

519,7(075)
1934

Б.В. Шамша, А.М. Гуржій, З.В. Дудар, В.М. Левикін

**МАТЕМАТИЧНЕ
ЗАБЕЗПЕЧЕННЯ
ІНФОРМАЦІЙНО-
КЕРУЮЧИХ
СИСТЕМ**

3919-39
013.1(013)
МЗ4
А.М.Гуржій, З.В.Дудар, В.М.Левикін, Б.В.Шамша

МАТЕМАТИЧНЕ ЗАБЕЗПЕЧЕННЯ ІНФОРМАЦІЙНО- КЕРУЮЧИХ СИСТЕМ

Затверджено Міністерством освіти і науки України
як підручник для студентів вищих навчальних закладів,
які навчаються за напрямом «Комп'ютерні науки»

ІАБОЧЕМЕНТ-2

Харків
«Компанія СМІТ»
2006

УДК 519.71
ББК 004.338.8
Г95

*Затверджено Міністерством освіти і науки України
як підручник для студентів вищих навчальних закладів,
які навчаються за напрямом «Комп'ютерні науки»
(лист № 14/18.2-1427 від 23.06.2004 р.)*

Видано за рахунок державних коштів. Продаж заборонено

Рецензенти:

А. І. Пушкар, професор, доктор економічних наук,
завідувач кафедри обчислювальної техніки і програмування
Харківського національного економічного університету;

І. А. Фурман, професор, доктор технічних наук,
завідувач кафедри автоматизації і комп'ютерних технологій
Харківського національного технічного університету
сільського господарства ім. П. Василенка

Гуржій А. М., Дудар З. В., Левикін В. М., Шамша Б. В.

Г95 Математичне забезпечення інформаційно-керуючих
систем: Підручник для студентів вищих навчальних закладів.— Харків: ТОВ «Компанія СМІТ», 2006.— 448 с.

ISBN 966-8530-49-7

У підручнику викладено теоретичні основи методів, застосовуваних при розробленні математичного забезпечення ІУС, розглянуто умови їх коректного використання для розв'язання завдань у різних предметних галузях. Подано приклади розв'язання задач за допомогою ППП «STATISTICA», «STATGRAPHICS PLUS».

Для студентів вищих навчальних закладів, аспірантів та викладачів дисциплін за напрямом «Комп'ютерні науки».

УДК 519.71
ББК 004.338.8

ISBN 966-8530-49-7

© А. М. Гуржій, З. В. Дудар,
В. М. Левикін, Б. В. Шамша, 2006
© ТОВ «Компанія СМІТ», 2006



ЗМІСТ

ПЕРЕДМОВА	5
РОЗДІЛ 1. СТРУКТУРА МАТЕМАТИЧНОГО ЗАБЕЗПЕЧЕННЯ ІНФОРМАЦІЙНО-УПРАВЛЯЮЧИХ СИСТЕМ	7
1.1. Основні визначення математичного забезпечення інформаційно-управляючих систем	7
1.2. Узагальнена структура математичного забезпечення	11
1.3. Задачі функціональних підсистем в ІУС для підприємства з дискретним характером виробництва	18
РОЗДІЛ 2. МЕТОДИ ОБРОБКИ ДАНИХ В ІНФОРМАЦІЙНО-УПРАВЛЯЮЧИХ СИСТЕМАХ	25
2.1. Дані. Типи даних	26
2.2. Вибіркові дані	29
2.3. Основні статистичні характеристики, що використовуються при розв'язанні задач в інформаційно-управляючих системах	39
2.4. Багатовимірна нормально розподілена генеральна сукупність	44
2.5. Довірчі області	47
2.6. Стислий аналіз методів обробки даних	56
РОЗДІЛ 3. КОРЕЛЯЦІЙНИЙ АНАЛІЗ	93
3.1. Коефіцієнти, що характеризують ступінь статистичного зв'язку ..	94
3.2. Побудова кореляційного рівняння	116
3.3. Приклад побудови кореляційної моделі курсу валют за фундаментальними факторами	123
3.4. Застосування кореляційного аналізу для оцінки показників ефективності роботи підприємств	150
РОЗДІЛ 4. РЕГРЕСІЙНИЙ АНАЛІЗ І ЙОГО ВИКОРИСТАННЯ У МАТЕМАТИЧНОМУ ЗАБЕЗПЕЧЕННІ ІНФОРМАЦІЙНО-УПРАВЛЯЮЧИХ СИСТЕМ	160
4.1. Поняття регресії	160
4.2. Побудова математичних моделей в ІУС з використанням регресійного аналізу	162

4.3. Методи зміщеного оцінювання.....	218
4.4. Робастні методи оцінювання.....	232
4.5. Приклад побудови регресійної моделі та її інтерпретація за даними таблиці значень показників виробничо-господарської діяльності підприємства з використанням ППП «STATISTICA»....	239
4.6. Приклад побудови моделі курсу валют за фундаментальними факторами з використанням ППП «STATISTICA»	285
РОЗДІЛ 5. ДИСПЕРСІЙНИЙ АНАЛІЗ	295
5.1. Теоретичні основи дисперсійного аналізу	295
5.2. Однофакторний дисперсійний аналіз	298
5.3. Приклад використання однофакторного дисперсійного аналізу.....	305
5.4. Рангові методи однофакторного дисперсійного аналізу.....	311
5.5. Однофакторний дисперсійний аналіз у пакетах прикладних програм	316
5.6. Двофакторний дисперсійний аналіз	322
РОЗДІЛ 6. КЛАСТЕРНИЙ АНАЛІЗ	331
6.1. Галузі застосування.....	331
6.2. Постановка задачі	332
6.3. Короткий аналіз методів кластеризації.....	334
6.4. Міра близькості між об'єктами або ознаками	335
6.5. Відстань між кластерами	344
6.6. Функціонали якості розбиття	347
6.7. Використання ППП для розв'язання задачі кластеризації.....	352
РОЗДІЛ 7. ВИКОРИСТАННЯ ДИСКРИМІНАНТНОГО АНАЛІЗУ ДЛЯ ОБРОБКИ ТЕХНІКО-ЕКОНОМІЧНОЇ ІНФОРМАЦІЇ В ІНФОРМАЦІЙНО-УПРАВЛЯЮЧИХ СИСТЕМАХ.....	374
7.1. Постановка задачі	374
7.2. Обчислення дискримінантних функцій і їх інтерпретація	381
7.3. Байєсівська процедура класифікації.....	387
7.4. Класифікація у випадку двох багатовимірних нормальних популяцій за невідомих параметрів	393
7.5. Покроковий дискримінантний аналіз.....	395
7.6. Оцінка застосовності дискримінантного аналізу при порушенні припущень	402
7.7. Використання ППП «STATISTICA» для розв'язання задач класифікації.....	404
ЛІТЕРАТУРА	424
ДОДАТОК	427

ПЕРЕДМОВА

Сучасне виробництво характеризується складністю технологічного процесу, наявністю великої кількості інформації, яка пов'язує різні аспекти соціально-економічної діяльності підприємств у різних предметних галузях (медицина, комерція, соціологія, наукові дослідження, технічні системи тощо). Відбувається різке збільшення документообігу та кількості службовців, зайнятих обробленням інформації. Постійно вдосконалюються технології збору, передавання та зберігання інформації, оскільки без ефективного та своєчасного оброблення вона втрачає свою цінність і стає мертвим вантажем.

За таких умов застосування інформаційно-управляючих систем (ІУС), базованих на сучасних досягненнях у галузі інформаційних технологій і математичних методах оброблення інформації, стає важливим фактором розвитку суспільного виробництва. Ефективність функціонування ІУС значною мірою залежить від якості розроблення математичного забезпечення (МЗ). Підсистема МЗ відноситься до забезпечувальних підсистем, спрямованих на розв'язання функціональних задач підприємства, і значною мірою зумовлюється структурою ІУС.

МЗ ІУС являє собою різноманітність задач у різних сферах людської діяльності, для розв'язання яких застосовуються економіко-математичні методи та алгоритми оброблення даних, напрямлених на вироблення та реалізацію керуючих впливів.

На цей час процес розроблення та проектування різних елементів МЗ, особливо задач, які стосуються питань побудови математичних моделей, їх інтерпретації та використання в управлінській діяльності, не доведений до інженерного рівня, тому цей процес є предметом науково-дослідницької діяльності, яка є вельми трудомісткою і потребує високої кваліфікації виконавців.

Метою написання цього навчального посібника є викладення основної частини методів, застосовуваних при розробленні МЗ ІУС, та їх

коректного використання для розв'язання завдань у різних предметних галузях.

Зауважимо, що арсенал методів і алгоритмів великий. Кожний алгоритм орієнтований, як правило, на вирішення конкретного класу питань і забезпечує ефективне розв'язання лише за дотримання певних теоретичних передумов, що становить перешкоду для їх коректного застосування. Вибір того чи іншого алгоритму для конкретного об'єкта ускладнений необхідністю проведення аналізу та оцінки їх ефективності у відповідності до класу об'єкта та його характеристик. Також слід зауважити, що більшість гіпотез, які розглядаються, базовані на дотриманні гіпотез щодо незалежності, сталості статистичних характеристик, нормальності законів розподілу помилок моделювання, залежних і незалежних змінних. У реальних умовах виконання таких передумов є малоімовірним. На цей час на заміну лінійної концепції оброблення даних прийшли методи, що ґрунтуються на нелінійному динамічному підході. Ці нові методи мають включати нелінійну динаміку, фрактали, еволюційні алгоритми та інші засоби, застосування яких забезпечує більш високу результативність оброблення реальних даних.

У навчальному посібнику вчинено спробу подати процедуру розв'язання задач в ІУС за допомогою пакетів прикладних програм «STATISTICA», «STATGRAPHICS PLUS» на інженерному рівні. Викладення методів розв'язання задач в ІУС за допомогою означених програмних засобів супроводжено великою кількістю прикладів.

РОЗДІЛ 1

СТРУКТУРА МАТЕМАТИЧНОГО ЗАБЕЗПЕЧЕННЯ ІНФОРМАЦІЙНО-УПРАВЛЯЮЧИХ СИСТЕМ

1.1. Основні визначення математичного забезпечення інформаційно-управляючих систем

Сучасні інформаційно-управляючі системи (ІУС) призначені для формування і реалізації ефективних управляючих впливів на об'єкт управління і являють собою складну ієрархічну структуру, що складається з багатьох підсистем, у рамках яких розв'язуються різноманітні функціональні задачі. Типи задач, методи їх розв'язання, алгоритми обробки інформації представлені в ІУС у вигляді математичного забезпечення.

Математичне забезпечення (МЗ) являє собою сукупність сучасних математичних методів, моделей, алгоритмів і програм, що забезпечують розв'язання функціональних задач у системі управління. Алгоритми пов'язані між собою певною структурою, яка відповідає специфіці об'єкта управління та сприяє реалізації функцій системи. Для кожного виробничого процесу та рівня його управління функції різні, тому як структура, так і набір алгоритмів можуть мати різний характер. МЗ реалізується у вигляді програм спеціального програмного забезпечення.

Методи алгоритмізації об'єктів і систем можна поділити на:

- ☒ теоретичні (аналітичні);
- ☒ експериментальні.

При розробленні алгоритмічних описів необхідно сформулювати мету та завдання управління. Виконуючи аналіз об'єкту управління (ОУ), необхідно визначити фізичні обмеження на параметри та обмеження, що накладаються умовами нормальної експлуатації. У процесі функціонування об'єктів їх характеристики з часом змінюються, що призводить до необхідності корегування та адаптації МЗ до існуючого стану об'єкта. У такому випадку ці задачі виконує обслуговуючий персонал.

У якості мов алгоритмічних описів застосовуються:

- ☒ опис за допомогою слів;
- ☒ принципові схеми процесів;
- ☒ структурні схеми алгоритмів функціонування;
- ☒ математичні описи.

Алгоритмічний опис найбільш показово може бути представлений графічно у вигляді структурної схеми алгоритму функціонування об'єкта, який відображає усі функціональні зв'язки між окремими частинами реального об'єкта.

Теорія та практика розробки математичного і програмного забезпечення ІУС показує, що найбільш перспективним шляхом розвитку є автоматизація проектування прикладного математичного забезпечення (ПМЗ) шляхом створення інтелектуальних систем обробки даних.

ПМЗ, що проектується, має не тільки відповідати своєму функціональному призначенню, але й задовольняти певним непрямым вимогам, таким як ефективність, надійність, модифікованість, адекватність, модульність.

Формалізація стратегії вибору структури та елементів МЗ значною мірою залежить від правильного вибору показників якості алгоритмічних та програмних модулів, їх кількісних оцінок, критеріїв вибору.

Модуль — це функціонально незалежний блок, що має зв'язок з іншими модулями за допомогою міжмодульного інтерфейсу. Модуль може бути представлений різноманітними алгоритмічними схемами пошуку рішень.

Алгоритм — це послідовність процедур переробки вихідних даних у потрібний результат. Сукупність вихідних даних, до яких застосовується алгоритм, що дозволяє за цими вихідними даними отримати потрібний результат, є областю застосування алгоритму.

Алгоритм є логічною структурою програми, її алгоритмічною схемою.

До основних якісних характеристик алгоритму слід віднести: точність, складність, надійність, універсальність, зв'язаність.

Функція складності може бути визначена як сума

$$C[q(n, m)] = a_l c_l [h(n, m)] + a_f c_f [z(n, m)],$$

де $c_l c_f$ — функції складності логічних та функціональних блоків відповідно; $a_l a_f$ — трудомісткість; n, m — показники розмірності задачі (обсяг вибірки та кількість вимірювань).

Під надійністю алгоритму розуміють здатність зберігати певні характеристики при зміні конкретних параметрів у постановці задачі.

Під універсальністю розуміють його здатність забезпечувати припустимі рішення для різноманітних класів задач. Ступінь універсальності алгоритму визначається кількістю класів $K(n_j)$, які містять усі задачі з вихідними даними. Характеристика n_j може знаходитися у неприпустимій області $\Omega_j \in \Omega$, де Ω_j — припустимий інтервал за координатою h_j , для якої алгоритм може бути ефективно застосовано.

Складність алгоритму C також залежить від кількості елементарних операцій N_a , необхідних для його виконання, від зв'язності (П), що визначає потрібний об'єм пам'яті під час реалізації алгоритму, від складності структури, складу та рівня операторів.

Важливим показником, що характеризує алгоритм, є зв'язність, яка поділяється на внутрішню, зовнішню і повну. Внутрішня зв'язність (П_а) оцінюється максимальною кількістю проміжних результатів, які необхідно запам'ятати під час виконання алгоритму, та кількістю вихідних даних. Зовнішня зв'язність (П_б) визначається кількістю вихідних даних. Загальна зв'язність П_а + П_б визначається довжиною запису алгоритму в прийнятому наборі операторів або числом команд у програмі, що реалізує довгий алгоритм.

Величина зв'язності алгоритму залежить від прийнятого числового методу розв'язання задачі. Бажано застосовувати такі методи числового розв'язання, які забезпечують найменшу зв'язність. При цьому точність результату, що отримується зменшується однаково.

Під час реалізації алгоритмів стикаються з обмеженнями, що накладаються на ЕОМ, які застосовуються для автоматизованих промислових об'єктів та мають невеликий обсяг пам'яті, обмежену продуктивність тощо.

Розв'язання задач в ІУС передбачає такі обов'язкові елементи: визначення задачі, розроблення математичної моделі, вибір обмежень та критерію ефективності або оптимізації, програмну реалізацію.

Коректний вибір задачі визначає успіх усієї подальшої роботи. Під час вибору задачі необхідно зважати на фактори: важливість та принципову можливість розв'язання задачі, існування різноманітних варіантів розв'язання.

Математична модель призначена для змістовного вибору задачі за допомогою математичних символів, вона являє собою аналітичну залежність між змінними, значення яких треба знайти в результаті розв'язання задачі, та вихідними даними. Складання математичної моделі — це багатоетапний процес. На першому етапі необхідно детально вивчити об'єкт, що моделюється. Скласти параметричну схему об'єкта. Визначити змінні, що контролюються, не контролюються, регулюються. Представити процес як об'єкт автоматичного управління. Розробити методику отримання даних, які б максимально відображали економічну та інформаційну сутність об'єкта. Далі необхідно оцінити тип моделі (динамічна чи статична). Визначити доцільність її використання для прогнозування, дослідження та управління. Тип моделі має відповідати класу об'єкта.

Клас об'єкта визначається за наступними показниками: детермінованість та стохастичність; одновимірність та багатовимірність; статичність та динамічність; стаціонарність та нестаціонарність тощо.

Після визначення класу моделі обирається її структура. Визначається метод побудови моделі. Особливу увагу треба приділяти дотриманню передумов та пропозицій методу, що використовується. Далі оцінюються параметри моделі. Проводиться оцінка адекватності та інтерпретація результатів побудови моделі. Перевіряються статистичні гіпотези та критерії, прийняті у методі ідентифікації, що використовується.

Після побудови моделі обираються обмеження та критерії оптимізації для цієї задачі. Далі обирається метод розв'язання оптимізаційної задачі.

1.2. Узагальнена структура математичного забезпечення

Структура математичного забезпечення формується в залежності від типу ІУС, функціональних задач, що вирішуються, засобів математичного забезпечення, що використовуються, та засобів автоматизації проектування МЗ ІУС.

На рис. 1.1 представлено узагальнену структуру МЗ ІУС. Незалежно від типу ІУС, у структурі МЗ повинні бути присутні такі елементи:

☑ засоби математичного моделювання економічних, інформаційних, технологічних, організаційних та інших процесів проектування та управління, що містять аналітичні, статистичні, імітаційні та інші моделі, що забезпечують функціонування об'єктів проектування та управління;

☑ методи вибору та засоби проектування;

☑ документація на МЗ ІУС;

☑ персонал та його функції.

Кожен блок задач, який представлено на рис. 1.1, складається з багатьох підзадач. Незважаючи на велику різноманітність задач, що розв'язуються на різних рівнях управління, майже в усіх випадках необхідно мати математичні моделі для розв'язання задач планування, управління, прогнозування та дослідження.

Специфіка функціонування інформаційно-управляючих систем визначає основний тип задач, що розв'язуються, та методів їх розв'язання.

Перелік функціональних задач, що вирішуються у банківських ІС, представлено на рис. 1.2. На рис. 1.3 подано структуру МЗ технічних систем. Перелік функцій, які необхідні для розв'язання основних задач у техніко-економічних інформаційно-управляючих системах, наведений нижче.

Для розв'язання практичних задач ідентифікації та управління в ІУС існує достатньо велика кількість алгоритмів. Однак, незважаючи на те, що арсенал алгоритмів досить великий, виникає чимало перепон для їх ефективного використання. Кожен алгоритм орієнтований, як правило, на розв'язання конкретного класу задач, та забезпечує ефективність розв'язання лише за деяких теоретичних припущень.

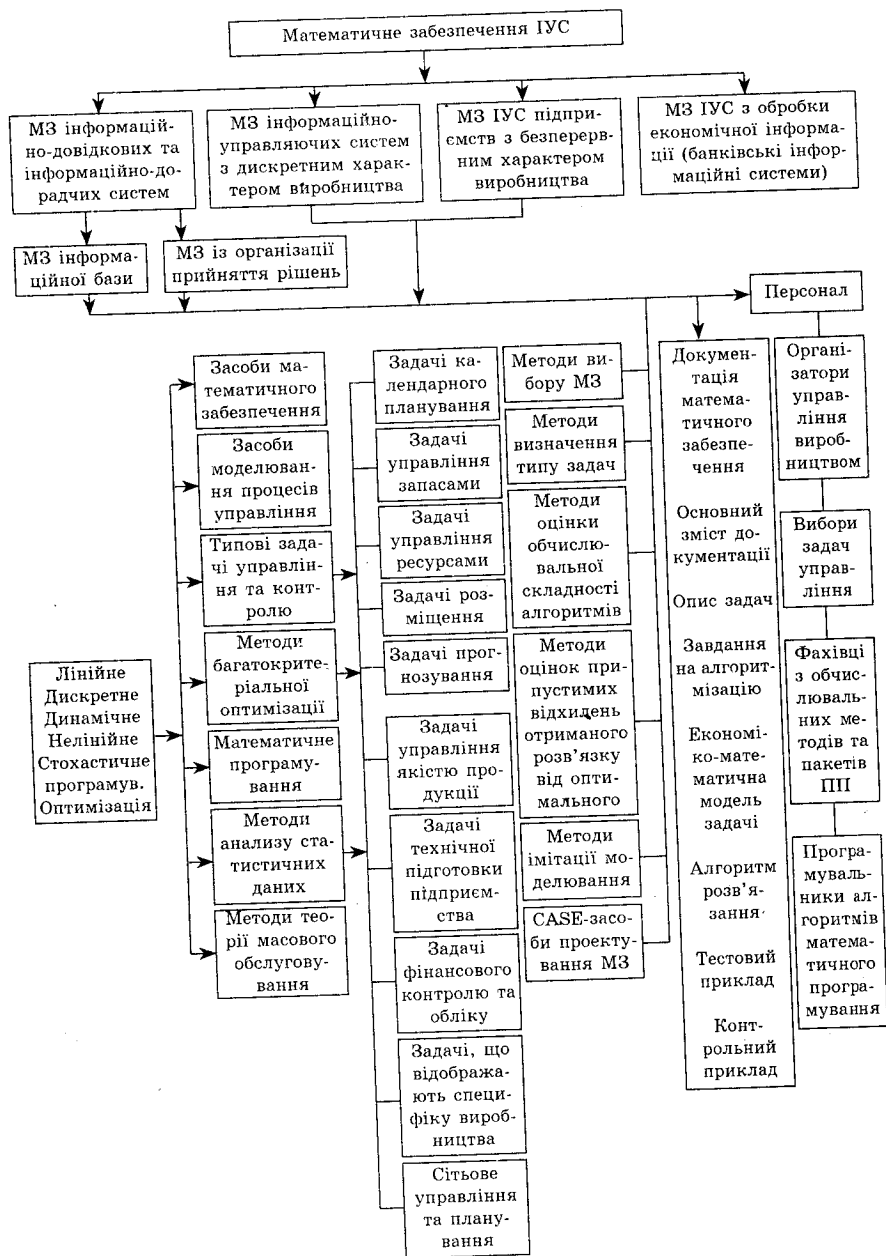


Рис. 1.1. Узагальнена структура МЗ ІУС

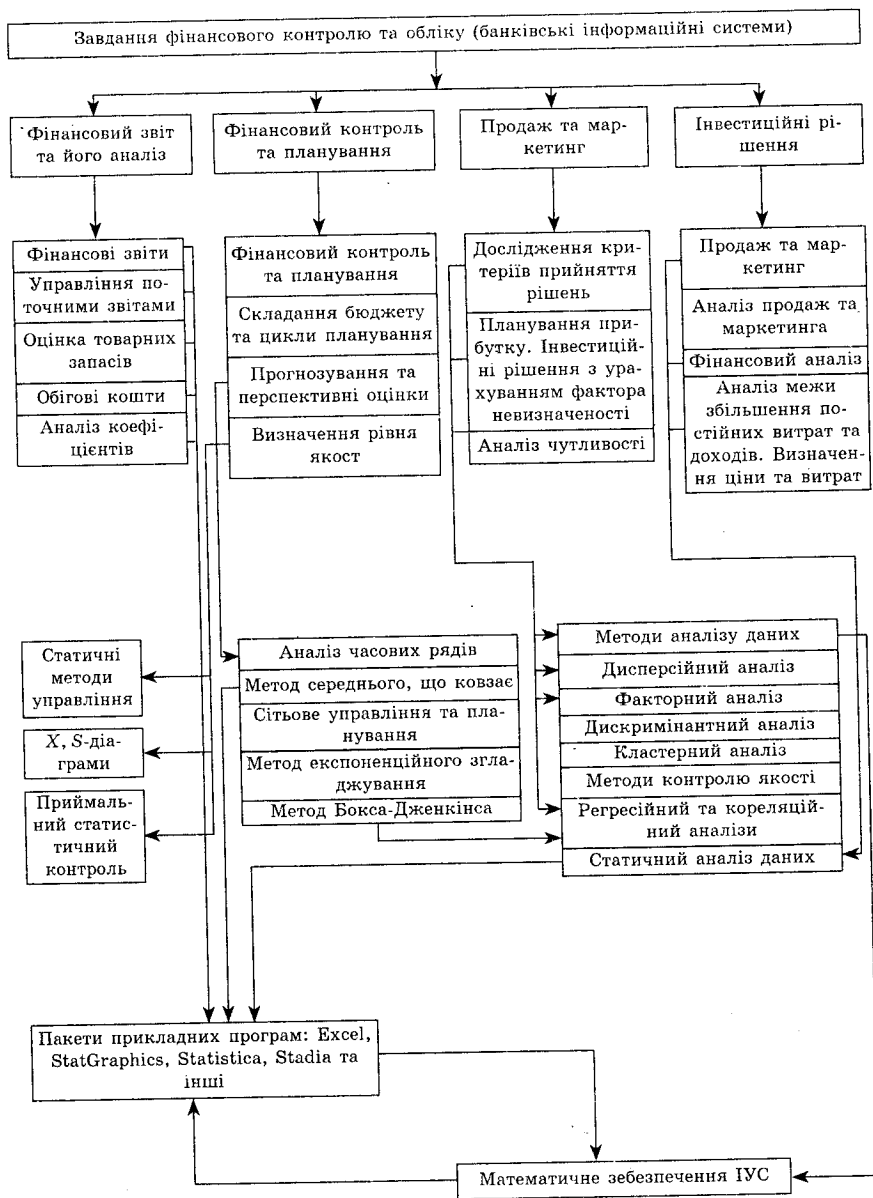


Рис. 1.2. Структура функціональних задач фінансового контролю та обліку на прикладі банківських ІС

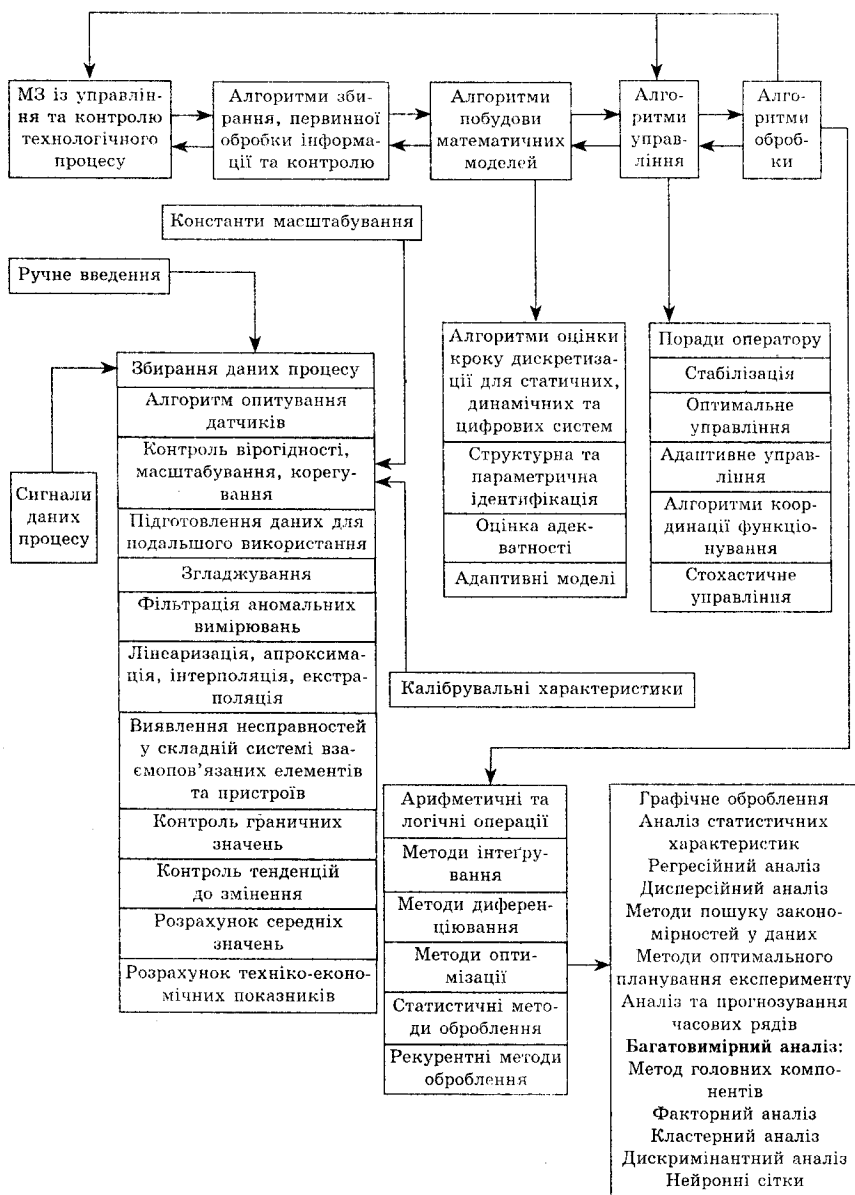


Рис. 1.3. Структура функціональних задач математичного забезпечення ІУС технологічним процесом

Вибір того чи іншого алгоритму для конкретного об'єкта являє певну складність, пов'язану з аналізом та оцінкою його ефективності в залежності від класу об'єкта.

При всій різноманітності найновітніших математичних, статистичних та інших методів та підходів до прийняття рішень в умовах невизначеності та часткової відсутності інформації про об'єкт, що досліджується, досить мало уваги приділено автоматизації вибору методу розв'язання задач в ІУС. Як правило, реальні процеси характеризуються великою кількістю різноманітних збурень, що не контролюються. Важливим моментом для визначення закономірності в поведінці кожного досліджуваного процесу є виявлення внутрішньої структури та властивостей вихідної інформації, що дозволить підійти до вибору необхідного методу її оброблення.

Сьогодні вибір методу розв'язання тієї чи іншої функціональної задачі здійснюють користувачі. Відсутність у структурі математичного забезпечення сучасних ІУС блоку автоматизованого вибору методу побудови математичної моделі процесу, що досліджується, зводить функції апаратно-програмних засобів до функцій первинного оброблення інформації, контролю, supervisory управління, діагностики несправностей, поточного ремонту тощо. За таких умов доцільно включити до структури МЗ ІУС комплекс програмних засобів, що забезпечує підтримку прийняття рішення щодо вибору методів розв'язання певних функціональних задач ІУС. Цей комплекс має входити як складова частина до комплексу засобів математичного забезпечення та являти собою систему взаємопов'язаних програмних модулів. Комплекс програмних засобів призначено для розв'язання задач статистичних випробувань, оцінки ефективності методів ідентифікації, зв'язку показників ефективності методів побудови моделі зі статистичними характеристиками об'єкта, що ідентифікується, вибору найбільш ефективного методу для конкретного об'єкта управління. Кожен модуль має містити взаємопов'язані програми, які забезпечують розв'язання конкретної задачі.

Нині існує тенденція впровадження у системи управління виробництвом, зокрема для формування управляючого впливу та формалізації оцінок ефективності прийняття рішень, сучасних засобів, базованих на технології та методах штучного інтелекту (ШІ). При цьому виробничі системи зі штучним інтелектом розглядаються не лише як системи якісно нового типу, але й як органічна ланка у структурі систем управління виробництвом.

Незважаючи на значні успіхи у галузі ІІІ, ще існує певний розрив між існуючими апаратними та програмними засобами ІІІ і можливостями їх практичного застосування на виробництві, що потребує подальших досліджень з їх вдосконалення з урахуванням конкретної галузі застосування.

Технології ІІІ охоплюють штучні нейронні сітки (НС), експертні системи (ЕС), системи підтримки прийняття рішень (СППР), методи, що використовують нечітку логіку (НЛ), генетичні алгоритми тощо. Ідеї, на яких вони базуються, істотно відрізняються від загальноприйнятих методів обчислень, оскільки вони імітують притаманні природному інтелекту шляхи вирішення проблем, тобто більш зрозумілі технологу, ніж «чистому» фахівцю з обчислювальної техніки; або імітують «природний», «генетичний» розвиток процесів. Наприклад, НС мають здатність до навчання, ЕС приймають рішення на основі наборів правил та досвіду експертів, а системи з нечіткою логікою оперують такими поняттями, як невизначеність та частково-приблизна істина.

Ці технології призначені для розв'язання складних задач, які або перевищують можливості загальноприйнятих алгоритмічних методів, або потребують для свого розв'язання надто великих матеріальних та часових витрат. Зважаючи на свою специфіку, методи, що лежать в основі зазначених технологій, доволі суперечливі й не завжди приймаються прихильниками традиційних підходів. Тим не менш, сьогодні ІІІ-технологія займає все більш важливе місце в арсеналі засобів розробників ІУС.

На рис. 1.4 представлено концептуальну модель вибору методу обробки даних для різних предметних галузей. На основі цієї моделі розробляються CASE-засоби проектування МЗ ІУС. У системах ІУС вирішується, як правило, два типи задач: інформаційні та оптимізаційні.

Інформаційні задачі не пов'язані з прийняттям управлінських рішень і не мають власного економічного критерію. Вони готують інформацію для задач управління та непрямо змінюють стан об'єкта.

Для розв'язання інформаційних задач в арсеналі МЗ ІУС необхідно мати такі методи аналізу та дослідження інформаційних потоків:

- ☒ аналіз потоків інформації за допомогою графічного методу (зокрема, теорії графів);
- ☒ метод функціонально-операційного аналізу;

- ☑ метод аналізу норм виробітку рішень;
- ☑ модуль-метод;
- ☑ метод матричного моделювання;
- ☑ метод семіотичного аналізу;
- ☑ метод схем інформаційних зв'язків планових розрахунків;
- ☑ метод реквізитів;
- ☑ метод послідовного аналізу задач управління;
- ☑ метод аналізу та оптимізації внутрішньозаводських документопотоків за допомогою транспортної моделі.

