. 3.43).

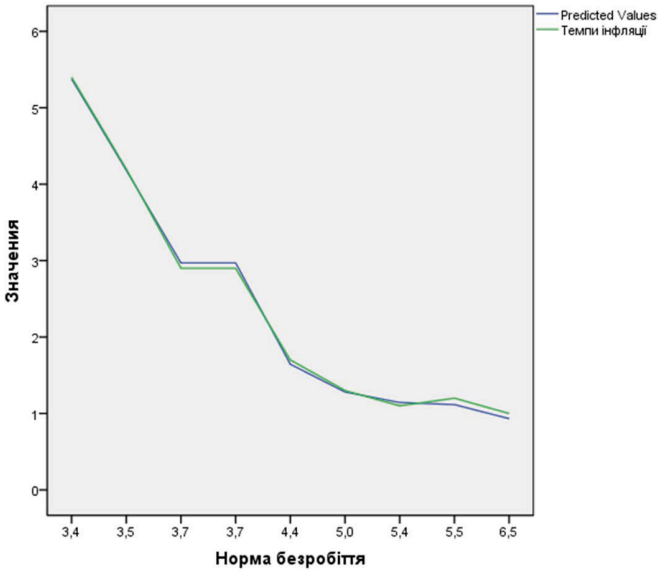


Рис. 3.43. Графіки побудованої моделі та початкових даних

Метод підгонки кривих

Метод підгонки кривих є одним з найбільш відомих методів прогнозування. Він полягає у визначенні кривої (групи кривих), яка з заданою точністю описує вихідний динамічний ряд [31, с. 8; 54, с. 42].

Основні види кривих підгонки:

– лінійна $\overline{y} = b_0 + b_1 x$;

– квадратична $\overline{y} = b_0 + b_1 x + b_2 x^2$;

– кубічна $\overline{y} = b_0 + b_1 x + b_2 x^2 + b_3 x^3$;

– логарифмічна $\overline{y} = b_0 \ln(x) + b_1$;

– експонентна $\overline{y} = b_0 \exp^{b_1 x}$;

– степенева $\overline{y} = b_0 x^{b_1}$;

– логістична $\overline{y} = \dfrac{1}{\dfrac{1}{b_0} + b_1 b_2^x}$;

– *S*-подібна $\overline{y} = \exp^{\left(b_0 + \frac{b_1}{x}\right)}$.

Приклад 3.16. За даними динамічного ряду (табл. 3.19) отримати прогнозні значення на наступних 3 періоди методом підгонки кривих.

Таблиця 3.19

Індекс людського розвитку Великобританії за 2008–2018 рр.

Рік	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Значення ІЛР	0,940	0,888	0,885	0,887	0,890	0,890	0,895	0,891	0,890	0,892	0,892

За зведенням моделі (рис. 3.44) можна зробити висновок, що оптимальною регресійною моделлю є кубічна (*max* R^2 = 0,73).

Сводка модели и оценки параметров

Зависимая переменная:ІЛР

Уравнение	Сводка для модели					Оценки параметра			
	R-квадрат	F	ст.св.1	ст.св.2	Знч.	Константа	b1	b2	b3
Линейный	,149	1,575	1	9	,241	,905	-,002		
Логарифмическая	,370	5,282	1	9	,047	,914	-,012		
Обратная	,679	19,074	1	9	,002	,882	,047		
Квадратичный	,430	3,018	2	8	,106	,928	-,012	,001	
Кубический	,730	6,297	3	7	,021	,965	-,043	,007	,000
Составная	,146	1,543	1	9	,246	,905	,998		
Степенная	,366	5,198	1	9	,049	,914	-,014		
S	,676	18,757	1	9	,002	-,126	,051		
Роста	,146	1,543	1	9	,246	-,100	-,002		
Экспоненциальная	,146	1,543	1	9	,246	,905	-,002		
Логистическая	,146	1,543	1	9	,246	1,105	1,002		

Рис. 3.44. Параметри регресійних моделей

Обчислимо за цією моделлю прогнозні значення ІЛР Великобританії на 2019–2021 рр.: точковий прогноз, верхня та нижня межі (рис. 3.45).

FIT_1	LCL_1	UCL_1
,90345	,86303	,94388
,90167	,86252	,94082
,89989	,86176	,93802

Рис. 3.45. Отримані прогнозні значення за оптимальною моделлю

Таблиця дисперсійного аналізу для оптимальної моделі (рис. 3.46).

Дисперсионный анализ

	Сумма квадратов	ст.св.	Средний квадрат	F	Знч.
Регрессия	,002	3	,001	6,297	,021
Остаток	,001	7	,000		
Итого	,002	10			

Рис. 3.46. Таблиця дисперсійного аналізу

Перевірка значущості рівняння регресії

Результати регресійного аналізу є достовірними (**Знч** < 0,021).

Розраховане зачення критерію Фішера $F_{емп} = 6,297$. На заданому рівні значущості 0,95 $F_{табл}(3; 7) = 4,35$. $F_{емп} > F_{табл}$, тому можна зробити висновок, що отримане рівняння регресії є значущим та адекватним, а модель – значущою.

Рівняння оптимальної регресійної моделі:

$$\overline{y} = 0,965 - 0,043x + 0,007x^2 + 0 \cdot x^3 .$$

З вірогідністю 95% Індекс людського розвитку Великобританії у 2019 р. становитиме від 0,86303 до 0,94388; у 2020 р. – від 0,86252 до 0,94082; у 2021 р. – від 0,86176 до 0,93802.

3.3.3. Логістична регресія

У регресійному аналізі передбачається, що результатний показник y_i є кількісною величиною, яка може набувати будь-яких значень на множині дійсних чисел. Проте в економічних і соціальних дослідженнях часто виникають завдання, в яких залежна змінна може набувати тільки цілочисельних значень [54, с. 57].

Залежно від типу змінних моделі з дискретними залежними змінними поділяють на моделі вибору серед скінченної кількості альтернативних варіантів і моделі порядкових даних. До останніх належать моделі, де Y – це, наприклад, кількість компаній, які стали банкрутами в поточному році, або кількість акцій, проданих окремою ТНК за досліджуваний період.

Залежно від кількості кінцевих варіантів, серед яких здійснюють вибір, розрізняють **моделі бінарного вибору** і **моделі множинного вибору**.

У моделях бінарного вибору результатний показник може набувати тільки двох значень – 0 або 1. Наприклад, результати голосування (за, проти), входження країни у міжнародне інтеграційне об'єднання (входить, не входить), рішення про купівлю товару (так, ні).

До моделей множинного вибору відносять моделі з невпорядкованими та впорядкованими альтернативними варіантами.

Бінарна логістична регресія

Досліджується залежність дихотомічної змінної (може набувати лише двох значень) від однієї або декількох незалежних змінних довільного типу.

При розв'язанні задачі класифікації на основі логістичної регресії застосовують ROC-аналіз. Останнім часом логістична регресія отримала поширення, наприклад, для розрахунку рейтингу політиків, при управлінні кредитними ризиками та в ряді інших прикладних завдань. Логістична регресія та ROC-аналіз входять у набір алгоритмів data mining (розділ 6).

Логістична регресія виражає статистичний зв'язок у вигляді залежності $P\{Y=1 \mid X\} = f(X)$, тобто прогнозується вірогідність події $\{Y = 1\}$, обумовлена значеннями незалежних змінних $X_1, X_2, ..., X_k$. Завданням логістичної регресії є замість передбачення значення бінарної змінної отримати прогноз неперервної змінної зі значеннями на відрізку [0, 1] при будь-яких значеннях незалежних змінних.

Логістична регресія виражає зв'язок між відгуком і факторами у вигляді залежності:

$$P\{Y = 1 \mid X_1, X_2, ..., X_k\} = \frac{e^{\hat{Y}}}{1 + e^{\hat{Y}}} = \frac{1}{1 + e^{-\hat{Y}}},$$

де $P\{Y = 1 \mid X_1, X_2, ..., X_k\}$ – ймовірність того, що очікувана подія відбудеться;

$e = 2,718\ldots$ – основа натуральних логарифмів;

$\hat{Y} = a_0 + a_1 X_1 + a_2 X_2 + ... + a_k X_k$ – лінійне рівняння регресії.

Назва «логістична регресія» походить від назви «логістичний розподіл» з функцією розподілу:

$$F(x) = \frac{e^x}{1 + e^x}. \qquad (3.4)$$

Модель, представлена цим видом регресії, є за суттю розподілом цього закону; аргумент функції – лінійна комбінація незалежних змінних.

Відношення ймовірності того, що подія відбудеться, до ймовірності того, що вона не відбудеться, $\frac{P}{1-P}$ називають **відношенням шансів** (ВШ).

З цим відношенням пов'язане ще одне представлення логістичної регресії. Розв'яжемо (3.4) відносно $\hat{Y}$ і отримаємо $\hat{Y} = \ln\left(\frac{P}{1-P}\right)$, де $P\{Y=1|X_1, X_2, ..., X_k\}$. Тоді відношення шансів можна записати у такому вигляді:

$$\frac{P}{1-P} = e^{a_0 + a_1X_1 + a_2X_2 + \ldots + a_kX_k} = e^{a_0} e^{a_1X_1} e^{a_2X_2} \cdots e^{a_kX_k} = e^{a_0} \left(e^{a_1}\right)^{X_1} \left(e^{a_2}\right)^{X_2} \cdots \left(e^{a_k}\right)^{X_k}.$$

Отже, якщо модель правильна, при незалежних змінних $X_1, X_2, ..., X_k$ зміна X_j на одиницю зумовить зміну відношення шансів у e^{a_j} разів.

Графік залежності, що пов'язує ймовірність події $P\{Y=1|\, X_1, X_2, ..., X_k\} = F(\hat{Y}) = \frac{e^{\hat{Y}}}{1+e^{\hat{Y}}}$ та величину $\hat{Y}$, представлено на рис. 3.47. Ця залежність має нелінійний характер, причому P не може виходити за межі діапазону [0, 1].

Якщо замість функції розподілу $F(Y) = \frac{e^{\hat{Y}}}{1+e^{-\hat{Y}}}$ вибрати функцію розподілу нормального закону $F(Y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Y} e^{\frac{-x}{2}} dx$, то замість логіт-моделі отримаємо близьку їй пробіт-модель.

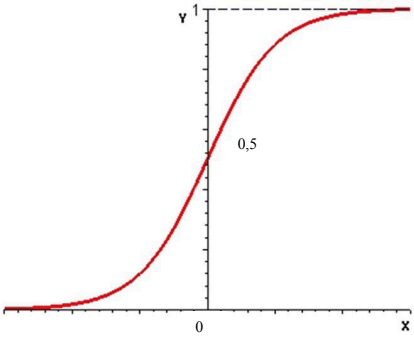


Рис. 3.47. Логістична крива

Probit- та logit-моделі застосовують при прогнозуванні бінарної величини для оцінювання якісних змінних, де використання лінійного

оцінювання ускладнене певними причинами. Наприклад, спрогнозувати шанси окремого політика (виграш або програш) відповідно до параметрів передвиборчої кампанії; встановити, чи впадуть обсяги продажів нижче від критичного рівня, якщо купівельна спроможність населення зменшиться на 5%; з'ясувати, чим інтереси тих, що віддали голос за відповідного кандидата, відрізняється від поведінки людей, що не проголосували за нього (результат голосування – залежна бінарна величина, політичні вподобання виборця, його стать, вік – фактори).

Встановити, у яких випадках потрібно застосовувати логіт-модель, а коли – пробіт-модель при малих вибірках неможливо, оскільки оцінки коефіцієнтів моделі $\hat{Y} = a_0 + a_1 X_1 + a_2 X_2 + ... + a_k X_k$ відрізняються практично постійним множником.

Існує декілька способів знаходження коефіцієнтів логістичної регресії. На практиці часто використовують метод максимальної правдоподібності. Його застосовують у математичній статистиці для отримання оцінок параметрів генеральної сукупності за даними вибірки. Основою методу є функція правдоподібності, яка виражає щільність імовірності (вірогідність) спільної появи результатів вибірки $Y_1, Y_2, ..., Y_k$:

$$L(Y_1, Y_2, ..., Y_k; \theta) = p(Y_1; \theta) \cdots p(Y_k; \theta).$$

Згідно з методом максимальної правдоподібності як оцінку Θ невідомого параметра θ беруть значення $\Theta = \Theta(Y_1, Y_2, ..., Y_k)$, яке максимізує функцію L.

Множинний логіт- чи пробіт-аналіз є логічним продовженням бінарного та виникає при розгляді вибору між більше, ніж двома альтернативами. Впорядкований логіт- чи пробіт-аналіз використовують при дослідженні альтернатив, які можна впорядкувати.

Мультиномінальна логістична регресія – це варіант логістичної регресії, при якій залежна змінна не є дихотомічною, а має більше двох категорій. Мультиномінальна логістична регресія придатна лише для категоріальних коваріат (незалежних змінних), причому важливо, чи є вони номінальними, чи порядковими. Не виключається можливість використання інтервальних коваріат.

Мультиномінальну регресію використовують для дослідження залежності номінальної результатної змінної. Порядкова регресія призначена для вивчення цільової порядкової змінної. Незалежні змінні мають бути категоріальними (номінальними або порядковими). Допускається застосування коваріат також з інтервальною шкалою.

Побудова логістичних моделей у програмі SPSS

Приклад 3.17. За умовними даними про 50 спеціалістів з розробки міжнародних проектів (табл. 3.20) встановити, чи можуть вибрані змінні (вік, наявність міжнародного досвіду, підвищення кваліфікації) бути використаними для прогнозування кількості нереалізованих ними проектів [31, с. 41; 54, с. 57].

Змінні моделі: Y – наявність реалізованих проектів; X_1 – вік спеціаліста; X_2 – досвід міжнародної співпраці (0 – немає досвіду роботи у міжнародних організаціях, 1– є досвід роботи у міжнародних організаціях); X_3 – підвищення кваліфікації (0 – не пройшов курси підвищення кваліфікації, 1 – пройшов перепідготовку).

Таблиця 3.20

Умовні дані для прикладу 3.17

№	Y	Вік, X_1	Досвід міжнародної співпраці, X_2	Проходження курсів підвищення кваліфікації, X_3
1	1	19	1	1
2	0	44	0	1
3	1	48	1	0
…	…	…	…	…
50	0	19	0	1

Структура файла даних (рис. 3.48).

Имя	Тип	Ширина	Десятич...	Метка
Y	Числовая	8	2	Нереалізовані проекти
X1	Числовая	8	2	Вік
X2	Числовая	8	2	Досвід міжнародної співпр...
X3	Числовая	8	2	Підвищення кваліфікації

Рис. 3.48. Структура файла даних для прикладу 3.8

Аналіз результатів

На нульовому кроці модель не побудована, всі «передбачені» значення Y дорівнюють одиниці, тому спостереження, в яких $Y = 1$, «передбачені» правильно, а спостереження, де $Y = 0$, – неправильно (рис. 3.49).

Таблица классификации[a,b]

Наблюденные			Предсказанные		
			Нереалізовані проекти		Процент корректных
			0	1	
Шаг 0	Нереалізовані проекти	0	0	25	,0
		1	0	25	100,0
	Общий процент				50,0

a. Модель включает константу.
b. Разделяющее значение = ,500

Рис. 3.49. Початковий крок реалізації бінарної логістичної моделі

На наступних кроках логістична модель буде передбачати шанси та ймовірність появи нереалізованих проектів на основі незалежних змінних рівняння регресії.

Якість наближення регресійної моделі оцінюють за допомогою функції правдоподібності. Мірою правдоподібності є від'ємне подвоєне значення логарифма цієї функції $-2\log(L)$ (рис. 3.50). Чим менше це значення, тим краще побудована бінарна логістична модель.

Сводка для модели

Шаг	-2 Log Правдоподобие	R квадрат Кокса и Снелла	R квадрат Нэйджелкерка
1	46,587[a]	,365	,487

a. Оценивание закончено на итерации номер 5, потому что оценки параметра изменились менее чем на ,001.

Рис. 3.50. Зведення для бінарної логістичної моделі

Як початкове значення для $-2\log$ використовують значення, отримане для регресійної моделі, яка містить тільки константи. Після додавання змінних впливу значення $-2\log$ дорівнює 46,587 (на 22,727 менше, ніж початкове). Таке зменшення величини означає покращення моделі. Різниця позначається як величина

χ-квадрат і є значущою (рис. 3.51). Це означає, що початкова модель після додавання змінних впливу стала значно кращою.

Объединенные тесты для коэффициентов модели

		Хи-квадрат	ст.св.	Знч.
Шаг 1	Шаг	22,727	3	,000
	Блок	22,727	3	,000
	Модель	22,727	3	,000

Рис. 3.51. Оцінки бінарної логістичної моделі

R-квадрат Кокса і Снелла та *R*-квадрат Нейджелкерка – псевдокоефіцієнти детермінації, отримані на основі відношення функцій правдоподібності моделей лише з константою та з усіма коефіцієнтами. Ці коефіцієнти рідко використовують при порівнянні моделей.

На основі моделі логістичної регресії можна будувати прогноз – відбудеться чи не відбудеться подія {*Y*–1}. Правило прогнозу, яке за замовчуванням використовується у процедурі логістичної регресії: якщо передбачена ймовірність події більша 0,5, вважають, що подія відбудеться; в протилежному разі роблять висновок, що подія не відбудеться.

За таблицею класифікації (крок не дорівнює нулю) визначають кількість правильно і неправильно передбачених відгуків у кожній з категорій аналізованої змінної та загальний відсоток коректних прогнозів (рис. 3.52).

Таблица классификации[a]

			Предсказанные		
			Нереалізовані проекти		
	Наблюденные		0	1	Процент корректных
Шаг 1	Нереалізовані проекти	0	20	5	80,0
		1	5	20	80,0
	Общий процент				80,0

a. Разделяющее значение = ,500

Рис. 3.52. Таблиця класифікації бінарної логістичної моделі

Аналіз змінних регресійної моделі, отриманої методом примусового включення (рис. 3.53).

Переменные в уравнении

		B	Стд.Ошибка	Вальд	ст.св.	Знч.	Exp(B)
Шаг 1[a]	X1	,014	,019	,557	1	,455	1,014
	X2(1)	-2,173	,763	8,115	1	,004	,114
	X3(1)	2,170	,763	8,092	1	,004	8,761
	Константа	-,744	1,031	,520	1	,471	,475

a. Переменные, включенные на шаге 1: X1, X2, X3.

Рис. 3.53. Параметри та оцінки логістичного рівняння регресії

B – коефіцієнти a_j регресійного рівняння;

Стд. Ошибка – міра мінливості коефіцієнтів a_j;

Вальд – критерій значущості коефіцієнтів регресії (чим вище його значення, тим вища значущість коефіцієнта регресії);

ст.св. – кількість ступенів вільності;

Знч. – значущість за критерієм Вальда (H_0: $a_j = 0$);

Exp(B) – відношення шансів e^{a_j} (використовують для інтерпретації результатів).

Спостережувану значущість коефіцієнтів обчислюють на основі статистики Вальда. Її універсальність дає можливість оцінити значущість не лише окремих змінних, а й категоріальних змінних загалом, незважаючи на те, що вони реструктуровані на індексні змінні. Статистика Вальда має розподіл *Xi*-квадрат. Кількість ступенів вільності дорівнює одиниці, якщо перевіряється гіпотеза про рівність нулю коефіцієнта при звичайній або індексній змінній, і для категоріальної змінної дорівнює кількості значень мінус одиниця (кількості відповідних індексних змінних). Квадратний корінь із статистики Вальда приблизно дорівнює відношенню величини коефіцієнта до його стандартної помилки (як і *t*-статистика у звичайній лінійній моделі регресії).

У таблиці коефіцієнтів (рис. 3.53) значущими є змінні X_2 та X_3. Змінна X_1 (вік) не ввійшла до рівняння.

У групі спеціалістів, в яких не було нереалізованих проектів, значно більше людей середнього віку, а в іншій групі надто багато наймолодших і найстарших осіб. Базуючись на цих результатах,

створюємо нову дихотомічну змінну X_4, яка дорівнює 0 для спеціалістів віком 21–65 років та одиниці в інших випадках.

Побудуємо рівняння логістичної регресії (модель 2), використовуючи нову змінну методом «примусового включення» та методами «включення» і «виключення».

Побудуємо рівняння логарифмічної регресії, яке містить змінну X_4.

Аналіз отриманих результатів

За трьома використаними методами отримано аналогічні моделі. Всі змінні в рівнянні є значущими.

Якість моделі 2 значно покращена, значення –2log у зведенні для моделі дорівнює 39,057 (рис. 3.54). Відсоток коректно передбачених подій збільшився з 80 до 86 (рис. 3.55).

Сводка для модели

Шаг	-2 Log Правдоподобие	R квадрат Кокса и Снелла	R квадрат Нэйджелкерка
1	39,057[a]	,454	,605

a. Оценивание закончено на итерации номер 6, потому что оценки параметра изменились менее чем на ,001.

Рис. 3.54. Зведення для логістичної моделі 2

Таблица классификации[a]

			Предсказанные		
			Нереалізовані проекти		
	Наблюденные		0	1	Процент корректных
Шаг 1	Нереалізовані проекти	0	21	4	84,0
		1	3	22	88,0
	Общий процент				86,0

a. Разделяющее значение = ,500

Рис. 3.55. Таблиця класифікації для логістичної моделі 2

Таблиця змінних рівняння, отримана методом включення (рис. 3.56):

Переменные в уравнении

		B	Стд.Ошибка	Вальд	ст.св.	Знч.	Exp(B)
Шаг 1[a]	X2	2,097	,646	10,529	1	,001	8,143
	Константа	-1,099	,471	5,431	1	,020	,333
Шаг 2[b]	X2	2,209	,758	8,494	1	,004	9,102
	X3	-2,209	,758	8,494	1	,004	,110
	Константа	-,104	,576	,033	1	,856	,901
Шаг 3[c]	X2	2,763	,938	8,679	1	,003	15,842
	X3	-2,199	,840	6,858	1	,009	,111
	X4	2,307	,915	6,355	1	,012	10,047
	Константа	-1,470	,885	2,759	1	,097	,230

Рис. 3.56. Таблиця «Змінні в рівнянні» для логістичної моделі 2

Побудовано рівняння регресійної логістичної моделі:

$$\hat{Y} = -1{,}47 + 2{,}76X_2 - 2{,}199X_3 + 2{,}307X_4 .$$

Якщо підставити відповідні значення незалежних змінних у це рівняння, результатом буде логарифм шансів не реалізувати проект. Щоб визначити шанси, необхідно піднести число *e* (основу натуральних логарифмів) до цього степеня. Розрахувати ймовірність того, що спеціаліст не реалізує свій проект, на основі інформації про досвід міжнародної співпраці, кваліфікацію та вік можна за допомогою співвідношення:

$$\text{ВШ} = \frac{P}{1-P},$$

де P – ймовірність не реалізувати проект.

Розв'яжемо це рівняння відносно P:

$$P = \frac{\text{ВШ}}{1+\text{ВШ}}$$

Використаємо його, щоб передбачити шанси та ймовірність реалізувати проект для спеціаліста, який не має досвіду міжнародної співпраці ($X_2 = 0$), не проходив курсів підвищення кваліфікації ($X_3 = 0$) та відноситься до вікової групи від 21 до 65 років ($X_4 = 0$):

$$Log(\text{ВШ}_{\text{нереаліз.проект}}) = \hat{Y} = -1{,}47 + 2{,}76X_2 - 2{,}199X_3 + 2{,}307X_4 =$$
$$= -1{,}47 + 2{,}76 \cdot 0 - 2{,}199 \cdot 0 + 2{,}307 \cdot 0 = -1{,}47;$$
$$\text{ВШ}_{\text{нереаліз.проект}} = \exp(-1{,}47) = 0{,}2299,$$

відповідна ймовірність не реалізувати проект

$$P_{\text{нереаліз.проект}} = \frac{0{,}2299}{1 + 0{,}2299} = 0{,}1869.$$

Аналогічно розрахуємо шанси та ймовірність не реалізувати проект для спеціаліста, який має досвід міжнародної співпраці ($X_2 = 1$):

$$Log(\text{ВШ}_{\text{нереаліз.проект}}) = \hat{Y} = -1{,}47 + 2{,}76X_2 - 2{,}199X_3 + 2{,}307X_4 =$$
$$= -1{,}47 + 2{,}76 \cdot 1 - 2{,}199 \cdot 0 + 2{,}307 \cdot 0 = 1{,}29;$$
$$\text{ВШ}_{\text{нереаліз.проект}} = \exp(1{,}29) = 3{,}6327;$$

$$P_{\text{нереаліз.проект}} = \frac{3{,}6327}{1 + 3{,}6327} = 0{,}7851.$$

Отже, шанси реалізувати проект залежно від наявності міжнародного досвіду спеціаліста з досвідом міжнародної співпраці збільшуються від 0,2299 до 3,6327 (майже в 16 разів). Зміна X_j на одиницю викликає зміну відношення шансів у e^{a_j} разів.

Кореляційна матриця (рис. 3.57) включається в протокол даних.

Якщо незалежні змінні $X_1, X_2, ..., X_k$ надто щільно пов'язані між собою (порушується умова їх незалежності), побудоване рівняння регресії може бути некоректним.

Корреляционная матрица

		Constant	X2	X3	X4
Шаг 1	Constant	1,000	-,729		
	X2	-,729	1,000		
Шаг 2	Constant	1,000	-,504	-,415	
	X2	-,504	1,000	-,301	
	X3	-,415	-,301	1,000	
Шаг 3	Constant	1,000	-,688	-,194	-,668
	X2	-,688	1,000	-,267	,467
	X3	-,194	-,267	1,000	-,156
	X4	-,668	,467	-,156	1,000

Рис. 3.57. Кореляційна матриця логістичної моделі

У результаті застосування логістичної регресії SPSS формує діаграму, в якій для позначення прогнозованої ймовірності використано два значення: 1 – проект не буде реалізовано і 0 – проект буде реалізовано (рис. 3.58). По горизонтальній осі відкладені значення прогнозної ймовірності, а по вертикальній осі – частоти. Окремі стовпці містять і 0, і 1, що свідчить про помилки у прогнозі. Числам 1 в правій частині діаграми і числам 0 в лівій її частині відповідають правильні пронози відносно реалізації проекту.

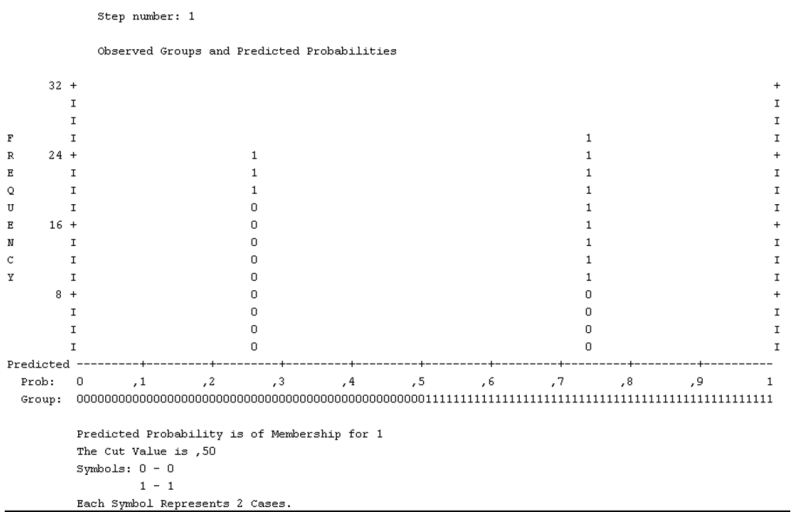


Рис. 3.58. Фактичне групування та прогнозування

За допомогою логістичної регресії можна прогнозувати лише ймовірність реалізації проекту конкретним спеціалістом залежно від значень (для цього спеціаліста) включених у модель змінних. Далі потрібно визначити **розділяюче значення** – поріг імовірності для прогнозування належності спеціаліста до групи тих, що мають реалізовані проекти чи не мають їх. За замовчуванням значення порогу встановлюється рівним 0,5.

Процедура ROC-аналізу корисна для оцінювання ефективності схем класифікації, які мають одну змінну з двома категоріями,

за якими класифікують об'єкти. Для представлення результатів бінарної класифікації використовують ROC-криву (рис. 3.60).

Із двох наявних класів один позначимо як клас з позитивними результатами, другий – з негативними. ROC-крива показує залежність кількості правильно класифікованих позитивних результатів від кількості неправильно класифікованих негативних результатів. У термінології ROC-аналізу перші називають **істинно позитивними**, другі – **помилково негативними**. При цьому передбачається, що у класифікатора є деякий параметр, варіюючи який отримують конкретне розбиття на два класи. Цей параметр часто називають **порогом** або **точкою відсікання**. Залежно від нього виходять різні величини помилок I та II роду. У логістичній регресії як параметр для розбиття на класи обирають розрахункове значення рівняння регресії (вільний член рівняння регресії). При цьому поріг відсікання змінюється від 0 до 1.

Для прикладу 3.20 таблиця класифікації (рис. 3.59) побудована на основі результатів класифікації за моделлю (поріг відсікання дорівнює 0,5) і за фактичною (об'єктивною) належністю спеціалістів до класів тих, що мали реалізовані проекти у минулому році (1,00) або не мали таких (0,00).

Таблица классификации[a]

Наблюденные			Предсказанные		
			Нереалізовані проекти		Процент корректных
			0	1	
Шаг 1	Нереалізовані проекти	0	21	4	84,0
		1	3	22	88,0
	Общий процент				86,0

a. Разделяющее значение = ,500

Рис. 3.59. Таблиця класифікації логістичної моделі

При проведенні ROC-аналізу використовують наступні позначення:

► ***TP*** (**True Positives**) – правильно класифіковані позитивні приклади **(істинно позитивні випадки)**;

► ***TN*** (**True Negatives**) – правильно класифіковані негативні приклади **(істинно негативні випадки)**;

► ***FN*** (**False Negatives**) – позитивні приклади, класифіковані як негативні **(помилка I роду)**; це так званий помилковий пропуск – подію, яка цікавить, помилково не виявляють (помилково негативні випадки);

► ***FP*** (**False Positives**) – негативні приклади, класифіковані як позитивні **(помилка II роду)**; помилкове виявлення, оскільки за відсутності події помилково виносять ухвалу про її наявність **(помилково позитивні випадки)**.

Що є позитивною подією, а що – негативною, залежить від конкретного завдання. У прикладі 3.17 прогнозуємо ймовірність впровадження проекту, тому позитивним результатом буде клас «проект реалізовано», негативним – «проект не реалізовано».

При аналізі частіше оперують не абсолютними показниками, а відносними – частками (rates), вираженими у відсотках.

Частка істинно позитивних випадків (True Positives Rate):

$$TPR = \frac{TP}{TP + FN} \cdot 100\% .$$

Частка помилково позитивних випадків (False Positives Rate):

$$FPR = \frac{FP}{TN + FP} \cdot 100\% .$$

Об'єктивну цінність кожного бінарного класифікатора визначають чутливість і специфічність моделі.

Чутливість (Sensitivity) – частка істинно позитивних випадків:

$$Se = TPR = \frac{TP}{TP + FN} \cdot 100\%$$

Специфічність (Specificity) – частка істинно негативних випадків, які були правильно ідентифіковані моделлю:

$$Sp = \frac{FP}{TN + FP} \cdot 100\% ,$$

$$FPR = 100\% - Sp.$$

Для прикладу 3.17:

$$Se = TPR = \frac{TP}{TP + FN} \cdot 100\% = \frac{21}{21 + 4} \cdot 100\% = 84\%,$$

$$Sp = \frac{FP}{TN + FP} \cdot 100\% = \frac{22}{22 + 3} \cdot 100\% = 88\%.$$

Модель з високою чутливістю часто дає істинний результат за наявності позитивного результату (виявляє позитивні приклади), а модель з високою специфічністю – за наявності негативного результату (виявляє негативні приклади).

Побудова ROC-кривої

Для кожного значення порогу відсікання, яке змінюється від 0 до 1, з кроком *dx* (наприклад, 0,01) розраховують значення чутливості *Se* і специфічності *Sp*. Як альтернатива порогом може бути кожне наступне значення змінної у вибірці.

Побудуємо графік залежності: по осі *Y* відкладемо чутливість *Se*, по осі *X* – величину 100% – *Sp* (або *FPR* – частку помилково позитивних випадків) (рис. 3.60).

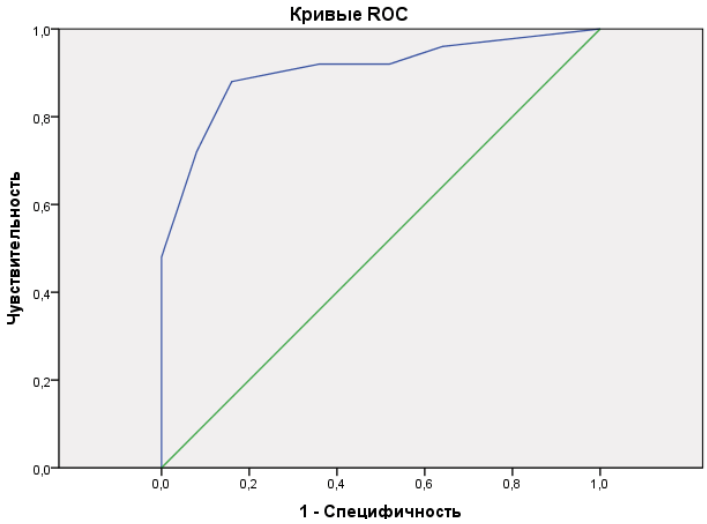


Рис. 3.60. ROC-крива для моделі 2

Для ідеального класифікатора графік ROC-кривої проходить через верхній лівий кут, де частка істинно позитивних випадків *TPR* складає 100% або 1 (ідеальна чутливість), а частка помилково позитивних прикладів *FPR* – 0. Тому чим ближче розташована крива до верхнього лівого кута, тим вища здатність прогнозної моделі. І навпаки, чим менший вигин кривої і чим ближче вона розташована до діагональної прямої, тим менш ефективна модель. Діагональна лінія відповідає «даремному» класифікаторові, тобто повній відсутності відмінності між двома класами.

Оцінка площі під ROC-кривими

Теоретично площа змінюється від 0 до 1, але, оскільки модель завжди характеризує крива, розташована вище від головної діагоналі, йдеться про зміни від 0,5 («даремний» класифікатор) до 1 («ідеальна» модель). Ця оцінка може бути отримана безпосередньо обчисленням площі багатокутника під експериментально отриманою ROC-кривою. Чисельний показник площі під кривою називають AUC (рис. 3.61).

Площадь под кривой

Тестовая переменная(ые):Предсказанная вероятность

Площадь	Стд. ошибка[a]	Асимптотическая Знч..[b]	Асимптотический 95% Доверительный интервал	
			Нижняя граница	Верхняя граница
,905	,045	,000	,817	,993

Рис. 3.61. Оцінка площі під ROC-кривою

Чим більший показник AUC, тим більшу прогностичну силу має модель. Однак AUC призначений здебільшого для порівняльного аналізу декількох моделей і не містить ніякої інформації про чутливість і специфічність моделі.

У науковій літературі визначена експертна шкала значень AUC, за якою можна робити висновки про якість моделі (табл. 3.21) [54, с. 86].

Таблиця 3.21

Оцінка якості моделі за значенням AUC

Інтервал AUC	0,9–1,0	0,8–0,9	0,7–0,8	0,6–0,7	0,5–0,6
Якість моделі	відмінна	дуже добра	добра	середня	незадовільна

Ідеальна модель володіє 100%-ю чутливістю та специфічністю. Проте на практиці досягти цього неможливо, також неможливо одночасно підвищити і чутливість, і специфічність моделі. Компроміс знаходять за допомогою порога відсікання, оскільки порогове значення впливає на співвідношення *Se* і *Sp*.

Поріг відсікання потрібний для того, щоб відносити нові приклади до одного з двох класів. Для визначення оптимального порога потрібно задати критерій, оскільки в різних завданнях різна оптимальна стратегія.

Критерії вибору порога відсікання:

– мінімальна чутливість (специфічність) моделі: для забезпечення чутливості тесту не менше 80% оптимальним порогом буде максимальна специфічність (чутливість), яка досягається при 80% чутливості (специфічності);

– максимальна сумарна чутливість і специфічність моделі; це значення порога, встановлене за замовчуванням;

баланс між чутливістю і специфічністю (*Se* ≈ *Sp*); порогом є точка перетину двох кривих: по осі *X* відкладають поріг відсікання, а по осі *Y* – чутливість або специфічність моделі (рис. 3.62).

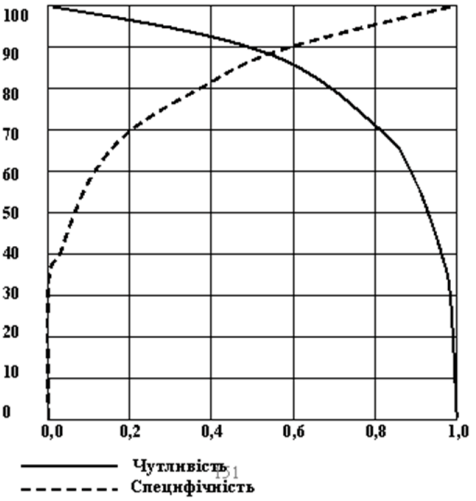


Рис. 3.62. «Точка балансу» між чутливістю та специфічністю

ПИТАННЯ ТА ЗАВДАННЯ ДЛЯ САМОПІДГОТОВКИ ТА САМОКОНТРОЛЮ

Теоретичні запитання

1. Визначення понять «прогноз» та «прогнозування».
2. Метод та методики прогнозування.
3. Основні функції прогнозування у МВ.
4. Принципи наукового прогнозування.
5. Пошуковий та нормативний прогнози.
6. Способи прогнозування.
7. Основні параметри прогнозів.
8. Основні групи факторів, що визначають становище держави на світовій арені.
9. Елімінування. Способи елімінування.
10. Методи експертних оцінок.
11. Коефіцієнт конкордації та коефіцієнт компетенції.
12. Кореляційний зв'язок між факторами.
13. Аналітичне групування.
14. Однофакторна лінійна регресія.
15. Коефіцієнт кореляції Пірсона.
16. *t*-критерій Стьюдента.
17. Модель множинної лінійної регресії.
18. Коефіцієнт детермінації.
19. *F*-критерій Фішера.
20. Параметричні та непараметричні тести.
21. Коефіцієнт асоціації та коефіцієнт контингенції.
22. Методи рангової кореляції.
23. Коефіцієнт кореляції рангів Спірмена.
24. Класи нелінійних регресій.
25. Метод підгонки кривих.
26. Логістична регресія.
27. Логіт- та пробіт-моделі.
28. Мультиноміальна логістична регресія.
29. Відношення шансів.
30. Чутливість та специфічність моделі.

Практичні завдання

Завдання 1. За даними спостережень (табл. 3.22) провести лінійну інтерполяцію наявної залежності обчислення значень для x = n, $x = n + 7$, x = $n + 11$ (n – номер варіанта).

Таблиця 3.22

Вихідні дані за варіантами для завдання 1

Ціна виробу	n **+ 10**	n **+ 13**	n **+ 17**	n **+ 19**	n **+ 21**	n **+ 23**	n **+ 27**	n **+ 29**
Кількість продажів	120	132	131	170	152	161	167	181

Завдання 2. Інтерполювати вихідні дані з попереднього завдання за допомогою кубічної сплайн-інтерполяції.

Завдання 3. Скласти таблицю значень функції на інтервалі від n до $n + 2$, $I = n\ldots. n + 13$, $x_i = n \cdot i + 10n$ (n – номер варіанта).

Завдання 4. Лінійне передбачення

Для попереднього завдання використати останні $n + 9$ точок, щоб обчислити коефіцієнти передбачення, та обчисли координати $n + 5$ точок вперед (n – номер варіанта).

Завдання 5. Побудувати інтерполяцію функції двох змінних $z(x, y) = nx^2 + ny$, $x = 0\ldots3$, $y = 1\ldots4$ (n – номер варіанта).

Побудувати графік поверхні, яку задають вихідні дані, та графік поверхні, яка задається інтерполяційною функцією.

Завдання 6. Проста лінійна регресія

За даними про щомісячні обсяги експорту-імпорту товарів за країнами світу в 2014 р. (за віріантом) провести порівняльний аналіз прогнозів експорту у наступні 3 роки, використовуючи метод підгонки кривих.

Завдання 6.1. За зведенням моделі визначити оптимальну регресійну модель та отримати для неї прогнозні значення.

Завдання 6.2. Для оптимальної моделі побудувати та пояснити таблицю дисперсійного аналізу.

Завдання 6.3. Записати рівняння оптимальної регресійної моделі та пояснити значення її коефіцієнтів.

Завдання 7. За умовними даними (табл. 3.23) побудувати регресійну модель, яка характеризує залежність обсягів продажів продукції на день Y (тис. грн.) від кількості днів рекламної кампанії X (дні).

Таблиця 3.23

Вихідні дані за варіантами для завдання 7

Варіант											
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	*X*	10	17	18	27	28	34	39	41	53	72
	Y	250	207	400	114	223	165	370	415	184	270
2	*X*	13	22	29	31	32	46	56	65	81	74
	Y	440	330	550	400	183	265	370	185	284	370
3	*X*	12	19	29	33	37	40	52	59	81	93
	Y	216	370	350	223	483	189	216	480	385	570
4	*X*	10	15	20	25	30	35	40	45	50	55
	Y	276	270	350	373	483	509	516	480	385	470
5	*X*	11	12	22	23	25	30	34	56	78	90
	Y	540	530	505	490	483	465	470	485	484	470
6	*X*	10	20	30	40	50	60	70	80	90	95
	Y	256	280	250	327	413	289	216	380	385	270
7	*X*	7	11	17	23	24	30	42	45	47	48
	Y	213	211	214	227	215	270	300	201	217	225
8	*X*	3	5	7	9	11	12	14	15	17	20
	Y	56	80	50	27	13	89	16	80	85	70
9	*X*	7	14	21	28	35	42	49	56	63	70
	Y	240	230	205	390	383	365	370	285	284	370
10	*X*	21	22	32	13	56	31	74	65	81	29
	Y	440	330	550	400	183	265	370	185	284	370
11	*X*	11	12	22	23	25	30	34	56	78	90
	Y	213	211	214	227	215	270	300	201	217	225
12	*X*	7	11	17	23	24	30	42	45	47	48
	Y	250	207	400	114	223	165	370	415	184	270
13	*X*	240	230	205	390	383	365	370	285	284	370
	Y	560	800	500	270	130	890	160	800	850	700

Продовження табл. 3.23

1	2	3	4	5	6	7	8	9	10	11	12
14	*X*	440	330	550	400	183	265	370	185	284	370
	Y	216	270	350	373	483	509	516	480	385	470
15	*X*	10	17	18	27	28	34	39	41	53	72
	Y	540	530	505	490	483	465	470	485	484	470
16	*X*	13	22	29	31	32	46	56	65	81	74
	Y	256	280	250	327	413	289	216	380	385	270
17	*X*	12	19	29	33	37	40	52	59	81	93
	Y	440	330	550	400	183	265	370	185	284	370
18	*X*	3	5	7	9	11	12	14	15	17	20
	Y	216	370	350	223	483	189	216	480	385	570
19	*X*	7	14	21	28	35	42	49	56	63	70
	Y	10	15	20	25	30	35	40	45	50	55
20	*X*	21	22	32	13	56	31	74	65	81	29
	Y	10	20	30	40	50	60	70	80	90	95

Завдання 8. Лінійна модель множинної регресії

За умовними даними про діяльність ТНК за 12 місяців минулого року (табл. 3.24) розробити проект модернізації ТНК, для чого необхідно: побудувати багатофакторну лінійну регресійну модель діяльності; визначити вплив факторних ознак на обсяг валової продукції; виявити найвпливовіші ознаки для визначення напрямків майбутньої модернізації.

X_1 – часовий фактор, порядковий номер місяця;

X_2 – фонди (тис. грн./робітника);

X_3 – фондовіддача (тис. грн. обсягу товарного продукту/тис. грн. основного фонду);

X_4 – продуктивність праці (тис. ум. од./робітника);

Y – валова продукція (тис. ум. од.).

Таблиця 3.24

Умовні дані для завдання 8

X_1	X_2	X_3	X_4	Y
1	328 + *n*	0,054 + 0,01·*n*	0,3 + 0,1·*n*	397 + 10·*n*
2	329 + *n*	0,101 + 0,01·*n*	0,6 + 0,1·*n*	670 + 10·*n*
3	329 + *n*	0,099 + 0,01·*n*	1,2 + 0,1·*n*	1209 + 10·*n*
4	347 + *n*	0,019 + 0,01·*n*	0,1 + 0,1·*n*	138 + 10·*n*
5	352 + *n*	0,065 + 0,01·*n*	0,3 + 0,1·*n*	378 + 10·*n*
6	370 + *n*	0,053 + 0,01·*n*	0,1 + 0,1·*n*	79 + 10·*n*
7	378 + *n*	0,178 + 0,01·*n*	2,3 + 0,1·*n*	1883 + 10·*n*
8	385 + *n*	0,174 + 0,01·*n*	2,6 + 0,1·*n*	2124 + 10·*n*
9	389 + *n*	0,298 + 0,01·*n*	5,5 + 0,1·*n*	5069 + 10·*n*
10	399 + *n*	0,195 + 0,01·*n*	2,4 + 0,1·*n*	2618 + 10·*n*

n – номер варіанта.

Завдання 8.1. Побудувати регресійну модель для прогнозування обсягів реалізації одного з видів продукції.

Завдання 8.2. Провести аналіз матриці коефіцієнтів парної кореляції.

Завдання 8.3. Побудувати лінійне рівняння регресії зі значущими факторами.

Завдання 8.4. Провести оцінювання параметрів регресії методом найменших квадратів.

Завдання 8.5. Пояснити отримані результати регресійного аналізу.

Завдання 9. У 10 компаніях вивчають взаємозв'язок між середньорічними витратами на рекламу X_1 (млн. грн.), рівнем затрат на проведення реклами X_2 (% до вартості реалізованої продукції) та вартістю реалізованої рекламної продукції Y (млн. грн.) (табл. 3.25).

Вважаючи, що між показниками існує лінійна залежність, визначити параметри рівняння регресії та оцінити адекватність обраної моделі.

Таблиця 3.25

Умовні дані для завдання 9

№ компанії	X_1	X_2	Y
1	3+*n*	4+*n*	20
2	3+*n*	3+*n*	25
3	5+*n*	3+*n*	20
4	6+*n*	5+*n*	30
5	7+*n*	10+*n*	32
6	6+*n*	12+*n*	25
7	8+*n*	12+*n*	29
8	9+*n*	11+*n*	37
9	9+*n*	15+*n*	36
10	10+*n*	15+*n*	40

n – номер варіанта.

Завдання 10. Для вихідних даних табл. 3.26 записати рівняння параболічної моделі регресії $y_i = a_0 + a_1 x_i + a_2 x_i^2 + \varepsilon_i$ та порівняти її з лінійною.

Таблиця 3.26

Умовні дані для завдання 10

x_i	1 + *n*	2 + *n*	3 + *n*	4 + *n*	5 + *n*	6 + *n*	7 + *n*	8 + *n*	9 + *n*	10 + *n*
y_i	3 + *n*	5 + *n*	7 + *n*	10 + *n*	12 + *n*	16 + *n*	18 + *n*	21 + *n*	24 + *n*	25 + *n*

n – номер варіанта.

Завдання 11. За умовними даними табл. 3.22 побудувати регресійну модель виробничої функції Кобба–Дугласа $Q = AK^{\alpha} \times L^{\beta}$, де A – виробничий коефіцієнт, що відображає пропорційність всіх функцій і змінюється при зміні базової технології (через 30–40 років);

Q – обсяг виробництва;

K – капітал;

L – праця;

α, β – коефіцієнти еластичності обсягу виробництва за витратами капіталу і праці.

Таблиця 3.27

Умовні дані для завдання 11

Рік	Обсяг виробництва (млн. дол.), *Q*	Затрати капіталу (млрд. дол.), *K*	Затрати праці (чол.), *L*
2006	94,7	0,2	5636
2007	112,4	0,3	5520
2008	224500	1160,6	4502
2009	302000	1160,6	4515
2010	486900	1159,3	5454
2011	283	2,4	4109
2012	392,2	2,4	3897
2013	493,9	2,5	3958
2014	696,4	1,6	4086

n – номер варіанта,

α = 0,5 (зростання витрат капіталу на 1% збільшує обсяг виробництва на 0,5%), β = 0,2.

Завдання 11.1. Використати заміну змінних.

Завдання 11.2. Використати нелінійну регресію.

Завдання 12. За умовними даними про експорт товарів у розрахунку на одну особу (табл. 3.28) побудувати регресійну модель, оцінити її статистичну значущість та адекватність.

Таблиця 3.28

Умовні дані для завдання 12

Рік	2012	2013	2014	2015	2016	2017	2018
Обсяг експорту товарів (млн. дол.)	84,6 + *n*	107,3 + *n*	109,2 + *n*	158,1 + *n*	137,1 + *n*	178,1 + *n*	201,7 + *n*

Завдання 12.1. Побудувати нелінійну регресійну модель.

Завдання 12.2. Оцінити коефіцієнт детермінації.

Завдання 12.3. Отримати прогнозні значення за нелінійною регресійною моделлю.

Завдання 12.4. Побудувати графіки регресійної моделі та початкових даних.

РОЗДІЛ 4

МОДЕЛЮВАННЯ РЯДІВ ДИНАМІКИ

4.1. Аналіз інтенсивності та тенденцій розвитку

При дослідженні динаміки міжнародних соціально-економічних та політичних процесів за допомогою методів моделювання вирішують важливі для дослідження МВ завдання. Серед інших оцінюють інтенсивність динаміки, виявляють та описують тенденції, структурні зрушення, стаціонарність і коливання рядів, виявляють значущі фактори впливу. Вивчення процесів соціально-економічного розвитку у часі проводять на основі аналізу рядів динаміки [19, с. 412].

Динамічний (часовий) ряд – сукупність значень статистичних показників, розташованих у хронологічному порядку, які характеризують зміну окремоого соціально-економічного чи політичного явища. Такий ряд представляють у вигляді таблиці чи графіка (табл. 4.1, рис. 4.1).

Таблиця 4.1.

Динаміка отриманих прохань про надання притулку в країнах ЄС

Роки	2010	2011	2012	2013	2014	2015	2016	2017
К-сть біженців, тис. чол.	261	310	336	432	627	1322	672	538

Виділяють такі елементи динамічного ряду:

► рівні ряду – числові значення відповідних статистичних показників, які характеризують об'єкт дослідження (абсолютні, відносні або середні величини);

► моменти (хронологічні дати) або інтервали часу, яким відповідають рівні ряду.

За ознакою часу виділяють такі динамічні ряди:

– інтервальні – характеризують агреговані розміри явищ за визначені періоди часу (доба, календарний місяць, декада, рік);

– моментні – характеризують стан явища на конкретний момент часу.

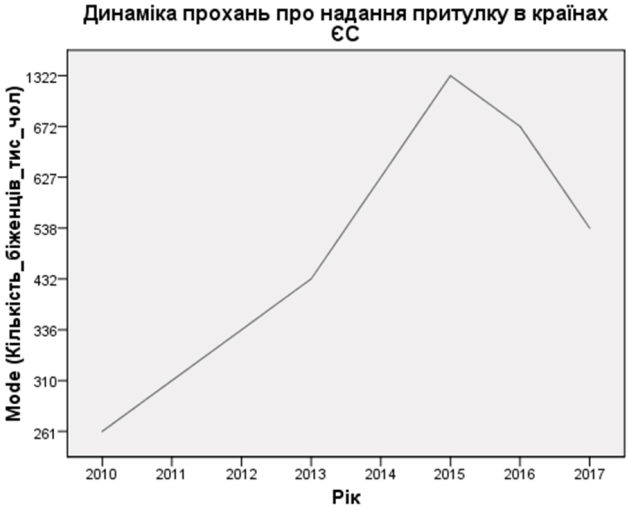


Рис. 4.1. Аналіз міграційних даних Євростату

Передумовою аналізу динамічних рядів є порівнянність їх рівнів. Необхідні вимоги зіставності часових рядів [18, с. 157]:

► рівність періодів, за які обчислено статистичні показники;

► однакова повнота охоплення досліджуваних частин явища;

► збіг територіальних меж явища;

► співвимірність рівнів ряду (однаковий масштаб та однакові одиниці виміру);

► єдине тлумачення одиниці об'єкта спостереження, єдина методологія розрахунку показників протягом досліджуваного періоду.

Завдання МВ, які вирішують за допомогою дослідження часових рядів:

– визначення параметрів інтенсивності зміни рівнів ряду між періодами спостереження;

– визначення середніх характеристик динамічного ряду за конкретний період;

– виявлення тенденцій розвитку досліджуваного соціально-економічного чи політичного явища загалом за визначений інтервал часу та на окремих його етапах зокрема;

– прогнозування динаміки явища.

Характеристики інтенсивності динаміки

Швидкість та інтенсивність розвитку процесів у МВ є мінливими (наприклад, динаміка темпів інфляції, міграційних процесів, курсів валют або темпів зростання ВВП). Для їх оцінювання використовують статистичні показники динамічних рядів, зокрема абсолютний приріст, відносний приріст, темп зростання і т. ін.

Основні характеристики часових рядів:

► показники зміни рівнів динамічного ряду (визначають напрям, швидкість та інтенсивність змін);

► середні характеристики рівнів динаміки.

Обчислення характеристик динаміки виконують за допомогою порівняння рівнів ряду [2, с. 67]. Рівень, який зіставляють, називають поточним (y_i), а рівень з яким зіставляють, – **базисним** (y_0). При порівнянні сукупності послідовних рівнів база порівняння може бути постійною чи змінною. Як постійну базу обирають частковий рівень ряду або початковий для розвитку досліджуваного явища рівень (наприклад, перше число місяця чи початок календарного року). Характеристики динаміки, обчислені відносно постійної бази (y_0), називають **базисними**. Якщо кожний рівень ряду порівнюють з попереднім (y_{i-1}), показники динаміки називають **ланцюговими**.

При зіставленні поточних рівнів обчислюють показники зміни рівнів ряду динаміки або їхні середні характеристики.

Показники зміни рівнів часового ряду:

– абсолютний приріст $\Delta_i = y_i - y_0$;

– темп зростання (зміни) $\text{Тз} = \frac{y_i}{y_0}$ у вигляді коефіцієнта чи у %;

– темп приросту $\text{Тпр} = \frac{\Delta_i}{y_0} = \text{Тз}\% - 100\%$ – відносна швидкість зростання (зменшення);

– абсолютне значення одного відсотка приросту $A = \frac{\Delta_i}{\text{Тз}}$ – сота частина рівня, взятого за базу порівняння (абсолютна величина, яка відповідає кожному відсотку приросту);

– коефіцієнт випередження $\text{Квип} = \frac{\text{Тз}_1}{\text{Тз}_2}$ – відношення темпів зміни двох динамічних рядів за однакові періоди.

Середні характеристики часового ряду відображають узагальнені, типові тенденції динаміки досліджуваного явища за певні періоди [72, с. 72].

Середні характеристики рівнів динаміки:

► середні рівні ряду (y_i – рівні рядів динаміки, n – кількість рівнів):

– для інтервального ряду з однаковими інтервалами (середня арифметична проста):

$$y_{сер} = \frac{y_1 + y_2 + \ldots + y_n}{n} = \frac{\sum_{i=1}^{n} y_i}{n};$$

– для моментного ряду з однаковими інтервалами (середня хронологічна):

$$y_{хр\,сер} = \frac{\frac{y_1 + y_2}{2} + \frac{y_2 + y_3}{2} + \ldots + \frac{y_{n-1} + y_n}{2}}{n-1} = \frac{\frac{y_1}{2} + y_2 + \ldots + y_{n-1} + \frac{y_n}{2}}{n-1};$$

– для ряду з неоднаковими інтервалами (середня арифметична зважена):

$$y_{сер} = \frac{y_1 f_1 + y_2 f_2 + \ldots + y_n f_n}{f_1 + f_2 + \ldots + f_n} = \frac{\sum_{i=1}^{n} y_i f_i}{\sum_{i=1}^{n} f_i};$$

► середній абсолютний приріст (середнє арифметичне з ланцюгових приростів):

$$\Delta y_{сер} = \frac{\sum_{i=1}^{n} \Delta y_i}{n-1};$$

► середній темп зростання (середня геометрична):

$$Тз_{сер} = \sqrt[n-1]{Тз_1 \cdot Тз_2 \cdots Тз_n} \cdot = \sqrt[n-1]{\frac{y_n}{y_1}} \cdot 100\%;$$

► середній темп приросту:

$$Тпр_{сер} = Тз_{сер} - 1 = Тз_{сер} - 100\% \cdot$$

Використання середніх величин є одним з основних інструментів оцінювання результатів міжнародних соціально-економічних та політичних процесів, пошуку резервів розвитку економік країн. Однак середні величини – це узагальнюючі показники, які нівелюють, не враховують індивідуальних відмінностей кількісних ознаках окремих одиниць сукупності, що можуть бути значущими для дослідження. Середня є достатньо надійною характеристикою ряду динаміки, якщо досліджуваний проміжок часу характеризують достатньо стабільні умови розвитку [4, с. 107].

Часовий ряд адекватно відображає об'єктивний закон динаміки економічного показника, якщо рівні цього ряду є порівнянними, однорідними, сталими та мають достатню сукупність спостережень.

Однорідність ряду динаміки означає відсутність нетипових, аномальних спостережень та викривлень тенденції.

Стійкість часового ряду відображає перевагу закономірності над випадковістю у зміні рівнів ряду. На графіках стійких часових рядів унаочнюється закономірність, а на графіках несталих рядів зміни послідовних рівнів є хаотичними, тому пошук закономірностей формування значень рівнів таких рядів недоцільний [59, с. 95].

Стаціонарний часовий ряд має постійну середню, а значення ряду коливаються навколо неї з деякою постійною дисперсією.

Аналіз рядів динаміки полягає у виявленні закономірностей варіації рівнів досліджуваного показника у часі, встановленні основної тенденції (тренда) розвитку явища – напряму його зміни (тенденції до зростання, стабільності чи зниження рівнів ряду). Рівні ряду динаміки формуються під сукупним впливом множини постійних (основних) факторів, які визначають конкретний вид тренда, і короткочасних факторів та різних випадкових обставин, що спричиняють відхилення фактичних значень рівнів ряду від тренда. Виявлення основної закономірності змін рівнів ряду передбачає її кількісне вираження без врахування випадкового впливу.

Встановлення основної тенденції у статистиці називають **вирівнюванням часового ряду**. Вирівнювання дає можливість представити вплив усіх основних факторів на варіацію рівнів ряду у вигляді функції часу:

$$y_t = f(t) + \varepsilon(t),$$

де $f(t)$ – тенденція, зумовлена дією постійних факторів, $\varepsilon(t)$ – величина, що визначає сукупний вплив випадкових факторів.

Найбільш поширені методи визначення основної тенденції динамічного ряду [72, с. 75]

Метод збільшення інтервалів (періодів) часу полягає в об'єднанні періодів та обчисленні агрегованих показників. За укрупнення інтервалів часу відхилення, спричинені дією випадкових факторів, нівелюються. Це дає змогу точніше встановити основну тенденцію розвитку явища.

Метод плинної (ковзної) середньої передбачає заміну початкового ряду динаміки рядом середніх: перший рівень нового ряду обчислюють як середню арифметичну просту (перша плинна середня) для перших за рахунком m рівнів початкового ряду; другий рівень нового ряду (другу плинну середню) розраховують як середню арифметичну просту для m рівнів, починаючи з другого рівня початкового ряду; потім з третього і т. д. Таким чином, при розрахунках середнього рівня ніби ковзають рядом динаміки від його початку до кінця, кожного разу відкидаючи один рівень на

початку і додаючи один наступний. На практиці використовують непарні інтервали. Наприклад, для $m = 3$:

$$\overline{y}_1 = \frac{y_1 + y_2 + y_3}{3};\ \overline{y}_2 = \frac{y_2 + y_3 + y_4}{3};\ \overline{y}_3 = \frac{y_3 + y_4 + y_5}{3}\ \text{і т. д.}$$

Аналітичне вирівнювання (згладжування) дає можливість не лише виявити тенденцію розвитку явища, а й кількісно оцінити її. Метод полягає у заміні фактичних рівнів динамічного ряду теоретичними, розрахованими на основі рівняння регресії (лінії тренда), за допомогою якого описують основну тенденцію. Обирають таку аналітичну функцію залежності рівнів динаміки від часу, яка найбільш точно описує коливний процес, та розраховують майбутні теоретичні значення рівнів ряду за відомими емпіричними даними.

Розрахунок параметрів математичних функцій здійснюють **методом найменших квадратів** (МНК), який полягає у відшуканні залежності, що найближче проходить до точок фактичних даних на графіку в осях координат «$t - y'$» (сума квадратів відхилень фактичних значень результатної ознаки у від вирівняних (теоретичних) значень y' є найменшою):

$$\sum (y - y')^2 \to \min .$$

За МНК визначають лінію, яка найбільше підходить для експериментальних даних та дає характеристику спрямованості досліджуваного явища. Здебільшого нею є парабола відповідного порядку.

Метод найменших квадратів

Приклад 4.1. Вирівнювання за прямою (парабола першого порядку).

Рівняння прямої має вигляд:

$$y' = a_0 + a_1 t,$$

де t – порядковий номер року чи іншого періоду;

y' – теоретичні рівні;

a_0 – початковий рівень;

a_1 – коефіцієнт регресії.

Розрахунок за прямою за методом найменших квадратів спрощується відповідним підбором способу розрахунку часу t таким чином, щоб $\sum t = 0$.

За таких умов розрахунок параметрів a_0 та a_1 здійснюють за формулами:

$$a_0 = \frac{\sum_{i=1}^{n} y_i}{n}, \quad a_1 = \frac{\sum_{i=1}^{n} yt}{\sum_{i=1}^{n} t^2},$$

де a_0 і a_1 – постійні параметри для підстановки їх у рівняння;
n – кількість членів ряду;
t – позначення одиниці часу.

Результати обчислень оформимо у вигляді табл. 4.2

Таблиця 4.2

Результати розрахунків для прикладу 4.1

Рівні ряду y	Умовний час t	$y \times t$	t^2	Вирівняні дані y'_t
14,0	–3	–42,0	9	14,77
14,9	–2	–29,8	4	14,57
14,5	–1	–14,5	1	14,37
14,7	0	0	0	14,17
14,3	1	14,3	1	13,97
14,0	2	28,0	4	13,77
12,8	3	38,4	9	13,57
$\sum y = 99{,}2$		$\sum xy = -5{,}6$	$\sum t^2 = 28$	

$$a_0 = \frac{\sum y}{n} = \frac{\sum 99{,}2}{7} = 14{,}17, \qquad a_1 = \frac{\sum yt}{\sum t^2} = \frac{-5{,}6}{28} = -0{,}2.$$

$y_1 = 14{,}17 + (-0{,}2) \times (-3) = 14{,}77;$
$y_2 = 14{,}17 + (-0{,}2) \times (-2) = 14{,}57;$
$y_3 = 14{,}17 + (-0{,}2) \times (-1) = 14{,}37;$
$y_4 = 14{,}17 + (-0{,}2) \times 0 = 14{,}17;$
$y_5 = 14{,}17 + (-0{,}2) \times 1 = 13{,}97;$
$y_6 = 14{,}17 + (-0{,}2) \times 2 = 13{,}77;$
$y_7 = 14{,}17 + (-0{,}2) \times 3 = 13{,}57.$

Вибір ліній для вирівнювання динамічних рядів:

► **пряма** – ланцюгові абсолютні прирости є умовно постійними (рівні ряду змінюються приблизно в арифметичній прогресії);

► **квадратична парабола** – зміна рівнів ряду відбувається з приблизно рівномірним прискоренням або уповільненням ланцюгових абсолютних приростів;

► **степенева функція** – рівні ряду динаміки виявляють тенденцію до сталості ланцюгових темпів зростання (зміна рівнів ряду динаміки відбувається у геометричній прогресії).