Розв'язання оптимальних задач в ІУС вимагає наявності у системі МЗ великого арсеналу методів побудови моделей та методів оптимізації.

Розглянемо короткий перелік задач, що розв'язуються в ІУС, для підприємств з дискретним характером виробництва.

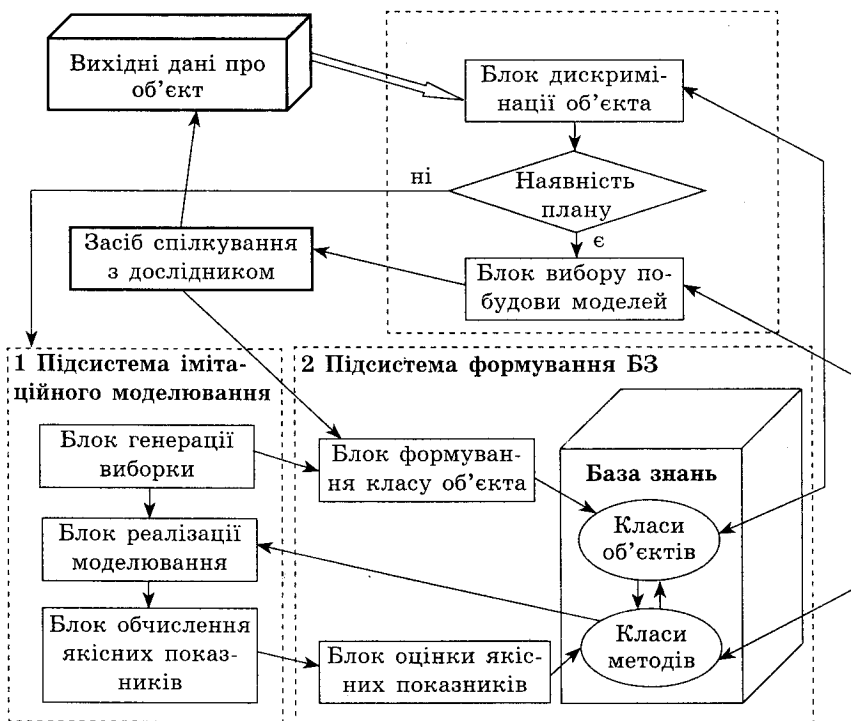


Рис. 1.4. Структурна схема інтелектуального комплексу (ІК) вибору методу побудови моделей

1.3. Задачі функціональних підсистем в ІУС для підприємства з дискретним характером виробництва

Задачі в інформаційно-управляючих системах розділяються на підсистеми таким чином:

- ☑ за циклами управління (для послаблення вимог до динамічних властивостей моделі);
- ☑ за виробничими ознаками (для послаблення вимог до нелінійних властивостей моделі).

З точки зору динамічних властивостей прийнято вирізняти такі задачі:

- ☑ прогнозування розвитку об'єкта управління;
 - ☑ перспективне планування об'єкта управління;
 - ☑ поточне планування діяльності об'єкта управління;
 - ☑ оперативне регулювання діяльності об'єкта управління;
 - ☑ облік, контроль та аналіз діяльності об'єкта управління.
- За виробничою ознакою вирізняють такі типи задач:
- ☑ управління НДР і ДКР;
 - ☑ управління формуванням виробничої структури конкретного об'єкта;
 - ☑ управління основним виробництвом;
 - ☑ управління допоміжно-обслуговуючими виробництвами конкретного об'єкта;
 - ☑ управління транспортними процесами;
 - ☑ управління виробничими потужностями та процесом використання основних фондів;
 - ☑ управління матеріально-технічним постачанням;
 - ☑ управління трудовими ресурсами;
 - ☑ управління збутом продукції;
 - ☑ управління фінансовими та грошовими коштами конкретного об'єкта;
 - ☑ управління розвитком та функціонуванням власне управляючої системи.
 - ☑ управління технічною підготовкою виробництва.

Розглянемо детальніше перелік задач, що класифікуються за виробничою ознакою.

Управління матеріально-технічним постачанням

Основною задачею матеріально-технічного постачання (МТП) є забезпечення виробництва необхідними матеріалами, комплектуючими та іншими виробами за дотримання режиму економії у витратах матеріальних ресурсів.

Задачі МТП можна розділити на групи:

- ☑ визначення потреби у матеріалах, комплектуючих та виробих на виробничу програму;
- ☑ розроблення плану постачання-складання заявок (не фонди) та розподіл фондів, що виділено;
- ☑ оперативний облік та управління рухом матеріальних ресурсів;
- ☑ бухгалтерський облік матеріальних ресурсів.

Перелік задач, що вирішуються в підсистемі управління матеріально-технічним постачанням:

- ☑ розрахунок потреби в матеріалах у звідній номенклатурі (по підприємству — на рік);
- ☑ розрахунок специфікованої потреби в комплектуючих виробих для замовлення (по підприємству — на рік);
- ☑ розрахунок специфікованої потреби в матеріалах для замовлення (по підприємству — на рік, півріччя);
- ☑ розрахунок специфікованої потреби в матеріалах на товарний випуск (по підприємству — на рік, півріччя);
- ☑ розрахунок специфікованої потреби в комплектуючих виробих на товарний випуск (по підприємству — на рік, півріччя, квартал);
- ☑ розрахунок потреби в матеріалах у звідній номенклатурі для виробництва запасних частин (по підприємству — на рік);
- ☑ розрахунок лімітів матеріалів (по цеху — на квартал, місяць);
- ☑ розрахунок лімітів комплектуючих виробів (по цеху — на квартал, місяць);
- ☑ облік забезпеченості матеріалами (по складу підприємства — за добу, місяць);
- ☑ облік забезпеченості комплектуючими виробами (по складу — підприємства за місяць);
- ☑ оперативний облік руху матеріалів (по складу підприємства — за добу, місяць);

- ☒ оперативний облік руху комплектуючих виробів (по складу підприємства — за добу, місяць);
- ☒ оперативний облік реалізації фондів за комплектуючими виробами (по складу підприємства — за добу, місяць);
- ☒ оперативний облік реалізації фондів за матеріалами (по складу підприємства — за добу, місяць);
- ☒ контроль за станом запасів матеріалів (по складу підприємства — за добу, місяць);
- ☒ контроль за станом запасів комплектуючих виробів (по складу підприємства — за добу, місяць);
- ☒ статистична звітність з матеріалів (по підприємству — за квартал).

Управління основним виробництвом

Управління основним виробництвом має забезпечити випуск за планово-обліковий період у складальних цехах такого набору виробів, який визначений відповідним комплексом замовлень. До таких **задач управління основним виробництвом** слід віднести:

- ☒ розрахунок розгорнутого плану потреби деталей, складальних одиниць на товарний випуск продукції (по підприємству — на рік, квартал, місяць);
- ☒ розрахунок плану потреби деталей, складальних одиниць на товарний випуск продукції підприємства (по цеху — на рік, квартал, місяць);
- ☒ розрахунок плану здавання та отримання деталей, складальних одиниць у натуральному виразі (по цеху — на рік, квартал, місяць);
- ☒ розрахунок оперативного плану здавання (отримання) деталей, складальних одиниць (по цеху — на місяць);
- ☒ розрахунок корегованого плану потреб деталей, складальних одиниць на змінений план по výroбах на товарний випуск продукції (по підприємству, цеху — на місяць);
- ☒ розрахунок оперативного календарного плану складання виробів (по цеху — на місяць);
- ☒ розрахунок оперативно-календарного плану виробництва (по дільницях заготовчих цехів — на місяць, добу);
- ☒ розрахунок оперативного плану здавання деталей, складальних одиниць (по дільниці цеха — на місяць);

- ☒ розрахунок оптимальних коопераційних планів-графіків (по цеху — на добу, зміну);
- ☒ розрахунок завантаження обладнання (по цехах, ділянках — на рік, квартал);
- ☒ оперативний облік виконання плану міжцехових поставок у натуральному виразі (по підприємству — за добу, з початку місяця).

Бухгалтерський облік

Бухгалтерський облік (БО) призначено для забезпечення своєчасного та повного обліку грошових коштів, товарно-матеріальних цінностей та основних коштів, що надходять у відповідну організацію, а також відображення їх руху; достовірного обліку витрат виробництва та обігу, виконання кошторису витрат реалізації продукції, виконання будівельно-монтажних та інших робіт, складання калькуляцій собівартості продукції, робіт та послуг; точного обліку та економічного аналізу результатів господарсько-фінансової діяльності організації; правильного нарахування та своєчасного перерахування платежів тощо.

Перелік задач, що вирішуються у підсистемі бухгалтерського обліку:

- ☒ облік праці та заробітної платні;
- ☒ облік матеріальних цінностей та розрахунків з постачальниками;
- ☒ облік готової продукції, її відвантаження та реалізації;
- ☒ облік основних коштів;
- ☒ облік браку у виробництві;
- ☒ облік грошових коштів;
- ☒ облік витрат на виробництво;
- ☒ звітний бухгалтерський облік та звітність, облік за іншими рахунками.

Задачі автоматизації управління технічною підготовкою виробництва:

- ☒ розрахунок складу та обсягу робіт з підготовки виробництва виробів;
- ☒ складання сітьової моделі плану виконання робіт з підготовки виробництва;
- ☒ планування підготовки виробництва за методом критичного шляху;

☑ розрахунок планових об'ємних показників технічної підготовки виробництва на рік, квартал (по виробу, підрозділу);

☑ розрахунок потрібної кількості ресурсів на технічну підготовку виробництва (по підрозділах на рік, квартал);

☑ оптимізація сіткових моделей планів підготовки виробництва;

☑ оперативний облік виконання планів підготовки виробництва та корегування планів;

☑ статистичний облік інформації за фактичною трудомісткістю робіт технічної підготовки виробництва;

☑ формування нормативів трудомісткості робіт технічної підготовки виробництва.

Управління збутом продукції

Основою функціонування підсистеми збуту є розробка оперативних календарних графіків відвантаження, що забезпечують своєчасне надходження сум за реалізовану продукцію.

Вихідними даними для розв'язання комплексу задач управління збутом продукції (УЗП) є такі відомості: найменування, вантажні адреси, розрахункові рахунки та інші реквізити споживачів, план постачань на рік з розбиттям по кварталах, відомості про постачання, відомості про виконання плану виробництва тощо.

До задач підсистеми управління збутом готової продукції можна віднести:

☑ формування плану постачань готової продукції (по підприємству — за рік, квартал, місяць);

☑ оперативно-календарне планування відвантаження продукції (по підприємству — на місяць, добу);

☑ облік руху готової продукції на складах підприємства (за добу, місяць);

☑ облік готової відвантаженої та реалізованої продукції (по підприємству — за добу, декаду, місяць);

☑ оперативний облік виконання плану відвантаження продукції (по складу підприємства — за добу, місяць);

☑ оперативний контроль за формуванням портфеля замовлень (по підприємству — на рік, квартал);

☑ складання статистичного звіту про виконання плану постачань продукції (по підприємству — за місяць, квартал, рік);

- ☒ оперативний облік виконання плану реалізації та прибутку (по підприємству — за добу, декаду, місяць);
- ☒ прогнозування сум реалізації та прибутку;
- ☒ нарахування податку з обігу.

Техніко-економічне планування — основа управління виробництвом

До основних задач техніко-економічного планування (ТЕП) належать:

- ☒ розрахунок виробничої програми;
- ☒ обчислення нормативної собівартості товарного випуску продукції;
- ☒ нормативні калькуляції на деталь, вузол, готовий виріб;
- ☒ розрахунок нормативної трудомісткості виробничої програми;
- ☒ визначення планової кількості виробничих працівників;
- ☒ розрахунок планового фонду заробітної платні виробничих працівників.

Короткий аналіз методів розв'язання функціональних задач в ІУС поданий у розділі 2.

Для розв'язання комплексу задач техніко-економічного планування необхідно розробити балансні моделі, моделі об'ємного планування, моделі календарного планування, моделі управління запасами.

Балансові моделі пов'язують кількісні потоки ресурсів та продукції, виробничі потужності підприємств та завантаження обладнання. Часто вони задаються системою лінійних рівнянь. Наявність критеріїв та обмежень зводить задачу прогнозування поведінки підприємства до задачі лінійного програмування, цілочислового програмування.

Задачі календарного планування зводяться до методів дискретної оптимізації. Залежно від конфігурації символів, пошук рішення може здійснюватися у вигляді послідовності символів, графіка Санта або матриці тощо. Для розв'язання задач з безперервними функціями використовуються методи послідовного конструювання.

Залежно від дерева варіантів, використовуються методи повного перебору, градієнтні методи, методи гілок та меж, методи пошуку з адаптацією, евристичні методи. Всі вони на основі

модельовання класифікуються за якісними ознаками, створюючи тим самим базу знань. Деякі методи математичного програмування, аналізу даних, оптимізації та галузі їх використання наведені нижче.

Контрольні запитання

1. Що являє собою математичне забезпечення?
2. Дайте визначення поняття «модуль».
3. Дайте визначення поняття «алгоритм».
4. Для чого призначена математична модель?
5. Які існують методи аналізу та дослідження інформаційних потоків?
6. Проаналізуйте узагальнену структуру МЗ ІУС.
7. За якими ознаками розбиваються задачі в ІУС на підсистеми?
8. Що є основною задачею МТП?
9. Що слід віднести до задач управління основним виробництвом?
10. Для чого призначений бухгалтерський облік?
11. Які задачі належать до задач автоматизації управління технічною підготовкою виробництва?
12. Що належить до основних задач ТЕП?

РОЗДІЛ 2

МЕТОДИ ОБРОБКИ ДАНИХ В ІНФОРМАЦІЙНО-УПРАВЛЯЮЧИХ СИСТЕМАХ

Для розв'язання функціональних задач в інформаційно-управляючих системах існує достатньо велика кількість методів та алгоритмів. Однак, незважаючи на те, що арсенал методів та алгоритмів досить великий, виникає чимало перепон для їх ефективного використання. Кожен метод орієнтований, як правило, на розв'язання конкретного класу задач і забезпечує ефективність розв'язання лише при виконанні певних теоретичних припущень.

Вибір методу оброблення даних для конкретного об'єкта являє певну складність, викликану необхідністю аналізу та оцінки його ефективності в залежності від класу об'єкта та його статистичних характеристик.

У зв'язку з нелінійним характером більшості методів оброблення даних в ІУС, наприклад алгоритмів ідентифікації та управління, їх аналітичне дослідження являє достатньо складну задачу. Багато методів потребують попередньої специфічної підготовки даних. Під час розв'язання задач моделювання та управління в ІУС отримання «чистих» вихідних даних є практично неможливим. Через наявність специфічних особливостей об'єктів управління, реальні дані містять шуми різного походження, мають нестационарні та нелінійні характеристики, що змінюються у часі. За таких умов лише ретельне попереднє оброблення даних може заощадити час та привести до успіху. Для цього необхідно

вирішити такі питання: збирання даних, їх фільтрація та аналіз, перетворення з метою зробити інформацію більш змістовною та зручною для методу обробки, що використовується. На цьому етапі важливо визначити сукупність інформативних змінних, що беруть участь у розв'язанні задач моделювання та прийнятті ефективних управляючих впливів. У зв'язку з цим необхідно глибоко розуміти сутність задачі та ретельно організовувати попередню обробку даних.

2.1. Дані. Типи даних

Дані в інформаційно-управляючих системах подаються у детермінованому та статистичному вигляді. Статистична інформація може бути представлена у вигляді абсолютних та відносних величин.

Абсолютні величини підрозділяються на індивідуальні та сумарні.

Перші з них характеризують розміри кількісних ознак. Цей вид показників слугує ґрунтом при статистичному зведенні для включення одиниць об'єкта у групи. На їх основі отримують абсолютні величини, з яких, у свою чергу, можна виділити показники чисельності сукупності та показники обсягу ознак сукупності.

Абсолютні величини — числа, що мають певну розмірність, одиниці вимірювань. На практиці натуральні одиниці виміру можуть бути складеними. Абсолютні величини вимірюються також у вартісних одиницях — цінах.

Вивчаючи економічні явища, статистика не може обмежуватись обчисленням тільки абсолютних величин. В аналізі статистичної інформації важливе місце посідають похідні узагальнюючі показники — середні та відносні величини [1].

Аналіз — це перш за все порівняння, співставлення статистичних даних. У результаті порівняння отримують якісну оцінку техніко-економічних явищ, яка подається через відносні величини — у статистиці вони є результатом ділення двох статистичних величин і характеризують кількісне співвідношення між ними.

Розраховуючи *відносні величини*, слід мати на увазі, що у чисельнику завжди знаходиться порівнюваний показник, який відображає те явище, яке вивчається, а у знаменнику — показник, з яким порівнюють, який приймають за основу або базу

порівняння. База порівняння виступає у якості такого собі вимірювача. Існують також *іменовані відносні величини*.

Якщо значення основи або бази порівняння приймається за одиницю (прирівнюється до одиниці), то відносна величина (результат порівняння) є коефіцієнтом та показує, у скільки разів величина, що вивчається, більша за базову. Розрахунок відносних величин у вигляді коефіцієнта застосовується у тому випадку, якщо величина, що порівнюється, істотно більша за ту, з якою вона порівнюється. Якщо значення основи або базу порівняння прийняти за 100 %, результат обчислення відносної величини подаватиметься також у відсотках.

Форма подання відносних величин залежить від кількісного співвідношення порівнюваних величин, а також від значеннєвого змісту отриманого результату порівняння. У тих випадках, коли порівнюваний показник більший за основу, відносна величина може бути виражена або коефіцієнтом, або у відсотках. Коли порівнюваний показник менший за основу, відносну величину краще виразити у відсотках. У кожному окремому випадку слід обирати ту форму подання відносних величин, яка є більш наочною та легше сприймається.

Розрахунок відносних величин може бути правильним лише за умови, що досліджувані показники, які порівнюються, є порівнянними [1, 2]. Причини, що зумовлюють непорівнянність показників, різноманітні, наприклад відмінності у методології збирання, оброблення статистичної інформації, у тривалості часових періодів, за які обчислені порівнювані показники, та ін. У таких випадках розрахунок відносних величин можна виконувати лише після приведення показників, що вивчаються, до порівнянного вигляду.

Показники структури можуть подаватися у частках (база порівняння приймається за 1).

Відносні величини інтенсивності показують, наскільки широко розповсюджене досліджуване явище у тому чи іншому середовищі. Вони характеризують співвідношення різнойменних, але пов'язаних між собою абсолютних величин.

На відміну від інших видів відносних величин, відносні величини інтенсивності завжди подаються іменованими величинами.

Розраховуються відносні величини інтенсивності шляхом ділення абсолютної величини досліджуваного явища на абсолютну

величину, що характеризує об'єкт середовища, у якому відбувається розвиток або розповсюдження явища. Відносна величина показує, скільки одиниць однієї сукупності припадає на одиницю іншої сукупності.

Однією з умов правильного використання статистичних показників є вивчення абсолютних та відносних величин у їх єдності. Якщо цієї умови не дотримано, можна прийти до хибного висновку. Тільки комплексне застосування абсолютних та відносних величин забезпечує всебічну характеристику явища, що вивчається.

Розглянемо методи отримання статистичних даних та форми статистичних спостережень в ІУС. Статистичні дані формуються в процесі статистичного спостереження, а потім піддаються систематизації, зведенню, обробці, аналізу та узагальненню. *Статистичне спостереження* являє собою факти, які характеризують явища суспільного життя та реєструються у відповідних облікових документах. Прикладом можуть слугувати встановлені форми звітності підприємств, записи відповідей громадян на запитання програми перепису населення, записи регістраторів для з'ясування задоволеності попиту населення товарами тощо.

Спостереження має проводитись відповідно до заздалегідь розробленої системи, плану, програми.

Статистичне спостереження здійснюється у двох формах: шляхом подання звітності (загальнодержавної або відомчої) та проведення спеціально організованих статистичних спостережень. Види вартісного спостереження розрізняються за часом реєстрації даних та за ступенем охоплення одиниць сукупності, що досліджується. За характером реєстрації даних у часі розрізняють спостереження безперервне, поточне та перервне (періодичне).

У процесі підготовки до проведення статистичного спостереження виникають програмно-методологічні та організаційні питання.

До першої групи належать питання, пов'язані з визначенням мети, об'єкта та одиниці спостереження, розробкою програми спостереження, проектування формулярів та тексту інструкцій, встановлення джерел та способів збирання даних.

Друга група містить відомості про орган спостереження, строки та місце проведення спостереження, складання попередніх списків одиниць, розташування та підготовку кадрів та ін.

Програмою статистичного спостереження називають перелік показників, що підлягають вивченню. З метою покращеного проведення статистичного спостереження розробляється організаційний план з урахуванням помилок спостереження, які виявляються у вигляді:

- ☒ помилок реєстрації;
- ☒ випадкових помилок;
- ☒ систематичних помилок;
- ☒ помилок репрезентативності.

Дані в ІУС можуть бути представлені у вигляді статистичних зведень, групувань, таблиць, виборок.

Статистичне зведення — це впорядкована система статистичних показників або узагальнених показників, яка дозволяє здійснити аналіз та прогнозування явищ, що вивчаються.

Групування — це процес об'єднання одиниць, що вивчаються, в окремі сукупності за однорідними ознаками.

Технологічне групування — виділення (відокремлення) з множини ознак однорідних.

2.2. Вибіркові дані

Вибірковий метод є найбільш прийнятним для розв'язання функціональних задач в ІУС, наприклад, для контролю якості, відвантаження готової продукції, планування тощо.

Статистична сукупність, що підлягає вивченню, з якої проводиться відбір частини одиниць, називається *генеральною сукупністю*. Відібрана з генеральної сукупності деяка частина одиниць, що піддається обстеженню, називається *вибірковою сукупністю* або *вибіркою*.

Значення вибіркового методу полягає в тому, що для мінімальної кількості одиниць, що обстежуються, проведення досліджування здійснюється у більш короткі строки та з мінімальними витратами праці та коштів. Це підвищує оперативність статистичної інформації, зменшує помилки реєстрації.

За дотримання правил наукової організації обстеження вибірковий метод дає достатньо точні результати, тому його доцільно застосовувати для перевірки даних суцільного обліку. Мінімальна кількість обстежуваних одиниць дозволяє провести дослідження більш ретельно та кваліфіковано. Так, під час перепису населення

практикуються вибіркові контрольні обходи для перевірки правильності записів суцільного спостереження.

В основу добору одиниць для обстеження покладений принцип рівних можливостей потрапляння у вибірку кожної одиниці генеральної сукупності. Саме в результаті дотримання цих принципів виключається утворення вибіркової сукупності лише за рахунок найкращих або найгірших зразків. Це запобігає появі систематичних (тенденційних) помилок та робить можливим проводити кількісну оцінку помилки представництва (репрезентативності).

Необхідність застосування вибіркового методу зумовлена тим, що у багатьох випадках статистичні дані являють собою лише деякі вибірки з більш широкої природно існуючої генеральної сукупності. При цьому виникають проблеми, пов'язані з надійністю цих вибірок для оцінок за ними значень невідомих параметрів розподілу значень деяких величин у відповідній генеральній сукупності, а у загальному випадку й для статистичних висновків різного типу, що отримуються на їх основі.

Важливе методологічне питання — «випадковий» характер вибіркового методу. В процесі імітаційного (статистичного) моделювання штучно генерується деяка сукупність можливих значень будь-яких величин із заданим законом розподілу ймовірностей, на основі якої оцінюються імовірнісні характеристики інших величин, що залежать від зазначених величин. Мета вибіркового методу полягає в оцінюванні параметрів цього розподілу за значно меншою імовірнісною сукупністю з цієї генеральної.

Для наближеного визначення невідомих параметрів розподілу випадкових величин за відомими емпіричними вибірковими даними К. Пірсоном було запропоновано *метод статистичного оцінювання*. Статистична оцінка являє собою функцію від наявних спостережень (статистику), яка дає найбільш точне, у певному розумінні, наближене значення для істинного значення відповідного параметра розподілу.

Фішер ввів поняття оптимальних оцінок й ефективності оцінювання, а також підкреслював важливість так званого методу максимальної правдоподібності як дуже важливого способу розробки найкращих оцінок. Тип статистичного виводу (висновку), що існував раніше, полягав у наступному твердженні: якби деяка нульова гіпотеза про властивості розподілу ймовірностей будь-якої змінної була правильною, то малоімовірно, що відпо-

відна цим властивостям (параметрам) рідкісна подія, що полягає у здійсненні певних значень цієї змінної, відбулася б у випадково отриманій вибірці.

Ступінь рідкості будь-якої події характеризується так званим *рівнем значущості*, який приймається зазвичай таким, що дорівнює п'яти відсоткам. При цьому кажуть, що ці рідкі події значуще розбігаються з нульовою гіпотезою на рівні п'яти відсотків, а їхня множина називається критичною областю п'яти-відсоткового розміру. Рівень значущості задає ступінь ризику відхилити вірну нульову гіпотезу (так звана помилка I роду). Помилки II роду полягають у прийнятті хибної нульової гіпотези, тобто у відхиленні вірної альтернативної гіпотези. Ймовірність відхилення гіпотези, тобто прийняття вірної альтернативної гіпотези, називається *потужністю критерію згоди даних з нульовою гіпотезою*.

Пrawdopodobність розподілу ймовірностей для даної сукупності спостережень називається ймовірність її отримання при даному розподілі. При цьому правдоподібність являє собою деяку статистику, тобто функцію від спостереження.

Способи визначення помилки вибірки при різних прийомах формування вибірових сукупностей та розповсюдження характеристик вибірки на генеральну сукупність становлять основний зміст статистичної методології вибірового методу.

Отримання даних вибіровим методом складається з низки послідовних етапів:

- ☑ обґрунтування (у відповідності до задач дослідження) доцільності застосування вибірового методу, параметра розподілу — значення, найбільш наближене до істини;

- ☑ складання програми проведення статистичного дослідження вибіровим методом;

- ☑ вирішення організаційних питань збирання й обробки вихідної інформації;

- ☑ встановлення частки вибірки, тобто частини одиниць генеральної сукупності, що підлягають обстеженню;

- ☑ обґрунтування способів формування вибірової сукупності;

- ☑ здійснення відбору одиниць з генеральної сукупності для їх обстеження;

- ☑ фіксація у відібраних одиницях (спробах) ознак, що вивчаються;

- ☑ статистична обробка отриманої у вибірці інформації з визначенням узагальнюючих характеристик ознак, що вивчаються;
- ☑ визначення кількісної оцінки помилки вибірки;
- ☑ розповсюдження узагальнюючих вибіркових характеристик на генеральну сукупність.

Генеральною сукупністю називають множину гіпотетично можливих спостережень, які б могли бути проведені у цьому комплексі умов. Оскільки у визначенні мова йде про гіпотетично можливі спостереження (або об'єкти), то генеральна сукупність є поняття абстрактне і її не слід змішувати (плутати) з реальними сукупностями, що підлягають статистичному дослідженню. Так, дослідивши навіть усі підприємства галузі, ми можемо розглядати їх як представників гіпотетично можливої більш широкої сукупності підприємств, які б могли функціонувати у рамках комплексу умов.

Генеральна сукупність може бути як кінцевою, так і нескінченною. Кінцева сукупність має місце, наприклад, при обстеженні сімейних бюджетів, коли вибірка формується із сукупності сімей, які фактично є у країні. Потім здійснюється спостереження за прибутками та видатками відібраних сімей.

Нескінченна генеральна сукупність спостерігається, наприклад, у наукових дослідженнях, коли нас цікавить усереднений результат великої кількості експериментів.

У найпростішому випадку генеральна сукупність є одновимірною випадковою величиною x із функцією розподілу $F(x) = P(x < x)$, яка визначає ймовірність того, що x набуде значення, менше за фіксоване дійсне число $x(x \in x)$.

У генеральній сукупності частка одиниць, які мають досліджувану ознаку, називається *генеральною часткою* p , а середня величина досліджуваної ознаки, що варіює, — *генеральною середньою* $\bar{x}$.

У вибірковій сукупності частку досліджуваної ознаки називають *вибірковою часткою*, або *частістю* w , а середню величину у вибірці — *вибірковою середньою* $\tilde{x}$.

Основна задача вибіркового обстеження полягає в тому, щоб на основі характеристик вибіркової сукупності (частоти w або середньої $\tilde{x}$) отримати вірогідні висновки про показники частки p або середньої $\bar{x}$ у генеральній сукупності.

Математична статистика виводить різні середні з формул степеневої середньої:

$$\bar{x} = \sqrt[z]{\frac{\sum x^z}{n}},$$

при $z = 1$ — середня арифметична, при $z = 0$ — середня геометрична, при $z = -1$ — середня гармонічна, при $z = -2$ — середня квадратична.

Вид середньої обирається, виходячи з аналізу досліджуваної сукупності.

Узагальнюючі показники у вибірці (w та $\bar{x}$) можуть тією чи іншою мірою відрізнятися від значень цих характеристик у генеральній сукупності (P та $\bar{x}$).

Можливі розбіжності між характеристиками вибіркової та генеральної сукупності вимірюються середньою помилкою вибірки μ .

У математичній статистиці доводиться, що значення середньої помилки вибірки визначаються за формулою

$$\mu = \sqrt{\frac{\sigma_0^2}{n}}. \quad (2.1)$$

Використання формули (2.1) має передумовою, що відома генеральна дисперсія σ_0^2 . Але при проведенні вибірових обстежень ці показники, як правило, невідомі. Застосування вибіркового методу як раз передбачає визначення характеристик генеральної сукупності.

На практиці для визначення середньої помилки зазвичай використовуються дисперсії вибіркової сукупності σ^2 . Ця заміна оснований на тому, що при дотриманні принципу випадкового відбору дисперсія достатньо великої вибірки прагне відобразити дисперсію у генеральній сукупності.

У математичній статистиці доводиться таке співвідношення між дисперсіями у генеральній та вибірковій сукупностях:

$$\sigma_0^2 = \sigma^2 \left(\frac{n}{n-1} \right). \quad (2.2)$$

З формули (2.2) бачимо, що дисперсія у вибірковій сукупності менша за дисперсію у генеральній сукупності на величину $\frac{n}{n-1}$.

Якщо n достатньо велике, то відношення $\frac{n}{n-1}$ наближене до одиниці. Наприклад, за $n = 100$ значення $\frac{n}{n-1} = 1,01$, а за $n = 500$ значення $\frac{n}{n-1} = \frac{500}{499} = 1,002$ і т. д.

У разі заміни генеральної дисперсії σ_0^2 вибірковою дисперсією σ^2 формула розрахунку середньої помилки записується так:

$$\mu = \sqrt{\frac{\sigma^2}{n}}.$$

Для показника середньої величини дисперсія кількісної ознаки у вибірці визначається за формулами:

$$\sigma_x^2 = \frac{\sum (x_i - \bar{x})^2}{n}, \quad \sigma_x^2 = \frac{\sum (x_i - \bar{x})^2 f_i}{\sum f_i},$$

де f_i — кількість ознак.

Оскільки при безповторному відборі чисельність генеральної сукупності N у процесі вибірки скорочується, то у формулу для розрахунку середньої помилки вибірки включають додатковий множник $\left(1 - \frac{n}{N}\right)$. Формула середньої помилки вибірки набуде такого вигляду:

$$\mu \approx \sqrt{\frac{\sigma^2}{n} \left(1 - \frac{n}{N}\right)}.$$

У математичній статистиці доводиться, що межі значень характеристик генеральної сукупності (p та $\bar{x}$) відрізняються від характеристик вибіркової сукупності (w та $\hat{x}$) на величину $\pm \mu$ лише з імовірністю 0,683. Це означає, що у 683 із 1000 випадків генеральна частка p та генеральна середня $\bar{x}$ будуть знаходитися у встановлених межах $\bar{x} = \hat{x} \pm \mu_x$. У решті ж випадків ($1000 - 683 = 317$) вони можуть вийти за ці межі.

Імовірність судження можна підвищити, якщо розширити межі відхилень, прийнявши за міру середню помилку вибірки, збільшену в t разів. Так, при подвоєному значенні μ (тобто за $t = 2$) імовірність судження досягає 0,954. Це означає, що тільки у 46 випадках із 1000 ($1000 - 954 = 46$) характеристики можуть

вийти за межі двох μ . При цьому розширюються й межі характеристик генеральної сукупності.

Якщо взяти, наприклад, потроєне μ , то імовірність судження підвищується до 0,997. При цьому тільки у трьох випадках зі 1000 характеристики генеральної та вибіркової сукупностей можуть відрізнятись більш ніж на 3μ . Таким чином, показники p та $\bar{x}$ генеральної сукупності за показниками частоти вибірки w та $\tilde{x}$ визначаються:

а) при вивченні частки альтернативної ознаки

$$p = w \pm t \cdot \mu_w; \quad (2.3)$$

б) при вивченні середньої величини кількісної ознаки

$$\bar{x} = \tilde{x} \pm t \cdot \mu_{\bar{x}}. \quad (2.4)$$

Множник t у формулах (2.3) та (2.4) (у статистиці він називається *коефіцієнтом довіри*) визначається у залежності від того, з якою довірчою імовірністю треба гарантувати результати вибіркового обстеження.

Вираження множника t для різних ступенів ймовірності представлено у вигляді функції

$$F(t) = \frac{1}{\sqrt{2\pi}} \int_{-t}^{+t} e^{-\frac{t^2}{2}} dt.$$

На практиці користуються готовими таблицями цієї функції, які обчислені для різних значень t щодо випадку нормально розподіленої сукупності. У табл. 2.1 наведені деякі їхні значення.

Таблиця 2.1

Кратність помилки t

Кратність помилки t	Імовірність $F(t)$	Кратність помилки t	Імовірність $F(t)$
	0,0000	2,0	0,9545
0,1	0,0797	2,5	0,9876
0,5	0,3829	2,6	0,9907
1,0	0,6827	3,0	0,9973
1,5	0,8664	4,0	0,999937

З табл. 2.1 бачимо, що зі збільшенням t функція $F(t)$ наближається до одиниці.

Практично при розв'язанні функціональних задач в ІУС зазвичай обмежуються значеннями t , що не перевищують двох-трьох одиниць. При цьому вибір тієї чи іншої довірчої ймовірності залежить від того, з яким ступенем вірогідності треба гарантувати результати вибіркового дослідження.

Отже, сенс середньої помилки вибірки полягає у такому.

Обчислені характеристики вибіркової частки w та вибіркової середньої $\bar{x}$ за своєю природою є випадковими величинами. Вони можуть набувати різних значень у залежності від того, які конкретні одиниці генеральної сукупності потраплять до вибірки. Це означає, що у кожному варіанті відбору буде різна помилка вибірки. При цьому кожен із можливих результатів вибірки, а отже, й кожна з можливих помилок вибірки мають певну ймовірність виникнення. Тому середня помилка вибірки, по суті, являє собою середню квадратичну величину з окремих помилок, зважену за ймовірністю їх виникнення.

Для практики вибірових обстежень важливо, що середня помилка застосовується для встановлення межі відхилень характеристик вибірки з відповідних показників генеральної сукупності не безвідносно. Лише з певним ступенем ймовірності можна стверджувати, що ці відхилення не перевищать величини $t \cdot \mu$, яка у статистиці називається *граничною помилкою вибірки*.

Гранична помилка вибірки Δ пов'язана з середньою помилкою вибірки μ відношенням:

$$\Delta = t \cdot \mu. \quad (2.5)$$

При цьому t як коефіцієнт кратності середньої помилки вибірки залежить від ймовірності, з якою гарантується величина граничної помилки вибірки.

Якщо у формулу (2.5) підставити конкретний зміст μ , то розрахунок граничної помилки вибірки при неповторному відборі можна здійснювати за такими алгоритмами:

а) частка альтернативної ознаки

$$\Delta_w = t \cdot \sqrt{\frac{w(1-w)}{n} \left(1 - \frac{n}{N}\right)};$$

б) середня величина кількісної ознаки

$$\Delta_x = t \cdot \sqrt{\frac{\sigma_x^2}{n} \left(1 - \frac{n}{N}\right)}.$$

При цьому слід мати на увазі, що при порівняно невеликому відсотку одиниць, що взяті до вибірки (до 5 %), множник $\left(1 - \frac{n}{N}\right)$ близький до одиниці. Тому на практиці при розрахунку величини граничної помилки вибірки (при неповторному відборі) множник $\left(1 - \frac{n}{N}\right)$ можна опустити і розрахунок проводиться за формулами повторного відбору, тобто

$$\Delta_x = t \cdot \sqrt{\frac{\sigma_x^2}{n}}.$$

У багатьох випадках в ІУС обробка даних проводиться на основі малої вибірки. Малою вибіркою називають несущільне статистичне обстеження, за якого вибіркова сукупність утворюється з порівняно невеликої кількості одиниць генеральної сукупності. Обсяг малої вибірки зазвичай складає від 4–5 одиниць до 30 одиниць.

Середня помилка малої вибірки $\mu_{м.в}$ обчислюється за формулою:

$$\mu_{м.в} \cong \left(\frac{\sigma_{м.в}^2}{n} \right)^{\frac{1}{2}},$$

де $\sigma_{м.в}^2$ — дисперсія малої вибірки.

З урахуванням генеральної сукупності маємо: $\sigma_0^2 = \sigma^2 \frac{n}{n-1}$.

Оскільки для малої вибірки $\frac{n}{n-1}$ має суттєве значення, то обчислення дисперсії малої вибірки проводиться з урахуванням так званого числа степенів вільності — числа незалежних вимірювань, які можуть використовуватися для оцінювання параметрів. Сенс такого визначення полягає у наступному. При обчисленні вибіркового середнього, дисперсії та інших характеристик до загальної кількості даних додаються ці точки. Так, при обчисленні лише середнього додається одна точка, при обчисленні дисперсії необхідно мати середнє, тому додаються дві точки. Якщо до вибірки додаються, наприклад, дві точки, отримані з тієї ж вибірки, сталість ряду зменшується саме на дві точки. Отже, число степенів вільності є первісна кількість членів вибірки за мінусом числа сталих, що виведені з цього ряду для визначення точки,

відносно якої визначаються відхилення. Для малих виборок ця поправка має суттєве значення. Так, для дисперсії число степенів вільності дорівнює кількості членів вибірки за мінусом одиниці, бо дисперсія визначається за відхиленням від середнього.

Гранична помилка малої вибірки $\Delta_{м.в}$ визначається за формулою $\Delta_{м.в} = t\mu_{м.в}$. При цьому значення коефіцієнта довіри t залежить не лише від заданої довірчої ймовірності, але й від кількості одиниць вибірки n . Для окремих значень t та n довірча ймовірність малої вибірки визначається за спеціальними таблицями Стюдента, в яких надані розподіли стандартизованих відхилень. Таблиці Стюдента наводяться у підручниках з математичної статистики.

Із збільшенням обсягу вибірки розподіл Стюдента наближається до нормального та при 20 вже мало відрізняється від нормального розподілу.

При проведенні малих вибірових обстежень важливо мати на увазі, що чим менший обсяг вибірки, тим більша різниця між розподілом Стюдента та нормальним розподілом. При мінімальному обсязі вибірки ($n = 4$) ця різниця вельми суттєва, що вказує на зменшення точності результатів малої вибірки. Для малої вибірки у якості довірчої ймовірності практично приймається значення 0,95 або 0,99.

Під час організації вибіркового обстеження слід мати на увазі, що розмір помилки вибірки насамперед залежить від чисельності вибіркової сукупності n . Середня помилка вибірки обернено пропорційна $\sqrt{n}$, тобто при збільшенні, наприклад, чисельності вибірки у чотири рази, її помилки зменшується вдвічі.

Збільшуючи чисельність вибірки, можна довести її помилку до як завгодно малих розмірів. Можна уявити, що при доведенні n до розмірів N помилка вибірки μ дорівнює нулю. Підвищення відсотка вибірки, як правило, призводить до збільшення обсягу дослідницької праці, викликає додаткові витрати матеріальних коштів. Але, з іншого боку, якщо обсяг вибірки не є достатнім, то результати можуть містити великі погрішності. Все це необхідно враховувати при організації вибіркового обстеження.

Необхідна чисельність вибірки визначається за формулою граничної помилки вибірки:

$$\Delta_x = t \sqrt{\frac{\sigma_x^2}{n}}.$$

Так, стосовно цієї формули обсяг необхідної вибірки можна отримати шляхом перетворень, розв'язуючи цю рівність відносно n . Звідси необхідна чисельність вибірки при розрахунку середньої величини кількісної ознаки (назвемо її n_x) буде виражена так:

$$\Delta_x^2 = t^2 \frac{\sigma_x^2}{n_x}, \quad n_x = \frac{t^2 \sigma_x^2}{\Delta_x^2}.$$

Так само виводять формулу для розрахунку чисельності вибірки при вибірковому обстеженні частки альтернативної ознаки (n_w):

$$\Delta_w = t^2 \frac{w(1-w)}{n_w}, \quad \text{звідси} \quad n_w = \frac{t^2 w(1-w)}{\Delta_w^2}.$$

2.3. Основні статистичні характеристики, що використовуються при розв'язанні задач в інформаційно-управляючих системах

Статистичні характеристики відображають кількісний та якісний аспекти досліджуваних предметних галузей та відіграють важливу управлінську функцію. За допомогою моментів 1-го, 2-го, 3-го і 4-го порядків, законів розподілу, кореляційних функцій, спектральних щільностей розв'язуються функціональні задачі як у часовій, так і в частотній областях. Зазначимо, що дані в ІУС можуть подаватися у вигляді безперервних k -вимірних випадкових величин, усі компоненти яких — безперервні (одновимірні) випадкові величини; дискретних k -вимірних випадкових величин, компоненти яких усі дискретні, та змішаних k -вимірних випадкових величин, серед компонент яких є як дискретні, так і безперервні випадкові величини. У цьому розділі основна увага приділяється статистичним характеристикам, що необхідні для глибокого розуміння кореляційного, регресійного, дисперсійного, кластерного і дискримінантного аналізів. Деякі теоретичні викладки та приклади запозичені з робіт [1–6].

Функцією розподілу випадкового вектора x називається детермінована невід'ємна величина, що визначається за формулою: $F(x) = P(\mathbf{x}_1 < x_1, \mathbf{x}_2 < x_2, \dots, \mathbf{x}_k < x_k) = P(\mathbf{x} < x)$, де x — k -вимірний вектор фіксованих дійсних чисел $x = (x_1, x_2, \dots, x_k)^T$.

Детермінована невід'ємна величина характеризується такими властивостями:

☑ $F(x) = 0$, якщо серед x_j наявна хоча б одна компонента, що дорівнює $-\infty$;

☑ $F(x) = 1$, якщо усі компоненти вектора x дорівнюють ∞ ;

☑ $F(x)$ задовольняє формулі обчислення ймовірності потрапляння випадкової точки у k -вимірний паралелепіпед з площинами, паралельними до координатних.

Функція розподілу $F(x)$ для безперервної k -вимірної випадкової величини є безперервною за визначенням. Безперервна k -вимірна випадкова величина має щільність розподілу ймовірностей $p(x) = (x_1, x_2, \dots, x_k) \geq 0$, що задовольняє умові

$$F(x) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \dots \int_{-\infty}^{x_k} p(x_1, x_2, \dots, x_k) dx_1 dx_2 \dots dx_k.$$

Щільність $p(x)$ має такі властивості:

$$\int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} p(x_1, \dots, x_k) dx_1 \dots dx_k = 1,$$

де k — загальне число (кратність) інтегралів.

Підсистема L ($1 \leq L < k$) компонент і додаткова підсистема $(k-L)$ компонент вектора x називаються *незалежними* (стохастично, ймовірно, статистично), якщо справедлива рівність:

$$F(x_1, x_2, \dots, x_k) = F(x_{j_1}, x_{j_2}, \dots, x_{j_L}) \cdot F(x_{j_{L+1}}, \dots, x_{j_k}).$$

Зокрема, компоненти x називаються незалежними, якщо

$$F(x_1, x_2, \dots, x_k) = F(x_1)F(x_2) \dots F(x_k).$$

У випадку незалежності стають справедливими аналогічні формули для добутків щільностей або ймовірностей маргінальних розподілів та збіг умовних розподілів із відповідними маргінальними.

Закон розподілу повністю характеризує генеральну сукупність, тобто остання може бути задана функцією розподілу або щільністю. Однак така вичерпна характеристика генеральної сукупності доволі складна і не потрібна для розв'язання багатьох практичних задач, у яких достатньо знати лише числові характеристики законів розподілу. Наприклад, для грубої характеристики одновимірної випадкової величини можна обмежитися її середнім значенням і величиною розкиду можливих значень.

Математичним сподіванням, або генеральною середньою дискретною випадкової величини x називається сума добутків можливих значень цієї величини на відповідні імовірності

$$Ex = \sum_{i=1}^N x_i p_i,$$

де p_i — ймовірність появи i -го можливого значення x_i випадкової величини x ($i = 1, 2, \dots, N$).

Для безперервної генеральної сукупності x зі щільністю $p(x)$ математичне сподівання дорівнює

$$Ex = \int_{-\infty}^{+\infty} xp(x)dx.$$

Аналогічно можна визначити математичне сподівання функції $\varphi(x)$ від випадкової величини x :

$$\text{дискретної } E\varphi(x) = \sum_{i=1}^N \varphi(x_i) p_i \text{ та}$$

$$\text{безперервної } E\varphi(x) = \int_{-\infty}^{+\infty} \varphi(x_i) p(x) dx.$$

Початковим v_k та центральним μ_k моментами k -го порядку скалярної випадкової величини x називаються математичні сподівання відповідно величин x_k та $(x - Ex)^k$, тобто $v_k = Ex^k$ та $v_k = E(x - Ex)^k$.

Звідси випливає, що початковий момент першого порядку є математичне сподівання випадкової величини $v_1 = Ex$.

Дисперсія σ_x^2 випадкової величини x є центральним моментом другого порядку $\sigma_x^2 = E(x - Ex)^2$.

Середнім квадратичним відхиленням називається додатний квадратний корінь з її дисперсії $\sigma_x = \sqrt{\sigma_x^2}$.

Одним з найважливіших результатів застосування багатовимірного аналізу є статистичний висновок про параметри або характеристики генеральної сукупності. Дано визначення основних генеральних характеристик.

Функція $Y_1 = g_1(x) = g_1(x_1, x_2, \dots, x_k)$ є одновимірною випадковою величиною, якщо вона матеріальна, однозначно визначена в усіх точках $x \in R_k$ (R_k — k -вимірний евклідов простір), за виключенням точок множини нульової імовірності.

Функція розподілу для неї обчислюється за формулою:

$$F_{y_1}(y_1) = F(y_1) = P(Y_1 < y_1) = P(g_1(x) < y_1).$$

Неважко також отримати властивості математичного сподівання, виходячи з визначення математичного сподівання функції $E(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k) = \beta_0 + \beta_1 E x_1 + \dots + \beta_k E x_k$, де $x_1, \dots, x_k$ — випадкові величини; $\beta_0, \beta_1, \dots, \beta_k$ — детерміновані величини; $E(x_1, x_2, \dots, x_k) = E x_1 \cdot E x_2 \cdot \dots \cdot E x_k$ — для взаємно незалежних випадкових величин $x_1, x_2, \dots, x_k$.

Моментом l -го порядку $l = l_1 + l_2 + \dots + l_k$ випадкового вектора x відносно постійного вектора c називається

$$E_{l_1, l_2, \dots, l_k} = E[(x_1 - c_1)^{l_1} (x_2 - c_2)^{l_2} \dots (x_k - c_k)^{l_k}].$$

Якщо $c_j = 0$, то моменти називаються *початковими*, якщо $c_j = E x_j$ — *центральними*.

На практиці буває достатньо обмежуватись моментами до другого порядку включно. Наведемо назви, позначення та формули деяких моментів:

☑ $E x_j^l = E(x_j)^l$ — початковий момент l -го порядку, математичне сподівання l -го степеня j -ї компоненти вектора x ;

☑ $E(x_i x_j)$, $i, j = 1 \div k$ — початковий змішаний момент другого порядку;

☑ $\sigma_{jj} = E x_j^2 - (E x_j)^2 = E(x_j - E x_j)^2 = \sigma_j^2$, де σ_{jj} — центральний момент другого порядку, дисперсія j -ї компоненти вектора x , $j = 1 \div k$;

☑ $\sigma_{ij} = E[(x_i - E x_i)(x_j - E x_j)] = E(x_i x_j) - (E x_i)(E x_j)$, де σ_{ij} — центральний змішаний момент другого порядку, коефіцієнт коваріації i -ї та j -ї компонент вектора x , $i, j = 1 \div k$ за $i \neq j$.

Коваріаційна матриця визначається як математичне сподівання добутку центрованого випадкового вектора на цей же транспонований вектор $\Sigma = E[(x - E x)(x - E x)^T]$. Матриця

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1j} & \dots & \sigma_{1k} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2j} & \dots & \sigma_{2k} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \sigma_{i1} & \sigma_{i2} & \dots & \sigma_{ij} & \dots & \sigma_{ik} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \sigma_{k1} & \sigma_{k2} & \dots & \sigma_{kj} & \dots & \sigma_{kk} \end{pmatrix}$$

називається *коваріаційною матрицею випадкового вектора x* .

Неважко помітити, що вона симетрична. Покажемо, що Σ невід'ємно визначена. Нехай $g_1(x) = c_1x_1 + c_2x_2 + \dots + c_kx_k = c^T x$, де c — будь-який детермінований вектор. Тоді дисперсія $Dg_1(x)$ (за визначенням математичного сподівання та дисперсії) буде дорівнювати

$$Dg_1(x) = E[g_1(x) - Eg_1(x)]^2 = c^T E[(x - Ex)(x - Ex)^T] c = c^T \Sigma c \geq 0.$$

Коефіцієнт коваріації нормованих випадкових величин називається *коефіцієнтом кореляції*, або *коефіцієнтом парної кореляції*

$$r_{ij} = E\left(\frac{x_i - Mx_i}{\sigma_i} \cdot \frac{x_j - Mx_j}{\sigma_j}\right) = \frac{\sigma_{ij}}{\sigma_i \sigma_j},$$

де σ_i, σ_j — середні квадратичні відхилення випадкових величин x_i та x_j .

Квадрат коефіцієнта кореляції називають *коефіцієнтом детермінації*.

Оскільки для будь-яких випадкових величин x_i та x_j

$$E\left(\frac{x_i - Ex_i}{\sigma_i} \pm \frac{x_j - Ex_j}{\sigma_j}\right)^2 = 2 \pm 2r_{ij} \geq 0,$$

то $-1 \leq r_{ij} \leq 1$.

Причому при $r_{ij} = \pm 1$ величини x_i та x_j пов'язані лінійною функціональною залежністю. Якщо ж x_i та x_j незалежні, то $r_{ij} = 0$.

Матриця

$$R = \begin{pmatrix} 1 & r_{12} & \dots & r_{1k} \\ r_{21} & 1 & \dots & r_{2k} \\ \dots & \dots & \dots & \dots \\ r_{k1} & r_{k2} & \dots & 1 \end{pmatrix}$$

називається *кореляційною*, вона симетрична та невід'ємно визначена.

Має місце очевидна властивість для будь-яких компонент x_i та x_j

$$E(x_i x_j) = Ex_i Ex_j + \sigma_{ij} = Ex_i Ex_j + r_{ij} \sigma_i \sigma_j.$$

Умовним математичним сподіванням функції $y_1 = g(x) = g_1(x_1, x_2, \dots, x_p, x_{p+1}, \dots, x_k)$ за умови (фіксуванні) $x_{p+1} = x_{p+1}$,

$x_{p+2} = x_{p+2}, \dots, x_k = x_k$ називається функція компонент $x_{p+1}, x_{p+2}, \dots, x_k$, що обчислюється за формулами (для дискретної та безперервної величин відповідно):

$$E y_1 / x_{p+1}, \dots, x_k = \left\{ \sum_{i_1} \dots \sum_{i_p} g_1(x_1^{(i_1)}, \dots, x_p^{(i_p)}, x_{p+1}, \dots, x_k) p_{i_1 \dots i_p / i_{p+1} \dots i_k}, \right. \\ \left. = \int_{-\infty}^{+\infty} (p) \dots \int_{-\infty}^{+\infty} g_1(x) p(x_1, x_2, \dots, x_p / x_{p+1}, \dots, x_k) dx_1 \dots dx_p. \right.$$

2.4. Багатовимірна нормально розподілена генеральна сукупність

При розгляді різних моделей статистичного аналізу часто передбачається нормальний розподіл усіх або деяких ознак генеральної сукупності. Кажуть, що безперервна k -вимірна випадкова величина розподілена нормально, якщо щільність розподілу має вигляд

$$p(x) = [(2\pi)^k |\Sigma|]^{-1/2} \exp \left\{ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right\}, \quad (2.6)$$

де $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \dots \\ \mu_k \end{pmatrix}$ — k -вимірний вектор математичних сподівань; Σ^{-1} —

матриця, обернена до коваріаційної матриці Σ розмірності $(k \times k)$; $|\Sigma|$ — визначник цієї матриці.

Нагадаємо, що матриця Σ є симетричною та додатно визначеною.

Таким чином, багатовимірний нормальний закон розподілу визначається вектором математичних очікувань μ та коваріаційною матрицею Σ , тобто $\left[k + \frac{k(k+1)}{2} \right]$ параметрами генеральної сукупності.

Приклад 2.1 [З]

Показати, що при $k = 1$ має місце одновимірний нормальний закон розподілу.

Розв'язання

Дійсно, при $k = 1$ $\Sigma = \sigma_{11} = \sigma^2$. Тоді $|\Sigma| = \sigma^2$, а обернена матриця $\Sigma^{-1} = \frac{1}{\sigma^2}$. Підставляючи знайдені значення у вираз (2.6), маємо

$$P(x) = \frac{1}{\sqrt{2\pi} \cdot s} \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\}.$$

Отримаємо щільність розподілу одновимірного нормального закону, що залежить від двох параметрів: математичного сподівання μ та середнього квадратичного відхилення σ .

Приклад 2.2 [З]

Прийнявши $k = 2$ у виразі (2.6), вивести щільність двовимірного нормального закону розподілу.

Розв'язання

При $k = 2$ маємо

$$(x - \mu) = \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}; \quad \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix},$$

де $\sigma_{12} = \sigma_{21}$.

Тоді, прийнявши $\sigma_{11} = \sigma_1^2$ та $\sigma_{22} = \sigma_2^2$, отримаємо

$$\Sigma^{-1} = \frac{1}{\sigma_1^2 \sigma_2^2 - \sigma_{12}^2} \begin{pmatrix} \sigma_2^2 & -\sigma_{12} \\ -\sigma_{21} & \sigma_1^2 \end{pmatrix}.$$

Звідки, згідно з виразом (2.6),

$$\begin{aligned} Q(x_1, x_2) &= (x - \mu)^T \Sigma^{-1} (x - \mu) = (x_1 - \mu_1, x_2 - \mu_2) \frac{1}{\sigma_1^2 \sigma_2^2 - \sigma_{12}^2} \times \\ &\times \begin{pmatrix} \sigma_2^2 & -\sigma_{12} \\ -\sigma_{12} & \sigma_1^2 \end{pmatrix} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix} = \frac{1}{\sigma_1^2 \sigma_2^2 - \sigma_{12}^2} \times \\ &\times ((x_1 - \mu_1)\sigma_1^2 - (x_2 - \mu_2)\sigma_{12} - \sigma_{12}(x_1 - \mu_1) + \sigma_1^2(x_2 - \mu_2)) \times \end{aligned}$$

$$\times \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix} = \frac{1}{\sigma_1^2 \sigma_2^2 - \sigma_{12}^2} \times \\ \times [(x_1 - \mu_1)^2 \sigma_1^2 - 2(x_1 - \mu_1)(x_2 - \mu_2)\sigma_{12} + (x_2 - \mu_2)^2 \sigma_2^2].$$

Розділивши чисельник та знаменник на $\sigma_1^2 \sigma_2^2$ та враховуючи, що $\frac{\sigma_{12}}{\sigma_1 \sigma_2} = r_{12} = r$, отримаємо

$$Q(x_1 x_2) = \frac{1}{1-r^2} \left[\frac{(x_1 - \mu_1)^2}{\sigma_1^2} - 2r \frac{(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1 \sigma_2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2} \right]. \quad (2.7)$$

Ураховуючи також, що

$$\frac{1}{\sqrt{(2\pi)^2 |\Sigma|}} = \frac{1}{2\pi \sqrt{\sigma_1^2 \sigma_2^2 - \sigma_{12}^2}} = \frac{1}{2\pi \sigma_1 \sigma_2 \sqrt{1-r^2}},$$

остаточно отримаємо

$$P(x) = \frac{1}{2\pi \sigma_1 \sigma_2 \sqrt{1-r^2}} \exp \left\{ -\frac{1}{2} Q(x_1, x_2) \right\}, \quad (2.8)$$

З виразів (2.7) та (2.8) випливає, що щільність двовимірного нормального закону розподілу визначається п'ятьма параметрами: математичними сподіваннями μ_1 та μ_2 випадкових величин x_1 та x_2 , їх середніми квадратичними відхиленнями σ_1 , σ_2 та коефіцієнтом кореляції r .

Визначена вище нормально розподілена генеральна сукупність є невинродженою. У випадку винродженості сукупності компоненти випадкового вектора x є лінійно залежними. Отже, частина з них являє собою лінійні комбінації решти компонент, що утворюють лінійно незалежну підсистему.

Можна показати, що найрізноманітніші умовні розподіли нормально розподіленої генеральної сукупності при фіксуванні різних підмножин компонент є нормальними.

Також можна показати, що необхідною й достатньою умовою взаємної незалежності компонент (або підмножин компонент) є дорівнювання нулю відповідних коефіцієнтів кореляції.

Найрізноманітніші квадратичні регресії є лінійними. Отже, у якості показника зв'язку можна ефективно застосовувати найрізноманітніші коефіцієнти кореляції (детермінації) та регресії.

2.5. Довірчі області

При малих обсягах виборок або підвиборок точкові оцінки можуть достатньо далеко відхилятися від оцінюваних параметрів, тому вводиться поняття інтервальної оцінки параметра генеральної сукупності. Визначення та поняття інтервального оцінювання можна перенести на випадок векторного параметра $\theta = (\theta_1, \theta_2, \dots, \theta_m)^T$ із заміною довірчого інтервалу довірчою областю у відповідному m -вимірному просторі.

Довірчою областю вектора параметрів θ генеральної сукупності називається випадкова область, що повністю визначається результатами спостережень, яка з близькою до одиниці довірчою ймовірністю (надійністю) γ містить невідоме значення вектора θ .

Очевидно, що існує нескінченна множина довірчих областей, що відповідають одному й тому самому значенню γ . Звичайно намагаються визначити довірчі області, що мають мінімальні розміри за цієї надійності γ . Часто цій умові задовольняють області, симетричні відносно вектора оцінок $\hat{\theta}$ параметрів θ .

Основну складність у побудові довірчої області являє визначення законів розподілів придатних статистик. На цей час ці питання достатньо добре розроблені лише для нормального розподілу випадкових величин, що спостерігаються.

2.5.1. Довірча область для вектора математичного сподівання

Нехай за результатами n спостережень з генеральної сукупності X з k -вимірним нормальним розподілом $N_k(\mu, \Sigma)$ знайдені вектор середніх $\bar{x}$ та незміщена оцінка $\hat{S}$ коваріаційної матриці Σ . Потрібно знайти з надійністю γ довірчу область для k -вимірного вектора генеральних середніх μ .

Припустимо, що коваріаційна матриця є відомою. Знайдемо таку придатну статистику, щоб її розподіл був відомий та за якою однозначно можна визначити довірчу область.

Нагадаємо, що для одновимірної нормально розподіленої генеральної сукупності довірчий інтервал для μ визначається за формулами:

$$|t| < \Phi^{-1}(\gamma), \quad t = \frac{\bar{x} - \mu}{\sigma} \sqrt{n},$$

де статистика t підкоряється стандартному закону $N_1(0, 1)$. Останню рівність можна переписати у вигляді

$$t^2 = n(x_1 - \mu_1)(\sigma_{11})^{-1}(x_1 - \mu_1),$$

припускаючи $x_1 = x$, $\mu_1 = \mu$ та $\sigma_{11} = \sigma^2$ при $k = 1$.

Отримана рівність узагальнюється на випадок k -вимірної сукупності таким чином:

$$t^2 = n(\bar{x} - \mu)^T \Sigma^{-1}(\bar{x} - \mu).$$

Відомо, що статистика t^2 при $k = 1$ має χ^2 розподіл з числом степенів вільності, яке дорівнює $v = 1$. Статистика t^2 для будь-якого k також розподілена за χ^2 з числом степенів вільності $v = k$.

Таким чином, з надійністю γ можна стверджувати, що вектор μ вкривається довірчою областю, яка задається нерівністю

$$n(\bar{x} - \mu)^T \Sigma^{-1}(\bar{x} - \mu) \leq (\chi^2)^{-1}(1 - \gamma).$$

Тепер припустимо, що коваріаційна матриця Σ невідома.

Щоб при $k = 1$ побудувати довірчий інтервал для μ , використовують статистику

$$t = \frac{\bar{x} - \mu}{\hat{s}} \sqrt{n},$$

яка має t -розподіл з $v = n - 1$ степенями вільності. Рівність можна переписати у еквівалентній формі

$$t^2 = n(\bar{x} - \mu)(\hat{S}^2)^{-1}(\bar{x} - \mu), \quad (2.9)$$

За аналогією будується статистика T^2 Хотеллінга, яку використовують при побудові довірчої області для вектора середніх μ :

$$T^2 = n(\bar{x} - \mu)\hat{S}^{-1}(\bar{x} - \mu), \quad (2.10)$$

де $\hat{S}^{-1}$ — матриця, обернена до коваріаційної матриці $\hat{S}$.

Ураховуючи, що F і T^2 розподілу пов'язані співвідношенням

$$T_{\alpha, k, n-k}^2 = \frac{k(n-1)}{n-k} F_{\alpha, k, n-k}, \quad (2.11)$$

отримаємо рівняння поверхні, що обмежує довірчу область k генеральних середніх з надійністю γ

$$(\bar{x} - \mu)^T \hat{S}^{-1} (\bar{x} - \mu) = \frac{k(n-1)}{n-k} F_{\alpha, k, n-k},$$

де $F_{\alpha, k, n-k}$ — точка F -розподілу, яка відповідає рівню значущості α та числам степенів вільності k та $(n-k)$.

Рівняння (2.10) та (2.11) визначають k -вимірний еліпсоїд (еліпсоїд при $k=2$) з центром $\bar{x}$, бо його ліва частина являє собою додатно визначену квадратичну форму відносно μ .

Приклад 2.3 [3]

У табл. 2.2 наводяться дані про фонд заробітної платні працівників централізованих бухгалтерій x_1 та товарообіг аптек x_2 , що обслуговуються.

Таблиця 2.2

Фонд заробітної платні, тис. грн	x_1	20	36	28	51	70	45	30	56
Товарообіг, млн грн	x_2	3,5	5,4	2,7	9,8	10,1	6,2	2,4	9,5

Знайти оцінки математичних сподівань, дисперсій та коефіцієнта кореляції, довірчу область для вектора математичних сподівань з надійністю $\gamma=0,95$.

Розв'язання

Знайдемо середні арифметичні $\bar{x}_1 = 42$ та $\bar{x}_2 = 6,2$ й перейдемо до центрованих величин $u_{ij} = x_{ij} - \bar{x}_j$.

$$U^T = \begin{pmatrix} -22 & -6 & -14 & 9 & 28 & 3 & -12 & 14 \\ -2,7 & -0,8 & -3,5 & 3,6 & 3,9 & 0 & -3,8 & 3,3 \end{pmatrix}.$$

Звідки

$$U^T U = \begin{pmatrix} -22 & -6 & \dots & 14 \\ -2,7 & -0,8 & \dots & 3,3 \end{pmatrix} \begin{pmatrix} -22 & -2,7 \\ -6 & -0,8 \\ \dots & \dots \\ 14 & 3,3 \end{pmatrix} = \begin{pmatrix} 1930 & 346,6 \\ 346,6 & 73,68 \end{pmatrix},$$

$$\hat{S} = \frac{1}{n-1} U^T U = \frac{1}{7} \begin{pmatrix} 1930 & 346,6 \\ 346,6 & 73,68 \end{pmatrix} = \begin{pmatrix} 275,714 & 49,514 \\ 49,514 & 10,526 \end{pmatrix}.$$

Отже, оцінки дисперсій та середніх квадратичних відхилень такі: $\hat{s}_1^2 = 275,714$, $\hat{s}_1 = 16,605$, $\hat{s}_2^2 = 10,526$, $\hat{s}_2 = 3,244$. Знайдемо вибірковий коефіцієнт кореляції

$$r = \frac{49,514}{16,605 \cdot 3,244} = 0,919.$$

Обернена матриця до $\hat{S}$ буде мати вигляд

$$\hat{S}^{-1} = \begin{pmatrix} 0,023365 & -0,109909 \\ -0,109909 & 0,6120157 \end{pmatrix}.$$

Тоді, згідно з рівнянням (2.11),

$$(42 - \mu_1 \quad 6,2 - \mu_2) \begin{pmatrix} 0,023365 & -0,109909 \\ -0,109909 & 0,6120157 \end{pmatrix} \begin{pmatrix} 42 - \mu_1 \\ 6,2 - \mu_2 \end{pmatrix} = \frac{2 \cdot 7}{8 \cdot 6} \cdot 5,14,$$

де $F(0,05, 2, 6) = 5,14$ знаходимо за таблицею для $\alpha = 1 - \gamma = 0,05$ та чисел степенів вільності $\nu_1 = 2$ та $\nu_2 = 6$.

Після перетворень отримуємо рівняння еліпса

$$0,0234(42 - \mu_1)^2 + 0,0242(42 - \mu_1)(6,2 - \mu_2) + \\ + 0,612(6,2 - \mu_2)^2 = 1,499,$$

яке визначає межі довірчої області для вектора $(\mu_1, \mu_2)^T$.

Нехай вектор x підкоряється k -вимірному розподілу $N_k(\mu, \Sigma)$, де Σ — коваріаційна матриця, C за $(k \times l)$ — матриця $(l \leq k)$ рангу l . Тоді $C^T X$ підкоряється l -вимірному нормальному розподілу $N_l(C^T \mu, C^T \Sigma C)$, бо лінійні комбінації нормально розподілених величин також розподілені нормально.

У цьому випадку статистика T^2 має вигляд

$$T^2 = n(C^T \bar{x} - C^T \mu)(C^T \Sigma C)^{-1}(C^T \bar{x} - C^T \mu) \quad (2.12)$$

зі значенням, що відповідає ймовірності $\alpha = 1 - \gamma$,

$$T_{\alpha, l, n-l}^2 = \frac{l(n-1)}{n-l} F_{\alpha, l, n-l}.$$

Оскільки вектор $C^T \mu$ містить l генеральних середніх, то на відміну від рівняння (2.11), тепер число степенів вільності дорівнює l та $(n-l)$. В окремому випадку, коли C — одинична матриця

порядку k , рівняння (2.12) зводиться до рівняння (2.11) та число степенів вільності стає рівним, як і раніше, k та $(n - k)$.

Використання лінійних комбінацій компонент вектора μ дозволяє розширити застосування статистики T^2 Хотеллінга при інтервальному оцінюванні у задачах порівняння.