Реальні часові ряди в економіці є динамічно нестабільними (нестаціонарними) [1, с. 358]. Кожен рівень часового ряду формується під впливом значної кількості факторів, які відображають закономірність і випадковість його формування.

При дослідженні соціально-економічних процесів часові ряди представляють як суму систематичної складової (середньої) та випадкового відхилення від неї:

$$y_t = f(t) + \varepsilon(t),$$

де $f(t)$ – невипадкова функція часу (детермінована складова), $\varepsilon(t)$ – випадкова (недетермінована) складова.

Розкладання часового ряду полягає в аналізі факторів, що впливають на значення його рівнів, у виділенні серед них головних (еволюційних і періодичних) та другорядних (випадкових).

Фактори, які формують регулярні коливання динамічного ряду

Сезонні – коливання, що мають періодичний або близький до нього характер упродовж року. Кількісним вимірником сезонних коливань є **індекс сезонності** (сезонна хвиля):

$$I_t^c = \frac{y_{сер_t}}{y_{сер}} \cdot 100\%,$$

де $y_{сер_t}$ – середній рівень ряду динаміки за інтервал часу t, $y_{сер}$ – середній рівень ряду динаміки за рік.

Для аналізу сезонних коливань необхідна наявність значень досліджуваного показника за декілька років, оскільки окремі коливання можуть бути зумовлені випадковими факторами і не мати сезонного характеру. Здебільшого це 3-річні дані за місяцями чи кварталами.

Циклічні (кон'юнктурні) коливання подібні на сезонні, але проявляються впродовж більш тривалих інтервалів часу. Їх можна пояснити впливом довготермінових циклів економічного, демографічного або астрофізичного характеру [20, с. 117].

Автокореляція (автокореляційна функція) – кореляція функції з самою собою, зміщеною на певну величину незалежної змінної. Автокореляцію використовують для знаходження закономірностей у ряді даних, наприклад, періодичність. Автокореляція – взаємозв'язок послідовних елементів часового ряду даних.

Причини автокореляції залишків:

– наявність помилок вимірювання у значеннях результатної ознаки;

– модель може не містити фактор, що спричиняє суттєвий вплив на результат, цей вплив відображається у залишках (наприклад, фактор часу);

– модель не враховує декілька другорядних факторів, загальний вплив яких на результат значний через збіг тенденцій їх зміни або фаз циклічних коливань.

Білий шум – часові ряди, рівні яких мають середню, що дорівнює нулю, сталу дисперсію та нульову коваріацію послідовних спостережень (нульову автокореляцію).

Систематичними компонентами часового ряду є його невипадкові компоненти – тренд, сезонність та циклічність [36, с. 244].

Процедуру оцінювання детермінованої складової часового ряду разом з усіма невипадковими компонентами називають **згладжуванням часового ряду**.

Випадкові фактори не підлягають вимірюванню, але характерні для кожного процесу МВ і визначають стохастичний харак-

тер його елементів. До випадкових факторів відносять помилки вимірювання, випадкові збурення і т. ін. Результат впливу випадкових факторів $\varepsilon(t)$ обчислюють як залишок або похибку, отриману після вилучення з часового ряду систематичних компонент.

Прогнозування передбачає збереження основних закономірностей у майбутньому – базується на екстраполяції.

Види екстраполяції:

► перспективна екстраполяція (на майбутнє);

► ретроспективна екстраполяція (на минуле).

Теоретичною основою поширення тенденції на майбутнє є інерційність основних соціальних, економічних та політичних процесів. Чим коротшим є термін екстраполяції, тим більш надійний і точний прогноз.

У процесі аналізу динамічних рядів у деяких випадках потрібно визначити окремі невідомі рівні всередині заданого ряду (провести інтерполяцію).

Цей метод базується на принципах, аналогічних до екстраполяції, але ступінь точності прогнозування очікуваного результату є значно вищим.

Інтерполяцію проводять у разі, коли метою згладжування є досягнення найбільшої близькості лінії тренда до фактичних рівнів часового ряду.

4.2. Наближення функцій. Інтерполяція

Нехай величина y є функцією аргументу x, тобто будь-якому значенню x з області визначення відповідає значення y.

На практиці бувають випадки, коли неможливо записати зв'язок між x та y у вигляді деякої залежності $y = f(x)$.

Найбільш поширеним випадком, коли вид зв'язку між параметрами x та y невідомий, є задання цього зв'язку у вигляді таблиці $\{x_i, y_i\}$. Це означає, що дискретній множині значень аргумента $\{x_i\}$ відповідає множина значень функції $\{y_i\}$ $(i = \overline{0,n})$. Цими значеннями можуть бути, наприклад, експериментальні дані.

На практиці можуть бути потрібні значення величини y і в інших точках, відмінних від вузлів x_i. Однак одержати ці значення можна тільки експериментальним шляхом, що не завжди зручно і вигідно.

З точки зору економії часу та засобів доцільно використовувати наявні табличні дані для наближеного обчислення шуканого параметра y при будь-якому значенні (з деякої області) визначального параметра x, оскільки точний зв'язок $y = f(x)$ невідомий.

Для цього розв'язують задачу про наближення (апроксимацію) функції: задану функцію $f(x)$ потрібно наближено замінити (апроксимувати) деякою функцією $\varphi(x)$ так, щоб відхилення $\varphi(x)$ від $f(x)$ у заданій області було найменшим. При цьому функцію $\varphi(x)$ називають **апроксимуючою**.

У випадку, коли функція $f(x)$ задана у вигляді таблиці значень, **задача апроксимації** полягає в такому: за табличними даними підібрати таку аналітичну залежність простої структури $\varphi(x)$, яка згладжуватиме особливості заданої експериментальної таблиці і якнайкраще відображатиме загальну тенденцію зміни $f(x)$ у середньому.

Основна мета апроксимації – одержати швидкий (економний) алгоритм обчислення значень $f(x)$ для значень x, яких не містить таблиця даних. Основне питання апроксимації – як вибрати $\varphi(x)$ і оцінити відхилення $\varphi(x)$ від $f(x)$.

На практиці $\varphi(x)$ вибирають із класу алгебраїчних поліномів (многочленів):

$$\varphi(x) = a_0 + a_1x + a_2x^2 + \ldots + a_mx^m.$$

Якщо початкова функція задана таблично (на множині окремих точок), апроксимацію називають **точковою**.

Якщо початкова функція задана на неперервній множині точок (наприклад, на відрізку $[a; b]$), апроксимацію називають **інтегральною** (неперервною).

Рівномірне наближення

У багатьох випадках, особливо при обробці експериментальних даних, середньоквадратичне наближення зручне, оскільки воно згладжує деякі неточності функції $f(x)$ і дає достатньо правильне

уявлення про неї. Однак іноді ставлять більш жорстку умову – у всіх точках деякого відрізка [a, b] модуль відхилення многочлена $\varphi(x)$ від $f(x)$ має бути меншим від заданого ε:

$$|f(x) - \varphi(x)| < \varepsilon.$$

У цьому випадку отримаємо **рівномірну апроксимацію** (рис. 4.2).

Абсолютним відхиленням Δ многочлена $\varphi(x)$ від функції $f(x)$ на відрізку [a, b] називають максимальне значення абсолютної різниці між ними на заданому відрізку:

$$\Delta = max\ |f(x) - \varphi(x)|.$$

Середньоквадратичне відхилення:

$$\overline{\Delta} = \sqrt{\frac{S}{n}}.$$

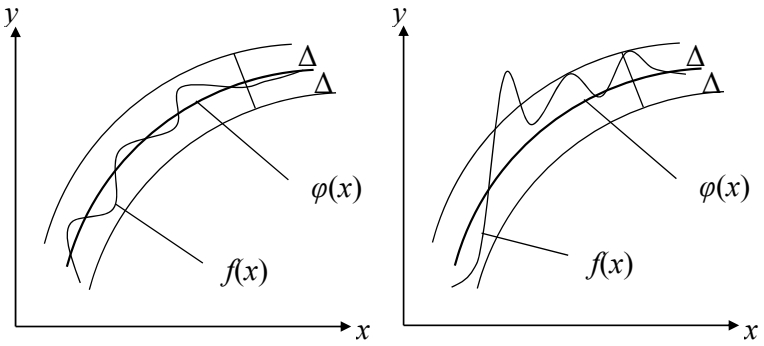


Рис. 4.2. Рівномірне наближення та середньоквадратичне відхилення

Існує також поняття **найкращого наближення функції** $f(x)$ многочленом $\varphi(x)$ фіксованого степеня m. У цьому разі коефіцієнти многочлена a_0, a_1, …, a_m вибирають таким чином, щоб на заданому відрізку [a, b] значення абсолютного відхилення Δ було мінімальним. Многочлен $\varphi(x)$ при цьому називають **многочленом найкращого рівномірного наближення**.

Одним з основних типів точкової апроксимації є **інтерполяція**: для заданої функції $y = f(x)$ будують функцію $\varphi(x)$, яка в заданих точках x_i $(i = \overline{0,n})$ набуває тих самих значень, що й функція $f(x)$:

$$\varphi(x_i) = f(x_i),$$

а в решті точок відрізка $[a; b]$ з області визначення $f(x)$ наближено представляє $f(x)$ з деякою похибкою. Точки x_i називають **вузлами інтерполяції**, а $\varphi(x)$ – **інтерполюючою функцією** (рис. 4.3). Найчастіше інтерполюючу функцію $\varphi(x)$ виражають через алгебраїчний многочлен степеня m.

Інтерполяцію в цьому разі називають **алгебраїчною**. Якщо використовують один многочлен $\varphi(x) = P_n(x)$ для інтерполяції функції $f(x)$ на всьому інтервалі зміни аргумента x ($m = n$, m – максимальний степінь інтерполяційного многочлена), інтерполяцію називають **глобальною**.

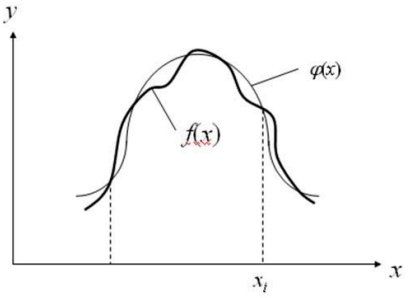


Рис. 4.3. Інтерполююча функція

Інтерполяційні многочлени можна будувати окремо для різних частин інтервалу зміни x. У цьому разі отримаємо **кускову** (**локальну**) **інтерполяцію**.

Здебільшого інтерполяційні многочлени використовують для апроксимації функцій у проміжних точках між крайніми вузлами інтерполяції ($x_0 < x < x_n$). Однак іноді їх використовують і для наближеного обчислення функції зовні інтервалу ($x < x_0, x > x_n$). Це наближення називають **екстраполяцією**.

При інтерполюванні основною умовою є проходження графіка інтерполяційного многочлена через задані значення функції у

вузлах інтерполяції. Виконання цієї умови в окремих випадках є недоцільним.

Наприклад, при великій кількості вузлів інтерполяції одержуємо високий степінь полінома у разі глобальної інтерполяції. Це пов'язано з певним недоліками – **осциляцією функції** (періодичними коливаннями).

Окрім того, табличні дані можуть містити похибки вимірювань, які будуть накопичені також в інтерполюючому многочлені. Отже, вибирають многочлен, графік якого близько проходить від заданих точок (рис. 4.4).

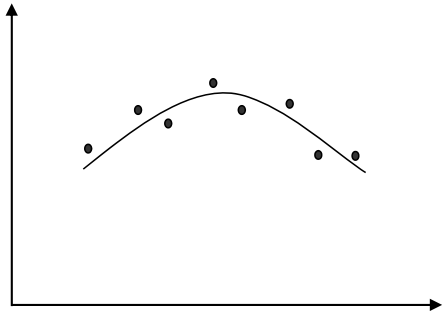


Рис. 4.4. Графік інтерполюючого многочлена

Поняття «близько» уточнюють при розгляді окремих видів наближення.

Середньоквадратичне наближення. Степінь полінома m значно менший від n. На практиці він становить не вище 5–6.

Мірою відхилення многочлена $\varphi(x)$ від заданої функції $f(x)$ на множині точок (x_i, y_i), $i = \overline{0,n}$ є величина S, яка дорівнює сумі квадратів різниць між значеннями многочлена та функції в заданих точках:

$$S = \sum_{i=0}^{n} \left[\varphi(x_i) - y_i\right]^2 .$$

При побудові апроксимуючого многочлена коефіцієнти a_0, a_1, …, a_m підбирають таким чином, щоб величина S була мінімальною. В цьому полягає ідея методу найменших квадратів.

Інтерполяція

Інтерполяційний поліном Лагранжа

Задача: знайти многочлен $P_n(x)$ степеня n, який у $n+1$ заданих точках $x_0, x_1, x_2, \ldots, x_n$ (вузлах інтерполяції) набуває значень $y_0, y_1, \ldots, y_n$ відповідно.

Для побудови $P_n(x)$ спочатку розглянемо допоміжні (фундаментальні) многочлени $Q_n^k(x)$ – многочлени n-го степеня відносно x, які задовольняють такі умови:

$$Q_n^k(x_i)=\begin{cases} 0, & \text{ïðè } i \neq k, \\ 1, & \text{ïðè } i = k, \end{cases} \quad \left(k=\overline{0,n}\right).$$

Ця властивість означає, що, наприклад, многочлен $Q_n^0(x)$ набуває в точці x_0 значення, що дорівнює одиниці, а в решті вузлів – нулю; многочлен $Q_n^1(x)$ у вузлі x_1 має значення 1, а в решті – нуль і т. д.

У загальному випадку многочлен $Q_n^i(x)$ у вузлі x_i набуває значення 1, а в решті вузлів – 0. Тоді шуканий многочлен:

$$P_n(x)=y_0Q_n^0(x)+y_1Q_n^1(x)+y_2Q_n^2(x)+\cdots+y_nQ_n^n(x).$$

Оскільки $x_0, x_1, x_2, \ldots, x_{k-1}, x_{k+1}, \ldots, x_n$ – нулі многочлена $Q_n^k(x)$, то:

$$Q_n^k(x)=c_k(x-x_0)(x-x_1)(x-x_2)\cdots(x-x_{k-1})(x-x_{k+1})\cdots(x-x_n).$$

Це інша форма запису полінома степеня n.

Визначаючи c_k з умови $Q_n^k(x)=1$, одержимо вираз для c_k (замість x підставляємо x_k):

$$c_k=\frac{1}{(x_k-x_0)(x_k-x_1)\cdots(x_k-x_{k-1})(x_k-x_{k+1})\cdots(x_k-x_n)}.$$

Запишемо явний вираз для допоміжних многочленів (коефіцієнтів Лагранжа):

$$Q_n^k(x)=\frac{(x-x_0)(x-x_1)(x-x_2)\cdots(x-x_{k-1})(x-x_{k+1})\cdots(x-x_n)}{(x_k-x_0)(x_k-x_1)\cdots(x_k-x_{k-1})(x_k-x_{k+1})\cdots(x_k-x_n)}=\prod_{i=0,i\neq k}^{n}\frac{x-x_i}{x_k-x_i}.$$

Запишемо інтерполяційний многочлен Лагранжа, враховуючи останню рівність:

$$P_n(x)=\sum_{k=0}^{n}y_k\left(\prod_{i=0,i\neq k}^{n}\frac{x-x_i}{x_k-x_i}\right)=\sum_{k=0}^{n}y_k\frac{(x-x_0)(x-x_1)(x-x_2)\cdots(x-x_{k-1})(x-x_{k+1})\cdots(x-x_n)}{(x_k-x_0)(x_k-x_1)\cdots(x_k-x_{k-1})(x_k-x_{k+1})\cdots(x_k-x_n)}.$$

Приклад 4.2. Записати інтерполяційний многочлен Лагранжа для функції, заданої табл. 4.3.

Таблиця 4.3

Дані для прикладу 4.2

x_i	–3	–1	1	2
y_i	8	6	4	18

Розв'язання

$$P_3(x) = y_0 \frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} + y_1 \frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)} +$$
$$+ y_2 \frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} + y_3 \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)}.$$

Підставимо значення $x_к$ та $y_к$ $(k = \overline{0,3})$.

$$P_3(x) = 8\frac{(x+1)(x-1)(x-2)}{(-3+1)(-3-1)(-3-2)} + 6\frac{(x+3)(x-1)(x-2)}{(-1+3)(-1-1)(-1-2)} +$$
$$+ 4\frac{(x+3)(x+1)(x-2)}{(1+3)(1+1)(1-2)} + 18\frac{(x+3)(x+1)(x-1)}{(2+3)(2+1)(2-1)}.$$

Шуканий поліном $P_3(x)=x^3 + 3x^2 - 2x + 2$.

Інтерполяційна формула Лагранжа має два суттєвих недоліки:

– формула громіздка, кожен доданок є многочленом n-го степеня;

– якщо з певної причини додаються вузли інтерполювання (наприклад, якщо отримана інтерполяційна формула неточна), то всі обчислення необхідно повторювати знову – жоден з доданків формули Лагранжа не зберігається.

Інтерполяційний поліном Ньютона

Розглянемо форму запису інтерполяційного полінома $P_n(x)$, яка допускає уточнення результатів інтерполяції послідовним додаванням нових вузлів. При цьому будемо використовувати поняття «розділені різниці функцій».

Нехай маємо функцію $f(x)$ і вузли інтерполяції x_i, $i = \overline{0,n}$ (не обов'язково рівновіддалені).

Розділеними різницями першого порядку називають величини, які мають зміст, наприклад, середніх швидкостей зміни функції:

$$f(x_i; x_j) = \frac{f(x_j) - f(x_i)}{x_j - x_i};$$

$$f(x_0; x_1) = \frac{f(x_1) - f(x_0)}{x_1 - x_0};$$

$$f(x_1; x_2) = \frac{f(x_2) - f(x_1)}{x_2 - x_1}.$$

Розділені різниці другого порядку визначаються співвідношеннями:

$$f(x_i; x_j; x_k) = \frac{f(x_j; x_k) - f(x_i; x_j)}{x_k - x_i};$$

$$f(x_0; x_1; x_2) = \frac{f(x_1; x_2) - f(x_0; x_1)}{x_2 - x_0};$$

$$f(x_1; x_2; x_3) = \frac{f(x_2; x_3) - f(x_1; x_2)}{x_3 - x_1}.$$

Розділена різниця *k*-го порядку визначається через розділені різниці (k–1)-го порядку за рекурентною формулою:

$$f(x_i; x_{i+1}; ...; x_{i+k}) = \frac{f(x_{i+1}; x_{i+2}; ...; x_{i+k}) - f(x_i; x_{i+1}; ...; x_{i+k-1})}{x_{i+k} - x_i}.$$

Розглянемо один вузол інтерполяції x_0.

З визначення розділеної різниці першого порядку $f(x; x_0)$ отримаємо:

$$f(x_0; x) = \frac{f(x) - f(x_0)}{x - x_0} = \frac{f(x_0) - f(x)}{x_0 - x};$$

$$f(x_0; x) = \frac{f(x) - f(x_0)}{x - x_0} = y_0 \Rightarrow f(x) = y_0 + (x - x_0) f(x; x_0)$$

Підставимо це значення у формулу для $f(x)$:

$$f(x) = y_0 + (x - x_0)[f(x_0; x_1) + (x - x_1) f(x; x_0; x_1)] =$$

$$y_0 + (x - x_0) f(x_0; x_1) + (x - x_0)(x - x_1) f(x; x_0; x_1).$$

Повторюючи цей процес для $n+1$ вузлів інтерполяції, отримаємо:

$$f(x) = y_0 + (x - x_0) f(x_0; x_1) + (x - x_0)(x - x_1) f(x_0; x_1; x_2) +$$
$$+ \ldots + (x - x_0)(x - x_1)\ldots(x - x_{n-1}) f(x_0; x_1; x_2; \ldots; x_{n-1}; x_n) +$$
$$+ (x - x_0)(x - x_1)\ldots(x - x_n) f(x; x_0; x_1; \ldots; x_n) =$$
$$= P_n(x) + (x - x_0)(x - x_1)\ldots(x - x_n) f(x; x_0; x_1; \ldots; x_n).$$

Оскільки $P_n(x)$ − інтерполяційний поліном для функції $f(x)$, то його значення у вузлах інтерполяції збігаються із значеннями функції $f(x)$ (збігаються також розділені різниці):

$$P_n(x_i) = f(x_i) = y_i, \ (i = \overline{0, n}),$$

оскільки залишковий член у цих вузлах становить:

$$R_n(x) = (x - x_0)(x - x_1)\ldots(x - x_n) f(x; x_0; x_1; \ldots; x_n) = 0$$

(x набуває значення x_0, x_1, …, x_n, тому один із співмножників завжди дорівнює 0, як і залишковий член у вузлах інтерполяції).

Отримаємо інтерполяційний поліном Ньютона з розділеними різницями:

$$P_n(x) = y_0 + (x - x_0) f(x_0; x_1) + \ldots + (x - x_0)(x - x_1)\ldots$$
$$\ldots(x - x_{n-1}) f(x_0; x_1; x_2; \ldots; x_{n-1}; x_n) =$$
$$= y_0 + \sum_{k=1}^{n} (x - x_0)(x - x_1)\ldots(x - x_{k-1}) f(x_0; x_1; \ldots; x_k) =$$
$$= y_0 + \sum_{k=1}^{n} \left(\prod_{i=0}^{k-1} (x - x_i) \right) f(x_0; x_1; \ldots; x_k).$$

Приклад 4.3. Знайти інтерполяційний поліном Ньютона для функції, заданої табл. 4.3.

Розв'язання

При $n = 3$ інтерполяційний поліном Ньютона буде мати вигляд:

$$P_n(x) = y_0 + (x - x_0) f(x_0; x_1) + (x - x_0)(x - x_1) f(x_0; x_1; x_2) +$$
$$+ (x - x_0)(x - x_1)(x - x_2) f(x_0; x_1; x_2; x_3).$$

Розділені різниці першого порядку:

$$f(x_0;x_1)=\frac{y_1-y_0}{x_1-x_0}=\frac{6-8}{-1+3}=-1;$$

$$f(x_1;x_2)=\frac{y_2-y_1}{x_2-x_1}=\frac{4-6}{1+1}=-1;$$

$$f(x_2;x_3)=\frac{y_3-y_2}{x_3-x_2}=\frac{18-4}{2-1}=14.$$

Розділені різниці другого порядку:

$$f(x_0;x_1;x_2)=\frac{f(x_1;x_2)-f(x_0;x_1)}{x_2-x_0}=\frac{-1+1}{1+3}=0;$$

$$f(x_1;x_2;x_3)=\frac{f(x_2;x_3)-f(x_1;x_2)}{x_3-x_1}=\frac{14+1}{2+1}=5.$$

Розділені різниці третього порядку:

$$f(x_0;x_1;x_2;x_3)=\frac{f(x_1;x_2;x_3)-f(x_0;x_1;x_2)}{x_3-x_0}=\frac{5-0}{2+3}=1.$$

j	x_j	y_j	*k* = 1	*k* = 2	*k* = 3
0	$x_0 = -3$	$y_0 = 8$	$\frac{6-8}{-1+3}=-1$		
1	$x_1 = -1$	$y_1 = 6$		0	
2	$x_2 = 1$	$y_2 = 4$	$\frac{4-6}{1+1}=-1$		
3	$x_3 = 2$	$y_3 = 18$	$\frac{18-4}{2-1}=14$	$\frac{14+1}{2+1}=5$	$\frac{5-0}{2+3}=1$

Шуканий поліном Ньютона:

$P_3(x) = 8 + (x + 3)(-1) + (x + 3)(x + 1)\cdot 0 + (x + 3)(x + 1)(x - 1)\cdot 1 =$

$= 8 - x - 3 + (x + 3)(x^2 - 1) = 5 - x + x^3 + 3x^2 - x - 3 = x^3 + 3x^2 - 2x + 2.$

4.3. Комп'ютерне моделювання динамічних процесів

Аналіз рядів динаміки у програмі «SPSS»

Метод підгонки кривих є одним з найбільш відомих методів прогнозування. Він полягає у визначенні кривої (групи кривих), яка з заданою точністю описує вихідний динамічний ряд [31, с. 8; 54, с. 42].

Основні види кривих підгонки:

– лінійна $\overline{y} = b_0 + b_1 x$;

– квадратична $\overline{y} = b_0 + b_1 x + b_2 x^2$;

– кубічна $\overline{y} = b_0 + b_1 x + b_2 x^2 + b_3 x^3$;

– логарифмічна $\overline{y} = b_0 \ln(x) + b_1$;

– експонентна $\overline{y} = b_0 \exp^{b_1 x}$;

– степенева $\overline{y} = b_0 x^{b_1}$;

– логістична $\overline{y} = \dfrac{1}{\dfrac{1}{b_0} + b_1 b_2^x}$;

– *S*-подібна $\overline{y} = \exp^{\left(b_0 + \frac{b_1}{x}\right)}$.

Приклад 4.4. За даними динамічного ряду (табл. 4.4) отримати прогнозні значення на наступних 3 періоди методом підгонки кривих.

Таблиця 4.4

Індекс людського розвитку Великобританії за 2008–2018 рр.

Рік	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Значення ІЛР	0,940	0,888	0,885	0,887	0,890	0,890	0,895	0,891	0,890	0,892	0,892

За зведенням моделі (рис. 4.5) можна зробити висновок, що оптимальною регресійною моделлю є кубічна (*max* R^2 = 0,73). Обчислимо за цією моделлю прогнозні значення ІЛР Великобританії на 2019–2021 рр.: точковий прогноз, верхня та нижня межі (рис. 4.6) [36, с. 184].

Сводка модели и оценки параметров

Зависимая переменная:ЛР

Уравнение	Сводка для модели					Оценки параметра			
	R-квадрат	F	ст.св.1	ст.св.2	Знч.	Константа	b1	b2	b3
Линейный	,149	1,575	1	9	,241	,905	-,002		
Логарифмическая	,370	5,282	1	9	,047	,914	-,012		
Обратная	,679	19,074	1	9	,002	,882	,047		
Квадратичный	,430	3,018	2	8	,106	,928	-,012	,001	
Кубический	,730	6,297	3	7	,021	,965	-,043	,007	,000
Составная	,146	1,543	1	9	,246	,905	,998		
Степенная	,366	5,198	1	9	,049	,914	-,014		
S	,676	18,757	1	9	,002	-,126	,051		
Роста	,146	1,543	1	9	,246	-,100	-,002		
Экспоненциальная	,146	1,543	1	9	,246	,905	-,002		
Логистическая	,146	1,543	1	9	,246	1,105	1,002		

Рис. 4.5. Параметри регресійних моделей

FIT_1	LCL_1	UCL_1
,90345	,86303	,94388
,90167	,86252	,94082
,89989	,86176	,93802

Рис. 4.6. Отримані прогнозні значення за оптимальною моделлю

Перевірка значущості рівняння регресії

Для оптимальної моделі побудуємо таблицю дисперсійного аналізу (рис. 4.7).

Дисперсионный анализ

	Сумма квадратов	ст.св.	Средний квадрат	F	Знч.
Регрессия	,002	3	,001	6,297	,021
Остаток	,001	7	,000		
Итого	,002	10			

Рис. 4.7. Таблиця дисперсійного аналізу

Результати регресійного аналізу є достовірними (**Знч** < 0,021).

Розраховане значення критерію Фішера $F_{емп} = 6{,}297$. На заданому рівні значущості 0,95 $F_{табл}(3; 7) = 4{,}35$. $F_{емп} > F_{табл}$, тому можна зробити висновок, що отримане рівняння регресії є значущим та адекватним, а модель – значущою.

Рівняння оптимальної регресійної моделі:

$$\overline{y} = 0{,}965 - 0{,}043x + 0{,}007x^2 + 0 \cdot x^3 .$$

З вірогідністю 95% Індекс людського розвитку Великобританії у 2019 р. становитиме від 0,86303 до 0,94388; у 2020 р. – від 0,86252 до 0,94082; у 2021 р. – від 0,86176 до 0,93802.

Дослідження часових рядів і прогнозування за моделлю ARIMA засобами «Statistica»

Для аналізу часових рядів у «Statistica» передбачено модуль **Time Series/Forecasting (Часові ряди/Прогнозування)**, призначений для побудуви простої моделі, що описує часовий ряд, згладжування, прогнозування майбутніх значень часового ряду на основі спостережуваних значень до визначеного моменту, проведення спектрального або Фур'є-аналізу ряду і т. ін. [35, с. 41; 52, с. 103].

Основні методи аналізу часових рядів у пакеті «Statistica»:

► **ARIMA** (autoregressive integrated moving average) – метод авторегресії та проінтегрованого ковзного середнього;

► **Interrupted time series analysis** – аналіз дискретного часового ряду (моделі інтервенції для ARIMA);

► **Exponential smoothing & forecasting** – експоненціальне згладжування та прогнозування;

► **Seasonal decomposition (1, 2)** – сезонна декомпозиція 1, 2 (квартальна, за місяцями);

► **Distributed lags analysis** – аналіз розподілених лагів (регресійна модель для двох часових рядів);

► **Spectral (Fourier) analysis** – спектральний аналіз (аналіз Фур'є).

► **ARIMA (авторегресія)** – важливий клас параметричних моделей, які широко застосовують у дослідженні часових рядів.

У «Statistica» є можливість здійснювати аналіз дискретної авторегресії, розглядати моделі ARIMA з інтервенцією – **аналіз дискретних часових рядів**. Необхідність у такому аналізі виникає, коли з деякого моменту різко змінюються умови зміни спостережуваного ряду.

Зовнішній вплив на ряд, спричинений різними факторами, може бути як короткоплинним (імпульсивним), так і тривалим

(стійким). У момент впливу часовий ряд різко змінюється, але далі знову описується авторегресійною моделлю.

У «Statistica» передбачені такі методи зовнішнього впливу:

– **експонеціальне згладжування і прогнозування** – згладжування спостережуваного ряду шляхом виділення з нього шуму і прогнозування майбутніх значеннь. У моделі можуть бути враховані різні види трендів і сезонність. Доступні моделі з адитивним і мультиплікативним шумами;

– **сезонна декомпозиція (1, 2)** – аналіз адитивних моделей часових рядів – часовий ряд x подають у вигляді:

$$X_t = f(t) + s(t) + u(t),\ t = 0, 1, 2,$$

де $f(t)$ – тренд (детермінована функція), $s(t)$ – сезонна складова, $u(t)$ – випадкова складова, а також мультиплікативні моделі часових рядів, у яких випадкова складова є множником:

$$X_t = f(t) \cdot s(t)\ u(t),\ t = 0, 1, 2;$$

– **аналіз розподілених лагів** – побудува регресії одного ряду на інший. Це важливо, наприклад, у разі, коли потрібно спрогнозувати значення залежного ряду на основі вимірювань зі зрушенням незалежного ряду (одні вимірювання випереджають інші). Такий аналіз називають **аналізом розподілених запізнювань** – один ряд запізнюється щодо іншого. Розподіляючи запізнювання незалежного ряду та надаючи їм різних ваг, можна якнайточніше наблизити значення залежного часового ряду.

Приклад 4.5. За даними про обсяг товарообігу (у млн. грн.) міжнародного холдингу за 24 місяці спрогнозувати обсяг товарообігу ТНК на наступних 3 місяці.

Підгонка авторегресійної моделі

Вибір вихідної моделі ARIMA. Перш, ніж підігнати до часового ряду авторегресійну модель, його необхідно зробити стаціонарним. Побудуємо графік за початковим рядом даних (рис. 4.8).

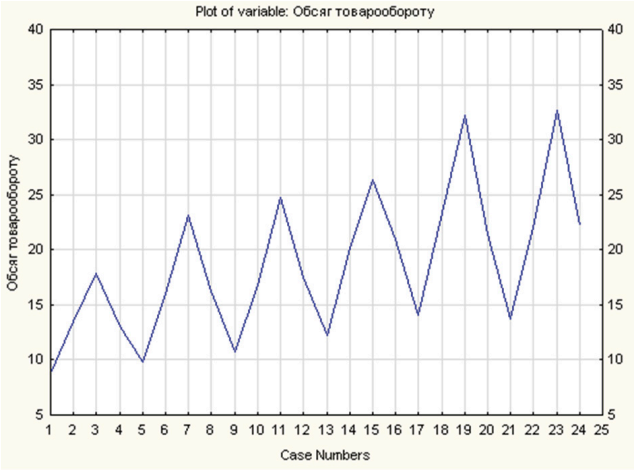


Рис. 4.8. Графічне представлення вихідних даних часового ряду

Спостережуваний ряд не є стаціонарним, не має різких стрибків, спостерігається тренд ряду, що виражається в плавному збільшенні його значень, і деяка сезонність, що виявляється в періодичності зростання обсягу виробництва. Динаміку збільшення обсягу виробництва холдингу можна розглядати як процес, що має регулярну складову.

Часовий ряд є **стаціонарним**, якщо значення процесу коливаються навколо окремого рівня, розмах коливань не збільшується і не зменшується з часом. Графік демонструє чергування зростаючих додатних і від'ємних послідовних значень.

Існують різні способи перетворення, що дають можливість зробити ряд більш стаціонарним. Модель ARIMA є найпростішою з можливих моделей. Вона представляє значення ряду, спостережувані в даний момент, у вигляді лінійної комбінації попередніх значень самого ряду (авторегресія) і лінійної комбінації значень часового ряду з незалежними значеннями (ковзне середнє).

Автокореляції та часткові автокореляції – це характеристики, які дають змогу визначити, на скільки близький ряд до ста-

ціонарного, а також зорієнтуватись у перетвореннях, які потрібно здійснити, щоб привести ряд до стаціонарного. Існує проста методика зробити висновок про близькість ряду до стаціонарного за автокореляціями та частковими автокореляціями – на кожному кроці можна виконати тільки одне перетворення значень виділеної змінної. Кінцева мета всіх перетворень – привести ряд до зручного для аналізу вигляду.

У «Statistica» є можливість виконати різні перетворення, які можна одержати застосуванням до ряду стандартних функцій [36, с. 262]. Новий часовий ряд утворюється перетворенням $f(x_{\text{новий}}) - f(x_{\text{старий}})$.

Можливі такі перетворення ряду:

– **Add constant** – додати константу до значень ряду;

– **Power** – піднести до степеня;

– **Inverse power** – знайти корінь степеня;

– **Natural log** – обчислити натуральний логарифм;

– **Exponent** – виконати експоненціальне перетворення;

– **Mean subtract (віднімання середнього)** – від значень ряду віднімається середнє значення, обчислене за всіма спостереженнями, або задане числове значення;

– **Standardize (стандартизувати)** – від значень ряду віднімається величина середнього значення ряду і результат ділиться на стандартне відхилення;

– **Trend subtract (віднімання тренда)** – від ряду віднімається лінійний тренд, параметри якого оцінюються або задаються;

– **Autocorr** (**автокореляції**) – лінійне перетворення, що дає можливість занулити автокореляції на вказаному лазі;

– ***N*-pls mov. aver**. – *N*-точкове ковзне середнє;

– ***N*-pls mov. median –** ковзна медіана;

– **Weighted –** усереднення нерівними вагами;

– **Prior –** обчислення за попереднім значенням ряду;

– **Simple exponential –** просте експоненціальне згладжування;

– **4253H Filter** – 4253H фільтр;

– **Difference** (**різниця**) – обчислення нового значення ряду x **за** формулою: $x = x - y(lag)$, де значення **lag – запізнювання**;

– **Residnalizing** – обчислення нового значення ряду за формулою: $x = x - (a + by(lag))$, де параметри a та b задають або оцінюють методом найменших квадратів;

– **Shift – зсув** ряду вперед **(forward)** або назад **(back)**.

– **Differencing, integrate (віднімання, додавання)** – обчислюються значення нового ряду x за формулою $x(t) = x(t) \pm x(t - lag)$.

Перетворення ряду до стаціонарного вигляду

Щоб зробити ряд даних прикладу 4.5 стаціонарним, після чого підібрати авторегресійну модель, спочатку потрібно послідовно застосувати кілька перетворень до часовою ряду обсягів товарів за місяцями. Для зменшення амплітуд коливань часових рядів використовують логарифмічне перетворення **Natural log (натуральний логарифм)**.

Побудуємо графік перетворених даних (графік прологарифмованого ряду). Його дисперсія значно менша, ніж дисперсія вихідного ряду (рис. 4.9), – після перетворення ряду його коливання істотно зменшилися.

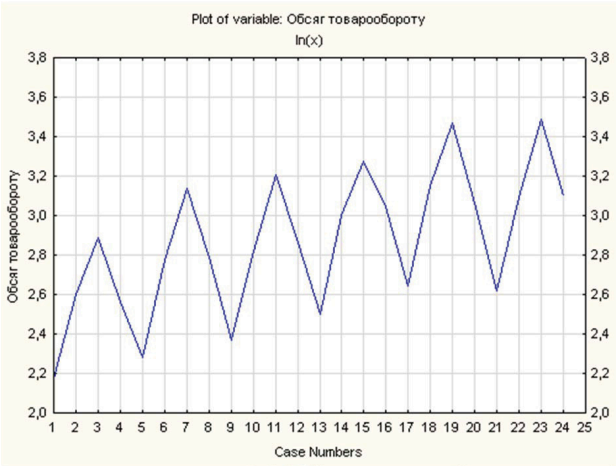


Рис. 4.9. Графічне представлення прологарифмованого часового ряду

Обчислимо основні описові статистики перетвореного ряду (рис. 4.10).

Descriptive Statistics (Ex5_data_new.sta)

Variable	Mean	Std.Dv.	Minimum	Maximum	First Case	Last Case	N
Обсяг товарообороту	18,72500	6,464098	8,800000	32,70000	1,000000	24,00000	24,00000
Обсяг товарообороту: ln(x)	2,87158	0,353315	2,174752	3,48738	1,000000	24,00000	24,00000

Рис. 4.10. Основні описові статистики початкового та перетвореного рядів

Після зменшення розкиду потрібно ідентифікувати параметри моделі для можливості подальшого оцінювання параметрів ARIMA. У моделі ARIMA використовуються такі типи параметрів: *p* – порядок авторегресії, *d* – порядок різниці, *q* – порядок ковзного середнього. Скорочено: модель ARIMA (*p, d, q*). Ідентифікувати модель ARIMA – це означає визначити ці параметри.

Нестаціонарність ряду є очевидною, наприклад, якщо в ряді яскраво виражений тренд. Особливо легко визначити візуально наявність монотонного тренда: логарифмічного, експонентного, лінійного, параболічного і т. ін. Із графіків на рис. 4.8 і 4.9 видно, що ряд має явну тенденцію до зростання значень при збільшенні номера спостереження, тобто спостерігається монотонний тренд.

Наявність тренда – перше свідчення про нестаціонарність аналізованого ряду.

Якщо тренд не виражений яскраво та немає інших особливостей ряду, що вказують на нестаціонарність, розглядають автокореляційну функцію, а точніше – вибіркову автокореляційну функцію.

Якщо автокореляційна функція не має тенденції до згасання, йдеться про нестаціонарність ряду.

Побудуємо автокореляційну функцію для 20 лагів. Лаг – це показник, що відображає відставання або випереджання в часі одного явища порівняно з іншим, пов'язаним з ним. З побудованого графіка (рис. 4.11) видно, що автокореляційна функція не має тенденції до згасання (такий висновок справедливий і для вихідного непрологарифмованого ряду).

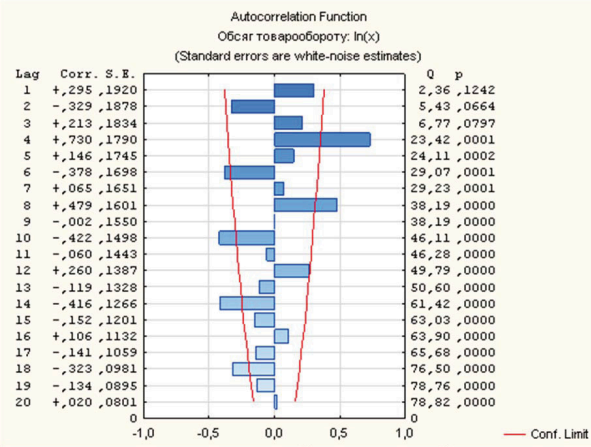


Рис. 4.11. Графік автокореляційної функції

На наступному етапі застосуємо до часового ряду різницеве перетворення $x = x - x(lag)$ для $lag = 1$.

Побудуємо графік для перетвореного ряду (рис. 4.12). Очевидно, що ряд став стаціонарним – немає тренда.

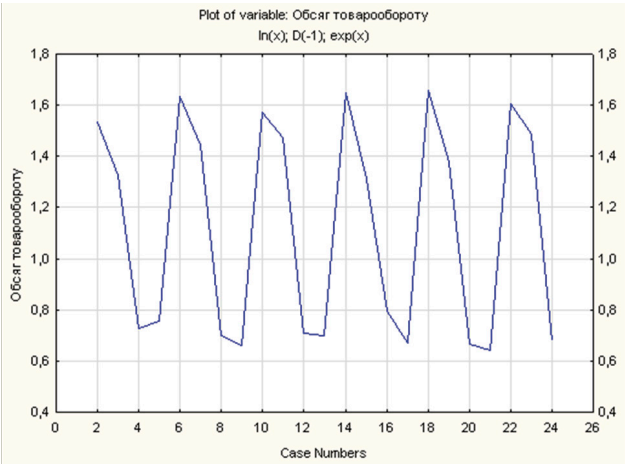


Рис. 4.12. Графік обсягу виробництва після застосування двох перетворень: *Ln*, *Differencing* з лагом 1

Експоненціальне згладжування та прогнозування

Експоненціальне згладжування є значно простішим методом, ніж ARIMA, і в деяких випадках дає можливість будувати прийнятні прогнози спостережуваних часових рядів.

Згладжування за методом ковзних середніх

Ця функція призначена для проведення згладжування за двома, трьома та більше точками.

Побудуємо графік початкових даних і результатів застосування процедури простого ковзного середнього (рис. 4.13).

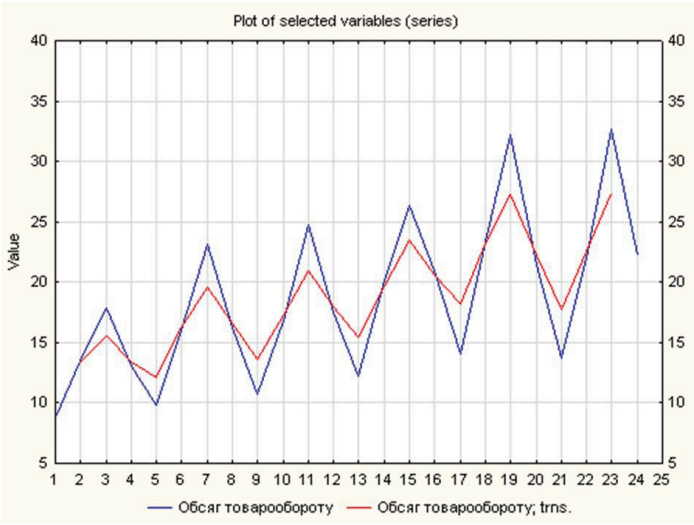


Рис. 4.13. Графік початкового ряду та згладженого за методом ковзних середніх

Перевірка ряду на наявність сезонної складової

Побудуємо **корелограму** – криву, що демонструє зміни взаємовпливу між спостереженнями залежно від часу. Отримані результати (рис. 4.14) свідчать, що початковий ряд даних має сезонну компоненту з періодом 4.

У «Statistica» є можливість виділити в ряді сезонну компоненту (**S**), задавши кількість сезонних індексів, тренд-циклічну компоненту (**TC**) і нерегулярну (випадкову) складову (**E**). Модель може бути мультиплікативною чи адитивною.

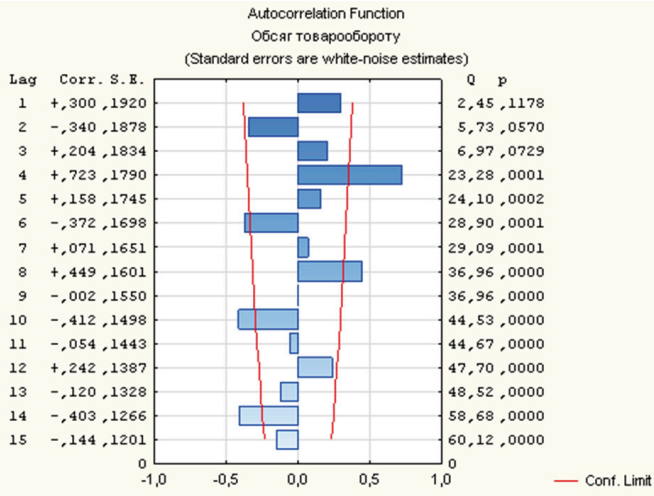


Рис. 4.14. Корелограма часового ряду з сезонною компонентою

Виконаємо процедуру сезонної декомпозиції. Результати розрахунків для мультиплікативної моделі виводяться у вигляді таблиці (рис. 4.15).

У другому стовпці (**Moving Averages**) наведені прості ковзні середні за чотирма точками часового ряду (без центрування), причому в таблиці значення ковзних середніх зміщені на 2 рядки: середнє арифметичне чотирьох перших точок $\frac{8,2+13,5+17,9+13}{4}=13,3$ визначається як значення ковзного середнього в третій точці і т. д.

У третьому і четвертому стовпцях (**Rations і Seasonal Factors**) розраховані відповідно відношення елементів вихідного ряду до ковзного середнього (у відсотках) і скориговані сезонні індекси.

У п'ятому стовпці (**Adjusted Series**) розраховано ряд, скоригований на сезонні індекси (ряд без сезонної складової). Він розраховується діленням вихідного ряду (X) на сезонні індекси та множенням результату на 100.

Seasonal Decomposition: Multipl. season (4) (Ex5_data_new.sta)
Обсяг товарообороту

Case	Обсяг товарообороту	Moving Averages	Ratios	Seasonal Factors	Adjusted Series	Smoothed Trend-c.	Irreg. Compon.
1	8,80000			64,6341	13,61510	13,24452	1,027980
2	13,50000			100,5211	13,43001	13,32584	1,007817
3	17,90000	13,30000	134,5865	138,4120	12,93241	13,48849	0,958773
4	13,00000	13,55000	95,9410	96,4328	13,48089	13,99768	0,963081
5	9,80000	14,17500	69,1358	64,6341	15,16227	14,87827	1,019088
6	16,00000	15,47500	103,3926	100,5211	15,91705	15,74827	1,010717
7	23,10000	16,27500	141,9355	138,4120	16,68931	16,35751	1,020284
8	16,20000	16,50000	98,1818	96,4328	16,79927	16,61287	1,011220
9	10,70000	16,70000	64,0719	64,6341	16,55473	16,80257	0,985250
10	16,80000	17,10000	98,2456	100,5211	16,71290	17,09837	0,977456
11	24,70000	17,42500	141,7504	138,4120	17,84527	17,63184	1,012105
12	17,50000	17,80000	98,3146	96,4328	18,14736	18,28803	0,992308
13	12,20000	18,62500	65,5034	64,6341	18,87548	18,87017	1,000282
14	20,10000	19,05000	105,5118	100,5211	19,99580	19,53439	1,023620
15	26,40000	19,92500	132,4969	138,4120	19,07349	20,16181	0,946021
16	21,00000	20,40000	102,9412	96,4328	21,77683	21,15357	1,029463
17	14,10000	21,22500	66,4311	64,6341	21,81511	21,98819	0,992128
18	23,40000	22,67500	103,1974	100,5211	23,27869	22,66250	1,027190
19	32,20000	22,77500	141,3831	138,4120	23,26388	22,63818	1,027639
20	21,40000	22,67500	94,3771	96,4328	22,19163	22,29553	0,995340
21	13,70000	22,32500	61,3662	64,6341	21,19624	22,07032	0,960396
22	22,00000	22,45000	97,9955	100,5211	21,88595	22,27927	0,982346
23	32,70000	22,65000	144,3709	138,4120	23,62512	22,84410	1,034189
24	22,20000			96,4328	23,02122	23,12651	0,995447

Рис. 4.15. Сезонна декомпозиція ряду для мультиплікативної моделі

У шостому стопці розрахована згладжена тренд-циклічна складова (**Smoothed Trend-c.**). Це результати згладжування ряду, скоригованого на сезонні індекси.

Відобразимо ці компоненти ряду на графіку (рис. 4.16).

Рисунок сезонної мінливості характерний для цього ряду. За допомогою сезонної складової можна скоригувати початковий ряд – вирахувати її з початкового ряду у випадку адитивної моделі та розділити значення початкового ряду на значення сезонної компоненти для мультиплікативної моделі.

Скоригований ряд більше не містить сезонних коливань (рис. 4.16). Після цього можна згладити отриманий ряд, щоб прибрати випадкову складову. Результатом згладжування ряду з сезонною поправкою є його тренд-циклічна компонента (**TC**). Вона показує загальний тренд і цикли, що присутні в ряді. Циклічна компонента відрізняється від сезонної компоненти тим, що цикли, як правило, мають тривалість більше одного сезону і не мають постійного періоду.

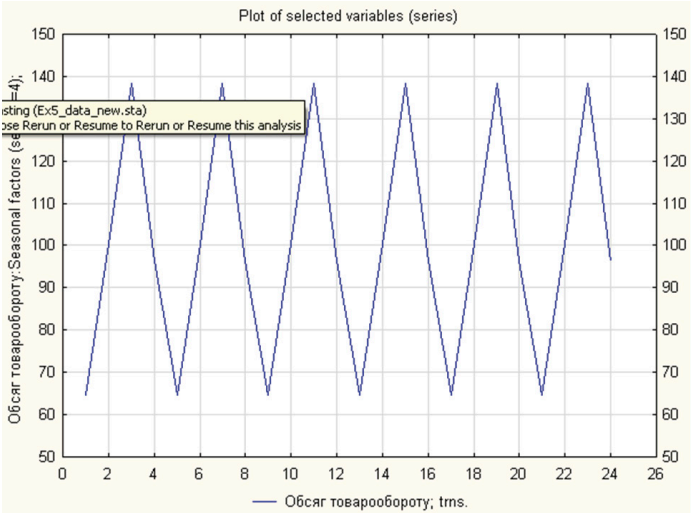


Рис. 4.16. Графік сезонної компоненти

Графік тренд-циклічної компоненти ряду та ряду з сезонною поправкою (рис. 4.17).

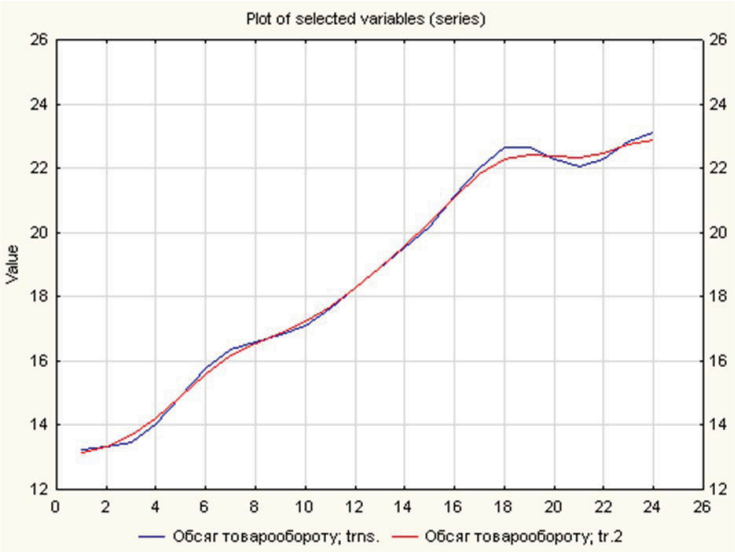


Рис. 4.17. Графік тренд-циклічної компоненти ряду та ряду з сезонною поправкою

Спрогнозуємо майбутні значення часового ряду на основі спостережуваних значень до поточного моменту. Для визначення моделі експоненціального згладжування задамо сезонний компонент, тренд і параметри згладжування [36, с. 276].

Оберемо лінійну модель для побудови тренда та розрахуємо оцінки параметрів рівняння регресії. В результаті отримаємо дві таблиці: з вихідними, прогнозованими значеннями ряду, залишками ряду та прогнозованими значеннями на 3 періоди спостереження вперед (рис. 4.18); з різними оцінками помилки згладжування, що може бути використана для найкращого підбору налаштувань згладжування (рис. 4.19).

Case	Обсяг товарообороту	Smoothed Series	Resids
16	21,00000	21,78943	-0,7894
17	14,10000	22,47723	-8,3772
18	23,40000	22,32248	1,0775
19	32,20000	23,12398	9,0760
20	21,40000	24,81609	-3,4161
21	13,70000	25,22483	-11,5248
22	22,00000	24,70744	-2,7074
23	32,70000	25,04472	7,6553
24	22,20000	26,49483	-4,2948
25		26,70697	
26		27,34860	
27		27,99023	

Рис. 4.18. Результати експоненціального згладжування

Model Number	Parameter grid search (Smallest abs. errors are highlighted) (Ex5_data_new.sta) Model: Linear trend, no season ; S0=-,204 T0=-,174 Обсяг товарообороту: Exp.smooth.resids.; **Alpha**	Gamma	Mean Error	Mean Abs Error	Sums of Squares	Mean Squares	Mean % Error	Mean Abs % Error
1	0,100000	0,100000	0,276427	4,611084	800,4229	33,35096	141,5028	-42,530
2	0,100000	0,200000	-0,048018	4,675648	816,8051	34,03354	238,8136	-158,406
3	0,100000	0,300000	-0,141048	4,716335	831,0322	34,62634	319,4597	-241,666
4	0,100000	0,400000	-0,108618	4,723907	845,2023	35,21676	385,3425	-306,125
5	0,100000	0,500000	-0,021240	4,753141	860,4602	35,85251	437,9903	-355,289
10	0,200000	0,100000	0,078632	4,913856	875,3846	36,47436	179,4691	-67,795
6	0,100000	0,600000	0,076717	4,784380	875,9408	36,49753	478,6617	-392,174
7	0,100000	0,700000	0,159726	4,826743	890,4613	37,10255	508,4239	-411,860
11	0,200000	0,200000	0,015162	5,012727	900,6662	37,52776	257,2949	-145,541
8	0,100000	0,800000	0,215431	4,888258	903,6022	37,65009	528,2114	-422,484

Рис. 4.19. Значення параметрів на сітці

Графіки вихідного ряду, ряду прогнозів і ряду залишків (рис. 4.20) демонструють, що ряд залишків є стаціонарним. Це свідчить про адекватність побудованої моделі лінійного згладжування.

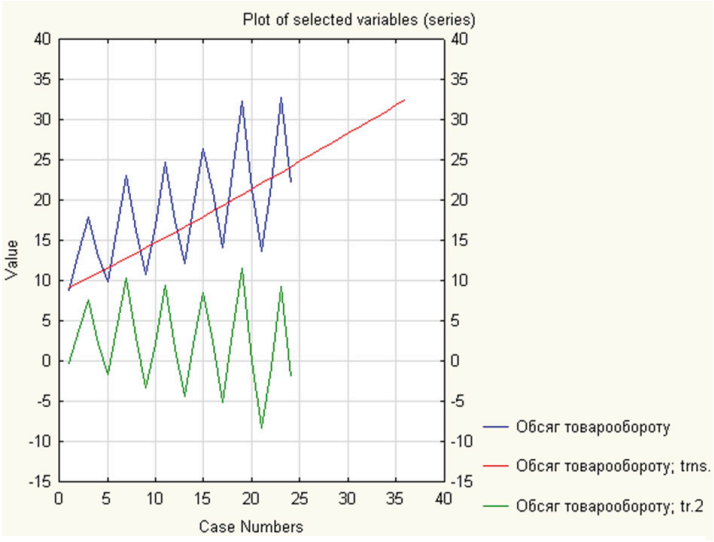


Рис. 4.20. Графіки вихідного ряду, ряду прогнозів і ряду залишків

ПИТАННЯ ТА ЗАВДАННЯ ДЛЯ САМОПІДГОТОВКИ І САМОКОНТРОЛЮ

Теоретичні запитання

1. Визначення поняття «часовий ряд».
2. Елементи динамічного ряду.
3. Класифікація динамічних рядів за ознакою часу.
4. Основні характеристики часових рядів.
5. Показники зміни рівнів часового ряду.
6. Середні характеристики рівнів динаміки.
7. Однорідність ряду динаміки.
8. Стійкість часового ряду.
9. Стаціонарний часовий ряд.
10. Визначення основної тенденції динамічного ряду: метод збільшення інтервалів, метод плинної (ковзної) середньої, аналітичне вирівнювання (згладжування).

11. Інтерполяція часових рядів.
12. Екстраполяція тренда.
13. Систематичні компоненти часового ряду.
14. Сезонні коливання динамічного ряду.
15. Циклічні (кон'юнктурні) коливання.
16. Автокореляція.
17. Поняття про білий шум.

Практичні завдання

Завдання 1. Побудувати графік вихідних даних часового ряду (за варіантом) та перевірити ряд на стаціонарність. За потреби виконати відповідні перетворення, щоб привести ряд до стаціонарного.

Завдання 2. Для зменшення амплітуд коливань часових рядів використати логарифмічне перетворення *Natural log*. Побудувати графік перетворених даних.

Завдання 3. Обчислити основні описові статистики початкового та перетвореного рядів.

Завдання 4. Застосувати до часового ряду різницеве перетворення та побудувати графік ряду після проведених двох перетворень: *Ln, Differencing* з лагом 1.

Завдання 5. Виконати згладжування за методом ковзних середніх. Побудувати графіки початкового ряду та згладженого за методом ковзних середніх.

Завдання 6. Перевірити ряд на наявність сезонної складової. Побудувати корелограму часового ряду з сезонною компонентою.

Завдання 7. Виділити в ряді сезонну компоненту, використовуючи адитивну та мультиплікативну моделі. Виконати сезонну декомпозицію ряду для мультиплікативної моделі. Побудувати графік сезонної компоненти.

При виявленні сезонної мінливості ряду скоригувати початковий ряд за допомогою сезонної складової. Побудувати графік початкового ряду та ряду з сезонною поправкою.

Завдання 8. Згладити отриманий ряд, щоб прибрати випадкову складову, – виділити тренд-циклічну компоненту. Побудувати графік тренд-циклічної компоненти ряду та ряду з сезонною поправкою.

Завдання 9. Виділити в початковому ряді нерегулярну (випадкову) складову. Побудувати графік.

Завдання 10. Спрогнозувати майбутні значення часового ряду на основі спостережуваних значень. Для визначення моделі задати сезонний компонент, тренд і параметри згладжування. Побудувати лінійну та експоненціальну моделі.

Завдання 11. Побудувати графіки вихідного ряду, ряду прогнозів і ряду залишків.

РОЗДІЛ 5

МЕТОДИ БАГАТОВИМІРНОГО МОДЕЛЮВАННЯ МІЖНАРОДНИХ ВІДНОСИН

Сучасні явища та процеси міжнародних відносин залежать від великої кількості факторів, які їх характеризують. Це зумовлює складнощі, пов'язані з дослідженням структури взаємозв'язків між факторами. У випадках, коли рішення приймають на підставі аналізу стохастичної, неповної інформації, використання методів багатовимірного статистичного аналізу є не лише виправданим, а й необхідним.

Багатовимірні статистичні методи дають можливість серед множини можливих математичних моделей обґрунтовано вибрати ту, яка якнайкраще відповідає початковим даним дослідження, що характеризують реальну поведінку досліджуваної сукупності об'єктів, оцінити надійність і точність висновків, зроблених на підставі обмеженого емпіричного матеріалу. Методи багатовимірного моделювання допомагають отримати всебічний опис даних, провести класифікацію, знайти закономірності та залежності між змінними.

5.1. Кластерний аналіз

При дослідженні МВ часто виникає необхідність багатовимірного групування об'єктів. Зокрема, такі завдання розглядають при формуванні рейтингу країн, вивченні суб'єктів ринку (споживачів та контрагентів), характеристик товарів і послуг, класифікації затрат і т. ін. Окрім того, міжнародні об'єднання на зразок СОТ, ЄС, НАТО також можна розглядати як угрупування країн за подібними ознаками.

Завдання розбиття множини елементів на кластери називають **кластерним аналізом**. Важливою перевагою кластерного аналізу поряд з іншими класичними методами дослідження МВ є можливість здійснювати розбиття об'єктів не за одним параметром, а за певним набором ознак. Такі дослідження можна проводити для множини початкових даних практично довільної природи. Це має велике значення, наприклад, при прогнозуванні кон'юнктури за наявності різнорідних показників, що ускладнюють застосування традиційних економетричних підходів. Кластерний аналіз дає можливість аналізувати достатньо великі обсяги даних і різко скорочувати, стискати великі масиви інформації, робити їх компактними та наочними [13, с. 159].

Кластерний аналіз – група методів, які використовують для класифікації об'єктів у відносно однорідні групи (**кластери**). Ці методи не є строгими із статистичної точки зору. Кластерний аналіз використовують на початковій стадії дослідження, коли поки не висуноto гіпотез відносно класів, в які об'єднують об'єкти.

Завдання кластеризації полягає в розподілі досліджуваної множини об'єктів на групи подібних об'єктів, які називають **кластерами** (англ. «claster» – «згусток, пучок, група, клас»). Розв'язком задачі класифікації є віднесення кожного з об'єктів даних до одного (чи декількох) із попередньо визначених класів і побудова у підсумку одним із методів класифікації моделі даних, що визначає розбиття множини об'єктів даних на класи.

Методика кластерного аналізу базується на понятті подібності об'єктів або їх ознак (залежно від мети групування). Розподіл

досліджуваної сукупності на кластери (групи) проводять шляхом вибору найбільш подібних одиниць спостережень.

Наприклад, для прогнозування попиту на продукцію ТНК часто проводять сегментацію цільових ринків (сегментацію споживачів). При цьому потенційних споживачів розглядають за сукупністю ознак: вік, стать, соціальний статус, освіта, інтереси, місце проживання, вподобання та ціннісні орієнтири і т. ін. Кластерний аналіз дає можливість сегментувати споживачів, об'єднуючи їх у групи з одночасним врахуванням вибраних для класифікації ознак (за принципом «знизу вгору»), або групувати ознаки (характеристики споживачів) в окремі кластери.

Виділяють агломеративні та ітеративні дивізійні методи кластерного аналізу.

Агломеративні методи кластеризації – ієрархічні методи, в яких на початковому етапі кожен об'єкт знаходиться в окремому кластері. На наступних етапах відбувається об'єднання об'єктів у більші кластери на підставі пониження деякого порогу, наприклад, збільшення відстані між об'єктами.

Ітеративні дивізійні методи кластеризації полягають у тому, що виконується розбиття об'єктів, об'єднаних в один або декілька великих кластерів, на фіксовану кількість кластерів, як правило, дрібніших.

Формальна постановка задачі кластеризації

Задано набір даних з такими властивостями:

► кожен примірник даних представлений чітким числовим значенням;

► клас для кожного конкретного примірника даних невідомий.

Знайти:

– спосіб порівняння даних між собою (міру подібності);

– спосіб кластеризації;

– розбиття даних за кластерами.

Формальне завдання кластеризації. Задано множину об'єктів даних I, кожен з яких представлений набором атрибутів (ознак). Потрібно побудувати відображення F множини елементів I

на множину кластерів C (кожному елементу з I поставити у відповідність елемент із C) (рис. 5.1).