За допомогою лінійних комбінацій можна, наприклад, знайти сукупні довірчі інтервали або перевірити гіпотезу відносно перших $(l \leq k)$ середніх значень генеральної сукупності. Для цього достатньо прийняти

$$C^T = \left[\begin{array}{cccccc} 1 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & 0 & \dots & 0 \end{array} \right] \Bigg|_l.$$

Оскільки C^T має розмірність $(l \times k)$ та ранг l , то вектор $C^T \mu$ має розмірність $(l \times 1)$ та містить l ($l \leq k$) генеральних середніх.

Щоб побудувати довірчий інтервал для генерального середнього μ_j ($j = 1, 2, \dots, k$), достатньо прийняти, що C_j є j -й стовпчик одиничної матриці розмірності k . Тоді $C_j^T \mu = \mu_j$ та статистика Хотеллінга

$$T^2 = \frac{n[C_j^T(\bar{x} - \mu)]^2}{C_j^T S C_j}$$

мають розподіл, який залежить від чисел степенів вільності l та $(n - 1)$.

Таким чином, згідно з рівнянням (2.12), довірча область для $C^T \mu$ з надійністю γ буде обмежена поверхнею

$$(C^T \bar{x} - C^T \mu)^T (C^T S C)^{-1} (C^T \bar{x} - C^T \mu) = \frac{l(n-1)}{n(n-l)} F_{\alpha, l, n-l}.$$

Зокрема, з надійністю γ довірчі межі для лінійної комбінації $C_j^T \mu$ (де C_j є стовпчик матриці C) визначаються як

$$C^T \bar{x} \pm \sqrt{\frac{1}{n} C_j^T S C_j F_{\alpha, l, n-l}}.$$

Приклад 2.4 [З]

За даними задачі 2.3, використовуючи лінійні комбінації, знайти з надійністю $\gamma = 0,95$ інтервальні оцінки генеральних середніх μ_1 та μ_2 .

Розв'язання

Для нашого прикладу $\alpha = 1 - \gamma = 0,05$, $\nu_1 = 1$, $\nu_2 = n - 1 = 7$. Згідно з таблицею F -розподілу, $F(0,05, 1, 7) = 5,59$.

Для побудови інтервальної оцінки середньої μ_1 приймемо $C_1 = (1, 0)^T$, так що $C_1^T \bar{x} = 42$, $C_1^T S C_1 = \hat{S}_1^2 = 275,714$. Тоді межі довірного інтервалу для μ_1 мають вигляд

$$\mu_1 \in \left[42 \pm \sqrt{\frac{1}{8} \cdot 275,714 \cdot 5,59} \right] = [42 \pm 13,880].$$

Звідси $28,120 \leq \mu_1 \leq 55,880$.

Для побудови інтервальної оцінки генерального середнього μ_2 приймаємо $C^2 = (0, 1)^T$, звідки $C_2^T \bar{x} = 6,2$, $C_2^T S C_2 = 10,526$.

Тоді з надійністю $\gamma = 0,95$ межі довірного інтервалу такі:

$$6,2 \pm \sqrt{\frac{1}{8} \cdot 10,526 \cdot 5,59} = 6,2 \pm 2,712.$$

Звідси $3,488 \leq \mu_2 \leq 8,912$.

2.5.2. Визначення сумісної довірчої області для математичного сподівання та дисперсії

Один з простих підходів у побудові багатовимірної довірчої області полягає у визначенні таких інтервалів $I_1(\theta_1)$, $I_2(\theta_2)$, ..., $I_m(\theta_m)$ для координат $\theta_1, \theta_2, \dots, \theta_m$ вектора параметрів θ , для яких імовірність одночасного викривання усіх $\theta_1, \theta_2, \dots, \theta_m$ відповідними інтервалами була б не менша за задане значення γ . Таким чином, мова йде про знаходження прямокутної довірчої області для вектора θ відповідної надійності, не меншої за γ .

Введемо подію $A_i = \{\theta_i \in I_i(\theta_i)\}$ ($i = 1, 2, \dots, m$) та означимо прямокутну область у m -вимірному просторі, яка утворена інтервалами $I_1(\theta_1)$, $I_2(\theta_2)$, ..., $I_m(\theta_m)$, через $I(\theta)$. На основі властивостей ймовірностей отримаємо

$$P(\theta \in I(\theta)) = P\left(\bigcap_{i=1}^m A_i\right) = 1 - P\left(\bigcup_{i=1}^m \bar{A}_i\right) \geq 1 - \sum_{i=1}^m P(\bar{A}_i),$$

де $\bigcap_{i=1}^m A_i$ — перетинання подій A_i ; $\bigcup_{i=1}^m \bar{A}_i$ — об'єднання подій $\bar{A}_i$, протилежних подіям A_i .

З цього випливає, що для визначення довірчої області достатньо знайти довірчі інтервали $I_1(\theta_1), I_2(\theta_2), \dots, I_m(\theta_m)$ для координат вектора θ , що відповідають надійності $\left[1 - \frac{1}{m}(1 - \gamma)\right]$. Тоді ймовірність усіх параметрів $\theta_1, \theta_2, \dots, \theta_m$ буде не менша за γ .

Нехай для нормальної генеральної сукупності x з невідомими параметрами μ та σ взято випадкову вибірку $x_1, x_2, \dots, x_n$ з обсягом n . Потрібно з надійністю γ знайти (сумісну) довірчу область для μ та σ , тобто для двовимірного вектора параметрів $\theta = (\mu, \sigma)^T$.

За даними вибірки знайдемо середню арифметичну $\bar{x}$ та середньоквадратичне відхилення s . Тоді довірчі інтервали окремо для математичного сподівання μ та середньоквадратичного відхилення σ , що відповідають надійності γ , мають вигляд

$$\bar{x} - t_\alpha \frac{s}{\sqrt{n-1}} \leq \mu \leq \bar{x} + t_\alpha \frac{s}{\sqrt{n-1}} \quad (2.13)$$

$$\text{та } \sqrt{\frac{ns^2}{\chi^2(\alpha/2, n-1)}} \leq \sigma \leq \sqrt{\frac{ns^2}{\chi^2(1-\alpha/2, n-1)}}, \quad (2.14)$$

де $\alpha = 1 - \gamma$, t_α та χ^2 знаходяться за таблицями розподілу Стьюдента і χ^2 для числа степенів вільності $\nu = n - 1$ та ймовірностей α і $\left(\frac{\alpha}{2}, 1 - \frac{\alpha}{2}\right)$ відповідно.

Довірчу область для вектора θ можна визначити формулою

$$\Delta_\alpha(q) = \left[(\mu, \sigma) : \begin{cases} \bar{x} - t \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + t \frac{\sigma}{\sqrt{n}} \\ \sqrt{\frac{ns^2}{\chi_2^2}} \leq \sigma \leq \sqrt{\frac{ns^2}{\chi_1^2}} \end{cases} \right], \quad (2.15)$$

де

$$\chi_2^2 = \chi^2 \left(\frac{\alpha}{2}, n-1 \right), \quad \chi_1^2 = \chi^2 \left(1 - \frac{\alpha}{2}, n-1 \right).$$

Ця область являє собою трапецію.

Для побудови такої області із заданою довірчою ймовірністю γ можна керуватися наступним. Оскільки при нормальному розподілі генеральної сукупності x оцінки $\bar{x}$ та s незалежні, причому $\bar{x}$ має нормальний розподіл $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$, а $\frac{ns^2}{\sigma^2}$ — розподіл χ^2 з $v = n - 1$ степенями вільності, то

$$P(\theta \in \Delta_\gamma(\theta)) = P\left\{ \left[|\bar{x} - \mu| \leq t^\sigma \right] \cdot \left[\sqrt{\frac{ns^2}{\chi_2^2}} \leq \sigma \leq \sqrt{\frac{ns^2}{\chi_1^2}} \right] \right\} = P\left[|\bar{x} - \mu| \leq \frac{t^\sigma}{\sqrt{n}} \right] \times \\ \times P\left[\sqrt{\frac{ns^2}{\chi_2^2}} \leq \sigma \leq \sqrt{\frac{ns^2}{\chi_1^2}} \right] = \Phi(t) P(\chi_1^2 \leq \chi^2 \leq \chi_2^2).$$

Визначивши t та χ_1^2 , χ_2^2 з умов

$$\Phi(t) = C\sqrt{\gamma}, \quad P(\chi_1^2 \leq \chi^2 \leq \chi_2^2) = \frac{\sqrt{\gamma}}{C},$$

де C — довільне число, що належить інтервалу $\left(\sqrt{\gamma}, \frac{1}{\gamma}\right)$, знай-

демо довірчу область вектора параметрів $\theta = \left(\frac{\mu}{\sigma}\right)$, що відповідає надійності γ .

Приклад 2.5

За результатами контролю $n = 14$ виробів знайдено, що середня висота сальникової камери дорівнює $\bar{x} = 88$ мм, а $s = 0,96$ мм. Потрібно визначити довірчу область для вектора параметрів з надійністю $\gamma = 0,95$.

Розв'язання

Для порівняння визначимо спочатку з надійністю $\gamma = 0,95$ довірчі інтервали для математичного сподівання μ та середньоквадратичного відхилення σ .

За таблицею t -розподілу для числа степенів вільності $v = n - 1 = 13$ та рівня значущості $\alpha = 1 - \gamma = 0,05$ знаходимо $t_\gamma = 2,160$. Згідно з виразом (2.13), маємо

$$88 - 2,160 \frac{0,96}{\sqrt{13}} \leq \mu \leq 88 + 2,160 \frac{0,96}{\sqrt{13}},$$

звідки $87,425 \leq \mu \leq 88,575$.

За таблицею χ^2 -розподілу для числа степенів вільності $v = 13$ та ймовірності $\frac{1-\gamma}{2} = 0,025$ знайдемо верхню межу $\chi_2^2 = 24,736$ довірчого інтервалу для χ^2 . Звідси нижня межа для σ дорівнює (2.14):

$$\sqrt{\frac{ns^2}{\chi_2^2}} = \sqrt{\frac{14 \cdot 0,96^2}{24,736}} = 0,722.$$

Для числа степенів вільності $v = 13$ та ймовірності $\frac{1+\gamma}{2} = 0,975$ знайдемо $\chi_1^2 = 5,008$. Звідси верхня межа σ має вигляд (2.14):

$$\sqrt{\frac{ns^2}{\chi_1^2}} = \sqrt{\frac{14 \cdot 0,96^2}{5,008}} = 1,605.$$

Таким чином, $0,722 \leq \sigma \leq 1,605$.

Щоб визначити довірчу область для вектора (μ, σ) з надійністю $\gamma = 0,95$, знайдемо за таблицею інтегральної функції Лапласа $t = 2,24$ за умови $\Phi(t) = \sqrt{\gamma} = 0,9747$.

За таблицею χ^2 -розподілу знаходимо

$$\chi_1^2 = \chi^2 \left(\frac{1 + \sqrt{\gamma}}{2}, n - 1 \right) = \chi^2(0,9873, 13) = 4,285,$$

$$\chi_2^2 = \chi^2 \left(\frac{1 - \sqrt{\gamma}}{2}, n - 1 \right) = \chi^2(0,0127, 13) = 27,090.$$

Тепер за формулою (2.15) знайдемо довірчу область:

$$88 - 0,599\sigma \leq \mu \leq 88 + 0,599\sigma, \quad 0,690 \leq \sigma \leq 1,735.$$

Для порівняння знайдемо прямокутну довірчу область з коефіцієнтом довіри, не меншим за $\gamma = 0,95$. Для цього достатньо за

формулами (2.13) та (2.14) знайти довірчі інтервали для μ та σ , що відповідають надійності $\gamma_1 = 1 - \frac{1-\gamma}{2} = 0,975$.

У результаті отримаємо:

$$\begin{cases} 87,357 \leq \mu \leq 88,643, \\ 0,690 \leq \sigma \leq 1,738. \end{cases}$$

2.6. Стислий аналіз методів обробки даних

Сьогодні спостерігається зростання зацікавленості до застосування сучасних математичних методів та моделей у різних предметних галузях (економіці, бізнесі, сфері управління, соціології) з метою вдосконалення стратегії поведінки системи. Визначення поведінки техніко-економічних систем у часі стає все більш необхідним. Для прогнозування розвитку потрібна здатність випереджати наслідки дії та створювати плани, які, скоріше, є «такими, що випереджають», ніж «такими, що виправляють». Окрім того, потрібно вміти аналізувати ситуації, які неможливо точно передбачити.

Проблеми прийняття рішень в ускладнених умовах посідають сьогодні особливе місце в інформаційних технологіях. Математичні методи стали широко застосовуватися для опису та аналізу складних техніко-економічних, соціальних та інших систем. Теорія оптимізації створила сукупність методів, які за наявності відомих та фіксованих параметрів допомагають приймати оптимальні рішення, використовуючи ЕОМ. Певні успіхи є і в тому випадку, коли параметри — випадкові величини з відомими законами розподілу.

Найбільш доступним методом дослідження якісних та кількісних показників ефективності використання методів обробки даних в автоматизованих системах керування є *імітаційне моделювання* [7].

Використання імітаційного моделювання для вибору методу обробки даних, а також для проведення експерименту з розробленими моделями дозволяє набутися глибоке розуміння досліджуваного процесу (об'єкта) та системи управління ним. Застосування методів імітаційного моделювання значно скорочує строки виконання робіт з проектування математичного та про-

грамного забезпечення систем управління (МЗ та ПЗ) та забезпечує значну економію матеріальних коштів, що виділяються на їх розроблення.

Як приклад розглянемо імітаційний підхід до проектування МЗ та ПЗ, оснований на аналізі ефективності використання алгоритмів ідентифікації та управління. Базою для використання такого підходу слугують набори алгоритмічних та програмних модулів розв'язання окремих задач управління.

Подамо процес проектування модульної системи як розробку такої формалізованої процедури F вибору методів M , алгоритмів A , програм P розв'язання K задач в АСУТП, яка приводить до вибору системи $S = \bigcup_{k=1}^K S_k$, що складається з підсистем S_k , $k = \overline{1, K}$.

Вибір кожної S_k зводиться до вибору такої підмножини M^k, A^k, P^k розв'язків задачі, яка задовольнятиме заданому критерію якості та статистичним характеристикам об'єкта. Подамо синтез S_k таким чином

$$\{M, A, P, H, E\} \xrightarrow{F} S_k \{M^k \in M, A^k \in A, P^k \in P\},$$

$$E^* = \text{extr}_{M^i \in M} \text{extr}_{A^{ij} \in A} \text{extr}_{P^{ijl} \in P} E(M^i, A^{ij}, P^{ijl}, H),$$

де M — клас методів; A — банк алгоритмів; P — бібліотека програм розв'язання окремих задач; H — характеристики задач; E — показники якості підсистеми S .

У зв'язку з цим вибір ефективного програмного модуля може бути поданий як послідовний вибір методу, алгоритму та програми. Тоді для проектування системи S необхідно розв'язати такі задачі:

☑ сформувати перелік K задач, що розв'язуються у системі S , тобто поставити задачу проектування підсистем S_k ;

☑ сформувати клас методів M , банк алгоритмів A , бібліотеку програм P розв'язання кожної k -ї задачі;

☑ виявити характеристики та показники E якості окремих модулів, за якими можна проводити порівняння альтернативних модулів;

☑ виявити специфічні властивості об'єкта H , які впливають на якість розв'язання задач в умовах функціонування системи;

☑ встановити математичну залежність між характеристиками якості модулів E та властивостями об'єкта H ;

☑ розробити формалізовану процедуру F автоматизованого проектування прикладного програмного забезпечення, тобто розробити критерії та алгоритми вибору найкращого модуля.

Для розв'язання задачі виявлення характеристик H та показників E якості окремих модулів відомо достатньо багато методів та алгоритмів, серед яких методи максимальної правдоподібності, мінімізації байесівського ризику та групового врахування аргументів, звичайний та зважений МНК, гребеневі та робасні процедури, рекурентні, нейронні сітки, генетичні алгоритми, адаптивні та інші процедури [4, 5, 6, 7, 8, 9].

Якість методів ідентифікації може бути визначена за такими показниками:

☑ середньоквадратична помилка оцінки коефіцієнтів моделі $L_1 = E\{(\hat{B} - B)'(\hat{B} - B)\}$, де $E\{\cdot\}$ — символ математичного сподівання; B — вектор істинних коефіцієнтів моделі; $\hat{B}$ — вектор оцінки коефіцієнтів моделі;

☑ середньоквадратична помилка оцінки коефіцієнтів, яка залежить від плану $L_2 = E\{(\hat{B} - B)'(X'X)(\hat{B} - B)\}$;

☑ максимальна координатна помилка оцінки коефіцієнта моделі $L_3 = \max(\hat{b}_j - b_j)$, де $j = \overline{1, m^*}$, m^* — кількість оцінюваних коефіцієнтів;

☑ відносна помилка оцінки коефіцієнта моделі

$$L_4 = \frac{1}{m^* \sum_{j=1}^{m^*} \frac{\hat{b}_j - b_j}{b_j^2}};$$

☑ загальна дисперсія $L_5 = \text{Tr}\{E\{[\hat{B} - E\{\hat{B}]\}'[\hat{B} - E\{\hat{B}]\}\}$;

☑ максимальна відносна координатна оцінка коефіцієнта моделі $L_6 = \max|(\hat{b}_j - b_j)/b_j|$, $j = \overline{1, m^*}$;

☑ зміщення оцінки коефіцієнта моделі

$$L_7 = (E\{\hat{B}\} - B)'(E\{\hat{B}\} - B).$$

Такий набір показників може використовуватись для будь-яких методів ідентифікації, являючись їхньою вичерпною характеристикою.

Вибір методів ідентифікації визначається метою дослідження, класом об'єкта, постановкою задачі, обсягом та способом подання апріорної інформації. Склад банку алгоритмів визначається, перш за все, специфікою об'єкта, а саме часовими та

статистичними характеристиками процесів, що відбуваються в об'єкті. Клас алгоритмів обирається проектувальником за результатами статистичного аналізу експериментальних даних. Вибір програм для включення в бібліотеку залежить від наявності готових програмних модулів та можливості програмної реалізації інших алгоритмів.

Основними показниками якості програми є такі: загальний час, витрачений на пошук розв'язку; загальний обсяг потрібної оперативної пам'яті ЕОМ; досяжна точність розв'язку.

Реалізація процедур автоматизованого вибору модулів передбачає можливість їх порівняльного аналізу за деяким критерієм якості. Обов'язковою умовою вибору показника якості для порівняння модулів є можливість кількісної оцінки критерію.

Для класу об'єктів, що розглядається, множина H формується з таких параметрів [4]: $h_1 = n$ — обсяг вибірки; $h_2 = p$ — кількість незалежних змінних; $h_3 = \lambda_{\min}/\lambda_{\max}$ — показник зумовленості матриці плану; $h_4 = \sigma_{ey}^2$ — дисперсія помилки у залежній змінній; $h_5 = b'b$ — довжина вектора істинних коефіцієнтів у моделі; $h_6 = \sigma_{ey}^2/\sigma_y^2$ — співвідношення дисперсій шум/сигнал для залежної змінної; $h_7 = \eta$ — рівень засміченості розподілу, якщо останній є сумішшю $f(\epsilon) = (1 - \eta)f_1(\epsilon) + \eta f_2(\epsilon)$; $h_8 = \tau_3$ — час згасання автокореляційної функції $R(\tau_3)$; $h_9 = \sigma_{xe}^2/\sigma_x^2$ — співвідношення дисперсії шум/сигнал для незалежних змінних; $h_{10} = \max r_{xe}$ — максимальна кореляція помилок зі змінними x .

Характеристики h_j можуть змінюватись у широких межах, однак для класу об'єктів, що розглядається, інтервали їхнього варіювання виявляються більш вузькими. На підставі проведених авторами досліджень реальних процесів визначені інтервали зміни за кожною h_j як пара

$$(h_{j\max} - h_{j\min}): h_1 = 20 - 200, h_2 = 5 - 50,$$

$$h_3 = 10^{-1} - 10^{-3} \quad (01 > h_3 = 10 - 1000), h_4 = 0 - 10,$$

$$h_5 = 1 - 100, h_6 = 0 - 60, h_7 = 0 - 20, h_8 = 0 - 50,$$

$$h_9 = 0 - 60, h_{10} = 0 - 0,5.$$

Інтервали дослідження необхідно прийняти більш широкими, ніж вони зазвичай зустрічаються на практиці. Це спричинено прагненням вивчити поведінку алгоритму в екстремальних ситуаціях на межах області зміни параметрів об'єкта Ω_H . Крім зазначених h_j , $j=1,10$, параметрами моделі є математичні сподівання,

дисперсії та інші статистичні характеристики випадкових величин та їх розподілів.

Розглянемо детально метод дослідження алгоритмів ідентифікації з метою кількісної оцінки показників їхньої якості. Для прийняття рішення про вибір найбільш преференційної альтернативи інформацію можуть надати лише експериментальні дослідження модуля на тестових задачах з відомими характеристиками [7].

Розглянемо постановку та розв'язання задачі машинної імітації при дослідженні програмних реалізацій алгоритмів ідентифікації. Нехай $E_l = \{e_s\}_{s=1}^s$ — сукупність визначених раніше показників якості l -го модуля. На кожний показник e_s впливає один або декілька параметрів $h_j \in H$. Для отримання експериментальної характеристики якості l -го модуля необхідно побудувати регресійні залежності $e_s = \psi_s(H) \forall s$, де $\psi(\cdot)$ — деяка детермінована функція.

Для отримання залежностей e_s здійснюють серію машинних експериментів за $h_j \in H$, де незалежними змінними виявляються h_j , залежними — кожна з e_s . Ця задача розпадається на підзадачі:

- ☑ вибір плану експерименту;
- ☑ синтез імітаційних моделей з параметрами h_j .

Ураховуючи, що основною вимогою до моделей, які синтезуються, є мінімум залишкової дисперсії та дисперсій оцінок коефіцієнтів, з огляду на малий обсяг апріорної інформації для проведення машинних експериментів слід звернутися до методів оптимального планування експерименту. Синтез імітаційної моделі гіпотетичного об'єкта з параметрами H базується на методах статистичного імітаційного моделювання, які дозволяють генерувати вибірки псевдовипадкових чисел із заданими характеристиками. Для отримання вибірки з заданими характеристиками необхідно розв'язати такі задачі.

Моделювання одновимірної випадкової вибірки. З огляду на стохастичну природу досліджуваних процесів використовуються *генератори псевдовипадкових чисел (ГПВЧ)*. Детальний аналіз існуючих ГПВЧ наведений у [7].

Помилка у залежній та незалежній змінних може мати імпульсний та флуктуаційний характер. Перший вид помилки характеризується як викид, другий — як постійно діюча випадкова помилка (помилка вимірювань).

Імітація викидів проводиться за схемою $x_j = x_j + d_{\tilde{e}_k}$, де $x_j = \{x_{tj}\}_{t=1}^h$ — вектор j -ї змінної, d — величина викиду, $\tilde{e}_k$ — нульовий вектор з одиницею на місці $\tilde{k}$ -ї компоненти.

Імітація помилок вимірювань у залежних і незалежних змінних здійснюється у відповідності до виразів

$$y^{(0)} = y^{(0)} + e, \quad x_j^{(0)} = x_j^{(0)} + \xi,$$

де $e = \{e_t\}_{t=1}^h$, $\xi = \{\xi_t\}_{t=1}^h$ — вектори помилок. Вектори імітуються як рівномірно або нормально розподілені псевдовипадкові числа. Однак реальні розподіли нечасто вдається описати ідеальною залежністю.

Найбільший інтерес становлять η — засмічені розподіли вигляду $f(\xi) = (1 - \eta)f_1 + \eta f_2(\xi)$, де $f_1(\cdot)$ та $f_2(\cdot)$ — відомі щільності розподілів (базова та засмічувальна). Реальні розподіли близькі до нормальних, тому в якості базового $f_1(\cdot)$ приймається нормальний розподіл з нульовим середнім та відомою дисперсією. У якості $f_2(\cdot)$ приймається розподіл з більшою, ніж у $f_1(\cdot)$, дисперсією. Це дозволяє подовжити «хвости» розподілу. Якщо параметр розташування розподілу $f_2(\cdot)$ прийняти таким, що дорівнює нулю, то матиме місце випадок симетричного засмічення. Якщо центр розподілу зміщений відносно нуля, то імітується асиметричне засмічення.

Моделювання багатовимірної вибірки передбачає імітацію потрібної кореляційної структури вихідних даних або, інакше, моделювання m одновимірних виборок із заданими взаємно кореляційними властивостями, які визначають ступінь зумовленості вихідних даних. Отримання матриці X із заданою зумовленістю зводиться до моделювання X з заданою кореляційною матрицею $\tilde{R}$, а останнє — до моделювання $\tilde{R}$ з заданими власними числами $\Lambda = \{\lambda_j\}$, $j = \overline{1, m}$. У загальному випадку імітація X здійснюється за схемою $Z = \bigcup \Lambda^{1/2} \in N(O, \Lambda) \Rightarrow X = ZC' \in N(O, \tilde{R})$, де $\bigcup \sim N(O, I)$, $\tilde{R} = C \wedge C'$, $N(\cdot)$ — нормальний закон розподілу; C — матриця власних векторів.

Обчислювальна схема моделювання X із заданою R з використанням модифікованого алгоритму Грама — Шмідта має вигляд:

1 етап. Генерація $\bigcup \sim N(O, I)$.

2 етап. Моделювання $\Lambda = \text{diad}\{\lambda_j\}$, виходячи з потрібних співвідношень.

3 етап. Моделювання C — матриці власних векторів з використанням модифікованого алгоритму Грама — Шмідта.

4 етап. Обчислення $\tilde{R} = C \wedge C'$.

5 етап. Отримання X : $X = U$.

Наведена схема дослідження може бути достатньо легко розповсюджена на алгоритми розв'язання різних задач. Незважаючи на те, що реалізація автоматизованого проектування алгоритмічного забезпечення ідентифікації вимагає значних обчислювальних витрат, часто цей підхід є єдиним, що дозволяє обґрунтовано підійти до вибору алгоритму. На практиці зустрічаються випадки, коли для ідентифікації моделей об'єкта обирається перший-ліпший алгоритм. Дослідження авторів показують, що для одних й тих самих експериментальних даних різні алгоритми приводять до різних моделей, які відрізняються не лише порядком коефіцієнтів, але й їхніми знаками. Цей та інші аналогічні факти свідчать, що для перевірки коректності роботи використовуваних алгоритмів необхідне проведення їх серйозного попереднього дослідження на різних тестових задачах.

Коректність статистичних методів — це наявність областей даних, що відповідають обмеженням застосовності метода для отримання оптимальних оцінок.

Одна з фундаментальних і разом з тим менш за все розроблених проблем сучасної статистики — взаємозв'язок статистичних методів, вибір найбільш ефективних серед них при розв'язанні тих чи інших задач. При цьому ефективними виявляються ті, що забезпечують отримання заданих достовірних оцінок при мінімумі витрат сил, коштів та часу.

Виділяють п'ять стандартних типів закономірностей у вихідних даних:

- ☒ асоціація;
- ☒ послідовність;
- ☒ класифікація;
- ☒ кластеризація;
- ☒ прогнозування.

Асоціація має місце в тому випадку, якщо декілька подій пов'язані одна з одною.

Якщо існує ланцюжок пов'язаних у часі подій, то говорять про *послідовність*. За допомогою *класифікації* виявляються ознаки, що характеризують групу, до якої належить той чи інший об'єкт. Це робиться за допомогою аналізу вже класифікованих об'єктів та формулювання деякого набору правил.

Кластеризація відрізняється від класифікації тим, що власне групи заздалегідь не задані. За допомогою кластеризації засоби Data Mining самостійно виділяють різні однорідні групи даних.

Основою для систем *прогнозування* слугує історична інформація, що зберігається у БД у вигляді часових рядів. Якщо вдається знайти шаблони, які адекватно відображують динаміку поведінки цільових показників, існує імовірність, що з їхньою допомогою можна передбачити й поведінку системи у майбутньому.

Основні методи статистичної обробки даних подані для користувача у вигляді *пакетів прикладних програм* (ППП). Як приклад найбільш потужних та поширених статистичних пакетів можна назвати SAS (компанія SAS Institute), SPSS (SPSS), STATGRAPICS (Manugistics), STATISTICA, STADIA та ін. У роботах [1, 2] наведений цікавий аналіз методів статистичної обробки даних, базований на хронології їх розвитку.

Практично будь-яка обробка даних починається з їх графічного аналізу. Застосування графіків у статистиці нараховує більш ніж двохсотрічну історію. Засновником графічного методу в статистиці комерційної діяльності вважають англійського економіста У. Плейфейра (1731–1798) [2]. У його роботі «Комерційний та політичний атлас» (1786 р.) уперше були застосоване графічне зображення статистичних даних (лінійні, стовпчикові, секторні та інші діаграми).

Графіки являють собою високорозвинену сукупність методів для пояснення, інтерпретації та аналізу числових даних за допомогою точок, ліній, площин, інших геометричних фігур та символів. Графічні методи мають особливу цінність для подання кількісних даних у простій та зрозумілій формі, а також для полегшення порівняння величин, тенденцій та залежностей. Вони мають низку інших переваг, зокрема наочність та стислість подання даних. Забезпечуване графічними методами вичерпне подання даних дозволяє виявляти приховані фактори, сприяє більш адекватному розумінню досліджуваного явища.

На статистичному графіку розрізняють такі основні елементи: поле графіка, графічний образ, просторові та масштабні орієнтири, експлікація графіка.

Поле графіка називають місце, на якому він виконується. Це аркуші паперу, географічні мапи, план місцевості тощо. Поле графіка характеризується його форматом (розмірами та

пропорціями сторін). Розмір поля графіка залежить від його призначення. Сторони поля статистичного графіка, зазвичай, знаходяться у певній пропорції. Прийнято вважати, що найбільш оптимальним для зорового сприйняття є графік, виконаний на полі прямокутної форми зі співвідношенням сторін від 1 : 1,3 до 1 : 1,5 (правило «золотого перерізу»). Інколи використовується й поле графіка з рівними сторонами, тобто квадратне.

Графічний образ — це символічні знаки, за допомогою яких зображуються статистичні дані. Вони вельми різноманітні: лінії, точки, плоскі (прямокутники, квадрати, круги тощо) та об'ємні геометричні фігури, силуети або рисунки предметів.

Одні й ті самі статистичні дані можна зобразити за допомогою різних графічних образів. Тому при побудові графіка важливий правильний добір графічного образу. Він має найбільш зрозуміло відображати досліджувані показники та відповідати основному призначенню графіка.

Просторові орієнтири визначають розміщення графічних образів на полі графіка. Вони задаються координатною сіткою або контурними лініями та розділяють поле графіка на частини, що відповідають значенням досліджуваних показників.

У статистичних графіках частіше за все застосовують систему прямокутних (декартових) координат, але можуть бути і графіки, побудовані за принципом полярних координат (кругові графіки).

У так званих статистичних мапах засобами просторової орієнтації виступають *географічні орієнтири* (контури суші або лінії річок, морів та океанів тощо). Просторові орієнтири дозволяють визначати розташування графічних образів на полі графіка.

Масштабні орієнтири статистичного графіка надають графічним образам кількісну значущість, яка передається за допомогою системи масштабних шкал.

Масштаб графіка — це міра переведення числової величини у графічну (наприклад, 1 см на мапі відповідає 100 км на місцевості). При цьому, чим довший відрізок лінії, що прийнята за числову одиницю, тим крупніший масштаб.

Масштабною шкалою є лінія, окремі точки якої читаються (у відповідності до прийнятого масштабу) як певні числа. На масштабній шкалі розрізняють: лінію — носій інформації, яка є опорою шкали; позначені на ній рисочками точки (розташовані за певним порядком); цифрові позначення чисел, які

відповідають окремим точкам. Шкала графіка може бути прямо- лінійною або криволінійною (наприклад, окружність (коло), що містить 360°).

Розрізняють також рівномірні та нерівномірні шкали.

Шкала є *рівномірною*, якщо однаковим графічним відрізкам відповідають однакові числові величини. Рівномірні арифметич- ні шкали використовують для побудови більшості статистичних графіків.

Нерівномірним шкалам відповідають нерівні числові значен- ня. Так, у логарифмічній шкалі графічні відрізки пропорційні не числам, а їх логарифмам. Однаковим графічним відрізкам на такій шкалі відповідають однакові різності логарифмів чисел (рис. 2.1).

Шкала, за якою відраховуються рівні досліджуваних показ- ників, як правило, починається з 0. Останнє число, що наносить- ся на шкалу, дещо перевищує максимальний рівень, відлік якого проводиться за цією шкалою. При побудові графіка допуска- ється розрив масштабної шкали. Цей прийом використовується для зображення статистичних даних, що мають значення лише у певних межах.

Експлікація графіка — це пояснення його змісту; містить за- головок графіка, пояснення масштабних шкал та окремих елемен- тів графічного образу. Заголовок графіка у стислій та чіткій формі пояснює основний зміст зображуваних даних. Окрім заголовка, на графіку наводять текст для читання графіка. Цифрові позначення шкали доповнюють зазначенням одиниць вимірювання.

Залежно від форми, графіки та діаграми поділяють таким чином: арифметичні та напівлогарифмічні графіки у прямокут- них координатах; смужкові та стовпчикові діаграми; графіки розподілу частот та до них подібні; статистичні мапи; інші гра- фічні засоби (такі, що містять діаграми кругові та розсіювання); діаграми рисункові та в тривимірній проекції.

Діаграма являє собою креслення, на якому статистична інформація зображується за допомогою геометричних фігур або символічних знаків. У техніці та економіці частіше за все

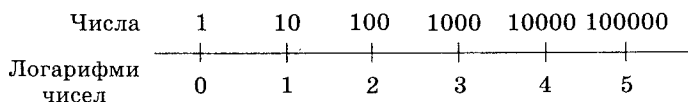


Рис. 2.1. Логарифмічна шкала

застосовують лінійні діаграми, побудовані в системі прямокутних координат.

Графіки в прямокутних координатах є найбільш поширеною формою графічного зображення статистичних даних. Причому на одному графіку можуть бути нанесені одночасно декілька серій даних, поєднаних одна з одною неперервними або перервними лініями. На осі абсцис відкладаються варіанти досліджуваного показника (або часові параметри), а на осі ординат — величини досліджуваного показника. За відмітками (точками) обох осей координат визначається положення кожного рівня на полі графіка. Послідовно поєднуючи точки відрізками ліній, отримують емпіричну лінію графіка. За виглядом цієї лінії можна судити про характер розвитку досліджуваного явища у просторі або у часі. Як приклад лінійної діаграми, наведемо графічне зображення зміни курсу валют (див. рис. 2.2).

Для побудови лінійної діаграми важливо правильно обрати масштаб (як за віссю абсцис, так і за віссю ординат). Від оптимального співвідношення цих масштабів залежить наочність (зрозумілість) статистичного графіка. Якщо масштаб шкали за віссю абсцис буде надто розтягнутий, то зміни рівнів у часі будуть недостатньо помітними. Однак при розтягнутому масштабі за віссю ординат коливання рівнів будуть надто різкими. Для виключення зорових ілюзій, що спотворюють уявлення про

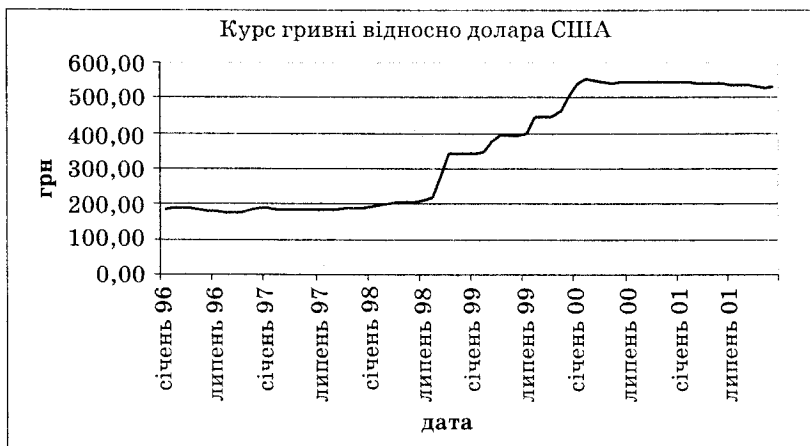


Рис. 2.2. Лінійна діаграма

реальний стан речей, доцільно побудувати координатну сітку з урахуванням правила «золотого перерізу», за якого формат поля графіка обирається у співвідношенні від 1 : 1,3 до 1 : 1,5. Необхідно суворо дотримуватися встановленого масштабу не лише за віссю ординат, але й за віссю абсцис. Якщо показання часу між окремими періодами (або датами) нерівновеликі, то при нанесенні на поле графіка точок досліджуваних рівнів на вісь абсцис необхідно строго дотримувати співвідношення між ними.

Важливою перевагою лінійних графіків є те, що на одному й тому самому полі графіка можна зобразити декілька показників, що дозволяє порівнювати та виявляти специфіку їх розвитку у часі або характер зміни одного показника за різними об'єктами у просторі або території. При цьому слід ураховувати, що кожен показник треба зображувати лінією певного вигляду (суцільна, пунктирна тощо) або певного кольору.

Небажано зображувати на одному лінійному графіку велику кількість показників, оскільки це призводить до втрати наочності. Якщо ж на одному графіку необхідно відобразити різнойменні показники, то можуть виникнути ускладнення, пов'язані з відмінністю у їхній розмірності.

Лінійні графіки іноді будуються з логарифмічною шкалою за віссю ординат. Координатну сітку, в якій за віссю абсцис нанесена шкала у рівномірному масштабі, прийнято називати *арифметичною*.

Графіки з рівномірною шкалою за віссю ординат надають достатньо наочне уявлення про зміни досліджуваних абсолютних показників. Арифметичні графіки мають арифметичну (рівномірну) шкалу — як за віссю абсцис, так і за віссю ординат.

У порівнянні з ними більші можливості мають напівлогарифмічні графіки, оскільки вони можуть наочніше відобразити залежність відносних змін однієї змінної від абсолютних змін іншої (*див.* рис. 2.3).

Особливістю напівлогарифмічного графіка є те, що одна з його осей координат має логарифмічну шкалу, а інша — арифметичну. Оскільки відносна зміна деякої величини пропорційна до абсолютної зміни її логарифма, то нахил напівлогарифмічного графіка цієї величини характеризує відносну його зміну в залежності від абсолютної зміни іншої величини.

Смужкові та стовпчикові діаграми призначені для порівняння величин за допомогою пропорційних до них за довжиною

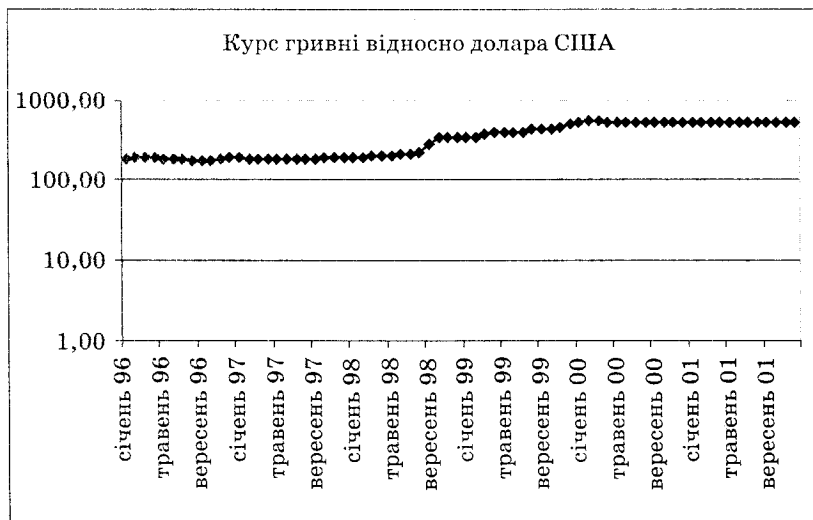


Рис. 2.3. Напівлогарифмічний графік

смуг або стовпчиків, що розташовуються горизонтально або вертикально. Стовпчикові діаграми зазвичай використовуються для графічного зображення рядів динаміки. Смушкові діаграми використовуються, наприклад, для зображення розміру або структури будь-яких однорідних явищ.

Для побудови стовпчикових діаграм використовується, як і в лінійних графіках, прямокутна система координат. При цьому кожне значення досліджуваного показника зображується у вигляді вертикального стовпчика, основа якого розміщується на осі абсцис. Ширина стовпчиків може бути довільною, але обов'язково однаковою для кожного стовпчика. Висота стовпчиків (у відповідності до прийнятого за віссю ординат масштабом) має строго відповідати зображуваним даним.

Кількість стовпчиків визначається числом досліджуваних показань (даних). Відстань між стовпчиками повинна бути однаковою. Біля основи стовпчиків подається назва досліджуваного показника. Рівні (величини), що характеризують значення зображуваних показників, поміщуються всередині кожного стовпчика.

У статистиці знаходять застосування й так звані *стрічкові* (смугові) *графіки*. На цих діаграмах основи стовпчиків розташовуються вертикально, а масштабна шкала наноситься на

горизонтальну вісь. За своєю формою стрічкова діаграма являє собою ряд смуг однакової ширини, що простираються за віссю абсцис. Довжина смуг (стрічок) відповідає значенням зображуваних показників. При побудові стрічкових діаграм дотримуються таких самих вимог, що й при побудові стовпчикових графіків (однакова ширина смуг, початок масштабної шкали від нульової позначки).

Лінійні, стовпчикові й смугові діаграми мають широке застосування у поданні статистичної інформації. При цьому лінійним та стовпчиковим діаграмам надається перевага, якщо у досліджуваних показниках виявляється загальна тенденція до зростання. Горизонтальні стрічки (смуги) більш наочні, якщо зображувані показники відображують результат (підсумок) функціонування (див. рис. 2.4).

Графічними формами зображення простих розподілів частот є полігон частот, гістограма та згладжена крива розподілу. Для побудови *полігону частот* вісь абсцис розбивається на інтервали значень величини, що аналізується. Відповідні частоти появи цих значень в кожному інтервалі проставляються у вигляді точок за віссю ординат навпроти його середини. Далі ці точки поєднуються прямими лініями.

Гістограма будується за допомогою вертикальних стовпчиків, основами яких слугують інтервали значень величини, що аналізується, а їх площі дорівнюють частотам появи цих значень



Рис. 2.4. Смуужкова діаграма

у відповідних інтервалах, які можуть бути рівної або нерівної довжини. Головна мета отримання згладжених кривих розподілу полягає в усуненні впливу випадкових факторів на характер розподілу. Для деяких цілей більш корисною у порівнянні з простим графіком розподілу частот виявляється крива накопичених (кумулятивних) частот. В останньому випадку частоти, які відповідають послідовним інтервалам, накопичуються, починаючи з одного з кінців розподілу (див. рис. 2.5).

Кругові діаграми застосовуються для зображення структури будь-якої сукупності даних діаграми. В цих діаграмах площа кола приймається за величину всієї досліджуваної статистичної сукупності, а площі окремих секторів відображають питому вагу (частку) її складових частин. При цьому оскільки площі секторів пропорційні їхнім центральним кутам, то для побудови секторної діаграми сума всіх кутів (360°) розподіляється пропорційно до питомої ваги окремих частин досліджуваної сукупності. При відсотковому поданні складу досліджуваної статистичної сукупності виходять зі співвідношення $1\% = 3,6^\circ$ (див. рис. 2.6).

Для вивчення статистичної інформації про комерційну діяльність на ринку товарів та послуг застосовують так звані *радіальні діаграми*. Будуються вони на базі полярних координат. Початком відліку в них слугує центр кола, а носіями масштаб-

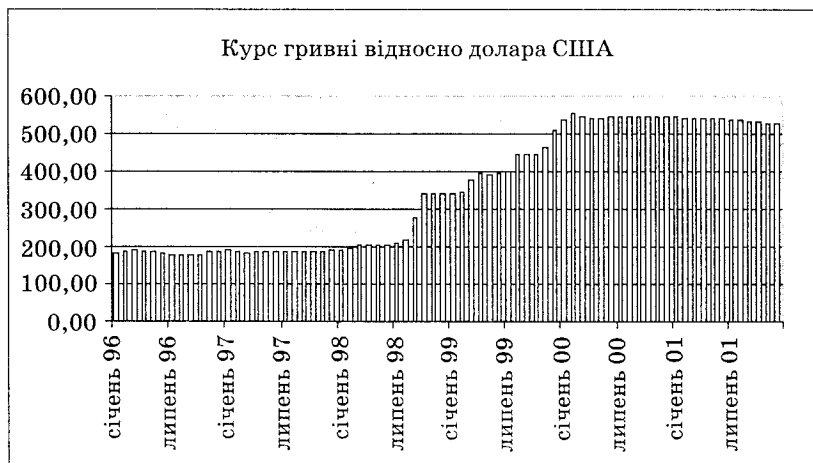


Рис. 2.5. Гістограма



Рис. 2.6. Кругова діаграма

них шкал є його радіуси. Зазвичай в основі радіальних діаграм лежать повторювані річні цикли з помісячними або поквартальними даними. Так, при вивченні річного циклу з помісячними даними коло ділять радіусами на 12 рівних частин. Кожному радіусу дається назва місяця року, а їх розташування подібно до циферблата годинника. На кожному радіусі у відповідності зі встановленим масштабом наносяться точки, що відповідають досліджуваному за кожен місяць даним. Отримані таким чином точки поєднують між собою лініями. В результаті виходить спіралеподібна лінія, що характеризує внутрішньорічні цикли комерційної діяльності.

Діаграми розсіювання дозволяють зобразити в графічній формі степінь і тип залежності між двома змінними. *Рисункові, діаграми* використовуються, головним чином, через наочність, хоча вони рідко мають більшу інформативність, ніж традиційні графічні засоби. В останні роки на практиці отримали розповсюдження графіки і діаграми у тривимірній аксонометричній проекції з дотриманням законів перспективи, які, безперечно, наочні й зрозумілі.

Різновидом графіків слід вважати також макети, ескізи, схеми та взагалі будь-які засоби і моделі для наочного й ефективного представлення в цілому неозорих реальних об'єктів та процесів.

Особливим видом графіка накопичених частот є *крива концентрації*, або *крива Лоренца*. Цей графік використовується для зображення ступеня нерівномірності розподілу таких величин, як багатство, дохід, обсяг роздрібної торгівлі тощо. Чим більше крива розподілу відхиляється від прямої лінії, тим в більшій мірі виявляється нерівномірність розподілу.

Для зображення статистичних даних застосовуються різноманітні мапи. *Штриховані мапи* застосовуються для зображення ступеня відносної виразності явищ у чітко розмежованих територіальних одиницях — район, країна, область і т. п. Чим сильніше виражене явище, тим густіше штрихування.

Символьні карти, на відміну від штрихованих карт, застосовуються для зображення абсолютних частот або величин, що характеризують які-небудь явища. Через розмір, число, щільність розташування та затінення символу можна виразити певну величину. Причому для різних явищ застосовуються різні форми символів.

Для зображення статистичних даних можуть також застосовуватися *графіки ізоліній*. Якщо лінії на цих графіках проведені через точки з однаковою величиною або інтенсивністю явища, то вони називаються ізометричними. Якщо ж ці лінії зв'язують між собою рівні відносини двох явищ, то ці графіки називаються ізоплотами.

Особливим сучасним різновидом є *комп'ютерні графіки* (свого роду графіки графіків), представлені в канонічному вигляді або у вигляді дерев цілей.

Нарешті, графіки, зокрема в техніці, — це різновид технології виготовлення машин і устаткування; на графічній основі будуються найсучасніші комп'ютери і комп'ютерні програми; у графічній формі подаються комп'ютерні «продукти», наприклад виготовлені автоматизованим способом проекти; власне кажучи, графіками є також комп'ютерні дані — як вхідні, так і вихідні.

В останні роки отримав розвиток новий науковий напрямок в обробці даних (Computational Intelligence), що вивчає взаємозв'язок між нейронними сітками, генетичними алгоритмами і нечіткими системами.

У складних багаторівневих ієрархічних системах управління мають місце різні види невизначеностей. При розв'язанні багатьох функціональних задач оперативно-диспетчерського управ-

ління в АСК, як наприклад, оперативного і перспективного планування, розміщення, керування запасами тощо, користуються спрощеними моделями зі строгими обмеженнями, що знижує результати досліджень і часто сприяє невірним рішенням. У такій ситуації останнім часом пропонується використовувати апарат теорії нечітких множин для прийняття ефективних рішень. Основні додатки цього підходу знаходяться в таких галузях, як штучний інтелект, лінгвістика, пошук інформації, процеси прийняття рішень, розпізнавання образів, медична діагностика, психологія, право, економіка й інших галузях людської діяльності.

У 1965 р. Л. Заді представив першу роботу з нечітких множин [11]. Ця робота була присвячена опису явищ і понять, що мають багатозначний і неточний характер. При розв'язанні більшості прикладних задач в АСК використовується числова (кількісна) і якісна інформація. Другий вид інформації найчастіше представляється у формі бази нечітких правил, які можна успішно використовувати для розв'язання більшості функціональних задач в АСК. Успішно теорія нечітких множин знаходить застосування й у стохастичних системах. Це пов'язано з труднощами отримання точних багатовимірних законів розподілу й інших імовірнісних характеристик. У теорії статистичного оцінювання крім точечного визначення параметра вказується й інтервальна оцінка. Для отримання функції, що характеризує припустимість кожного значення всередині заданого інтервалу на основі статистичної інформації можливе успішне використання теорії нечітких множин. З'являється можливість отримати нечітку величину, для якої за максимумом значення функції отримати точечну оцінку.

Для поводження (використання) з неточно відомими величинами зазвичай застосовується апарат теорії ймовірностей. Однак випадковість пов'язана з невизначеністю, що стосується приналежності деякого об'єкта до звичайної множини. Це розходження між нечіткістю і випадковістю призводить до того, що математичні методи нечітких множин зовсім не схожі на методи теорії ймовірностей. Вони в багатьох відношеннях простіші внаслідок того, що поняттю імовірнісної міри в теорії ймовірностей відповідає більш просте поняття функції приналежності в теорії нечітких множин. З цієї причини навіть у тих випадках, коли невизначеність у процесі прийняття рішень може бути представлена імовірнісною моделлю, зазвичай зручніше оперувати з нею

методами теорії нечітких множин без залучення апарата теорії ймовірностей.

Підхід на основі теорії нечітких множин виявляється, по суті справи, альтернативою загальноприйнятим кількісним методам аналізу систем. Такий підхід дає наближені, але в той же час ефективні способи опису поведінки систем, настільки складних і погано визначених, що вони не піддаються точному математичному аналізу. Теоретичні ж основи цього підходу цілком точні й строгі в математичному сенсі й не є самі по собі джерелом невизначеності. У кожному конкретному випадку ступінь точності розв'язку може бути узгоджений з вимогами задачі й точністю наявних даних. Подібна гнучкість складає одну з важливих рис методу, що розглядається.

Нейронні сітки представляють великий клас систем, архітектура яких має аналогію (як тепер відомо, досить слабку) з побудовою нервової тканини з нейронів. В одній з найбільш розповсюджених архітектур, багат шаровому персептрону з зворотним (оберненим) поширенням помилки, імітується робота нейронів у складі ієрархічної сітки, де кожен нейрон більш високого рівня з'єднаний своїми входами з виходами нейронів розташованого нижче шару. На нейрони найнижчого шару подаються значення параметрів, на основі яких потрібно приймати рішення, прогнозувати розвиток ситуації тощо.

Ці значення розглядаються як сигнали, що передаються у наступний шар, послабляючись або підсилюючись в залежності від числових значень (ваг), що приписуються міжнейронним зв'язкам. У результаті на виході нейрона найверхнього шару виробляється деяке значення, що розглядається як відповідь — реакція всієї сітки на введені значення вхідних параметрів. Для того щоб сітку можна було застосовувати надалі, її передусім треба «натренувати» на отриманих раніше даних, для яких відомі і значення вхідних параметрів, і правильні відповіді на них. Тренування полягає у підборі ваг міжнейронних зв'язків, що забезпечують найбільшу близькість відповідей сітки до відомих правильних відповідей.

Дерева розв'язання (decision trees) являються одним з найбільш популярних підходів до розв'язання задач Data Mining. Вони створюють ієрархічну структуру правил класифікуючого типу «ЯКЩО... ТО...» (if-then), яка має вигляд дерева. Для ухвалення рішення, до якого класу віднести деякий об'єкт або

ситуацію, потрібно відповісти на питання, що стоять у вузлах цього дерева, починаючи з його кореня. Питання мають вигляд «значення параметра A більше x ?». Якщо відповідь позитивна, здійснюється перехід до правого вузла наступного рівня, якщо негативна — до лівого вузла; потім знову впливає питання, пов'язане з відповідним вузлом.

Популярність підходу пов'язана ніби з наочністю і зрозумілістю. Але дерева розв'язань принципово не здатні знаходити «кращі» (найбільш повні і точні) правила в даних. Вони реалізують спрощений принцип послідовного перегляду ознак і «чіпляють» фактично осколки дійсних закономірностей, створюючи лише ілюзію логічного висновку.

Разом з тим, більшість систем використовують саме цей метод. Найвідомішими з них є See5/C5.0 (RuleQuest, Австралія, <http://www.rulequest.com/>), Clementine (Integral Solutions, Велика Британія, <http://www.spss.com/clementine/>), SIPINA (University of Lyon, Франція), IDIS (Information Discovery, США), KnowledgeSEEKER (ANGOSS, Канада).

Еволюційне програмування знаходить усе більш широке застосування при розв'язанні функціональних задач ІУС.

Проілюструємо сучасний стан даного підходу на прикладі системи PolyAnalyst — вітчизняної розробки, що одержала сьогодні загальне визнання на ринку DataMining. У цій системі гіпотези про вид залежності цільової змінної від інших змінних формуються у виді програми на деякій внутрішній мові програмування. Процес побудови програм будується як еволюція у світі програм (цим підхід трохи схожий на генетичні алгоритми). Коли система знаходить програму, що більш-менш задовільно виражає шукану залежність, вона починає вносити в неї невеликі модифікації і відбирає серед побудованих дочірніх програм ті, які підвищують точність. Таким чином, система «ви-рощує» кілька генетичних ліній програм, що конкурують між собою в точності вираження шуканої залежності. Спеціальний модуль системи PolyAnalyst перекладає знайдені залежності з внутрішньої мови системи на зрозумілу користувачеві мову (математичні формули, таблиці тощо).

Інший напрямок еволюційного програмування пов'язано з пошуком залежності цільових змінних від інших у формі функцій якогось визначеного виду. Наприклад, в одному з найбільш вдалих алгоритмів цього типу — методі групового обліку

аргументів (МГОА) — залежність шукають у формі поліномів. Сьогодні з систем, що продаються в Росії, МГОА реалізований у системі NeuroSbell компанії Ward Systems Group.

Генетичні алгоритми треба розглядати скоріше як могутній засіб розв'язання різноманітних комбінаторних задач та задач оптимізації. Проте генетичні алгоритми ввійшли зараз у стандартний інструментарій методів DataMining, тому вони й включені в цей огляд.

Перший крок при побудові генетичних алгоритмів — це кодування вихідних логічних закономірностей у базі даних, які іменують хромосомами, а весь набір таких закономірностей називають популяцією хромосом. Далі для реалізації концепції відбору вводиться спосіб зіставлення різних хромосом. Популяція обробляється за допомогою процедур репродукції, змінюваності (мутацій), генетичної композиції. Ці процедури імітують біологічні процеси. Найбільш важливі серед них: випадкові мутації даних в індивідуальних хромосомах, переходи (кроссінговер) і рекомбінація генетичного матеріалу, що міститься в індивідуальних батьківських хромосомах (аналогічно гетеросексуальній репродукції), і міграції генів. Під час роботи процедур на кожній стадії еволюції виходять популяції зі все більш досконалими індивідуумами.

Генетичні алгоритми зручні тим, що їх легко розпаралелювати. Наприклад, можна розбити покоління на кілька груп і працювати з кожною з них незалежно, обмінюючись час від часу декількома хромосомами. Існують також й інші методи розпаралелювання генетичних алгоритмів.

Генетичні алгоритми мають низку недоліків. Критерій відбору хромосом і використовувані процедури являються евристичними і зовсім не гарантують знаходження «кращого» рішення. Як і в реальному житті, еволюцію може «заклинити» на якій-небудь непродуктивній гілці. І, навпаки, можна навести приклади, як двоє безперспективних батьків, які будуть виключені з еволюції генетичним алгоритмом, виявляються здатними зробити високо-ефективного нащадка. Це стає особливо помітно при розв'язанні високорозмірних задач зі складними внутрішніми зв'язками.

Прикладом може слугувати система GeneHunter фірми Ward System Group.

Алгоритми обмеженого перебору були запропоновані в середині 60-х років М. М. Бонгардом для пошуку логічних закономірнос-

тей у даних. З тих пір вони продемонстрували свою ефективність при розв'язанні великої кількості задач з різних галузей.

Ці алгоритми обчислюють частоти комбінацій простих логічних подій у підгрупах даних. Приклади простих логічних подій: $X = a$; $X < a$; $X > a$; $a < X < b$ та ін., де X — будь-який параметр, a та b — константи. Обмеженням слугує довжина комбінації простих логічних подій (у М. Бонгарда вона дорівнювала 3). На підставі аналізу обчислених частот робиться висновок про корисність тієї чи іншої комбінації для встановлення асоціації в даних, для класифікації, прогнозування тощо.

Найбільш яскравим сучасним представником цього підходу є система WizWhy виробництва WizSoft. Хоча автор системи Абрахам Мейдан не розкриває специфіку алгоритму, покладеного в основу роботи WizWhy, за результатами ретельного тестування системи були зроблені висновки про наявність у ній обмеженого перебору (вивчалися результати, залежності часу їхнього одержання від числа аналізованих параметрів тощо).

Автор WizWhy стверджує, що його система виявляє всі логічні if-then-правила в даних. Насправді це, звичайно, не так. По-перше, максимальна довжина комбінації в if-then-правилі в системі WizWhy дорівнює 6, і, по-друге, з самого початку роботи алгоритму проводиться евристичний пошук простих логічних подій, на яких потім будується весь подальший аналіз. Зрозумівши ці особливості WizWhy, неважко було запропонувати найпростішу тестову задачу, яку система взагалі не змогла розв'язати. Інший момент — система видає рішення за прийнятний час тільки для порівняно невеликої розмірності даних.

Проте система WizWhy є на сьогоднішній день одним з лідерів на ринку продуктів DataMining, і не безпідставно. Система постійно демонструє більш високі показники при розв'язанні практичних задач, ніж усі інші алгоритми.

Регресійний аналіз (РА) — один з широко розповсюджених методів побудови математичних моделей за статистичними даними у інформаційно-управляючих системах і застосовується в ситуаціях, коли одна досліджувана змінна має очікуване значення, що залежить від значень інших змінних. Передбачається, що ця залежність належить до певного функціонального типу з декількома невідомими параметрами. Наприклад, можна припустити, що при відповідних умовах очікувана оцінка на іспиті є лінійною функцією від тривалості підготовки до нього. В цьому випадку

ця функція характеризується двома параметрами: нахилом до осі абсцис і точкою перетину з віссю ординат.

При розгляді регресійного зв'язку тільки одна зі змінних вважається випадковою, значення інших змінних розглядаються або як фіксовані, якщо вони контролюються в ході експерименту, або як відібрані деяким чином з великої кількості можливих значень. Залежність між очікуваним значенням змінної (називаної також залежною, прогнозованою чи регресійною змінною) і не випадковими змінними (називаними також аргументами регресії, незалежними змінними, предикторами або регресорами) називається регресійним зв'язком.

Таким чином, *регресійний зв'язок* — це напрямлений зв'язок між змінними, у якому очікуване значення однієї випадкової змінної співвідноситься з не випадковими значеннями інших змінних. Цей тип зв'язку відрізняється від інших типів статистичного зв'язку, особливо від кореляційних зв'язків. Кореляція являє собою зв'язок між двома або більшим числом випадкових змінних і може бути подана як залежність ступеня варіації однієї змінної від варіації іншої змінної або декількох змінних.

Регресійний зв'язок, хоча і співвідносить одну випадкову змінну з деякими іншими не випадковими змінними, у багатьох випадках може бути застосований і коли його аргументи є випадковими. У цьому випадку може бути визначена умовна регресія щодо значень, що спостерігаються, його випадкових аргументів. При цьому очікуване значення однієї випадкової змінної співвідноситься зі значеннями, що спостерігаються, інших випадкових змінних. Не слід ігнорувати передумови й припущення, в основі яких лежить регресійний аналіз. Незважаючи на довгу історію, РА постійно розвивається і знаходить застосування в різних предметних галузях.

Розвиток РА пов'язаний з розробкою нових напрямків, що спрямовані на подолання труднощів, зумовлених порушеннями передумов класичної процедури.

При застосуванні РА необхідно велику увагу приділяти аналізу зумовленості матриць системи нормальних рівнянь, отриманих методом МНК. У випадку наявності мультиколінеарності доцільно перейти до методів зміщеного оцінювання.

Порушення нормальності розподілу помилки моделі призводить до необхідності використання різних типів робасних оцінок, непараметричних методів.

Поняття робасності методів запропоновано Дж. Боксом (1953). За допомогою цих методів можуть бути розроблені оцінки та критерії перевірки статистичних гіпотез, сталі у відношенні до різного роду викривлень вихідних даних у вибірках, порушенню вихідних припущень щодо виду розподілу даних (тих, що розглядаються як значення випадкових величин), а також до інших відхилень від умов, що задають стохастичні моделі.

Наявність помилок у рівнях факторів, їх корельованість та неоднорідність вимагає від дослідника вміння своєрідного підходу до використання РА.

Класифікації й групування являють собою систему супідрядних понять про властивості об'єктів, що відповідають назвам класів, тобто множність усіх потенційно чи реально існуючих об'єктів, що володіють даними властивостями. На відміну від класу група об'єктів — це яка-небудь частина реально існуючих об'єктів з даного класу. Під класифікацією як операцією розуміється визначення найменування класу з деякої їхньої взаємопов'язаної сукупності, до якого належать який-небудь індивідуум або об'єкт, виходячи з його властивостей, що спостерігаються. Групування — це яка-небудь сукупність розкласифікованих деяким чином об'єктів. Коли характеристики об'єктів являють собою числові змінні, приписування об'єктів до визначених груп іноді називається дискримінацією, а відповідна йому функція від значень, що спостерігаються, цих змінних — дискримінантною.

Проблема класифікації виникає, коли дослідник не може безпосередньо віднести певного індивідуума до якої-небудь категорії, а повинен встановити цю категорію виходячи з вимірювань, що характеризують його, відповідей на задані йому питання або інших особливостей. У багатьох випадках можна припустити, що існує обмежена множина статистичних сукупностей, одним з елементів яких може виявитися даний індивідуум, і що для кожної сукупності визначені статистичні розподіли властивостей індивідуумів, що складають їх. При цьому індивідуум, що класифікується, розглядається як випадкове спостереження з однієї з цих сукупностей. Потрібно визначити, до якої сукупності відноситься даний індивідуум, що володіє певними характеристиками.

Р. Л. Фішер у 1936 р. розглянув випадок двох сукупностей і вперше запропонував лінійну дискримінантну функцію, що

являє собою лінійну комбінацію вимірів. У якості значення цієї функції можуть бути прийняті, наприклад, одиниця, якщо спостереження належить до першої сукупності, і нуль, якщо воно належить до другої сукупності. Невідомі параметри дискримінантної функції можуть бути визначені на підставі лінійної регресійної моделі за спостереженнями результатів вимірювань і двома сукупностями.

Методи факторного аналізу призначені для розв'язання таких трьох задач:

☑ виявлення залежностей між об'єктами і між змінними шляхом скорочення розмірності, тобто числа вимірювань, що використовуються при їхньому моделюванні;

☑ виявлення лінійних залежностей між змінними, а також між об'єктами;

☑ встановлення прихованих факторів, що впливають на досліджувані змінні.

Найбільш важливе в застосуванні факторного аналізу — це скорочення розмірності. Якщо в заданій матриці даних кожен об'єкт характеризується деяким числом координат (змінних), що безпосередньо спостерігаються, то скорочення розмірності виражається в скороченні цього числа таким чином, щоб деяке менше число схованих факторів ідентифікувало ці об'єкти приблизно з тією же точністю, тобто мали приблизно ту ж інформативність, що і спочатку задані координати. Аналогічно можуть бути скорочені координати, що описують у цій матриці кожен досліджувану змінну.

Дуже схожою з факторною моделлю з формальної точки зору, але відмінною від неї за цілями є модель **методу головних компонент**, розроблена Гарольдом Хотеллінгом у 1933 р. [2]. У факторній моделі кожна досліджувана змінна подається у вигляді суми лінійної комбінації випадкових схованих факторів помилки. Передбачається, що ці випадкові помилки для різних змінних мають нормальні розподіли з нульовими середніми та з різними дисперсіями, які незалежні одне від одного і від факторів. Фактори також передбачаються нормально розподіленими і незалежними друг від друга, що мають нульові середні й одиничні дисперсії. Коефіцієнти в лінійній комбінації факторів називаються навантаженнями (вагами). У цій моделі потрібно оцінити невідомі дисперсії випадкових помилок і матрицю навантажень досліджуваних змінних на фактори.

Метод головних компонент полягає у відборі декількох найбільш інформативних компонент (координат). Історія застосування цього методу показує, що він зазвичай використовувався для скорочення розмірності за об'єктами, у той час як факторний аналіз — в основному за досліджуваними змінними. У моделі методу головних компонент не робиться ніяких припущень про тип розподілів, що характеризують фактори, а тільки передбачається, що усі вони мають нульові середні й одиничні дисперсії. На відміну від моделі факторного аналізу в цій моделі кожна досліджувана змінна подається у вигляді лінійної комбінації факторів без випадкових помилок.

Основною перевагою моделі факторного аналізу в порівнянні з моделлю методу головних компонент є те, що зміна масштабів (ділення або множення на деякий коефіцієнт) вимірювань деяких чи всіх досліджуваних змінних виражається в пропорційних змінах елементів матриці навантажень. Метод головних компонент з цього погляду не забезпечує їхню незалежність від таких змін масштабу в досліджуваних змінних. Крім того, цей метод застосовується, коли всі досліджувані змінні виміряні за кожним з об'єктів.

Методи аналізу стохастичних часових рядів. З чисто емпіричної точки зору динамічний ряд являє собою послідовність числових даних, у якій кожен елемент пов'язується з певним моментом (періодом) часу. В загальному випадку, якщо одночасно в кожен момент часу спостерігається відразу декілька змінних, наприклад показників економічного розвитку країни, вони можуть бути розглянуті як послідовність векторних величин, названих багатовимірним рядом динаміки.

З теоретичної точки зору ряд динаміки являє собою послідовність фіксованих значень впорядкованих у часі випадкових величин, які у своїй сукупності називаються стохастичним процесом. При цьому ряд динаміки називається також реалізацією стохастичного процесу.

Аналізуючи ряди динаміки, необхідно враховувати такі дві характерні риси. По-перше, між спостереженнями явно передбачається статистична залежність, а по-друге, як правило, навіть у принципі їх неможливо незалежно відтворити хоча б ще один раз. Наприклад, фактична зміна обсягу валового суспільного продукту країни по роках унікальна і не може бути, як правило, повторена з абсолютною точністю. Тому більшість задач аналізу

рядів динаміки припускає отримання висновків про властивості відповідного стохастичного процесу лише за даними про одну його реалізацію.