Відображення F

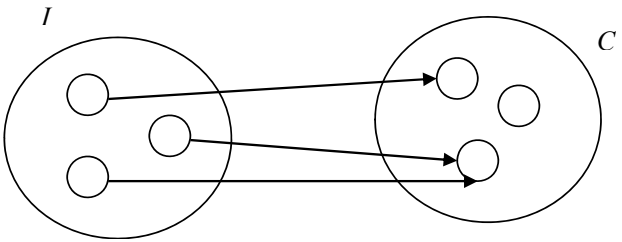


Рис. 5.1. Відображення множини *I* у множину *C*

При відображенні I в C кожен елемент з множини I має один і тільки один образ у множині C.

Відображення F задає модель даних, що є розв'язком задачі. Якість розв'язку визначає кількість правильно класифікованих об'єктів даних. Множину I запишемо у вигляді:

$$I = \{i_1, i_2, \ldots, i_j, \ldots, i_n\},$$

де i_j – досліджуваний об'єкт.

Кожен з об'єктів характеризує набір ознак:

$$i_j = \{x_1, x_2, .., x_h, \ldots, x_m\}.$$

Кожна змінна x_h може набувати значень із заданої множини:

$$x_h = \left\{v_h^1, v_h^2, \ldots\right\}.$$

Задача кластеризації полягає в побудові множини:

$$\mathrm{C} = \{c_1, c_2, \ldots, \mathrm{c}_k, \ldots, c_g\},$$

де c_k – кластер, що містить подібні між собою об'єкти з множини I:

$$c_h = \left\{ i_p, j_p \;\middle|\; i_j \in I,\ i_p \in I,\ d(i_j, i_p) < \sigma \right\},$$

де σ – величина, що визначає міру близькості для введення об'єктів в один кластер, $d(i_j, i_p)$ – міра близькості між об'єктами (відстань).

Якщо відстань $d(i_j, i_p)$ менша від заданого значення σ, елементи вважають близькими та поміщають в один кластер. В іншому разі елементи вважають такими, що відрізняються один від одного, і їх поміщають у різні кластери. Більшість традиційних алгоритмів, на основі яких розв'язують задачу кластеризації, представляє вихідні дані у вигляді матриці відмінності D, рядки та стовпці якої відповідають елементам множини I.

Елементами матриці є значення $d(i_j, i_p)$ в рядку j та стовпці p. На головній діагоналі значення дорівнюють нулю:

$$D = \begin{pmatrix} 0 & d(i_1, i_2) & d(i_1, i_n) \\ d(i_2, i_1) & 0 & d(i_2, i_n) \\ d(i_n, i_1) & d(i_n, i_2) & 0 \end{pmatrix}.$$

Відстані між об'єктами припускають їх представлення у вигляді точок m-вимірного простору R^m.

Найбільш відомі міри близькості:

– відстань Евкліда:

$$d_2(x_i, x_j) = \sqrt{\sum_{t=1}^{m} (x_t - x_j)^2};$$

– відстань за Хеммінгом (середня різниця за координатами):

$$d_H(x_i, x_j) = \sqrt{\sum_{t=1}^{m} |x_t - x_j|};$$

– відстань Чебишева (використовують у разі, коли потрібно визначити два об'єкти як «різні», якщо вони розрізняються за одним виміром):

$$d_\infty(x_i, x_j) = \max_{1 \le t \le m} |x_t - x_j|;$$

– пікова відстань (припускає незалежність між випадковими змінними):

$$d_L(x_i, x_j) = \frac{1}{m} \sum_{t=1}^{m} \frac{|x_t - x_j|}{x_t + x_j}.$$

Кожну з наведених мір відстані можна вибирати лише за наявної інформації про характер даних, до яких застосовують кластеризацію.

Результатом кластерного аналізу є набір кластерів, що містять елементи початкової множини об'єктів. Кластерна модель має описувати як кластери, так і належність кожного об'єкта до одного з них.

Для невеликої кількості об'єктів, що характеризуються двома змінними, результати кластерного аналізу зображають графічно [6, с. 165]. Елементи представляють точками, кластери розділяють прямими, які описують лінійними функціями (рис. 5.2). Якщо кластери не можна розділити прямими, зображають лінії, які описуються нелінійними функціями. Якщо елемент може належати до кількох кластерів, використовують Віденські діаграми (рис. 5.3).

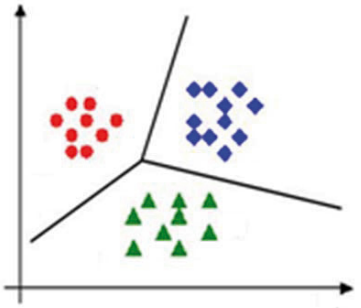


Рис. 5.2. Розподіл об'єктів на кластери прямими лініями

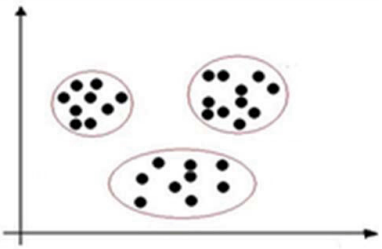


Рис. 5.3. Розподіл об'єктів на кластери з використанням Віденських діаграм

Деякі алгоритми кластерного аналізу не лише дають змогу зарахувати елемент до одного з кластерів, а визначають вірогідність його належності. В цьому разі зручніше представляти результат їх роботи у вигляді таблиці. У ній рядки відповідають елементам початкової множини, стовпці – кластерам, а в комірках вказують вірогідність належності елемента до кластера.

Ряд алгоритмів кластеризації будує ієрархічні структури кластерів. У таких структурах верхній рівень відповідає множині об'єктів (одному кластеру). На кожному наступному рівні він ділиться на декілька підкластерів доти, поки кластери не відповідатимуть окремим об'єктам (рис. 5.4). Такі діаграми називають **дендрограмами** (dendrograms).

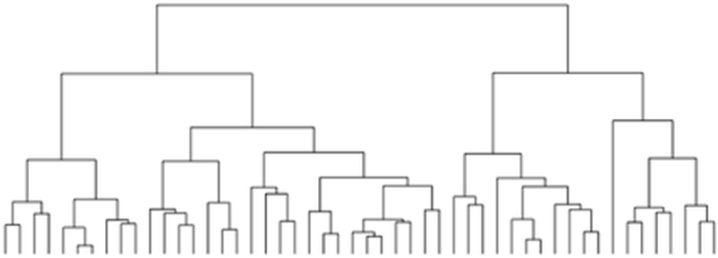


Рис. 5.4. Приклад дендрограми

Дендрограма – відображення послідовного процесу кластеризації. Це найбільш популярний спосіб представлення результатів кластеризації.

Базові методи кластеризації поділяють на ієрархічні та неієрархічні. Перші будують дендрограми «знизу вверх» (агломеративні) або «зверху вниз» (дивізимні) [13, с. 167].

Агломеративні алгоритми

На першому кроці множину I представляють як множину кластерів:

$$c_1 = \{i_1\},\ c_2 = \{i_2\},\ \ldots,\ c_p = \{i_p\},\ \ldots,\ c_m = \{i_m\}.$$

На наступному кроці вибирають два найбільш близьких один до одного об'єкти (наприклад, c_j, c_p) та об'єднують в один загальний кластер. Нову множину утворюють $m - 1$ кластерів:

$$c_1 = \{i_1\}, c_2 = \{i_2\}, \ldots, c_p = \{i_p, i_q\}, \ldots, c_m = \{i_m\}.$$

Повторюючи процес, отримують послідовну множину кластерів, що складається з $m - 2$, $m - 3$, $m - 4$ і т. д. кластерів. У кінці процедури буде сформовано кластер, що складається з m об'єктів і збігається з початковою множиною I.

Для визначення відстані між кластерами можна обирати різні способи. Залежно від цього отримують алгоритми з різними властивостями.

Найбільш популярні методи визначення відстані між кластерами:

► відстань між найближчими сусідами – найближчими об'єктами кластерів;

► відстань між найбільш віддаленими сусідами;

► відстань між центрами кластерів, або центроїдний метод (центр об'єднаного кластера обчислюють як середнє центрів об'єднуваних кластерів без урахування їх обсягів);

► метод медіан – центроїдний метод, на основі якого центр об'єднаного кластера обчислюють як середнє всіх об'єктів;

► середня відстань між кластерами;

► середня відстань між усіма об'єктами пари кластерів з урахуванням відстаней усередині кластерів;

► за методом Уорда як відстань між кластерами використовують приріст суми квадратів відстаней об'єктів до центрів кластерів, які отримують у результаті їх об'єднання.

Найбільш популярним з неієрархічних алгоритмів є алгоритм k-середніх і його різновиди.

Метод ***k*-середніх** – метод кластерного аналізу, метою якого є розподіл m спостережень (з простору R^n) на k кластерів. При цьому кожне спостереження відноситься до кластера, центр (центроїд) якого є якнайближчим до нього.

Як міру близькості використовують відстань Евкліда:

$$\rho(x, y) = \|x - y\| = \sqrt{\sum_{p=1}^{n} \left(x_p - y_p\right)^2} ,$$

де $x, y \in R^n$.

Розглянемо ряд спостережень $(x^{(1)}, x^{(2)}, \ldots, x^{(m)})$, $x^{(m)} \in R^n$.

Метод k-середніх розділяє m спостережень на k ($k \le m$) груп $S = \{S_1, S_1, \ldots, S_k,\}$ таким чином, щоб мінімізувати сумарне квадратичне відхилення точок кластерів від центроїдів цих кластерів:

$$\min\left[\sum_{i=1}^{k} \sum_{x^{(j)} \in S_i} \left\|x^{(j)} - \mu_i\right\|^2\right],$$

де $x^{(i)} \in R^n$, $\mu_i \in R^n$, m – центроїд для кластера S_i.

Якщо міра близькості до центроїда визначена, розбиття об'єктів на кластери зводиться до визначення центроїдів цих кластерів. Кількість кластерів k задає дослідник.

Розглядають початковий набір k середніх (центроїдів) $\mu_1, \ldots, \mu_k$ кластерів $S_1, \ldots, S_k$ відповідно. На першому етапі центроїди кластерів вибирають випадково або за визначеним правилом (наприклад, обирають центроїди, які б максимізували початкові відстані між кластерами).

Описова точка μ_i для кластера i є центром тяжіння кластера. У разі одновимірних даних (наприклад, результати голосування) центр тяжіння є середнім арифметичним значень точок у кластері.

Для багатовимірних даних, де кожна точка має декілька вимірів, центр тяжіння матиме однакову кількість компонент, кожна з яких буде середнім арифметичним відповідних складових точок кластера.

Парна агломеративна кластеризація

Спостереження відносять до тих кластерів, чиє середнє (центроїд) до них якнайближче. Кожне спостереження належить тільки до одного кластера, навіть якщо його можна віднести до двох і більше кластерів.

Після цього центроїд кожного *i*-го кластера переобчислюють за правилом:

$$\mu_i = \frac{1}{s_i} \sum_{x^{(j)} \in S_i} x^{(j)}.$$

Існує близько 100 різних алгоритмів кластеризації, проте найчастіше використовують ієрархічний кластерний аналіз та кластеризацію методом *k*-середніх. Алгоритм *k*-середніх полягає в переобчисленні на кожному кроці центроїда для кожного кластера, отриманого на попередньому кроці.

Алгоритм зупиняється, коли значення μ_i не змінюються: $\mu_i^{\text{крок } t} = \mu_i^{\text{крок } t+1}$.

Неправильний вибір кількості кластерів *k* може привести до некоректних результатів. При використанні методу *k*-середніх важливо спочатку провести перевірку відповідної кількості кластерів для конкретного набору даних.

Приклад 5.1. Методом *k*-середніх провести кластерний аналіз країн за такими вихідними даними:

– множина об'єктів $I = \{i_1, i_2, \ldots, i_j, \ldots, i_n\}$ = {Фінляндія, Португалія, Греція, Словенія, Ліхтенштейн, Болгарія,…};

– набір ознак $i_j = \{x_1, x_2, .., x_h, \ldots, x_m\}$ = {ІРЛП, Індекс задоволеності життям, Індекс економічної свободи, Індекс конкурентоспроможності};

– кожна змінна x_h може набувати значень із заданої множини:

► x_1 (ІРЛП) = {0, 0,01, …, 1};

► x_2 (Індекс задоволеності життям) = {–100, –99, …, 100};

► x_3 (Індекс економічної свободи) = {0, 1, …, 100};

► x_4 (Індекс конкурентоспроможності) = {0, 1, …, 10].

У результаті проведення кластерного аналізу у середовищі статистичного пакета «SPSS» отримано модель даних, що визначає розбиття множини об'єктів даних на класи (рис. 5.5–5.7) [32, с. 100; 54, с. 112].

Initial Cluster Centers

	Cluster	
	1	2
ІЛР	,778	,895
Індекс_задоволеності_життям	143,3	260,0
Індекс_економічної_свободи	51,90	90,10
Індекс_конкурентноспроможності	58,510	74,118

Рис. 5.5. Обчислені центри кластерів методом *k*-середніх

Cluster Membership

Case Number	Cluster	Distance
1	1	13,551
2	.	.
3	.	.
4	.	.
5	.	.
6	3	28,297
7	.	.
8	.	.
9	.	.
10	3	19,773
11	.	.
12	2	18,326

Рис. 5.6. Таблиця розподілу об'єктів за кластерами (фрагмент результатів)

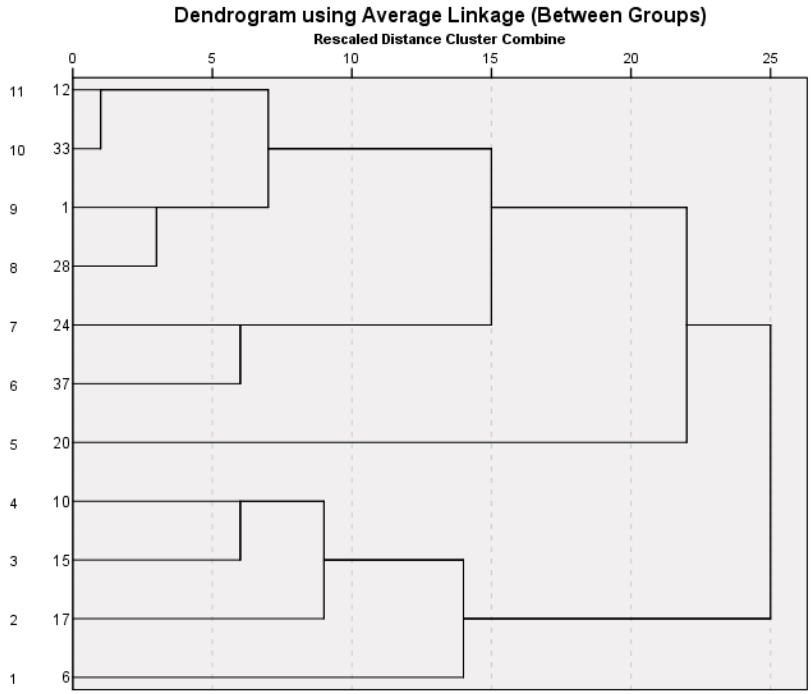


Рис. 5.7. Дендрограма з використанням методу міжгрупових зв'язків

Кластерний аналіз засобами «Statistica»

Приклад 5.2. За даними результатів офіційних статистичних досліджень для країн світу за 2017 р. провести кластерний аналіз об'єктів (країн) за змінними: *кількість мігрантів (International migrant stock), кількість біженців (Estimated refugee stock), рівень доходів (Income Rate), ВВП (GDP), рівень інфляції (Inflation Rate), рівень безробіття (Unemployment rate), Індекс Джині (Gini Index), прямі іноземні інвестиції (Foreign direct investment), Індекс глобального тероризму (Global terrorism Index), кількість поточних криз (Numbers of ongoing crisis and disasters)* [35, с. 77; 52, с. 136].

Результати кластеризації методом *k*-середніх

У результаті емпіричних досліджень виділено 3 кластери. Найбільший вплив на розподіл держав на групи має змінна прямі іноземні інвестиції (рис. 5.8). Перший кластер сформували дві країни – США та Нідерланди (рис. 5.9). Їх характеризують у кілька разів більші, ніж в інших групах, середні значення змінних *кількість мігрантів* та *прямі іноземні інвестиції,* високий рівень *ВВП* та *біженців*. У цих країнах також низькими є *рівень інфляції, рівень безробіття* та *Індекс Джині*. Ці дві високорозвинені країни найбільш привабливі для біженців та мають найвищий рівень мігрантів, які переміщуються в пошуках кращого життя (рис. 5.9). Хоча ці країни є одними з найдорожчих місць для проживання на планеті, емігранти відчувають себе фінансово безпечно і залишаються в них на довготривалий період.

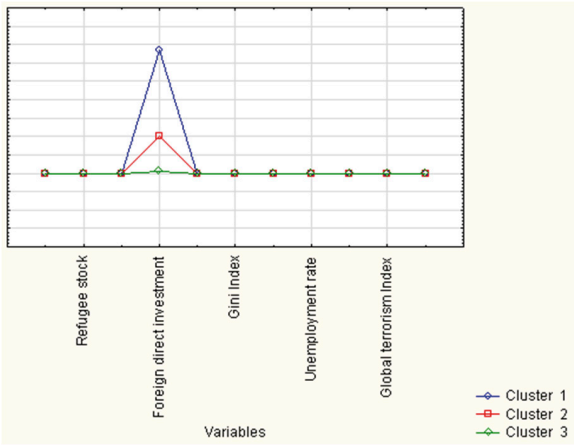


Рис. 5.8. Графік середніх для кожного кластера

	Distance
NLD	5,771922E+09
USA	5,771922E+09

	Distance
BRA	9,528041E+09
CHN	1,936458E+10
DEU	5,560975E+09
CHE	1,012187E+10
GBR	5,846299E+09

Рис. 5.9. Члени кластерів 1 та 2

До другого кластера ввійшли Бразилія, Китай, Німеччина, Швейцарія та Великобританія (рис. 5.9). Для країн цього кластера середні більшості аналізованих змінних мають середні серед трьох виділених груп значення. Країни другого кластера є найбільш безпечними для проживання – середні значення *Індексу глобального миру* та *кількості поточних криз* є найменшими серед усіх кластерів. Однак і рівень *прямих іноземних інвестицій* є найменшим у цій групі (рис. 5.10).

Усі інші країни світу ввійшли до третього кластера. Їх характеризує максимальне середнє значення *Індексу глобального миру*, що свідчить про низький рівень соціального захисту та безпеки, мілітаризацію та конфлікти в цих країнах. Індекс глобального миру розглядає «гармонію, досягнуту відсутністю насильства або страху перед насильством». Він оцінює так званий негативний мир. Держави, що входять до третього кластера, мають найбільші *рівень інфляції, рівень безробіття* та *кількість поточних криз* і найменші серед всіх аналізованих країн *ВВП* та *Індекс глобального тероризму*. Ця найбільша група країн найменш приваблива для біженців. Проте і середня кількість мігрантів, як і *прямі іноземні інвестиції*, для країн третього кластера є найменшими з-поміж інших груп (рис. 5.10).

Variable	**Cluster No. 1**	Cluster No. 2	Cluster No. 3
International migrant stock	2,591674E+07	5,049656E+06	1,268342E+06
Refugee stock	4,649555E+05	3,722462E+05	1,523679E+05
Global Peace Index	1,937000E+00	1,843400E+00	1,957474E+00
Foreign direct investment	3,356847E+11	1,018587E+11	6,008534E+09
Income Rate	1,010000E+02	1,020000E+02	1,028462E+02
Gini Index	3,485000E+01	3,782000E+01	3,582051E+01
Inflation Rate	1,900000E+00	2,260000E+00	5,083333E+00
Unemployment rate	4,650000E+00	5,620000E+00	8,510257E+00
Ongoing crisis and disasters	4,306000E+03	3,063000E+03	4,971167E+03
Global terrorism Index	3,920500E+00	3,480600E+00	2,527974E+00
GDP	5,717497E+04	3,924292E+04	2,388981E+04

Рис. 5.10. Середні кластерів

5.2. Факторний аналіз

При дослідженні складних економічних та політичних систем часто немає можливості безпосередньо виміряти величини, які визначають їх властивості (фактори). Окрім того, є невідомими кількість та зміст цих факторів. Але можуть бути виміряними інші величини, що залежать від них. Якщо невідомий фактор впливає на декілька вимірюваних ознак, вони виявляють певний зв'язок (наприклад, кореляція) між собою. З огляду на це загальна кількість факторів може бути значно меншою, ніж кількість вимірюваних ознак. Для виявлення таких факторів використовують факторний аналіз. Зменшення кількості факторів може бути необхідним також для забезпечення точності подальшого аналізу даних, скорочення ресурсів пам'яті ЕОМ та часу їх обробки, візуалізації отриманих результатів і т. ін.

Формальний опис задачі факторного аналізу

Задано масив p-вимірних спостережень [8, с. 137]:

$$X_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \dots \\ x_{ip} \end{pmatrix}, \; i = 1, 2, \dots 1 \; n,$$

де n – кількість спостережень (об'єктів, станів);

p – кількість параметрів, які характеризують кожне спостереження.

Необхідно представити результати у вигляді нового масиву:

$$Z_i = \begin{pmatrix} z_{i1} \\ z_{i2} \\ \dots \\ z_{ip'} \end{pmatrix}, \; i = 1, 2, \dots 1 \; n,$$

де p' є значно меншим, ніж p. Компоненти вектора Z_i називають **факторами**. На практиці прагнуть забезпечити виконання однієї з умов $p' = 0{,}1 \dots 0{,}25$ або $p' = 1 \dots 3$.

Традиційно перший етап факторного аналізу – це вибір нових ознак, які є лінійними комбінаціями вихідних факторів і відображають більшу частку загальної мінливості вихідних даних. Вони зберігають основну частину інформації, яку містили початкові дані. Другий етап – обертання факторів з метою спрощення їх інтерпретації. Вихідною інформацією для дослідження методами факторного аналізу є кореляційна матриця, побудована із застосуванням коефіцієнта кореляції Пірсона для кількісних ознак. Основною вимогою до цієї матриці є її додатна напіввизначеність (усі головні мінори невід'ємні), з чого випливає невід'ємність усіх її власних значень.

Методами факторного аналізу вирішують три основні групи проблем:

► пошук передбачуваних неявних закономірностей, зумовлених впливом зовнішніх або внутрішніх факторів на досліджуваний процес;

► виявлення та вивчення статистичного зв'язку ознак з факторами або головними компонентами;

► стискування інформації шляхом представлення процесу за допомогою узагальнених факторів або головних компонент, кількість яких є меншою за кількість початкових ознак (параметрів), але достатньою для забезпечення відтворення кореляційної матриці з необхідною точністю.

Розрізняють два основних типи факторного аналізу: *R*- та *Q*-техніку. *R*-техніка розроблена британським психологом Р.-Б. Кеттеллом і передбачає розрахунок коефіцієнтів кореляції між параметрами (ознаками), які утворюють матрицю вихідних даних. Її використовують для зменшення кількості параметрів. *Q*-техніку запропонував британський психолог В. Стефенсон у 1935–1936 рр. За її допомогою вивчають кореляцію між об'єктами або їх станами; застосовують для зменшення кількості об'єктів. З формального погляду при застосуванні *R*-техніки шукають кореляцію між стовпцями таблиці спостережень, а при використанні *Q*-техніки – між її рядками (табл. 5.1).

Таблиця 5.1

Загальний вигляд таблиці спостережень для факторного аналізу

Номери об'єктів (станів)	Параметри об'єктів (станів)			
	1	**2**	**…**	***p***
1	x_{11}	x_{12}	…	x_{1p}
2	x_{21}	x_{22}	…	x_{2p}
…	…	…	…	…
n	x_{n1}	x_{n2}	…	x_{np}

При дослідженні одного об'єкта, якщо значення його ознак вимірюють у різні моменти часу, використовують *P*-техніку. Понад 50% усіх завдань, що передбачають застосування факторного аналізу, вирішують за допомогою *R*-техніки.

Основними методами факторного аналізу є методи головних компонент, головних факторів, максимальної правдоподібності та центроїдний. Усі вони ґрунтуються на припущенні, що досліджувана залежність є лінійною. Вихідні дані мають підпорядковуватися багатовимірному нормальному розподілу, але центроїдний метод є достатньо стійким до відхилень від такої закономірності.

Мета факторного аналізу – зменшення кількості змінних та визначення структури взаємозв'язків між ними (класифікація даних). З формального погляду його мета – одержати матрицю факторного відображення. Її рядки – це координати векторів, що відповідають *n* змінним у p'-вимірному факторному просторі. Близькість цих векторів між собою свідчить про взаємну залежність змінних. Якщо кількість факторів перевищує одиницю, здійснюють обертання матриці факторного відображення для одержання більш простої її структури.

Однією з проблем, які виникають при застосуванні факторного аналізу, є необхідність знаходження власних значень кореляційної матриці. Якщо вона вироджена (квадратна матриця, визначник якої дорівнює нулю), ця задача може бути нерозв'язною. Для матриць високого порядку може відбуватися втрата значущості у процесі обчислень. В окремих випадках проблему виродженості можна зняти виключенням лінійно залежних параметрів.

Метод Якобі дає змогу визначити власні значення для вироджених кореляційних матриць. Однак при цьому частина їх, яка дорівнює різниці між порядком та рангом матриці, матиме значення, що не перевищують обчислювальної похибки. Завдяки цьому метод головних компонент є стійкішим до аналізу відповідних даних, ніж метод максимуму правдоподібності. Водночас він є гіршим за останній з погляду можливості отримання точної оцінки загальності й досягнення повного відтворення кореляційної матриці.

Обов'язкові умови факторного аналізу:

– всі досліджувані ознаки є кількісними;

– кількість ознак (факторів) принаймні вдвічі більша, ніж кількість змінних;

– однорідність вибірки;

– симетричний розподіл вихідних змінних.

Реалізація факторної моделі у «Statistica»

Приклад *5.3*. Визначити основні фактори сталого розвитку країн світу за офіційними статистичними даними країн світу за 2017 р.: *Індекс людського розвитку* (*Human Development Index, HDI*); *очікувана тривалість навчання дітей шкільного віку* (*Expected years of schooling*); *середня тривалість навчання дорослого населення* (*Mean years of schooling, MYS*); *ВВП на душу населення* (*GDP per capita*); *Індекс Джині* (*Gini Index*); *Індекс тероризму* (*Ranked of terrorism*); *Індекс корупції* (*Rank of corruption, RC*); *Індекс торгівлі* (*Enabling Trade Index, ETI*); *доступність і якість транспортної інфраструктури* (*Availability and quality of transport infrastructure*); *наявність і використання інформаційно-комунікаційних технологій ІКТ* (*ICT*); *Індекс соціального прогресу* (*Social Progress Index*); *доступність знань* (*Access to Basic Knowledge*, *ABK*); *охорона здоров'я* (*Health and Wellness*); *особиста свобода і свобода вибору* (*Personal Freedom and Choice*, *PFC*); *толерантність* (*Tolerance*, *T*); *Індекс глобальної конкурентоспроможності* (Global Competitiveness Index); *Індекс економічної свободи* (*Ranking of the world by economic freedom*, *REF*); *Глобальний індекс нерівності* (*The Global Peace Index*

records a less peaceful and more unequal world); *кількість населення* (*Population*); *міграція* (*Migrants*); рейтинг країн за *Індексом щасливої планети* (*HPI Rank*); *Індекс щасливої планети* (*Happy Planet Index*); *роки щасливого життя* (*Happy Life Years, HLE*); *екологічний слід* (*Footprint*, *F*); *нерівність можливостей* (*Inequality of Outcomes*, *IO*); *очікувана тривалість життя з урахуванням нерівності* (*Inequality-adjusted Life Expectancy*, *IALE*); *благополуччя з урахуванням нерівності* (*Inequality-adjusted Wellbeing*, *IAW*); *Індекс щасливої планети* (*Happy Planet Index*).

Величини, які визначають значення та характеристики (фактори) сталого розвитку (прогресу), неможливо виміряти безпосередньо. Крім того, на сьогодні остаточно не визначено не лише перелік, а й оптимальну кількість та реальний зміст цих факторів. Насправді їх кількість може бути в рази меншою, ніж кількість вимірюваних ознак, які використовують для оцінювання стандартів життя.

Для пошуку передбачуваних, неочевидних закономірностей, зумовлених впливом зовнішніх (внутрішніх) факторів на динаміку показників сталого розвитку, виявлення та вивчення статистичного зв'язку виділених ознак з суттєвими факторами використаємо один із методів факторного аналізу – **метод головних компонент** [36, с. 286].

Завданням факторного аналізу є об'єднання великої кількості ознак, які характеризують економічний (політичний) процес чи об'єкт, у меншу кількість штучно створених на їх основі факторів. Отримана у результаті система факторів має описувати вибіркові дані не гірше, ніж початкова, та бути найбільш зручною з точки зору змістовної інтерпретації [46, с. 21; 34, с. 162; 14, с. 61].

Інтерпретація результатів факторного аналізу

Основні результати факторного аналізу виражаються у факторних навантаженнях, факторних полях, факторних вагах та власних значеннях факторів.

Факторні навантаження (factor loadings) – коефіцієнти кореляції кожної з аналізованих змінних із кожним з виділених факторів. Чим тісніший зв'язок змінної з фактором, тим більшим є її факторне

навантаження. Додатний знак факторного навантаження вказує на прямий зв'язок змінної з фактором, від'ємний – на обернений.

Для всіх факторів вказано навантаження кожної початкової змінної, яке відображає величину проекції змінної на факторну координатну вісь (рис. 5.11). Факторні навантаження можна інтерпретувати як кореляції між відповідними змінними та факторами – чим вище навантаження за модулем, тим ближче фактор до вихідної змінної. Факторні навантаження відображають найбільш важливу інформацію для інтерпретації отриманих факторів.

У результаті застосування факторного аналізу для вивчення вагомих факторів впливу на глобальний сталий розвиток виділено 2 фактори. З першим із них найтісніше пов'язані *Індекс людського розвитку, нерівність доходів, очікувана тривалість життя з урахуванням нерівності, очікувана тривалість навчання дітей шкільного віку та доступність знань*. На другий виділений фактор найбільше впливають показник корупції і *ВВП на душу населення* для заданого мінімального факторного навантаження 0,7. Виділені фактори можна трактувати як соціальний (фактор 1) та економічний (фактор 2) фактори сталого розвитку, що підтверджує і графік факторних навантажень (рис. 5.12).

Variable	Factor 1	Factor 2
HDI	0,884516	0,435184
Happy Life Years	0,687130	0,661293
Footprint	0,565247	0,594708
Inequality of Outcomes	-0,873233	-0,418710
Inequality-adjusted Life Expectancy	0,853880	0,393271
Inequality-adjusted Wellbeing	0,520201	0,690018
GDP/capita ($PPP)	0,402834	0,799481
Expected years of schooling	0,284345	0,143061
Mean years of schooling	0,883171	0,299606
Corruption	0,423172	0,822170
Enabling Trade Index	0,578123	0,703026
Transport infrastructure	0,490810	0,582846
ICT	0,796148	0,543051
Access to Basic Knowledge	0,925891	0,174293
Health and Wellness	0,053029	0,622168
Personal Freedom and Choice	0,489358	0,748425
Tolerance and Inclusion	0,272274	0,791815
Economic freedom	0,432982	0,657479
Expl.Var	7,108605	6,377558
Prp.Totl	0,394922	0,354309

Рис. 5.11. Таблиця факторних навантажень

Щоб встановити значущість виділених факторів, проведено аналіз власних значень факторів (рис. 5.13). Його використовують, щоб встановити, який із факторів найбільш значущий. **Власні значення (eigenvalues)** – це дисперсії, які пояснюються факторами.

Отримані результати:

► **Eigenvalue** – **дисперсії** кожного фактора;

►**% Total Variance** – **відсоток від загальної дисперсії** для кожного фактора. Для прикладу 5.1 перший фактор пояснює приблизно 27% дисперсії, а другий фактор – 12% дисперсії;

►**Cumul Eigenvalue** – накопичена або **кумулятивна дисперсія** факторів;

►**Cumulative % – накопичений відсоток від загальної дисперсії.**

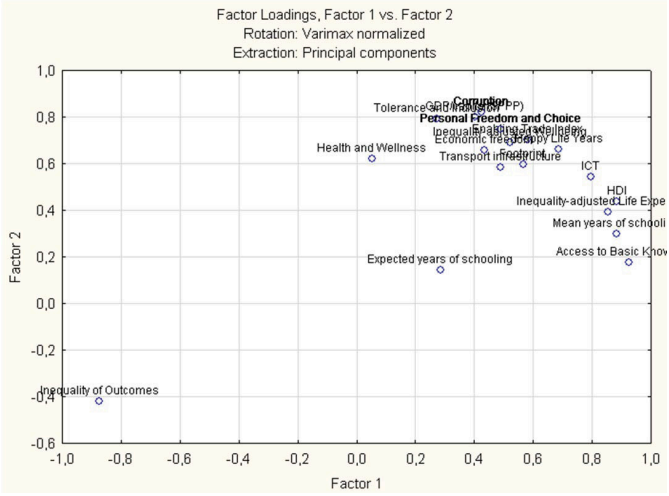


Рис. 5.12. Графічне представлення факторних навантажень

Value	Eigenvalue	% Total variance	Cumulative Eigenvalue	Cumulative %
1	12,15230	67,51280	12,15230	67,51280
2	1,33386	7,41033	13,48616	74,92313

Рис. 5.13. Власні значення виділених факторів

Для аналізу важливими є такі показники:

– накопичений відсоток дисперсії обох факторів (75%), який визначає, наскільки повно вдалося описати початкову сукупність даних за допомогою виділених факторів. Чим вищий цей показник, тим більшу частину масиву даних вдалося факторизувати;

– відсоток загальної дисперсії для кожного фактора, який вказує на значущість цього фактора. Чим більший відсоток дисперсії пояснює фактор, тим він більш значущий і тим більше змінних він вміщає.

У прикладі 5.3 перший виділений фактор пояснює 68% дисперсії, другий – 7%. Разом вони описують 75% дисперсії, тобто три чверті масиву даних (рис. 5.13). Це означає, що проведена факторизація неповна, існують й інші фактори, менш значущі, однак теж достатньо важливі. Зокрема, тіньова економіка, що не врахована в запропонованій моделі, свобода віросповідання, якість питної води, задоволеність роботою, ментальність тощо.

Для проведення якісного факторного аналізу потрібно встановити, скільки факторів необхідно виділити, щоб вони максимально повно описували весь масив даних та були достатньо значущими (рис. 5.14).

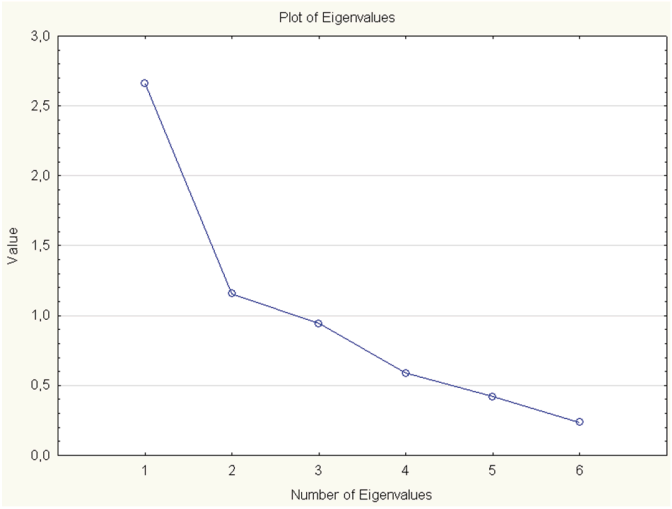


Рис. 5.14. Графік власних значень

Критерій кам'янистого насипу – графік, на якому в порядку спадання зображують власні значення кожного виділеного фактора. Ордината точки на графіку, де зменшення власних значень зліва направо максимально сповільнюється (2), вказує на оптимальну кількість факторів, якими можна обмежитись для проведення адекватного аналізу.

Модулі факторного аналізу програми «Statistica» автоматично використовують критерій Кайзера, за яким для аналізу відбирають лише фактори, власні значення яких більші за 1.

Критерій Кайзера іноді зберігає занадто багато факторів, а критерій кам'янистого насипу – занадто мало. Однак кожен з цих критеріїв дає відповідний результат, якщо їх застосовувати в нормальних умовах (невелика кількість факторів та велика кількість змінних і об'єктів). Часто використовують обидва критерії послідовно – відсікають зовсім незначущі фактори з допомогою критерію Кайзера, а потім до тих факторів, які залишилися, застосовують критерій кам'янистого насипу.

Переважно досліджують декілька рішень з різною кількістю факторів, а потім обирають одне, найбільш осмислене.

У результаті проведеного аналізу отримано таку факторну модель:

$$F_1=0{,}88\times HDI+0{,}87\times IO+0{,}85\times IAW+0{,}88\times MYS+0{,}79\times ICT+0{,}93\times ABK;$$

$$F_2=0{,}82\times GDP+0{,}82\times RC+0{,}7\times ETI+0{,}75\times PFC+0{,}8\times T,$$

де IO – нерівність можливостей (у %), IAW – благополуччя з урахуванням нерівності, MYS – середня тривалість навчання дорослого населення, ICT – наявність і використання ІКТ, ABK – доступність знань, RC – Індекс корупції, ETI – Індекс торгівлі, PFC – особиста свобода і свобода вибору, T – толерантність.

Виділено два фактори, які можна трактувати як соціальну та економічну компоненти сталого розвитку відповідно. Вони є лінійними комбінаціями аналізованих індикаторів і відображають більшу частку загальної мінливості досліджуваних ознак, тому зберігають основну частину інформації, яку містили початкові дані.

Результати факторного аналізу будуть якісними, якщо вдасться їх інтерпретувати відповідно до смислу показників, які характеризують виділені фактори. Цей етап діяльності потребує від дослідника досвіду аналізу та чіткого уявлення про аналізовані змінні. Включення в набір змінних якомога більшої кількості різноманітних показників не покращить результатів дослідження. Факторний аналіз не дає якогось нового знання, він лише допомагає виявити структуру заданих змінних. При підборі змінних для факторного аналізу потрібно керуватися їх змістовним наповненням.

5.3. Дискримінантний аналіз

Дискримінантний аналіз (ДА) є розділом багатовимірного статистичного аналізу, що охоплює методи класифікації багатовимірних спостережень за принципом максимальної схожості за наявності навчальних вибірок. Його застосовують в економіці, психології, соціології, політиці та інших науках [54, с. 196].

На відміну від кластерного аналізу (п. 5.1), у ДА нові кластери не утворюються. Лише формулюється правило, за яким об'єкти, що підлягають класифікації, відносять до однієї з існуючих (навчальних) підмножин (класів) на основі порівняння величини функції дискримінанта об'єкта, що класифікується, з деякою константою дискримінації. Наприклад, у досліджуваній системі з'являється новий об'єкт, який характеризує та сукупність ознак, що вивчаються. Необхідно встановити, до якого класу він належить з найбільшою ймовірністю.

Дискримінантну функцію в загальному випадку записують у вигляді лінійної комбінації початкових показників, коефіцієнти якої підбирають з умови найбільших відмінностей функції між відомими класами.

Сфери застосування дискримінантного аналізу:

► статистичний аналіз і моделювання залежностей між окремими ознаками з врахуванням розподілу неоднорідних статистичних сукупностей на однорідні групи (класи);

► вибір змінних, які якнайкраще розрізняють (дискримінують) утворювані сукупності (дві або більше);

► класифікація нових об'єктів на основі виявлених залежностей;

► складання карти сприйняття;

► прогнозування поведінки нових об'єктів спостереження шляхом їх зіставлення з поведінкою об'єктів навчальних підмножин;

► уточнення результатів класифікації, отриманих методом кластерного аналізу і т. ін.

Наприклад, ТНК при укладанні угод класифікує своїх контрагентів за певними ознаками на надійних і ненадійних. За допомогою ДА можна визначити, до якої з раніше виявлених сукупностей (навчальних вибірок) можна віднести нових бізнес-партнерів. При проведенні маркетингових досліджень можна виявити відмінні характеристики споживачів товарів, які реагують на відповідний вид реклами.

Математичний опис задачі дискримінантного аналізу

Розглянемо множину M, яка складається з n об'єктів спостереження. Кожний i-й об'єкт множини M описує сукупність p значень дискримінантних змінних (ознак) x_j $\left(i = \overline{1,n},\ \ j = \overline{1,p}\right)$. При цьому множина об'єктів M містить q $\left(q \geq 2\right)$ навчальних підмножин M_k розмірності n_k кожна і підмножину M_0 об'єктів, які підлягають дискримінації, k – номер підмножини $\left(k = \overline{1,q}\right)$.

Потрібно встановити правило (лінійну або нелінійну дискримінантну функцію $f(x)$) розподілу m об'єктів підмножини M_0 з відповідними ознаками за підмножинами M_k. Вибір вигляду дискримінантної функції $f(x)$ залежить від геометричного розташування розділяючих класів у просторі дискримінантних змінних. Геометрична інтерпретація постановки завдання ДА на прикладі двох навчальних підмножин M_1 і M_2 $\left(q = 2\right)$ представлена на рис. 5.15.

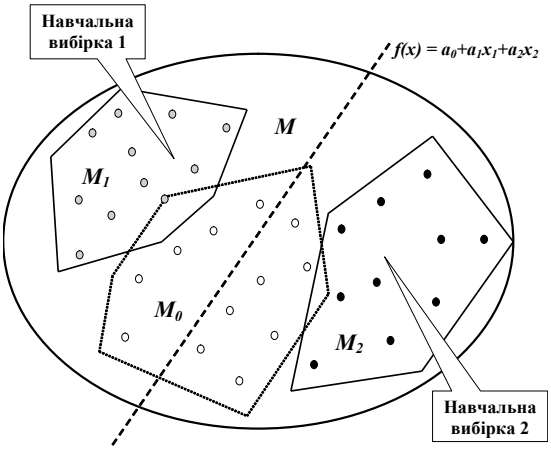


Рис. 5.15. Геометрична ілюстрація постановки завдання ДА $(q = 2)$

Найчастіше використовують лінійну форму дискримінантної функції, яку представляють у вигляді скалярного добутку вектора дискримінантних множників $A = (a_1, a_2, ..., a_p)$ і вектора дискримінантних змінних $X^T = (x_{i1}, x_{i2}, ..., x_p)$:

$$F_i = AX_i^T$$

або

$$F_i = a_1 x_{i1} + a_2 x_{i2} + \cdots a_p x_{ip},$$

де X^T – транспонований вектор дискримінантних змінних x_{ij} (значень *j*-х ознак у *i*-го об'єкта спостережень).

Основні припущення теорії дискримінантного аналізу:

► множина M_0 об'єктів розбита на декілька $(q \geq 2)$ навчальних підмножин (класів) M_k, які відрізняються від інших груп змінними x_{ij};

► у кожній підмножині M_k знаходиться принаймні два об'єкти спостереження $(n_k \geq 2)$, причому всі об'єкти множини *M* мають належати одній з підмножин (класів) M_k;

► кількість n об'єктів спостереження має перевищувати кількість p дискримінантних змінних $(0 < p < n-2)$ не менше, ніж на дві одиниці. Кількість p вибирають на підставі логічного аналізу початкової інформації;

► дискримінантні змінні x_{ij} мають бути виміряними за інтервальною шкалою або шкалою відношень. **Інтервальну шкалу** використовують для кількісного опису відмінностей між властивостями об'єктів. Її задають точкою відліку (наприклад, середня величина, мода, медіана) і одиницею виміру. **Шкала відношень** (окремий випадок інтервальної шкали) дає можливість зіставити кількісні характеристики різних об'єктів;

► між дискримінантними змінними існує лінійна незалежність (відсутня мультиколінеарність), тобто жодна змінна не має бути лінійною комбінацією інших змінних;

► усередині кожної підмножини M_k виконується нормальний закон розподілу дискримінантних змінних x_{ij} при фіксованих значеннях інших змінних.

Якщо ці умови не виконуються, розглядають питання про доцільність використання дискримінантного аналізу для класифікації нових спостережень.

Основні проблеми ДА – відбір дискримінантних змінних і вибір вигляду дискримінантної функції. Для отримання найкращих відмінностей навчальних підмножин можна використовувати критерії послідовного відбору змінних або покроковий дискримінантний аналіз. Після визначення набору дискримінантних змінних вирішують питання про вибір вигляду дискримінантної функції (лінійної або нелінійної).

Дискримінантними змінними можуть бути не лише початкові (спостережувані) ознаки, а й головні компоненти або головні фактори, виділені у факторному аналізі.

Геометрична інтерпретація дискримінантних функцій

Геометрично дискримінантні змінні представляють у вигляді осей p-вимірного евклідового простору, в якому кожен i-й об'єкт спостереження є точкою цього простору з координатами x_{ij}.

Скупчення точок у просторі утворюють q навчальних підмножин (класів), відмінності та взаємне положення яких можна визначити за допомогою їхніх центроїдів.

Центроїд – уявна точка підмножини, координати якої обчислюють як середні значення змінних усередині цієї підмножини. Використовують для опису відмінностей між підмножинами також визначення належності до них нових об'єктів.

Центроїди характеризують положення k-ї підмножини у просторі розмірністю q–1, тобто на одиницю меншої за кількість підмножин (класів).

На рис 5.16 у координатах двох дискримінантних змінних x_1 і x_2 зображені дві підмножини M_1 і M_2 (навчальні вибірки) множини M, всередині яких точками позначені належні їм об'єкти.

Положення кожного i-го об'єкта k-ї підмножини характеризують дві дискримінантних змінних $x_{i1}^{(k)}$, $x_{i2}^{(k)}$. Підмножини M_1 і M_2 розділені лінійною комбінацією дискримінантних змінних x_1 та x_2 вигляду:

$$f(x) = a_1 x_1 + a_2 x_2 . \qquad (5.1)$$

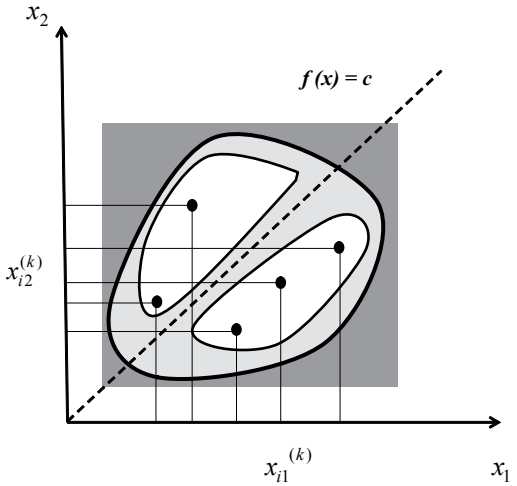


Рис. 5.16. Геометричне представлення дискримінантних змінних x_1 , x_2

Дискримінантні множники a_1, a_2 канонічної дискримінантної функції (5.1) для *k*-ї підмножини дають можливість перейти від двовимірного простору початкових показників до одновимірного, забезпечуючи при цьому мінімальну помилку класифікації.

Поворот осей до збігу осі x_1 з лінією, яка розділяє підмножини, дає можливість перейти до нової системи координат (рис. 5.17). Нова система координат з осями x_1' і x_2' розділяє проекції на вісь x_2' об'єктів, які належать різним підмножинам, таким чином, що підмножини M_1 і M_2 знаходяться по різні сторони від осі x_1'. Межа, яка розділяє підмножини M_1 і M_2, збігається з віссю x_1'. Вона задана функцією $\overline{f}(x) = 0{,}5\left(\overline{f}^{(1)}(x) + \overline{f}^{(2)}(x)\right)$, рівновіддаленою від $\overline{f}^{(1)}$ і $\overline{f}^{(2)}$.

Величину $\overline{f}(x)$ називають **константою дискримінації**. З рис. 5.16 і 5.17 видно, що об'єкти, розташовані над прямою $\overline{f}(x)$, знаходяться ближче до центра підмножини M_1, тому їх відносять до першої групи. Об'єкти, розташовані нижче від цієї прямої, містяться ближче до центра підмножини M_2, тому їх відносять до другої групи. Такий вибір межі $\overline{f}(x)$ забезпечує мінімальну ймовірність помилки класифікації.

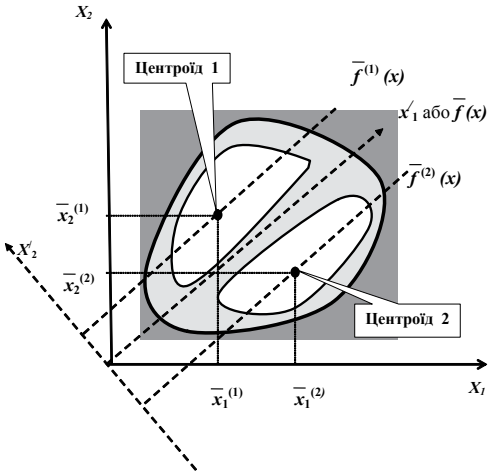


Рис. 5.17. Геометрична інтерпретація центрів (центроїдів) двох навчальних підмножин M_1 і M_2

Лінійна функція (5.1) є проекцією *i*-го об'єкта на деяку (одновимірну) дискримінанту вісь. Якщо середнє значення *j*-ї ознаки в об'єктів *k*-ї підмножини позначити через $\overline{x}_j^{(k)}$, то середні значення функцій $\overline{f}^{(1)}$ і $\overline{f}^{(2)}$ у підмножинах визначаються за формулами:

$$\overline{f}^{(1)}(x) = a_1\overline{x}_1^{(1)} + a_2\overline{x}_2^{(1)};$$

$$\overline{f}^{(2)}(x) = a_1\overline{x}_1^{(2)} + a_2\overline{x}_2^{(2)}.$$

Геометрично функції $\overline{f}^{(1)}$, $\overline{f}^{(2)}$ є двома паралельними прямими, які проходять через центри (центроїди) підмножин (рис. 5.17).

Лінійну дискримінантну функцію не завжди можна використовувати для представлення розділяючої прямої (поверхні) між підмножинами. Наприклад, якщо навчальні підмножини не є опуклими, то лінійна дискримінантна функція не забезпечує мінімальних помилок класифікації.

Якщо навчальні підмножини близько розташовані одна до одної, ймовірність помилкової класифікації нових об'єктів зростає у міру їх віддалення від центрів обох множин. У цьому разі переглядають набір дискримінантних змінних.

Для графічного зображення центроїдів і окремих об'єктів спостереження у просторі дискримінантних функцій будують **карти сприйняття**. Вони візуалізують кількісні дані про схожість об'єктів спостереження у формі графіків (одно- чи двокоординатних).

При використанні однієї дискримінантної функції точки, які відповідають об'єктам, обчисленим за цією функцією, розташовують вздовж деякої прямої в однокоординатній системі. В цьому разі точки характеризують частку функції, яка відноситься до цього спостереження. Недолік такого способу відображення в тому, що при великій кількості точок втрачається інформація про щільність їх розподілу. У такому разі будують гістограми для кожної групи. Це дає можливість проводити відносне порівняння розподілу груп.

При побудові декількох дискримінантних функцій центроїди та об'єкти відображають на графіках у двохкоординатній системі в про-

сторі значень двох функцій. Якщо функцій більше двох, такий графік дуже інформативний, оскільки перші дві функції залишаються найбільш важливими порівняно з іншими. Групові центроїди виділяють символами (зірочка і т. ін.) на тлі інших об'єктів спостережень.

Карти сприйняття дають можливість візуально оцінити ступінь відмінності груп і взаємне розташування їх центроїдів. Прямі лінії в полі графіка розділяють «території» відповідних груп (рис. 5.18).

Критерії порівняння вибірок за кількома ознаками

Першим критерієм порівняння вибірок є **коефіцієнт детермінації**. Для кількох груп загальну дисперсію ознаки можна представити у вигляді суми міжгрупової та внутрішньогрупової дисперсій.

Мірою мінливості в цьому разі будуть суми квадратів відхилень спостережень від відповідних середніх:

$$SS_x = SS_u + SS_e, \tag{5.2}$$

де SS_x – сума квадратів відхилень спостережень від загального середнього, що характеризує загальну мінливість;

SS_u – міжгрупова дисперсія (сума квадратів відхилень групових середніх від загального середнього), що характеризує мінливість між групами;

$SS_{e:}$ – внутрішньогрупова дисперсія (сума квадратів відхилень спостережень від групових середніх), що характеризує мінливість усередині груп.

Розділимо обидві частини рівняння (5.2) на SS_x:

$$1 = \frac{SS_u}{SS_x} + \frac{SS_e}{SS_x} \ .$$

Відношення $\frac{SS_u}{SS_x}$ називають **коефіцієнтом детермінації** та позначають η^2. Він показує, у скільки разів мінливість спостережень між групами перевищує загальну мінливість.

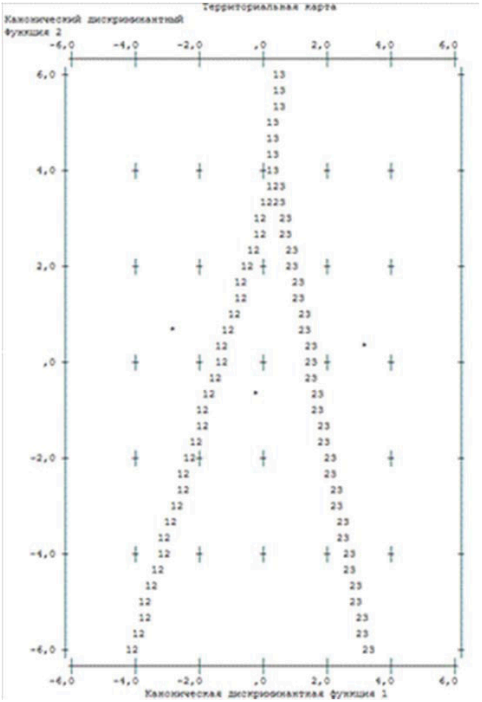


Рис. 5.18. Карта сприйняття для двох дискримінантних функцій

Значення цього коефіцієнта може знаходитись у межах $0 \le \eta^2 \le 1$. Якщо всі групові середні дорівнюють загальному середньому, $SS_u = 0$ і $\eta^2 = 0$ (середньогрупові значення показника x у групах однакові). Якщо всередині груп немає ніякої мінливості, $SS_e = 0$ і $\eta^2 = 1$, а це означає, що різним значенням змінної x відповідають різні класи. Чим ближче η^2 до одиниці, тим краща дискримінаційна здатність змінної x.

Квадратний корінь з коефіцієнта детермінації η^2 називають **емпіричним кореляційним відношенням**.

Другим критерієм є **характеристика λ (власне значення)**. Вона показує, у скільки разів мінливість між групами перевищує мінливість всередині груп:

$$\lambda = \frac{SS_u}{SS_e}.$$

Лямбда характеризує частку дисперсії оцінок дискримінантної функції, яка не обумовлена відмінностями між групами. Якщо середні для всіх груп рівні, то $\lambda = 1$ і зменшується зі зростанням різниць середніх значень. Чим більше λ, тим краще підібрана дискримінантна функція.

Обидва критерії пов'язані співвідношенням:

$$\eta = \frac{\lambda}{1+\lambda}.$$

Якість класифікації в канонічному дискримінантному аналізі оцінюють за однією з таких характеристик:

1. **Відносний відсотковий зміст** показує, на скільки відсотків ця функція слабша за інші:

$$\frac{\lambda_i}{\sum_{j=1}^{p}\lambda_j}.$$

2. **Коефіцієнт канонічної кореляції** показує, яка частина загальної мінливості дискримінантної функції пояснюється різницею між групами:

$$\eta_i = \sqrt{\frac{\lambda_i}{1+\lambda_i}}.$$

3. **Критерій Фішера** розраховують за формулою:

$$F = \lambda \frac{n-q}{q-1}$$

і порівнюють з табличними значеннями $F_{\alpha,q-1,n-q}$ на вибраному рівні значущості α (переважно α = 0,01 або 0,05) з кількістю ступенів вільності $q - 1$ та $n - q$ (p – кількість змінних, q – кількість класів) або оцінюють за рівнем значущості α.

Рівні значущості характеризують імовірність того, що відмінності між групами є випадковими. Дискримінантну функцію вважають значущою на заданому рівні значущості α з кількістю ступенів вільності v, якщо для неї фактичні значення критерію χ^2 перевищують табличні $\chi^2_{\alpha,v}$. Замість цього можна використовувати рівень значущості – ймовірність того, що статистика χ^2 при нульовій гіпотезі (незначущості дискримінантної функції) випадково досягне обчисленого рівня. Якщо для функції $\alpha < 0{,}01$, знайдена дискримінантна функція значуща.

4. **Статистика Уїлкса** – в аналіз вводять ту змінну, яка на цьому кроці має найменшу Λ-статистику. Цей критерій відбору вважають найкращим. Він оцінює відносний внесок залишкової дисперсії. Його перевага в тому, що критерій враховує не лише відмінності між класами, а й однорідність кожного класу (міру скупчення об'єктів навколо центроїдів).

Оцінка інформативності показника після статистики Уїлкса аналогічна оцінці значущості членів регресійної моделі за допомогою коефіцієнтів часткової кореляції.

Якщо l – загальна кількість дискримінантних функцій з ненульовими λ_i, то

$$\Lambda_0 = \frac{1}{1+\lambda_1} \cdot \frac{1}{1+\lambda_2} \cdot \frac{1}{1+\lambda_3} \cdot \ldots \cdot \frac{1}{1+\lambda_l}$$

є мірою залишкової мінливості з урахуванням всіх дискримінантних функцій, тобто Λ_0 оцінює дискримінаційну здатність усієї системи функцій.

Далі оцінюють дискримінаційну здатність системи без першої, найбільш важливої функції:

$$\Lambda_1 = \frac{1}{1+\lambda_2} \cdot \frac{1}{1+\lambda_3} \cdots \frac{1}{1+\lambda_l}.$$

Ця величина більша, ніж Λ_0. Чим ближче Λ_1 до одиниці, тим менша дискримінаційна здатність системи функцій, що залишилась.

Далі обчислюють

$$\Lambda_2 = \frac{1}{1+\lambda_3} \cdots \frac{1}{1+\lambda_l}$$

і так до Λ_{l-1}.

Λ_j оцінюють послідовно за критерієм Пірсона:

$$\chi_j^2 = -\left(n - 1 - \frac{p+q}{2}\right) \ln \Lambda_j ,$$

який порівнюють з табличним значенням $\chi^2_{\alpha,\nu}$. У цьому разі n – кількість спостережень, p – кількість змінних, q – кількість класів, $\nu = (p-j)(q-j-1)$ – кількість ступенів вільності.

Послідовний відбір змінних дає можливість знайти оптимальну кількість показників, які мають такі самі (чи кращі) дискримінантні можливості, що й повний набір початкових змінних. Чим менше показників, тим легше інтерпретувати результати аналізу.

Оскільки змінні, відібрані в модель, є сильними дискримінаторами і можуть корелювати між собою (нести одну і ту саму інформацію), то після кожного виключення (включення) змінних перераховують Λ-статистику Уїлкса та оцінюють значущість змін у цій характеристиці. Оскільки Λ-статистика Уїлкса є мірою залишку моделі (мірою невизначеності), бажано, щоб вона набувала найменшого значення.

Дискримінантний аналіз у середовищі «Statistica»

Приклад 5.4. За даними прикладу 5.3 провести дискримінантний аналіз *Індексу людського розвитку* (*ІЛР*) – одного з найбільш авторитетних рейтингів, за допомогою яких традиційно вимірюють рівень сталого розвитку окремої країни. За величиною цього показника 193 країни світу розподіляють на чотири групи: країни з найбільш високим рівнем розвитку (very high), високим (high), середнім (medium) та низьким (low) [10].

Щоб перевірити узгодженість такої класифікації з основними показниками сталого розвитку країн, проведено дискримінантний аналіз ІЛР [19, с. 168] за такими показниками: кількість років щасливого життя, екологічний слід, нерівність можливостей, очікувана тривалість життя з урахуванням нерівності, налагоджений добробут з

урахуванням нерівності, ВВП на душу населення, середня тривалість навчання дорослого населення, Індекс корупції, Індекс торгівлі, доступність і якість транспортної інфраструктури, наявність і використання ІКТ, доступність знань, охорона здоров'я, особиста свобода і свобода вибору, толерантність, Індекс економічної свободи.

З класифікаційної матриці можна зробити висновок, що розподіл країн на групи за рівнями ІЛР є коректним для ознак, вибраних для аналізу (рис. 5.19). Лише три країни неправильно віднесені до виділених груп: 2 країни у групі «високий» рівень розвитку та 1 країна у групі «середній», що становить менше 3% від загальної кількості спостережень. Отже, показники, виділені як значущі при проведенні канонічного аналізу, з високою достовірністю можуть бути використані також для класифікації країн за рівнем ІЛР.

Group	Percent Correct	very high p=,33333	high p=,29730	medium p=,18018	low p=,18919
very high	94,5946	35	2	0	0
high	100,0000	0	34	0	0
medium	100,0000	0	0	20	0
low	95,2381	0	0	1	20
Total	97,3214	35	36	21	20

Рис. 5.19. Класифікаційна матриця

Для побудови класифікаційних функцій розподілу країн за рівнями ІЛР (найвищий, високий, середній, низький) обчислено коефіцієнти класифікаційних функцій класів (рис. 5.20).

	Classification Functions; grouping: HDI Rank			
Variable	**very high p=,33333**	high p=,29730	medium p=,18018	low p=,18919
Inequality-adjusted Life Expectancy	31,54	31,43	30,73	29,04
Corruption	0,61	0,40	0,31	0,27
Access to Basic Knowledge	1,61	1,70	1,63	1,23
Mean years of schooling	11,55	9,92	8,51	7,92
Economic freedom	1,10	1,51	1,62	1,51
Enabling Trade Index	20,69	15,72	17,52	21,29
Personal Freedom and Choice	0,30	0,40	0,55	0,50
ICT	12,42	10,91	7,61	5,85
GDP/capita ($PPP)	-0,00	0,00	0,00	0,00
Inequality of Outcomes	19,13	18,60	17,94	17,72
Footprint	6,49	5,58	4,81	4,30
Happy Life Years	-39,28	-40,82	-41,51	-39,43
Inequality-adjusted Wellbeing	291,94	301,21	304,37	289,54
Constant	-1620,10	-1578,19	-1505,00	-1360,33

Рис. 5.20. Коефіцієнти класифікаційних функцій класів

Специфікація дискримінантної моделі має вигляд:

very high =–1620,1+31,54×*IALE*+0,61×*RC*+1,61×*ABK*+
+11,55×*MYS*+1,1×*REF*+ 20,69×*ETI*+0,3×*PFC*+12,42×*ICT*+
+19,13×IO+6,49×*F*+39,28×*HLE*+291,94×*IAW*;

high =–1578,19+31,43×*IALE*+0,4×*RC*+1,7×*ABK*+9,92×*MYS*+
+1,51×*REF*+15,72×*ETI*+ 0,4×*PFC*+10,91×*ICT*+18,6×*IO*+5,58×*F*+
+40,28×*HL*E+301,21×*IAW*;

medium =–1505+30,78×*IALE*+0,31×*RC*+1,63×*ABK*+8,51×*MYS*+
+1,62×*REF*+ 17,52×*ETI*+0,55×*PFC*+7,61×*ICT*+17,94×*IO*+4,81×*F*+
+41,51×*HLE*+304,37×*IAW*;

low =–1360,33+29,04×*IALE*+0,27×*RC*+1,23×*AB*K+7,92×*MYS*+
+1,51×*REF*+21,29×*ETI*+0,5×*PFC*+5,85×*ICT*+17,72×*IO*+4,3×*F*+
+39,43×*HLE*+289,54×*IAW*,

де *IALE* – очікувана тривалість життя з урахуванням нерівності, *RC* – Індекс корупції, *ABK* – доступність знань, *MYS* – середня тривалість навчання дорослого населення, *REF* – Індекс економічної свободи, *ETI* – Індекс торгівлі, *PFC* – особиста свобода і свобода вибору, *ICT* – наявність і використання ІКТ, *IO* – нерівність можливостей, *F* – екологічний слід, *HLE* – кількість років щасливого життя, *IAW* – благополуччя з урахуванням нерівності.