Ці дві особливості рядів динаміки — внутрішня залежність і неможливість відтворення — змушують при розробці моделей і методів їхнього аналізу робити ряд додаткових апріорних припущень про їхні імовірнісні властивості. Однак ці моделі і методи не мали б практичної цінності, якби ряди динаміки, що реально спостерігаються, не володіли низкою закономірностей.

Одна з форм цього роду закономірностей, що спостерігається в деяких рядах динаміки, — *стаціонарність*. Властивість стаціонарності для ряду динаміки означає, що його імовірнісні характеристики не залежать від часу спостереження. Зокрема, його математичні сподівання та дисперсія не змінюються в часі, а автоковаріація між спостереженнями залежить лише від тривалості часу між ними (лага).

Відступ від властивості стаціонарності може виражатися, наприклад, у зміні в часі математичних сподівань або дисперсій. Однак при цьому вихідний процес може бути перетворений таким чином, що процес, який вийшов у результаті, буде стаціонарним. Наприклад, якщо математичні сподівання спостережень змінюються в часі, то різниця між спостереженням і його математичним сподіванням може бути розглянута як стаціонарний процес з нульовою середньою.

Сфера застосування стаціонарних моделей може бути розширена й іншим шляхом, якщо припустити, що спостереження ряду динаміки отримані, виходячи з деякої сукупності значень стаціонарного процесу, що не спостерігається безпосередньо. Наприклад, можна припустити, що різниця між поточним і попереднім спостереженням являє собою стаціонарний процес. При цьому кожне спостереження — це сума значень стаціонарного процесу.

Одним з перших методів дослідження рядів динаміки був *метод екстраполяції*, що полягав у припущенні рівномірної зміни рівнів ряду, нанесенні спостережень на графік і їхньому інтуїтивному аналізі. Однак без об'єктивних доказових процедур цей метод міг легко призвести до помилкових висновків про нібито існуючі цикли. Це особливо виявлялося при використанні рядів динаміки в прогнозуванні.

При дослідженні рядів динаміки за допомогою регресійного аналізу головна увага звертається на систематичні зміни в їхніх

рівнях. Кожне спостереження при цьому розглядається як сума деякого полінома, що залежить від часу, однієї чи декількох тригонометричних функцій з відомим періодом, а також стаціонарного процесу (випадкової помилки) з нульовим математичним сподіванням. Оцінка невідомих параметрів зазвичай здійснюється за допомогою *методу найменших квадратів*. При аналізі рядів динаміки за допомогою *корелограм* основна увага приділяється виявленню залежності між спостереженнями в різні моменти часу. Корелограма ряду динаміки являє собою графік коефіцієнтів кореляції (автокореляції) між поточними спостереженнями і попередніми з визначеними лагами в часі. Якщо корелограма має різко виражений максимум або яку-небудь іншу особливість, то можна стверджувати про наявність певної закономірності.

Обидва ці підходи все ще зберігають практичну цінність і є частиною більш досконалих методів, таких, як спектральний аналіз і авторегресійні моделі.

Особливий вид — екстремальні ряди динаміки, що відкривають можливість працювати з незвичайною інформацією, зокрема з робасними оцінками.

Методи лінійного програмування [2]. Це методи розв'язання систем лінійних рівнянь стосовно до розподілу ресурсів розвитку конкретних економічних систем (підприємств, галузей, груп галузей, матеріального виробництва і народного господарства в цілому). Велика розмірність досліджуваних систем (і відповідно моделей) визначає необхідність рішення потрібних систем лінійних рівнянь за допомогою сучасних ЕОМ. Тому іноді досліджувані методи називаються ще *комп'ютерними методами лінійного програмування*.

Методи розв'язання систем лінійних алгебраїчних рівнянь розбиваються на дві групи. До першої групи належать так звані точні чи прямі методи — алгоритми, що дозволяють отримати розв'язання системи за кінцеве число арифметичних дій. Сюди відноситься відоме правило Крамера знаходження розв'язку за допомогою визначників, метод Гаусса (метод виключень) і метод прогону. Другу групу складають наближені методи, зокрема ітераційні методи розв'язання систем лінійних рівнянь. Пряме обчислення визначника, що використовує його явне вираження через елементи матриці, рідко застосовується на практиці через його трудомісткість. Один з найефективніших способів обчислення визначників заснований на методі Гаусса.

Постановка задачі лінійного програмування була запропонована Л. В. Канторовичем. У 1940 р. він же разом з М. К. Гавуріним розробив метод потенціалів для розв'язання часткового випадку загальної задачі лінійного програмування — транспортної задачі. Надалі теорія лінійного програмування одержала розвиток у роботах Дж. фон Неймана, Д. Гейла, Г. Куна і Л. Таккера, Г. Ланцига й ін.

Лінійне програмування являє собою важливу складову частину загальної теорії математичного програмування. Математичне програмування — це поняття, застосовуване для позначення певних математичних методів, призначених для рішення проблеми пошуку мінімуму або максимуму деякої функції при певних обмеженнях, які зазвичай виражені у виді нерівностей.

До задачі лінійного програмування в ІУС може бути зведена, наприклад, задача максимізації якого-небудь ефекту, зокрема прибутку, від виготовлення деякої сукупності різних товарів за заданих умов їхнього виробництва. Передбачаються заданими: величина прибутку від виробництва і реалізації одиниці кожного виду товарів; обсяги наявних ресурсів різних видів, а також витрати кожного виду ресурсу на виробництво одиниці кожного виду товару. Потрібно визначити розміри або інтенсивності виробництва товарів, за яких максимізується сумарний прибуток, а також не порушуються обмеження на обсяги наявних ресурсів. Сукупність математичних виражень, за допомогою яких моделюється ця задача, називається *лінійною програмою*. Заданими змінними величинами (параметрами) у цій моделі є розміри виробництва кожного виду товару. Рівняння, що характеризує залежність загального обсягу прибутку від параметрів моделі, називається цільовою функцією. Сукупність заданих не від'ємних параметрів, що задовольняють всім обмеженням моделі, називається припустимим рішенням, яке також називається оптимальним, якщо при ньому досягається максимальне значення цільової функції.

Обмеження на обсяги наявних ресурсів можуть бути представлені у вигляді рівностей, якщо в модель задачі ввести нові невід'ємні параметри, що виражають наявність або відсутність вільних залишків невикористаних ресурсів за їх видами. Тому ці змінні можуть бути названі вільними і відмінність від уже розглянутих заданих змінних моделі. Введення вільних змінних не змінює задані параметри ні в оптимальному розв'язку задачі, ні в її припустимих розв'язках.

У перетвореній зазначеним образом моделі припустимим розв'язком називається сукупність невід'ємних заданих і вільних змінних, що задовольняють представленим у вигляді рівностей обмеженням на обсяги наявних ресурсів. Якщо в цьому розв'язку мається не менше додатних елементів, ніж число видів ресурсів (або відповідних обмежень), то він називається базисним. Відповідно до теорії лінійного програмування, якщо ця задача має припустимий розв'язок з деяким відповідним йому обсягом прибутку, то вона має і базисне розв'язання з не меншим обсягом прибутку. Ця властивість, таким чином, дозволяє обмежити пошук оптимальних розв'язків рамками множини базисних рішень.

Один з основних результатів теорії лінійного програмування полягає в тому, що з будь-якою лінійною програмою, називаною прямою, може бути зв'язана інша програма, називана двоїстою до неї. Залежність між прямою і двоїстою до неї лінійною програмою характеризується низкою важливих математичних теорем і змістовних економічних інтерпретацій.

Слід також зазначити, що прибуток, одержуваний від виробництва одиниці кожного виду товару, може бути розглянутий як ефективність приросту виробництва даного виду товару, тобто як приріст сумарного ефекту від виробництва всіх товарів, що припадають на одиницю приросту виробництва даного товару.

Двоїста задача полягає в мінімізації її цільової функції, що являє собою сумарний ефект від споживання невикористаних залишків запасів ресурсів при фіксованому виробництві товарів. Іншими словами, потрібно забезпечити найбільш ефективне використання наявних обсягів різних видів ресурсів. Відповідно до теорем теорії подвійності, якщо і пряма, і двоїста задача мають припустимі рішення, то цільова функція прямої задачі — сумарний ефект від виробництва товарів — не перевищує сумарного ефекту від виробничого споживання наявних ресурсів. При цьому обидві ці задачі мають оптимальні розв'язки, за яких ці ефекти збігаються.

Якщо при оптимальному розв'язанні прямої (двоїстої) задачі одне з обмежень двоїстої (прямої) задачі є строгою нерівністю, то відповідна цій нерівності змінна в прямій (двоїстій) задачі дорівнює нулю. Якщо ж змінна в прямій (двоїстій) задачі є додатною, то відповідне їй обмеження в двоїстій (прямій) задачі є рівністю.

У термінах уже розглянутої економічної інтерпретації зміст цих теорем полягає у такому. Якщо деякий ресурс, що має місце,

не використовується повністю, якщо відповідне йому обмеження в прямій задачі — строга нерівність, то ефективність приросту його виробничого споживання дорівнює нулю. Звідси, отже, можна зробити висновок, що збільшення запасу цього ресурсу не призведе до збільшення прибутку. Якщо деяке обмеження в двоїстій задачі виконується як нерівність, відволікання ресурсів для виробництва відповідного йому товару призведе до зниження за рахунок прибутку від виробництва всіх інших товарів на величину більшу, ніж одержуваного при цьому прибутку, від додаткового цього товару. Отже, цей товар додатково виробляти не вигідно.

Задачі нелінійного програмування характеризуються нелінійністю цільової функції, або функцій, що входять в обмеження, чи одночасно обох цих типів функцій. Необхідність застосування цих моделей часто виникає в ІУС і економічних дослідженнях. Розглянутий приклад аналізу виробництва за допомогою лінійного програмування був заснований на наступних двох вельми жорстких вимогах. Величина прибутку, що отримується у результаті виробництва додаткової одиниці товару, вважається постійною і, отже, незалежною від обсягу виробництва. Крім того, величини додаткових витрат ресурсів на виробництво кожної додаткової одиниці товару також вважаються постійними, незалежними від обсягу виробництва. Отже, розглянута модель лінійного програмування не враховує ані можливість зниження середнього прибутку внаслідок зниження попиту, ані можливість збільшення додаткових витрат ресурсів, яке необхідно для збереження постійного приросту продукції, внаслідок відомого явища зниження економічної ефективності при збільшенні масштабів господарської діяльності на незмінній науково-технічній базі. Нелінійне програмування й призначене для таких випадків, коли необхідна модифікація моделей для більш реального відображення економічних явищ.

У 1951 році Г. Кун і Л. Таккер виявили й обґрунтували умови (названі їхніми іменами), яким повинне задовольняти оптимальне розв'язання задачі нелінійного програмування. Ці умови узагальнюють умови двоїстої задачі лінійного програмування.

Динамічне програмування було розроблене головним чином Р. Беллманом у 1957 р. Воно включає сукупність важливих методів, які можуть бути застосовані до ситуацій, пов'язаних з послідовним прийняттям рішень щодо переведення економічної системи

з одного стану в інший. З математичної точки зору ці методи засновані на принципі оптимальності Беллмана, відповідно до якого оптимальна послідовність рішень характеризується тим, що, починаючи з будь-якого даного в ній рішення, усі наступні рішення повинні бути оптимальними стосовно стану, що виник у результаті прийняття цього рішення, яке б воно не було. Багато проблем, пов'язаних з дослідженням операцій, можуть бути сформульовані як динамічні програми (моделі). Ці моделі схожі з моделями математичного програмування, оскільки вони також характеризуються максимізацією (або мінімізацією) деякої функції при обмеженнях типу нерівності невід'ємних змінних.

Метод розпізнавання образів (1954) — різновид методу типологічного групування одночасно за двома і більше ознаками, здійснюваного, як правило, за допомогою ЕОМ: застосовується при наявності великих масивів вихідної первинної інформації, що характеризує значну кількість одиниць спостереження. Образ — це тип, до якого належить кожна одиниця. Тип визначається за сукупністю ознак; чим більше таких ознак, тим точніше встановлюється образ. У вибіркових спостереженнях образ визначається шляхом ідентифікації даних, якими оперують (у межах даних генеральної сукупності). Збіжність даних є гарантія правильного визначення образу. Точність встановлення образу вимірюється помилками вибірки (середньої і частки): мала помилка означає високу точність; велика — низьку точність. Вибіркові оцінки обмежуються, як правило, одновимірними вимірюваннями, метод розпізнавання образів, як самостійний метод, — багатовимірними.

Задачею стохастичного програмування називається задача, в якій ряд параметрів, що входять у цільову функцію й обмеження, являють собою випадкові величини. Такі задачі виникають у багатьох практичних ситуаціях, наприклад, коли потрібно спланувати виробництво при невизначеному попиті на продукцію.

У найпростішому випадку, коли в задачі стохастичного програмування відсутні обмеження, вона може бути розв'язана методом стохастичної апроксимації, запропонованим у 1951 р. американськими статистиками Роббінсом і Монро. У 1952 р. Кіфер і Вольфовіц модифікували процедуру стохастичної апроксимації Роббінса — Монро для визначення значення аргументу, за якого невідома функція регресії досягає свого максимального значення. При цьому передбачається, що вимірюванню доступні лише

значення самої цієї функції, що спотворюються деякою випадковою помилкою. Метод стохастичного програмування може бути застосований, наприклад, при моделюванні міжгалузевого балансу, коли доцільно припустити, що матриця коефіцієнтів прямих витрат внаслідок помилок вимірювання є випадковою. При цьому фактичний випуск кінцевої продукції також є випадковим і коливається щодо її планового випуску, тому для забезпечення більшої пропорційності у виробництві валового суспільного продукту за галузями народного господарства як критерій оптимальності може бути прийнята мінімізація математичного сподівання суми квадратів відхилень фактичного випуску кінцевої продукції від її планового випуску.

Підходи до розв'язання багатьох статистичних задач засновані на розробці параметричної моделі з визначеним типом розподілу — такий, наприклад, як лінійна модель з нормально розподіленими помилками. Класичний підхід до отримання рішення за даних умов полягає в оцінці параметрів цієї моделі деяким оптимальним образом. Однак цей підхід може бути пов'язаний з труднощами, оскільки в більшості випадків модель являє собою лише деяку апроксимацію реальності й у точності строго її не відображує. Для того щоб стверджувати, що оптимальна оцінка для ідеалізованої моделі близька до оптимальної і для більш точної моделі, необхідно, щоб ця оцінка мала властивість безперервності при зміні умов моделі, таким чином, щоб невеликі зміни в умовах моделі призводили лише до незначних змін оптимальної оцінки.

Однак багато відомих класичних процедур отримання оптимальних оцінок не забезпечують їхньої безперервності є чутливими до всякого роду змінам у вихідних припущеннях, у тому числі і до незначних, і тим самим виступають як не сталі. Робасна оцінка має властивість нечутливості і тим самим при незначній зміні умов і припущень, що лежать в основі ідеалізованої моделі, не реагує на поводження модифікованої моделі, тобто вказує на її сталість. Область цього не реагування називається робасністю. Аналогічна ситуація, наприклад, виникає, коли переходять від ідеалізованої лінійної моделі до часто більш реалістичній нелінійній моделі. Для цього до лінійної моделі додають невелике нелінійне відхилення, аналізують його вплив на раніше отримане рішення й у разі потреби одержують нові, більш сталі рішення.

Використовувані при моделюванні багатьох статистичних задач рівняння в більшості випадків включають нормально роз-

поділені помилки, аналогічні лінійним вираженням у рівняннях ідеалізованих моделей фактично нелінійних залежностей. При цьому перехід від нормального розподілу до розподілу, близькому до нормального, може бути розглянутий як аналог додавання до лінійного виразу невеликого нелінійного відхилення.

Відхилення від вихідних припущень моделі можуть виникнути внаслідок грубих помилок, що виникають через невірні вимірювання, при запису або при переписуванні даних або через інші невідомі причини. Округлення й групування даних може також призвести до відхилень від умов, що визначають модель.

До сьогодні робасні методи отримали розвиток головним чином у додатку до забезпечення сталості при незначних відхиленнях від припущень про тип розподілу випадкових величин. Для цього випадку мають добре розвинені теорія як для отримання оцінок параметрів розподілу, так і для перевірки статистичних гіпотез. Однак значною мірою ще не розроблена теорія для робасних процедур обробки статистичних даних, що забезпечують сталість одержуваних за їх допомогою рішень при відхиленнях від припущень про незалежність випадкових величин.

Найбільш складний випадок робасності — визначення самого типу розподілу тоді, коли іншими шляхами зробити це неможливо, наприклад при надмалих вибірках, що складаються з двох-трьох точок.

Кластерний аналіз (кінець 1950-х рр.). Розвиток методів кластерного аналізу пов'язаний з появою потужних комп'ютерів. Техніка кластерного аналізу перебуває на початковій стадії розвитку. Дані, використовувані для кластерного аналізу, характеризують ряд властивостей об'єктів деякої їхньої сукупності. Його ціль полягає у групуванні цих об'єктів за деяким невеликим числом класів, названих кластерами.

Дані для кластерного аналізу можуть бути представлені у вигляді прямокутного масиву, рядки якого відповідають аналізованим об'єктам, а стовпчики — досліджуваним змінним. Зазвичай ці змінні двоїсті і визначають лише наявність або відсутність деяких властивостей. Для кожного об'єкта може бути визначена відстань (у геометричному сенсі) від комбінації значень цих змінних, яка характеризує його, до «нульової» точки, яка відповідає нульовим значенням змінних, а також відстань до будь-якого іншого об'єкта. *Кластерами* називаються групи об'єктів, близькі одна до одної у геометричному сенсі. Наприклад,

кластерний аналіз дозволяє виділити з деякої множини підприємств ті, які характеризуються специфічними умовами виробництва, або підрозділити їх за результатами виробництва на передові, середні і відстаючі.

Кластерний аналіз зазвичай може бути поділений на такі етапи:

- ☑ вибір об'єктів, що характеризують їх змінні для включення в статистичну сукупність;
- ☑ визначення мір близькості між парами об'єктів;
- ☑ групування об'єктів за кластерами;
- ☑ інтерпретація й аналіз результатів групування.

Мартингальні методи аналізу випадкових соціально-економічних процесів (60–80-ті роки XX в.). Вони засновані на моделюванні приросту (збільшення) випадкових процесів, умовні математичні сподівання яких $X(t_s) - X(t_i)$ байдужі до минулого та за заданих значень $X(s)$ дорівнюють нулю. При цьому одиницею спостереження є не окремі одинарні випадкові величини, а їх множини, що інтерпретуються в категоріях приросту як власне мартингали. Звідси і назва — теорія мартингального аналізу випадкових масових процесів.

Найбільш розповсюджена галузь застосування розглянутих методів — процеси масового обслуговування, у тому числі (в останнє десятиліття) окремі соціально-економічні процеси (попит та пропозиція, потоки транспорту, контроль якості, суспільна думка й ін.).

Мартингальні методи використовуються для аналізу процесів, що безупинно змінюються в часі. Однак вони застосовні й для аналізу випадкових дискретних процесів, математичні сподівання приросту яких, як правило, невід'ємні (субмартингали). Можливість застосування мартингальних методів для дослідження соціально-економічних процесів пов'язана з поширенням у статистичній науці визнання того факту, що фактори, що керують цих процесів, можуть бути розглянуті як випадкові. При цьому під випадковими розуміються такі фактори, вплив окремо узятих одиниць яких на загальний стан процесу в силу їх численності не має передбачуваного характеру. Тому, хоча дія кожної з цих одиниць у кінцевому рахунку є визначеною, немає можливості враховувати в моделі в явному вигляді ізольовано кожну з них, а доцільно розглядати їх сукупну дію, що і виступає як випадкова. Ця властивість спільної дії великого числа

детермінованих процесів використовується при комп'ютерній генерації псевдовипадкових чисел на основі складних ітераційних детермінованих алгоритмів. Як відомо, справедлива і зворотна закономірність: сумарна, незалежна одна від одної дія великого числа стохастичних процесів набуває більш-менш детермінованого характеру.

У ІУС, де мають місце явища з кількісно невимірюваними станами, отримали поширення марковські ланцюжки (процеси), що представляють особливий різновид випадкових масових процесів. Процес називається *марковським*, якщо умовна імовірність будь-якого можливого значення його стану в більш пізній момент часу щодо поточного і попередніх станів дорівнює умовній імовірності цього значення відносно тільки поточного стану і не залежна від усіх попередніх. Звідси поняття випадкових процесів наслідків. Взагалі, соціальні процеси не задовольняють у точності вимозі, висуваній до марковського процесу. Тому цей тип процесу може бути розглянутий лише як деяке наближення (апроксимація) для реальних процесів.

Марковські моделі отримали поширення в багатьох галузях суспільних наук. Застосування цих моделей для аналізу процесів професійної мобільності і планування трудових ресурсів розглянуто, наприклад, у роботах Д. Бартолом'ю. Марковські моделі використовуються також для планування системи освіти, у демографії, при вивченні міграції населення, землекористування і розміщення промислових підприємств, а також для різних систем масового обслуговування, зокрема виробничих систем, систем медичного і побутового обслуговування населення, судових систем і систем бухгалтерського обліку.

У практиці аналітичних статистичних робіт стохастичні методи застосовуються у фізиці, агрономії, біології, медицині. Однак немає сумніву, що в ІУС і економіки (контроль за якістю продукції і технології виробництва, держприймання, вивчення громадської думки тощо) вони мають ще більш значні перспективи застосування.

Контрольні завдання і вправи

1. Оцінити граничну величину помилки вибірки для даних, наведених у додатку 1, табл. 1.1.
2. Для задачі 1 оцінити кількість даних, необхідних для отримання вибірки з погрешністю 0,05.

3. Оцінити одномірну щільність розподілу для даних, приведених у додатку 2, табл. 2.1.

4. Визначити основні статистичні характеристики ряду для задачі 1.

5. Оцінити з надійністю $\gamma = 0,95$ статистичні характеристики для даних, наведених у табл. 2.1.

6. За даними, наведеними у табл. 2.1, знайти довірчу область для векторів математичного сподівання, дисперсії.

7. Для даних табл. 1.1 визначити довірчу область для вектора математичного сподівання параметрів з надійністю 0,95.

8. Визначити довірчу область для середньоквадратичного відхилення, використовуючи дані, наведені в табл. 1.1.

9. Побудувати графіки за даними табл. 1.2.

РОЗДІЛ 3

КОРЕЛЯЦІЙНИЙ АНАЛІЗ

Цей розділ присвячений методам кореляційного аналізу, які дозволяють дослідити та оцінити щільність зв'язку між двома або більше безперервними змінними. Розглядаються ступені залежності однієї змінної від іншої або декількох змінних. Ступені статистичного зв'язку досліджуються шляхом обчислення коефіцієнтів парної та множинної кореляції, часткової кореляції, авто- та взаємно кореляційних функцій. Під час вивчення не кількісних ознак або безперервних кількісних даних з невідомими законами розподілу використовується рангова кореляція.

Для оцінки статистичного зв'язку між двома та більше часовими послідовностями в роботі [28] пропонується використовувати природи змінних. Зауважимо, що методи кореляційного та регресійного аналізів тісно переплітаються між собою. Тому основною задачею кореляційного аналізу є дослідження щільності кореляційного зв'язку між випадковими змінними процесу та отримання коефіцієнта рівняння регресії через парні коефіцієнти кореляції. Задовільні результати при застосуванні кореляційного аналізу можливо отримати за виконання таких передумов:

☑ процеси X та Y , що досліджуються, мають бути випадковими. У цьому випадку можливо навести статистичні висновки стосовно коефіцієнтів кореляції між X та Y ;

☑ вхідні та вихідні випадкові змінні мають підкорятися багатовимірному нормальному закону розподілу ймовірності. Ця передумова передбачає наявність лінійного статистичного

зв'язку та інтерпретацію коефіцієнта кореляції у ймовірнісному розумінні;

☑ вхідні випадкові змінні мають бути функціонально не пов'язані між собою;

☑ часткові виміри з кожної з вхідних змінних мають бути статистично незалежними;

☑ усі вхідні змінні мають вимірюватися зі зневажливо малою помилкою.

Процес має бути керованим, тобто істотні зміни вхідних змінних мають викликати істотні зміни цільових функцій.

Кореляційний аналіз застосовується лише для отримання лінійних математичних моделей.

3.1. Коефіцієнти, що характеризують ступінь статистичного зв'язку

3.1.1. Коефіцієнт кореляції

Формула для коефіцієнта кореляції включає двовимірні дані, починаючи з першого вимірювання (X_1, Y_1) та закінчуючи останнім вимірюванням (X_n, Y_n) . Наприклад, X_1 може бути обсягом збуту компанії IBM, а Y_1 — її чистим прибутком. Розглядаючи кожен стовпчик чисел окремо, можна, наприклад, вираховувати звичайне стандартне відхилення вибірки для значень X , тобто знайти S_x ; аналогічно S_y являє стандартне відхилення лише для значень Y . Формула для обчислення коефіцієнта кореляції також включає суму попарних добутків відхилень X та Y від середнього, поділену на $n - 1$.

Формула коефіцієнта кореляції має вигляд:

$$r_{xy} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2 \right]^{\frac{1}{2}}}, \quad i = \overline{1, n}.$$

Доданки у чисельнику виражають взаємодію двох змінних та визначають знак кореляції (додатний чи від'ємний). Якщо, наприклад, між змінними існує сильний додатний взаємозв'язок (збільшення однієї змінної пов'язано зі збільшенням другої) кое-

фіцієнт кореляції буде додатним числом. У цьому випадку відхилення від середнього величин X та Y матимуть однаковий знак. Аналогічно, якщо між змінними існує від'ємний взаємозв'язок (збільшення однієї змінної пов'язано зі зменшенням другої), усі доданки у чисельнику дають від'ємне значення кореляції.

Знаменник просто нормує чисельник таким чином, що коефіцієнт кореляції виявляється легко інтерпретованим числом (тобто таким, що не має розмірності) в діапазоні від -1 до 1 . Оскільки у чисельнику виразу присутній добуток двох змінних, має сенс перетворити його на «чисте» число, розділивши на добуток співмножників, що включають ці змінні. Якщо не виконати таке ділення, буде важко інтерпретувати фізичний сенс статистичного зв'язку. Якщо б, наприклад, X та Y вимірювались у доларах, тоді чисельник мав би таку незвичайну розмірність, як «долари у квадратах».

Чисельник у виразі для коефіцієнта кореляції являє собою коваріацію X на Y . Незважаючи на те, що іноді він використовується як самостійна характеристика (наприклад, в теорії фінансів для опису сумісної зміни курсів акцій на двох біржах), зручніше користуватися коефіцієнтом кореляції. Кореляція та коваріація являють, власне, одну й ту саму інформацію (за умови, якщо також відомі часткові стандартні відхилення), однак кореляція представляє цю інформацію у більш зручній формі.

Формула для розрахунку коефіцієнта кореляції *симетрична* відносно X та Y . Кореляція X з Y — це те саме, що й кореляція Y з X , тобто не має значення, яка з двох змінних буде вказана першою. Це ствердження справедливе для кореляції, але не для регресії. Коефіцієнт кореляції показує ступінь статистичного лінійного зв'язку. Якщо дві випадкові величини пов'язані детермінованим зв'язком, то коефіцієнт кореляції може дорівнювати нулю.

Залежні випадкові величини у загальному випадку можуть бути корельовані або не корельовані. Щоб показати, що залежні випадкові величини можуть бути некорельовані, розглянемо найпростіший приклад. Припустимо X — випадкова перервна величина з трьома можливими значеннями — $1, 0, 1$, кожного з яких вона набуває з імовірністю $1/3$. Очевидно, що випадкова величина $Y = X^2$ залежить від випадкової величини X , оскільки її значення повністю визначається, коли стає відомим значення випадкової величини X . Таким чином, випадкові величини X та

$Y = X^2$ завжди залежні. Покажемо, що у цьому прикладі їх кореляційний момент дорівнює нулю, тобто що вони не корельовані. У цьому випадку $n = 3$.

Можливі значення випадкової величини $Y = X^2$ дорівнюють квадратам відповідних можливих значень випадкової величини X :

$$x_1 = -a, \quad x_2 = 0, \quad x_3 = a, \quad p_1 = p_2 = p_3 = \frac{1}{3}$$

та аналогічно Y :

$$y_1 = x_1^2 = a^2, \quad y_2 = x_2^2 = 0, \quad y_3 = x_3^2 = a^2.$$

Знаходимо математичні сподівання випадкових величин X та Y :

$$E_x = x_1 p_1 + x_2 p_2 + x_3 p_3 = -a \frac{1}{3} + 0 \frac{1}{3} + a \frac{1}{3} = 0$$

$$\text{та } E_y = y_1 p_1 + y_2 p_2 + y_3 p_3 = a^2 \frac{1}{3} + 0 \frac{1}{3} + a^2 \frac{1}{3} = \frac{2}{3} a^2.$$

Тепер обчислимо кореляційний момент випадкових величин X та Y :

$$\begin{aligned} k_{xy} &= (x_1 - E_x)(y_1 - E_y)p_1 + (x_2 - E_x)(y_2 - E_y)p_2 + (x_3 - E_x)(y_3 - E_y)p_3 = \\ &= -a \left(a^2 - \frac{2}{3} a^2 \right) \frac{1}{3} + 0 \left(0 - \frac{2}{3} a^2 \right) \frac{1}{3} + a \left(a^2 - \frac{2}{3} a^2 \right) \frac{1}{3} = 0. \end{aligned}$$

Таким чином, залежні випадкові величини X та Y некорельовані.

З іншого боку, в умовах наведеного прикладу залежні випадкові величини X та $Z = X^3$ корельовані, тобто що їх кореляційний момент k_{xz} не дорівнює нулю. Ми бачимо, що кореляційний момент випадкових величин може деякою мірою слугувати характеристикою залежності між ними. Якщо випадкові величини незалежні, то вони й некорельовані, їх кореляційний момент дорівнює нулю. Якщо випадкові величини корельовані, тобто їхній кореляційний момент не дорівнює нулю, то вони залежні, однак у загальному випадку кореляційний момент не є повною характеристикою залежності між випадковими величинами. Дійсно, як

впливає з прикладу, кореляційний момент може дорівнювати нулю, коли випадкові величини залежні.

В одиничному випадку нормально розподілених випадкових величин кореляційний момент є повною характеристикою залежності між випадковими величинами. Нормально розподілені випадкові величини залежні тоді і тільки тоді, коли вони корельовані. Нормально розподілені випадкові величини некорельовані тоді і тільки тоді, коли вони незалежні. Інтерпретація коефіцієнта кореляції в залежності від його величини представлена в табл. 3.1.

Таблиця 3.1

Інтерпретація коефіцієнта кореляції

Кореляція	Загальноприйнята інтерпретація	Деякі інші можливості
1	Ідеальний додатний взаємозв'язок. Усі точки даних мають розташовуватись строго на прямій лінії, що напрямлена вгору та праворуч	Відсутні
Близько до 1	Сильний додатний взаємозв'язок. Точки даних доволі щільно згруповані (з невеликим випадковим розкидом) біля прямої лінії, яка напрямлена вгору та праворуч	Точки даних розташовуються строго на кривій, що напрямлена вгору (нелінійна структура). Між точками даних взаємозв'язку, в основному, немає, але один викид даних (точка, що різко відхиляється) викривив кореляцію.
Близько до 0, але зі знаком «+»	Незначний додатний взаємозв'язок. Точки даних утворюють випадкову хмару з незначною орієнтацією вгору та праворуч	Кореляція викривлена наявністю у даних груп об'єктів, що по-різному взаємопов'язані між собою
0	Відсутність взаємозв'язку, цілком випадкова хмара, що не має орієнтації ані вгору, ані вниз при русі праворуч	Точки даних розташовуються строго на кривій, що має нахил вгору з одного боку та вниз — з іншого. Точки даних розташовуються строго на прямій лінії, але один викид даних (точка, що відхиляється) викривив кореляцію. Кореляція викривлена наявністю в даних груп об'єктів, що по-різному взаємопов'язані

Кореляція	Загальноприйнята інтерпретація	Деякі інші можливості
Близько до 0, але зі знаком «-»	Незначний від'ємний взаємозв'язок. Точки даних утворюють випадкову хмару з зовсім незначною орієнтацією вниз та праворуч	Точки даних розташовуються строго на кривій, що напрямлена вниз (нелінійна структура). Точки даних у цілому не утворюють будь-якої структури, але один викид даних (точка, що різко відхиляється) викривив кореляцію. Кореляція викривлена наявністю в даних груп об'єктів, що по-різному взаємопов'язані між собою
Близько до -1	Сильний від'ємний взаємозв'язок. Точки даних щільно згруповані (з невеликим випадковим розкидом) біля прямої лінії, що напрямлена вниз та праворуч	
-1	Ідеальний від'ємний взаємозв'язок. Усі точки даних мають розташовуватись строго на прямій лінії, що напрямлена вниз та праворуч	Відсутні
Не визначена	Точки даних розташовуються строго на горизонтальній або на вертикальній лінії	Недостатньо даних (менше $n = 2$ різних пар значень X та Y)

3.1.2. Множинний та частковий коефіцієнти кореляції

Множинний коефіцієнт кореляції (МКК) є мірою ступеня лінійного зв'язку між залежною змінною Y та незалежними змінними $x = \{x_{jl}\}$, $j, l = \overline{1, p}$, де p — кількість незалежних змінних. Геометрична інтерпретація МКК — це косинус кута між векторами Y_i та $\hat{Y}_i$, де $i = \overline{1, n}$, де n — кількість значень y_i та $x_{i,j}$, що виміряні. З вузької точки зору це проста кореляція між значеннями Y_i , що виміряні, та значеннями $\hat{Y}_i$, що передбачені з рівняння регресії.

Квадрат МКК, або коефіцієнт детермінації (КД), можна підрахувати за формулою

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

Коли $\sum_{i=1}^n (y_i - \bar{y})^2 = 0$, то $R^2 = 1$. Тому $0 \leq R^2 \leq 1$.