Це система рівнянь є лінійними комбінаціями виділених для дослідження ознак, які оптимально розподілять аналізовані групи країн за рівнями значень ІЛР.

Діаграма розсіювання канонічних значень ілюструє внесок кожної дискримінантної функції в розподіл країн за групами ІЛР (рис. 5.21).

Оцінки дискримінантного аналізу підтвердили висновки, отримані у результаті попередньо проведеного факторного аналізу: рівень сталого розвитку країни значною мірою пов’язаний із соціальними показниками, зокрема нерівністю можливостей, якістю освіти, розвитком інформаційних технологій, особистою свободою і свободою вибору. Серед економічних показників найбільш вагомим є Індекс торгівлі.

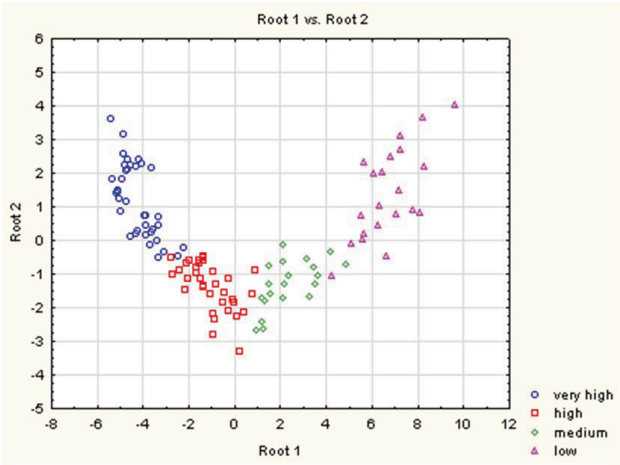


Рис. 5.21. Діаграма розсіювання канонічних значень

5.4. Канонічний аналіз

Загальні положення

Метод канонічної кореляції призначений для аналізу залежностей між двома наборами змінних. Якщо обчислення попарних кореляцій між змінними дає можливість встановити залежності між окремими парами змінних, то за допомогою методу канонічного аналізу визначають залежність між двома наборами загалом.

Наприклад, потрібно дослідити залежність між прогнозами соціальних змін, які опубліковані у трьох виданнях, та реальними змінами, які відображаються п'ятьма статистичними показниками. У такому разі метою дослідження є виявлення взаємозв'язку між двома множинами [96, с. 619].

Для обчислення канонічних коренів знаходять власні значення попарних кореляцій. Ці значення виражають частку дисперсії, яка пояснюється кореляцією між відповідними канонічними змінними. Частку обчислюють відносно дисперсії канонічних змінних. Знаходять стільки власних значень, скільки є змінних у меншому з двох досліджуваних наборів.

Власні значення послідовно обчислюють. Спочатку визначають ваги, які максимізують кореляцію між зваженими сумами за двома множинами змінних, і знаходять відповідне їм значення першого кореня. Далі обчислюють наступну пару канонічних змінних, які мають максимальну кореляцію і не корелюють з попередніми парами, з подальшим обчисленням значення канонічного кореня.

Квадратні корені з отриманих власних значень можна трактувати як коефіцієнти кореляції. Ці корені стосуються канонічних змін, тому їх називають **канонічними кореляціями**. Відповідно до власних значень послідовно добуті канонічні кореляції утворюють спадну послідовність. Змістовне тлумачення допускають не лише найбільші кореляції, а й наступні. Канонічні корені оцінюють один за одним у порядку спадання величини. Для подальшого аналізу залишають лише статистично значущі корені.

Після встановлення значущості коренів виникає проблема їх тлумачення. Кожен корінь є двома зваженими сумами, що відповідають двом наборам даних. Ці ваги трактують аналогічно до часткових кореляцій. Їх називають **канонічними вагами**. Канонічні ваги тлумачать аналогічно до вагових коефіцієнтів факторів.

Чим більша вага за абсолютним значенням, тим значніший внесок кожної змінної в значення канонічної змінної. Розгляд канонічних ваг дає можливість дослідити, як конкретні змінні в кожній множині впливають на зважену суму, тобто канонічну змінну. Канонічні ваги можуть бути використані для обчислення значень канонічних змінних. Для цього достатньо додати вхідні змінні з відповідними коефіцієнтами.

Іншим способом тлумачення канонічних коренів є розгляд звичайних кореляцій між канонічними змінними (або факторами). Ці кореляції також називають **канонічними навантаженнями** факторів. Змінні, сильно корельовані з канонічною змінною, значною мірою пояснюються нею.

Канонічні значення відповідають унікальному внескові кожної змінної у зважену суму або канонічну змінну. Навантаження

канонічних факторів відображають повну кореляцію між відповідними змінними та зваженою сумою. Можливі ситуації, коли при близьких до нуля канонічних вагах відповідні навантаження змінних дуже великі або, навпаки, при великих канонічних вагах навантаження малі. Такі випадки важко тлумачити. Ця ситуація може виникати за наявності двох щільно пов'язаних, що майже дублюють одна одну, змінних. При обчисленні ваг для зважених сум за кожною множиною до цієї суми буде включено тільки одну з цих двох змінних. Якщо більша вага буде приписана одній із змінних, внесок іншої змінної вважають несуттєвим. При цьому звичайні кореляції між існуючими сумарними значеннями двох канонічних змінних (навантаження факторів) можуть виявитися суттєвими в обох факторів.

Дисперсія. Коефіцієнти канонічної кореляції відповідають кореляції між зваженими сумами за двома множинами даних. Вони не відображають інформації про те, яку частину мінливості (**дисперсії**) кожен канонічний корінь пояснює в змінних.

Інформацію про частку дисперсії можна отримати з навантажень канонічних факторів. Вони є кореляцією між канонічними змінними та початковими змінними у відповідній множині. Піднесені до квадрата кореляції відображають частку дисперсії, що пояснюється кожною змінною. Для кожного кореня можна обчислити середнє значення цих часток. При цьому отримують середню частку мінливості, поясненої в цій множині на основі відповідної змінної.

Канонічна кореляція при піднесенні до квадрату дає частку дисперсії, що є загальною для сум за кожною множиною (канонічною зміною). Якщо помножити цю частку на частку добутої дисперсії, отримаємо **міру надлишковості** множини змінних, тобто величину, яка відображає, наскільки є надлишковою одна множина змінних при заданій іншій множині.

Надлишковість першої множини при заданій другій множині та надлишковість другої множини змінних при заданій першій множині обчислюють за рівностями:

$$Надлишковість = \frac{\sum навантаження^2_{лів}}{p} R_c^2;$$

$$Надлишковість = \frac{\sum навантаження^2_{прав}}{q} R_c^2,$$

де p – кількість змінних у першій (лівій) множині змінних, q – кількість змінних у другій (правій) множині, величина R_c^2 – квадрат відповідної канонічної кореляції.

Для отримання загального коефіцієнта надлишковості додають надлишковості за всіма значущими коренями.

За великих обсягів вибірки незначні канонічні кореляції (наприклад, 0,3) можуть виявитися **статистично значущими**. Для обчислення надлишковості цей коефіцієнт підносять до квадрата. Отримаємо незначну величину, яка свідчить про незначну частку мінливості змінних. Це потрібно враховувати при з'ясуванні, наскільки реальна мінливість в одній множині змінних пояснюється другою множиною.

Застосування критеріїв для перевірки значущості канонічної кореляції базується на припущенні, що змінні у вибірці мають багатовимірний нормальний розподіл. Рекомендується використовувати достатньо великі вибірки для отримання достовірних оцінок навантажень канонічних факторів. Зокрема, спостережень має бути у 20 (40–60) разів більше, ніж кількість досліджуваних змінних. Однак, як свідчить практика, при значних кореляціях між даними навіть малі обсяги вибірки (наприклад, 50) дають можливість у більшості випадків виявити ці кореляції.

Наявність викидів може мати значний вплив на величину коефіцієнтів кореляції. При збільшенні обсягів вибірки вплив невеликої кількості викидів нівелюється. Перед проведенням процедури

канонічного аналізу рекомендують виявити значні викиди, наприклад, за допомогою діаграми розсіювання.

Вимагається, щоб змінні в обох множинах не були повністю надлишковими. Наприклад, при включенні однієї і тієї самої змінної двічі в одну із множин формується надлишковість, при якій незрозуміло, яку ж вагу приписати цій змінній. Крім того, при надлишковості спостерігається сильна корельованість між спостережуваними змінними.

У такому разі проблематичним є обчислення відповідної оберненої матриці, що порушує процедуру обчислення канонічної кореляції. Такі кореляційні матриці називають **погано обумовленими**.

Замість розгляду звичайних сум за множинами доцільно розглядати **зважені суми**, щоб ваги, віднесені до окремих доданків, відповідали реальній структурі змінних:

$$a_1 y_1 + a_2 y_2 + \cdots + a_p y_p = b_1 x_1 + b_2 x_2 + \cdots b_q x_q,$$

де p та q – кількість змінних першої та другої множини відповідно.

Реалізація канонічного аналізу у програмі «Statistica»

Приклад 5.5. Для країн ЄС-28 дослідити залежності між *Індексом людського розвитку, Індексом задоволеності життям* і *міжнародним рейтингом щастя* (перша група показників) та якістю освіти, *впевненістю у судовій системі, загальними витратами на охорону здоров'я* і *рівнем безробіття* (друга група показників).

Змінні, що використовуються в аналізі:

Y_1 – *Індекс людського розвитку (Human Development Index, HDI);*

Y_2 – *рейтинг щастя (Ranking of Happiness, RH);*

Y_3 – *Індекс економічної свободи (Index of economic freedom, IEF);*

X_1 – Індекс гендерного розвитку (*Gender Development Index*, *GDI*);

X_2 – якість освіти, % задоволених (*Education quality*, *EQ*);

X_3 – *здоров'я та якість догляду,* % задоволених (*Health care quality*, *HCQ*);

X_4 – *стандарт життя,* % задоволених (*Standard of living*, *SL*);
X_5 – *відчуття безпеки,* % стверджувальних відповідей (*Feeling safe*, *FS*);
X_6 – *ідеальна робота,* % стверджувальних відповідей (*Ideal job*, *IJ*);
X_7 – *загальні витрати на охорону здоров'я*, у % від ВВП (*Health expenditure, HET*).

Попередньо проведений аналіз показників дає можливість висунути гіпотезу про те, що ознаки Y_1, Y_2, Y_3 є результатними (залежними) ознаками. Ознаки X_1, X_2, X_3, X_4, X_5, X_6, X_7 – факторні (незалежні). При цьому не заперечується наявність також зворотних зв'язків між цими групами змінних.

Завдання:

– оцінити величину канонічної кореляції між двома групами показників;

– перевірити статистичну значущість;

– виявити внутрішні латентні властивості досліджуваних індикаторів (канонічні змінні), пояснити їхній економічний зміст та дати їм кількісну оцінку;

– виявити можливі практичні напрямки оцінювання канонічних змінних для кожного об'єкта спостереження.

Для аналізу використано дані відповідних показників за 2017 р. (рис. 5.22).

	1 Human Development Index	2 Gender Development Index	3 Education quality	4 Health care quality	5 Standard of living	6 Feeling safe	7 Ideal job	8 Ranking of Happiness	9 Health expenditure	10 Index of economic freedom
Австрія	0,885	0,943	75	89	84	81	84	7,200	11,2	71,7
Бельгія	0,890	0,975	83	89	81	72	73	6,937	10,6	68,4
Болгарія	0,782	0,991	42	38	37	54	51	4,218	8,4	65,9
Великобританія	0,907	0,965	65	77	79	79	71	6,867	9,1	76,4
Греція	0,865	0,961	45	35	36	62	59	4,857	8,1	53,2
Данія	0,923	0,977	75	85	91	80	79	7,527	10,8	75,3
Естонія	0,861	1,030	52	51	46	65	50	5,429	6,4	77,2

Рис. 5.22. Фрагмент файла даних

Аналіз результатів

Отримано таблиці загальних підсумків дослідження, які стосуються r_1, y_1, x_1 (рис. 5.23) [39, с. 35].

Canonical Analysis Summary (canonical)
Canonical R: ,92768
Chi?(21)=58,878 p=,00002

N=28	Left Set	Right Set
No. of variables	3	7
Variance extracted	100,000%	76,5856%
Total redundancy	70,1633%	56,2250%
Variables: 1	Human Development Index	Gender Development Index
2	Ranking of Happiness	Education quality
3	Index of economic freedom	Health care quality
4		Standard of living
5		Feeling safe
6		Ideal job
7		Health expenditure

Рис. 5.23. Таблиця загальних підсумків канонічного аналізу

Канонічне значення *R*. Отримане канонічне значення R достатньо велике (0,92); статистична значущість також висока (p < 0,01). Отримане R (рис. 5.23) відноситься до першого (найбільш значущого) канонічного кореня. Цю величину можна інтерпретувати як кореляцію між зваженими сумами змінних у першій та другій множинах.

Загальна частка дисперсії ознак першої групи ($Y_1 – Y_3$), виділена за допомогою першої канонічної змінної y_1, становить 100%. Загальна частка дисперсії ознак другої групи ($X_1 – X_4$) канонічної змінної x_1 становить 76,59%.

Загальна втрата для змінних першої групи дорівнює 70,16%, для змінних другої групи – 56,22%. Це означає, що 70,16% варіації *ІЛР, рейтингу щастя* та *Індексу економічної свободи* визначаються змінами аналізованих індикаторів. Водночас змінні $Y_1 – Y_3$ детермінують 56,22% варіації досліджуваних індикаторів $X_1 – X_4$.

Отримані результати свідчать про достатньо високу точність побудованої канонічної моделі: менше 24% дисперсії змінних $Y_1 – Y_3$ залежать від інших, неврахованих у моделі факторів.

Характеристичні корені (**Eigenvalues**) – таблиця квадратів коефіцієнтів канонічної кореляції r_1^2, r_2^2, r_3^2 (рис. 5.24). Кількість характеристичних коренів дорівнює кількості змінних у меншій із множин: min(3, 7) = 3.

	Eigenvalues (canonical)		
Root	**Root 1**	Root 2	Root 3
Value	0,860583	0,407925	0,216591

Рис. 5.24. Таблиця характеристичних коренів

Перевірка значущості канонічних коренів (рис. 5.25)**.**

	Chi-Square Tests with Successive Roots Removed (canonical)					
Root Removed	**Canonicl R**	Canonicl R-sqr.	Chi-sqr.	df	p	Lambda Prime
0	0,927676	0,860583	58,87791	21	0,000019	0,064667
1	0,638690	0,407925	16,51679	12	0,168765	0,463837
2	0,465394	0,216591	5,24817	5	0,386371	0,783409

Рис. 5.25. Перевірка значущості канонічних коренів

У першому стовпці наведені канонічні коефіцієнти кореляції r_1, r_2, r_3, у другому – їх квадрати (характеристичні корені λ_1, λ_2, λ_3) матриці, в третьому – розрахункові значення для послідовно виділених канонічних коренів, у четвертому – кількість ступенів вільності статистики χ^2, у п'ятому – *p*значущість, у шостому – значення Λ'-статистики Уїлкса.

Послідовний критерій значущості. Спочатку розглядаються всі три канонічні змінні разом (без видалення коренів). Отримане значення є суттєво значущим (*p*-рівень < 0,0000019).

Далі перший (найбільш значущий) корінь видаляється і визначається статистична значущість двох інших. Значення (у другому рядку таблиці результатів) не є значущим (*p*-рівень = 0,168765). Можна на основі цього зробити висновок, що тільки перший корінь є статистично значущим і має бути вивчений більш детально.

Якщо б значення при другому застосуванні критерію теж було значущим, потрібно було б розглянути також третій рядок таблиці результатів, щоб перевірити, чи значущий третій корінь.

Факторна структура та надмірність

Далі будемо розглядати лише перший канонічний корінь. Потрібно надати йому смислову інтерпретацію, тобто дослідити, як він корелює зі змінними у двох початкових множинах. Ці кореляції називають **навантаженнями канонічних факторів** або **структурними коефіцієнтами**.

Факторна структура лівої множини (рис. 5.26).

Variable	Factor Structure, left set (canonical)		
	Root 1	Root 2	Root 3
Human Development Index	-0,876207	-0,075255	-0,476024
Ranking of Happiness	-0,998663	0,018449	0,048284
Index of economic freedom	-0,530666	0,842895	-0,089001

Рис. 5.26. Факторна структура лівої множини

Перші дві змінні Y_1 та Y_2 (ІЛР та рейтинг щастя відповідно) мають значне навантаження на перший канонічний фактор, тобто сильно корелюють з ним.

Як міра надмірності використовується середнє значення дисперсії, що пояснюється першим коренем. Фактично це сума квадратів навантажень канонічних факторів, поділена на 3 (кількість змінних у першій множині).

Побудуємо таблицю результатів (рис. 5.27).

Factor	Variance Extracted (Proportions), left set (canonical)	
	Variance extractd	Reddncy.
Root 1	0,682224	0,587111
Root 2	0,238825	0,097423
Root 3	0,078950	0,017100

Рис. 5.27. Таблиця частки видобутої дисперсії для лівої множини

Перший канонічний корінь «витягує» в середньому близько 68% дисперсії зі змінних другої множини. Якщо помножити це значення на частку дисперсії, загальної між канонічними змінними в двох множинах (на R-квадрат), отримаємо число у другому стовпці таблиці результатів (надмірність).

Отже, задаючи значення змінних у правій множині, можна пояснити близько 59% дисперсії у змінних лівої множини, на основі значення першого канонічного кореня.

Факторна структура правої множини

Лише змінна X_4 (*стандарт життя*) має значне навантаження на перший канонічний фактор, тобто сильно корелює з ним (рис. 5.28).

	Factor Structure, right set (canonical)		
Variable	Root 1	Root 2	Root 3
Gender Development Index	0,396222	0,624681	0,120897
Education quality	-0,831935	-0,057076	0,243792
Health care quality	-0,878036	-0,024235	-0,008036
Standard of living	-0,976523	-0,011211	-0,046380
Feeling safe	-0,770337	-0,230262	-0,395266
Ideal job	-0,644437	-0,293817	-0,393491
Health expenditure	-0,679621	-0,596383	-0,200460

Рис. 5.28. Факторна структура правої множини

Навантаження, що відповідають змінній *Індекс гендерної рівності*, набагато менше за інші. Можна зробити висновок, що значна кореляція між змінними у двох множинах (на підставі відомостей про перший корінь) імовірно є наслідком залежності між станом охорони здоров'я, якістю освіти, високими стандартами життя, безпекою та індикаторами, що визначають якість життя (ІЛР та позицією у рейтингу щастя). Отримані результати пояснює також загальна кореляційна матриця початкових змінних (рис. 5.29).

	Correlations, left set with right set (canonical)						
N=28	Gender Development Index	Education quality	Health care quality	Standard of living	Feeling safe	Ideal job	Health expenditure
Human Development Index	-0,378872	0,624961	0,716644	0,804566	0,724792	0,625117	0,625495
Ranking of Happiness	-0,356996	0,775540	0,812978	0,903512	0,702072	0,584725	0,618094
Index of economic freedom	0,136234	0,368726	0,419531	0,476614	0,271637	0,175370	0,021809

Рис. 5.29. Кореляційна матриця

Якщо розглядати якість життя як пояснюючу змінну, можна зробити висновок, що вона впливає на ІЛР та стан щастя, але не впливає (або впливає набагато менше) на економічну свободу (рис. 5.28).

Перший канонічний корінь пояснює майже 57% дисперсії у змінних правої множини (рис. 5.30). Змінюючи значення змінних

лівої множини, на підставі першого канонічного кореня можна пояснити близько 50% дисперсії в інших змінних.

	Variance Extracted (Proportions), right set (canonical)	
Root Variable	**Variance extractd**	Reddncy.
Root 1	0,577751	0,497202
Root 2	0,127031	0,051819
Root 3	0,061074	0,013228

Рис. 5.30. Таблиця частки видобутої дисперсії правої множини

Обчислимо навантаження канонічних факторів для обох вихідних множин (рис. 5.31).

	Canonical Weights, left set (canonical)		
Variable	**Root 1**	Root 2	Root 3
Human Development Index	-0,095846	-0,259202	-1,88334
Ranking of Happiness	-0,923439	-0,395300	1,76225
Index of economic freedom	0,011655	1,171898	-0,20672

	Canonical Weights, right set (canonical)		
Variable	**Root 1**	Root 2	Root 3
Gender Development Index	-0,153288	0,71516	-0,15906
Education quality	-0,204511	-0,30982	1,12847
Health care quality	-0,060806	0,51016	0,03624
Standard of living	-0,755556	0,55717	0,08961
Feeling safe	-0,037711	0,02816	-1,17529
Ideal job	0,013735	0,53714	-1,08335
Health expenditure	-0,116521	-1,20473	0,70974

Рис. 5.31. Таблиці канонічних ваг

Це три пари канонічних змінних y_1 і x_1, y_2 і x_2, y_3 і x_3 (стовпці на рис. 5.31). Розглядаємо лише першу пару канонічних змінних, найбільш щільний взаємозв'язок якої $r_1 = 0{,}923439$.

Статистичні ваги перших канонічних змінних знаходяться в межах від 0,011655 до 0, 923439.

Мінімальне значення статистичної ваги відповідає Z_{y_3} ($\beta_3 = 0{,}011655$), тому фактор y_3 (*Індекс економічної свободи*) дає найменший внесок у пояснення варіації результатних ознак другої групи. Для підвищення якості моделі змінну y_3 можна спробувати виключити (рис. 5.32).

Variable	Canonical Weights, left set (canonical) Root 1	Root 2
Human Development Index	-0,095591	-1,89978
Ranking of Happiness	-0,917421	1,66632

Variable	Canonical Weights, right set (canonical) Root 1	Root 2
Gender Development Index	-0,158172	0,01352
Education quality	-0,201747	1,02370
Health care quality	-0,064201	0,15520
Standard of living	-0,759218	0,21744
Feeling safe	-0,038610	-1,13571
Ideal job	0,009480	-0,92655
Health expenditure	-0,108017	0,40625

Рис. 5.32. Таблиці канонічних ваг після виключення змінної y_3

Оскільки модель не покращилась (факторні навантаження для кожної з груп не збільшились), виключення змінної y_3 є недоцільним.

У загальному випадку модель вважають якісною, коли канонічний коефіцієнт кореляції є значущим, факторні ваги всіх канонічних змінних додатні та достатньо великі.

Оскільки для r_1 $p < 0{,}1$, а для r_2, r_3 $p > 0{,}1$ (рис. 5.25), можна зробити висновок про те, що для першого канонічного коефіцієнта кореляції нульову гіпотезу H_0: $r_1 = 0$ відхиляємо практично із 100% надійністю. Для другого і третього канонічних коренів нульову гіпотезу H_0: $r_1 = 0$ приймаємо – вони статистично незначущі.

Перший канонічний коефіцієнт кореляції $r_1 = 0{,}927676$ значущий. Відповідні йому канонічні змінні є значущими:

$$Z_y = 0{,}095846\, y_1 + 0{,}923439\, y_2 + 0{,}011655\, y_3;$$
$$Z_x = 0{,}153288\, x_1 + 0{,}204511\, x_2 + 0{,}060806\, x_3 + 0{,}755556\, x_4 + $$
$$+ 0{,}037711\, x_5 + 0{,}013735\, x_6 + 0{,}116521\, x_7.$$

Це латентні показники, якісне тлумачення яких здійснюють аналогічно до пояснення головних компонент або загальних факторів.

Для графічного відображення канонічних значень побудуємо діаграму розсіювання (рис. 5.33).

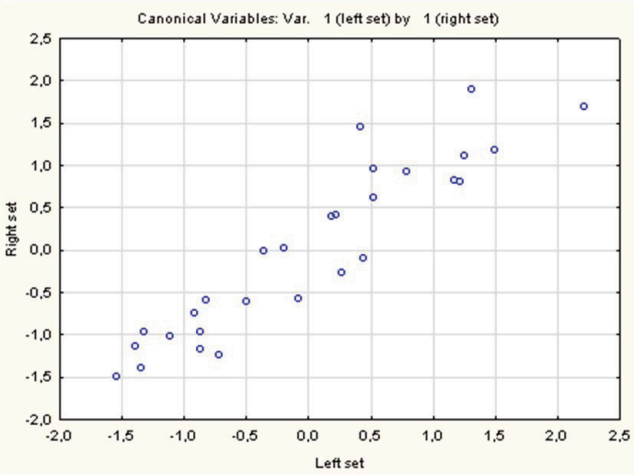


Рис. 5.33. Діаграма розсіювання канонічних змінних

На отриманому графіку немає різко виражених викидів. Крім того, відхилення від регресійної лінії не утворюють будь-яких характерних обрисів (наприклад, розташовуючись у вигляді *U* або *S* навколо лінії регресії).

Можна зробити висновок, що ніяких помітних порушень основних припущень канонічного аналізу не спостерігається.

За допомогою цього графіка можна наочно перевірити розбиття спостережень на групи. Такі кластери (групи) можуть виникати, якщо аналізована вибірка неоднорідна за природою.

5.5. Аналіз відповідності

Аналіз відповідності (correspondence analysis) – це розвідувальний метод аналізу, призначений для візуального та чисельного дослідження структури таблиць зв'язаності великої розмірності. В обчислювальному плані метою аналізу відповідності є представлення відстаней між точками у просторі меншої розмірності [13, с. 561].

Сьогодні аналіз відповідності інтенсивно застосовують у різних сферах, зокрема в соціології, політиці, економіці, маркетингу,

медицині, управлінні. Відомі застосування методу в археології, аналізі текстів, де важливим є аналіз структур даних.

Приклади застосування методу аналізу відповідності:

► дослідження соціальних груп населення в різних регіонах за статтями витрат за кожною групою;

► дослідження результатів голосування в ООН з принципових питань (1 – за, 0 – проти, 0,5 – утрималось). У 1967 р. досліджували 127 країн за 13 важливими питаннями. Результати аналізу показали, що за першим чинником країни чітко поділились на дві групи: одна – з центром США, друга – з центром СРСР (двополюсна модель світу). Інші чинники можна інтерпретувати як ізоляціонізм, неучасть у голосуванні і т. ін.;

► дослідження імпорту товарів (назва товару – рядок таблиці, країна-виробник – стовпець);

► дослідження текстів: для виявлення анонімного автора скандальної книги про одну з президентських кампаній журнал «New-Yorker» запропонував експертам-лінгвістам проаналізувати тексти 15 можливих авторів і текст анонімного видання. Тексти були представлені рядками таблиці зв'язності: у рядку i вказували частоту j появи i-го слова. Методом аналізу відповідності був визначений найбільш вірогідний автор скандального тексту.

Застосування аналізу відповідності у МВ пов'язане з дослідженням структури складних таблиць, що містять індикаторні змінні, які показують наявність або відсутність відповідної ознаки. Такі таблиці мають велику розмірність і дослідження їх структури є нетривіальним завданням.

Завдання візуалізації складних об'єктів можуть бути вирішені за допомогою аналізу відповідності. Зображення – це багатовимірна таблиця. Завдання полягає в знаходженні площини, за допомогою якої можна максимально точно відтворити початкове зображення.

Математичне обґрунтування методу

Аналіз відповідності використовує специфічну інтерпретацію статистики χ^2 за Пірсоном.

Метод призначений для дослідження таблиць зв'язності. Критерієм якості відтворення багатовимірної таблиці у просторі меншої розмірності є значення статистики χ^2. Неформально аналіз відповідності можна розглядати як факторний аналіз категоріальних даних або як метод скорочення розмірності.

Рядки або стовпці початкової таблиці представляють точками простору, між якими обчислюють відстань χ^2 (аналогічно до того, як обчислюють статистику χ^2 для порівняння спостережуваних та очікуваних частот).

Далі шукають простір невеликої розмірності (переважно двовимірний), в якому обчислені відстані спотворюються мінімально і в цьому сенсі максимально точно відтворюють структуру початкової таблиці зі збереженням зв'язків між ознаками.

Вихідною інформацією для застосування методу аналізу відповідності є таблиці спряженості (зв'язаності) (рис. 5.34).

	GNI low	GNI lower-middle	GNI middle	GNI high
HPI very high	0	10	2	1
HPI high	0	13	4	2
HPI upper middle	1	8	4	9
HPI middle	2	13	6	3
HPI lower-middle	7	13	4	3
HPI low	5	6	2	2
HPI very low	6	8	1	2

Рис. 5.34. Таблиця спряженості між аналізованими ознаки

Так, 4 перших числа кожного рядка таблиці (маргінальні частоти) можна розглядати як координати рядка в 4-вимірному просторі. Формально можна обчислити відстані χ^2 між цими точками (рядками таблиці).

При таких маргінальних частотах можна відобразити ці точки в просторі розмірності 3 (кількість ступенів свободи дорівнює 3).

Очевидно, що чим менша відстань, тим більша схожість між групами, і, навпаки, чим більша відстань, тим більша відмінність.

Припустимо, що можна знайти простір меншої розмірності (наприклад, розмірності 2) для представлення точок-рядків, що зберігає майже всю інформацію про відмінності між рядками.

Ефективність такого підходу не є наглядною для таблиць невеликої розмірності, проте він корисний для великих таблиць, що виникають, зокрема, в маркетингових дослідженнях.

Наприклад, застосуємо цей період при голосуванні за 17 кандидатів при аналізі виборчих кампаній за інформацією про виділені переваги 100 респондентами. У результаті застосування аналізу відповідності можна представити 17 претендентів (точок) на площині. Аналізуючи розташування точок, можна виявити закономірності при голосуванні за конкретного кандидата, які будуть корисними при проведенні передвиборчої кампанії.

Аналіз відповідності у середовищі «Statistica»

Приклад 5.6. Провести аналіз взаємозв'язку між *Індексом щасливої планети* (*Happy Planet Index, HPI*) та *валовим національним доходом* (*Gross national income, GNI*).

Країни світу було розділено за обраними для дослідження показниками на групи (табл. 5.2).

Термінологія, яку використовують в аналізі відповідності

Mass (Маса). Спостереження в таблиці нормують: обчислюють відносні частоти для таблиці. Сума всіх елементів таблиці має дорівнювати 1 (кожен елемент ділять на загальну кількість спостережень). Отримана стандартизована таблиця показує, як розподілена маса за комірками таблиці або за точками простору. Суми за рядками і стовпцями у матриці відносних частот називають **масою** рядка і стовпця відповідно [13, с. 561].

Таблиця 5.2

Вихідні дані для прикладу 5.6

Happy Planet Index		Gross national income	
Ranking	**Range**	**Ranking**	**Range**
very high	>36,1	high	>30000
high	36,1–32,7	middle	5000–30000
upper-middle	32,6–28,8	lower-middle	1000–5000
middle	28,7–24,8	low	<1000
lower-middle	24,7–21		
low	20,7–16,8		
very low	<16,8		

Quality (Якість). У контексті аналізу відповідності це якість представлення певної точки-рядка в системі координат, визначеній кількістю вимірів. Якість точки визначають як відношення квадрата відстані від точки до початку координат в обраній кількості вимірів. За квадратом відстані від початку координат у просторі визначають максимальну кількість вимірів. Низька якість означає, що вибрана кількість вимірів не представляє відповідний рядок (або стовпець).

Relative inertia (Інерція). Інерцію визначають як значення χ^2 Пірсона для таблиці 2×2, поділене на загальну кількість спостережень. Відносна інерція представляє частку загальної інерції, що належить цій точці, і не залежить від вибраної користувачем розмірності. Часткове рішення може достатньо добре представляти точку (висока якість), але та сама точка може здійснювати дуже малий внесок у загальну інерцію (точка-рядок, елементами якого є відносні частоти, має схожість з деяким рядком, елементи якого середні за всіма рядками).

Row & column profile (Інерція та профілі рядків і стовпців). Якщо рядки і стовпці таблиці повністю незалежні (наприклад, *Індекс щасливої планети* не залежить від в*алового національного доходу*), то елементи таблиці можуть бути відтворені за допомогою сум за рядками і стовпцями або, як називають за термінологією аналізу відповідності, за допомогою **профілів** рядків і стовпців.

Relative Dim. *n* (Відносна інерція для кожної розмірності). Це відносний внесок певної точки-рядка у величину інерції, визначений відповідною розмірністю.

Cosine' (Косинус квадрат) – якість, або квадратичні кореляції з кожною розмірністю. Цей стовпець містить якість для кожної точки, визначену відповідною розмірністю. Косинус квадрат можна інтерпретувати як «кореляцію» між відповідними точкою та розмірністю. Ця величина є квадратом косинуса кута, утвореного цією точкою і відповідною віссю.

Відповідно до формули обчислення χ^2 для таблиць спряженості 2×2 очікувані частоти таблиці, в якій стовпці і рядки незалежні, обчислюють перемножуванням відповідних профілів стовпців і рядків з діленням отриманого результату на загальну суму. Відхилення від очікуваних величин (при гіпотезі про повну незалежність змінних за рядками і стовпцями) впливає на значення статистики χ^2.

Метрика координатної системи. У певних випадках термін «відстань» використовують для позначення відмінностей між рядками та стовпцями матриці відносних частот, які відповідно представляють у просторі меншої розмірності в результаті використання методів аналізу відповідності. Насправді відстані, представлені у вигляді координат у просторі відповідної розмірності, – це не лише відстані Евкліда, обчислені за відносними частотами стовпців і рядків, а деякі зважені відстані.

Графічний аналіз результатів – найважливіша частина аналізу. Так, горизонтальна вісь відповідає максимальній інерції. На графіку вказується відсоток загальної інерції, що пояснюється конкретним власним значенням. Перетин двох осей – центр тяжіння спостережуваних точок, що відповідає середнім профілям. Якщо точки належать одному і тому самому типу (є або рядками, або стовпцями), то чим менша відстань між ними, тим щільніший зв'язок. Для того, щоб встановити зв'язок між точками різного типу (між рядками і стовпцями), потрібно розглянути кути між ними з вершиною в центрі тяжіння.

Загальне правило візуального оцінювання міри залежності:

– 2 довільних точки різного типу потрібно з'єднати відрізками прямих з центром тяжіння (точка з координатами 0; 0);

– якщо утворений кут гострий, рядок і стовпець додатно корельовані;

– якщо кут тупий, кореляція між змінними зворотна;

– якщо кут прямий, кореляція відсутня.

Результати аналізу відповідності

Аналіз відповідності можна розглядати як розкладання статистики χ^2 на компоненти з метою визначення простору найменшої розмірності, що дає можливість представити відхилення від очікуваних величин [95].

Обчислені власні значення (Eigenvalues) надають відомості про кількість вимірів, достатню для якісного представлення інформації таблиці даних. Перший вимір «витягує» 60% від загальної інерції. Додавання другого виміру збільшує «пояснену» інерцію до 95,8% (рис. 5.35).

Eigenvalues and Inertia for all Dimensions (corespond
Input Table (Rows x Columns): 7 x 4
Total Inertia=,24766 Chi?=33,930 df=18 p=,01288

Number of Dims.	Singular Values	Eigen-Values	Perc. of Inertia	Cumulatv Percent	Chi Squares
1	0,385338	0,148485	59,95433	59,9543	20,34247
2	0,297937	0,088766	35,84150	95,7958	12,16100
3	0,102040	0,010412	4,20417	100,0000	1,42647

Рис. 5.35. Власні значення та інерція для всіх вимірів

Найбільші відхилення між спостережуваними та очікуваними частотами виявлено у групах країн, в яких зафіксовано рівень *HPI,* вищий від середнього, і високий рівень *GNI* та країн з дуже низьким *HPI* і низьким *GNI*. А реальна кількість країн з рівнем *Happy Planet Index,* вищим від середнього, та високим рівнем *валового національного доходу* на 5 більша, ніж очікуваних. Кількість країн, в яких зафіксовано дуже низький рівень *Happy Planet Index* та низький рівень *валового національного доходу* на 3 більша, ніж за гіпотези про незалежність (рис. 5.36).

Критерій χ^2 Пірсона є об'єктивною оцінкою близькості емпіричних розподілів до теоретичних. Досягнутий рівень значущості 0,01 – результати є статистично значущими. Кількість ступенів вільності $df = 18$, $\chi^2 = 33{,}9$. $\chi^2(0{,}01;18) = 34{,}8$. Розраховане значення критерію більше за табличне. Отже, можна стверджувати, що передбачені значення є достатньо близькими до спостережених (рис. 5.37).

	Observed minus Expected Frequencies (corresponc Input Table (Rows x Columns): 7 x 4 Total Inertia=,24766 Chi?=33,930 df=18 p=,01288			
	GNI low	GNI lower-middle	GNI middle	**GNI high**
HPI very high	-1,99270	3,26277	-0,18248	-1,08759
HPI high	-2,91241	3,15328	0,81022	-1,05109
HPI upper middle	-2,37226	-3,40146	0,30657	5,46715
HPI middle	-1,67883	0,56204	1,97080	-0,85401
HPI lower-middle	2,86131	-0,99270	-0,53285	-1,33577
HPI low	2,70073	-1,77372	-0,51825	-0,40876
HPI very low	3,39416	-0,81022	-1,85401	-0,72993

Рис. 5.36. Таблиця відхилень очікуваних частот від спостережених

	Contributions to Chi-Square (corespondent_HPI_GNI_1) Input Table (Rows x Columns): 7 x 4 Total Inertia=,24766 Chi?=33,930 df=18 p=,01288				
	GNI low	GNI lower-middle	GNI middle	GNI high	Total
HPI very high	1,99270	1,580130	0,015258	0,56661	4,15470
HPI high	2,91241	1,009799	0,205799	0,36210	4,49011
HPI upper middle	1,66880	1,014776	0,025446	8,46053	11,16955
HPI middle	0,76613	0,025398	0,963980	0,18924	1,94475
HPI lower-middle	1,97819	0,070426	0,062637	0,41152	2,52278
HPI low	3,17229	0,404709	0,106654	0,06937	3,75301
HPI very low	4,42097	0,074511	1,204398	0,19517	5,89504
Total	16,91149	4,179748	2,584173	10,25454	33,92995

Рис. 5.37. Таблиця внесків комірок у χ^2

Мета аналізу відповідності полягає в тому, щоб підсумувати відхилення від очікуваних частот не в абсолютних, а у відносних одиницях.

Аналіз рядків і стовпців. Рядки та стовпці таблиці даних можна представити точками у просторі меншої розмірності, який максимально точно відтворює схожість (і відстані) між відносними частотами для стовпців/рядків таблиці.

Результати графічного аналізу відповідності

Аналіз відповідності передбачає знаходження простору меншої розмірності, який повністю представляє дані таблиці (рис. 5.38–5.40). При цьому критерієм якості є нормований χ^2, або інерція. Використання у дослідженні одновимірного простору дало можливість пояснити 74,39% інерції таблиці (рис. 5.38).

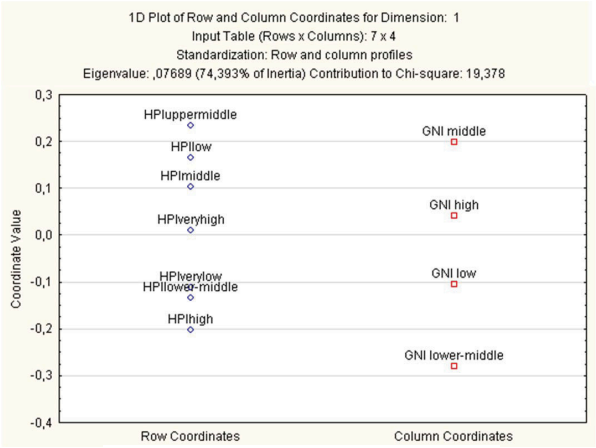


Рис. 5.38. 1-D графік координат рядків і стовпців

Дві розмірності також дають можливість пояснити 74,39% інерції. Очевидною перевагою двовимірного простору є те, що рядки, які відображаються у вигляді близьких точок, близькі один до одного і за відносними частотами. Аналізуючи розташування точок за першою віссю, можна помітити, що країни з дуже низьким рівнем *HPI* та країни з рівнем *HPI,* нижчим від середнього, відносно близькі за координатами (рис. 5.39 – 5.40).

Таблиця відносних частот (частоти стандартизовані так, що їх сума за кожним рядком дорівнює 100%) унаочнює схожість даних двох груп за рівнями *Happy Planet Index* та *Gross national income* (рис. 5.41–5.42).

Остаточною метою аналізу відповідності є інтерпретація векторів в отриманому просторі нижчої розмірності. Одним із способів, який може допомогти в інтерпретації отриманих результатів, є представлення на діаграмі стовпців. Можна вважати, що перша вісь дає градацію рівня *Happy Planet Index*. Отже, велику міру схожості між країнами з дуже низьким рівнем *HPI* та країнами з рівнем *HPI,* нижсим від середнього, можна пояснити наявністю в цих групах великої кількості країн з низьким рівнем *Gross national income* (рис. 5.42).

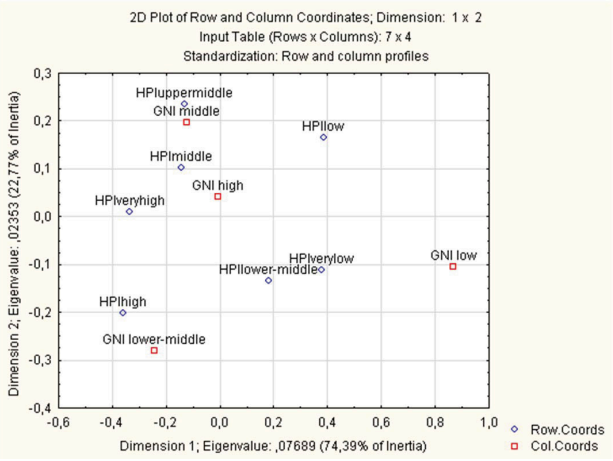


Рис. 5.39. 2-D графік координат рядків і стовпців

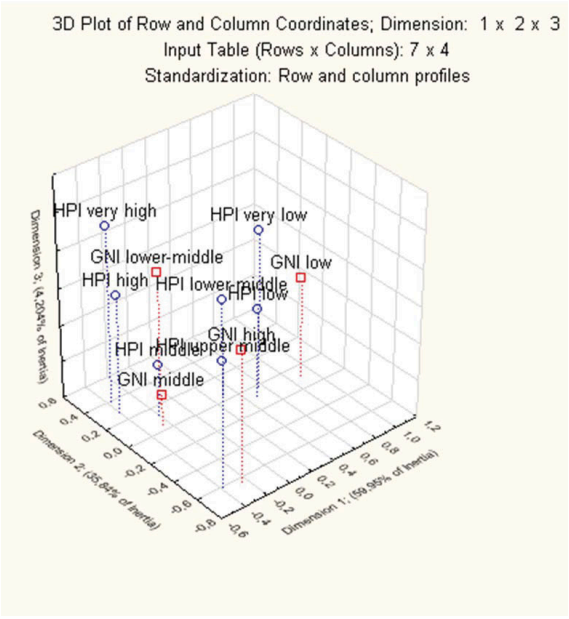


Рис. 5.40. 3-D графік координат рядків і стовпців

	Percentages of Row Totals (corespondent_HPI_GNI_all) Input Table (Rows x Columns): 7 x 4 Total Inertia=,24766 Chi?=33,930 df=18 p=,01288				
	GNI low	GNI lower-middle	GNI middle	GNI high	**Total**
HPI very high	0,00000	76,92308	15,38462	7,69231	100,0000
HPI high	0,00000	68,42105	21,05263	10,52632	100,0000
HPI upper middle	4,54545	36,36364	18,18182	40,90909	100,0000
HPI middle	8,33333	54,16667	25,00000	12,50000	100,0000
HPI lower-middle	25,92593	48,14815	14,81481	11,11111	100,0000
HPI low	33,33333	40,00000	13,33333	13,33333	100,0000
HPI very low	35,29412	47,05882	5,88235	11,76471	100,0000

Рис. 5.41. Таблиця відносних частот за рядками

	Percentages of Column Totals (corespondent_HPI_ Input Table (Rows x Columns): 7 x 4 Total Inertia=,24766 Chi?=33,930 df=18 p=,01288			
	GNI low	GNI lower-middle	GNI middle	GNI high
HPI very high	0,0000	14,0845	8,6957	4,5455
HPI high	0,0000	18,3099	17,3913	9,0909
HPI upper middle	4,7619	11,2676	17,3913	40,9091
HPI middle	9,5238	18,3099	26,0870	13,6364
HPI lower-middle	33,3333	18,3099	17,3913	13,6364
HPI low	23,8095	8,4507	8,6957	9,0909
HPI very low	28,5714	11,2676	4,3478	9,0909

Рис. 5.42. Таблиця відносних частот за стовпцями

Для оцінювання якості отриманого рішення використовують спеціальні статистики. Всі або більшість точок мають бути правильно представлені – відстані між ними в результаті застосування процедури аналізу відповідності не мають спотворюватися.

Результати обчислення статистик за наявними координатами рядків та стовпців для двовимірного рішення представлені на рис. 5.43–5.44.

Row Name	Row Number	Coordin. Dim.1	Coordin. Dim.2	Mass	Quality	Relative Inertia	Inertia Dim.1	Cosine? Dim.1	Inertia Dim.2	Cosine? Dim.2
	Row Coordinates and Contributions to Inertia (corespondent_HPI_GNI_all) Input Table (Rows x Columns): 7 x 4 Standardization: Row and column profiles									
HPI very high	1	-0,360316	0,398786	0,094891	0,903832	0,122449	0,082967	0,406230	0,170002	0,497603
HPI high	2	-0,392589	0,286045	0,138686	0,998420	0,132335	0,143955	0,652190	0,127836	0,346230
HPI upper middle	3	-0,377065	-0,603259	0,160584	0,996833	0,329195	0,153762	0,280039	0,658356	0,716794
HPI middle	4	-0,199316	0,123866	0,175182	0,679608	0,057317	0,046870	0,490264	0,030279	0,189344
HPI lower-middle	5	0,300355	0,044983	0,197080	0,987162	0,074353	0,119738	0,965506	0,004493	0,021657
HPI low	6	0,489533	-0,083797	0,109489	0,985866	0,110611	0,176706	0,957800	0,008661	0,028065
HPI very low	7	0,574690	-0,016338	0,124088	0,953191	0,173742	0,276002	0,952421	0,000373	0,000770

Рис. 5.43. Координати і внесок рядків в інерцію

Column Name	Column Number	Coordin. Dim.1	Coordin. Dim.2	Mass	Quality	Relative Inertia	Inertia Dim.1	Cosine? Dim.1	Inertia Dim.2	Cosine? Dim.2
Column Coordinates and Contributions to Inertia (corespondent_HPI_GNI_all) Input Table (Rows x Columns): 7 x 4 Standardization: Row and column profiles										
GNI low	1	0,892549	-0,089130	0,153285	0,999106	0,498424	0,822394	0,989241	0,013718	0,009865
GNI lower-middle	2	-0,105089	0,209985	0,518248	0,936599	0,123188	0,038545	0,187594	0,257434	0,749005
GNI middle	3	-0,252739	0,039097	0,167883	0,582131	0,076162	0,072222	0,568526	0,002891	0,013605
GNI high	4	-0,248603	-0,633474	0,160584	0,993515	0,302227	0,066839	0,132592	0,725957	0,860923

Рис. 5.44. Координати і внесок стовпців в інерцію

Якість висока для всіх груп країн, окрім країн з середнім *Happy Planet Index*, для яких якість є достатньою. Отже, обрана кількість вимірів добре представляє всі рядки та стовпці початкової таблиці даних.

Графічний аналіз результатів дослідження зв'язку між рівнями *Happy Planet Index* та *Gross national income* (рис. 5.45) підтвердив, зокрема, наявність прямого зв'язку між групами країн з дуже низьким рівнем *HPI* та низьким рівнем *Gross national income*; країн з рівнем *HPI,* нижчим від середнього, і країн з низьким рівнем *Gross national income*. Між країнами з високим рівнем *HPI* та високим рівнем *Gross national income* виявлено зворотний зв'язок, що дає підстави стверджувати, що існує рівень доходу, при якому рівень щастя не збільшується зі зростанням кількості грошей.

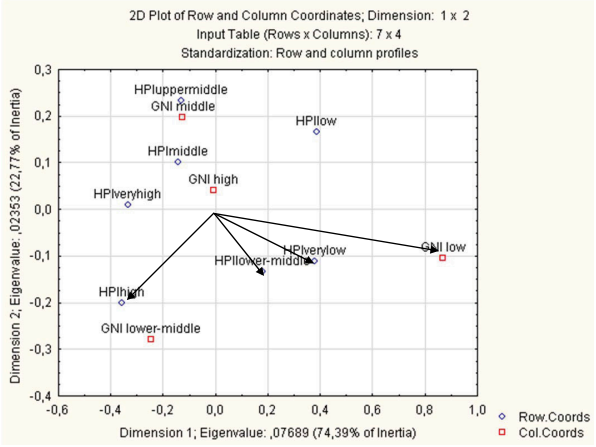


Рис. 5.45. Візуальне оцінювання залежності між точками різного типу

5.6. Аналіз виживання

Аналіз виживання (Survival analysis) – статистичний аналіз, розроблений для вивчення, оцінювання та порівняння часу, що минув до настання деякої події (припинення діяльності організації, розпад міжнародного об'єднання) [13, с. 561]. Методи аналізу виживання були розроблені для медичних, біологічних досліджень і страхування, але згодом почали широко застосовуватися в соціальних і економічних науках. Наприклад, для оцінювання терміну придатності зброї чи обмундирування, порівняння термінів функціонування фірми у конкретній галузі залежно від вибраного фактора.

Основні поняття аналізу виживання

Нехай t (невід'ємна випадкова величина) – час очікування до настання деякої події. Цю подію називають «смерть», а час її очікування – «час виживання».

Виживання (**Survive**) $S(t)$ – ймовірність «прожити» час більше t з моменту початку спостереження.

Це ймовірність ненастання події до визначеного моменту часу:

$$S(t) = P(T > t) = 1 - P(T < t),$$

де $P(T > t)$ – ймовірність «пережитого» часу t.

Події можуть бути різними, не тільки загибель об'єкта.

Властивості функції $S(t)$:

– $S(t) = 1$, якщо $t = 0$: на початку дослідження очікувана подія не відбулося в жодного зі спостережуваних. Імовірність «дожити» до цього моменту дорівнює 1;

– $S(t) = 0$, якщо $t = \infty$: в кінці дослідження подія відбулася у всіх спостережуваних. Імовірність «дожити» до цього моменту становить 0.

Графік функції виживання $S(t)$ – крива виживання, що відображає ймовірність пережити будь-який з моментів часу t. Він описує залежність імовірності «дожити» від часу. Час може бути виміряний у будь-яких відомих одиницях (дні, місяці і т. ін).

Якщо графік функції виживання гострий, то виживаність низька – очікувана подія настала швидко у всіх випробовуваних. Якщо графік пологий, виживаність вважається високою – має минути значна кількість часу, щоб очікувана подія настала у всіх випробовуваних (рис. 5.46).

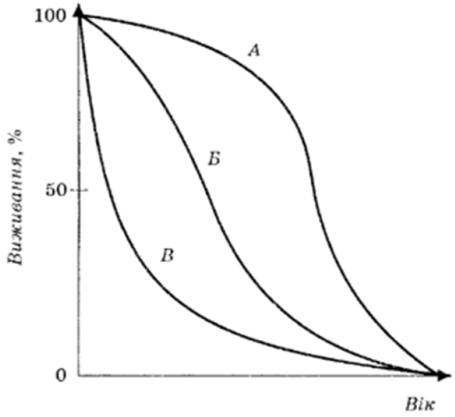


Рис. 5.46. Три типи кривих виживання

Криву виживання використовують, зокрема, для визначення медіани виживання та інших процентилів часу життя. Час, до якого доживе половина випробовуваних, називають **медіаною виживання**. Якщо подія не настала у половини випробовуваних взагалі, то медіану визначити неможливо. У таких випадках визначають час, до якого «дожили» (подія не настала) три чверті всіх випробовуваних (75%). При порівнянні двох і більше кривих за медіаною можна оцінити виживаність у різних групах.

Типи даних в аналізі виживання

Методи аналізу виживання дають можливість вивчати неповні (цензуровані) дані. Такі спостереження є типовими, коли досліджувана величина представляє час до настання деякої критичної події, а тривалість спостереження обмежена за часом. Наприклад, при вивченні «виживання» нових підприємств або часу «життя» продуктів автомобільного виробництва неможливо дочекатися

того моменту, коли всі досліджувані підприємства перестануть працювати, а автомобілі остаточно вийдуть з ладу.

Спостереження, які містять неповну інформацію, називають **цензурованими** (censored) **спостереженнями**. Дослідження називають **повним** (complete), якщо всі дані нецензуровані.

Типи цензурованих даних:

► після завершення часу експерименту об'єкти дослідження «знищуються», їм присвоюється тривалість «життя», що дорівнює тривалості експерименту;

► після досягнення появи події у необхідної кількості об'єктів (наприклад, у половини) інші об'єкти дослідження «знищуються»;

► об'єкт може вийти з поля зору дослідника, тоді тривалість його життя обчислюється як тривалість його участі в експерименті: від його появи в експерименті до зникнення;

► для об'єктів, у яких подія так і не настала після завершення дослідження, час «життя» обчислюється від їх появи в експерименті до завершення дослідження.

Можна використовувати тільки повні часи життя, але в такому разі для дослідження залишилось би дуже мало спостережень і отримані оцінки були б неточними. Використання поряд з повними спостереженнями неповних або цензурованих є головною особливістю методів аналізу виживання [13].

Таблиці часу життя

Якщо об'єкти в процесі дослідження не вибувають (всі дані повні), то оцінка виживання $S(t)$ визначається як відношення кількості тих, що пережили момент t, до обсягу вибірки n (за класичною формулою ймовірності).

Для побудови кривої виживання будують таблиці часу життя (дожиття).

Основними способами побудови таблиць часу життя є методи Катлера–Едерера і Каплана–Мейєра. Метод Катлера–Едерер застосовують для великих наборів даних. У цьому разі час розбивають на

інтервали. Метод Каплана–Мейєра застосовують для малих вибірок. У цьому разі фіксують моменти, в які відбулася хоча б одна подія.

Метод Каплана–Мейєра

Часовий проміжок експерименту розбивають на «моменти» (інтервали) таким чином, щоб час кожної події «смерть» або «цензурування» належав за можливості до різних інтервалів. На кожному з них обчислюють частку тих, що «пережили» такий момент:

$$q_i = 1 - \frac{d_{t_i}}{n_{t_i}},$$

де – кількість «померлих» на момент t_i, n_{t_i} – кількість спостережуваних до моменту t_i.

Кумулятивне виживання на досліджуваному моменті є добутком частки кількості тих, що «пережили» цей момент, і «виживання», отриманого на попередньому моменті:

$$S(t_i) = \prod_{j=1}^{i} \left(1 - \frac{d_{t_j}}{n_{t_j}}\right).$$

Результати розрахунків представляють у вигляді таблиць і графіків.

Приклад 5.7. Проводилось дослідження виходу на аграрний ринок 10 нових підприємств. Отримано такі дані: 8 фірм пропрацювали відповідно по 3, 5, 7 (2 фірми), 8, 11, 12 (2 фірми) місяців з початку дослідження. Окрім того, одне підприємство через 6 місяців змінило профіль роботи й одне увійшло до складу ТНК через 9 місяців після початку дослідження (ці дані цензуровані).

Очікувана подія – *підприємство припинило свою діяльність*. Одиниця часу – *1 місяць*. Моменти: *3, 5, 7, 8, 11, 12*.

Отже, часи «життя»: 3, 5, 6+, 7, 7, 8, 9+, 11, 12, 12.

Побудова таблиці часу життя за методом Каплана–Мейєра (табл. 5.3)

Таблиця 5.3

Таблиця часу життя за методом Каплана–Мейєра

Момент часу t_i	Кількість спостережуваних до моменту t n_{t_i}	Кількість «померлих» до моменту t d_{t_i}	Частка тих, що пережили момент t_i $q_i = 1 - \frac{d_{t_i}}{n_{t_i}}$	Виживання $S(t)$
3	10	1	1 – 0,1 = 0,9	0,9
5	9	1	1 – 0,111 = 0,89	0,9×0,89 = 0,8
7	7	2	1 – 0,28 = 0,72	0,9×0,8×0,72=0,57
8	5	1	1 – 0,2 = 0,8	0,9×0,89×0,72×0,8=0,46
11	3	1	1 – 0,33 = 0,67	0,9×0,89×0,72×0,8×0,67=0,3
12	2	2	1 – 1 = 0	0

Результати, отримані в таблиці, представляють у вигляді графіка (кривої виживання) (рис. 5.47).

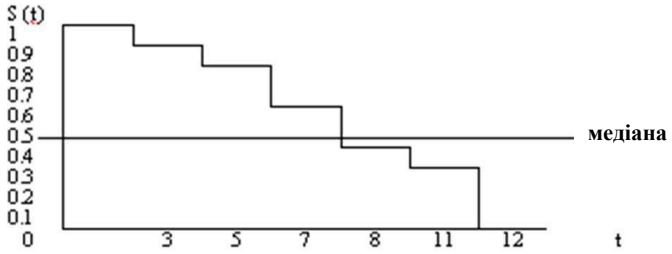


Рис. 5.47. Крива виживання

Порівняння кривих виживання (табл. 5.4).

Методи аналізу виживання використовують переважно до тих самих статистичних завдань, що й інші методи. Однак їх застосовують до цензурованих (неповних) даних. Як функцію розподілу в методах аналізу виживання використовують функцію виживання, яка є ймовірністю того, що об’єкт проживе час, що становить більше t.

Таблиця 5.4

Побудова таблиці часу життя за методом Катлера–Едерера

Момент часу t_i		К-сть спостережуваних до моменту t n_{t_i}	Кількість «померлих» до моменту t d_{t_i}	Частка тих, що пережили момент t_i $q_i = 1 - \frac{d_{t_i}}{n_{t_i}}$	Виживання S(t)
Cencored					
3	+	10	1	1 – 0,1 = 0,9	0,9
5	+	9	1	1 – 0,111 = 0,89	0,9×0,89 = 0,8
6	–	8			
7	+	7	2	1 – 0,28 = 0,72	0,9×0,8×0,72=0,57
8	+	5	1	1 – 0,2 = 0,8	0,9×0,89×0,72×0,8=0,46
9	–	4			
11	+	3	1	1 – 0,33 = 0,67	0,9×0,89×0,72×0,80,67= 0,3
12	+	2	2	1 – 1 = 0	0

Загалом цензуровані спостереження є типовими, коли досліджувана величина представляє час до настання деякої критичної події, а тривалість спостереження обмежена в часі. Цензуровані спостереження застосовують у багатьох сферах діяльності.

Наприклад, їх застосовують у соціальних науках – при вивченні інтенсивності вибуття студентів з вищого навчального закладу (час до вибуття), динаміки чисельності працівників в окремих організаціях, при дослідженні періоду володіння мандатом депутата Верховної Ради України, членства в коаліціях і т. ін.

В економіці методи аналізу виживання застосовують у страхуванні, при вивченні «виживання» нових підприємств або часу «життя» продуктів. У завданнях контролю якості типовим є вивчення «виживання» елементів виробів під навантаженням (аналіз термінів відмов). В актуарній математиці (вивчає питання, пов'язані з оцінюванням ризиків) як об'єкт дослідження використовують таблиці смертності, які містять дані про смертність за обрані інтервали часу осіб окремих категорій.

Приклад 5.8. Дослідити динаміку «виживання» в кризових умовах 30 вітчизняних банків за період з 1 січня 1990 р. по 3 листопада 2016 р. за такими вхідними даними (рис. 5.48).

	1 Name Bank	2 Month_1	3 Day_1	4 Year_1	5 Month_2	6 Day_2	7 Year_2	8 censored	9 term existence	10 ownership	11 independence
1	Інпромбанк	3	23	1993	3	7	2013	complete	20	1	1
2	Юнікон	2	8	2013	4	28	2016	censored	3	2	1
3	Приват Банк	3	19	1992	11	3	2016	censored	14	6	1
4	Фінансова ініціатива	7	4	2004	6	24	2015	censored	11	9	1
5	Державний Ощадний банк України	12	31	1991	11	3	2016	censored	25	2	1
6	А-Банк	6	3	2007	11	3	2016	censored	9	2	1
7	Індустріалбанк	7	4	1990	11	3	2016	censored	26	4	0
8	РОДОВІД Банк	1	23	1990	2	26	2016	censored	26	2	0
9	Промінвестбанк	8	26	1992	11	3	2016	censored	23	6	1
10	РАДАБАНК	7	12	1993	11	3	2016	censored	21	5	1
11	Інтеграл-банк	9	30	1994	11	26	2015	complete	21	2	1

Рис. 5.48. Фрагмент файла даних

Організація файла

Файл даних містить для кожного з досліджуваних банків таку інформацію: назву, дату створення та дату банкрутства або кінцеву дату спостереження (місяць, день та рік в окремих змінних), термін існування банку до «смерті» чи завершення терміну спостереження (*term existence*), тип власності (*ownership*): 1– акціонерне товариство, 2 – ПАТ, 3 – ПАТ «Комерійний банк» і т. ін. У змінній *independence* вказано період заснування банку: 1 – після проголошення незалежності України, 0 – до проголошення. Змінна *censored* є індикатором цензурованих спостережень (значення *complete* встановлюється лише для банків, про які точно відомо, що вони припинили своє існування – «померли»).

Після завершення дослідження є банки, які «вижили» протягом усього періоду спостереження, зокрема серед тих, де призначили тимчасову адміністрацію, але які поки не ліквідовані, а також серед тих, які стали філіями інших банків до завершення експерименту. Недоцільно було б втрачати зібрану про них інформацію, оскільки більшість цих банків «вижили» протягом часу спостереження. Такі спостереження використовують у дослідженні як цензуровані.

Найбільш природним способом опису виживання у вибірці є побудова таблиць і розподілів часу життя (Life tables & Distributions). Техніка таблиць часу життя – один з найстаріших методів аналізу даних про «виживання» (часу відмов).

Таблицю виживання (рис. 5.49) можна розглядати як «розширену» таблицю частот. Однак звичайну таблицю частот будують за повними спостереженнями. У таблиці «життя» враховані як повні, так і неповні спостереження. Ідея таблиць «життя» або «виживання» полягає в тому, щоб обчислити найпростіші статистики та описати час «виживання» об'єктів [95].

Interval	Interval Start	Mid Point	Interval Width	Number Entering	Number Withdrwn	Number Exposed	Number Dying	Proportn Dead	Proportn Survivng	Cum.Prop Survivng	Problty Density
Intno.1	0,000	437,227	874,4545	30	0	30,00000	0	0,016667	0,983333	1,000000	0,000019
Intno.2	874,455	1311,682	874,4545	30	1	29,50000	0	0,016949	0,983051	0,983333	0,000019
Intno.3	1748,909	2186,136	874,4545	29	0	29,00000	0	0,017241	0,982759	0,966667	0,000019
Intno.4	2623,364	3060,591	874,4545	29	2	28,00000	0	0,017857	0,982143	0,950000	0,000019
Intno.5	3497,818	3935,045	874,4545	27	1	26,50000	2	0,075472	0,924528	0,933036	0,000081
Intno.6	4372,273	4809,500	874,4545	24	0	24,00000	1	0,041667	0,958333	0,862618	0,000041
Intno.7	5246,727	5683,955	874,4545	23	0	23,00000	2	0,086957	0,913044	0,826676	0,000082
Intno.8	6121,182	6558,409	874,4545	21	1	20,50000	0	0,024390	0,975610	0,754791	0,000021
Intno.9	6995,636	7432,864	874,4545	20	0	20,00000	4	0,200000	0,800000	0,736381	0,000168
Intno.10	7870,091	8307,318	874,4545	16	3	14,50000	5	0,344828	0,655172	0,589105	0,000232
Intno.11	8744,546	9181,772	874,4545	8	6	5,00000	1	0,200000	0,800000	0,385965	0,000088
Intno.12	9619,000			1	1	0,50000	0	1,000000	0,000000	0,308772	

Рис. 5.49. Фрагмент таблиці виживання

Для цього часову вісь – область можливих часів настання критичних подій («смертей», відмов і т. ін.) – розбивають на визначену кількість інтервалів. У системі «Statistica» кількість інтервалів часової осі за замовчуванням дорівнює 12. Для кожного інтервалу обчислюють кількість і частку об'єктів, які на початку розглянутого інтервалу були «живими», кількість і частку об'єктів, які «померли» на цьому інтервалі, а також кількість і частку об'єктів, які були вилучені (withdrawn) або цензуровані в кожному інтервалі (рис. 5.48).

При аналізі таблиць виживання потрібно враховувати, що на конкретному часовому інтервалі спостереження може бути або цензурованим (у банку введена тимчасова адміністрація), або спостерігається фатальний результат (банк припинив свою діяльність).

У стовпці **Number Entering** вказано кількість об'єктів, які були «живими» на початку досліджуваного часового інтервалу.

Стовпець **Number Withdrwn** містить кількість цензурованих на кожному інтервалі об'єктів (вилучених із спостереження, позначка censored).

У стовпці **Number Exposed** (кількість досліджуваних) виведено кількість об'єктів, які були «живими» на початку досліджуваного часового інтервалу, мінус половина від кількості вилучених.

Стовпець **Number Dying** містить кількість об'єктів, які «померли» на цьому інтервалі (позначка complete).

Proporth Dead (частка «померлих») – відношення кількості об'єктів, які «померли» у поточному інтервалі, до кількості об'єктів, досліджуваних на цьому інтервалі.

Proporth Survivng (частка об'єктів, що «вижили») – одиниця мінус частка «померлих».

Cump. Prop. Survivng – кумулятивна частка об'єктів, що «вижили», або функція виживання. Це ймовірність того, що об'єкт «переживе» поточний інтервал. Вона дорівнює добутку часток об'єктів, що «вижили», в усіх попередніх інтервалах.

Problty Density – щільність імовірності «смерті» на цьому інтервалі: від функції виживання на цьому інтервалі віднімається функція виживання на наступному інтервалі та ділиться на тривалість (довжину) інтервалу, відображену у третьому стовпці таблиці (**Interval Width**).

До даних можна підганяти основні сімейства розподілів, використовуючи звичайний метод найменших квадратів або дві модифікації методу зважених найменших квадратів. Щоб вибрати найбільш відповідне сімейство розподілів, спочатку розглянемо модель з експоненціальним розподілом. Оцінювання згоди проводиться за допомогою критерію χ-квадарт (рис. 5.50).

Якщо критерій значущий, роблять висновок, що підігнаний розподіл значно відрізняється від даних спостереження. З огляду на це відхиляють експоненціальне сімейство розподілів і роблять висновок, що воно не узгоджується з даними.

	Parameter Estimates, Model: Exponential (vuzuvanna_laba) Note: Weights: 1=1., 2=1./V, 3=N(I)*H(I)						
Estimatn Method	**Lambda**	Variance Lambda	Std.Err. Lambda	Log-Likelhd.	Chi-Sqr.	df	p
Weight 1	0,000121	0,000000	0,000034	-59,3719	33,91009	10	0,000192
Weight 2	0,000032	0,000000	0,000012	-58,4324	32,03098	10	0,000398
Weight 3	0,000088	0,000000	0,000021	-56,9557	29,07759	10	0,001215

Рис. 5.50. Оцінки для сімейства експоненціальних розподілів та значення критерію χ-квадрат

З таблиці результатів (рис. 5.50) випливає, що жоден метод підгонки не дає експоненціального розподілу задовільної згоди. Щоб переконатись у правильності отриманого результату, побудуємо графік функції виживання (рис. 5.51). На отриманих графіках жодна з експонент не апроксимує спостережувану функцію виживання адекватно. Оцінена функція виживання значно відхиляється від апроксимуючих функцій виживання. З графіка оцінювання щільності можна зробити висновок, що до 2000 р. (перших 10 років спостереження) ймовірність «смерті» для банків була мінімальною. Найвищою вона була після 2012 р. (рис. 5.51).

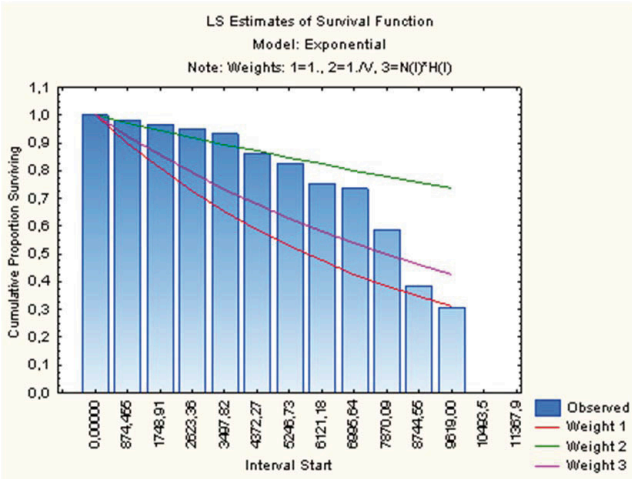


Рис. 5.51. Графік функції виживання

Однією з основних характеристик, що описують перебіг процесу в аналізі виживання, є інтенсивність відмов, або функція миттєвого ризику. Ця функція є важливою прогностичною характеристикою, яка описує «перебіг хвороби». Формально функція миттєвого ризику дорівнює ймовірності того, що об'єкт «помре» в наступному інтервалі спостереження за умови, що спочатку він був живим.

Графік функції (рис. 5.52) ризику наочно демонструє, що на початку досліджуваного періоду ризик «смерті» дуже малий, потім

він коливається і суттєво зростає наприкінці спостереження. Саме функцію ризику використовують для прогностичних цілей.

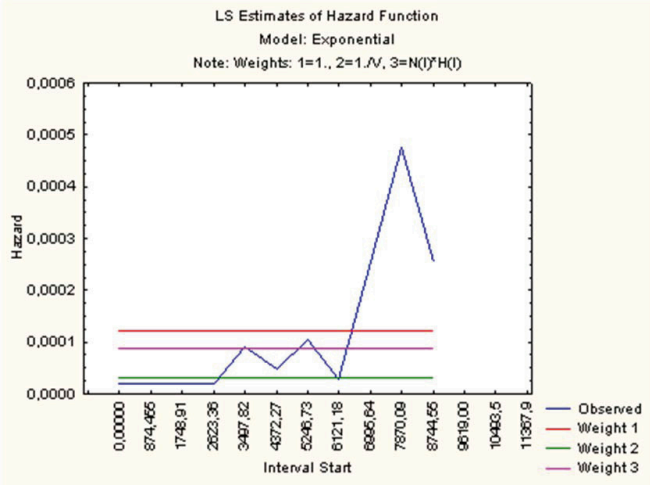


Рис. 5.52. Графік функції ризику

Проте реально в цьому разі отримуємо лише оцінку ризику. Відповідно важливою є точність отриманих оцінок. Оцінкам з великою похибкою не можна довіряти. Наприклад, непереконливими є оцінки, похибка яких має той самий порядок, що й ці оцінки. Отримані оцінки з великою похибкою (**Str. Err.**) потрібно відхилити (рис. 5.53). Це надзвичайно важливий принцип аналізу даних. Відомо, що для отримання надійних оцінок параметрів і помилок у таблицях життя потрібно, як мінімум, 30 спостережень.

Медіана очікуваного часу життя (Median Life Exp.) – це моменти часу, в яких функція виживання дорівнює $\frac{1}{2}$. Наприклад, з першого рядка таблиці випливає, що банк з імовірністю $\frac{1}{2}$ проживе більше 8253 дні після заснування (рис. 5.53).

Якщо банк «пережив» перший інтервал – 8253 дні після створення, то медіана його часу «життя» дорівнює 7415 – банк проіснує наступних 7415 днів і т. д. У загальному випадку таблиця часів життя дає уявлення

про розподіл відмов або «смертей», якщо спостережень доволі багато. Однак для прогнозу часто необхідно знати форму функції виживання. Для цієї мети використовують різні класичні розподіли.

Interval	Cum.Prop Survivng	Problty Density	Hazard Rate	Std.Err. Cum.Surv	Std.Err. Prob.Den	Std.Err. Haz.Rate	Median Life Exp	Std.Err. Life Exp
Intno.1	1,000000	0,000019	0,000019	0,000000	0,000027	0,000027	8253,660	392,9633
Intno.2	0,983333	0,000019	0,000020	0,023373	0,000027	0,000028	7415,079	389,6748
Intno.3	0,966667	0,000019	0,000020	0,032773	0,000027	0,000028	6576,497	386,3584
Intno.4	0,950000	0,000019	0,000021	0,039791	0,000027	0,000029	5737,915	386,4178
Intno.5	0,933036	0,000081	0,000090	0,045745	0,000055	0,000063	4899,973	390,1107
Intno.6	0,862618	0,000041	0,000049	0,063882	0,000040	0,000049	4177,082	378,9880
Intno.7	0,826676	0,000082	0,000104	0,070611	0,000056	0,000073	3379,988	371,0084
Intno.8	0,754791	0,000021	0,000028	0,080719	0,000029	0,000040	2720,445	944,2318
Intno.9	0,736381	0,000168	0,000254	0,082842	0,000078	0,000126	1950,264	932,6456
Intno.10	0,589105	0,000232	0,000476	0,093436	0,000092	0,000208	1748,909	0,0000
Intno.11	0,385965	0,000088	0,000254	0,095680	0,000082	0,000253	874,455	0,0000
Intno.12	0,308772			0,103083				

Рис. 5.53. Крайній правий фрагмент таблиці виживання

Найбільш важливі такі сімейства розподілів: експоненціальне, Вейбулла і Гомперца. Ці розподіли мають невідомі параметри, які оцінює програма. Процедура оцінювання параметрів заснована на методі найменших квадратів. Для проведення оцінювання може бути застосована модель лінійної регресії, оскільки всі перераховані сімейства розподілів можуть бути «зведені до лінійних» (щодо параметрів) за допомогою відповідних перетворень.

Такі перетворення в окремих випадках приводять до того, що дисперсія залишків залежить від інтервалів (дисперсія різна на різних інтервалах). Щоб врахувати це, в алгоритмах підгонки додатково використовуються оцінки зважених найменших квадратів двох типів.

Оцінювання Каплана–Мейєра

Одним із завдань в аналізі виживання є оцінювання функції виживання, тобто ймовірність того, що об'єкт «проживе» певний час після «операції».

Для цензурованих спостережень функцію виживання можна оцінити безпосередньо, не використовуючи таблицю часів життя. Такий метод вперше запропонували Каплан і Мейєр у 1958 році [13].

Для хронологічних подій застосування таке оцінювання функції виживання:

$$S(t) = \prod \left(\frac{n-j}{n-j+1} \right)^{\delta(j)},$$

де $S(t)$ – оцінка функції виживання, n – загальна кількість подій (обсяг вибірки), j – порядковий (хронологічний) номер окремої події, $\delta(j) = 1$, якщо j-та подія означає відмову («смерть») і $\delta(j) = 0$, якщо j-та подія означає втрату спостереження (індикатор цензурування), $\prod$ – добуток за всіма спостереженнями j, які завершились до моменту t.

Оцінювання Каплана–Мейєра функції виживання (рис. 5.54) побудовано за даними прикладу 5.8.

Case Number	Time	Cumulatv Survival	Standard Error
2+	1175,000		
30+	3285,000		
6+	3441,000		
25	3862,000	0,962963	0,036345
4+	4007,000		
19	4052,000	0,924444	0,051398
18	4784,000	0,885926	0,062032
29	5960,000	0,847407	0,070284
28	6015,000	0,808889	0,076923
13+	6821,000		
1	7289,000	0,768444	0,083032
26	7441,000	0,728000	0,087962
24	7652,000	0,687556	0,091904
11	7727,000	0,647111	0,094981
17	7989,000	0,606667	0,097275
23	8001,000	0,566222	0,098841
21	8090,000	0,525778	0,099713
15+	8312,000		

Рис. 5.54. Оцінки Каплана–Майєра функції виживання

З таблиці можна зробити висновок, наприклад, що ймовірність того, що банк пропрацює більше 3862 днів, дорівнює 0,96. Ймовірність того, що банк буде існувати більше 4052 днів, становить 0,92 і т. д.

У першому стовпці таблиці вказані номери спостережень, для яких у даний момент часу відбулася деяка подія (банкрутство).

Знак + означає, що об'єкт цензурований (вилучений зі спостереження). Стандартна похибка функції виживання достатньо мала (порівняно з похибками для таблиць часів життя). Побудуємо графік функції виживання за методом Каплана–Мейєра (рис. 5.55).

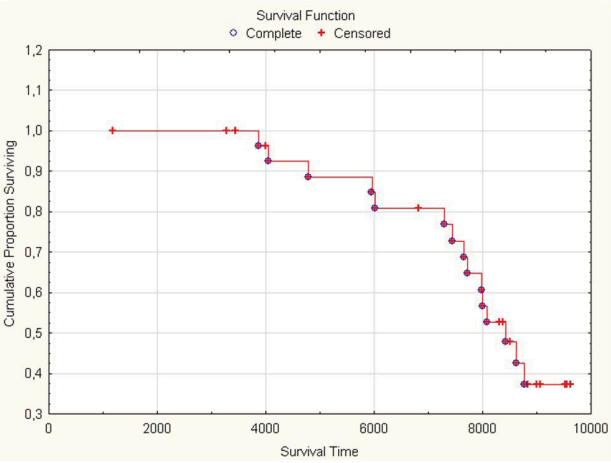


Рис. 5.55. Графік функції виживання за методом Каплана–Майєра

Для зручності інтерпретації на графіку повні спостереження позначені точками, неповні – хрестиками.

Перевага методу Каплана–Мейєра (порівняно з методом таблиць життя) полягає в тому, що оцінки не залежать від розбиття часів життя на інтервали. Отже, не потрібно розбивати часову вісь на інтервали.

Порівняння виживання у групах

Приклад 5.9. Для початкового файла даних порівняти динаміку «виживання» банків залежно від періоду заснування банку: 1 – після проголошення незалежності України, 0 – до проголошення незалежності України (змінна *independence*).

Для порівняння виживання в групах є кілька критеріїв: варіант непараметричного критерію Вілкоксона, запропонований для неповних спостережень Геханом і Пето, *F*-критерій Кокса і логарифмічний ранговий критерій [13]. Більшість цих критеріїв обчислюють відповідні нормальні наближення, які можуть

бути використані для статистичної перевірки відмінностей між групами. Однак критерії дають змогу отримати надійні результати лише при доволі великих обсягах вибірок. Для прикладу 5.9 жоден з критеріїв не дав значущих результатів (рис. 5.56).

Log-Rank Test (vuzuvanna_laba)
WW = ,71465 Sum = 6,8301 Var = ,48787
Test statistic = 1,023154 p = ,30624

Survival Time	Group	Score
1175,0+	2,0000	0,00000
3441,0+	2,0000	0,00000
3862,0	2,0000	0,91667
4784,0	2,0000	0,82576
6821,0+	2,0000	-0,17424
7289,0	1,0000	0,71465
7441,0	2,0000	0,58965
7652,0	2,0000	0,44679
7727,0	2,0000	0,28012
8090,0	2,0000	0,08012
8621,0	2,0000	-0,16988
9074,0+	2,0000	-1,16988
9530,0+	2,0000	-1,16988
9555,0+	2,0000	-1,16988

Рис. 5.56. Результати застосування логарифмічного рангового критерію

У таких випадках інформативними є візуальні методи аналізу. Графіки унаочнюють відмінності між групами (рис. 5.57).

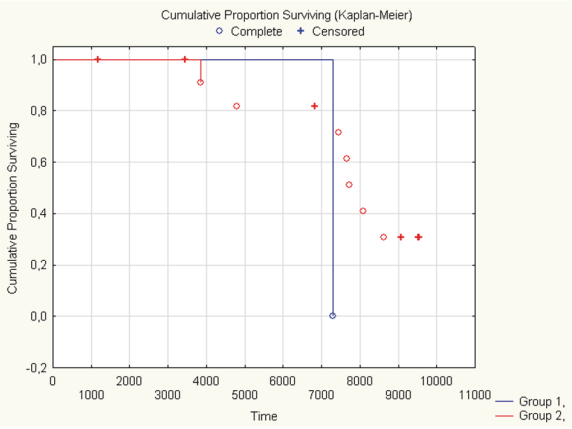


Рис. 5.57. Графік відмінностей між групами

Регресійні моделі в аналізі виживання

Більш складним завданням є оцінювання функції миттєвого ризику, яка є ймовірністю «летального результату» за малий проміжок часу за умови, що на початку досліджуваного періоду об'єкт був «живим». Це важлива характеристика прогнозу розвитку події.

Безпосереднє оцінювання функції миттєвого ризику може потребувати великої кількості спостережень. У таких випадках застосовують спеціальні моделі, одна з яких – модель Кокса пропорційних ризиків, або, мовою теорії надійності, модель пропорційних інтенсивностей [13]. Дослідження полягає у встановленні факту зв'язку окремих змінних зі спостережуваними часами життя. У разі наявності такої залежності потрібно оцінити її чисельно.

У таких дослідженнях не можна безпосередньо використовувати класичну регресію, оскільки часи «життя» не є простими лінійними функціями від відповідних регресорів. Відповідно аналіз методами множинної регресії може привести до помилкових висновків, наприклад, не виявити важливі регресори. Окрім того, знову виникає проблема неповних спостережень, оскільки окремі спостереження можуть бути незавершеними.

В аналізі виживання передбачено чотири загальних регресійних моделі для неповних даних:

► модель пропорційних інтенсивностей Кокса (Proportional hazard (Cox) regression);

► експоненціально-регресійна модель (Exponential regression);

► нормальна лінійна регресійна модель (Normal regression);

► логнормальна лінійна регресійна модель – модифікація нормальної моделі (Lognormal regression).

Для кожної з цих моделей у програмі «Statistica» можна обчислити оцінки максимальної правдоподібності (Maximum likelihood estimations).

Модель Кокса

Модель пропорційних інтенсивностей, або пропорційних ризиків Кокса, – це найбільш загальна регресійна модель, в якій передбачається, що функція інтенсивності має вигляд $h(t) = h_0(t)\, y(z_1, ..., z_m)$. Множник $h_0(t)$ називають **базовою функцією інтенсивності**.

Модель може бути параметризована, наприклад, у такому вигляді:

$$h[(t),(z_1, z_2, ..., z_m)] = h_0(t) e^{b_1 z_1 + ... + b_m z_m}.$$

У правій частині формули – добуток двох функцій, причому кожна з них залежить від своєї множини змінних. Функцію $h_0(t)$ можна розглядати як функцію інтенсивності за умови рівності нулю всіх коваріат. Вона не залежить від змінних z (коваріат). Другий співмножник залежить від змінних z, які, можливо, залежать від t. Отже, функція миттєвого ризику в моделі Кокса представлена у вигляді добутку двох співмножників, один з яких характеризує об'єкт, інший – базову функцію миттєвого ризику.

Приклад 5.10. За даними прикладу 5.8 дослідити залежність банкрутства банку від терміну його існування (до закриття чи завершення спостереження – змінна *term existence*) та типу власності (змінна *ownership*).

Результати проведеного регресійного аналізу за моделлю Кокса є статистично значущими (рис. 5.58). Для досліджуваної вибірки банкрутство банків залежить від терміну його існування та типу власності.

Dependent Variable: Survival times in days (vuzuvanna_laba)
Censoring var.: censored
Chi? = 17,2131 df = 2 p = ,00018

N=30	Beta	Standard Error	Beta 95% lower	Beta 95% upper	t-value	Wald Statist.	p	Risk ratio	Risk ratio 95% lower	Risk ratio 95% upper
term existence	-0,227320	0,057367	-0,339757	-0,114883	-3,96258	15,70201	0,000074	0,796666	0,711943	0,891470
ownership	-0,332290	0,164749	-0,655193	-0,009388	-2,01695	4,06808	0,043709	0,717279	0,519342	0,990656

Рис. 5.58. Таблиця результатів регресійного аналізу за моделлю Кокса

Регресори є незалежними між собою, отже, явище мультиколінеарності не спостерігається (рис. 5.59).

	Parameter Correlations (vuzuvanna_laba)	
Variable	**term existence**	ownership
term existence	1,000000	0,245252
ownership	0,245252	1,000000

Рис. 5.59. Таблиця кореляційного аналізу регресорів

Середні значення незалежних змінних і стандартні похибки можна переглянути у таблиці (рис. 5.60).

	Means and Standard Deviations			
variable	**mean**	st. dev.	minimum	maximum
term existence	18,533	6,318	3,000	26,000
ownership	3,467	2,013	1,000	9,000
No.days	7109,967	2261,952	1175,000	9619,000

Рис. 5.60. Таблиця середніх значень регресорів та стандартні похибки

Побудуємо графік функції виживання для випадку, коли всі незалежні змінні дорівнюють своїм середнім значенням (рис. 5.61).

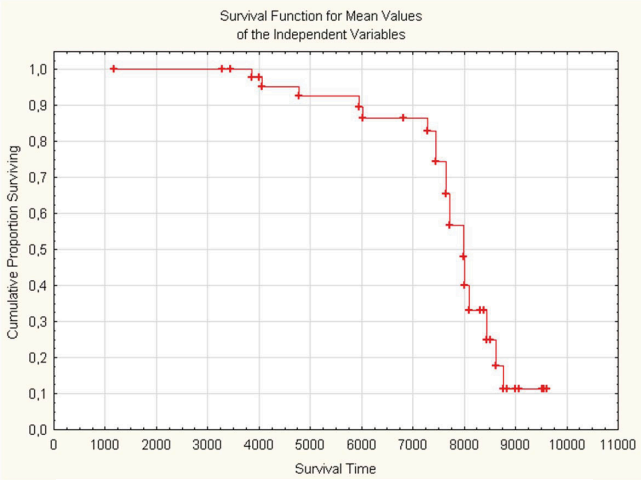


Рис. 5.61. Графік функції виживання для середніх значень агресорів

Експоненціальна регресія:

$$S(t) = e^{a+b_1 z_1 + \ldots + b_m z_m},$$

де $S(z)$ – час «життя», а – невідома константа, b_i – параметри регресії.

Для оцінювання адекватності моделі використовують критерій згоди χ-квадрат. Статистика χ-квадрат може бути обчислена як функція логарифма правдоподібності для моделі з усіма оціненими параметрами (L_1) і логарифма правдоподібності моделі, в якій всі коваріати перетворюються в 0 (L_0).

Якщо значення χ-квадрат значуще, відкидають нульову гіпотезу і роблять висновок, що незалежні змінні значущо впливають на тривалість життя.

Одним із способів перевірки адекватності експоненціальної моделі є побудува залишків часів «життя» та їх порівняння зі значеннями стандартних експоненціальних порядкових статистик (рис. 5.62).

N=30	Dependent Variable: Survival times in days Censoring var.: censored Chi? = 2,87172 df = 2 p = ,23793 **Beta**	Standard Error	t-value
term existence	0,060697	0,043075	1,409099
ownership	0,144808	0,152604	0,948914
Constant	7,915422	0,921335	8,591257

Рис. 5.62. Таблиця результатів критерію χ-квадрат

Результати застосованого критерію згоди χ-квадрат виявились незначущими (рис. 5.63). Отже, доходимо висновку, що дані дослідження не відповідають експоненціальному розподілу, про що свідчать також графіки залишків часів життя та експоненціального розподілу.

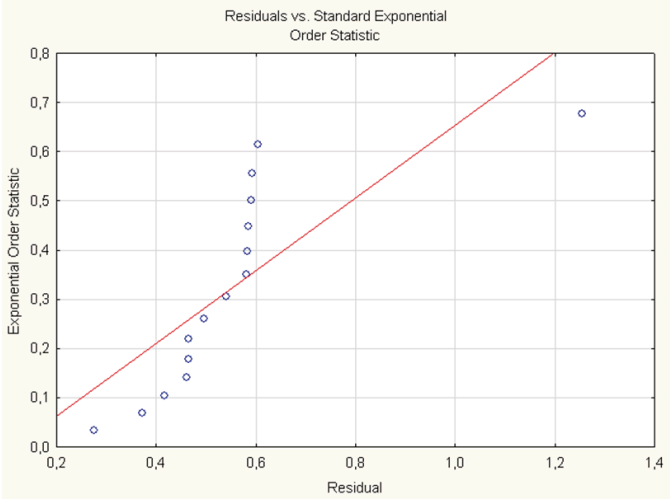


Рис. 5.63. Графіки залишків часів життя та експоненціального розподілу

Нормальна і логнормальна регресії

У нормальній моделі передбачається, що часи життя (або їх логарифми) мають нормальний розподіл. Модель збігається зі звичайною моделлю множинної регресії та може бути записана у такому вигляді:

$$t = a + b_1 z_1 + b_1 z_1 + ... + b_m z_m,$$

де t – час життя.

Якщо обирають модель логнормальної регресії, то t замінюють на ln.

Модель нормальної регресії особливо корисна, оскільки часто дані можна перетворити в приблизно нормальні за допомогою відповідного перетворення. У деякому сенсі це найбільш загальна параметрична модель (на відміну від моделі пропорційних інтенсивностей Кокса, яка є непараметричною).

ПИТАННЯ ТА ЗАВДАННЯ ДЛЯ САМОПІДГОТОВКИ ТА САМОКОНТРОЛЮ

Теоретичні запитання

1. Визначити основні поняття кластерного аналізу.
2. Формальна постановка задачі кластеризації.
3. Міри близькості.
4. Факторний аналіз.
5. Математичний опис задачі факторного аналізу.
6. Етапи факторного аналізу.
7. Обов'язкові умови факторного аналізу.
8. Метод головних компонент.
9. Базові положення дискримінантного аналізу.
10. Сфери застосування дискримінантного аналізу.
11. Формальне представлення задачі дискримінантного аналізу.
12. Основні припущення теорії дискримінантного аналізу.
13. Основні поняття канонічного аналізу.
14. Власні значення в канонічному аналізі.
15. Оцінювання канонічних ваг у канонічному аналізі.
16. Аналіз відповідності.
17. Математичне обґрунтування аналізу відповідності.
18. Визначити основні поняття аналізу виживання.
19. Функція виживання та її властивості.
20. Типи даних в аналізі виживання.

Практичні завдання

Завдання 1. Провести факторний аналіз (не проводячи поворотів) за вихідними даними за варіантом:

Варіант	Змінні	Варіант	Змінні	Варіант	Змінні
1	Y_2, Y_3, X_4, X_5	4	Y_2, Y_3, X_{14}, X_{15}	7	Y_1, Y_2, X_{10}, X_{11}
2	Y_2, Y_3, X_8, X_9	5	Y_2, Y_3, X_{16}, X_{17}	8	Y_1, Y_2, X_{12}, X_{13}
3	Y_2, Y_3, X_6, X_7	6	Y_2, Y_3, X_4, X_5	9	Y_1, Y_2, X_{14}, X_{15}

Завдання 1.1. Отримати матрицю парних коефіцієнтів кореляції.

Завдання 1.2. Визначити три найбільш значущих фактори та пояснити їх економічний зміст за матрицею факторних навантажень:

Y_1 – продуктивність праці;

Y_2 – індекс зниження собівартості продукції;

Y_3 – рентабельність;

X_4 – працемісткість одиниці продукції;

X_5 – питома вага робітників у складі промислово-виробничого персоналу (ПВР);

X_6 – питома вага закуплених виробів;

X_7 – коефіцієнт змінності обладнання;

X_8 – премії та винагороди на одного працівника;

X_9 – питома вага втрат від браку;

X_{10} – фондовіддача;

X_{11} – середньорічна чисельність ПВР;

X_{12} – середньорічна вартість основних виробничих фондів;

X_{13} – середньорічний фонд заробітної плати;

X_{14} – фондоозброєність праці;

X_{15} – оборотність нормованих оборотних коштів;

X_{16} – оборотність ненормованих оборотних коштів;

X_{17} – невиробничі витрати.

Варіанти завдань

	Y1	Y2	Y3	X4	X5	X6	X7	X8
1	9.26	204.20	13.26	0.89	0.34	1.73	0.31	0.28
2	9.44	209.60	10.16	0.93	0.33	0.99	0.15	0.25
3	12.11	223.54	13.72	1.33	0.17	1.73	0.14	0.47
4	10.81	236.70	12.83	0.68	0.32	0.47	0.18	1.53
5	9.33	62.00	10.63	0.89	0.36	1.73	0.31	0.21
6	9.87	53.10	9.12	1.53	0.33	1.33	0.17	0.13
7	8.17	172.10	25.95	1.12	0.15	0.97	0.26	0.38
8	9.12	56.50	23.39	0.99	0.32	1.82	0.29	0.38
9	5.88	52.60	14.68	1.65	0.31	0.68	0.26	0.20
10	6.30	46.60	10.05	0.56	0.15	1.80	0.28	0.35
11	6.19	53.20	13.89	0.58	0.17	1.19	0.25	0.20
12	5.46	30.10	9.68	1.53	0.15	0.97	0.49	0.20
13	6.50	146.40	10.03	0.70	0.16	1.15	0.26	0.17
14	6.61	18.10	9.13	1.77	0.15	0.02	0.28	0.25
15	4.32	13.60	5.37	0.74	0.17	0.06	0.17	0.16
16	7.37	89.80	9.86	1.08	0.34	1.39	0.17	0.21
17	7.02	62.50	12.62	1.15	0.34	0.08	0.31	0.19
18	8.25	46.30	5.02	0.97	0.34	0.77	0.18	1.24
19	8.15	103.47	21.18	1.12	0.19	0.77	0.31	0.43
20	8.72	73.30	25.17	0.99	0.19	1.08	0.18	0.14
21	6.64	76.60	19.40	0.58	0.34	0.93	0.31	0.29
22	8.10	73.01	21.00	1.03	0.34	0.10	0.15	0.43
23	5.52	32.30	6.57	1.24	0.15	0.11	0.28	0.17
24	9.37	198.54	14.19	0.89	0.19	1.44	0.18	0.21
25	13.17	598.12	15.81	0.68	0.34	0.48	0.14	0.42
26	6.67	71.69	5.20	1.03	0.19	1.24	0.18	1.19
27	5.68	90.63	7.96	0.73	0.32	0.77	0.29	1.87
28	5.19	82.10	17.50	0.73	0.19	0.93	0.30	0.15
29	10.02	76.20	17.16	0.85	0.33	0.13	0.27	0.03
30	8.16	119.47	14.54	1.03	0.34	1.73	0.29	0.24
31	3.78	21.83	6.21	0.47	0.36	0.77	0.14	0.93
32	6.45	48.40	12.08	0.56	0.33	0.16	0.29	0.13
33	10.38	173.50	9.39	0.89	0.32	0.74	0.44	0.27
34	7.65	74.10	9.28	0.99	0.15	1.95	0.14	0.17
35	8.77	68.60	11.44	1.95	0.16	0.58	0.29	0.24
36	7.00	60.80	10.31	1.03	0.16	1.77	0.18	0.19
37	11.06	355.60	8.65	0.01	0.20	0.70	0.44	0.29
38	9.02	264.81	10.88	0.02	0.15	0.74	0.31	0.25

	X9	X10	X11	X12	X13	X14	X15	X16	X17
1	0.89	0.14	112216.00	166.19	9889.98	6.40	167.29	10.08	17.72
2	1.80	0.30	37631.94	186.10	22123.47	7.80	92.88	14.76	18.39
3	1.53	0.31	45178.00	220.45	10787.25	9.76	159.01	6.45	26.46
4	0.60	0.18	76688.00	169.30	10272.25	7.90	93.96	21.83	22.37
5	1.39	0.37	7361.00	39.93	55268.00	5.35	173.88	11.94	28.13
6	1.24	0.19	84496.00	40.41	45322.00	9.96	162.30	12.60	17.55
7	1.77	0.41	114132.00	102.96	12657.25	4.50	88.56	11.52	21.79
8	0.09	0.36	7801.00	37.02	57564.00	4.88	101.16	8.28	19.52
9	0.52	0.41	84504.00	45.94	118239.88	3.46	167.29	11.52	23.85
10	0.80	2.06	35852.00	40.07	64362.00	3.62	140.76	32.40	21.88
11	0.74	0.41	43244.00	45.44	69647.88	3.56	128.52	11.52	25.68
12	0.05	0.24	6358.00	41.08	49844.00	5.65	177.84	17.28	18.13
13	1.03	0.40	47378.00	136.14	22497.50	4.28	114.48	16.20	25.74
14	1.48	0.21	4210.00	42.39	6920.00	8.83	93.24	13.36	21.21
15	0.73	0.36	3572.50	37.39	5736.00	8.52	126.72	17.28	22.86
16	0.36	0.49	54544.00	101.78	47266.00	7.22	91.27	9.72	16.38
17	0.13	0.43	91264.00	47.91	72080.00	4.82	69.12	16.20	13.21
18	0.46	0.44	5975.00	32.61	83704.00	5.47	66.24	24.88	14.41
19	0.29	0.18	64044.00	103.73	107636.00	6.23	67.16	14.76	13.44
20	1.87	2.24	34328.00	38.95	67592.00	4.25	50.40	7.56	13.69
21	0.47	0.30	58424.00	81.32	99812.00	5.38	70.89	8.64	16.66
22	0.34	0.15	83240.00	67.75	75680.00	5.88	72.00	8.64	15.06
23	0.27	0.17	6462.00	59.66	44196.00	9.27	97.20	9.00	20.09
24	0.80	2.30	114896.00	107.81	20898.50	4.36	80.28	14.76	15.91
25	0.97	0.31	21791.47	512.62	28946.00	10.31	51.48	10.08	18.27
26	1.39	0.44	83568.00	53.53	74687.88	4.72	105.12	14.76	14.44
27	0.16	0.18	68976.00	80.83	8631.25	4.18	128.52	10.38	22.88
28	0.15	0.39	67663.88	59.42	31314.00	3.13	94.68	14.76	15.50
29	1.15	2.60	34428.00	36.96	64752.00	4.02	85.32	20.52	19.35
30	0.21	0.45	127256.00	91.88	8206.25	5.20	76.32	14.46	16.95
31	0.89	0.45	6265.00	17.16	44676.00	2.72	153.00	24.88	30.53
32	1.15	2.25	33192.00	27.29	65188.00	3.12	107.34	11.16	17.78
33	0.13	0.49	127983.88	184.33	22697.00	10.38	90.72	6.45	22.09
34	0.33	0.14	41368.00	58.42	68104.00	5.65	82.44	9.72	18.29
35	0.64	0.18	33556.00	59.31	65616.00	6.67	79.12	3.24	26.05
36	0.93	0.29	124560.00	49.87	127344.00	5.93	120.96	6.45	26.20
37	0.14	0.50	110548.00	391.27	7919.00	11.89	84.60	5.40	17.26
38	0.13	0.26	95968.00	258.61	14314.75	8.30	85.32	6.12	18.95

Завдання 1.3. На основі розв'язку завдання 1.2 зробити поворот факторів за методом Варімакс нормалізований.

Завдання 1.4. Визначити три найбільш значущих фактори і дати їм економічну інтерпретацію за матрицею факторних навантажень.

Завдання 2. За даними досліджень Програми розвитку ООН (Development Report 2018) за значенням Індексу людського розвитку країни поділено на такі групи:

– Very high human development (дуже високий ІЛР);
– High human development (високий ІЛР);
– Medium human development (середній ІЛР);
– Low human development (низький ІЛР).

Завдання 2.1. За відомими даними складових ІЛР для 40 країн класифікувати 2 країни (за варіантом).

Варіанти завдань

Варіант	№ країн, які поділено на групи	№ країн, які потрібно класифікувати
1	1–12, 94–105, 138–140, 176–188	Україна, Росія
2	*Країни ЄС + Колишній СРСР	США, Японія
3	4–13, 72–81, 133–142, 175–184	50, 117
4	10–19, 60–69, 110–119, 150–159	74, 170
5	25–34, 55–64, 120–129, 160–169	3, 82
6	32–41, 72–81, 132–141, 172–181	89, 188
7	4–13, 96–105, 134–143, 179–188	1, 111
8	15–24, 80–89, 127–136, 167–176	100, 180
9	17–29, 87–96, 115–124, 173–162	30, 130

Пояснити результати дискримінантного аналізу.

Завдання 3. Провести аналіз відповідності для 40 країн світу за такими змінними:

<table>
<tr><th>1</th><th colspan="2">Змінна</th><th>Діапазон можливих значень</th></tr>
<tr><td>2</td><td colspan="2">Форма державного правління</td><td>1 – монархія; 2 – парлам. монархія; 3 – республіка; 4 – парлам. респ.; 5 – презид. Респ.; 6 – презид.-парлам. респ.</td></tr>
<tr><td rowspan="9">3</td><td>Варіант</td><td>Назва змінної</td><td rowspan="9">Значення змінної розбити на кілька категорій (не менше 5)

Наприклад:
Рівень ВВП на душу населення (у дол.)
1 – вище 50000; 2 – 25000–50000;
3 – 15000–25000; 4 – 15000–10000;
5 – нижче 10000</td></tr>
<tr><td>1</td><td>Загальний експорт (2015)</td></tr>
<tr><td>2</td><td>Загальний імпорт (2015)</td></tr>
<tr><td>3</td><td>Рейтинг країн за якістю життя (2015)</td></tr>
<tr><td>4</td><td>Рейтинг країн за рівнем корупції (2015)</td></tr>
<tr><td>5</td><td>Міграційний рейтинг країн ЄС (2015)</td></tr>
<tr><td>6</td><td>Міграційний рейтинг областей України (2015)</td></tr>
<tr><td>7</td><td>Політичний режим (для країн ЄС)</td></tr>
<tr><td>8</td><td>Індекс людського розвитку (2014)</td></tr>
</table>

Завдання 4. Провести кластерний аналіз за даними попереднього завдання.

Завдання 5. Для країн світу дослідити залежності між двома групами змінних згідно з метою дослідження.

Завдання 5.1. Оцінити канонічне значення R.

Завдання 5.2. Обчислити канонічні корені та перевірити їх значущість.

Завдання 5.3. Проаналізувати факторну структуру та надмірність.

Завдання 5.4. Оцінити канонічні коефіцієнти кореляції. За потреби провести виключення незначущих змінних.

Завдання 5.5. Записати значущі канонічні змінні, що відповідають значущому коефіцієнту кореляції.

Отримані результати пояснити.

Завдання 6. Дослідити динаміку «виживання» групи підприємств (не менше 30) самостійно обраної галузі (ТНК, кластери економіки тощо) у заданих умовах за власним вибором (наприклад, економічна криза, зміна оподаткування, перехід на нову технологію виробництва тощо) і за визначений період (відповідно до дослідження).

РОЗДІЛ 6

МОДЕЛІ DATA MINING ДЛЯ АНАЛІЗУ МВ

Інтелектуальний аналіз даних (ІАД, data mining), або розвідка даних, – це процес, мета якого – виявляти нові кореляції, тенденції, шаблони, зв'язки та категорії у результаті просіювання великих обсягів даних з використанням методик розпізнавання шаблонів і статистичних та математичних методів. Сутність і мета data mining – пошук у великих інформаційних масивах даних неочевидних, об'єктивних, корисних на практиці закономірностей [6, с. 10].

Класична статистика оперує усередненими характеристиками вибірки, які часто є фіктивними величинами. Відповідно її методи є ефективними переважно для перевірки сформульованих гіпотез, тоді як формулювання гіпотези в окремих випадках виявляється складним і працемістким завданням. Сучасні технології data mining аналізують інформацію з метою автоматичного пошуку шаблонів, характерних для окремих фрагментів неоднорідних багатовимірних даних.

Data mining – сукупність великої кількості різних методів виявлення знань. Вибір методу часто залежить від типу наявних даних і характеру результатної інформації.

При розвідці даних багаторазово виконують операції та перетворення над «сирими» даними (відбір ознак, стратифікація – розміщення шарами, кластеризація, візуалізація та регресія), призначеними для знаходження:

– інтуїтивно зрозумілих людині структур, які краще розкривають суть процесів;

– моделей передбачення результатів або характеристик реальних ситуацій на основі аналізу історичних або суб'єктивних даних.

За допомогою методів data mining можна побудувати модель, яка призведе до радикального покращення фінансового та ринкового становища компанії чи навіть позиції держави на міжнародній арені.

Завдання data mining:

► **класифікація (Classification)** – виявлення ознак, які характеризують групи об'єктів досліджуваного набору даних (класи); за цими ознаками новий об'єкт можна віднести до одного з класів;

► **кластеризація (Clustering)** – поділ об'єктів на групи;

► **асоціація (Associations)** – пошук закономірностей між пов'язаними подіями у наборі даних;

► **послідовність (Sequence), або послідовна асоціація (sequential association)**, – пошук часових закономірностей між транзакціями з метою встановлення закономірностей між подіями, пов'язаними в часі; послідовність визначає висока ймовірність ланцюжка хронологічних подій;

► **прогнозування (Forecasting)** – на основі особливостей історичних даних оцінюють майбутні значення показників; застосовують методи математичної статистики, нейронні мережі і т. ін.;

► **визначення відхилень (Deviation Detection), аналіз відхилень або викидів** – виявлення й аналіз даних, які найбільше відрізняються від загальної множини даних; відшукання нехарактерних шаблонів;

► **оцінювання (Estimation)** – прогнозування неперервних значень ознак;

► **аналіз зв'язків (Link Analysis)** – знаходження залежностей у наборі даних;

► **візуалізація (Visualization, Graph Mining)** – створення графічного образу аналізованих даних; передбачає використання графічних методів, які демонструють наявність закономірностей у сукупностях даних;

► **підбиття підсумків (Summarization)** – опис конкретних груп об'єктів за допомогою аналізованого набору даних.

Сфера застосування data mining необмежена.

6.1. Дослідження структур даних

Інтерактивне буріння (drill-down)

Першим кроком у багатьох проєктах розвідки даних є їхнє інтерактивне дослідження з метою отримання первинного уявлення про типи аналізованих змінних і можливі взаємозв'язки між ними. У системі «Statistica» та її модулі «Видобування даних» зокрема передбачено широкий набір методів розвідувального аналізу та методів графічного аналізу (графічне або візуальне видобування даних). Інтерактивне буріння надає користувачеві інструменти аналізу, які поєднують графічні та розвідувальні методи та дають можливість швидко визначати розподіли змінних і зв'язки між ними, виявити спостереження, що належать до специфічних груп даних [95].

Термін «буріння» в контексті видобування даних повністю розкриває можливості цього методу: користувач може відбирати спостереження з великого набору даних шляхом виділення в ньому підгруп, які характеризуються визначеними значеннями або діапазонами значень змінних.

Для виконання буріння можуть бути вибрані категоріальні та неперервні змінні. Способи розбиття неперервних змінних на групи:

– задати кількість категорій, на які потрібно розбити множину;

– вказати довжину кроку, щоб отримати впорядковані категорії;

– вибрати межі для змінних.

Модуль «Інтерактивне буріння» дає можливість не лише «бурити в глибину» базу даних (вибирати групи спостережень за допомогою послідовного ускладнення умов вибору), а й здійснювати «буріння вгору». Існує можливість на будь-якому етапі дослідження відмінити умову, накладену на одну з вибраних раніше змінних. У процесі подальшої обробки даних програма буде використовувати лише ті дані, які задовольняють умови, що залишилися.

Перевагами інтерактивного буріння є можливість автоматичного коригування додаткових результатів у процесі буріння: можна проводити такі види аналізу на вибраній підмножині спостережень:

► обчислення описових статистик і таблиць частот;

► побудова діаграм розмаху, які представляють розподіл неперервних змінних;

► формування матричних діаграм розсіювання, які дають уявлення про зв'язки між неперервними змінними;

► виконання будь-якого іншого виду аналізу, який дає можливість провести система «Statistica» за допомогою витягування вибраної підмножини спостережень.

Приклади використання інтерактивного буріння:

– отримати список покупок, зроблених покупцями з різними демографічними характеристиками;

– вивчити ефективність окремих ліків для людей різних вікових категорій і т. ін.;

– «витягнути» групу людей, які за результатами дослідження ринку, проведеного за методом буріння, будуть купувати новий продукт компанії.

Приклад 6.1. Файл даних містить офіційні статистичні дані – значення основних соціальних індикаторів для країн НАТО за 2017 р. (рис. 6.1). Дослідити структуру даних за номінальними змінними *Індекс соціального розвитку* та *Індекс демократії*.

Країни НАТО	1 Індекс соціального розвитку	2 Індекс щастя населення	3 Індекс якості життя	4 Індекс глобалізації	5 Індекс демократії	6 Індекс процвітання	7 Індекс тероризму	8 Індекс благодійності	9 Індекс верховенства закону	10 Індекс ефективності системи освіти
Бельгія	високий	6,891	7,51	91,75	середній	74,24	4,66	35	0,77	75,7
Великобританія	високий	6,714	7,01	87,26	високий	76,92	5,1	50	0,81	84,8
Данія	високий	7,522	8,01	88,37	високий	77,06	1,51	44	0,89	84,2
Ісландія	високий	7,504		67,9	високий	76,06	0,13	46	0,74	
Італія	середній	5,964	7,21	82,19	середній	66,2	2,75	30	0,65	53,8
Канада	високий	7,316	7,81	86,51	високий	77,03	2,96	54	0,81	79,6
Люксембург	високий	6,863		84,21	високий	75,71	0,11	38	0,76	
Нідерланди	високий	7,377	7,94	92,84	високий	77,33	2,41	51	0,85	81,6
Норвегія	високий	7,537	8,09	83,5	високий	79,25	0,1	45	0,89	75,3
Португалія	високий	5,195	6,92	85,04	середній	69,55	0,1	26	0,72	56,6
США	високий	6,993	7,38	79,73	середній	72,83	5,43	56	0,73	100

Рис. 6.1. Фрагмент файла даних прикладу 6.1

Статистика для змінних буріння (рис. 6.2–6.3).

	Frequency table of Індекс демократії N Total: 29, Selected: 29			
	Count	Cumulative Count	Percent(%)	Cumulative Percent(%)
середній	13	13	44,82759	44,8276
високий	9	22	31,03448	75,8621
низький	7	29	24,13793	100,0000

Рис. 6.2. Таблиця частот для змінної *Індекс демократії*

Буріння дає можливість проводити різного роду аналіз усіх інших змінних лише на вибраній підмножині спостережень. З гістограми (рис. 6.3) видно, що лише в 9 країнах НАТО рівень демократії доволі високий. Дослідимо детальніше цю групу (рис. 6.4–6.5).

У результаті проведеного аналізу встановлено, що у всіх країнах НАТО з високим *Індексом демократії Індекс соціального розвитку* також високий.

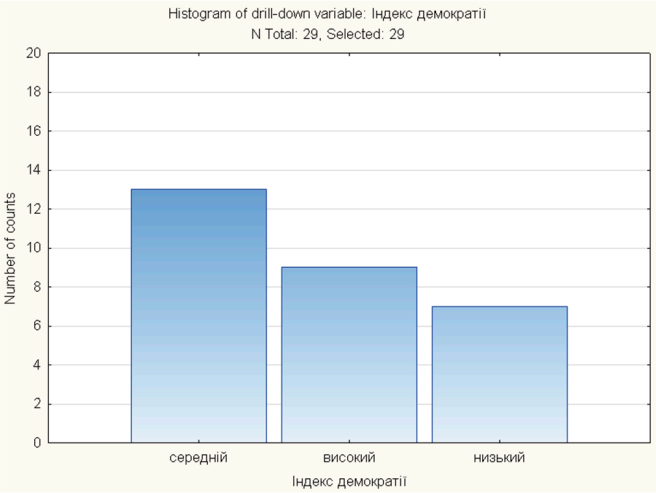


Рис. 6.3. Гістограма змінної *Індекс демократії*

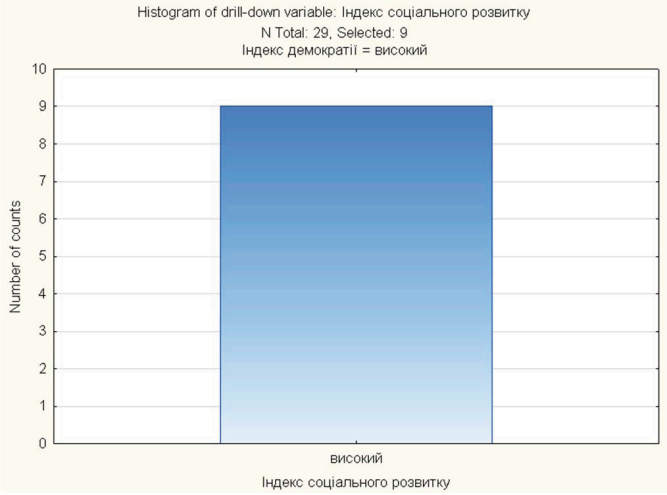


Рис. 6.4. Гістограма змінної *Індекс соціального розвитку* для групи країн з високим рівнем демократії

	Frequency table of Індекс соціального розвитку N Total: 29, Selected: 9 Індекс демократії = високий			
	Count	Cumulative Count	Percent(%)	Cumulative Percent(%)
високий	9	9	100,0000	100,0000

Рис. 6.5. Таблиця частот змінної *Індекс соціального розвитку* для групи країн з високим рівнем демократії

6.2. OLAP-системи. Пошук асоціативних правил

Багатовимірна модель даних

Вимір – це послідовність значень одного з аналізованих параметрів. Наприклад, для параметра «час» – це послідовність календарних днів, для параметра «регіон» – це може бути список міст [6, с. 54].

Множинність вимірів припускає представлення даних у вигляді багатовимірної моделі. За вимірами у багатовимірній моделі відкладають параметри, що відображають аналізовану предметну сферу.

Багатовимірне концептуальне представлення – це множинна перспектива, що складається з декількох незалежних вимірів, за якими можуть бути проаналізовані окремі сукупності даних (рис. 6.6). Одночасний аналіз за декількома вимірами визначають як **багатовимірний аналіз**.

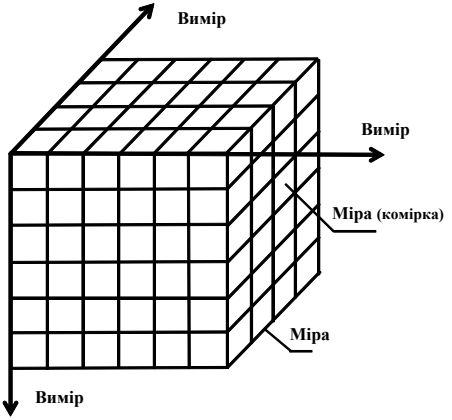


Рис. 6.6. Представлення даних у вигляді гіперкуба

Кожен вимір може бути представлений у вигляді ієрархічної структури. Наприклад, вимір «виконавець» може мати такі ієрархічні рівні: «ТНК – підрозділ – відділ – працівник». Окрім того, деякі виміри можуть мати декілька видів ієрархічного представлення. Наприклад, вимір «час» може містити дві ієрархії з такими рівнями: «рік – квартал – місяць – день» і «тиждень – день».

На перетинах осей вимірів (Dimensions) розташовують дані, які кількісно характеризують аналізовані факти, – **міри**. Наприклад, обсяги продажів, виражені в одиницях продукції або у грошовому еквіваленті, залишки на складі, витрати і т. ін.

Багатовимірну модель даних можна представити як гіперкуб, ребрами якого є виміри, а комірками – міри.

Операції, визначені для гіперкуба

Зріз – формування підмножини багатовимірного масиву даних, що відповідає єдиному значенню одного або декількох елементів вимірів, які не входять у цю підмножину. Наприклад, при виборі елемента «виграш» виміра «стратегія» зрізом даних є підкуб, в який входять усі інші виміри. Дані, що не увійшли до сформованого зрізу, пов'язані з тими елементами виміру «стратегія», які не були вказані як визначальні (наприклад, «коаліція», «обмеження», «угода» і т. ін.). Якщо розглядати термін «зріз» з позиції кінцевого користувача, то найчастіше його роль відіграє двовимірна проекція куба (рис. 6.7).

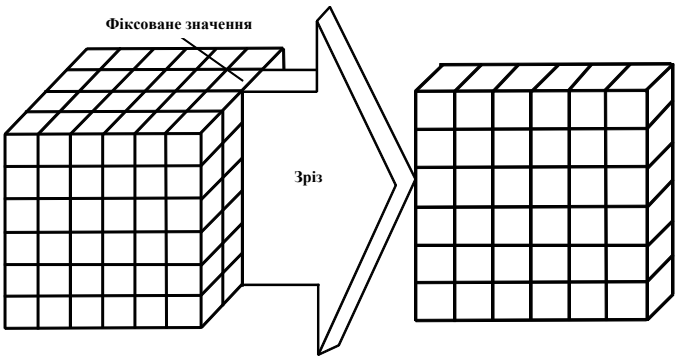


Рис. 6.7. Операція зрізу

Обертання – зміна розташування вимірів на сторінці звіту. Наприклад, операція обертання може полягати в перестановці місцями рядків і стовпців таблиці або переміщенні аналізованих вимірів у стовпцях чи рядках створюваного звіту. Обертанням куба даних є переміщення позатабличних вимірів на місце вимірів, представлених на сторінці звіту, і навпаки. При цьому позатабличний вимір стає новим виміром рядка або стовпця (рис. 6.8).

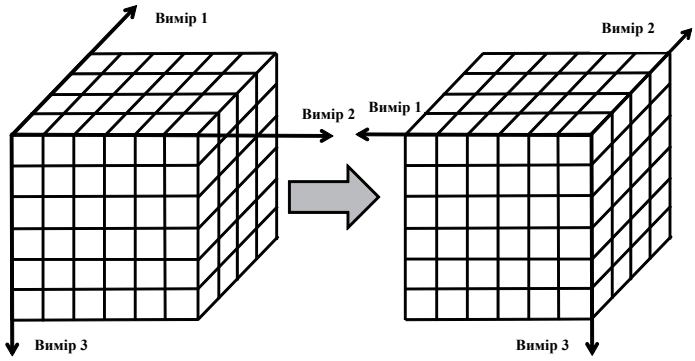


Рис. 6.8. Операція обертання

Консолідація і деталізація – операції, які визначають відповідно перехід вгору за напрямком від детального представлення даних до агрегованого, і навпаки (рис. 6.9, 6.10). Напрям деталізації (узагальнення) може бути заданий як за ієрархією окремих вимірів, так і згідно з іншими зв'язками, встановленими у межах вимірів або між вимірами. Наприклад, якщо при аналізі даних про обсяги імпорту країн ЄС виконати операцію деталізації для виміру «регіон», у звіті будуть відображені елементи «Австрія», «Бельгія», «Болгарія», «Великобританія» і т. ін. У результаті подальшої деталізації елемента «Великобританія» будуть відображені елементи «Англія», «Шотландія», «Уельс» і т. ін.

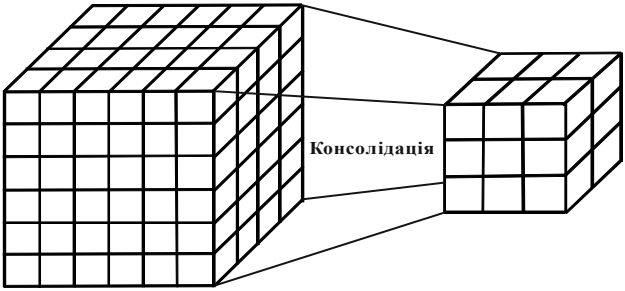


Рис. 6.9. Операція консолідації

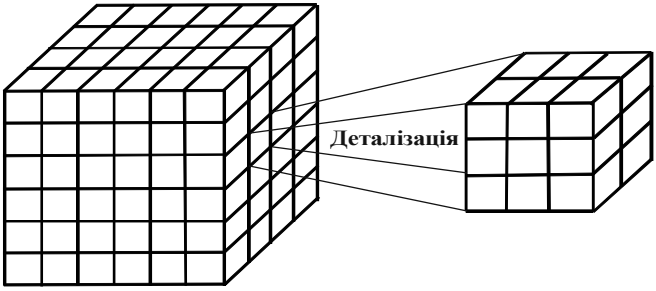


Рис. 6.10. Операція деталізації

OLAP (On-Line Analytical Processing) – технологія оперативної аналітичної обробки даних, що використовує методи і засоби для збору, зберігання і аналізу багатовимірних даних з метою підтримки процесів ухвалення рішень.

Основне призначення OLAP-систем – підтримка аналітичної діяльності, довільних запитів користувачів-аналітиків. Мета OLAP-аналізу – перевірка сформульованих гіпотез.

Пошук асоціативних правил

Одним із найбільш поширених завдань аналізу даних є визначення наборів об'єктів, що часто зустрічаються, у великій множині наборів. Опишемо це завдання в узагальненому вигляді. Для цьо-

го позначимо об'єкти, що становлять досліджувані набори, через множину:

$$I = \{i_1, i_2, ..., i_j, ..., i_n\},$$

де j_i – об'єкти, що входять в аналізовані набори, n – загальна кількість об'єктів. У сфері торгівлі, наприклад, такими об'єктами є товари, представлені в прайс-листі.

Набори об'єктів з множини *I*, які зберігають у базі даних (БД) та аналізують, називають **транзакціями**. Транзакцію можна описати як підмножину множини *I:*

$$T = \{i_j | i_j \in I\}.$$

Такі транзакції у торгівлі відповідають наборам товарів, придбаних споживачами. Їх зберігають у БД у вигляді товарного чеку або накладної.

Набір транзакцій, інформація про які доступна для аналізу, опишемо множиною:

$$D = \{T_1, T_2, ..., T_r, ..., T_m\},$$

де m – кількість доступних для аналізу транзакцій.

Приклад 6.2.

D = {{чіпси, вода, пиво}, {кокоси, вода, горіхи },
{горіхи, кокоси, чіпси, кокоси, вода}, {кокоси, горіхи, кокоси }}.

Для використання методів data mining множину *D* можна представити у вигляді табл. 6.1

Таблиця 6.1

Умовні дані для прикладу 6.2

Номер транзакції	Номер товару	Назва товару	Номер транзакції	Номер товару	Назва товару
0	1	Чіпси	2	2	Кокоси
0	3	Вода	2	1	Чіпси
0	4	Пиво	2	2	Кокоси
1	2	Кокоси	2	3	Вода
1	3	Вода	3	2	Кокоси
1	5	Горіхи	3	5	Горіхи
2	5	Горіхи	3	2	Кокоси

Множину транзакцій, в які входить об'єкт ji, позначимо:

$$D_{j_i} = \left\{ T_r \;\middle|\; j_i \in T_r;\;\; j = \overline{1,n},\; r = \overline{1,m} \right\} \subseteq D.$$

Для прикладу 6.2 множиною транзакцій, що містять об'єкт вода, є множина:

$$D_{\text{вода}} = \{\{\text{чіпси, вода, пиво}\}, \{\text{кокоси, вода, горіхи}\}, \{\text{горіхи, кокоси, чіпси, кокоси, вода}\}\}.$$

Довільний набір об'єктів (itemset) позначимо $F = \left\{ i_j \middle| i_j \in I;\;\; j = \overline{1,n} \right\}$. Наприклад, F = {кокоси, вода}.

Набір, що складається з k об'єктів, називають k-елементним набором. F = {кокоси, вода} – 2-елементний набір.

Множину транзакцій, в які входить набір F, позначимо $D_F = \left\{ T_r \middle| F \subseteq T_r;\;\; r = \overline{1,m} \right\} \subseteq D$.

Для прикладу 6.2 $D_{\{\text{кокоси, вода}\}}$ = {{кокоси, вода, горіхи}, {горіхи, кокоси, чіпси, кокоси, вода}}.

Відношення кількості транзакцій, в яку входить набір F, до загальної кількості транзакцій називають **підтримкою набору**:

$$Supp(F) = \frac{|D_F|}{D}.$$

Для набору {кокоси, вода} підтримка дорівнює 0,5, тобто цей набір входить у дві транзакції (з номерами 1 і 2), а всього транзакцій 4.

При проведенні дослідження аналітик може задати мінімальне значення підтримки аналізованих наборів $Supp_{\min}$. Набір називають **частим**, якщо значення його підтримки більше від мінімального заданого значення підтримки:

$$Supp(F) > Supp_{\min}.$$

При пошуку асоціативних правил потрібно знайти множину всіх частих наборів:

$$L = \left\{ F \middle| Supp(F) > Supp_{\min} \right\}.$$

Для прикладу 6.1 частими наборами при $Supp_{\min}$ = 0,5 є:
$Supp$({чіпси}) = 0,5;

Supp({чіпси, вода}) = 0,5;
Supp({кокоси}) = 0,75;
Supp({кокоси, вода}) = 0,5;
Supp({кокоси, вода, горіхи}) = 0,5;
Supp({кокоси, горіхи}) = 0,75;
Supp({вода}) = 0,5;
Supp({горіхи}) = 0,75.

Аналіз асоціативних правил у модулі «Statistica Data Miner»

Приклад 6.3. Файл даних містить інформацію про чистий прибуток та рекламні витрати за 2005–2015 рр. (за місяцями та середніми значеннями за рік) компанії «Procter & Gamble» – одного з найбільших рекламодавців у світі (рис. 6.11). Потрібно визначити можливі залежності між змінними.

Вихідна таблиця складається з 26 неперервних змінних та 11 спостережень.

	9 вересень (прибуток)	10 жовтень (прибуток)	11 листопад (прибуток)	12 грудень (прибуток)	13 середнє (прибуток)	14 січень (реклама)	15 лютий (реклама)	16 березень (реклама)
2005	5,56	5,81	5,52	5,81	**5,5**	0,41	0,41	0,42
2006	6,19	6,2	6,34	6,3	**5,9**	0,48	0,5	0,51
2007	6,51	7,01	6,9	7,34	**6,5**	0,55	0,57	0,61
2008	6,98	7,12	6,45	6,5	**6,7**	0,57	0,58	0,6
2009	5,3	5,66	5,89	6,29	**5,4**	0,53	0,55	0,56
2010	5,98	6,02	6,36	6,21	**6,2**	0,7	0,71	0,73
2011	6,33	6,32	6,27	6,46	**6,4**	0,67	0,7	0,73
2012	6,72	6,94	6,92	6,98	**6,6**	0,75	0,77	0,78
2013	7,79	7,62	8,12	8,42	**7,8**	0,72	0,75	0,79
2014	8,31	8,31	8,72	9,01	**8,2**	0,61	0,63	0,65
2015	7,89	7,19	7,64	7,59	**8,0**	0,71	0,72	0,73

Рис. 6.11. Фрагмент файла даних

Описовий аналіз

Графік середньорічного чистого прибутку компанії «Procter & Gamble» (рис. 6.12).

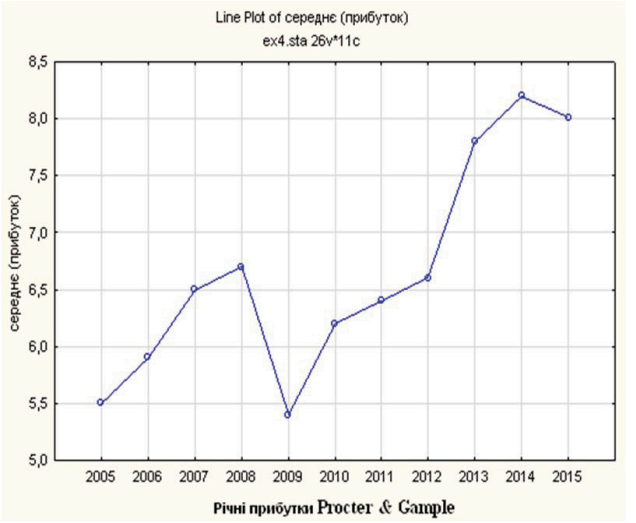


Рис. 6.12. Лінійний графік значень змінної *середнє* (*прибуток*)

На графіку (рис. 6.13) чітко виділений тренд до збільшення прибутків компанії, крім періоду з 2008 по 2009 р. (що можна пояснити світовою економічною кризою) та 2015 р. Рекламні витрати компанії також характеризує тренд до зростання до 2013 р., за винятком кризового періоду 2008–2009 рр. (рис. 6.14).