Квадрат МКК або КД можна обчислити за дисперсіями багатовимірного нормального розподілу $Y, x_1, \dots, x_p$.

$$R_{y, x_1, \dots, x_p}^2 = \frac{(\sigma_y^2 - \sigma)^2}{\sigma_y^2},$$

де σ^2 — дисперсія умовного розподілу. Величина $R = \sqrt{R^2}$ є чисельним коефіцієнтом кореляції.

Статистичний критерій з оцінки значущості коефіцієнта детермінації визначається за F -критерієм.

$$F_{\text{розр}} = \frac{R^2}{1 - R^2} \cdot \frac{n - p}{n - 1}.$$

Якщо $F_{\text{розр}} > F_m(\alpha, p, n - p - 1)$, то можна вважати, що коефіцієнт детермінації значущий, тобто незалежні змінні достатньо добре пояснюють стохастичну залежність $y = f(x, e)$. Отже, квадрат МКК дорівнює частці дисперсії Y , що «пояснена» регресійною залежністю $y = f(x, e)$.

Величина $(1 - R_{y, x_1, \dots, x_p}^2)^{1/2}$ є часткою відхилення Y , яка залишилася «непоясненою» залежністю $y = f(x, e)$. Наприклад, нехай $R = 0,9$, залишається 44 % непоясненого стандартного відхилення Y .

У багатьох пакетах прикладних програм (ППП) МКК обчислюється за таблицею дисперсійного аналізу $R_{y, x_1, \dots, x_p} = \sqrt{\frac{SSD}{SST}}$, де

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2, \quad v_T = n - 1;$$

$$SSD = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2, \quad v_D = 1;$$

$$SSR = \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad v_R = n - 2.$$

Ця величина може бути обчислена і як додатний квадратний корінь із коефіцієнтів детермінації R^2 , МКК використовується як міра лінійної залежності Y від X . Чим ближче $R_{y, x_1, \dots, x_p}$ до 1, тим сильніша залежність.

Друге застосування оцінки МКК — квадрат цього коефіцієнта оцінює частку дисперсії Y , яка пояснена лінійною регресією Y по $x_1, \dots, x_p$.

Визначимо чисельний коефіцієнт кореляції R між однією величиною Y та двома іншими x_1, x_2 за формулою

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

$$\text{Отримуємо } R \geq |r_{yx_1}|, \quad R \geq |r_{yx_2}|, \quad R \geq |r_{yx_1/x_2}|, \quad R \geq |r_{yx_2/x_1}|.$$

Звідси випливає, що R може тільки збільшитися, якщо в модель включати додаткові змінні x_j ($j=1, p$), та не збільшитися, якщо з ознак провадити тільки виключення.

Квадрат МКК можна визначити як частку остаточної дисперсії Y , яка «пояснена» додаванням змінної x_k до набору $\{x_1, \dots, x_{k-1}\}$. Нехай l, h, d — три різні змінні з множини $\{y, x_1, \dots, x_p\}$, то МКК першого порядку визначається виразом

$$\rho_{lh,d} = \frac{r_{lh} - r_{ld} \cdot r_{hd}}{\sqrt{(1 - r_{ld}^2)} \sqrt{(1 - r_{hd}^2)}},$$

де r_{lh}, r_{ld}, r_{hd} — парні коефіцієнти кореляції.

Послідовно застосовуючи рекурентну формулу

$$\rho_{lh,cd} = \frac{\rho_{lh,c} - \rho_{ld,c} \rho_{hd,c}}{\sqrt{(1 - \rho_{ld,c}^2)} \sqrt{(1 - \rho_{hd,c}^2)}},$$

де C — будь-яка підмножина змінних, що залишилися, можна отримати часткові коефіцієнти кореляції будь-якого порядку.

Частковий коефіцієнт кореляції є показником наявності лінійного статистичного зв'язку між двома змінними (l, h) без урахування впливу решти випадкових змінних (c, d).

Порівняння часткового коефіцієнта детермінації та відповідного парного коефіцієнта детермінації дозволяє визначити наявність інших змінних, які впливають на цей зв'язок.

Так, якщо частковий коефіцієнт детермінації менший, ніж відповідний парний коефіцієнт детермінації, то взаємозалежність між двома величинами зумовлена впливом випадкових змінних, що фіксуються. З іншого боку, якщо частковий коефіцієнт де-

термінації більший за відповідний парний, то величини, що фіксуються, послаблюють цей зв'язок.

Класичний підхід оцінки статистичного лінійного зв'язку передбачає наявність вихідних даних, які отримані у статичному режимі. Для розв'язання багатьох функціональних задач в ІУС цього недостатньо. Інтерес становлять наявність та використання показників, що характеризують динаміку складних багатозв'язкових систем.

3.1.3. Рангова кореляція

Під час вивчення не кількісних ознак або кількісних ознак з безперервними та невідомими законами розподілу класичні підходи кореляційного аналізу або не застосовні, або не ефективні [8]. Для вивчення щільності зв'язку в цих випадках використовують методи непараметричної статистики, серед яких найбільш поширені методи рангової кореляції.

Розглянемо варіаційний ряд для ознаки x :

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(j)}, \dots, x_{(n)}.$$

Нагадаємо, що *рангом спостережуваного значення x_j ознаки x* називається номер цього спостереження у варіаційному ряді (тобто $R(x_j) = j$) за умови, що нерівності — суворі. Якщо ж у варіаційному ряді зустрічаються однакові члени, то в якості однакових (зв'язних) рангів береться середня арифметична відповідних номерів.

Наприклад, суми балів, що їх набрали студенти за виконання двох контрольних завдань, були такими: 5, 10, 8, 7, 9, 10, 5, 5. Варіаційний ряд має вигляд: 5, 5, 5, 7, 8, 9, 10, 10.

Ранг трьох студентів, які потрапили у початок ряду, дорівнює $(1 + 2 + 3)/3 = 2$ або $R(5) = 2$. Далі $R(7) = 4$, $R(8) = 5$, $R(9) = 8$, $R(10) = (9 + 10)/2 = 9,5$. Очевидно, що ранги можуть не бути натуральними числами.

Для вимірювання зв'язків між ознаками, значення яких можна впорядкувати за ступенем прояву ними властивостей, що аналізуються, застосовують *коефіцієнти рангової кореляції*. Вивчимо спочатку парний зв'язок, тобто зв'язок між двома ознаками x та y . Припустимо, є вибірка обсягу n з безперервно розподіленої двовимірної генеральної сукупності (xy) : (x_1, y_1) ... (x_n, y_n) .

З безперервності розподілу виходить, що множина виборок з деякими однаковими значеннями за x або за y має міру нуль. Отже, можна вважати, що у ранжуванні значень x та y зв'язні ранги відсутні. Під час вивчення зв'язку між x та y можна припускати, що вибірка впорядкована за x , тоді такий вибірки відповідає така матриця (підстановка):

1	2	...	i	...	n
$R(1)$	$R(2)$	...	$R(i)$		$R(n)$

В ній перший рядок складається з рангів спостереження x , а другий — з рангів y .

Замість вивчення зв'язку між x та y за допомогою кількісних значень x та y досліджуватимемо цей зв'язок, використовуючи відповідні ранги. Очевидно, що жорсткому (функціональному) додатному зв'язку між x та y відповідає підстановка:

1	2	...	i	...	n
1	2	...	i	...	n

жорсткому від'ємному зв'язку — підстановка:

1	2	...	n
n	$n - 2$	...	1

Решта $(n - 2)$ підстановок отримуються при тому чи іншому ступені зв'язку, власним стохастичним (статистичним) зв'язком.

Для вимірювання ступеня подібності між перестановками $(1, 2, \dots, n)$ та $(R(1), R(2), \dots, R(n))$ ступень зв'язку між x та y назовемо *інверсією* (безладом) між елементами перестановки $R(i)$ та $R(j)$ (другого рядку підстановки), якщо $R(i)$ стоїть лівіше $R(j)$ та більше за $R(j)$. Якщо ж за тієї ж умови $R(i)$ менше $R(j)$, то кажуть, що елементи $R(i)$ та $R(j)$ інверсії не утворюють або утворюють порядок.

Як міру зв'язку беруть різницю між сумами чисел порядків N та чисел Q , що утворені елементами другого рядку підстановки.

Керуючись комбінаторними підрахунками, можна визначити ймовірності отримання перестановок з заданою мірою зв'язку. Для підстановок з чотирьох елементів розглянемо розрахункову таблицю [8]:

Таблиця 3.2

Число порядків N	Число інверсій Q	Міра подібності S_k	Підстановки	Ймовір- ність
0	6	-6	4321	1/24
1	5	-4	3421, 4231, 4322	3/24
2	4	-2	3412, 4132, 4213, 2431, 3241	5/24
3	3	0	3214, 2413, 4123, 3142, 1432, 2341	6/24
4	2	2	2143, 1423, 2314, 3124, 1342	5/24
5	1	4	2134, 1324, 1243	3/24
6	0	6	1234	1/24

Неважко помітити (та довести для загального випадку), що сума чисел порядків N та інверсій Q дорівнює сумі номерів перестановки, тобто $1 + 2 + \dots + n = (n + 1)/2$; розподіл ймовірностей симетричний відносно центра S_k , що дорівнює нулю; якщо сума номерів парна, то всі можливі значення міри S_k парні (якщо ж $n(n + 2)/2$ — непарне число, то всі S_k — непарні числа). Із зазначеного випливає, що таблиці для розв'язання задач перевірки гіпотез відносно міри подібності (або зв'язку) можна давати для невід'ємних значень S_k (парних або непарних).

Коефіцієнт рангової кореляції Кендалла визначається шляхом нормування випадкової величини S_k , іншими словами, шляхом ділення на $n(n - 1)/2$:

$$r_k = \frac{2S_k}{n(n-1)} = 1 - \frac{4Q}{n(n-1)} = \frac{4N}{n(n-1)} - 1.$$

Він змінюється у межах від -1 до $+1$, які відповідають найжорсткішим значенням міри подібності — від'ємного та додатного зв'язку між x та y . Нуль відповідає відсутності зв'язку між ознаками у генеральній сукупності. Як звичайний парний коефіцієнт кореляції $r_{xy}(\rho_{xy})$, коефіцієнт кореляції Кендалла не є абсолютним вимірником зв'язку (якщо він дорівнює нулю, то x та y можуть виявитися залежними).

За великих обсягів n вибірки і незалежності x та y можна використовувати нормальний закон розподілу r_k з параметрами $Mr_k = 0$ та

$$Dr_k = \frac{2(2n+5)}{9n(n-1)}.$$

Інший коефіцієнт рангової кореляції, запропонований Спірменом, використовує міру подібності з урахуванням ваги рангів:

$$r_c = 1 - \frac{6S_c}{n^3 - n}, \quad S_c = \sum_{i=1}^n (R(i) - i)^2.$$

Цей коефіцієнт можна отримати за формулою парного коефіцієнта кореляції r_{xy} для вибірки:

$$(i, R(i)), \quad i = \overline{1, n}.$$

За великих обсягів вибірки n і незалежності x та y коефіцієнт рангової кореляції Спірмена підпорядковується нормальному закону розподілу з параметрами $Mr_c = 0$ та $Dr_c = 1/(1 - n)$.

У теорії доводиться, що випадкові величини r_c приблизно в півтора рази більші за r_k при більших (великих) n , якщо тільки їх квадрати не надто близькі до одиниці. Під час вивчення зв'язків між кількістю ознак, що вимірюються за порядковою шкалою, кількість яких більша за два, застосовують міру подібності відповідної кількості ранжувань (перестановок). Міра є сумою квадратів відхилень сум рангів спостережень (об'єктів) від їх спільного середнього рангу

$$S_w = \sum_{i=1}^n (S_i - \bar{S})^2 = \sum_{i=1}^n S_i^2 - \frac{\left(\sum_{i=1}^n S_i\right)^2}{n}, \quad \bar{S} = \frac{\sum_{i=1}^n S_i}{n}, \quad S_i = \sum_{j=1}^K R_{ij},$$

де R_{ij} — ранг i -го спостереження за j -ою ознакою, $i = \overline{1, n}$, $j = \overline{1, k}$.

Показником узгодженості визначається коефіцієнт конкордації Кендалла, що обчислюється за формулою

$$W = \frac{12S_w}{k^2(n^3 - n)}.$$

Цей коефіцієнт може змінюватись від нуля (абсолютна неузгодженість) до одиниці (повний збіг усіх ранжувань), що легко перевіряється.

Доведено, що за відсутності зв'язку між k ознаками за більших значень n ($n > 7$) випадкові величини $k(n-1) \cdot W = \frac{12S_w}{kn(n+1)}$ наближено розподілені як χ^2 з числом степенів вільності $v = n - 1$.

За $k = 2$ виходить наступне співвідношення між W та r_c : $W = 0,5(r_c + 1)$.

За малих значень n та k існують таблиці розподілу S_k , S_c та S_w , придатні для перевірки гіпотези незалежності ознак. Наведемо фрагменти цих таблиць за умови $n = 10$, $\alpha_1 = P(|S| \geq S_1)$ (табл. 3.3) та за умови $n = 5$, $k = 3$, $\alpha = P(|S| \geq S_w)$ (табл. 3.4).

Таблиця 3.3

S_k	17	19	21	23	25	27	29	31
α_k	0,156	0,108	0,072	0,046	0,028	0,0166	0,0092	0,0046
S_c	248	258	268	278	288	298	308	
α_k	0,144	0,096	0,060	0,034	0,0174	0,0072	0,0022	

Таблиця 3.4

S_w	50	56	60	62	66	74	78	82	86
α_w	0,163	0,096	0,063	0,056	0,038	0,015	0,0053	0,0028	0,0006

Приклад 3.1

На основі даних про стаж роботи x десяти програмувальників та часу розв'язання кожним із них певної контрольної задачі обчислити коефіцієнти рангової кореляції Кендалла та Спірмена і перевірити значущість цих коефіцієнтів, якщо $\alpha = 0,01$.

Таблиця 3.5

№ з/п	Стаж x , міс.	Час розв'язання y , хв
1	32	12
2	15	24
3	16	23
4	18	21
5	20	20
6	28	9
7	21	11
8	29	10
9	23	15
10	17	16

Розв'язання

Розташувавши пари спостережень (x, y) за порядком зростання x , отримуємо таку підстановку:

1	2	3	4	5	6	7	8	9	10
10	9	6	8	7	3	5	1	2	4

Обчислимо міру подібності S_k , приписуючи до числа інверсій, що утворюються елементом другої перестановки, знак мінус:

-1	-1	-1	-1	-1	-1	-1	-1	-1	=	-9
	-1	-1	-1	-1	-1	-1	-1	-1	=	-8
		-1	1	-1	-1	-1	-1	-1	=	-3
			-1	-1	-1	-1	-1	-1	=	-6
				-1	-1	-1	-1	-1	=	-5
					-1	-1	+1	+1	=	0
						-1	-1	-1	=	-3
							1	+1	=	2
								1	=	1
								S_k	=	-31

Обчислимо коефіцієнт рангової кореляції Кендалла:

$$r_k = \frac{2S_k}{n(n-1)} = \frac{2(-31)}{10 \cdot 9} = -\frac{31}{45} = -0,6889.$$

Перевіримо гіпотезу $H_0: \rho_k = 0$ проти альтернативи $H_1: \rho_k \neq 0$. При $n = 10$ значенню $|S_k| = 31$ відповідає рівень значущості $\alpha = 0,0046$. Цей рівень менший за $0,01$. Отже, гіпотеза H_0 відкидається з ймовірністю помилки, меншої за $0,01$, тобто генеральний коефіцієнт рангової кореляції ρ_k значно відрізняється від нуля.

Обчислимо коефіцієнт r_c за умовою попереднього прикладу. Отримаємо

$$S_c = (10-1)^2 + (9-2)^2 + (6-3)^2 + (8-4)^2 + (7-5)^2 + (3-6)^2 + \\ + (5-7)^2 + (1-8)^2 + (2-9)^2 + (4-10)^2 = 306,$$

$$r_c = 1 - \frac{6 \cdot 306}{3 \cdot 330} = 1 - \frac{306}{165} = 20,8545.$$

За $n = 10$ та $S_c = 308$ з фрагменту таблиці отримуємо, що коефіцієнт рангової кореляції Спірмена є значущим на рівні, меншому за $0,0092$.

Таким чином, зв'язок між стажем програмувальника та часом, що витрачений на розв'язання задачі, можна вважати доведеним (чим більший стаж, тим менші в середньому витрати часу на розв'язання задачі).

Приклад 3.2 [3]

Трьом суддям належало надати місця п'яти спортсменам. Кожен суддя присудив кожному спортсменові місце (від першого до п'ятого) відповідно до ступеня переважності. Рішення суддів наведені у таблиці.

Таблиця 3.6

Спортсмен	Результат спортсмена	Місце	Рішення суддів
А	2	1	3
Б	3	3	2
В	1	2	1
Г	5	4	4
Д	4	5	5

Треба встановити, чи є у наведених результатах деякий порядок переваги, спільний для всіх, і якщо є, то визначити реальні місця, що займає кожен спортсмен, відповідно до дійсного ступеня продемонстрованої майстерності.

Розв'язання

Обчислимо суму квадратів відхилень S_w та коефіцієнт конкордації W .

Таблиця 3.7

Спортсмен	Сума місць S_i	Квадрати
А	6	36
Б	8	64
В	4	16
Г	13	169
Д	14	196
Усього	45	481

$$S_w = \sum_{i=1}^5 S_i^2 - \frac{(\sum S_i)^2}{n} = 481 - \frac{45^2}{5} = 76,$$

$$W = \frac{12S_w}{k^2(n^3 - n)} = \frac{12 \cdot 76}{3^2(5^3 - 5)} = 0,844.$$

Оскільки значенню $S_w = 76$ відповідає рівень значущості α , що перебуває в межах інтервалу $(0,0053; 0,015)$, гіпотезу про

неузгодженість думки суддів слід відкинути з ймовірністю помилки, меншою за 0,015. Висновок про остаточне розташування спортсменів по місцях можна зробити, керуючись, наприклад, сумами номерів S_i .

Так, у нашому прикладі спортсмен А має посісти друге місце, спортсмена Б слід поставити на третє, спортсмена В — на перше місце, потім йдуть спортсмени Г та Д.

За наявності об'єктів, що не розрізняються за ознаками (зв'язними рангами), засоби і формули обчислення дещо змінюються.

Для обчислення коефіцієнта рангової кореляції Кендалла та інших зручно скористатися таблицею сполученості ознак x і y :

Таблиця 3.8

$X \backslash Y$	1	...	j	...	b	n_{i*}
1	n_{11}	...	n_{1j}	...	n_{1b}	n_{1*}
...	...	...	...	...	...	...
i	n_{i1}	...	n_{ij}	...	n_{ib}	n_{i*}
...	...	...	...	...	...	...
a	n_{a1}	...	n_{aj}	...	n_{ab}	n_{a*}
n_{*j}	n_{*1}	...	n_{*j}	...	$n - b$	n_{**}

У таблиці n_{ij} — число (частота) об'єктів, спостережень, що мають i -й рівень (ранг) за ознакою x та j -й рівень (ранг) за ознакою y ; $n_{i*} = \sum_j n_{ij}$ — кількість об'єктів, що мають i -й рівень за ознакою x ; $n_{*j} = \sum_i n_{ij}$ — кількість об'єктів, що мають j -й рівень за ознакою y ; $n_{**} = \sum_i \sum_j n_{ij} = \sum_i n_{i*} = \sum_j n_{*j} = n$ — загальна кількість об'єктів (обсяг вибірки).

Обчислюються такі величини:

$$A = \sum_{i=1}^a \sum_{j=1}^b n_{ij} \left(\sum_{k>i} \sum_{l>j} n_{kl} \right), \quad B = \sum_{i=1}^a \sum_{j=1}^b n_{ij} \left(\sum_{k>i} \sum_{l<j} n_{kl} \right),$$

$$T_1 = \frac{1}{2} \sum_{i=1}^a n_{i*} (n_{i*} - 1), \quad A_{ij} = \sum_{k>i} \sum_{l>j} n_{kl} + \sum_{k<i} \sum_{l<j} n_{kl},$$

$$T_2 = \frac{1}{2} \sum_{j=1}^b n_{*j} (n_{*j} - 1), \quad B_{ij} = \sum_{k>i} \sum_{l<j} n_{kl} + \sum_{k<i} \sum_{l>j} n_{kl}.$$

Коефіцієнт рангової кореляції Кендалла обчислюється для квадратних таблиць сполученості, тобто для $a = b$, за формулою:

$$r_k = (A - B) / \left\{ [0,5n(n-1) - T_1][0,5n(n-1) - T_2] \right\}^{\frac{1}{2}}.$$

Коефіцієнт рангової кореляції Спірмена обчислюється за такою формулою

$$r_c = \frac{12 \sum_i \sum_j n_{ij} \left[\sum_{k<i} n_{k*} + (n_{k*}/2) - (n/2) \right] \left[\sum_{l<j} n_{*l} + (n_{*l}/2) - (n/2) \right]}{\left\{ \left[n^3 - n - \sum_i (n_{i*}^3 - n_{i*}) \right] \left[n^3 - n - \sum_j (n_{*j}^3 - n_{*j}) \right] \right\}^{\frac{1}{2}}}.$$

При $n > 10$ для перевірки гіпотези $H_0: \rho_c = Q$ проти $H_1: \rho_c \neq Q$ на рівні значущості α можна використовувати статистику r з числом степенів вільності $v = n - 2$ (або статистику Стьюдента

$$t = \sqrt{\frac{1 - r_c^2}{n - 2}} \quad \text{з } v = n - 2).$$

Для прямокутних таблиць сполученості впорядкованих значень ознак за $a \neq b$ має місце коефіцієнт зв'язності Стюарта, що визначається за формулою $r_{\text{ст}} = 2 \min(a, b)(A - B)/(n^2(\min(a, b) - 1))$, при цьому для достатньо великих значень n

$$\sigma_{r_{\text{ст}}} = \frac{2 \min(a, b)}{n^3(\min(a, b) - 1)} \left[n^2 \sum_i \sum_j n_{ij} (A_{ij} - B_{ij})^2 - 4n(A - B)^2 \right]^{\frac{1}{2}}.$$

За наявності зв'язних рангів формула для обчислювання коефіцієнта конкордації Кендалла модифікується:

$$r_w = \frac{12S_w}{k^2(n^3 - n) - k \sum_{j=1}^k T_j}, \quad T_j = \sum_{l=1}^{m_j} (n_{lj}^3 - n_{lj}),$$

де m_j — число груп зв'язних рангів у ознаки x_j , а n_{lj} — число об'єктів (рангів), що утворюють l -у групу зв'язних рангів (якщо ранги, що не розрізняються, відсутні, то $m_j = n$, $n_{lj} = l$, $l = \overline{1, n}$; отже, $T_j = 0$).

3.1.4. Оцінка щільності статистичного зв'язку для часових рядів

Кореляційний аналіз часто використовується при розв'язанні різних функціональних задач в АСК, особливо в економічних дослідженнях — для вивчення статистичного зв'язку між змінними, що подані часовими рядами. Незважаючи на те, що він не дозволяє відокремити прямі причинно-наслідкові зв'язки від непрямих або навіть просто від випадкових збігів коливань змінних, цей аналіз цінний тим, що забезпечує можливість надати кількісну означеність ступеню сполученості коливань цих рядів, підтвердити чи поставити під сумнів ті або інші гіпотези.

Економічні дослідження потребують особливої теорії кореляційного аналізу, яка спеціально призначена для вивчення сполученості (або зв'язності) коливань нестационарних процесів, поданих єдиною реалізацією.

Щоб отримати загальне уявлення про усереднені кореляційні властивості двох нестационарних рядів, було розроблено спеціальний статистичний показник кореляції, що ґрунтується на використанні приростів змінних. Для цього модифікований коефіцієнт кореляції представляється таким чином (для спрощення будемо вважати лаг $\tau = 0$) [1]:

$$R_{\text{mod}} = \frac{\frac{1}{T} \sum_{t=2}^T \Delta x_{1t} \Delta x_{2t}}{\frac{1}{T-1} \sum_{t=2}^T |\Delta x_{1t} \Delta x_{2t}|} = \frac{\sum_{t=2}^T \Delta x_{1t} \Delta x_{2t}}{\sum_{t=2}^T |\Delta x_{1t} \Delta x_{2t}|}, \quad (3.1)$$

де $\Delta x = x_{it} - x_{it-1}$, $i = 1, 2$.

Знаменник формули (3.1) грає роль нормуючого коефіцієнта. За такого вимірника ступеня корельованості двох рядів зазначений вище недолік усувається і не відбувається ніякої деформації вихідних даних. Кожен доданок у чисельнику та знаменнику є незмінним, не залежить від будь-яких суб'єктивних процедур або бажань дослідника.

Зазначимо, що усереднення за часом замість усереднення за реалізаціями тут залишається. Загалом, таке усереднення ґрунтується на гіпотезі про постійність кореляційного зв'язку двох рядів упродовж всього вибіркового періоду. Це припущення,

очевидно, також занадто сильне для економічних нестационарних процесів.

Обчислення коефіцієнта (3.1) можливе для не випадкових функцій, бо його визначення не пов'язане з будь-якими гіпотезами ймовірнісного характеру. Зокрема, легко помітити таку особливість запропонованого коефіцієнта кореляції: модифікований коефіцієнт кореляції будь-яких двох функцій, що монотонно зростають або монотонно спадають, завжди дорівнюватиме одиниці. У цьому випадку доцільно перейти до аналізу різниць другого (або навіть більш високого) порядку. Коефіцієнт, що отримується, можна назвати модифікованим коефіцієнтом кореляції другого (або більш високого) порядку.

Для нестационарних часових рядів статистичний зв'язок може ставати чи то слабше, чи то сильніше, і якщо знайти спосіб виявити динаміку цього зв'язку, то можна отримати додаткову цікаву інформацію про взаємозв'язки змінних, яка є корисною у змістовному аналізі досліджуваних явищ кореляції у часі. У роботі [28] запропоновано рекурентний коефіцієнт кореляції вигляду

$$r_t(d) = \frac{s_t}{d_t}, \quad (3.2)$$

де $s_t = (1 - \alpha)s_{t-1} + \alpha(\Delta x_{1t}\Delta x_{2t})$, $t = 1, 2, \dots, T$, $d_t = (1 - \alpha)d_{t-1} + \alpha|\Delta x_{1t}\Delta x_{2t}|$, $0 < \alpha < 1$.

Як бачимо, s_t та d_t є експоненціально-зваженими ковзними середніми добутоків приростів та абсолютних добутоків приростів двох рядів. Параметр α — *стала згладжування*, або параметр адаптації, показує, як змінюється сила кореляційного зв'язку на вибіркового періоді згладжування.

Початкові значення s_0 та d_0 у роботі [28] пропонується визначати як прості арифметичні середні добутоків та абсолютних значень добутоків приростів, що обчислюються на основі T_0 найбільш ранніх спостережень вибірки, $0 < T_0 \leq T - 1$. Тоді $r_0 = s_0/d_0$ буде початковим коефіцієнтом кореляції, що співпадає з модифікованим коефіцієнтом кореляції, який розраховано за (3.1) при $T = T_0$. Величина α може обиратися з вказаного діапазону $[0, 1]$, виходячи з конкретного застосування обчислюваного коефіцієнта кореляції.

Якщо обчислення такого коефіцієнта є самостійною задачею, то можна запропонувати такий спосіб визначення найкращого α .

Розглядатимемо $r_{t-1}(\alpha)$ як прогноз співпадіння коливання рядів x_{1t} та x_{2t} в момент t , а

$$e_t = \frac{\Delta x_{1t} \Delta x_{2t}}{|\Delta x_{1t} \Delta x_{2t}|} - r_{t-1}(\alpha) \quad -$$

як помилку цього прогнозу на той самий момент. Тоді можна ввести функцію втрат

$$Q_v(\alpha) = \sum_{t=2}^T |e_t|^v, \quad (3.3)$$

де $v = l$ або 2 , та визначати оптимальне значення α шляхом мінімізації функції $Q_v(\alpha)$ за α . Мінімум легко знайти методом спроб різних α з інтервалу $[0, 1]$. За необхідності так само методом спроб встановлюється й найкраще значення T_0 . У цьому випадку функція втрат розглядається як функція двох змінних $Q_v(T_0, \alpha)$.

Можливий й інший варіант коефіцієнта кореляції

$$r_t = (1 - \alpha)r_{t-1} + \alpha \frac{\Delta x_{1t} \Delta x_{2t}}{|\Delta x_{1t} \Delta x_{2t}|}. \quad (3.4)$$

Тут у якості r_α можна взяти відношення $r_0 = c_0/q_0$, де

$$c_0 = \frac{1}{T_0} \sum_{j=1}^{T_0} \Delta x_{1j} \Delta x_{2j}, \quad q_0 = \frac{1}{T_0} \sum_{j=1}^{T_0} |\Delta x_{1j} \Delta x_{2j}|.$$

Такий коефіцієнт називають *коефіцієнтом знакової кореляції приростів*. Далі, якщо діяти так само, як і в попередньому випадку, неважко обчислити помилки e_t та побудувати критерій $Q_v(T_0, \alpha)$, використовуючи (3.3) і (3.4). Оптимальні значення T_0 і α відповідатимуть мінімуму Q .

У класичному кореляційному аналізі відрхунок відхилень змінної ведеться від математичного сподівання. Однак на практиці таке відхилення являє інтерес лише в окремих випадках. Зокрема, в економічних дослідженнях набагато частіше важливий знак відхилення від попереднього значення ряду (тобто знак прирощення) або від нульового рівня. Тоді й відхилення доцільно розглядати відносно відповідних величин.

Загалом, в залежності від задачі, аналізу можуть піддаватися відхилення від моди, медіани, від будь-якого заданого рівня. Назвемо такий рівень критичним, маючи на увазі, що відхилення від нього у різні боки мають істотні якісні відмінності. Означимо такий рівень для ряду x_t через E_x . Нехай у цьому випадку по-

трібно встановити, чи існує взаємозв'язок між відхиленнями однієї змінної, наприклад $(x_t - E_x)$, з відхиленнями іншої $(z_t - E_z)$, і якщо існує, то наскільки він сильний. Візьмемо у якості вимірювача кореляційного зв'язку коефіцієнт

$$r_E = \frac{E[(x_t - E_x)(z_t - E_z)]}{E[(x_t - E_x)(z_t - E_z)]}$$

і розглянемо його властивості. Цілком зрозуміло, що він знаходиться у звичайних межах $-1 \leq r_E \leq 1$. Його оцінкою може слугувати кореляційний коефіцієнт:

$$\bar{r}_E = \frac{\sum (x_t - E_x)(z_t - E_z)}{\sum [(x_t - E_x)(z_t - E_z)]}.$$

Якщо $E_x = 0$ та $E_z = 0$, то коефіцієнт r_E характеризує силу взаємозв'язку між відхиленнями відповідних змінних від нуля. За $E_x = x_{t-1}$ та $E_z = x_{t-1}$ отримуємо розглянутий вище модифікований коефіцієнт кореляції, який також можна назвати *коефіцієнтом кореляції приростів*. Однак відзначимо, що при $E_x = M(x)$ та $E_z = M(z)$, де M — символ математичного сподівання, коефіцієнт r_E не є ідентичним класичному коефіцієнту кореляції через різні нормуючі множники (знаменники).

Ніяких обмежень для використання кореляційного коефіцієнта немає. Він може застосовуватися для змінних, що мають зміщений (коли $E_t = M(x)$) та несиметричний розподіл відносно критичного рівня E_x . Однак стаціонарність та ергодичність змінних передбачаються. Якщо ж ці припущення надто сильні, можна побудувати адаптивний коефіцієнт кореляції за аналогією з (3.2), де прирощення слід замінити на відхилення від критичних значень.

Розглянуті коефіцієнти кореляції відображають силу взаємозв'язку коливань двох рядів у цілому. Така характеристика може виявитися надто загальною і не завжди задовольнить дослідника. Поставимо запитання більш конкретніше: наскільки сильний зв'язок, наприклад, додатних відхилень x_t з відхиленнями z_t ? Для його вимірювання введемо поняття *умовного коефіцієнта кореляції*.

Визначимо його як

$$r_c(x_t, z_t | x_t \geq E_x) = \frac{\sum (x_t - E_x)(z_t - E_z)}{\sum [(x_t - E_x)(z_t - E_z)]}, \quad x_t \geq E_x,$$

де сумуються тільки ті добутки, в яких $x_t \geq E_x$. Так само можна обчислити коефіцієнт взаємозв'язку від'ємних відхилень ряду x_t з відхиленням z_t . Природно, що в обох випадках коефіцієнти кореляції можуть набувати неоднакові значення. Таким чином, особливості взаємозв'язку двох рядів тепер не переходять за одним-єдиним значенням коефіцієнта кореляції, як це передбачається в класичній теорії, а представляються більш дробово двома коефіцієнтами.

Переваги такого підходу найбільш яскраво виявляються тоді, коли одна змінна (наприклад, x_t) є контрольованою, а інша (z_t) — безпосередньо недоступною, але корельованою з першою. Тоді можна обрати таке критичне значення E_x , при якому утримання x_t в області $E_x \leq x_t$ (або $x_t \leq E_x$) мало б найбільш сприятливі наслідки (тобто найбільше значення коефіцієнта кореляції) для утримання z_t в межах $z_t \leq E_z$ (або, навпаки, $z_t \geq E_z$). У загальному випадку умови при обчисленні коефіцієнта кореляції можуть, звісно, мати довільний вигляд

$$r_c(x_t, z_t | A \leq x_t - E_x \leq B) = \frac{\sum (x_t - E_x)(z_t - E_z)}{\sum |(x_t - E_x)(z_t - E_z)|}, \quad (3.5)$$

$$A \leq x_t - E_x \leq B,$$

аби вистачило статистичних даних для отримання стійких оцінок коефіцієнтів.