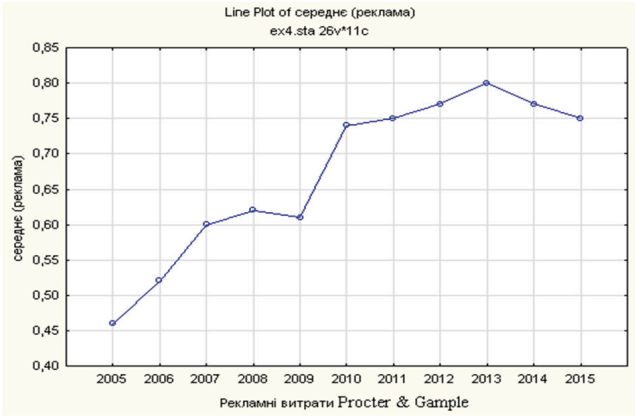


Рис. 6.13. Лінійний графік значень змінної *середнє* (*реклама*)

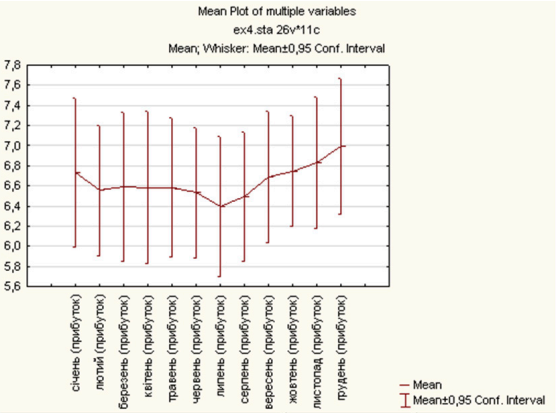


Рис. 6.14. Лінійний графік середньомісячних прибутків

Як видно з графіка, найбільша мінливість середньомісячних прибутків спостерігається для літніх (спадання) та зимових (зростання) місяців (рис. 6.14). Для витрат характерне зменшення обсягів коштів на початку року та у літній період (рис. 6.15).

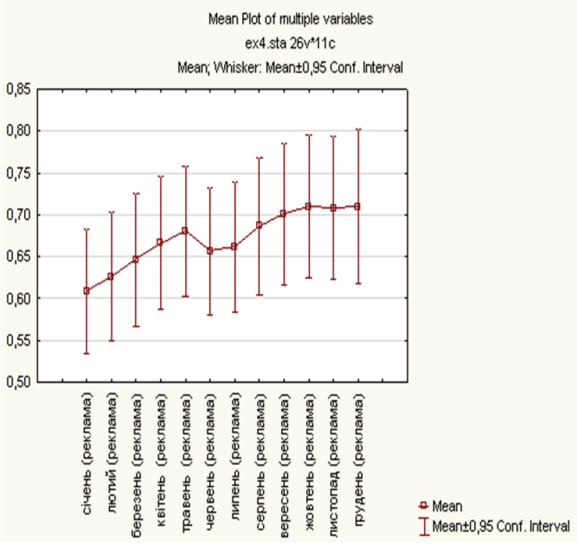


Рис. 6.15. Лінійний графік середньомісячних витрат на рекламу

Для виявлення прихованих залежностей та взаємозв'язків у даних потрібно використати асоціативні правила. Для цього необхідно перетворити початкову таблицю даних. Обчислимо середні прибутки та рекламні витрати компанії за весь період спостережень.

Якщо прибутки компанії за окремий місяць конкретного року (наприклад, червень 2013 р.) більші від середньомісячних прибутків за весь період спостережень (червень 2005–2015 рр.) (рис. 6.16), замінимо це значення в комірці початкової таблиці на 1 і на 0 – у протилежному разі. Аналогічну процедуру необхідно виконати також для змінних, в яких зберігається інформація про видатки компанії на рекламу.

Після перекодування даних отримаємо нову таблицю (рис. 6.17).

Descriptive Statistics (ex4.sta)

Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
січень (прибуток)	11	6,730909	5,140000	9,110000	1,103018
лютий (прибуток)	11	6,558182	5,280000	8,430000	0,965317
березень (прибуток)	11	6,592727	4,690000	8,510000	1,100446
квітень (прибуток)	11	6,581818	4,700000	8,230000	1,127349
травень (прибуток)	11	6,584545	4,950000	8,230000	1,027350
червень (прибуток)	11	6,532727	5,340000	8,040000	0,960241
липень (прибуток)	11	6,399091	5,110000	7,970000	1,039302
серпень (прибуток)	11	6,490000	5,320000	8,160000	0,954117
вересень (прибуток)	11	6,687273	5,300000	8,310000	0,973880
жовтень (прибуток)	11	6,745455	5,660000	8,310000	0,819541
листопад (прибуток)	11	6,830000	5,520000	8,720000	0,970773
грудень (прибуток)	11	6,991818	5,810000	9,010000	1,004777
січень (реклама)	11	0,609091	0,410000	0,750000	0,110404
лютий (реклама)	11	0,626364	0,410000	0,770000	0,114479
березень (реклама)	11	0,646364	0,420000	0,790000	0,118429
квітень (реклама)	11	0,666364	0,440000	0,820000	0,119103
травень (реклама)	11	0,680909	0,450000	0,830000	0,115798
червень (реклама)	11	0,656364	0,430000	0,780000	0,113426
липень (реклама)	11	0,661818	0,440000	0,780000	0,115743
серпень (реклама)	11	0,686364	0,470000	0,840000	0,122904
вересень (реклама)	11	0,700909	0,480000	0,870000	0,125016
жовтень (реклама)	11	0,710000	0,490000	0,890000	0,127279
листопад (реклама)	11	0,708182	0,510000	0,910000	0,127579
грудень (реклама)	11	0,710000	0,510000	0,940000	0,138058

Рис. 6.16. Описові статистики

	1 січень (прибуток)	2 лютий (прибуток)	3 березень (прибуток)	4 квітень (прибуток)	5 травень (прибуток)	6 червень (прибуток)	7 липень (прибуток)	8 серпень (прибуток)	9 вересень (прибуток)	10 жовтень (прибуток)
2005	0	0	0	0	0	0	0	0	0	0
2006	0	0	0	0	0	0	0	0	0	0
2007	0	0	0	0	0	0	0	0	0	1
2008	1	1	1	1	1	1	0	0	1	1
2009	0	0	0	0	0	0	0	0	0	0
2010	0	0	0	0	0	0	0	0	0	0
2011	0	0	0	0	0	1	1	0	0	0
2012	0	0	1	1	0	0	0	0	1	1
2013	1	1	1	1	1	1	1	1	1	1
2014	1	1	1	1	1	1	1	1	1	1
2015	1	1	1	1	1	1	1	1	1	1

Рис. 6.17. Фрагмент перетворених даних

Аналіз ***Association Rules*** **(*асоціативні правила*)** базується на побудові асоціативних правил зв'язку між спостережуваними процесами. Асоціативні правила дають можливість знаходити закономірності між пов'язаними подіями.

Визначимо, які закономірності можна знайти у даних прикладу 6.3 за допомогою аналізу асоціативних правил.

Оберемо такі параметри аналізу асоціативних правил (рис. 6.18):

– Support (підтримка);
– Confidence (ймовірність);
– Correlation (кореляція).

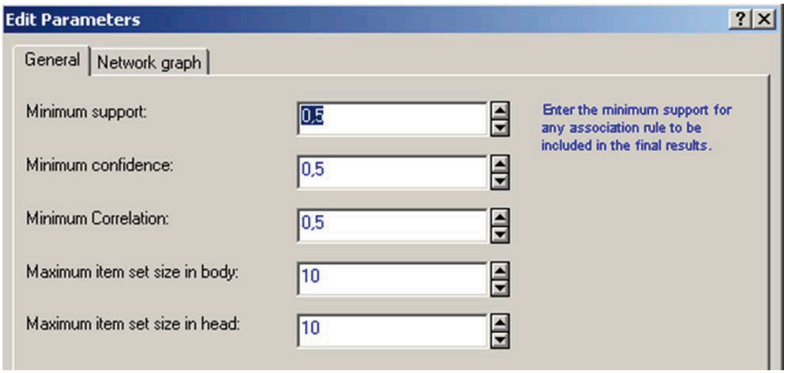


Рис. 6.18. Вікно параметрів асоціативних правил

Результати проведеного аналізу асоціативних правил представлено на рис. 6.19.

Summary of association rules
Min. support = 35,0%, Min. confidence = 50,0%, Min. correlation = 50,0%
Max. size of body = 10, Max. size of head = 10

	Body	==>	Head	Support(%)	Confidence(%)	Correlation(%)
1	квітень (реклама) == 1,	==>	червень (прибуток) == 1,	35,48387	66,66667	64,73389
3	травень (реклама) == 1,	==>	липень (прибуток) == 1,	40,32258	71,42857	74,70179
4	листопад (реклама) == 1,	==>	січень (прибуток) == 1,	37,09677	65,71429	69,82533
5	грудень (реклама) == 1,	==>	лютий (прибуток) == 0,	37,09677	65,71429	68,72565

Рис. 6.19. Асоціативні правила

Аналіз дав змогу виявити асоціативні правила, які задовольняють обмеження на мінімальні значення рівня підтримки, достовірності та кореляції.

Наприклад, якщо у квітні витрати компанії на рекламу були більшими, ніж 0,6 млрд. дол., то у червні її чистий прибуток становив більш, ніж 6,53 млрд. дол. з імовірністю 0,66 (рис. 6.19).

Використання асоціативних правил дає можливість на етапі аналізу розвідкою знайти можливі приховані залежності та зв'язки. Отримані результати можна використовувати для побудови регресійних і прогнозних моделей, проведення кластеризації й т. ін.

6.3. Дерева класифікації

Дерева класифікації (Classification Trees), що відомі також за загальною назвою **«дерева рішень» (Decision Trees**, **DT)**, – це структури даних, які дають можливість інтерпретувати шаблони даних з метою їх розпізнавання. Це ієрархічні структури, які складаються із вузлів прийняття рішень за оцінкою значень визначених змінних для передбачення належності кожного спостереження до відповідного класу [5].

Дерева класифікації можуть бути використані як компонент, що відповідає за прийняття рішень, у ході вирішення складного економічного, соціального чи політичного завдання. Кожна ситуація представляється у вигляді множини атрибутів, тому в кожній конкретній ситуації дерева рішень можуть запропонувати найкращий спосіб дій.

Дія алгоритмів побудови дерев рішень базується на застосуванні методів регресійного і кореляційного аналізу. Один з найпопулярніших алгоритмів цього сімейства – CART (Classification and Regression Trees), який полягає в ієрархічному поділі даних кожної гілки дерева на дві дочірні гілки. При цьому подальший поділ залежить від того, наскільки багато вихідної інформації вона описує. Поділ проводиться на основі найбільш високого коефіцієнта кореляції між фактором, за яким відбувається поділ, і параметром, який надалі потрібно передбачити. Коефіцієнт кореляції обчислюється за даними, які описує гілка.

Дерево рішень має один особливий вузол – **кореневий**. Це основа дерева. З нього можна перейти по дереву у будь-який інший вузол. Вузли, які знаходяться наприкінці будь-якого ланцюжка гілок, називаються **листковими вузлами**. Кожен рівень дерева можна розглядати як одне із рішень. Вузол прийняття рішень забезпечує перевірку умови, а кожна гілка означає один із можливих варіантів.

За суттю, при кожній перевірці умови відбувається сортування вибірки даних таким чином, що кожен елемент даних визначається як такий, що відповідає лише одній гілці. У результаті об'єднання таких перевірок у деяку ієрархію фактично організується процес розбиття всіх даних на щораз менші частини. Цей процес відбувається доти, поки не буде досягнутий листковий вузол (рис. 6.23).

Кількість можливих способів представлення рішень може бути доволі великою. Відповідно вибір способу розбиття залежить від типу атрибуту, що перевіряється, а також від операції, яка використовується при перевірці умови.

На практиці пакетна обробка дає можливість виявити найкращий спосіб розбиття. Переглядаються всі вибірки та оцінюється кількість сторонніх включень у наборі щодо кожного атрибута. Набір, який містить сторонні включення, містить достатньо різні змінні відгуку. Для вимірювання кількості сторонніх включень необхідно визначити ентропію (Entropy). Алгоритми CART використовують індекс Джині,

– один з основних економетричних індексів. Може бути використана також ентропія (Information Gain – інформаційний приріст), розроблена Шеноном в теорії інформації [10].

Загальноприйняте визначення кількості сторонніх включень для значень із множини S (training data)

$$Entropy(S) = \sum_{I=1}^{C} \left(- p(I) log_2 p(I) \right), \tag{6.1}$$

де C – кількість виходів,

$$p(I) = \frac{|S_i|}{|S|}, \tag{6.2}$$

де S_i – підмножина множини S, для якої аналізований атрибут A має значення i;

$|S_i|$ – кількість елементів, які містяться в S_i;

$|S|$ – кількість елементів, які містяться в S.

Ентропія дорівнює нулю, якщо всі зразки належать до однієї і тієї самої категорії. Вона дорівнює нулю, якщо величини представлені у співвідношенні 50% щодо 50%.

Щоб визначити, який атрибут найкраще вибрати для здійснення класифікації, недостатньо лише обчислити ентропію. Потрібно обчислити приріст інформації, що вимірює очікуване зменшення ентропії:

$$Gain(S, A) = Entropy(S) - \sum \left(\left(\frac{|S_V|}{S} \right) \times Entropy(S_V) \right) \tag{6.3}$$

При визначенні оптимального атрибуту для розподілу спостережень за формулою (6.3) перевага надається атрибутам з великою кількістю можливих значень. Щоб нівелювати цю особливість, застосовують відношення:

$$Gain\ Ratio(S, A) = \frac{Gain(S, A)}{SplitInfo(S, A)}, \tag{6.4}$$

де $SplitInfo(S, A)$ – ентропія (кількість сторонніх включень) у результаті розбиття значень із множини S за значеннями атрибута A.

Відмінності між змінними відгуку та вимірюваннями оцінюють за індексом Джині:

$$Gain\ Index = 1 - \sum \left(p(I)^2 \right). \quad (6.5)$$

Дерева рішень є основою ефективного методу обчислення оцінок для невідомих вибірок. Структура даних, що є їх основою, – проста та невелика, а для проходження по дереву потрібно небагато зусиль. Дерева рішень здебільшого застосовують для завдань аналізу неочевидних закономірностей даних. З їх допомогою усувається проблема масової обробки інформації в реальному часі.

Приклад 6.4. Провести дослідження країн ЄС-28 за такими показниками за 2018 р.: *рівень гендерного розвитку, якість освіти* (у % задоволених), *якість медичної допомоги* (у % задоволених), *відчуття безпеки* (% стверджувальних відповідей), *стандарт життя* (1 – низький, 2 – середній, 3 – високий) та *Індекс людського розвитк*у (для країн ЄС – високий та дуже високий) (рис. 6.20).

	Gender Development Index	Education quality	Health care quality	Standard of living	Feeling safe	HDI rank
Австрія	0,943	75	89	3	81	very hight human development
Бельгія	0,975	83	89	3	72	very hight human development
Болгарія	0,991	42	38	1	54	hight human development
Великобританія	0,965	65	77	2	79	very hight human development
Греція	0,961	45	35	1	62	very hight human development
Данія	0,977	75	85	3	80	very hight human development

Рис. 6.20. Фрагмент файла даних

Ініціалізуємо проект data mining за допомогою процедури дерева класифікації з категоріальними та порядковими предикторами (рис. 6.21).

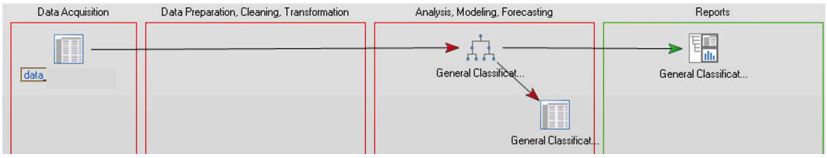


Рис. 6.21. Вікно проекту Data Miner

Результати проведеної класифікації представлені у вигляді структури дерева, яке розгалужується на ліву та праву гілки, що містять по п'ять вузлів (рис. 6.22).

Tree structure 1 (example_clasif)
Dependent variable: **Standard of living**
Options: Categorical response, Tree number 1

Node #	Left branch	Right branch	Size of node	N in class низькі	N in class середні	N in class високі	Selected category	Split variable	**Split constant**
1	2	3	28	11	10	7	низькі	**Health care quality**	64,5
2	4	5	13	11	2	0	низькі	**Health care quality**	48,5
4	6	7	5	3	2	0	низькі	**Feeling safe**	57,5
6			2	2	0	0	низькі		
7			3	1	2	0	середні		
5			8	8	0	0	низькі		
3	8	9	15	0	8	7	середні	**Health care quality**	83,0
8	10	11	9	0	8	1	середні	**Gender Development Index**	1,0
10			8	0	8	0	середні		
11			1	0	0	1	високі		
9			6	0	0	6	високі		

Рис. 6.22. Структура дерева класифікації

У стовпці **Node** вказано номер вузла, у стовпці **Size of node** – кількість об'єктів у відповідному вузлі.

Ліва вершина містить 2-й, 4-й, 6-й, 8-й та 10-й вузли, права – 3-й, 5-й, 7-й та 9-й. З першого рядка таблиці (рис. 6.22) з'ясовуємо, що в першій вершині кількість країн ЄС з низькими стандартами життя становить 11, з середніми – 10 і 7 країн класифіковано як такі, що мають високий рівень стандартів життя. З вершини 1 виходять дві гілки (права та ліва) з відповідними вершинами 2 та 3.

Умова розділення країн ЄС за вершинами 2 та 3 така: якщо значення якості медичної допомоги менше або дорівнює 64,5, то стандарти життя низькі. 11 країн ЄС з низькими стандартами життя, 2 – з середніми.

За вершинами 6 та 7 країни розподіляються таким чином: якщо значення відчуття безпеки менше або дорівнює 57,5, то стандарти життя низькі. 3 країни ЄС з низькими стандартами життя, 2 – з середніми.

Для 15 країн ЄС, рівень якості медичної допомоги яких не є низьким (вершина 3), проводиться розподіл за вершинами 8 та 9: якщо значення якості медичної допомоги менше або дорівнює 83,0, то стандарти життя середні. 8 країн ЄС з середніми стандартами життя, 7 – з високими.

Правило розподілу об’єктів за вершинами 10 та 11 (для країн зі значенням якості медичної допомоги ≤ 83): якщо значення *Індексу гендерного розвитку* менше або дорівнює 0,99, то стандарти життя середні. 8 країн ЄС з середніми стандартами життя, 1 – з високими.

Інтерпретація результатів значно спрощується за допомогою граф дерева класифікації (рис. 6.23) [42, с. 63].

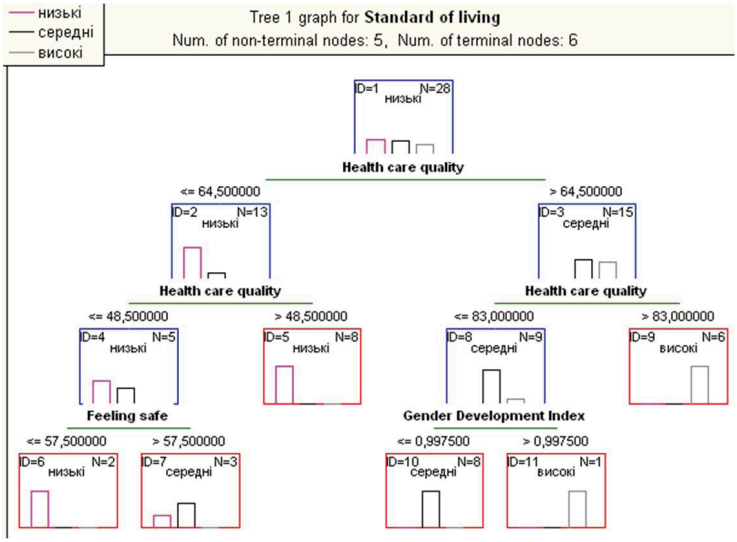


Рис. 6.23. Дерево класифікації за стандартами життя

У таблиці результатів класифікації (рис. 6.24) представлена інформація про те, скільки спостережень кожного з класів віднесено за результатами кластеризації до відповідного класу (за рядками матриці), про склад початкових класів (за стовпцями матриці) та обсяг вибірки.

Result of terminal nodes 1 (data_1)
Dependent variable: **Standard of living**
Options: Categorical response, Tree number 1

Node #	**Class 1**	Class 2	Class 3	Gain
6	2	0	0	2,000000
7	1	2	0	3,000000
5	8	0	0	8,000000
10	0	8	0	8,000000
11	0	0	1	1,000000
9	0	0	6	6,000000

Рис. 6.24. Вікно результатів класифікації

У результаті проведеної класифікації неправильно класифікованим виявився лише один об'єкт, який належить до групи 1, а опинився у групі 2 (рис. 6.25, 6.26).

Classification matrix 1 (data_1)
Dependent variable: **Standard of living**
Options: Categorical response, Tree number 1, Analysis sample

	Observed 1	Observed 2	Observed 3
Predicted 1	10,00000		
Predicted 2	1,00000	10,00000	
Predicted 3			7,000000

Рис. 6.25. Передбачені та спостережені об'єкти

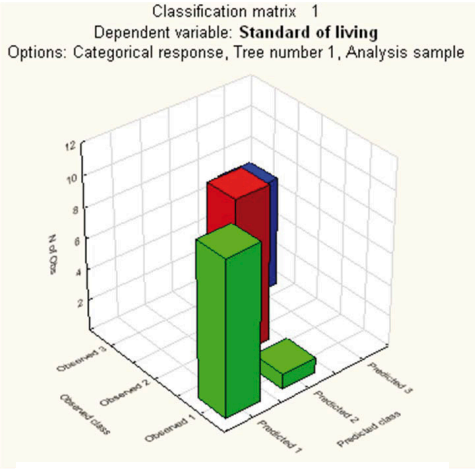


Рис. 6.26. 3D-діаграма матриці класифікації

Метод дерев класифікації є гнучким засобом для передбачення користувачем належності спостережень до відповідних класів. Застосування цього методу в «Statistica Data Miner» дає можливість проводити класифікацію одразу за декількома змінними та різними способами, що полегшує аналіз та підвищує достовірність результатів.

6.4. Нейронні мережі

Ідея нейронних мереж виникла в результаті спроб змоделювати поведінку живих істот, що сприймають дії зовнішнього середовища і навчаються на власному досвіді. Такого роду ідеї на перетині різних галузей знання характерні для науки сучасного часу.

Основним є поняття **нейронів** – спеціальних нервових клітин, здатних сприймати, перетворювати і поширювати сигнали. Нейрон має кілька каналів введення інформації (**дендрити**) і один канал її виведення (**аксон**). Аксони нейрона з'єднуються з дендритами інших нейронів за допомогою **синапсів**. При збудженні нейрон посилає сигнал своєму аксону. Через синапси сигнал передається іншим нейронам, які відповідно можуть збуджуватися або, навпаки, перебувати у стані гальмування [13, с. 561].

На основі природних біологічних образів створюють відповідні математичні моделі. Нейрон збуджується, коли сумарний рівень його вхідних сигналів перевищує **поріг збудження**, або активації. Інтенсивність сигналу, отримуваного нейроном, залежить від активності синапсів.

Нейрон отримує сигнали через кілька вхідних каналів. Кожен сигнал проходить через з'єднання (синапс), що має певну інтенсивність (вагу), яка відповідає синаптичній активності нейрона.

Поточний стан нейрона визначається рівністю:

$$u_i = \sum_{j=1}^{n} u(i,j)x(j) + \omega(i,0), \qquad (6.6)$$

де $x(j)$, $j = 1, 2, \ldots, N$ – вхідні сигнали. Коефіцієнти $\omega(i,j)$ називають **вагами** синаптичних зв'язків, додатні значення яких відповідають

збуджуючим синапсам, від'ємні значення – гальмівним. Якщо $\omega(i,j) = 0$, вважають, що зв'язок між нейронами і та j відсутній. Величину $\omega(i,0)$ називають **пороговим значенням**.

Отриманий нейроном сигнал перетворюється за допомогою функції активації, або **передавальної функції** f, у вихідний сигнал:

$$y_i = f(u_i). \tag{6.7}$$

Це одна з перших моделей нейрона, запропонована Маккаллоком і Пітсом у 1943 р. [13]. Існує й стохастична модель нейрона, в якій вихідний сигнал є випадковою величиною, що набуває двох значень, які відповідають гальмуванню або збудженню.

З математичної точки зору в моделі нейрона використовують нелінійне перетворення вектора $x(1), x(2), \ldots, x(N)$ у вихідний сигнал y_i. Функція активації f у виразі (6.7) – це деяка нелінійна функція, що моделює процес передачі збудження. Найпростіший приклад такої функції – це індикаторна, або стрибкоподібна, функція, що визначається рівністю: $f(u) = 1$, якщо $i >= 0$, $f(u) = 1$, якщо $i < 0$.

Якщо вибрати функцію f вигляду:

$$f(u) = \frac{1}{1 + e^{-bu}} \quad ,$$

де $b > 0$, отримаємо так званий **сігмоїдний нейрон**.

Об'єднані між собою нейрони утворюють мережу. З математичної точки зору ця мережа задається складним багатовимірним перетворенням, складеним з простих перетворень. За допомогою таких найпростіших перетворень можна наближати дуже складні багатовимірні функції та оцінювати складні залежності.

Виходи нейронів з'єднуються з входами інших нейронів. Сигнал від одного нейрона передається іншим нейронам (нейрон інформує про свій стан інші нейрони). З математичної точки зору це перетворення вхідних значень X на вході мережі у значення Y на виході. У термінах біології входи і виходи відповідають сенсорним і руховим нервами. Крім вхідних і вихідних нейронів, у мережі можуть бути присутніми також проміжні (приховані) шари

нейронів. Найпростіші мережі мають структуру прямої передачі сигналу: сигнали проходять від входів через приховані елементи і в результаті надходять на вихідні елементи.

Кожний нейрон як елемент мережі описується своїм набором параметрів (формули 6.6, 6.7). Вхідний шар служить для введення значень вхідних змінних, вихідний шар – для виведення результатів. Приховані вихідні нейрони з'єднані зі всіма елементами попереднього шару. Послідовність шарів та їх з'єднань називають **архітектурою мережі**.

При роботі мережі на вхідні елементи подаються значення вхідних змінних (вхідний сигнал), потім збуджуються нейрони першого проміжного шару, далі – другого проміжного шару і т. д. У результаті перетворений сигнал поступає на вихідний шар.

Перетворення сигналу

Послідовно для кожного нейрона в мережі обчислюється значення активації. Від зваженої суми виходів елементів попереднього шару віднімається порогове значення. Потім значення активації перетворюються за допомогою передавальної функції і в результаті отримують вихід нейрона, що надходить на вхід нейронів, з якими він з'єднаний.

Навчання мережі

Нейронні мережі використовують у задачах класифікації, прогнозування та побудови нелінійних залежностей (нелінійна регресія). Однак для цього мережу потрібно навчити.

Людина завжди намагалась зрозуміти, як організоване мислення, коли людський мозок, що складається з найпростіших нейронів, осягає найглибші закономірності. Процес отримання знань моделюється за допомогою нейронних мереж.

Відомо, що знання формуються послідовно, – вони не надаються у завершеному вигляді, а здобуваються за допомогою навчання. Цей принцип використаний у нейронних мережах. Після побудови моделі нейрона і нейронної мережі потрібно запропонувати модель навчання.

Формально співвідношення (6.6) і (6.7) задають просте перетворення величин з різними функціями *f*. Нехай маємо складне перетворення *F* вхідного набору даних (надходить на вхід мережі) у вихідний набір (спостерігається на виході мережі).

Потрібно реалізувати перетворення *F* за допомогою нейронної мережі. Математичною мовою це означає, що потрібно наблизити невідому складну функцію найпростішими перетвореннями, що задаються рівняннями (6.6) та (6.7). Відомо, що така мережа існує, але не вказується, як саме її налаштувати. Можна використати загальний підхід, пов'язаний з навчанням, тобто послідовним отриманням знань, покаранням – за неправильну відповідь і заохоченням за правильну [13].

Спочатку потрібно визначити архітектуру мережі, тобто встановити кількість нейронів і зв'язки між ними, вибрати конкретну синаптичну функцію, яка моделює процес передачі збудження.

Розіб'ємо дані на дві частини: навчальну та контрольну вибірки.

Загальна ідея полягає в тому, що спочатку на вхід мережі подається навчальна вибірка з відомими результатами – величини X та спостережувані відгуки $Y = F(X)$.

Змінюючи ваги $\omega(i, j)$ та значення порогу активації, для кожного нейрона налаштовують мережу (знаходять якнайточніше значення функції F).

Далі на тестовій вибірці тестують побудовану мережу (мережі). Наприклад, у задачі класифікації можна задати умову, щоб мережа правильно класифікувала не менше 90% спостережень. У задачі прогнозування – прагнути до того, щоб точність прогнозу на визначену кількість кроків уперед була не нижче від заданої. Якщо мережа пройшла тестування, її можна використовувати для аналізу даних, будувати прогноз або проводити класифікацію. Цей процес можна ефективно організувати лише за допомогою комп'ютерних технологій через його працемісткість і складність перетворення даних.

У цьому процесі є певна похибка, пов'язана, наприклад, з вибором навчальної вибірки і ризиком застосування мережі на реаль-

них даних. Однак такі самі похибки виникають при застосуванні будь-яких математичних методів на практиці. Нейронні моделі є синтезом різних методів, які можна реалізувати лише за допомогою комп'ютерних технологій.

Приклад 6.5. Використовуючи автоматизовані нейронні мережі «Statistica», провести регресійний аналіз ІЛР та його складових на основі даних для країн ЄС-28 за 2017 р. (рис. 6.27).

У «Statistica» можна побудувати нейромережу для виконання різних видів аналізу: регресію, класифікацію, прогнозування часових рядів (з неперервною чи категоріальною залежною змінною), кластерний аналіз [95].

	1 Human Development Index (HDI)	2 Life expectancy at birth	3 Expected years of schooling	4 Mean years of schooling	5 Gross national income (GNI) per capita, $
Австрія	0,885	81,4	15,7	10,8	43,869
Бельгія	0,890	80,8	16,3	11,3	41,187
Болгарія	0,782	74,2	14,4	10,6	15,596
Великобританія	0,907	80,7	16,2	13,1	39,267
Греція	0,865	80,9	17,6	10,3	24,524
Данія	0,923	80,2	18,7	12,7	44,025
Естонія	0,861	76,8	16,5	12,5	25,214
Ірландія	0,916	80,9	18,6	12,2	39,568
Іспанія	0,876	82,6	17,3	9,6	32,045

Рис. 6.27. Фрагмент файла даних

Архітектуру мережі можна аналізувати на основі ітерацій алгоритму та фіксації помилок моделей (рис. 6.28). Для регресії використовують середньоквадратичну помилку, для класифікації – відсоток правильно класифікованих спостережень.

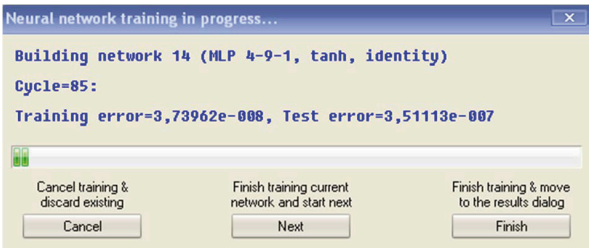


Рис. 6.28. Навчання нейронної мережі

Аналіз результатів

У вікні результатів можна проаналізувати отримані рішення. Програма відбере кращі мережі та покаже якість рішення (рис. 6.29).

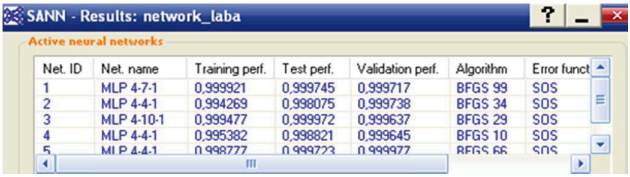


Рис. 6.29. Побудовані нейронні мережі

Далі обирають найкращу мережу для кожного конкретного дослідження (активують модель).

Одним із способів перевірки є порівняння спостережуваних значень і передбачених результатів. Це може бути порівняння емпіричних даних і передбачених значень для вибраної мережі (наприклад, для навчальної та тестової вибірок (рис. 6.30)) або перегляд матриці похибок класифікації на тестовій вибірці (для завдань класифікації).

Найкращі мережі зберігають у форматі PMML з метою подальшого використання, наприклад, для автоматичної побудови прогнозів.

Case name	**Human Development Index (HDI) Target**	Human Development Index (HDI) - Output 1. MLP 4-4-1
Австрія	0,885027	
Бельгія	0,890263	0,867611
Болгарі	0,781675	0,867677
Великобританія	0,906698	0,867154
Греція	0,865187	0,867849
Данія	0,923328	0,867485
Ірланді	0,915542	0,867632
Іспанія	0,876124	0,867924
Кіпр	0,850000	0,867452
Латвія	0,818766	0,867412
Литва	0,839422	0,867406
Люксембург	0,891852	0,867601
Мальта	0,838978	0,867740
Німеччина	0,916080	0,867907
Польща	0,842678	0,867920
Португалія	0,830095	0,867466

Рис. 6.30. Таблиця спостережуваних і передбачених значень

6.5. Споживчий кредитний скоринг

Скоринг – система оцінювання кредитоспроможності на основі кредитних історій банків. Визначає ймовірність дефолту потенційного позичальника на основі його соціально-демографічних характеристик. На основі бази даних «непривабливих» і «привабливих» кредитів фінустанова за допомогою статистичних інструментів може виявити фактори, що впливають на здатність і бажання клієнта повернути борг [88].

У процесі надання кредиту банки зацікавлені у вивченні платоспроможності майбутнього споживача кредиту. Метою скорингу є моделювання чи передбачення ймовірності, з якою претендент на кредит може бути віднесений до «привабливих» або «непривабливих» клієнтів [94].

За допомогою методів «Statistica Data Miner» можна побудувати кредитно-скорингову модель для ідентифікації входів або предикторів (незалежних змінних), які відокремлюють ризикованих клієнтів від усіх інших. Інтелектуальні методи успішно застосовують на тестових даних. У подальшому вони можуть бути використані для передбачення нових ризикованих клієнтів.

Приклад 6.6. Побудувати математичну (статистичну) модель, за допомогою якої на основі кредитних історій позичальників банк може визначити їх солідність (кредитоспроможність).

Файл даних містить 1000 спостережень за 20 змінними (предикторами), в яких збережена інформація про колишніх клієнтів, що були позичальниками в одному з німецьких банків (табл. 6.2, рис. 6.31) [82].

Таблиця 6.2

Структура початкового файла:

№ з/п	Змінна	Опис	Категорії	Значення
1	2	3	4	5
1	**Solvency**	Кредитоспроможність	кредитоспроможні	1
			некредитоспроможні	0
2	**Balance**	Баланс поточного рахунка	немає поточного рахунка	1
			немає балансу або дебету	2
			0 <= ... < 200 DM	3
			... >= 200 DM або розрах. рахунок, хоча б на 1 рік	4
3	**Duration**	Тривалість у місяцях (за термінами)	<=6	10
			6 < ... <= 12	9
			12 < ... <= 18	8
			18 < ... <= 24	7
			24 < ... <= 30	6
			30 < ... <= 36	5
			36 < ... <= 42	4
			42 < ... <= 48	3
			48 < ... <= 54	2
			> 54	1
4	**Payment**	Кредитна історія у цьому банку	нестабільна виплата попередніх кредитів	0
			проблематичний поточний рахунок / існують додаткові кредити в інших банках	1
			немає попередніх кредитів	2
			немає ніяких проблем з поточними кредитами в цьому банку	3
			попередні кредити в цьому банку повернено	4
5	**Purpose of credit**	Ціль кредиту	новий автомобіль	1
			автомобіль б/у	2
			предмети меблів	3
			телевізор	4
			побутова техніка	5
			ремонт	6
			освіта	7
			відпустка	8
			перекваліфікація	9
			бізнес	10
			інше	0

Продовження табл. 6.2

1	2	3	4	5
6	**Amount**	Сума кедиту, в DM	<=500	10
			500 < ... <= 1000	9
			1000 < ... <= 1500	8
			1500 < ... <= 2500	7
			2500 < ... <= 5000	6
			5000 < ... <= 7500	5
			7500 < ... <= 10000	4
			10000 < ... <= 15000	3
			15000 < ... <= 20000	2
			> 20000	1
7	**Value of savings or stocks**	Сума заощаджень, у DM	< 100	2
			100 <= ... < 500	3
			500 <= ... < 1000	4
			>= 1000	5
			немає заощаджень	1
8	**Has been employed by current employer for**	Досвід роботи в поточній сфері	безробітний	1
			<= 1 року	2
			1 <= ... < 4 років	3
			4 <= ... < 7 років	4
			>= 7 років	5
9	**Install-ment**	Ставка, у % від наявного доходу	>= 35	1
			25 <= ... < 35	2
			20 <= ... < 25	3
			< 20	4
10	**Marital Status / Sex**	Сімейний стан /стать	чоловік: розлучений / проживає окремо	1
			жінка: розлучена / проживає окремо / заміжня	2
			чоловіки: вільний	2
			чоловік: одружений / вдівець	3
			жінка: вільна	
11	**Further debtors / Guaran-tors**	Інші дебітори / гаранти	немає	1
			співзаявник	2
			поручитель	3

Продовження табл. 6.2

1	2	3	4	5
12	**Living in current household for**	Кількість років проживання за цією адресою	< 1	1
			1 <= ... < 4	2
			4 <= ... < 7	3
			>= 7	4
13	**Most valuable available assets**	Найцінніше майно	право власності на будинок або землю	4
			договір страхування житла / життя	3
			машина / інше	2
			немає активів	1
14	**Age**	Вік	<=25	1
			26 <= ... <= 39	2
			40 <= ... <= 59	3
			60 <= ... <= 64	4
			>= 65	5
15	**Further running credits**	Боргові зобов'язання	в інших банках	1
			в торгових точках	2
			немає боргів	3
16	**Type of apartment**	Тип квартири	орендована квартира	2
			власна квартира	3
			будинок	1
17	**Number of previous credits at this bank**	Кількість попередніх кредитів у цьому банк	один	1
			два або три	2
			чотири або п'ять	3
			шість і більше	4
18	**Occupation**	Сфера діяльності	безробітний / некваліфікований, без постійного місця проживання (ПМП)	1
			некваліфікований, з ПМП	2
			кваліфікований працівник / молодший державний службовець	3
			вища посадова особа / підприємець	4
19	**Number of persons entitled to mainte-nance**	Кількість утриманців	від 0 до 2	2
			3 і більше	1
20	**Telephone**	Телефон	немає	1
			є	2
21	**Foreign worker**	Резидент	так	1
			ні	2

	1 Creditability	2 Balance	3 Duration	4 Payment	5 Purpose	6 Amount	7 Value of savings or stocks	8 Has been employed by current employer for	9 Installment	10 Marital Status / Sex	11 Further debtors / Guarantors	12 Living in current household for	13 Most valuable available assets	14 Age	15 Further running credits
583	0	4	24	2	3	3621	2	5	2	3	1	4	3	31	3
584	0	1	12	2	4	741	2	1	4	2	1	3	2	22	3
585	0	1	12	2	2	7865	1	5	4	3	1	4	4	53	3
586	1	1	24	2	1	2910	1	4	2	3	1	1	4	34	3
587	1	1	18	4	0	5302	1	5	2	3	1	4	4	36	3
588	1	1	36	2	2	3620	1	3	1	3	3	2	2	37	3
589	1	1	18	2	3	3509	1	4	4	2	3	1	1	25	3
590	1	2	12	2	2	3017	1	2	3	2	1	1	1	34	3
591	1	1	12	2	2	1657	1	3	2	3	1	2	1	27	3
592	1	1	8	4	10	1164	1	5	3	3	1	4	4	51	1
593	0	1	36	4	2	6229	1	2	4	2	2	4	4	23	3
594	0	1	24	1	0	1193	1	1	1	2	2	4	4	29	3
595	1	1	30	0	2	4583	1	3	2	1	3	2	1	32	3
596	1	1	36	4	2	5371	1	3	3	3	3	2	2	28	3
597	1	1	12	2	2	708	1	3	2	3	3	3	2	38	3

Рис. 6.31. Фрагмент файла даних для прикладу 6.6

Для кожного споживача визначена бінарна характеристика «кредитоспроможність» (змінна *Creditability*). Ця змінна містить інформацію про «привабливість» кожного з клієнтів. У наборі даних 70% клієнтів класифіковано як «привабливі», 30% – як «непривабливі». Клієнти, які прострочили оплату на 90 днів, можуть бути віднесені до категорії з високим ризиком. Відповідно клієнти без прострочення платежу можуть бути віднесені до категорії з низьким ризиком. Інші критерії визначення «привабливих» і «непривабливих» клієнтів враховують перевищення кредитного ліміту, кількості прострочених місяців оплати або функції від цих чи інших змінних.

Підготовка даних

Модуль «Statistica Data Miner» дає можливість найпростішим чином застосувати до даних потужні методи моделювання і виявити переваги результатних моделей, пов'язані з їх прогностичними й описовими можливостями. Дані – це головний ресурс для виявлення прихованих закономірностей між змінними. Перед застосуванням інструментів видобування даних потрібно належним чином підготувати початкові дані.

У протилежному разі може бути ефект «сміттєпроводу». У підсумку будуть прийняті неправильні стратегічні рішення, а будь-яка помилка може привести до катастрофічних втрат. Для прийняття найбільш результативного можливого рішення важливо попередньо підготувати дані і відповідно збільшити точність моделі.

Особливості підготовки даних:

► оцінювання значень основних описових статистик (середніх, максимальних і мінімальних значень, квантилів тощо);

► наявність або відсутність у даних значень, які суттєво відхиляються від загалу (викидів);

► наявність або відсутність у групах «привабливих» і «непривабливих» клієнтів пропущених значень у даних;

► потреба у попередніх перетвореннях даних;

► необхідність відсіювання ознак (зменшення кількості змінних).

Відсіювання ознак

У «Statistica Data Miner» передбачені такі ефективні інструменти data mining: попередня обробка, чистка та фільтрація даних для ефективного відбору ознак з тисяч (або навіть мільйонів) можливих предикторів; автоматичне оптимальне поєднання, об'єднання декількох джерел даних з вирівнюванням даних, залежних від декількох критеріїв; агрегування даних; обробка пропущених (відсутніх) значень; видалення повторюваних записів, викидів і т. ін.

Для зниження працемісткості завдання можна зменшити розмірність масиву даних. Доступна в «Statistica Data Miner» процедура **Feature Selection and Variable Screening (Чистка та фільтрація даних)** автоматично знаходить важливі предиктори, які дають можливість правильно провести класифікацію клієнтів (рис. 6.32).

Рис. 6.32. Процедура чистки та фільтрації даних у вікні браузера процедур Data Mining

Щоб виділити найбільш значущі при побудові прогнозу змінні, побудуємо електронну таблицю значущості (рис. 6.33) та графік предикторів (рис. 6.34).

Результати виконання аналізу Загальні дерева класифікації та регресії

Predictor importance 1 (kredit_scoring)
Dependent variable: **Creditability**
Options: Continuous response, Tree number 1

	Variable rank	Importance
Balance	100	1,000000
Payment	85	0,846986
Purpose of credit	82	0,816829
Value of savings or stocks	79	0,792796
Amount of credit in DM	74	0,743971
Duration	66	0,662649
Has been employed by current employer for	65	0,646208
Age	56	0,555160
Most valuable available assets	53	0,533928
Marital Status / Sex	52	0,516305
Further debtors / Guarantors	47	0,472759
Installment in % of available income	32	0,316788
Type of apartment	31	0,307759
Occupation	30	0,300461
Foreign worker	29	0,291224
Further running credits	28	0,284943
Number of previous credits at this bank	27	0,265068
Living in current household for	25	0,253355
Number of persons entitled to maintenance	13	0,127389
Telephone	6	0,062831

Рис. 6.33. Таблиця значущості предикторів для залежної змінної *Creditability*

За таблицею та графіком значущості предикторів для залежної змінної *Creditability* роблять висновки про дискримінантну вагу кожного з предикторів: чим більший ранг відповідного предиктора, тим більша його значущість. Змінні *Balance of current account* (баланс поточного рахунка), *Payment of previous credits* (оплата попередніх кредитів), *Purpose of credit* (ціль кредиту) і *Further debtors / Guarantors* (інші дебітори / гаранти) виділені як найважливіші предиктори (рис. 6.34).

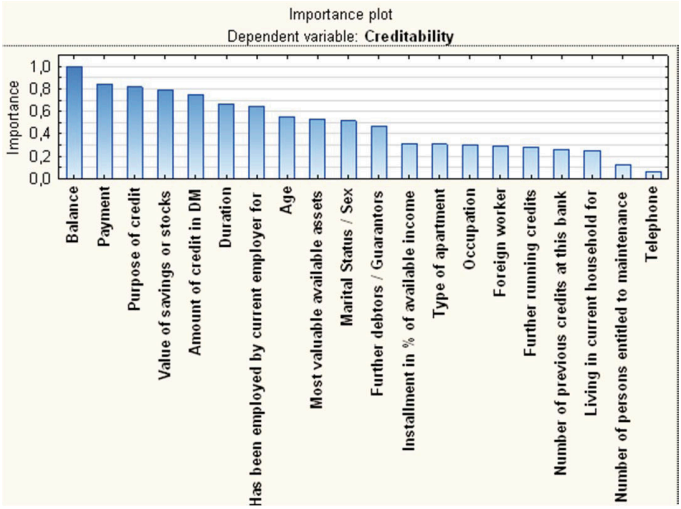


Рис. 6.34. Графік значущості предикторів для залежної змінної *Creditability*

Ці предиктори будуть надалі перевірені з використанням таких засобів видобування даних та алгоритмів машинного навчання (Data Mining and Machine Learning Algorithms) системи «Statistica Data Miner»:

– Standard Classification Trees with Deployment (C And RT) – загальні дерева класифікації та регресії;

– Standard Classification CHAID with Deployment (C And RT) – загальні CHAID-моделі;

– Boosting Classification Trees with Deployment – використання дерев класифікації та регресії;

– Intelligent Problem Solver with Deployment – інтелектуальне розв'язання завдань;

– Support Vector Machine with Deployment (Classification) – метод опорних векторів (класифікація);

– MARSplines for Classification with Deployment – багатовимірні адаптивні сплайни (MAP-сплайни).

У сфері кредитного скорингу найбільш популярним інструментом для передбачення рівня ризику позичальників є метод класифікації. Однак використання паралельно різних інструментів дає можливість виявити приховані залежності та підтвердити попередні висновки.

Етапи підготовки / аналізу даних

1. Поділ вихідної множини даних на дві підмножини (34% спостережень – для тестування і 66% – для побудови моделі) за допомогою методу Split Input Data into Training and Testing Samples (Classification) (Розбиття вхідних даних на навчальну та тестувальну вибірки (Класифікація)) (рис. 6.35).

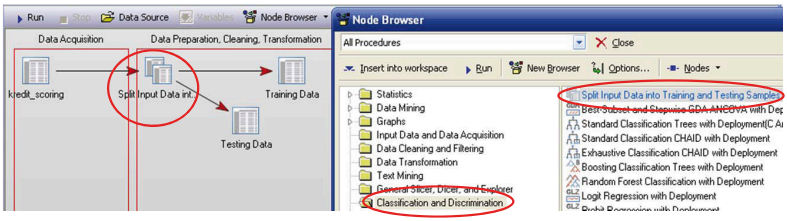


Рис. 6.35. Браузер процедури розбиття вхідних даних

2. Метод Stratified Random Sampling (Стратифікована випадкова вибірка) використовують для вилучення однакової кількості спостережень для обох типів: «привабливих» і «непривабливих» клієнтів (рис. 6.36).

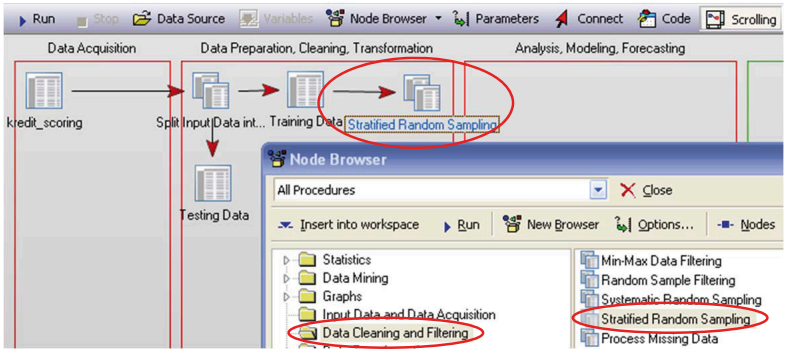


Рис. 6.36. Браузер процедури стратифікованої випадкової вибірки

3. Інструмент Feature Selection (Відсіювання ознак) використовують для виявлення найкращих предикторів, які класифікують клієнтів на «привабливих» і «непривабливих» (рис. 6.37).

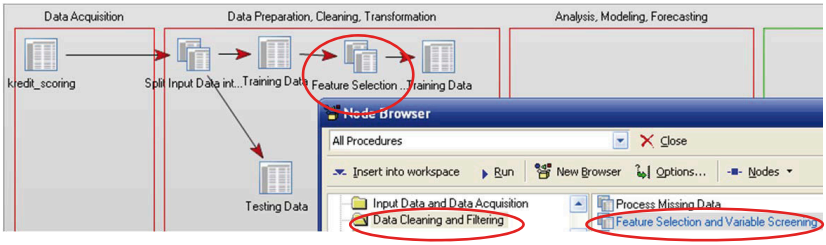


Рис. 6.37. Браузер процедури відсіювання ознак

У результаті застосування процедури відсіювання ознак кількість можливих предикторів зменшилась з 20 до 11.

4. Застосування різних прогнозних моделей Machine Learning algorithms (Алгоритми машинного навчання) для визначення та розуміння взаємозв'язків між змінними (рис. 6.39). Для здійснення різних видів аналізу над одними і тими самими даними попередньо виконують процедуру створення дублікатів даних, що підлягають дослідженню, – Multiple Copies of Data Source (Множинне копіювання джерела даних) (рис. 6.38).

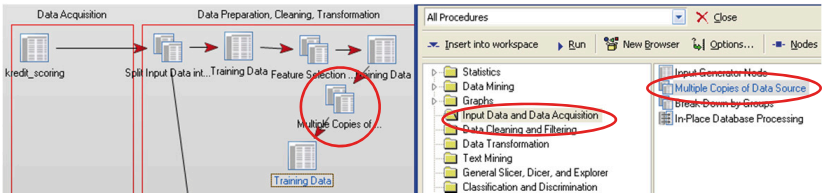


Рис. 6.38. Браузер процедури тиражування джерела даних

5. Для вибору оптимальної моделі прогнозування використовують порівняльні інструменти: Lift Charts, Gain Charts, Cross tabulation (Ліфтові карти, Підсилюючі карти) і т. ін. (рис. 6.40).

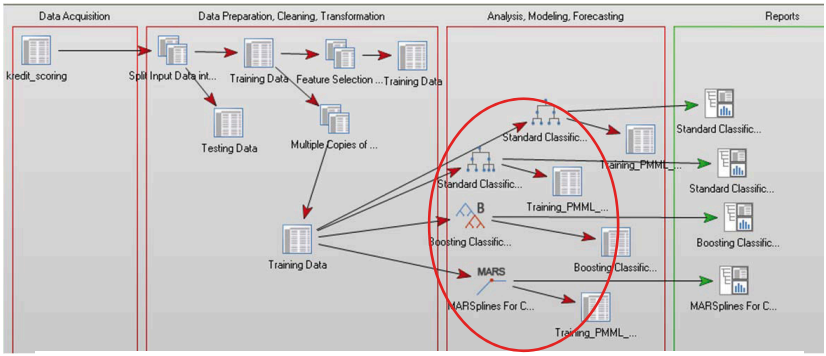


Рис. 6.39. Браузер процедур застосування прогнозних моделей для навчання

6. Для оцінювання точності прогнозу застосовують модель для накладень тестової вибірки («hold-out» sample). Ковзний контроль або крос-перевірка чи крос-валідація (cross-validation, CV) – це процедура емпіричного оцінювання узагальнюючої здатності алгоритмів, які навчаються на прецедентах.

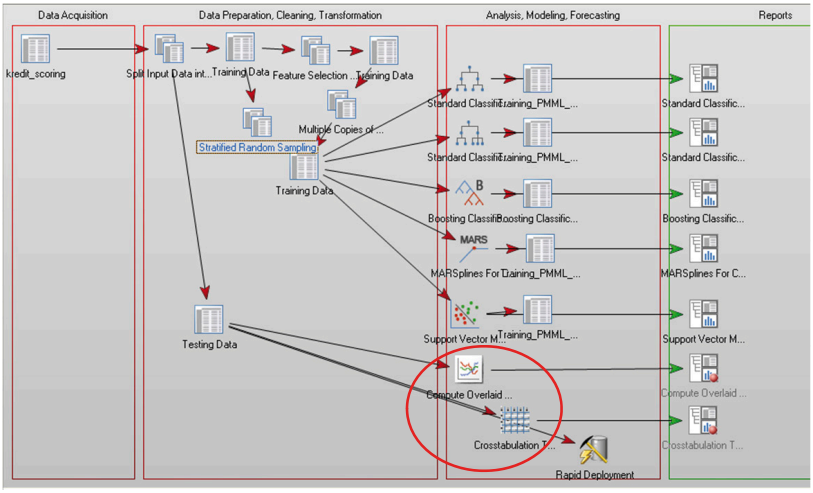


Рис. 6.40. Робочий простір Data Miner після завершення аналізу

Аналіз результатів

Дерева рішень – потужний засіб для класифікації та прогнозування (рис. 6.41). Перевагою дерев рішень є те, що вони можуть бути представлені графічно, що робить їх особливо легкими для сприйняття.

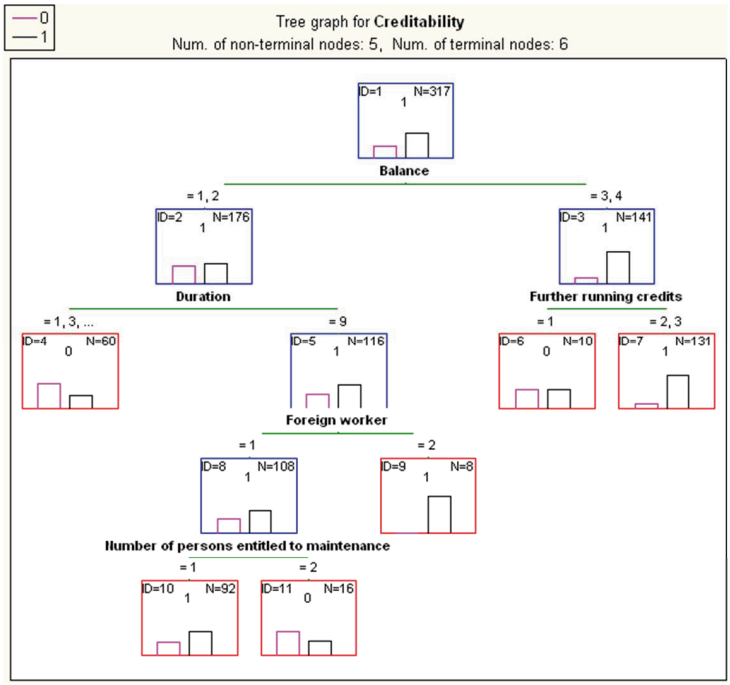


Рис. 6.41. Дерево рішень CHAID для кредитоспроможності

CHAID-алгоритм створив дерево з шістьма термінальними вершинами як результат шести if-then-умов для передбачення «привабливих» / «непривабливих» клієнтів. Термінальні вершини (термінальне листя) – це елементи, в яких подальший поділ не приведе до збільшення точності рішення. За допомогою інструменту Стратифікована випадкова вибірка (Stratified Random Sampling) сформовано навчальну множину даних (317 спостережень) з однаковою пропо-

рцією «привабливих» і «непривабливих» клієнтів. Процес побудови дерева починається від верхньої (кореневої) вершини.

На першому етапі крайня права вершина результату від першого поділу, яка містить 141 випадок, поділяється на основі предиктора *Further running credits* (наступне використання кредитів) ще на дві вершини. На наступному етапі утворюються дві вершини, які містять 10 та 131 спостереження відповідно. Це вершини з більшістю спостережень, що відповідають «привабливим» клієнтам. Оскільки подальший поділ цих вершини не може збільшити точність прогнозу, вони стають термінальними. Крайня ліва вершина, яка містить 176 випадків, поділяється на основі предиктора *Duration* (тривалість) ще на дві вершини і т. д.

Правила рішень можна отримати, рухаючись по шляху до кожної термінальної вершини.

Матриця класифікації – CHAID-модель

Матриця класифікацій дає змогу порівняти класифікації, які є насправді, з передбаченими класифікаціями (тими, яких більшість у відповідному термінальному вузлі) з метою підсумовування точності класифікації для різних вихідних значень. Програма обчислює матрицю передбачених і спостережених частот вихідних значень для тестової множини, які відображені в таблиці та на гістограмі (рис. 6.42, 6.43).

	Observed 0	Observed 1
Predicted 0	185,0000	10,0000
Predicted 1	14,0000	459,0000

Рис. 6.42. Матриця передбачених та спостережених частот вихідних значень для тестової множини

Матриця класифікації містить кількість спостережень, які були класифіковані коректно (головна діагональ матриці), і ті, які класифікували неправильно. Підсумкова модель може правильно передбачити кредитну належність з 96,4% точності (185 + 459) / (185 + 10 + 14 + 459). Основна мета аналізу – зменшення частки «непривабливих» позичальників, передбачених як «привабливі».

Відсоток правильного передбачення для категорії «непривабливих» клієнтів становить 94,87%.

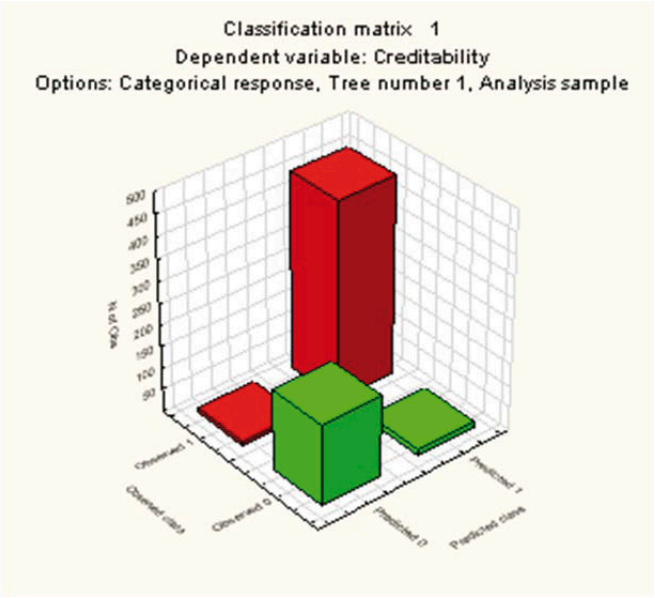


Рис. 6.43. 3D-діаграма матриці класифікації

Порівняльне оцінювання моделей

Більш ефективно провести експерименти з різними методами в процесі видобутку або моделювання даних, ніж покладатися на один із цих методів. Різні інструменти допоможуть зрозуміти проблему загалом або верифікують попередні висновки.

Карта підсилення

Карта підсилення (gain chart, карта виграшів) забезпечує візуалізацію результатної корисної інформації, отриманої за однією чи більше статистичними моделями. Карта підсилення унаочнює виграш при прогнозуванні з використанням статистичних моделей порівняно із застосуванням тільки основної статистичної інформації (лише кількість відгуків у звичайній вибірці).

Карти підсилення (рис. 6.44) будують для множинних прогнозних моделей на основі навчальних моделей за допомогою процедури Compute Overlaid Lift Charts системи «Statistica Data Miner» (рис. 6.40).

Ця карта демонструє, що модель Boosting Trees with Deployment (дерева класифікації та регресії) є найкращою серед доступних моделей для прогнозування результатів. Враховуючи два верхніх дециля (2/10), можна передбачити правильно приблизно 40% спостережень у групі «непривабливих». Можна також зробити висновок, що основна модель є мірою для вимірювання корисності відповідних класифікаційних моделей.

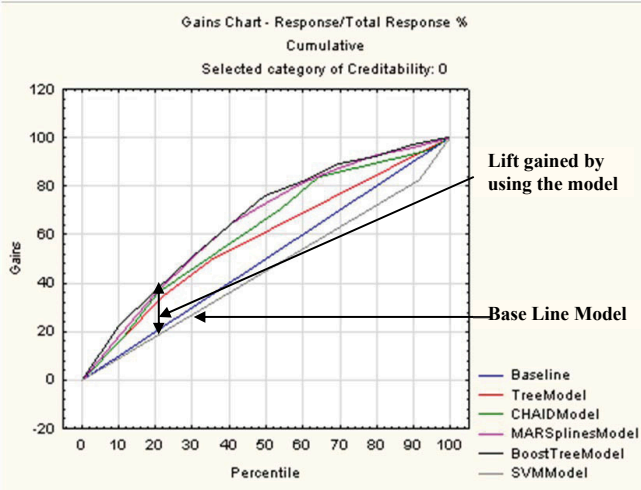


Рис. 6.44. Карта підсилення для Creditability = 0

Ліфтова карта

Ліфтова карта також демонструє перевагу прогнозних моделей порівняно з основною статистичною моделлю, тобто відсотком заданого відгуку в основній вибірці. По вертикальній осі вказується відношення правильно передбачених відгуків до кількості відгуків в основній моделі для заданого процентиля (відношення підсилюючого чи ліфтового значення, пов'язаного з використанням конкретної моделі).

Вибір моделі для прогнозу

На останньому етапі вибирають найкращу модель і застосовують її до нових даних для прогнозування «привабливих / «непривабливих» клієнтів. Найкращою вважають модель, яка продемонструвала найкращу точність на тестових даних порівняно з іншими моделями.

Для подальшого використання прогнозних моделей у «Statistica» потрібно зберегти PMML-код отриманого оптимального моделювання. Далі його можна використовувати у «Statistica Data Miner» для прогнозування / класифікації кредитного ризику нової групи претендентів. Далі передбачені / класифіковані претенденти можуть бути відсортовані за ймовірністю, отриманою при прогнозуванні, що дасть можливість завчасно прийняти рішення про відмову у видачі споживчого кредиту.

6.6. Автоматична класифікація текстів

У сучасному світі близько 80% всієї інформації створюється та зберігається у текстовому форматі. У таких даних прихована значна частка інформації, оскільки неструктурованість текстів не дає можливості повною мірою застосовувати алгоритми data mining. Для вирішенням цієї проблеми використовують методи інтелектуального аналізу неструктурованих текстів. Text mining, або text analytics, – це пошук нових чи існуючих фактів шляхом застосування методів обробки природної мови та статистичного навчання [13]. Методи інтелектуального аналізу неструктурованих текстів виникли та розвиваються на перетині декількох прикладних сфер: інтелектуальний аналіз даних, обробка природних мов, пошук інформації, витягування інформації й управління знаннями. Виявлення знань у тексті – нетривіальний процес пошуку насправді нових, потенційно корисних і зрозумілих шаблонів у неструктурованих текстових даних.

Автоматична класифікація текстів засобами «Statistica Data Miner»

Приклад. 6.7. Провести класифікацію повідомлень, опублікованих наприкінці 2017 р. Всесвітнім економічним форумом – швей-

царською неурядовою організацією щорічних зустрічей у Давосі. На зустрічі запрошують провідних керівників бізнесу, політичних лідерів, видатних мислителів і журналістів. Предметом обговорення є найбільш гострі світові проблеми, включаючи охорону здоров'я і охорону довкілля [100].

Тексти цих повідомлень збережені у файлах з розширенням doc. Документи можуть бути збереженими в інших форматах (наприклад, .pdf, .ps, .txt тощо). Приклад одного з таких файлів зображено на рис. 6.45.

Завдяки способу відбору статей усі документи автоматично були класифіковані на 3 групи: вибори президента США, міграційна криза та штучний інтелект. Мета аналізу полягає в тому, щоб автоматично класифікувати статті, пов'язані з політикою.

Загальні можливості методів, за допомогою яких можна автоматично класифікувати тексти великих обсягів на декілька груп, є дуже широкими. «Statistica» підтримує розподіл результатів текстового видобування. [5].

Immigrants don't harm economies. In fact, the opposite may be true

Immigrants have become a major scapegoat in recent years for sputtering Western economies.

From the U.K.'s jarring "Brexit" from the European Union to Donald Trump's infamous walland more recent proposal to apply "extreme vetting" to those wishing to enter the U.S., many politicians have found success by casting immigrants as a threat to the physical, social and economic welfare of natives.

In short, Americans (and our European brethren) are unhappy, and many are convinced immigration brings harm. A recent poll found that almost two-thirds of Americans think immigration, including the legal kind, "jeopardizes the United States."

While it has become a popular notion in the West that immigrants jeopardize the job prospects of natives, over 30 years of economic research (including my own) gives strong reason to believe otherwise.

And in fact, the opposite may be more likely: There's evidence immigrants actually promote more economic growth.

Рис. 6.45. Фрагмент текстового файла

Щоб помістити дані в таблицю «Statistica», потрібно виконати процедуру Web-Crawling (рис. 6.46).

	1 URLs	2 Root	3 Class	4 Checked files
16	D:\Economic\MEVMINT_PPR\pract\Text_Miner\text\23.doc	D:\Economic\MEVMINT_PPR\pract\Text_Miner\text	no	
17	D:\Economic\MEVMINT_PPR\pract\Text_Miner\text\24.doc	D:\Economic\MEVMINT_PPR\pract\Text_Miner\text	no	
18	D:\Economic\MEVMINT_PPR\pract\Text_Miner\text\25.doc	D:\Economic\MEVMINT_PPR\pract\Text_Miner\text	no	
19	D:\Economic\MEVMINT_PPR\pract\Text_Miner\text\26.doc	D:\Economic\MEVMINT_PPR\pract\Text_Miner\text	no	
20	D:\Economic\MEVMINT_PPR\pract\Text_Miner\text\27.doc	D:\Economic\MEVMINT_PPR\pract\Text_Miner\text	no	
21	D:\Economic\MEVMINT_PPR\pract\Text_Miner\text\28.doc	D:\Economic\MEVMINT_PPR\pract\Text_Miner\text	yes	
22	D:\Economic\MEVMINT_PPR\pract\Text_Miner\text\29.doc	D:\Economic\MEVMINT_PPR\pract\Text_Miner\text	yes	
23	D:\Economic\MEVMINT_PPR\pract\Text_Miner\text\3.doc	D:\Economic\MEVMINT_PPR\pract\Text_Miner\text	yes	
24	D:\Economic\MEVMINT_PPR\pract\Text_Miner\text\30.doc	D:\Economic\MEVMINT_PPR\pract\Text_Miner\text	yes	

Рис. 6.46. Таблиця даних, отримана за допомогою технології Web-Crawling

Змінна *URLs* містить повні посилання на текстові файли даних, у змінній *Root* зберігаються посилання на директорію, в якій є текстові файли.

Додамо до таблиці даних змінну *Class*, яка містить інформацію про те, як класифікували кожен документ (належить чи не належить до групи політика). Крім того, у файлі даних передбачена змінна *Checked files*, яка пізніше буде використана під час крос-перевірки результатної моделі для обчислення її адекватності та точності.

Збережемо таблицю даних з розширенням .data. Завдання аналізу полягає у побудові моделі, яка б автоматично визначала документи, що належать до категорії **політика**.

Далі потрібно вибрати один із файлів, який буде використано як стандартний список стоп-слів (неінформативних). Це список винятків, який містить слова, що не будуть враховані при індексуванні під час обробки документів.

Результати Text Mining (рис. 6.47–6.51).

Term-document frequency matrix

	986 worker	987 world	988 world?	989 worri	990 would	991 write	992 written	993 wrong	994 wrote	995 year	996 yet	997 york	998 you?r	999 young	1000 youtub
48	4	2			1				1						
49	8			1						4				3	
50		4			2			1		2		1			
51							1			1					
52	8	1					1			2	1				1
53	2				27					1					
54	4									3		1		1	
55	1	5	1	2	7					4	1		1	1	
56										1			2		
57	3	2	2		1					6	1				
58							1					2			
59		5	1		2					1					
60				1	5			1	1	1				3	1

Рис. 6.47. Таблиця частот для слів аналізованих текстових файлів

Text Miner results: may be used as Input spreadsheet for subsequent analyses

	1 URLs	2 Root	3 Class	4 Checked files	5 ?a	6 ?but	7 ?i	8 ?if	9 ?the	10 ?this	11 ?we	12 abil	13 abl	14 ac
1	D:\Economic\MEVMNT_	D:\Economic\MEVMNT_	yes											
2	D:\Economic\MEVMNT_	D:\Economic\MEVMNT_	yes										1	
3	D:\Economic\MEVMNT_	D:\Economic\MEVMNT_	yes											
4	D:\Economic\MEVMNT_	D:\Economic\MEVMNT_	yes				1		1	1		1	1	
5	D:\Economic\MEVMNT_	D:\Economic\MEVMNT_	yes							3				
6	D:\Economic\MEVMNT_	D:\Economic\MEVMNT_	yes			1	1	1	1	1	1		3	
7	D:\Economic\MEVMNT_	D:\Economic\MEVMNT_	yes				1							
8	D:\Economic\MEVMNT_	D:\Economic\MEVMNT_	yes											
9	D:\Economic\MEVMNT_	D:\Economic\MEVMNT_	yes										1	
10	D:\Economic\MEVMNT_	D:\Economic\MEVMNT_	yes		1		4	1						
11	D:\Economic\MEVMNT_	D:\Economic\MEVMNT_	yes			2			2		1	3		
12	D:\Economic\MEVMNT_	D:\Economic\MEVMNT_	yes		1							1		1

Рис. 6.48. Початковий файл з частотами появи слів у текстових файлах

Далі потрібно використати процедури вибору та відсіювання змінних для визначення підмножини з 1000 слів, які були виділені для побудови майбутньої моделі.

Результати вибору впливових предикторів за темою політика представлено на рис. 6.49–6.50.

	Best predictors for cate	
	Chi-square	p-value
privat	14,00000	0,000912
topic	12,00000	0,002479
use	16,87481	0,004743
go	11,89300	0,007759
specif	6,42857	0,011230
hous	8,68571	0,012999
low	6,00000	0,014306
propos	6,00000	0,014306
th	6,00000	0,014306
percent	12,00000	0,017351

Рис. 6.49. Критерій χ-квадрат для найкращих предикторів

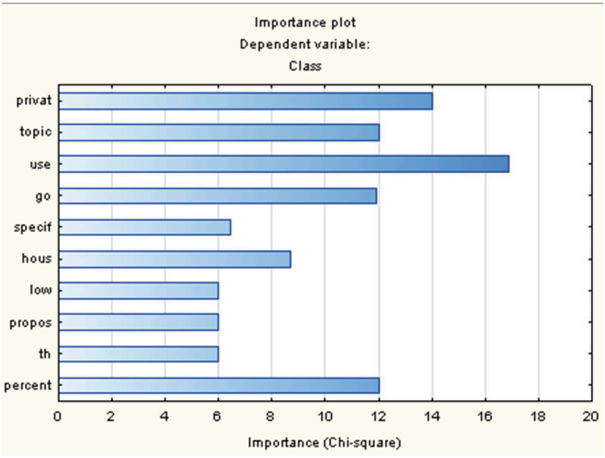


Рис. 6.50. Графік значущості 10 найбільш значущих неперервних предикторів

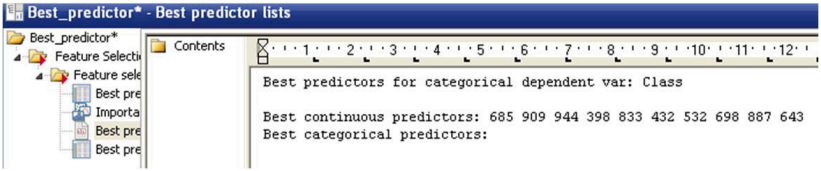


Рис. 6.51. Перелік номерів 10 найбільш значущих предикторів

Інтелектуальний аналіз текстів у середовищі відкритої програмної платформи для інтелектуального аналізу даних «RapidMiner 8.0»

Приклад. 6.8. Провести контент-аналіз передвиборчих меседжів кандидата у президенти США Д. Трампа. Інформаційну базу становлять останні 1869 передвиборчих повідомлень Д. Трампа, які він надіслав своїм потенційним виборцям через соціальну мережу мікроблогів Twitter [99].

Засобами «Statistica Data Miner» виявлено 10 основних предикторів у аналізованих повідомленнях (рис. 6.52). Словами, які найчастіше використовував Д. Трамп у своїх повідомленнях, були такі: Клінтон (Clinton), Обама (Obama), підтримка (support), засоби масової інформації (media), безпечний (safe), демократ (democrat), судження (judgment), небезпека (danger), Америка.

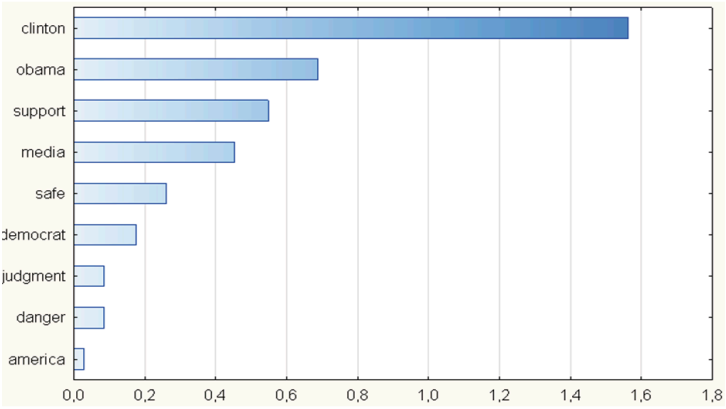


Рис. 6.52. Основні предиктори передвиборчих меседжів Трампа

Для генерування асоціативних зв’язків, які можуть існувати між словами в аналізованих повідомленнях, до неструктурованих

текстових даних послідовно застосовано відповідні процеси та оператори «RapidMiner» (рис. 6.53):

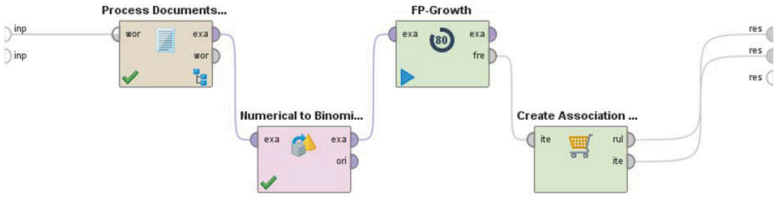


Рис. 6.53. Оператори, застосовані для генерування асоціативних правил

– Process Documents from Files – генерування векторів слів з текстової колекції, що зберігається у різних файлах, за допомогою таких операторів (рис. 6.54):

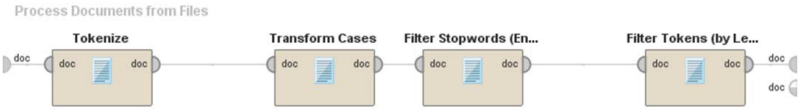


Рис. 6.54. Оператори процесу Documents from Files

- ► Tokenize – розбиття тексту документа на послідовність токенів (лексем);
- ► Transform Cases – приведення всіх символів у текстовому документі у відповідність до одного регістру;
- ► Filter Stopwords – видалення з документа стоп-слів;
- ► Filter Tokens (by Lenght) – фільтрування токенів відповідно до їхньої довжини (кількості символів, які вони містять);

– Numerical Binominal – перетворення числових атрибутів на біноміальні;

– FP-Growth – визначення всіх об'єктів, що часто використовуються;

– Create Associations Rules – генерування кортежу асоціативних правил з аналізованої послідовності частих кортежів.

Асоціативні правила є структурами «якщо, то» (if-then). Для кожної з них обчислюються такі оцінки (рис. 6.55) [5]:

Premises	Conclusion	Support	Confidence	LaPlace	Gain	p-s	Lift	Convicti...
america	make	0.378	0.839	0.950	-0.523	0.203	2.159	3.805
great	make	0.365	0.874	0.963	-0.470	0.203	2.249	4.869
great	america, make	0.365	0.874	0.963	-0.470	0.207	2.314	4.955
great	america	0.371	0.887	0.967	-0.465	0.183	1.971	4.884
make	great	0.365	0.940	0.983	-0.412	0.203	2.249	9.630
make	america, great	0.365	0.940	0.983	-0.412	0.221	2.534	10.408
america, make	great	0.365	0.967	0.991	-0.391	0.207	2.314	17.385
make	america	0.378	0.972	0.992	-0.400	0.203	2.159	19.699
america, great	make	0.365	0.985	0.996	-0.376	0.221	2.534	41.766
great, make	america	0.365	1	1	-0.365	0.201	2.221	∞

Рис. 6.55. Загальні оцінки моделі для генерації асоціативних правил (фрагмент отриманих результатів)

– підтримка (*Support*(*F*)) – відношення кількості аналізованих неструктурованих текстових повідомлень, які містять набір, або кортеж (слово/словосполучення) *F*, до загальної кількості досліджуваних меседжів. Кортеж називають **частим**, якщо значення його підтримки більше за мінімальне значення підтримки, задане користувачем: $Support(F) > Support_{min}$. При пошуку асоціативних правил потрібно знайти множину всіх частих кортежів:

$$L = \{Support(X) > Suppor_{\min}\};$$

– ймовірність (Confidence) – це умовна ймовірність того, що у меседжі буде міститись набір *X* (Premises), якщо у цьому текстовому документі присутній кортеж *Y* (Conclusion), виражена як $\{X \to Y\}$;

$$Confidence\{X \to Y\} = \frac{Support(X,Y)}{Support(X)};$$

– ймовірність Лапласа (*LaPlace*) – отримані оцінки побудованої моделі для генерування асоціативних правил мають високу ймовірність (> 0,95).

– корисність (приріст) інформації (Information gain) – аналізовані інформаційні масиви містять багато атрибутів, значна кількість яких може бути несуттєвою або навіть надмірною для конкретного дослідження. У теорії інформації, започаткованій К. Шенноном, обсяг інформації вимірюється кількістю бітів (0 та 1), необхідних для зберігання або представлення одного символу в повідомленні. Ця теорія розглядає також поняття значення, або «інформаційного контенту», повідомлень. Такий підхід передбачає мінімізацію кількості атрибутів текстів, необхідних для ідентифікації окремого кортежу. Атрибут з найвищою корисністю інформації (Gain) використовують для генерації правила асоціації. За значенням приросту інформації цього атрибута оцінюють корисність інформації інших атрибутів повідомлень [5]:

$$Gain(A) = info(D) - infoA(D),$$

де D – масив текстових даних, $info(D)$ – обсяг масиву D, виміряний у бітах, A – атрибут, $infoA(D)$ – кількість інформації, яку містить атрибут A (у бітах);

– частка кортежу в інформаційному масиві (sample proportion, *p-s*);

– підвищення інтересу (*Lift*) – відношення частоти появи кортежу X у текстових повідомленнях, які містять обидва кортежі X та Y, до частоти появи кортежу Y загалом у всіх аналізованих текстових повідомленнях:

$$Lift\{X \to Y\} = \frac{Support(X,Y)}{Support(X) \times Support(Y)}.$$

Значення *Lift*, що становить більше 1, означає, що кортеж Y з високою ймовірністю буде міститись у текстовому повідомленні, якщо в ньому зберігається кортеж X;

– ступінь наслідування правил (Conviction):

$$Conviction\{X \to Y\} = \frac{1 - Support(Y)}{1 - Confidence\{X \to Y\}}.$$

За допомогою побудованої моделі отримано такі асоціативні правила (рис. 6.56):

```
Association Rules
[america] --> [great, make] (confidence: 0.811)
[america] --> [great] (confidence: 0.823)
[america] --> [make] (confidence: 0.839)
[great] --> [make] (confidence: 0.874)
[great] --> [america, make] (confidence: 0.874)
[great] --> [america] (confidence: 0.887)
[make] --> [great] (confidence: 0.940)
[make] --> [america, great] (confidence: 0.940)
[america, make] --> [great] (confidence: 0.967)
[make] --> [america] (confidence: 0.972)
[america, great] --> [make] (confidence: 0.985)
[great, make] --> [america] (confidence: 1.000)
```

Рис. 6.56. Згенеровані асоціативні правила (фрагмент отриманих результатів)

Отже, наприклад, з імовірністю 0,985 можна стверджувати, що при появі у текстовому повідомленні набору слів {America, great} у цьому меседжі буде присутнім і кортеж {make}. А наявність у тексті кортежу {great, make} забезпечує вірогідність (ймовірність = 1) появи у повідомленні слова «America». Отримані асоціативні правила дали змогу виявити у масиві з 1869 неструктурованих текстових повідомлень Трампа у Twitter його основне передвиборче гасло: «Make America great again» («Зробимо Америку величною знову»). Ця фраза ідентифікує всіх американців як єдиний народ, націю понад націями. Д. Трамп використав її як дієвий інструмент психологічного нав'язування, що змусило багатьох людей зробити несвідомий вибір.

Граф побудованої моделі представляє не лише значення підтримки та ймовірності для кожного із згенерованих асоціативних правил, а й унаочнює, як зв'язки різних атрибутів пов'язані між собою (рис. 6.57).

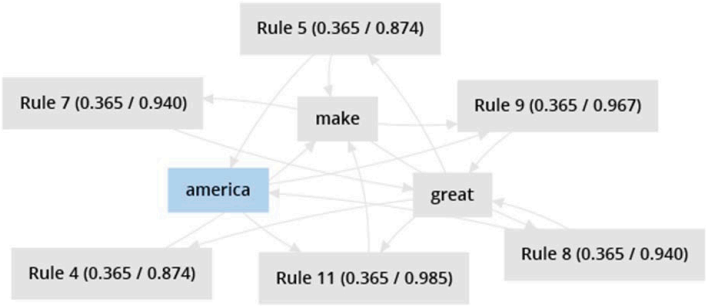


Рис. 6.57. Графічне представлення зв'язків між словами у передвиборчих меседжах Д. Трампа (фрагмент результатів)

Сучасні методи data mining дають змогу побудувати модель, яка здатна допомогти управлінцям не лише радикально покращити фінансове та ринкове становище компанії, а й виявити неочевидні закономірності, що можуть змінити розстановку політичних сил в окремо взятій країні чи зміцнити позиції окремої держави на міжнародній арені.

ПИТАННЯ І ЗАВДАННЯ ДЛЯ САМОПІДГОТОВКИ ТА САМОКОНТРОЛЮ

Теоретичні запитання

1. Сутність і мета інтелектуального аналізу даних.
2. Визначених поняття «інтерактивне буріння».
3. Основні можливості інтерактивного буріння.
4. OLAP-технологія.
5. Багатовимірна модель даних.
6. Вимір. Представлення виміру у вигляді ієрархічної структури.
7. Операції над гіперкубом: зріз, обертання, консолідація і деталізація.
8. Завдання пошуку асоціативних правил.
9. Поняття про дерева класифікації.
10. Матриця та діаграма класифікації.
11. Основні поняття нейромереж.

12. Навчання нейронної мережі.
13. Види аналізу, які реалізують засобами нейромереж.
14. Поняття «скориг».
15. Дерево рішень. Матриця класифікації.
16. Основні поняття та методи Text Analytics.
17. Завдання Text Mining.
18. Етапи аналізу текстових документів.
19. Прийоми видалення неінформативних слів у Text Mining.
20. Основні результати інтелектуального аналізу текстів.

Практичні завдання

Завдання 1. Провести власні дослідження (зібрати інформацію) для групи країн за вибором (не менше 20 спостережень, наприклад, країни ЄС) щодо політики, яку вони проводять щодо резонансних подій, явищ або процесів міжнародного значення (не менше 8 змінних). Наприклад, ставлення кожної з країн до війни на сході України, до анексії Криму, до Brexit, до політики Д. Трампа, до економічних санкцій США проти Росії (введення/посилення), до випробування ядерної зброї КНДР і т. ін.

Усі змінні є категоріальними і мають однаковий набір можливих значень. Наприклад, усі змінні можуть містити лише такі варіанти: підтримує, не підтримує, надає допомогу, готова надати допомогу, дружнє ставлення, є прихильником, толерантне ставлення і т. ін.

Створити файл даних за таким зразком:

	Ставлення до Brexit	**Ставлення до анексії Криму**	**…**	**Ставлення до посилення економічних санкцій США проти Росії**
Країна 1	підтримує	не підтримує		готова надати допомогу
Країна 2	толерантне ставлення	надає допомогу		не підтримує
…	…	…	…	…
Країна 20	є прихильником	готова надати допомогу		є прихильником

Провести аналіз змінних методом інтерактивного буріння. Пояснити отримані результати.

Завдання 2. За значеннями змінних 1–10 за 2005–2017 рр. та значеннями змінних 11–20 за 2005–2017 рр. для 40 країн світу визначити можливі залежності між змінними, використовуючи аналіз асоціацій.

Варіанти завдань

Варіант	Змінні 1–10	Змінні 11–20
1	ІЛР (2008–2017 рр.)	Індекс освіти (2008–2017 рр.)
2	ІЛР (2008–2017 рр.)	Ймовірна тривалість життя при народженні (2008–2017 рр.)
3	ІЛР (2008–2017 рр.)	Валовий національний дохід на душу населення (2008–2017 рр.)
4	ІЛР (2008–2017 рр.)	Середня тривалість навчання дорослого населення (2008–2017 рр.)
5	ІЛР (2008–2017 рр.)	Очікувана тривалість навчання дітей шкільного віку (2008–2017 рр.)
6	Населення (2008–2017)	Кількість дітей до 5 років (2008–2017 рр.)
7	Населення (Population) (2008–2017 рр.)	Кількість людей після 65 років (2008–2017 рр.)
8	ВНП на душу населення (2005–2017 рр.)	Загальний експорт–імпорт (2008–2017 рр.)
9	ВНП на душу населення (2008–2017 рр.)	Мінімальна зарплата (2008–2017 рр.)

Приклад файла даних:

Країни	ІЛР 2008	…	ІЛР 2017	ІО 2008	…	ІО 20017
Україна						
Австрія						
…						
Швеція						

Завдання 3. Провести класифікацію 40 країн світу (за варіантом), самостійно обравши номінальну залежну та не менше 6 незалежних (як числових, так і номінальних) змінних.

Провести інтерпретацію отриманих результатів

Завдання 4. Використовуючи автоматизовані нейронні мережі «Statistica», побудувати оптимальну регресійну модель на основі даних файла data.xls (за варіантом).

Дослідження провести для всіх країн, для яких відомі значення таких показників (відповідно до варіанта):

№ варіанта	Залежна змінна	Незалежні змінні
1	2	3
1	Gini Index	Life expectancy at birt, Mean years of schooling, GDP per capit, Ranked of terrorism, Trade Index
2	Gini Index	GDP per capita, Trade Index, Availability and quality of transport infrastructure, ICT, Social Progress Index
3	Gini Index	Access to Basic Knowledge, Health and Wellness, Personal Freedom and Choice, Tolerance
4	Gini Index	Global Competitiveness Index, economic freedom, The Global Peace Index, Population, Migrants
5	Gini Index	GDP per capit, ICT, economic freedom, Social Progress Index, Migrants
6	The Global Peace Index	Gini Index, Life expectancy at birt, GDP per capit, Ranked of terrorism, Trade Index
7	The Global Peace Index	GDP per capit (current international $), ICT, economic freedom, Ranked of terrorism, Migrants
8	The Global Peace Index	Gini Index, GDP per capita, Trade Index, ICT, Social Progress Index
9	The Global Peace Index	Global Competitiveness Index, economic freedom, The Global Peace Index, Population, Migrants
10	The Global Peace Index	GDP per capit, ICT, economic freedom, Social Progress Index, Migrants
11	Social Progress Index	Life expectancy at birt, Mean years of schooling, GDP per capit, Ranked of terrorism, Trade Index
12	Social Progress Index	GDP per capita, Trade Index, Availability and quality of transport infrastructure, ICT, HDI
13	Social Progress Index	Access to Basic Knowledge, Health and Wellness, Personal Freedom and Choice, Tolerance
14	Social Progress Index	Global Competitiveness Index, economic freedom, The Global Peace Index, Population, Migrants
15	Social Progress Index	GDP per capit, ICT, economic freedom, HDI, Migrants
16	Global Competentitiveness Index	Gini Index, Life expectancy at birt, GDP per capit, Ranked of terrorism, Trade Index

1	2	3
17	Global Competentitiveness Index	GDP per capit, ICT, economic freedom, Ranked of terrorism, Migrants
18	Global Competentitiveness Index	Gini Index, GDP per capita, Trade Index, ICT, Social Progress Index
19	Global Competentitiveness Index	HDI, economic freedom, The Global Peace Index, Population, Migrants
20	Global Competentitiveness Index	GDP per capit, ICT, economic freedom, Social Progress Index, Migrants
21	The Global Peace Index	Life expectancy at birt, Mean years of schooling, GDP per capit, Ranked of terrorism, Trade Index

Провести інтерпретацію результатів.

Завдання 5 (для групи з трьох студентів). Сформувати інформаційну базу, яка містить не менш ніж по 20 текстових файлів (завершених за змістом; які *не є частинами одного документа*; на обсяг обмеження не накладаються) за трьома темами, пов'язаними з МВ.

Провести класифікацію шести «нових» текстів (по 2 для кожного студента групи), для яких відомо, що кожен з них належить до одного з попередньо виділених класів (за тематикою) документів.

Перелік рекомендованих тематик

1. Євросепаратизм.
2. Від'єднання Каталонії від Іспанії.
3. «Цифрове» втручання Росії у Brexit
4. «Цифрове» втручання Росії у вибори Д. Трампа.
5. Політика Д. Трампа.
6. Технології big data (використання у соцмережах).
7. Економічні санкції проти Росії.
8. Випробування ядерної зброї КНДР.
9. Зміни в ЄС: Brexit та можливі нові члени-учасники.
10. Гібридні війни.
11. Мовне питання в Україні.
13. Фейкові новини.

РОЗДІЛ 7

МОДЕЛЮВАННЯ ПРОЦЕСУ ПРИЙНЯТТЯ ОПТИМАЛЬНИХ РІШЕНЬ У МВ

7.1. Теорія ігор як інструмент дослідження міжнародних відносин

При вирішенні завдань МВ в умовах тенденції до глобалізації часто аналізуються ситуації, в яких прагнуть реалізувати власні інтереси дві чи більше конкуруючих сторін, що мають різні, здебільшого протилежні, цілі. Такі ситуації називають конфліктними. Побудовою математичних моделей для дослідження конфліктних ситуацій займається теорія ігор.

Вона виникла у 40-х рр. XX ст. як прикладний аспект математичного моделювання і сьогодні займає важливе місце серед сучасних теорій та концепцій міжнародних відносин. Основоположниками теорії ігор є Дж. фон Нейман, Д. Льюїс, Дж. Неш, Л. Шеплі, М. Шубік та ін. [13, с. 114]. З 1944 р. теорія ігор застосовується не лише в математиці, а й в політичній та економічній сферах. Переважно її застосовували для вивчення проблематики військово-стратегічної галузі, зокрема для аналізу гонки озброєнь у 1950–1970-х рр. (аналіз відносин США та колишнього СРСР).

Серед засновників сучасної теорії ігор виокремимо А. Сміта, Д. Г'юма, К. Менгера, Дж. Мілля, М. Вебера, К. Поппера та ін. [49, с. 137]. Сьогодні теорію ігор застосовують для аналізу та прогнозування політичних, соціальних, біологічних й економічних явищ і процесів. Наприклад, для моделювання результатів виборчих кампаній, пошуку оптимальних рішень при вирішенні військових і політичних конфліктів та міжнародних економічних суперечок. Парадигму теорії ігор використовують сучасні науки, серед яких соціологія, інформатика, статистика, економетрія, управління і т. ін.

У науковій літературі подано багато різних визначень теорії ігор. Її трактують як теорію раціональної поведінки людей, інтереси яких не збігаються [73, с. 17], як «науку про стратегічне мислення» [80, с. 34], як «…теорію математичних моделей для прийняття оптимальних рішень в умовах конфлікту» [16, с. 16] або як «…розділ прикладної математики, який досліджує моделі прийняття рішень в умовах незбігання інтересів сторін (гравців)» [25, с. 68].

На сьогодні не існує загальноприйнятого визначення поняття «теорія ігор». Актуальним залишається також питання подолання міждисциплінарних бар'єрів та розробки єдиної, цілісної теорії прийняття рішень, елементи якої можуть бути застосовані у різних конфліктних ситуаціях [50, с. 163].

З точки зору МВ теорію ігор можна розглядати як моделювання раціональної поведінки декількох гравців (акторів міжнародної арени), які поділяють спільний ризик у разі прийняття неефективного рішення. Водночас рішення кожного з них залежить від рішень, прийнятих іншими гравцями. Логіка гравців та їх рішення невідомі іншим учасникам гри.

Сьогодні теорію ігор активно використовують для дослідження взаємодії акторів політичної арени, зокрема для пошуку оптимальних рішень (вибору можливих стратегій поведінки політичних гравців) в умовах невизначеності (нестабільної політичної обстановки), та як потужний апарат вивчення суспільно-економічних процесів [57, с. 6].

Традиційно у теорії ігор визначають дві великих групи ігор: **ігри з нульовою сумою, або антагоністичні** (виграш однієї сторони еквівалентний програшу іншої), та **кооперативні ігри** (передбачають формування коаліцій) [17, с. 115]. У політиці та економіці ігри з нульовою сумою застосовують рідко. Одним із небагатьох прикладів є моделювання президентських виборів С. Брамса [76]. Окремі фахівці у сфері міжнародних відносин вважають недоліком методу гри з нульовою сумою стратегічний характер теорії ігор та неможливість описати динаміку гри.

Реальні міжнародні процеси моделюють за допомогою ігор з ненульовою (змінною) сумою, в яких учасники можуть досягти обопільних вигод чи обопільно програти. Це можна пояснити, зокрема, відсутністю достовірної, актуальної та повної інформації про стратегію партнера або відставанням від розвитку подій одного із гравців і т. ін.

Кооперативні ігри припускають можливість прийняття всіма гравцями обопільного рішення (розподілу виграшу), на який погодяться всі учасники. Прикладом застосування теорії коаліційних ігор у політиці є ціна Шеплі, яка визначає відносну оцінку сили кожного із гравців у ситуації, що склалася [93]. Її застосовують для вивчення різних політичних проблем, зокрема для визначення відносної сили кожного зі штатів під час виборів президента США [87], дослідження зваженої виборчої системи [91], з'ясування ефекту «вагона, що віддаляється» (ефекту приєднання до більшості в процесі формування коаліції) [76], аналізу прийняття рішень у Раді безпеки ООН [90]. Для обчислення відносної величини сили кожного з гравців при розв'язанні коаліційних ігор використовують також індекс Баншафа [75, с. 319].

Коаліційні ігри застосовують для вироблення стратегій ведення переговорів та при дослідженні розвитку політичних процесів у коаліціях [17, с. 116]. Для формалізації процесів формування міжнародних об'єднань, відображення діяльності політичних партій та опису основних принципів формування урядів держав з парла-

ментською формою правління використовують модель В. Райкера кооперативної гри *n* осіб з нульовою або скінченною сумою та трансферабельною корисністю, яка передбачає лише мінімальні виграші коаліцій [91].

У випадках, коли різні сторони політичних чи економічних суперечок не можуть досягти спільної згоди, використовують некооперативні ігри. Одним із прикладів є «рівновага за Нешем». Для вирішення ситуації в грі використовують стан рівноваги учасників (жоден із гравців в односторонньому порядку не може поліпшити свій результат). Найбільш відомою некооперативною грою вважають гру двох осіб з ненульовою сумою («дилема в'язня»), яку застосовують для вивчення політичних ситуацій з чітко вираженою конфронтацією та тенденцією до співпраці (змішані мотиви поведінки). Гравці, що приймають рішення, не мають можливості координувати свої дії [74].

Для дослідження політичних процесів за умов неповноти наявних даних, якими характеризуються сучасні міжнародні відносини, використовують концепцію ігор з неповною інформацією [84, с. 160] в умовах невизначеності [78]. Однак за наявності значної кількості факторів застосування простих ігор для дослідження складних політичних процесів є недоцільним, оскільки ймовірність отримання значущих результатів у таких ситуаціях невисока.

Перераховані вище концепції передбачають можливе кількісне вираження раціональної поведінки людей (здатність гравця усвідомлювати свої цілі та обирати для їх досягнення найбільш ефективні з можливих варіантів дій). Для виконання цієї умови вподобання раціонального суб'єкта мають підпорядковуватись чіткій ієрархії, що майже неможливо, оскільки в житті кожної людини є багато цілей і не завжди вдається виокремити одну, пріоритетну. На практиці для окремого індивіда цей постулат не виконується ніколи.

Теорія ігор має декілька основних моделей, які відображають різноманітність раціональної поведінки людей. У міжнародних відносинах, як і в інших соціальних системах, гравець (раціональ-

ний учасник) робить вибір з обмеженої множини варіантів (стратегій). Йому достатньо усвідомлювати пріоритетність цілей лише у своїй сфері діяльності (МВ) та обирати власні стратегії для реалізації найвигіднішого сценарію саме в цій галузі. Для формування множини варіантів, з яких відбувається раціональний вибір, використовують уявлення та постулати теорій і концепцій міжнародних відносин.

Інтереси держави чи інших інституцій не завжди збігаються з інтересами окремих осіб, які є їх представниками. Крім відносно обмеженої кількості цілей соціальних інституцій, в їхній діяльності менше мінливості та невизначеності. У теорії ігор проблема невизначеності поведінки людини вирішується орієнтацією на найбільш несприятливий сценарій.

Пріоритети соціальних інституцій, переважно держав, вивчають різні теорії міжнародних відносин. Зокрема, реалісти вважають, що державам вигідніше прагнути відносних переваг над партнером. Інституціоналісти спонукають до співробітництва, що сприяє збільшенню власних здобутків. У цьому сенсі всі теорії міжнародних відносин, за винятком постпозитивістських, базуються на гіпотезі про раціональну поведінку міжнародних акторів. Не досліджуючи закономірності та особливості такої поведінки, ці теорії пропонують свої системи ієрархії для раціонального гравця.

Постулати про раціональність поведінки людей не є тривіальними для дослідження міжнародних відносин. Прагнення обмежити та підпорядкувати системі дії індивідуальних акторів є обґрунтованим, оскільки більшість наукових проблем міжнародних відносин не стосується ймовірної поведінки окремих індивідів.

Суть теорії ігор щодо дослідження МВ полягає у виборі оптимальної стратегії раціональної поведінки в конфліктних ситуаціях. Метою теорії ігор є визначення найкращого з можливих варіантів стратегій для кожного гравця [41, с. 456].

Стратегія гравця – сукупність домовленостей, які обмежують вибір його дій на кожному етапі залежно від ситуації, що скла-

лася. Конфлікт з точки зору теорії ігор – це ситуація, в якій зіштовхуються протилежні інтереси двох чи більше гравців, що прагнуть досягти різних цілей. Кожна зі сторін-учасників конфліктної ситуації може впливати на хід подій, але не має можливості повністю ним управляти. Конфліктні ситуації виникають під час вирішення різних політичних та економічних, зокрема міжнародних, проблем (дотримання міжнародних угод, розподіл ресурсів, конкуренція за ринки збуту і т. ін).

У процесі гри її учасники здійснюють ходи. **Ходом** гравця називають реалізацію однієї з вибраних ним дій, передбачених правилами. **Грою** називають формалізовану спрощену модель конфліктної ситуації, яку використовують для її вивчення. Теорія ігор визначає для різних типів конфліктних ситуацій оптимальні стратегії раціональної поведінки учасників – стратегії гравців, що забезпечують рівновагу у грі. Оптимальні стратегії гравців забезпечують кожному з них конкретний виграш. Виконання гравцем ходу, не передбаченого узгодженою стратегією, може тільки зменшити його виграш.

Ігри розрізняють за такими ознаками:

– кількість учасників;

– характеристики платіжних функцій, які визначають виграш кожного гравця залежно від його поведінки та поведінки інших учасників конфлікту;

– інформація про ситуацію, що склалася, відома іншим гравцям;

– правила, що обмежують вибір стратегії поведінки учасників;

– можливості укладання угод між різними сторонами конфлікту і формування коаліцій;

– розуміння понять «рівновага» та «справедливе вирішення гри».

Прикладом використання теорії ігор для дослідження економічних конфліктних ситуацій є побудова математичних моделей типових для ринкової економіки явищ конкуренції. Найпростіший варіант – боротьба двох конкурентів за ринок збуту (парна гра). У

складних іграх беруть участь багато суперників, які вступають при цьому між собою в постійні або тимчасові коаліції (множинна, або коаліційна гра).

Ходи гравців поділяють на два види: індивідуальні (свідомий вибір гравцем одного з можливих варіантів дій) та випадкові (випадково вибрана дія). Розв'язком гри є вибір кожним із гравців оптимальної стратегії – один із гравців отримує максимальний виграш, тоді як інший дотримується своєї стратегії. Такі стратегії називають **оптимальними**. Для оптимальних стратегій виконується умова стійкості: відмова кожного із гравців від своєї стратегії є збитковою для нього.

Якщо гра повторюється багато разів, гравці можуть бути зацікавленими не у виграші (програші) в кожній конкретній партії, а в середньому виграші (програші) в усіх партіях.

Науковці виокремлюють стратегічну та статистичну теорії ігор [41, с. 460].

Характеристики стратегічної теорії ігор:

► активні дії обох учасників;

► дії обох гравців є раціональними з погляду їхніх інтересів;

► повна невизначеність у виборі стратегії кожною із сторін;

► обидва гравці діють на підставі детермінованої інформації, визначеної матрицею витрат.

Характеристики статистичної теорії ігор:

– природа «нерозумна» (перший гравець) є неактивним учасником і не протидіє максимальному виграшу другого гравця;

– статистик (другий гравець) намагається виграти гру в уявного противника (природи);

– часткова невизначеність у виборі стратегії;

– природа розвивається і діє відповідно до об'єктивно існуючих законів;

– наявність можливості у статистика поступового вивчення законів, зокрема на основі статистичного експерименту.

Із позиції ігрового підходу експеримент у міжнародних відносинах пов'язаний з моделюванням та науково обґрунтованою «симуляцією» конкретних ситуацій та процесу прийняття рішень. Теорія ігор, зокрема, моделює вибір сторонами стратегій поведінки та результати їхніх взаємодій. Така модель умовно розглядає окремі міжнародні ситуації, гіпертрофуючи вплив окремого фактора або навпаки, абстрагуючись від нього. Це дає змогу експериментувати з конкретними факторами розвитку ситуацій.

Приклад 7.1. ТНК займається виготовленням виробів з хутра, реалізація яких залежить від природних кліматичних умов, зокрема погодних (табл. 7.1). Визначити оптимальну стратегію реалізації продукції ТНК, спрямовану на отримання максимального прибутку.

Таблиця 7.1

Основні виробничі показники ТНК (умовні дані)

Вид продукції	Собівартість одиниці, дол. *(c)*	Ціна, дол. *(p)*	Обсяг реалізації, од. (*q*)	
			Тепла погода	Холодна погода
Виріб № 1	1800	3200	1200	500
Виріб № 2	1000	2000	1000	2500

Перший гравець – природа (попит на продукцію). Стратегії першого гравця (множина станів природи):

$$v = (v_1, v_2),$$

де v_1 – попит на продукцію ТНК у теплу погоду, v_2 – попит на продукцію у холодну погоду.

Другий гравець – статистик (ТНК). Стратегії другого гравця (можливі варіанти розвитку подій):

$$\alpha = (\alpha_1, \alpha_2),$$

де α_1 – обсяги реалізації продукції за теплої погоди, α_2 – обсяги реалізації продукції за холодної погоди.

Другий гравець займає активну позицію і може оцінювати наслідки кожного варіанта реалізації продукції залежно від стану природи.

Аналітична функція прибутку:

$$P(v, a) = q(p - c).$$

Стратегії гравців:

– очікуваний прибуток в умовах реалізації стратегії природи v_1 (попит на продукцію ТНК у теплу погоду):

$$P = 1200(3200 - 1800) + 1000(2000 - 1000) = 2\,680\,000 \text{ дол.};$$

– очікуваний прибуток при виконанні стратегії природи v_2 (попит на продукцію в прохолодну погоду):

$$P = 500(3200 - 1800) + 2500(2000 - 1000) = 3\,200\,000 \text{ дол.};$$

– якщо ТНК обере стратегію α_1 (обсяги реалізації продукції за теплої погоди), вона зможе продати всі вироби № 1 і тільки частину виробів № 2 (100 із 2500 од.) та отримати прибуток:

$$P = 1200(3200 - 1800) + 1000(2000 - 1000) - (2500 - 1000) \times \\ \times (2000 - 1000) = 1\,180\,000 \text{ дол.};$$

– якщо ТНК обере стратегію α_2 (обсяги реалізації продукції за холодної погоди), вона продасть усі вироби № 2 і частину виробів № 1 (500 із 1200 од.) та отримає прибуток:

$$P = 500(3200 - 1800) + 2500(2000 - 1000) - (1200 - 500) \times \\ \times (3200 - 1800) = 2\,220\,000 \text{ дол.}$$

Побудуємо матрицю гри, або платіжну матрицю (табл. 7.2).

Таблиця 7.2

Матриця прибутку *P*(*v*, *a*) для прикладу 7.1 (*у* дол.)

	v_1	v_2	min за рядками
α_1	2 680 000	1 180 000	1 180 000
α_2	2 220 000	3 200 000	2 220 000
max за стовпцями	2 680 000	3 200 000	–

Із платіжної матриці визначимо, що перший гравець «природа» ні за яких варіантів розвитку подій не отримає прибуток, що менший, ніж 1 180 000 дол. Проте за збігу погодних умов з обраною стратегією виграш ТНК становитиме 2 680 000 дол. або 3 200 000

дол. Якщо другий гравець (ТНК) буде постійно застосовувати стратегію α_1, а гравець «природа» – стратегію v_2, то виграш зменшиться до 1 180 000 дол.

Тенденція до зменшення прибутків буде спостерігатись і в разі, коли гравець ТНК буде постійно застосовувати стратегію α_2, а гравець «природа» – стратегію v_1.

Максимальний прибуток ТНК може одержати, якщо буде почергово застосовувати стратегії α_1 та α_2. Таку стратегію називають **змішаною**, а її складові α_1 та α_2 – **чистими стратегіями** [41, с. 463].

Оптимізація змішаної стратегії дасть можливість другому гравцю (статистику) завжди отримувати середнє значення виграшу незалежно від стратегії першого гравця (природи).

Позначимо частоту застосування ТНК стратегії α_1 через k. Відповідно частота застосування ним стратегії α_2 становитиме $1 - k$.

Якщо другий гравець застосує оптимальну змішану стратегію, то при використанні першим гравець кожної зі стратегій α_1 та α_2 (тепла та холодна погода відповідно) ТНК отримає однаковий середній прибуток:

$$2\,680\,000\,k + 1\,180\,000\,(1-k) = 2\,220\,000\,k + 3\,200\,000\,(1-k);$$

$$k = \frac{2020000}{2480000} = \frac{101}{124};$$

$$k \approx 0{,}8;$$

$$1-k = 1-\frac{2020000}{2480000} = 1-\frac{101}{124} = \frac{24}{124};$$

$$1-k \approx 0{,}2.$$

Перевіримо цю гіпотезу. При застосуванні стратегії v_1 (тепла погода) гравця «природа» середній прибуток ТНК становитиме:

$$2\,680\,000\,k + 1\,180\,000\,(1-k) \approx 2\,680\,000 \times 0{,}8 + 1\,180\,000 \times 0{,}2 \approx$$
$$\approx 2\,400\,000 \text{ дол.}$$

При реалізації стратегії v_2 (прохолодна погода) гравця «природа» середній прибуток ТНК становитиме:

$$2\,220\,000\,k + 3\,200\,000\,(1-k) \approx 2\,220\,000 \times 0{,}8 + 3\,200\,000 \times 0{,}2 \approx$$
$$\approx 2\,400\,000 \text{ дол.}$$

Отже, гравець ТНК, застосовуючи чисті стратегії α_1 та α_2 у відношенні 101 до 24 (приблизно 4 до 1), матиме оптимальну змішану стратегію, що забезпечить йому у будь-якому разі середній прибуток у розмірі 2 400 000 дол. Середній платіж, який одержують при реалізації оптимальної стратегії, називають **ціною гри**.

Кількість виробів № 1 та виробів № 2, які ТНК має випускати для одержання максимального прибутку, становить:

(1200 виробів № 1 + 1000 виробів № 2) × 0,8 + (500 виробів № 1 + 2500 виробів № 2) × 0,2 = 1060 виробів № 1 + 1300 виробів № 2.

Оптимальна стратегія – виробництво 1060 виробів № 1 та 1300 виробів № 2, при виборі якої ТНК отримає середній прибуток у розмірі 2 400 000 дол.

Теорію ігор застосовують для вивчення політичних та економічних явищ і процесів як на макрорівні при розробці математичних моделей, в яких враховують інтереси різних держав чи галузей економіки, так і на рівні підприємства для вибору оптимальних рішень при вирішенні питань дефіциту економічних ресурсів, відносин з контрагентами, розподілу прибутків і т. ін. Суттєвим обмеженням такого підходу є використання єдиного критерію – виграшу (прибутку) як характеристики ефективності, тоді як при вирішенні більшості економічних завдань традиційно застосовують декілька таких показників.

Приклад 7.2. Оман та Ємен є єдиними у світі виробниками сирої нафти, попит на яку обчислюється за функцією:

$$p(x) = 14 - x,$$

де p – ціна в сотнях євро, x – загальна кількість сирої нафти в мільйонах барелів. Витрати на виробництво x_i млн. барелів $i = (0, j)$ становлять:

$$C(x) = 2x_i.$$

Припустимо, що кожна країна може виробляти три, чотири або шість мільйонів барелів.

Побудувати матрицю виплат та визначити сальдо Неша в грі.

Розв'язання

Обчислимо прибутки гравців, визначивши виграш:

$$\Pi(p, x_0, x_j)_i = p(x) \times x_i - C(x_i),$$

$$p(x) = p(x_0 + x_j),$$

$$\pi_0 = [14 - (3 + 6)] \times 3 - 2 \times 3 = 9,$$

$$\pi_j = [14 - (3 + 6)] \times 6 - 2 \times 6 = 18.$$

Отримаємо матрицю прибутку (платіжну матрицю) (рис. 7.1, 7.2).

		Ємен		
		3	4	6
Оман	3			9,18
	4			
	6			

Рис. 7.1. Фрагмент матриці прибутку

Знайдемо оптимальні стратегії (підкреслено) кожного гравця на всі допустимі стратегії його опонента. Пари стратегій з обома підкресленими виплатами відображають залишки Неша (рис. 7.2).

		Ємен		
		3	4	6
Оман	3	18,18	15,20	9,18
	4	20,15	16,16	8,12
	6	18,9	12,8	0,0

Рис. 7.2. Матриця прибутку для прикладу 7.2

Можливі рішення для Оману:

► якщо Ємен виробляє 3 млн. барелів нафти, то найкраща відповідь Омана – 4 млн. барелів, оскільки 20 > 18;

► якщо Ємен виробляє 4 млн. барелів нафти, то найкраща відповідь Омана – 4 млн. барелів, оскільки 16 > 15 і 16 > 12;

► якщо Ємен виробляє 6 млн. барелів нафти, то найкраща відповідь Омана – 3 млн. барелів, оскільки 9 > 8 і 9 > 0.

Можливі рішення для Ємену:

– якщо Оман виробляє 3 млн. барелів нафти, то найкраща відповідь Ємену – 4 млн. барелів, оскільки 20 > 18;

– якщо Оман виробляє 4 млн. барелів нафти, то найкраща відповідь Ємену – 4 млн. барелів, оскільки 16 > 15 16 > 12;

– якщо Оман виробляє 6 мільйони барелів нафти, то найкраща відповідь Ємену – 3 млн. барелів, оскільки 9 > 8 і 9 > 0.

Існує лише одна рівновага Неша: (4, 4).

Приклад 7.3. Двоє виробників сигарет мають вирішити, чи рекламувати свою продукцію, чи ні. Якщо обидва продукти однаково відомі, частка ринку кожного з них становить 50%. Якщо одна компанія здійснює рекламну кампанію, а інша – ні, то перша збільшує свою частку ринку до 75%.

Загальна кількість проданих сигарет, незалежно від розміру рекламної діяльності обох компаній, становить 40 млн. За кожну продану пачку виробник отримує прибуток у розмірі 1 євро. Припустимо, що рекламна кампанія коштує 5 млн. євро.

1. Визначити баланс Неша в цій грі. Чи буде законна заборона реклами сигарет мати сенс з точки зору виробника?

2. Як зміняться результати, якщо загальний попит на сигарети не є постійним, а збільшується на 10 млн. з кожною рекламною кампанією?

Розв’язання:

$$0{,}25 \times 40 = 10,$$

$$0{,}75 \times 40 - 5 = 25.$$

		Фірма 2	
		рекламувати	не рекламувати
Фірма 1	рекламувати	15, 15	25, 10
	не рекламувати	10, 25	20, 20

Рис. 7.3. Баланс Неша для прикладу 7.3

Якщо обидві фірми продовжать рекламувати сигарети, їх виграш становитиме (25,25). Заборона реклами приведе до виграшу (20,20) (рис. 7.4).

$$0{,}5 \times 60 - 5 = 25,$$

$$0{,}25 \times 50 = 12{,}5,$$

$$0{,}75 \times 50 - 5 = 32{,}5.$$

		Фірма 2	
		рекламувати	не рекламувати
Фірма 1	рекламувати	25, 25	32,5, 12,5
	не рекламувати	12,5, 32,5	20, 20

Рис. 7.4. Баланс Неша для випадку продовження реклами

7.2. Теорія катастроф у вивченні МВ

Термін «катастрофа» означає стрибкоподібні зміни, які виникають при плавних змінах значень параметрів. Теорію катастроф вважали науковим «переворотом» у математиці. Сьогодні моделі теорії катастроф застосовують в економіці, соціології, психології, лінгвістиці, природознавстві, техніці та інших галузях.

Обґрунтованість теорії катастроф істотно залежить від обґрунтованості початкових уявлень. В одних випадках теорії

повністю підтверджуються експериментом. В інших, зокрема у біології, психології та соціальних науках (наприклад, у додатках до теорії поведінки біржових гравців), як початкові передумови, так і висновки мають переважно евристичне значення [3, с. 16].

Найбільш неочікуваною для дослідника є ситуація, в якій невеликі, поступові зміни параметрів ведуть до несподівано різкої, обвальної зміни поведінки системи.

Основні положення теорії катастроф

Простим типом катастрофи є катастрофа «складка» (рис. 7.5).

Передбачається, що система спочатку перебуває в точці *A* на нижній гілці складчастого різноманіття. Зі зростанням змінної *p* змінна *x* збільшується також так, що система переходить через точку *B* і досягає точки *C*. У цій точці змінна *p* перетинає особливість S_1 і система здійснює «катастрофічний» стрибок на верхню гілку різноманіття в точку *C*. Подальше зростання змінної p відводить систему далі за точку *D*.

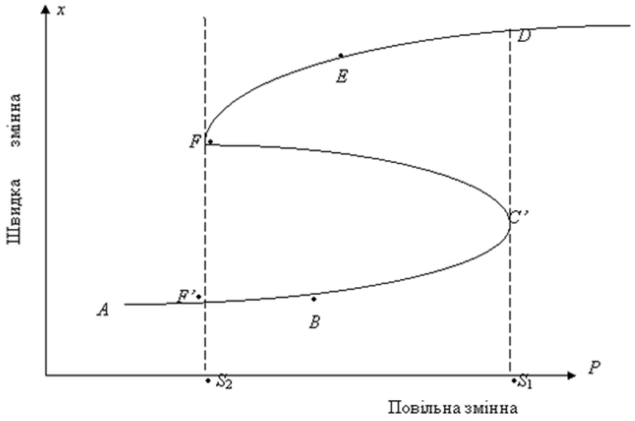


Рис. 7.5. Катастрофа «складка»

Якщо ж значення змінної *p* далі починає зменшуватись, система продовжує рухатись уздовж верхньої гілки різноманіття через точку *E* до точки *F*. У цій точці змінна *p* перетинає особливість

S_2 і система здійснює «катастрофічне» повернення на нижню гілку різноманіття в точку F', після чого подальші зміни змінної p ведуть систему або до точки A, або до точки B доти, поки вона знову не перетне особливість S_1.

Проста катастрофа «складка» ілюструє властивість бімодальності, представлену двома гілками складчастого різноманіття, і властивість розривності, що відображається різкими стрибками з однієї гілки на іншу, в особливостях S_1 і S_2. Гістерезис (явище, за якого поведінка системи істотно залежить від передісторії процесу) ілюструється тим, що траєкторія системи при зменшенні p після перетину особливості відрізняється від траєкторії, якою рухається система при збільшенні p. Конкретна форма функції, що зв'язує x та p на різноманітті, не важлива. Потрібно лише, щоб у проекціях x на p зберігалася особливість типу складки. Найпростіша з еквівалентних функцій, що представляють катастрофу «складка», задається многочленом третього степеня:

$$f(x, p) = -(x^3 - x + p).$$

Однією з найбільш популярних моделей теорії катастроф є катастрофа «зборка» (рис. 7.6).

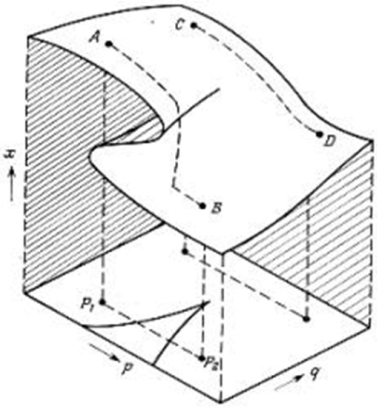


Рис. 7.6. Катастрофа «зборка»

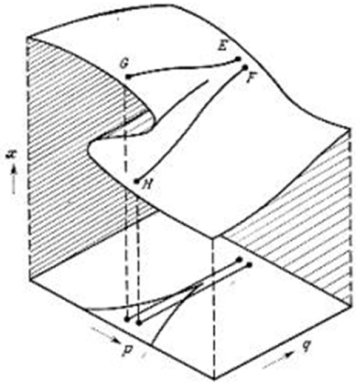
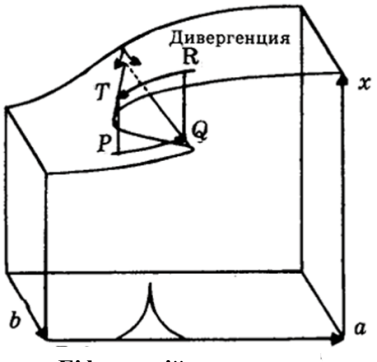


Рис. 7.7. Ілюстрація дивергенції при зборці

Дивергенція (лат. «divergence» – «відхиляюсь») – розходження ознак у споріднених організмів у процесі їх еволюції. Виникає внаслідок пристосування організмів до різних умов середовища.

Біфуркація (лат. «bifurcation» – «роздвоєний») – у широкому значенні різноманітні якісні перебудови чи метаморфози різних об'єктів при зміні параметрів, від яких вони залежать.



Біфуркаційна крива

Рис. 7.8. Якісні особливості катастрофічної поведінки системи

На рис. 7.8 продемонстровані якісні особливості катастрофічної поведінки систем. По осях *a* і *b* відкладені значення незалежних змінних, по осі *x* – залежної. Можливим положенням системи відповідає поверхня катастроф. Проекція цієї поверхні на площину (*a, b*) задає біфуркаційну криву.

Припустимо, що неперервній зміні значень параметрів *a* і *b* відповідає рух по кривій *RT*. У точці *T* відбувається катастрофа – система стрибком переходить з верхньої поверхні на нижню в точку *P*.

Кожному значенню параметрів *a* і *b* всередині біфуркаційної кривої відповідають два різних стани системи (бімодальність). На поверхні катастроф можна спостерігати явище гістерезису, коли поведінка системи істотно залежить від передісторії процесу. Наприклад, при зміні стану системи вздовж кривої *RT* відбувається стрибок з верхньої поверхні на нижню – з точки *T* в точку *P*. Однак при русі вздовж кривої *PQ* стрибок з нижньої поверхні на верхню відбудеться не в точці *P*, а в точці *Q*.

Є приклади застосування теорії катастроф для дослідження динаміки порушень режиму у в'язниці Гартрі впродовж 1972 р. [58]. За допомогою факторного аналізу було виділено два основних чинники, що впливають на безлади: напруженість (почуття розчарування і безвиході, важке становище); роз'єднаність (взаємне відчуження, відсутність спілкування, розбиття на два табори).

Аналіз підтвердив, що зі зростанням напруженості підвищується вірогідність хвилювань, а збільшення роз'єднаності пов'язане з характером хвилювань – вони стають раптовішими і жорсткішими (рис. 7.9).

Динаміка системи відповідає моделі катастрофи «зборка» [58]. З рис. 7.9 визначимо, що при низьких значеннях роз'єднаності система прагне до стійкого положення помірного хвилювання, але при високому рівні роз'єднаності вона змінює своє положення стрибком з нижньої поверхні на верхню і назад.

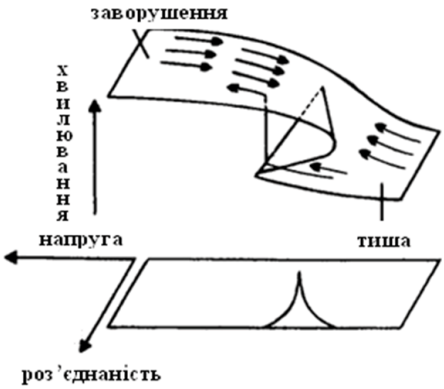


Рис. 7.9. Модель хвилювань у в'язниці

Розглянемо модель прийняття рішення про впровадження конкретного нововведення. Припустимо, що інновація приймається організацією, якщо оцінка прибутку, отриманого від впровадження нововведення, висока, і відкидається при низькій оцінці прибутку. Якщо оцінка набуває проміжного значення, то новинка може бути як знехтувана, так і прийнята. В останньому разі організація збирає додаткову інформацію про новинку, щоб точніше оцінити майбутній прибуток. Для вирішення цього завдання можна використати модель катастрофи «зборка» (рис. 7.10) [98].

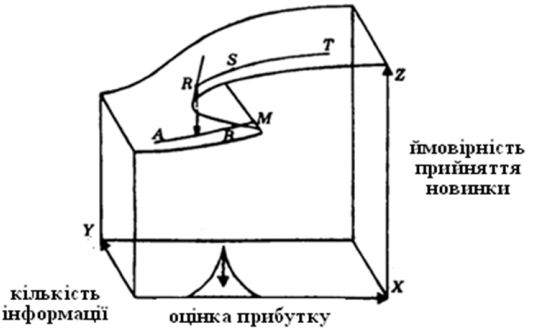


Рис. 7.10. Модель прийняття інновацій

Проекція поверхні катастроф на площину *XY* (рис. 7.11).

Рис. 7.11. Проекція поверхні катастроф

Кожній точці поза заштрихованою областю ризинка відповідає тільки одне рішення. Кожній точці всередині заштрихованої області відповідають два значення залежної змінної *Z*. Яке саме рішення буде прийнято, залежить від передісторії. Вертикальна пряма перетинає поверхню катастроф у трьох точках, але проміжне значення *Z* вважається неприпустимим.

Якщо керівництво було готове застосувати нововведення в точці *T* (рис. 7.10), то, рухаючись уздовж осі *X* (знижуючи оцінку прибутку, наприклад, до 1 млн. грн.), організація може впровадити новинку. Якщо організація відхилила новинку в точці *A*, то, перейшовши в точку *B* і збільшивши оцінку прибутку до 1 млн. грн., як і в точці *S*, вона не змінить рішення – діє інерція установки, кліше.

Перейдемо з точки *B* в точку *M* – оцінка прибутку зросте до 1,2 млн. грн. Далі незначна зміна оцінки до 1,21 млн. грн. приводить до різкої зміни рішення – інновація приймається.

При високому ступені інформованості (значення *Y* велике) і збільшенні параметра *X* стрибків не відбувається і система функціонує плавно.

Розглянемо в цій моделі петлю гістерезису (*A*, *M*, *T*, *R*, *A*).

Гістерезис (грецьк. «hysteresis» – «відстаючий») – властивість систем, що полягає в їх миттєвому відгуку на застосовані до них дії, що залежить у тому числі від їх поточного стану. Поведінка системи на інтервалі часу багато в чому визначається її передісторією. Для гістерезису характерне явище «насичення», а також неоднаковість траєкторій між крайніми станами (рис. 7.12).

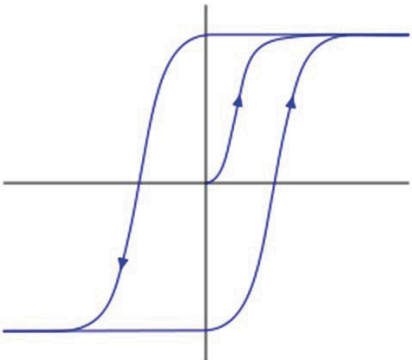


Рис. 7.12. Петля гістерезису

У цьому разі явище гістерезису (запізнювання) пояснюється інерційним сприйняттям менеджерів.

Хрестоматійний приклад гістерезису в оптичному сприйнятті

У верхньому ряду четверте ліворуч зображення (рис. 7.13) сприймається з однаковою ймовірністю як фігура дівчини і як чоловіче обличчя. Розпізнавання зображень усередині «дзьоба», виділеного штриховою лінією, залежить від напряму перегляду відповідного ряду – зліва направо або справа наліво. Поекспериментувавши з малюнком, можна дослідити особливості бістабільного сприйняття – явища, яке може бути описане моделлю катастрофи «зборка».

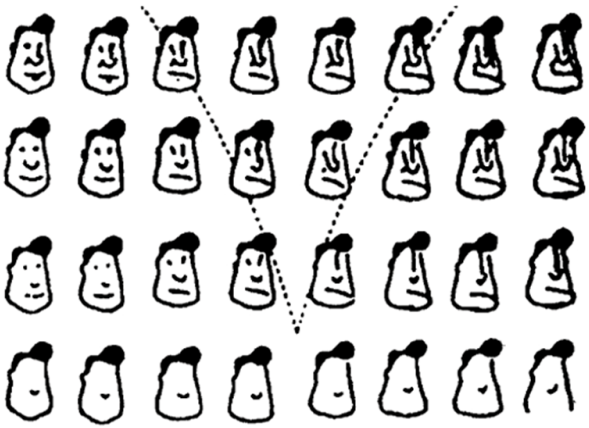


Рис. 7.13. Бістабільність сприйняття

Одне з основних понять сучасної нелінійної науки – біфуркація. У математиці біфуркація означає зміну кількості або стійкості рішень певного типу для моделі, що описує систему при зміні керуючих параметрів. У точці біфуркації система ніби робить вибір, який визначає її подальшу еволюцію. Поняття «біфуркація» описує процес переходу поступових кількісних змін керуючих параметрів в якісну зміну стану системи.

Так, доцільно розглянути два типи соціальних процесів [44]. У першому з них події мають позаособистісний характер, оскільки учасники процесу практично позбавлені права вибору. Можна вважати, що люди є частинками у броунівському русі велетенських соціальних процесів (розвиток громадських формацій, класові, національні рухи). Другий тип соціальних процесів пов'язаний з подіями, які здійснюються через свідомість людей і за допомогою цієї свідомості. «Людина опиняється перед можливістю вибору поведінки і незмінно співвідносить свої дії з образом цілі, уявленням про результати» [44]. Таким чином, там, де соціальний процес постає як множина альтернатив, вибір між якими здійснюється інтелектом і волею людини, потрібний пошук нових і складніших форм і моделей причинності.

Базуючись на ідеях синергетики (теорії самоорганізації в системах різноманітної природи), Ю. Лотман пропонує розглядати соціальний процес як багатофакторний потік. При досягненні точки біфуркації рух ніби зупиняється перед вибором подальшого шляху [44]. З цієї точки може виходити декілька рівноймовірнісних стійких траєкторій розвитку. У цьому моменті соціального процесу люди мають можливість здійснити вибір і змінити хід процесу.

Прикладом біфуркаційної діаграми історичного процесу є теорія розвитку давніх цивілізацій, яка може бути проілюстрована моделлю, представленою на рис. 7.14 [48].

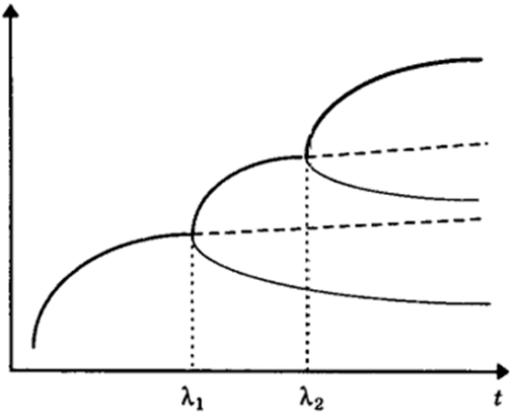


Рис. 7.14. Біфуркації в історичному процесі

По осі ординат відкладаються реальні доходи на душу населення, а по осі абсцис – час. Нехай з часом унаслідок зміни клімату й екології врожайність зернових зменшується. Нестача продовольства спричиняє зростання соціальної напруженості. Розростається криза і суспільство підходить до точки біфуркації (точка λ_1). Відповісти на «виклик історії» можна двома способами. Перший спосіб – зменшення потреб, жорсткий курс щодо сусідів (нижня гілка) (рис. 7.14). Другий спосіб – колонізація заморських територій, що перебувають на нижчій стадії розвитку. Наступний вибір (точка λ_2) пов'язаний із рішенням або бути торговою державою, або перейти до прямого управління колоніями.

7.3. Цифрові технології big data для прийняття оптимальних рішень

Термін **«великі дані» («big data»)** використовують для позначення аналізу даних у прогностичній аналітиці, інтелектуальному аналізі даних, управлінні взаємовідносинами з клієнтами. Це поняття застосовують у сферах, де критично важливою є якісна обробка великих обсягів даних, пов'язаних з постійним збільшенням швидкості їхнього потоку в організаційному процесі. Інформаційні ресурси та інформаційні технології сьогодні є одними із вирішальних чинників панівного становища у суспільстві, бізнес-середовищі та на політичній арені [85].

Великі дані – новий рубіж для інновацій, конкуренції та продуктивності. Цим терміном позначають набори даних, розмір яких перевершує можливості типових баз даних (БД) для введення, зберігання, обробки та аналізу інформації [47].

Проте великі дані означають дещо більше, ніж просто аналіз великих інформаційних масивів. Проблема не в тому, що сучасні ТНК, суспільні та державні організації обробляють значні обсяги даних, а в тому, що більша їх частина представлена в форматі, що не відповідає традиційному структурованому формату баз даних. Це веб-журнали, відеозаписи, текстові документи, машинний код або, наприклад, геопросторові дані. Все це зберігається в безлічі різноманітних сховищ, іноді навіть за межами установи. У результаті організації можуть мати доступ до великого обсягу власних даних і не володіти необхідними інструментами, щоб встановити взаємозв'язки між цими даними та сформулювати на їх основі вагомі висновки. Крім того, інформація постійно оновлюється і за допомогою традиційних методів аналізу даних неможливо ефективно обробляти великі інформаційні масиви.

За суттю поняття «big data» передбачає роботу з даними великих обсягів і різного складу, які отримують з різних джерел, з метою з підвищення ефективності прийняття управлінських рішень

та зростання конкурентоспроможності. Це поняття об'єднує техніки і технології, які «витягують» сенс з даних на екстремальній межі практичності.

У сучасному світі існує безліч джерел великих даних. Наприклад, дані, що неперервно надходять з вимірювальних пристроїв, потоки повідомлень із соціальних мереж, метеорологічні дані, дані супутникового дистанційного спостереження, потоки інформації про місце перебування абонентів мереж стільникового зв'язку, пристроїв аудіо- та відеореєстрації тощо. Масове поширення перерахованих вище технологій і принципово нових моделей використання різного роду пристроїв та інтернет-сервісів є відправною точкою для проникнення великих даних майже в усі сфери діяльності людини, передусім у науково-дослідну, комерційний сектор і державне управління. Сучасний «цифровий всесвіт» («digital universe») охоплює як цифрові зображення і відео, завантажені з мобільних телефонів, наприклад, на «Youtube», так і HD- відео, передане через мережі провайдерів. Це як корпоративні дані, що генеруються бізнес-додатками, так і дані, які створює великий адронний коллайдер. Значну частку даних, вироблених до 2020 р., згенерують не люди, а машини в ході взаємодії між собою та іншими мережами даних. Це можуть бути, наприклад, сенсори та інтелектуальні пристрої, які обмінюватимуться потоками інформації з іншими девайсами.

Інформаційні технології big data генерують щораз більше даних відповідно до свого застосування, що дає можливість збирати й отримувати інформацію, яка раніше була недоступною [97]. Сьогодні накопичено величезні обсяги даних, які піддаються аналітичній обробці і, таким чином, забезпечують нові можливості у плануванні бізнесу, політичної кар'єри та визначенні становища гравців на світовій арені.

Розвиток комунікацій робить можливим інтерактивний доступ до інформації масових користувачів, а також спрощує зв'язки між людьми. Нова технологія здійснює потужний вплив на сферу

конкуренції. Інформаційні системи дають можливість координувати зусилля у віддалених один від одного географічних регіонах. Усебічне розповсюдження інформаційних технологій створює важливу проблему для виконавців – дуже багато інформації. Ця проблема визначає нові форми використання інформаційних технологій для них – накопичення та аналіз потоків інформації.

Інструменти великих даних для інноваційного маркетингу

Разом зі стрімким накопиченням інформації швидкими темпами розвиваються також технології аналізу даних. Ще кілька років тому було можливо, наприклад, лише сегментувати спільноту на групи за схожими вподобаннями. Сьогодні можна будувати моделі для кожного окремого індивіда в режимі реального часу, аналізуючи, наприклад, його переміщення мережею Інтернет для пошуку конкретного «товару». Інтереси споживача можуть бути проаналізовані і згідно з побудованою моделлю виведена відповідна реклама або конкретні пропозиції [89]. Модель також можна налаштовувати і перебудовувати в режимі реального часу, що було неможливим нещодавно. Зокрема, це стало трендом у сфері телекомунікацій, де широко використовують технології для визначення фізичного розташування стільникових телефонів та їх власників. У найближчому майбутньому рекламна інформація в торгових центрах та онлайн- сервісах буде враховувати інтереси конкретних осіб і відображатись на моніторах у місцях появи кожного окремого потенційного споживача.

Big data завжди мали важливе значення для реклами з часів своєї появи, однак розвиток цих технологій дав можливість широкому загалу рекламодавців використовувати переваги інсайтів про потенційних споживачів, які раніше не були доступними. Big data допомагають рекламодавцям точніше визначати свою аудиторію, вкладати менше коштів у рекламу, пропонуючи її лише тим користувачам, які відгукнуться на неї. Наприклад, проксіміті-маркетинг за допомогою сервісів визначення місця розташування відстежує знаходження потенційного покупця за допомогою його мобільного

девайсу і надсиляє абоненту пропозиції або рекламні повідомлення відповідно до його інтересів.

Раніше організація з виготовлення бренда або окремого продукту передбачала створення форумів і поширення новин серед друзів і рідних. Однак зараз користувачі соціальних мереж створюють набагато більший контент. Надіслані ними тексти, фотографії, відеофайли та інші матеріали можуть мати велике значення для просування бренда і зростання обсягу продажів. Сьогодні користувачі створюють контент без заохочення, однак підтримка торгової марки допоможе створити команду лідерів думок.

Крім того, big data дають можливість здійснювати глибший аналіз, допомагаючи рекламодавцям точніше відстежувати рівень конверсії та інші фактори. Експерти прогнозують, що кількість інструментів великих даних буде зростати пропорційно до розповсюдження інформації про переваги застосування технологій big data.

Водночас існують ситуації, коли впровадження нових технологій може бути недоцільним. Наприклад, розріджені дані, що дають важливе розуміння дійсності, у деяких випадках є набагато цінішими, ніж великі дані, що описують безліч несуттєвої інформації.

Сьогодні великі обсяги інформації потребують опрацювання майже в усіх галузях. Їх використання відкриває нові можливості для різних сегментів споживачів. Автоматизація обробки великих даних – це друга технологічна революція після механізації процесів [81].

В останні роки технології аналізу big data стали загальнодоступними або зовсім відкритими. Апаратне та програмне забезпечення можна купити дуже дешево. Конкурентною перевагою є досвід і знання, необхідні для роботи з big data. Майже всі сервіси, доступні сьогодні в мережі Інтернет, базуються на аналізі великих даних. Нові технології big data застосовуються в найрізноманітніших галузях – від пошукових сервісів, перекладів та навігації до використання закономірностей, виявлених в інтернет-середовищі, в роздрібних офлайн-продажах і навіть політичній рекламі [92].

Важко знайти галузь, для якої проблематика великих даних не була б актуальною. Вміння оперувати великими обсягами даних, аналізувати взаємозв'язки між ними і приймати зважені рішення, з одного боку, дає потенціал для компаній з різних вертикалей для збільшення показників прибутковості та підвищення ефективності. З іншого боку, це чудова можливість для додаткового заробітку партнерів вендорів-інтеграторів і консультантів.

Незважаючи на незначний термін існування сектору big data, на сьогодні відомі високі результати ефективного використання цих технологій, що підтверджуються реальними прикладами. Один з найвищих показників отримано в енергетиці. За оцінюванням експертів, аналітичні технології big data здатні на 99% підвищити точність розподілу потужностей генераторів, а охорона здоров'я США завдяки big data може заощадити до 300 млрд. дол. [85].

Найскладнішим питанням залишається те, які результати при цьому передбачається отримати. Навіть банки не займаються аналізом великих обсягів даних у виробничих масштабах, оскільки основна їх частина належно організована і зберігається на мейнфреймах. Найдальше в цьому напрямку просунулися мережі роздрібної торгівлі. Вони з успіхом використовують дані з датчиків радіочастотної ідентифікації, систем постачання і купівельних карт. У багатьох інших галузях тільки починають шукати способи «видобування» та аналізу неявних даних.

Технології big data успішно реалізуються в різних індустріях, зокрема у банках, телекомі, енергетиці, медицині, керуванні міською інфраструктурою та у великій політичній грі. При всій різноманітності завдань вендорські рішення у сфері big data поки не набули яскраво вираженої галузевої спрямованості. Сьогодні ринок знаходиться на початковій стадії активного формування.

Більшість компаній збирає, записує й аналізує дані. Проте багато з них щодо big data зазнають складних ділових та інформаційно-технологічних проблем. Однак рішення на основі big data можуть вдосконалити процеси їх прийняття і підвищити конку-

рентоспроможність. Провідні світові компанії сьогодні отримують реальні стратегічні переваги від накопиченої інформації.

Інтернет є першою індустрією, що оперує big data. Абсолютно все, що виконує користувач онлайн, нікуди не зникає. Кожен «лайк» чи замовлення товару залишає цифровий слід у мережі. Навіть якщо користувач нічого не замовив, а лише переглянув інформацію, кілька подальших днів на різних сайтах, які він буде відвідувати, відображатимуться рекламні повідомлення з пропозицією придбати річ, якою він цікавився. Такий прийом називають **контекстною рекламою**. Ще більше інформації про себе, самі того не підозрюючи, користувачі залишають у соціальних мережах. Кожен проставлений «лайк» – дуже цінна інформація, на основі якої можна створити точний психологічний портрет конкретної людини. Аналіз кількох десятків чи сотень таких «лайків» дає можливість отримати інформацію про прибутки користувача, його сімейний стан, релігійні переконання, емоційну стабільність, сексуальну орієнтацію, колір шкіри та політичні погляди. Це терабайти інформації. І не відомо хто і як може її використати.

Уся зібрана інформація використовується для створення індивідуальних профілів потенційних покупців з метою адресної розсилки відповідної реклами, відстеження можливих загроз державній безпеці, проведення виборчих кампаній та розробки новітніх технологій маніпулювання суспільною свідомістю.

Сьогодні комунікація, як і передача інформації, є таргетованими. Вона індивідуалізована до кожної окремої особистості. Потенційний споживач ніколи не буде отримувати рекламу товарів і послуг, якими він не цікавився.

Big data – це величезні обсяги неструктурованої інформації про користувачів: історія відвідувань сайтів, інтереси, покупки тощо. Інформація деперсоналізована і повністю анонімна, але для ефективної реклами більшого і не потрібно. Кожен електронний пристрій має свій унікальний ID у комп'ютерній мережі, за яким можна відстежувати не лише його географічне місце розташуван-

ня у конкретний момент часу, а й зафіксувати (та проаналізувати) всі виконані його користувачем дії. Всі візити на сайти, перегляди профілів інших користувачів, електронні листи, поширення, «лайки», бесіди та події у різних соціальних мережах, інформація з платіжних карток назавжди залишають цифровий «слід». Система відстежує, аналізує та пов'язує інформацію про конкретного користувача з різних цифрових пристроїв, якими він користувався, – персональних комп'ютерів, мобільних телефонів, смартфонів тошо. Цієї інформації достатньо, щоб за допомогою технологій big data показувати конкретному споживачу оголошення та ролики з рекламними меседжами, які вплинуть саме на нього. Великі дані та цифрові технології відкривають нові можливості для підвищення ефективності використання інформації великих обсягів при проведенні цільових рекламних кампаній.

Завдання будь-якого рекламодавця – проаналізувати, використовуючи сучасні способи обробки даних, споживчу аудиторію, розділити її на сегменти за інтересами, захопленнями, споживацькими вподобаннями і запропонувати кожному бажане. На основі реакції користувача і його подальших дій система, заснована на технології big data, самостійно зможе «навчатися» за отриманими результатами: покращувати знання про користувача, формувати нові знання та уточнювати персоналізовану інформацію про споживачів. У маркетингу таку технологію називають «programmatic». На сьогодні у США на ній базується більше 60% реклами в Інтернеті [81].

Саме в сегментації полягає успіх реклами великих сучасних компаній, які змогли правильно розподілити аудиторії, зрозуміти їх бажання і запропонувати кожному потенційному споживачу саме те повідомлення, яке він хоче бачити і чути. Протягомі останніх кількох років великі бренди й онлайн-магазини об'єднують розрізнені дані з різних джерел і використовують класичні методи сегментації та таргетування.

Однак навіть найпотужніша сучасна інформаційна система не здатна проаналізувати життя мільйонів людей і змусити їх зробити

те, чого вони не хочуть. Потрібно враховувати тисячі чинників: від культурних особливостей до традицій у кожній конкретній сім'ї. Сучасні рішення можуть вплинути на вибір речі, але ніяк не змусити її купити. Великі компанії, які мають у своєму користуванні особисті дані, споживачів переживають за ставлення до них своїх клієнтів і дуже ретельно дотримуються умов деперсоналізації й анонімізації, щоб можна було «підлаштовувати Інтернет» під браузер, але не під конкретну особу. Крім того, користувач може заблокувати трекінг будь-якого рекламного банера у мережі.

Прорив у рекламних технологіях можливий завдяки нестандартному підходу до інтернет-маркетингу на основі технологій big data, психолого-поведінкового аналізу та персоналізованої реклами. Комп'ютерний алгоритм, розроблений польським психологом М. Козинським, дає можливість лише за кілька десятків «лайків» у соцмережах обчислити психотип будь-якої людини і згенерувати максимально співзвучне його цінностям і потребам персоналізоване рекламне послання [83].

Сьогодні існують спроби розробки спеціальних математичних моделей для більш детального аналізу великих обсягів інформації про потенційних споживачів і побудови на їх основі надточного таргетингу не тільки за географічним місцем розташування, намірами та інтересами, а й за їхнім психотипом і поведінковими характеристиками. На основі консолідованих даних з урахуванням отриманих знань маркетологи будують детально персоналізовану комунікацію з кожною з груп споживачів залежно від їхніх потреб, настроїв, купівельних можливостей, уподобань, психологічних особливостей як в офлайн-, так і в онлайн-середовищі. Для кожного окремого споживача розробляється практично індивідуальний меседж.

Сам собою такий підхід не новий, а фахівці з реклами завжди прагнули точніше зорієнтувати створені повідомлення на більш вузькі сегменти. Колись ці групи визначали соціологи на основі опитувань і анкетувань, а результати екстраполювали на аудиторію загалом. Сьогодні їх роль дедалі частіше виконує аналітика циф-

рових даних, математичні моделі та штучний інтелект, що сукупно дає можливість сегментувати аудиторію на зовсім іншому рівні точності.

Близько 73% людей віком від 20 до 35 років беруть участь у прийнятті рішень щодо придбання продуктів або послуг у своїх компаніях. Спочатку вони вивчають усі можливі варіанти в мережі Інтернет. Більше 70% пошуків розпочинаються з загального пошуку, а не з пошуку конкретних брендів. Ще до купівлі покупці мають величезну кількість інформації про репутацію компанії, технічні характеристики продукту та відгуки інших користувачів про переваги та недоліки товару. Основними джерелами інформації є соціальні медіа та спеціалізовані веб-сайти. Дослідження Forrester 2017 свідчать, що споживачі надають перевагу невеликим інформаційним повідомленням та візуалізації, а не телефонним дзвінкам [83].

Застосування цифрових технологій зменшує витрати користувачів, заощаджує час, забезпечує об'єднання різних джерел інформації, організовує, інформує та забезпечує доступ до раніше важкодоступних даних і висновків – усіх «елементів вартості», які мають прямий взаємозв'язок із придбанням та використанням продукту.

Сьогодні спостерігається зміна поведінки споживачів. Більше 80% з них при виборі покупок ураховують екологічні, соціальні та благодійні вподобання компаній-постачальників. Перші враження про бренд важливі також на ринку, тому компанії пропагують свій імідж через власні веб-сайти, користувацькі форуми та короткі тематичні дослідження, які надають споживачам онлайн-інформацію про обрані ними елементи вартості [83].

У світі реклами збір і сегментування даних сьогодні виконують DMP-платформи (Data Management Platform), що дають можливість бачити користувачів Інтернету в розрізі сотень (і навіть

тисяч) сегментів, компонувати їх з різною логікою для визначення найбільш ефективних для бренда комбінацій.

Цифрові «сліди» своїх переваг і потреб споживачі щодня залишають у мережі, здійснюючи пошукові запити, банківські платежі, користуючись GPS-навігацією і проходячи психологічні тести, що містять багато особистих питань. Необмежений доступ до масиву цієї інформації відкриває широкі можливості для впливу та маніпуляції.

Термін «big data» сьогодні відомий широкому загалу. Завдяки цій технології інформація, поширена в мережі Інтернет, та події, що відбуваються офлайн, залишають цифровий відбиток. Зберігається все – відомості про операції через електронні платіжні системи, дані про запити в пошуковій системі «Google», лайки у соціальних мережах, розмови у чатах, дзвінки зі «Skype» та «Viber», звичайна прогулянка зі смартфоном тощо.

Уся зібрана інформація використовується для створення індивідуальних профілів потенційних покупців з метою адресної розсилки відповідної реклами, відстеження можливих загроз державній безпеці, проведення виборчих кампаній та розробки новітніх технологій маніпулювання суспільною свідомістю. Big data стали для людства значним досягненням та великою небезпекою.

Сильні цього світу активно формують так зване сервітивне суспільство, в якому пропонують комфортне життя, привабливі позики, нереальну можливість отримати все й одразу. Технології big data відкривають для великого бізнесу та політики нові інструменти таємного впливу на аудиторію з метою управляти їхньою поведінкою, вчинками та навіть вибором. Використання новітніх методів таргетування сприяє формуванню психологічної залежності у споживачів. Основна ціль – утримати у своїй владі людину, яка не зможе самостійно забезпечити своє подальше існування. Така ідеологія сучасних ТНК спрямована на отримання великих надприбутків.

Технології big data як інструмент політичної стратегії

Технології інтернет-маркетингу на основі big data застосовують сьогодні відомі світові політичні літери. Як свідчить досвід, перемогу отримує той, хто використовує новітні методи аналітики та інформаційних технологій. Протягом двох останніх десятиліть Інтернет перетворив громадянське суспільство на безпрецедентний рух, стимулюючи колективні дії до радикально нового рівня самовизначення громадян. Демократичний вибір сьогодні щораз рідше реалізується за допомогою виборчих бюлетенів. Він щоденно виявляється онлайн. Інтернет спричинив демократизаційний ефект і здійснив вплив на недостатньо представлені та маргіналізовані групи (обмежені в отриманні соціальних послуг, програм і соціальних привілеїв), що живуть під авторитарними режимами. Соціальні спільноти стали незалежною альтернативою жорстко контрольованим засобам масової інформації.

Щораз більше членів суспільства сьогодні залучається до участі в публічному дискурсі та висловлюється за свої суспільні інтереси. Інтернет зробив вирішальною думку громадськості. Це особливо актуально, коли існують конфлікти та розбіжності між різними інтересами політичних і громадських груп. Інтернет є дієвим ресурсом здійснення надзвичайного тиску на політиків груп людей з високим ступенем мотивації, які мають можливість висловити своє невдоволення кожного разу, коли їхні погляди не збігаються з політикою можновладців.

Демократичні органи, як правило, обираються на період від трьох до п’яти років, однак думки громадян змінюються щодня й інколи ці перепади настрою зростають у величезних масштабах. Часто тисячі людей починають «твітити» один і той самий предмет у той самий день, а політична різноманітність електорату є надзвичайно динамічною та кардинально полярною. У таких випадках політичні діячі не можуть досягти консенсусу, здатного задовольнити всіх.

Навіть якщо владі не вдається врахувати цифрове вираження настроїв громадян, вважаємо помилкою не пов'язувати інтернет-контент з реальними політичними ситуаціями. Так, Brexit є не лише політичною катастрофою, а й прикладом поєднання неконтрольованої сили Інтернету з тим явищем, що пересічні люди втратили контроль над політикою, яка формує їхнє життя. Коли люди розуміють, що їхні демократичні представники більше не відстоюють їх інтереси, вони в мережі Інтернет шукають однодумців, організовуючись у громадські рухи.

З огляду на останні події (перемога цифрових виборчих кампаній на підтримку Д. Трампа та Brexit) популярність правих популістських настроїв зростає як в Європі, так і в США. Ці рухи підживлюються популістським гнівом, відродженням націоналізму та глибокою ворожнечею до іммігрантів [100]. Люди, які поділяли праві популістські ідеї, але ніколи не були достатньо сміливими, щоб їх відкрито висловити, тепер мають можливість зв'язатися з іншими однодумцями в соціальних мережах та використовувати Інтернет для висловлення своїх думок. Вони стають більш впевненими та енергійними, тому що простежують, що інші поділяють їхні переконання. Це пов'язано з тим, що посилення зв'язків між людьми, які поділяють спільні погляди, робить попередні переконання більш екстремальними. Індивід відчуває свою належність до спільноти і отримує здатність реалізувати те, що вважав немислимим раніше. Так, голосування за Brexit для багатьох стало голосуванням за відновлення своєї політичної незалежності та повернення втраченої національної ідентичності.

Велика різноманітність та доступність цифрового контенту забезпечує користувачам можливість вибору для спілкування груп людей зі схожим світоглядом – послідовників і друзів. Вони підтримують однодумців, посилюючи відповідно впевненість у правоті власних поглядів, які є неефективними для здорового демократичного дискурсу. І хоча соціальні медіа-платформи,

такі як «Facebook» і «Twitter», як правило, мають можливості забезпечувати політичне різноманіття поглядів, дослідження підтверджують, що контент-фільтри (програмне забезпечення для управління доступністю вмісту чи ресурсами), які вони іноді створюють, посилюються алгоритмами персоналізації платформ, які базуються на соціальних мережах і попередньо висловлених власних ідеях користувачів. Це означає, що замість того, щоб створити ідеальний тип цифрової опосередкованої «громадської думки», яка дозволить громадянам висловловити власні занепокоєння та поділитися своїми надіями, Інтернет фактично загострює ідеологічний конфлікт інтересів між протилежними поглядами. Зокрема, Інтернет використовує фрагментацію думок, дозволяючи особам, які є найбільш пристрасними, мотивованими та відвертими, знаходити людей з протилежними поглядами і вступати з ними у дебати. Таке явище відбулось у соціальних мережах під час проведення референдуму про вихід Великобританії з ЄС [100].

Аналогічним чином актуальні глобальні події, такі як стихійні лиха, терористичні напади або загрози навколишньому середовищу (наприклад, викид радіації), також можуть вплинути на громадську думку та спонукати прийняття поспішних політичних рішень з потенційно нестабільними наслідками. Політики ризикують прийняти важливі політичні рішення, засновані на поточних емоційних всплесках населення або швидких популярних думках, а не обрати оптимальну стратегію для країни чи світу. Наприклад, важливі та далекосяжні рішення, такі як Brexit, мають бути зваженими і схваленими більшістю громадян (2/3) протягом кількох років. Відповідно критично важливим завданням для тих, хто формує політику, є навчитися відрізняти моменти, коли начебто популярний рух фактично відображає загальну волю більшості та коли це лише відлуння гучної, але незначної меншості.

Нова форма цифрової опосередкованої політики є основним компонентом четвертої промислової революції. Інтернет нині використовується для встановлення порядку денного з висхідного рівня, надання громадянам можливостей висловлювати власні думки в мережевій публічній сфері та організовувати колективні дії. Зокрема, соціальні медіа змінили характер політичної агітації та продовжуватимуть відігравати важливу роль у майбутніх виборах і політичних кампаніях у всьому світі.

Однак ця технологія також може бути платформою для конфліктів і зловмисної агітації правих популістів, які не мають функціональних можливостей для демократичного дискурсу, тоді як сучасні системи управління вразливі до емоційних сплесків і популістських рухів, що виникають у соціальних мережах. Brexit засвідчив, що ця швидкісна технологія відкрита для всіх і може бути використана для впливу на громадську думку та результати референдумів чи виборів.

Так, під час останніх виборів президента США пропаганда, дезінформація та спам у «Facebook» і «Instagram» охопили приблизно 146 млн. американських громадян – майже половину населення країни. З 1 вересня до 15 листопада 2016 р. облікові записи «Twitter», зареєстровані в Росії, створили приблизно 1,4 млн. автоматичних повідомлень, пов'язаних із виборами, які загалом отримали майже 288 млн. користувачів. В «Instagram» претендента у президенти США Д. Трампа налічувалось чимало послідовників, однак більшість з них була російськими ботами [100].

Російські боти виявилися головними шанувальниками «Твіттер» Д. Трампа. Про це повідомили співробітники соцмережі, в якій президент США, як відомо, є активним дописувачем. В останні місяці передвиборчої кампанії саме російські фейкові користувачі пів мільярда разів передруковували, часто скандальні,

повідомлення Д. Трампа і його образи на адресу конкурентки – Г. Клінтон. Таким чином, вони підвищували популярність мільярдера в Інтернеті.

У рамках розслідування російського впливу на вибори, яке проводив Конгрес США, з'ясували також, що пов'язані з Росією інтернет-тролі створили понад 300 підробних сторінок, на яких видавали себе за політиків і журналістів, розповсюджували фейкові новини та навіть збирали акції протестів [100].

Сучасні соціальні медіа можуть впливати на результати виборів. Ті самі засоби масової інформації, які пропагували свободу та демократію у всьому світі, наприклад, у період Арабської весни 2011 р., тепер використовуються як ідеальний інструмент для маніпулювання думками та поширення ненависті, непорозумінь і розбрату.

Цифрові засоби масової інформації сьогодні використовують як інструмент політичної агітації. У будь-якій точці світу кандидати та партії мають можливості звернутися до широкої аудиторії. Політичні лідери можуть ретранслювати свої повідомлення у режимі реального часу. Крім того, соціальні медіа забезпечують платформу зворотного зв'язку: кожен середньостатистичний виборець має можливість брати участь в обговоренні, поширювати повідомлення, виявляти підтримку чи незгоду з програмою політика навіть зі свого смартфону.

Велика політика – це гра одного актора на публіку, а не лише перегони між партіями. Боти з великою кількістю прихильників, які поширюють фальшиві новини в закритих групах соціальних мереж, є ідеальними каналами для дезінформації. Деякі з цих новин спеціально розроблені для отримання політичних переваг окремими партіями чи кандидатами. Наприклад, американські ЗМІ повідомили, що велика кількість македонських підлітків на політичних майданчиках США розповсюджує через «Facebook» консервативні новини. 44% американців отримують новини з «Facebook», а

Д. Трамп, обраний президентом, може заплатити визначену ціну за такі операції [100].

Медіа-маніпулятори продають свої історії за допомогою соціальних мереж. Навіть коли дезінформація розвіяна, вона продовжує формувати думки та ставлення людей до обговорюваних подій. Таке відкрите маніпулювання свідомістю мас може зруйнувати демократію.

Так, під час референдуму за вихід Великобританії з ЄС у мережі «Facebook» з'явилися повідомлення про оплачену рекламну атаку, спрямовану на конкретних виборців у певних виборчих округах. З цією так званою «темною рекламою» ототожнювали саме консерваторів. Цей факт був порушенням основних правил щодо прозорості кампанії та націлювання на виборців.

Якщо раніше ніхто серйозно не сприймав інформацію про те, що пост на «Facebook» чи повідомлення у «Twitter» може спричинити коливання на виборах, сьогодні цілі пропагандистських стратегій у соціальних мережах спрямовані на реальні проблеми, такі як імміграція, економічна криза, тероризм та соціальна нерівність. Їхньою метою є прихований вплив на думки та поведінку мільйонів людей і спонукання до вибору певного політичного порядку чи точки зору.

Сучасні компанії-розробники програмного забезпечення, що базується на цифрових технологіях впливу, пропонують програмні рішення в різних сферах надання послуг, у тому числі «інтелектуального управління»: поглиблений аналіз виборчої карти, висвітлення основної політичної динаміки, оцінювання ваг політичних сил; адаптація програми до очікувань виборців; сегментація виборців та цілеспрямована комунікація; аналізу статей балансу тощо.

Такі компанії працюють у галузі аналізу даних. Вони надають підтримку своїм клієнтам у процесі збору, обробки та аналізу стратегічної інформації; пропонують певні інноваційні, високопродук-

тивні та надійні рішення для аналізу великих даних з метою збереігання лише інформації, необхідної безпосередньо для прийняття рішень.

Програмні розробки для підтримки стратегічних рішень і здійснення політичного впливу, запропоновані цими компаніями, не мають адекватної бази знань у демократичній сфері, тому неможливим є забезпечення отримання надійних результатів. Якість використовуваних даних невідома, а анонімність їх неможливо перевірити. Для референдуму Brexit та виборів Д. Трампа було використано не лише законно отримані дані, більша частина яких надійшла з «Facebook», а й різні технічні методи для відновлення великих даних без їх законного придбання і навіть без згоди користувачів Інтернету. Загалом проблема полягає не лише у фальсифікації результатів виборів президента США чи проведення референдуму про вихід Великобританії з ЄС. Це набагато ширше питання прихованого застосування особистих даних і технологій big data для здійснення впливу на процес демократичного вибору суспільства [77].

Технології великих даних відкривають широкі можливості як для великого бізнесу, так і для політики. Сьогодні тільки розпочинається розробка нових ефективних методик використання інструментів big data для підсвідомого впливу на цільову аудиторію з метою управляти поведінкою, вчинками та навіть вибором спільноти. У багатьох галузях здійснено лише перші спроби знайти способи «видобування» та аналізу неявних даних.

Великі дані та цифрові технології відкривають нові можливості для раціонального використання інформації великих обсягів та розвитку демократичного суспільства. Водночас вони можуть бути використані і як інструменти латентного психологічного впливу на споживачів та спонукати їх до небажаних вчинків і нераціонального вибору.

ПИТАННЯ ТА ЗАВДАННЯ ДЛЯ САМОПІДГОТОВКИ І САМОКОНТРОЛЮ

1. Теорія ігор як прикладний аспект математичного моделювання.
2. Основні поняття теорії ігор: хід гравця, гра, стратегія, ціна гри.
3. Ознаки класифікації ігор.
4. Ігри з нульовою сумою.
5. Кооперативні та некооперативні ігри.
6. Коаліційні ігри.
7. Застосування теорії ігор до дослідження подій МВ.
8. Поняття рівноваги.
9. Характеристики стратегічної теорії ігор.
10. Особливості статистичної теорії ігор.
11. Основні положення теорії катастроф.
12. Типи катастроф.
13. Застосування теорії катастроф до вивчення суспільних подій.
14. Визначення поняття «bid data».
15. Сфери застосування технологій великих даних.

Практичні завдання

Завдання 1. Альфонс і Бруно – єдині клієнти приватного банку «Brothers Layman». Вони поклали на депозит у цьому банку по 10 000 євро кожен. Банк вклав ці суми у довгостроковий проект. Якщо обидва клієнта залишать гроші в банку до кінця проекту, Альфонс і Бруно отримають по 11 000 євро кожен.

Якщо гроші будуть повернені до кінця проекту, проект припиняється, а вкладені суми втрачаються. У цьому разі банк може отримати лише 15 000 євро з інших ліквідних активів для задоволення вимог. Якщо лише один клієнт поверне свої гроші, він отримає 10 000 євро, тоді як інший матиме залишок у розмірі 5000 євро. Якщо обидва клієнти захочуть забрати свої гроші до кінця проекту, кожен отримає 7500 євро, половину наявних коштів банку.

Побудувати матрицю виграшів та визначити баланс Неша.

Завдання 2. Сукупний попит на шоколадні боби визначається функцією:

$$p = 252 - 8x,$$

де p – ціна в центах, x – необхідна кількість товару у кг.

Постачальниками є дві компанії «W & Wund» (фірма 1) та «Cutiesauf» (фірма 2), які виробляють однакову продукцію. Обидві компанії мають однакову функцію витрат:

$$K_i(x_i) = 2x_i^2,$$

де x_i – кількість виробленого товару в кг:

$$X = x_{W\&W} + x_{Cuties}.$$

Побудувати матрицю виграшів за умови можливих стратегій «співпраця» та «відсутність співпраці» для кожної з фірм.

Визначити оптимальні обсяги виробництва, ринкову ціну на шоколадні боби та прибуток компанії, які врівноважуються у разі одночасної конкуренції за обсягом.

ЛІТЕРАТУРА

1. Айвазян С. А. Прикладная статистика : уч. пособ. [для вузов] : в 2 т. Т. 2 : Основы эконометрики. 2-е изд., испр. Москва : ЮНИТИ-ДАНА, 2001. 432 с.

2. Аллен Р. А. Экономические индексы. Москва : Статистика, 1980. 321 с.

3. Арнольд В. И. Теория катастроф. Москва : Наука, 1990. 128 с.

4. Афанасьєв В. І. Метод середніх в економічних розрахунках. Москва : Фінанси і статистика, 1996. 224 с.

5. Барсегян А. А. Технология данных Data Mining, Visual Mining, Text Mining, OLAP. Санкт-Петербург : БХВ-Петербург, 2007. 384 с.

6. Барсегян А. А. и др. Анализ данных и процессов : уч. пособ. 3-е изд., перераб. и доп. Санкт-Петербург : БХВ-Петербург, 2009. 512 с.

7. Басков А. Я., Туленков И. В. Методология научного исследования : уч. пособ. Киев : МАУП, 2002. 216 с.

8. Бахрушин В. Є. Методи аналізу даних : навч. посіб. Запоріжжя : КПУ, 2011. 268 с.

9. Беккер Й., Вилков В., Таратухин В., Кугелер М., Роземанн М. Менеджмент процессов Москва : Эксмо, 2008. 384 с.

10. Біловодська О. А., Грищенко О. Ф. Конспект лекцій з дисципліни «Системний аналіз і прийняття інноваційних рішень». Суми : Сумський державний університет, 2010. 106 с.

11. Болюх М. А. та ін. Економічний аналіз : навч. посіб. / за ред. акад. НАНУ, проф. М. Г. Чумаченка. 2-ге вид., переробл. і допов. Київ : КНЕУ, 2003. 556 с.

12. Боришполец К. П. Методы политических исследований : уч. пособ. Москва : Аспект Пресс, 2005. 221 с.

13. Боровиков В. Statistica. Искусство анализа данных на компьютере: для профессионалов. 2-е изд. Санкт-Петербург : Питер, 2003. 688 с.

14. Буреева Н. Н. Многомерный статистический анализ с использованием ППП «STATISTICA» : уч.-метод. материалы по программе повышения квалификации «Применение программных средств в научных исследованиях и преподавании математики и механики». Нижний Новгород, 2007. 112 с.

15. Василенко О. А., Сенча І. А. Математично-статистичні методи аналізу у прикладних дослідженнях : навч. посіб. Одеса : ОНАЗ ім. О. С. Попова, 2011. 166 с.

16. Васильев В. А. Модели экономического обмена и кооперативные игры. Новосибирск : Изд-во НГУ, 1984. 96 с.

17. Вовк Р. В. Моделювання міжнародних відносин : навч. посіб. Київ : Знання, 2012. 246 с.

18. Глинский В. В., Ионин В. Г. Статистический анализ : уч. пособ. 3-е изд., перераб. и доп. Москва : ИНФРА-М ; Новосибирск : Сибирское соглашение, 2002. 241 с.

19. Головач А. В., Єріної А. М., Козирєва О. В. Статистика : підруч. / за ред. А. В. Головача. Київ : Вища школа, 1993. 623 с.

20. Гольдберг А. М. и др. Общая теория статистики : учеб. / под ред. А. М. Гольдберга, В. С. Козлова. Москва : Финансы и статистика, 1985. 367 с.

21. Горбатенко В. П., Бутовська І. О. Політичне прогнозування : навч. посіб. Київ : МАУП, 2005. 152 с.

22. Горкавий В. К., Ярова В. В. Математична статистика : навч. посіб. Київ : ВД «Професіонал», 2004. 384 с.

23. Григорків В. С. Моделювання економіки : підруч. Чернівці : Чернів. нац. ун-т ім. Ю. Федьковича, 2019. 360 с.

24. Грисенко М. В., Чугаєв О. А. Кількісні методи аналізу міжнародних економічних відносин : навч. посіб. Київ : Ін-т міжнар. відносин КНУ ім. Т. Шевченка, 2012. 235 с.

25. Губко М. В., Новиков Д. А. Теория игр в управлении организационными системами. Москва : ИПУ, 2005. 138 с.

26. Диференціальні моделі. Стійкість : навч. посіб. [для студ. вищ. навч. закл.] / за ред. А. М. Самойленка. Київ : Вища школа, 2000. 331 с.

27. Економіко-математичне моделювання : навч. посіб. за заг. ред. В. В. Вітлінського. Київ : КНЕУ, 2008. 536 с.

28. Економічний аналіз : навч. посіб. / за ред. акад. НАНУ, проф. М. Г. Чумаченка. Вид. 2-ге, переробл. і допов. Київ : КНЕУ, 2003. 556 с.

29. Электронный учебник компании «StatSoft». URL : www.statosphere.ru.

30. Іванов В. Ф. Контент-аналіз: методологія і методика дослідження ЗМК : навч. посіб. / за ред. А. З. Москаленко. Київ : КНУ ім. Т. Шевченка, 1994. 112 с.

31. Ковальчук О. Я. Прогнозування міжнародних соціально-економічних процесів: комп'ютерне моделювання в SPSS : метод. вказівки для виконання практичних завдань з курсу «Прогнозування міжнародних соціально-економічних процесів». Тернопіль : ТНЕУ, 2016. 59 с.

32. Ковальчук О. Я. Методичні вказівки для виконання практичних завдань з курсу «Математичне моделювання та прогнозування міжнародних відносин і технології підтримки прийняття рішень». Тернопіль : ТНЕУ, 2015. 57 с.

33. Ковальчук О. Я. Методичні вказівки для виконання практичних завдань з курсу «Інтелектуальний аналіз даних». Тернопіль : ТНЕУ, 2014. 108 с.

34. Ковальчук О. Я. Конспект лекцій з дисципліни «Інтелектуальний аналіз даних». Тернопіль : ТНЕУ, 2015. 109 с

35. Ковальчук О. Я. Методичні вказівки для виконання практичних завдань з курсу «Статистичний аналіз даних». Тернопіль : ТНЕУ, 2015. 88 с.

36. Ковальчук О. Я. Математичне моделювання і прогнозування в міжнародних відносинах : навч. посіб. Тернопіль : ТНЕУ, 2016. 423 с.

37. Ковальчук О. Я. Математичне моделювання сталого розвитку : моногр. Тернопіль: ТНЕУ, 2017. 245 с.

38. Ковальчук О. Я. Методичні вказівки для виконання практичних завдань з курсу «Кількісні методи в міжнародних відносинах». Ч. 1. Тернопіль, 2017. 88 с.

39. Ковальчук О. Я. Навчально-методичні матеріали для самостійного виконання практичних завдань з курсу «Кількісні методи в міжнародних відносинах». Ч. 2. Тернопіль, 2018. 85 с.

40. Кондіус І. С. Конспект лекцій за темою «Прогнозування соціально-економічних процесів» (частина 1 навчально-методичного комплексу «Прогнозування соціально-економічних процесів») : метод. матеріали з питань самостійної роботи із спеціальною літературою. Т. 1. Севастополь : Севастоп. центр перепідготовки та підвищення кваліфікації, 2013. 76 с.

41. Купалова Г. І. Теорія економічного аналізу : навч. посіб. Київ : Знання, 2008. 639 с.

42.Лабораторний практикум з навчальної дисципліни «Бізнес-статистика» для студентів спеціальності 8.03050601 «Прикладна статистика» денної форми навчання / укл. О. В. Раєвнєва, І. В. Чанкіна, Л. А. Гольтяєва. Харків : Вид. ХНЕУ, 2013. 68 с.

43. Ланде Д. В., Фурашев В. М., Юдкова К. В. Основи інформаційного та соціально-правового моделювання : навч. посіб. Київ : НТУУ «КПІ», 2014. 220 с.

44. Лотман Ю. М. Культура и взрыв. Москва : Гнозис, 1992. 270 с.

45. Лугінін О. Є., Білоусова С. В., Білоусов О. М. Економетрія : навч. посіб. Київ : Центр уч. л-ри, 2005. 252 с.

46. Ляшенко О. М., Ковальчук О. Я. Багатовимірний аналіз даних у системі STATISTICA. Методичні вказівки для виконання практичних завдань з курсу «Інформаційно-аналітична діяльність у міжнародних відносинах». Тернопіль : ТНЕУ, 2016. 87 с.

47. Майер-Шенбергер В., Кукьер К. Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим. Москва : Манн, Иванов и Фербер, 2014. 240 с.

48. Малинецкий Г. Г. Математические основы синергетики. Хаос, структуры, вычислительный эксперимент. 6-е изд. Москва : Либроком, 2009. 312 с.

49. Маловіан І. Ф. Політична гра в умовах демократизації суспільства: особливості та складові. *Вісник СевНТУ*. Серія «Політологія». 2013. Вип. 145. С. 137–140.

50. Маляренко Т. Теорія ігор у дослідженні конфліктів. *Освіта регіону.* 2009. № 4. С. 162–168.

51. Мальський М. З., Мацях М. М. Теорія міжнародних відносин : підруч. 3-тє вид., переробл. і допов. Київ : Знання, 2007. 461 с.

52. Мамчич Т. І., Оленко А. Я., Осипчук М. М., Шпортюк В. Г. Статистичний аналіз даних з пакетом STATISTICA : навч.-метод. посіб. Дрогобич : Вид. фірма «Відродження», 2006. 207 с.

53. Мармоза А. Т. Теорія статистики : підруч. 2-ге вид., переробл. та допов. Київ : Центр уч. л-ри, 2013. 529 с.

54. Многомерный статистический анализ в экономических задачах: компьютерное моделирование в SPSS : уч. пособ. / под ред. И. В. Орловой. Москва : Вузовский учеб., 2009. 310 с.

55. Моторин Р. М. Міжнародна економічна статистика : підруч. Київ : КНЕУ, 2004. 324 с.

56. Опря А. Т. Статистика : навч. посіб. Київ : Центр уч. л-ри, 2012. 448 с.

57. Палеха Ю. І., Герасимчук В. І., Шиян О. М. Соціологія : навч. посіб. Київ : ЄУФІМБ, 2003. 283 с.

58. Постон Т., Стюарт И. Теория катастроф и ее приложения. Москва : Мир, 1980. 607 с.

59. Присенко Г. В., Равікович Є. І. Прогнозування соціально-економічних процесів. Київ : КНЕУ, 2005. 382 с.

60. Растригин Л. А., Еренштейн Р. Х. Метод колективного распознавания. Москва : Энергоиздат, 1981. 80 с.

61. Руденко В. М. Математична статистика : навч. посіб. Київ : Центр уч. л-ри, 2012. 304 с.

62. Рябушкин Т. В., Яковлева Н. И., Ефимова М. Р., Ипатова И. М. Общая теория статистики. Москва : Финансы и статистика, 1981. 251 с.

63. Саати Т. Л. Математические модели конфликтных ситуаций / под ред. И. А. Ушакова. Москва : Сов. радио, 1977. 304 с.

64. Советов Б. Я., Яковлев С. А. Моделирование систем : учеб. [для бакалавров]. 7-е изд. Москва : Юрайт, 2012. 343 с.

65. Станжицький О. М., Таран Є. Ю., Гординський Л. Д. Основи математичного моделювання : навч. посіб. Київ : Київ. ун-т, 2006. 96 с.

66. Стеценко І. В. Моделювання систем : навч. посіб. Черкаси : ЧДТУ, 2010. 399 с.

67. Туронок С. Г. Политический анализ: курс лекций. Москва : Дело, 2005. 359 с.

68. Хусаїнов Д. Я., Харченко І. І., Шатирко А. В. Введення в моделювання динамічних систем : навч. посіб. Київ : КНУ ім. Т. Шевченка, 2011. 135 с.

69. Цыганков П. А. Теория международных отношений : уч. пособ. Москва : Гардарики, 2003. 590 с.

70. Шеннон Р. Имитационное моделирование систем – искусство и наука. Москва : Мир, 1986. 418 с.

71. Шиян А. А. Теорія ігор: основи та застосування в економіці та менеджменті : навч. посіб. Вінниця : ВНТУ, 2009. 164 с.

72. Штефан С. В. Основи прикладної статистики : метод. поради. Київ : Нац. ун-т ім. Т. Шевченка, 2002. 92 с.

73. Aumann R. J. Lectures on Game Theory. San Francisco : Westview Press, 1989. 120 p.

74. Axelrod R. Effective Choice in the Prisoner's Dilemma. *Journal of Conflict Resolution.* 1980. № 24. P. 3–25.

75. Banzhaf J. Weighted Voting Does'nt Work: A Mathematical Study. *Rutgtrs Law Review*. 1965. № 19. P. 317–343.

76. Brams S. J., Riker W. H. Models of Coalition Formation in Voting Bodies. *Mathematical Applications in Political Science.* University Press of Virginia, 1972.

77. Cadwalladr C. The great British Brexit robbery: how our democracy was hijacked. URL: https://www.theguardian.com.

78. Calvert R. Models of Imperfect Information in Politics. New York, 1986.

79. Chacon H., Roy A. Piecewise Solutions to Big Data. *Working Papers from College of Business, University of Texas at San Antonio.* URL : http://interim.business.utsa.edu.

80. Dixit A., Nalebuff B. Thinking Strategically: The Competitive Edge in Business, Politics and Everyday Life. N.Y. : Norton, 1991. 394 p.

81. *Electro Magnetic Compatibility.* URL: https://support.emc.com.

82. *Institut Fur Statistik.* URL: http://www.statistik.lmu.de/.

83. *Green plum.* URL: www.greenplum.com.

84. Harsanyi J. Games of Incomplete Information Played by Bayesian Players. *Management Science.* 1967. № 14. P. 159–182.

85. How Digital Natives Are Changing B2B Purchasing. *Harvard Business Review*. URL: https://hbr.org.

86. Hove N. Commentary: Big Data, a Big Decision. *Foresight: The International Journal of Applied Forecasting.* 2017. Is. 45. P. 24–26.

87. Mann I., Shapley L. The apriori Voting Strength of the Electoral College. Game Theory and Related Approaches to Social Behavior; ed. By L. Shubik. N.Y., 1964.

88.Nisbet R., Miner G., Elder J. Handbook of statistical analysis and data mining applications. Academic Press, 2009. 864 p.

89. Oakes N. Bloomberg Business. URL: http://www.bloomberg.com/research/stocks/private.

90. Rapoport A., Chamma A. Prisoner's Dilemma: A Study in Conflict an Cooperation. Univercity of Michigan Press, 1965.

91. Riker W. H., Shapley L. S. Weighted Voting: A Mathematical Analysis for Instrumental Judgments. Representation ed. By Pennok J. W. New York, 1968.

92. Salesforce boots SAP from customer-wrangling software top slot. *The Register*. URL: http://www.channelregister.co.uk.

93. Shapley L. Shubik M. A Method of Evaluating the Destribution of Power in a Committee System. *American Political Science Review*. 1954. № 40.

94. *Stack Exchange*. URL**:** http://stats.stackexchange.com.

95. *StatSoft's Electronic Statistics Textbook*. URL: http://www.statsoft.com.

96. Stevens J., Pituch K. Applied multivariate statistics for the social sciences. New York : Taylor & Francis Group, 2016.

97. *Tableau Software***.** URL**:** www.tableausoftware.com.

98. Terence A., Richard O. William O. Bearden The relationships among consumer satisfaction, involvement, and product performance: A catastrophe theory application. *Behavioral Science*. 1995. № 40. P. 104–132.

99. *Tweets from @realdonaldtrump.* URL**:** http://varianceexplained.org/r/trump-tweets/.

100. *Weforum.* URL**:** https://www.weforum.org.

ГЛОСАРІЙ ТЕРМІНІВ

Абсолютні статистичні величини – узагальнюючі показники, отримані у результаті первинного спостереження, які характеризують розміри, рівень і обсяг сукупності у конкретних умовах місця та часу.

Автоматичне анотування – скорочення тексту зі збереженням його сенсу. Результат містить найбільш значущі речення у тексті.

Агломеративні методи кластеризації – ієрархічні методи, в яких на початковому етапі кожен об'єкт знаходиться в окремому кластері. На подальших етапах відбувається об'єднання об'єктів у більші кластери на підставі пониження деякого порогу, наприклад, збільшення відстані між об'єктами.

Аналіз виживання – статистичний аналіз для вивчення оцінки та порівняння часу, що минув до настання деякої події

Аналіз за допомогою індикаторів – виокремлення найважливіших факторів (індикаторів) поведінки учасників МВ.

Аналіз кореляцій – встановлення факту наявності або відсутності залежності між двома параметрами.

Аналіз регресії – встановлення виду залежності (рівняння) між відгуком (залежною змінною) та факторами (незалежними змінними).

Аналіз тенденцій – виявлення закономірності розвитку явища в часі.

Аналіз трендів – виділення трендів у наборах документів на визначений період.

Аналітичне групування – метод дослідження взаємозв'язків, за якого всі спостереження поділяють на групи за величиною факторної ознаки і для кожної групи обчислюють середні значення результатної ознаки.

Апроксимація – наближене представлення одних математичних об'єктів іншими, простішими.

Асиметрія – ступінь відхилення графіка розподілу частот від симетричного вигляду відносно середнього значення.

Біфуркація – у широкому сенсі різні якісні перебудови чи метаморфози різних об'єктів при зміні параметрів, від яких вони залежать.

Багатовимірне концептуальне представлення (multidimensional conceptual view) – множинна перспектива, що складається з декількох незалежних вимірів, уздовж яких можуть бути проаналізовані окремі сукупності даних.

Варіація – коливання, мінливість або різноманітність значень ознаки окремих одиниць статистичної сукупності.

Великі дані – набори даних, розмір яких перевершує можливості типових баз даних для введення, зберігання, обробки та аналізу інформації.

Вибірка – сукупність об'єктів, вибраних випадковим чином з генеральної сукупності, на основі дослідження яких роблять висновок про генеральну сукупність.

Вивчення документів – аналіз офіційних документів державних, політичних, релігійних та громадських організацій щодо зовнішньо-економічної діяльності.

Відносні статистичні величини – похідні (обчислені) показники, які характеризують кількісне зіставлення статистичних даних.

Генеральна сукупність – множина всіх об'єктів, що підлягають дослідженню.

Гістерезис – властивість систем, їх миттєвий відгук на застосовані до них дії, що залежить у тому числі від їх поточного стану; поведінка системи на інтервалі часу багато в чому визначається її передісторією.

Дельфійський метод – систематичне та контрольоване обговорення конкретної проблеми кількома анонімними експертами у сфері МВ.

Дерева рішень – це структури даних, які дають можливість інтерпретувати шаблони даних з метою їх розпізнавання.

Дескриптивні моделі використовують для опису різних процесів, об'єктів.

Деталізація – операція, яка визначає перехід вниз у напрямі від агрегованого (up) представлення даних до детального (down).

Дивергенція – розходження ознак у споріднених організмів у процесі їх еволюції. Виникає за пристосування організмів до різних умов середовища.

Динамічний (часовий) ряд – сукупність значень статистичних показників, розташованих у хронологічному порядку, які характеризують зміну конкретного соціально-економічного чи політичного явища.

Дискримінантний аналіз – розділ багатовимірного статистичного аналізу, що охоплює методи класифікації багатовимірних спостережень за принципом максимальної схожості за наявності навчальних вибірок.

Експеримент – верифікація та перевірка достовірності побудованих гіпотез.

Екстраполяція – знаходження невідомих рівнів наприкінці чи на початку динамічного ряду відповідно до обраних індикаторів за визначеними часовими інтервалами.

Ексцес – міра гостро- або плосковершинності графіка розподілу досліджуваної ознаки.

Елімінування – усунення впливу всіх, окрім одного, факторів на величину результатного показника.

Емпірична лінія зв'язку відображає форму взаємозалежності між ознаками.

Зворотний кореляційний зв'язок між ознаками – зі збільшенням факторної ознаки величина відгуку має тенденцію до зменшення.

Івент-аналіз – аналіз публічної інформації з метою визначення та систематизації дій у міжнародних відносинах.

Імітація – відтворення за допомогою комп'ютерної програми процесу функціонування складної системи в часі.

Ітеративні дивізійні методи кластеризації – розбиття об'єктів, об'єднаних в один або декілька великих кластерів, на фіксовану кількість кластерів, здебільшого дрібніших.

Інтерполяція – знаходження показника (приблизний розрахунок) у середині ряду на основі закономірностей розвитку явища за досліджуваний період.

Інформаційна модель – інформація про характерні властивості та стани об'єкта, його взаємозв'язки із зовнішнім середовищем.

«Катастрофа» – стрибкоподібні зміни, які виникають при плавних змінах значень параметрів.

Класифікація (classification) – визначення для кожного документа однієї або декількох заздалегідь заданих категорій, до яких він належить. Частковий випадок – визначення тематики документа.

Кластерний аналіз – група методів, які використовують для класифікації об'єктів у відносно однорідні групи (кластери).

Когнітивне картування – вивчення особливостей мислення (понять та категорій) осіб, які приймають рішення у сфері зовнішньоекономічної політики.

Консолідація (Drill Up) – операція, яка визначає перехід вгору у напрямі від детального (down) представлення даних до агрегованого (up).

Контент-аналіз – виявлення та аналіз специфічних характеристик інформаційних масивів з метою встановлення прихованого змісту чи сенсу, на якому хоче наголосити автор.

Кореляційний зв'язок – зі зміною факторної ознаки змінюється середнє значення відгуку.

Крива розподілу – графічне зображення (неперервна лінія) змін частот варіаційного ряду, функціонально пов'язаних зі зміною варіант.

Матеріальна модель відтворює геометричні та фізичні властивості об'єкта-оригіналу, має реальне втілення.

Метод асоціацій та аналогій – ідентифікація асоціативних зв'язків між основними поняттями у множині документів.

Метод експертних оцінок – виявлення думок експертів у галузі МВ щодо конкретної проблеми.

Метод «мозкового штурму» – оперативна колективна генерація нестандартних ідей незалежних фахівців з різних предметних галузей, що полягає в інтенсивному аналізі проблемної ситуації, продукуванні думок, їх критиці та формулюванні контрідей, угрупуванні, відборі та оцінювання висловлених ідей.

Моделювання у міжнародних відносинах – метод наукового дослідження, пов'язаний з абстрагованим відображенням реальних явищ, ситуацій та процесів МВ за допомогою ідеалізованих об'єктів і зв'язків між ними.

Модель – ідеальний чи матеріальний образ (замінник) досліджуваних об'єктів та зв'язків між ними.

Мультиколінеарність – сильна взаємна корельованість.

Навігація за текстом (text-base navigation) – переміщення по документах відносно тем і значущих термінів.

Нормативний прогноз розробляють на основі заздалегідь визначених цілей.

Обчислювальний експеримент – методологія дослідження, заснована на вивченні математичної (інформаційної) моделі за допомогою логіко-математичних комп'ютерних алгоритмів.

Одиниця статистичної сукупності – окремий елемент статистичної сукупності, який є носієм ознак, значення яких підлягають реєстрації.

Оптимізаційні моделі призначені для керування об'єктами та прийняття рішень.

Побудова сценаріїв – визначення ймовірного розвитку подій МВ на основі аналізу конкретної ситуації.

Помилкова кореляція – завищення міри впливу незалежних змінних на результатну ознаку.

Порівняння у теорії міжнародних відносин – зіставлення спільних та відмінних рис окремих явищ і процесів, що відбуваються на міжнародній арені.

Порівняння паралельних рядів – зіставлення ряду значень факторної ознаки та ряду відповідних значень результатної ознаки.

Пошуковий прогноз – умовне перенесення на майбутнє закономірностей розвитку об'єкта дослідження в минулому і сучасному за умов збереження наявних тенденцій.

Прогальний аналіз – формування судження про відсутні факти, фактори, причини.

Прогнозні методи – наукові передбачення майбутніх явищ та процесів у міжнародних економічних відносинах.

Прогнозування – процес передбачення майбутнього стану предмета чи явища на основі систематичного аналізу інформації про минуле та сучасне, про якісні та кількісні характеристики його розвитку.

Прямий кореляційний зв'язок – закономірність розвитку явища, за якої збільшення величини факторної ознаки зумовлює зростання величини результатної ознаки.

Результатні ознаки (відгуки) характеризують наслідки (змінюються під дією факторних ознак).

Рейтинговий метод – встановлення ступеня популярності особи (організації, угрупування), її діяльності, програм, планів, політики на конкретний момент часу; визначення місця актора політичної арени серед собі подібних, визначене шляхом голосування, соціологічних опитувань, анкетування.

Розпізнавання образів (об'єктів, сигналів, ситуацій, явищ, подій або процесів) – ідентифікація об'єкта чи визначення його параметрів за зображенням (оптичне розпізнавання) або аудіозаписом (акустичне розпізнавання) і т. ін.

Ряд розподілу – впорядкований розподіл одиниць досліджуваної сукупності на групи за однією варіативною ознакою.

Середня величина – узагальнюючий показник, що характеризує типовий рівень варіативної ознаки, в розрахунку на одиницю однорідної сукупності, який може не збігатися з жодним з індивідуальних значень ознаки.

Системний метод дослідження міжнародних відносин – розгляд об'єкта дослідження в його єдності та цілісності.

Спостереження – суб'єктивна фіксація реальних процесів та явищ МВ, яка полягає в планомірному, науковоорганізованому збиранні даних для подальшого дослідження.

Статистична гіпотеза – припущення щодо виду або параметрів закону статистичного розподілу, якому підпорядковується генеральна сукупність.

Статистична закономірність – кількісна закономірність змін у просторі та часі масових явищ і процесів МВ, які складаються з множини елементів (одиниць сукупності).

Статистична сукупність – впорядкована множина однорідних матеріальних об'єктів, кожен з яких володіє спільними властивостями, умовами та факторами існування й динаміки.

Статистичний метод дослідження МВ – оцінювання кількісних параметрів масових суспільних явищ у причинно-наслідковому зв'язку з їхнім якісним змістом.

Статистичний показник – узагальнююча кількісна характеристика властивостей сукупності загалом чи її частин зокрема щодо конкретних умов місця і часу в поєднанні з їхньою якісною визначеністю (економічним змістом).

Стохастичний (статистичний) зв'язок – між зміною факторної та результатної ознак немає однозначної відповідності.

Теорія ігор – теорія математичних моделей прийняття оптимальних рішень у конфліктних ситуаціях.

Узагальнення – сукупність дій зі зведення окремих фактів в єдине ціле з метою виявлення загальних ознак і закономірностей, властивих досліджуваному явищу чи процесу.

Узагальнюючі показники – числові статистичні характеристики досліджуваної сукупності загалом чи її частин зокрема.

Факторний аналіз застосовують за необхідності обмеження кількості індикаторів (змінних).

Факторні (незалежні) ознаки характеризують фактори (зумовлюють зміни інших, пов'язаних із ними ознак).

Функціональний зв'язок – однозначна відповідність між змінами факторної ознаки та величини результату.

Цензуровані спостереження – спостереження, які містять неповну інформацію.

OLAP (On-Line Analytical Processing) – технологія оперативної аналітичної обробки даних, що використовує методи і засоби для збору, зберігання й аналізу багатовимірних даних з метою підтримки процесів ухвалення рішень.

НАВЧАЛЬНЕ ВИДАННЯ

Ольга Ярославівіна Ковальчук

МАТЕМАТИЧНЕ МОДЕЛЮВАННЯ ТА ПРОГНОЗУВАННЯ В МІЖНАРОДНИХ ВІДНОСИНАХ

Підручник

Редактор Оксана Бойчук
Комп'ютерне верстання Любові Верней

Підписано до друку 07. 11. 2019 р.
Формат 60x84 $^{1}/_{16}$. Гарнітура Times.
Папір офсетний. Друк на дублікаторі.
Умов. друк. арк. 23,9. Облік.-вид. арк. 21,7.
Зам. № У263-19. Тираж прим. 100.

Видавець та виготовлювач:
Тернопільський національний економічний університет
вул. Бережанська, 2, м. Тернопіль 46004

Свідоцтво про внесення суб'єкта видавничої справи до Державного реєстру видавців ДК № 3467 від 23.04.2009 р.

Видавничо-поліграфічний центр «Економічна думка ТНЕУ»
вул. Бережанська, 2, м. Тернопіль 46004
тел. (0352) 47-58-72
E-mail: edition@tneu.edu.ua