Доводом на користь застосування умовного коефіцієнта кореляції (3.5) можуть слугувати такі міркування. При низьких значеннях амплітуди відхилень кореляційний зв'язок може не простежуватись через вплив випадкових перешкод, тобто бути неістотним. При великих відхиленнях визначеність цього зв'язку зростає майже до детермінованості. Тоді за великої амплітуди відхилень однієї змінної різко зростає ступінь зумовленості області знаходження (наприклад, знак приросту) другої змінної. Легше визначити і лаг (запізнювання у часі) однієї змінної відносно іншої. Можна вважати, що в цьому випадку зростає потужність критерію.

Задаючи значення A та B , можна виділити будь-яку область амплітуд відхилень першої змінної ($x_t - E_x$), що нас цікавить, та обчислити коефіцієнт кореляції зі змінною z_t для цієї області. Такий підхід відкриває шлях до досконального вивчення взаємозв'язку рядів, виявлення специфічних локальних кореляцій-

них властивостей. Можна сказати, що кореляційні властивості постають у «розщепленому» вигляді — їх виражає множина коефіцієнтів.

У роботі [3] розглянуто ще один метод дослідження сили сполученості двох змінних на основі оцінки умовної ймовірності їхнього зв'язку. Коефіцієнт кореляції визначається як

$$\begin{aligned} r_p(x_t, z_t \mid A \leq x_t - E_x \leq B, C \leq z_t - E_z \leq D) = \\ = p(C \leq z_t - E_z \leq D, A \leq x_t - E_x \leq B), \end{aligned} \quad (3.6)$$

де p — ймовірність того, що $C \leq z_t - E_z \leq D$ за умови $A \leq x_t - E_x \leq B$.

Звідси виходить, що областю існування коефіцієнта кореляції буде відрізок $[0, 1]$. Вона відрізняється від області існування розглянутих вище коефіцієнтів кореляції, яка охоплює відрізок $[-1, 1]$.

Визначення (3.6) теоретичне. Практичну оцінку на основі статистичних даних можна отримати як відношення

$$\bar{r}_p(x_t, z_t \mid A \leq x_t - E_x \leq B, C \leq z_t - E_z \leq D) = \frac{n_{z|x}}{N_x},$$

де $n_{z|x}$ — число спостережень, коли $z_t - E_z$ опинялося у межах $C \leq z_t - E_z \leq D$ при $A \leq x_t - E_x \leq B$; N_x — загальне число спостережень у вибірці, що задовольняють умові $A \leq x_t - E_x \leq B$.

Введені визначення кореляційних коефіцієнтів мають низку переваг у порівнянні з класичним, але, звісно, не претендують на те, щоб повністю замінити собою останнє. Навпаки, у кожній конкретній задачі треба використовувати той тип кореляційного коефіцієнта, який є найадекватнішим меті, що поставлена.

У запропонованому модифікованому коефіцієнті кореляції враховуються прирощення змінних разом з їхнім знаком. Це рівнозначне до того, що замість відхилень від математичного сподівання тут розглядаються відхилення від раніше досягнутих значень. Ідея відмовитись від математичного сподівання як від рівня відліку відхилень, та замінити його на деякий критичний (по суті, довільний, заданий дослідником) рівень, знайшла застосування і в інших розглянутих визначеннях. Адаптивний коефіцієнт кореляції дозволяє простежувати динаміку кореляційного зв'язку у часі. Для дослідження кореляції знаків прирощувань змінних побудовано *адаптивний коефіцієнт знакової кореляції*. Практичне використання цього підходу — для короткочасного

прогнозування валютних курсів. Звісно, цей метод може застосовуватись і на ринку цінних паперів.

Умовний коефіцієнт кореляції, як і ймовірнісний, можна назвати *парціальним коефіцієнтом кореляції*, бо вони обидва здатні відображати окремі властивості двох досліджуваних часових рядів, які спостерігаються за певних умов. Два останніх підходи дозволяють проводити диференційний аналіз взаємозв'язку змінних, виявляти, переходячи від одного діапазону амплітуд відхилень до іншого, неоднорідність зв'язку у різних піддіапазонах та надають досліднику більше інформації для міркування та використання у аналітичній та прогнозній діяльності. За необхідності можуть бути побудовані адаптивні варіанти цих коефіцієнтів. Однією з перспективних галузей застосування подібних підходів є побудова моделей, що перемикаються, тобто моделей з перемиканням у критичні моменти часу з одного режиму роботи на інший. Кожен режим можна представити у моделі спеціальним регресійним рівнянням, а для визначення критичних моментів перемикання режимів доцільно проводити попередній кореляційний аналіз. Критичними є моменти, коли змінна перетинає (в наших означеннях) рівень E_t , A або B .

Розглянуті підходи дають можливість отримувати значно більш докладнішу інформацію про взаємозв'язки змінних, ніж це пропонує класична теорія, та відкривають нові перспективи у пізнанні складних динамічних процесів.

3.2. Побудова кореляційного рівняння

Поставлено задачу: обчислити за даними лінійну апроксимуючу функцію у вигляді

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x_1 + \hat{b}_2 x_2 + \dots + \hat{b}_k x_k + e. \quad (3.7)$$

Залежність (3.7) є регресійним рівнянням. Щоб отримати кореляційну залежність, необхідно перевірити нормальність закону розподілу і всі дані нормалізувати за формулами:

$$t(x_{ji}) = \frac{x_{ji} - \bar{x}_j}{\sigma\{x_j\}}, \quad t(y_i) = \frac{y_i - \bar{y}}{\sigma\{y_i\}},$$

де j — номер вхідної змінної, $j = 1, \dots, p$; $\sigma\{x\}$ та $\sigma\{y\}$ — середньоквадратичні відхилення.

Нові змінні мають однакові математичні сподівання та дисперсії, тобто

$$\bar{t}_j = \bar{t}_y = 0 \quad (j=1,2,\dots,p), \quad \sigma^2\{t_j\} = \sigma^2\{t_y\} = 1.$$

Тоді рівняння регресії набуває вигляду:

$$t_y = \hat{\beta}_1 t_1 + \hat{\beta}_2 t_2 + \dots + \hat{\beta}_p t_p. \quad (3.8)$$

Коєфіцієнти $\beta_1, \dots, \beta_k$ дозволяють судити про порівняльну силу впливу кожної із вхідних змінних на цільову функцію y . У (3.8) немає вільного члена. Оцінки $\hat{\beta}_j$ відрізняються від оцінок $\hat{b}_j$ — у регресійному аналізі, але вони, за умови дотримання передумов 1–5, дозволяють судити про відносну силу впливу кожної з вхідних змінних x_j на Y .

Щоб визначити оцінки вектора β_i , методом МНК визначимо

$$J = \sum_{i=1}^n (t_{yi} - \hat{t}_{yi})^2 \rightarrow \min. \quad (3.9)$$

Для обчислення $\hat{\beta}_i$ використовуємо

$$\frac{\partial J}{\partial \beta_i} = 0. \quad (3.10)$$

Підставляючи (3.9) в (3.10) та прирівнюючи окремі похідні до нуля, маємо j рівнянь з j невідомими. Отримаємо:

$$\left\{ \begin{array}{l} \sum_{i=1}^n 2(t_{yi} - \hat{\beta}_1 t_{1i} - \hat{\beta}_2 t_{2i} - \dots - \hat{\beta}_p t_{pi})(-t_{1i}) = 0, \\ \dots\dots\dots \\ \sum_{i=1}^n 2(t_{yi} - \hat{\beta}_1 t_{1i} - \hat{\beta}_2 t_{2i} - \dots - \hat{\beta}_p t_{pi})(-t_{pi}) = 0. \end{array} \right. \quad (3.11)$$

Якщо всі j рівнянь (3.11) розділити на $(n - 1)$ і скоротити на 2, то при кожному $\hat{b}_j$ у якості множників будуть парні коефіцієнти кореляції.

$$\tau_{jl} = \tau\{X_j, X_l\} = \frac{1}{n-1} \sum_{i=1}^n t_{ji} \dots t_{li}.$$

Вільним членом у кожному з n рівнянь є коефіцієнт кореляції між кожним i -входом та виходом.

Приймемо до уваги

[illegible]

За експериментальними даними розраховуємо r_{jl} та підставляємо в (3.12). У матричній формі систему рівнянь (3.12) запишемо у вигляді:

$$A = KB, \quad (3.13)$$

де

$$A = \begin{vmatrix} r_{y1} \\ r_{y2} \\ \vdots \\ r_{yp} \end{vmatrix}; \quad K = \begin{vmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{vmatrix}; \quad B = \begin{vmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_p \end{vmatrix}.$$

Розв'яжемо рівняння (3.13) відносно B , отримаємо

$$K^{-1}A=K^{-1}KB, \quad K^{-1}K=E=1, \quad B=K^{-1}A.$$

Обернену матрицю K отримаємо послідовним обчисленням кожного її члена A (j — номер рядка, l — номер стовпчика).

Через визначники прямої матриці K , тобто

$$k_{jl} = \frac{\Delta_{jl}}{\Delta} (-1)^{j+1},$$

де Δ — головний визначник матриці K , а Δ_{jl} — мінор головного визначника Δ , що отриманий шляхом викреслення j -го рядка та l -го стовпчика головного визначника Δ , знак мінора враховується співмножником $(-1)^{j+l}$.

Коефіцієнти $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$ показують силу впливу усіх x_j на Y , а знак β_j — напрямок впливу j -ї змінної. Фізичний сенс коефіцієнтів β полягає у такому: кожен із $\hat{\beta}_j$ показує, на яку частину σ_y змінилося б середнє значення функції відгуку y , якби вхідна змінна x_j змінилася на σ_{xj} , а решта вхідних змінних залишалися б при цьому стабілізованими на своїх рівнях.

Для практичного використання рівняння (3.12) необхідно оцінити статистичну значущість оцінок коефіцієнтів $\hat{\beta}_j$, використовуючи t -критерій Стюдента, оскільки Y — вектор, розподілений за нормальним законом, а β_j є лінійною комбінацією Y і, отже, теж розподілений за нормальним законом.

З цією метою використовують

$$t = \sqrt{n-2} \frac{r_{ji}}{\sqrt{1-r_{ji}^2}}$$

з числом степенів вільності $n-2$. Якщо виявиться, що $t < t_{\text{кр}} = t_q$, то коефіцієнти r_{ji} або r_{yj} визнаються такими, що незначно відрізняються від нуля.

Середньоквадратична помилка у визначенні оцінок $\hat{\beta}_j$ знаходиться за формулою

$$S\{\beta_j\} = \sqrt{\frac{1 - \hat{R}^2}{n - (p+1)}} \cdot \alpha_{jj},$$

де $\hat{R}^2 = \sum_{j=1}^p \hat{\beta}_j r_{yj}$, α_{jj} — діагональний член оберненої матриці кореляційної матриці K^{-1} .

Довірчі інтервали для коефіцієнтів β_j визначаються таким чином:

$$\left[\hat{\beta}_j - t_{\text{кр}} S\{\hat{\beta}_j\} < \beta_j < \hat{\beta}_j + t_{\text{кр}} S\{\hat{\beta}_j\} \right].$$

Оцінки коефіцієнтів $\hat{b}_0, \hat{b}_1, \dots, \hat{b}_k$ рівняння у натуральному масштабі розраховуються за формулами

$$\hat{b}_j = \frac{S\{y\}}{S\{x_j\}} \hat{\beta}_j, \quad \hat{b}_0 = \bar{y} - \sum_{j=1}^k \hat{b}_j \bar{x}_j.$$

Таким чином отримують математичну модель у натуральному масштабі. Модель представляється у вигляді рівняння регресії, яке показує, як змінюється положення середнього значення виходу зі зміною вхідних змінних.

Оцінку щільності статистичного зв'язку здійснює коефіцієнт множинної кореляції $\hat{R}$. Вважається нормальним, якщо $\hat{R} = 0,8-0,9$.

Статистичну значущість коефіцієнта множинної кореляції оцінюють шляхом порівняння $\hat{R}$ зі середньоквадратичною помилкою $S\{\hat{R}\}$. За $R=0$ $\hat{S}[R] = \frac{1}{\sqrt{N-1}}$ і якщо виконується нерів-

ність $\frac{\hat{R}}{S\{\hat{R}\}} > 3$ ($\alpha = 1 - P = 0,01$), тоді $\hat{R}$ відрізняється від нуля

з ймовірністю $P = 0,99$ ($\alpha = 1 - P = 0,01$).

Межі довірчих інтервалів для R визначають за допомогою Z -критерію Фішера.

Якщо вдається отримати середньоквадратичну помилку експериментально $S_b\{x\}$ за паралельними дослідями, то можна перевірити адекватність за F -критерієм Фішера.

$$F = \frac{S_{\text{ост}}^2\{y\}}{S_b^2\{y\}}, \quad S_{\text{ост}}^2\{y\} = \frac{1}{n-p} \sum_{j=1}^n (y_{ji} - \bar{y}),$$

де p — число членів регресії.

За дотримання умови

$$F < F_{\text{табл}}(n-k, \zeta-1, \alpha),$$

рівняння регресії вважається адекватним.

На основі вищезазначеного, запропонуємо алгоритм проведення кореляційного аналізу. На першому етапі перевіримо гіпотезу про випадковий характер змінних X та Y .

Задача перевірки випадковості може бути розбита на *дві підзадачі*:

1) перевірка випадковості власне величин X та Y ;

2) перевірка випадковості вибірки, тобто припущення про відсутність істотного зміщення середньої величини у часі.

Перша підзадача розв'язується шляхом використання критерію серій. З цією метою послідовність спостережень величин X та Y представляють у вигляді послідовності нулів та одиниць, де одиницею позначають значення, яке перевищує середнє або медіану, та нулем — власне значення, менше за медіану. Після позначення вектор спостережень перетворюється у послідовність серій $\{r_i\}$, де r_i — кількість елементів одного виду, що йдуть поспіль, i — номер серії. Доведено, що при $N > 10$ розподіл величини r наближається до нормального з такими характеристиками:

$$\bar{r} = \frac{N+1}{2}, \quad S_r = \frac{1}{2}\sqrt{N-1}.$$

Тоді з ймовірністю 0,954 теоретичне число серій r буде знаходитися в межах

$$\frac{N+1}{2} - \sqrt{N-1} \leq r_T \leq \frac{N+1}{2} + \sqrt{N-1}.$$

Якщо фактичне значення r_{ϕ} потрапляє у зазначені межі, то змінну, яка розглядається, можна вважати випадковою величиною.

Серією називається послідовність спостережених значень, перед якими та після яких розташовані спостережені значення іншої категорії. Якщо послідовність N спостережень представляє незалежні спостережені значення однієї й тієї ж випадкової величини, тобто ймовірність знаків (+) та (-) не змінюється від одного спостереження до іншого, то вибірковий розподіл числа серій у послідовності є випадкова величина r з середнім значенням

$$\mu_r = \frac{2N_1N_2}{N} + 1 \quad (1)$$

та дисперсією

$$\sigma_r^2 = \frac{2N_1N_2(2N_1N_2 - N)}{N^2(N-1)}. \quad (2)$$

Тут N_1 — число спостережень зі знаком (+), N_2 — число спостережень зі знаком (-). Коли $N_1 = N_2 = N/2$, співвідношення (1) та (2) набувають вигляду:

$$\mu_r = (N/2) + 1, \quad \sigma_r^2 = \frac{N(N-2)}{4(N-1)}.$$

У [10] наведено таблицю, що містить дані про 100-відсоткові точки функції розподілу для критеріїв серій. Гіпотезу можна перевірити при будь-якому заданому рівні значущості α шляхом співставлення числа серій з граничними значеннями $r_{n, 1-\alpha/2}$ та $r_{n, \alpha/2}$, де $n = N/2$. Якщо спостережене число серій виходить за межі цього інтервалу, гіпотезу слід відкинути при рівні значущості α .

Приклад 3.3

Нехай Y складається з таких чисел ($N = 20$):

5,5 5,7 6,8 5,4

5,1 5,0 5,5 6,8

5,7 6,5 4,9 5,8

5,2 5,4 5,4 6,9

4,8 5,8 5,9 5,5

Визначимо, чи є незалежні спостереження випадковими.

Розв'язання

Оберемо $\alpha = 0,05$. Медіаною для $N = 20$ буде величина $x = 5,6$. Складемо послідовність з (+) та (-).

--	+	--	+	-	+	-	+++	--	+	-	+++	-
1	2	3	4	5	6	7	8	9	10	11	12	13

Число серій дорівнює 13. Область прийняття гіпотези визначається інтервалом $r_{10, 1-\alpha/2} < r \leq r_{10, 1-\alpha/2}$. З таблиці [10] при $\alpha = 0,05$ знаходимо $r_{10, 1-\alpha/2} = r_{10, 0,975} = 6$ та $r_{10, \alpha/2} = r_{10, 0,025} = 15$. Гіпотеза приймається, бо $r = 13$ входить до інтервалу.

Щоб розв'язати другу підзадачу, використовується *метод послідовних різниць*.

Елементи вибірки $\{y_i\}$, що досліджується, розташовуються за порядком отримання спостережень, для них обчислюються вибіркові середні та дисперсія

$$\bar{y} = \frac{1}{N} \sum_{i=1}^N y_i, \quad \hat{S}^2 = \frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{y})^2.$$

Визначають різниці Δy між сусідніми спостереженнями

$$\Delta y_1 = y_2 - y_1, \quad \Delta y_2 = y_3 - y_2, \quad \dots, \quad \Delta y_{N-1} = y_N - y_{N-1},$$

та математичне сподівання квадрату різниці

$$M[\Delta y] = \frac{1}{N-1} \sum \Delta y_i^2 = 2\hat{c}^2,$$

де $\hat{c}^2$ — оцінка генеральної дисперсії.

Фактична величина критерію випадковості вибірки

$$t_{\phi} = \frac{\hat{c}^2}{S^2}.$$

Відомо, що при $N > 20$ та випадковій вибірці t_{ϕ} розподілена нормально. Виходячи з 1 %-го рівня значущості, маємо $F(t_T) = 0,99$ та $t_T = 2,58$. Тоді теоретичне значення критерію

$$t_T = 1 - \frac{2,58}{\sqrt{N+1}}.$$

За умови $t_{\phi} < t_T$ для конкретного N гіпотеза випадковості відкидається.

На другому етапі попередньо визначимо наявність лінійної залежності між X та Y за *діаграмою розсіювання*, яка являє собою графічне відображення точок $(X_1, Y_1), \dots, (X_n, Y_n)$ на площину XY . Принцип побудови діаграм розсіювання з використанням ППП STATISTICA викладено в розділі 4.

На наступному етапі необхідно перевірити, чи підкоряються всі вхідні та вихідні випадкові змінні багатовимірному нормальному закону розподілу ймовірності. Принцип перевірки з використанням ППП STATISTICA викладено в розділі 4.

Функціональну зв'язаність вхідних випадкових змінних між собою оцінимо за кореляційною матрицею. Статистичну незалежність окремих вимірювань за кожною з вхідних змінних оцінимо за автокореляційними функціями. Далі будується рівняння регресії і проводиться інтерпретація результатів. Визначається значущість парних та множинних коефіцієнтів кореляції із зазначенням точкових та інтервальних оцінок. Аналізуються окремі коефіцієнти кореляції.

3.3. Приклад побудови кореляційної моделі курсу валют за фундаментальними факторами

Аналіз фундаментальних факторів проведемо на прикладі офіційного курсу гривні відносно до долара США, що встановлений Національним банком України, за період з січня 1996 року до грудня 2001 року (середнє значення за період); дані отримані з Інтернету.

Позначимо:

Y_CURS	— курс валют;
X1_VVP	— валовий внутрішній продукт, визначений у фактичних цінах;
X2_DOHOD	— грошові доходи населення, млн. грн;
X3_RASHOD	— грошові витрати та заощадження, млн. грн;
X4_InCenR	— індекс цін виробників промислової продукції;
X5_InCenP	— індекс цін споживчого ринку;
X6_DefBud	— розмір дефіциту зведеного бюджету;
X7_MO	— гроші поза банком;
X8_M1 = MO+	— кошти у розрахункових і поточних рахунках, визначені у національній валюті;
X9_M2 = M1+	— кошти, визначені у національній валюті, та валютні кошти;
X10_KrNac	— відсоткові ставки банків;
X11_DepNac	— відсоткові ставки банків у депозиті.

Щоб побудувати кореляційну модель, використаємо пакет прикладних програм STATISTICA. Зауважимо, що модуль «Кореляційний аналіз» у пакеті відсутній. Однак використання модулів **Basic Statistic / Tables** та **Multiple Regression** дозволить розв'язати усі статистичні задачі кореляційного аналізу.

На першому етапі роботи з ППП STATISTICA введемо в систему вихідні дані. Робоче вікно системи STATISTICA реалізовано згідно зі стандартами програм, що працюють у середовищі **Windows**, і схоже на вікна інших прикладних цього середовища, таких як **MS Word**, **MS Excel** та ін.

Робоче вікно має таку структуру: верхній заголовок STATISTICA; нижче — рядок меню, панель інструментів і робоча область, що займає більшу частину вікна. В робочу область виводяться всі документи системи, які отримуються у процесі аналізу.

Вихідні дані в системі STATISTICA організовано у вигляді електронної таблиці. Якщо ви маєте досвід роботи з **MS Excel**, то вам буде легко працювати з електронними таблицями в STATISTICA. Таблиці з вихідними даними в STATISTICA мають особливу назву, щоб не плутати їх з іншими таблицями системи.

Стовпчики електронної таблиці з вихідними даними називаються **Variables** (Змінні), а рядки **Cases** (Спостереження). У якості змінних зазвичай виступають досліджувані величини, а спостереження — це значення, яких набувають змінні в окремих вимірюваннях.

Система може працювати як з числовими, так і з текстовими даними, що, безумовно, важливо в практичних статистичних дослідженнях. Загалом, електронні таблиці можуть містити числову та текстову інформацію.

Підведіть курсор миші до пункту меню **File** (Файл) та клацніть лівою кнопкою. У випадному меню, яке з'явиться на екрані, оберіть **Open Data** (Відкрити дані).

Команди з випадного меню можна викликати також за допомогою певної комбінації «гарячих» клавіш. Наприклад, щоб обрати команду **Open Data** (Відкрити дані) з меню **File** (Файл) використовуйте комбінацію CTRL+O. Після натискання цих клавіш або курсора миші на команді **Open Data** (Відкрити дані) на екрані з'явиться діалогове вікно, в якому обирається файл даних (рис. 3.1).

З системою STATISTICA постачається велика кількість файлів, що містять цікаві дані. Ці файли знаходяться в каталозі **Statistica\examples\datasets**.

Продемонструємо на простих прикладах, як створюються файли даних у STATISTICA.

Вихідне положення — основне робоче вікно STATISTICA. У верхній частині вікна розташований рядок меню.

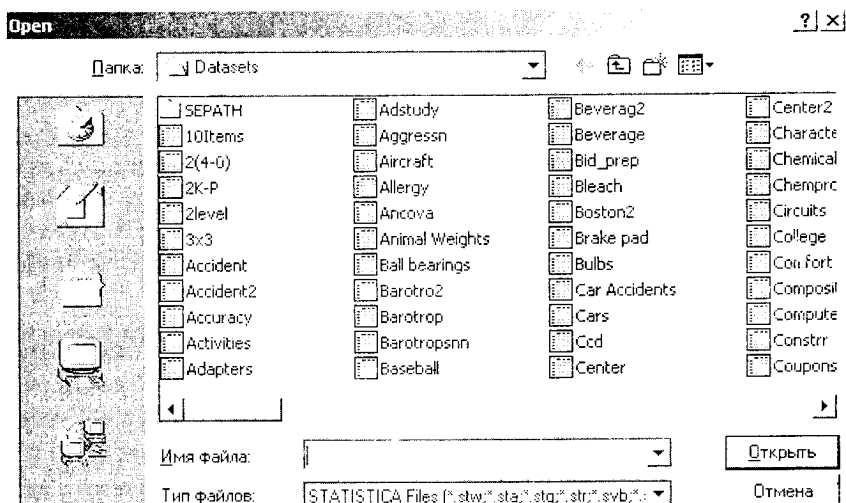


Рис. 3.1. Відкриття файла з каталогу stat\examples

Створення нового файлу даних починається з вибору у рядку меню пункту **File** (Файл) (знаходиться у лівому кінці). Підведіть курсор миші до слова **File** (Файл) та клацніть лівою кнопкою. У випадному меню, яке з'явилося на екрані після курсору миші, оберіть команду **New** (Нові дані).

Приклад 3.4

Створення файлу REKLAMA1.STA

У таблиці наведені дані про розмір рекламного оголошення та його вартість при розміщенні у газеті [2].

Таблиця 3.9

Вартість рекламних оголошень

Довжина, мм	Ширина, мм	Площа, мм ²	Ціна, грн
47	35	1645	1446
47	73	3431	2768
47	111	5217	3974
47	149	7003	5147
47	209	9823	6290
47	225	10575	7537
47	263	12361	8828
47	301	14147	10260

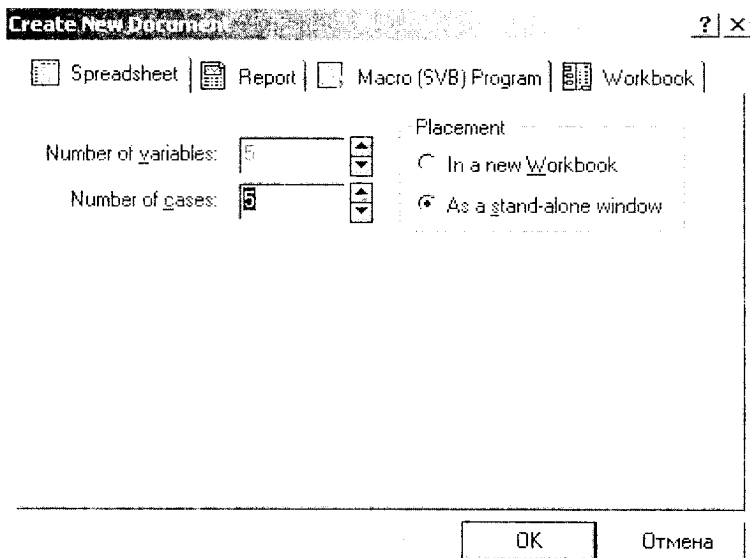
Створимо файл STATISTICA з цими даними.

Розв'язання

Крок 1. Створення електронної таблиці. Оберіть команду **New** (Нові дані) з меню **File** (Файл). Ця команда доступна також з комбінації клавіш CTRL+N. У діалоговому вікні, що з'явилося, **Create New Document** оберіть закладку **Spreadsheet** для задання ім'я таблиці, збережіть його під ім'ям **reklama.sta**. Створення файлу саме з таким ім'ям показано на рис. 3.2.

Натисніть кнопку **OK** у правому верхньому кутку вікна. **STATISTICA** автоматично відкриє порожню електронну таблицю з ім'ям **reklama.sta**, яка з'явиться на екрані. Цією таблицею можна користуватися як сторінкою у нотатнику та вносити до неї необхідні дані.

У заголовку вікна електронної таблиці автоматично відображається ім'я та розмір файлу (**reklama.sta 5v*5с**). Розмір таблиці за замовчуванням — 10 × 10 (10 змінних з іменами **VAR1**, **VAR2**, **VAR3**, ..., **VAR10** та 10 пронумерованих спостережень).



	1 Var1	2 Var2	3 Var3	4 Var4	5 Var5
1					
2					
3					
4					
5					

Рис. 3.2. Створення нового файла

Зробимо в таблиці необхідну кількість рядків та стовпчиків. Нам треба, щоб у таблиці було 4 змінні та 8 спостережень.

Крок 2. Настроювання таблиці. Проведемо настроювання розмірів таблиці — створимо необхідну кількість змінних та спостережень. Для даних рекламного оголошення потрібні чотири змінні: ширина рекламної об'яви, довжина, площа, ціна. (Вважатимемо, що реклама має стандартну прямокутну форму.)

Натисніть кнопку **Vars Variables** (Змінні) на панелі інструментів, оберіть команду **Delete Variables** (Видалення змінних)

вказіть діапазон змінних, що видаляються, як показано на рис. 3.3. Натисніть кнопку **OK**.

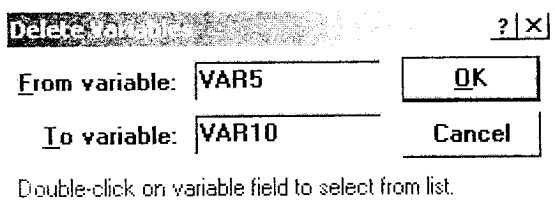
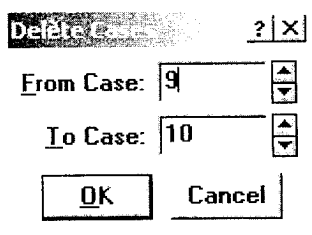


Рис. 3.3. Вікно видалення змінних

Необхідне число спостережень — 8. У створеній таблиці число спостережень дорівнює 10. Два зайвих спостереження з таблиці слід видалити.

Щоб видалити зайві спостереження, скористайтесь кнопкою **Cases** (Спостереження) та командою **Delete** (Видалення). У вікні, що з'явилося, зробіть установки, як показано на рисунку:



Зазначте діапазон спостережень, що видаляються, у діалоговому вікні **Delete Cases** (Видалити спостереження) і натисніть кнопку **OK**. Тепер електронна таблиця виглядає таким чином, як на рис. 3.4.

Крок 3. Підготовка таблиці до введення даних, заголовка таблиці та імен змінних. Двічі клацніть мишею на білому полі в таблиці під словами **Data: REKLAMA1.STA 4V*8C**. На екрані з'явиться вікно **Data File Header** (Заголовок файлу даних), в якому можна дати заголовок таблиці та додаткову інформацію про дані. Введіть заголовок таблиці «ЦІНА РЕКЛАМИ» та додаткову інформацію про знижки на рекламу, як показано на рис. 3.4. Ця додаткова інформація завжди буде для вас досяжною.

Data: REKLAMA				
	1	2	3	4
	VAR1	VAR2	VAR3	VAR4
1				
2				
3				
4				
5				
6				
7				
8				

Рис. 3.4. Порожня таблиця з чотирма змінними та вісьмома спостереженнями

Натисніть кнопку **OK** у діалоговому вікні **Data File Header** (Заголовок файла даних). Тепер таблиця виглядає, як показано на рис. 3.5 і майже готова до введення даних, однак її можна дещо вдосконалити: вивести імена змінних, які відображають сенс записів, та специфікувати їх.

Data: REKLAMA				
	1	2	3	4
	VAR1	VAR2	VAR3	VAR4
1				
2				
3				
4				
5				
6				
7				
8				

Рис. 3.5. Таблиця з заголовком

Задайте імена змінних. Двічі клацніть на імені змінної **VAR1** в електронній таблиці. На екрані з'явиться вікно специфікації **VAR1**. У полі **Name** (Ім'я) напишіть: «ДОВЖИНА». Довжина імені має не перевищувати 8 символів.

За замовчуванням змінні відображаються у форматі восьми значущих цифр та трьома розрядами після десяткової точки. Для даних реклами достатньо 5 значущих цифр, а розряди після коми не потрібні. Ця установка проводиться у рамці **Display Format** (Формат відображення) шляхом задання параметра **Desimals** (Число розрядів після десяткової точки), що дорівнює нулю. Інші установки виконайте згідно з рис. 3.6. Натисніть кнопку **OK**.

Variable 2 ? X

Name: ШИРИНА **MD code:** -9999

Display Format

Column width: 5 **Decimals:** 0

Category:	Representation:
Number	1000; -1000
Date	1 000; -1 000
Time	1000; {1000}
Scientific	1 000; {1 000}
Currency	
Percentage	

OK **Cancel** << >>

All Specs **Text Values** **Values/Stats** **Graphs**

Long name (label, link, or formula with Functions):

Examples: Label: Gross income in 1991 Formulas: = v1 + v2 ; comment
 Link: @Excellc:\file.xls!r2c2:r4c4 = {v1>0}*AGE + v3

Рис. 3.6. Найменування змінних

Те саме зробить для змінної **VAR2**, надайте їй ім'я **ШИРИНА**, змінній **VAR3** надайте ім'я **ПЛОЩА**, змінній **VAR4** — ім'я **ЦІНА**. Для двох останніх змінних у рамці **Display Format** (Формат відображення) задайте **Decimals** (Число розрядів після десяткової

точки), що дорівнює нулю, а в рамці **Column width** (Ширина стовпчика) — восьми (рис. 3.8).

	Цена рекламы			
	1	2	3	4
	Длина	Ширина	Площадь	Цена
1				
2				
3				
4				
5				
6				
7				
8				

Рис. 3.7. Таблиця з новими іменами змінних

Тепер таблиця готова до того, щоб вводити дані.

Крок 4. Введення даних в електронну таблицю. Оскільки даних небагато, будемо вводити їх з клавіатури. Введіть дані, як показано на рис. 3.8. Дані вводяться тільки у колонки **ДОВЖИНА**, **ШИРИНА**, **ЦІНА**. Колонка **ПЛОЩА** залишається порожньою.

Уведемо дані у колонку **ПЛОЩА**. Зробимо це прямим обчисленням площі прямокутника за довжиною та шириною.

Двічі клацніть на імені змінної **ПЛОЩА** в електронній таблиці. На екрані з'явиться вікно специфікації змінної **ПЛОЩА** (див. рис. 3.9).

У полі **Long names** (Довгі імена) запишіть формулу $=v1*v2$. Площа прямокутника дорівнює добутку довжини на ширину. Натисніть кнопку **ОК**.

Площа рекламних оголошень буде обчислена та занесена у комірки змінної **VAR3**. Заповнена таблиця вартості рекламних оголошень у газеті з'явиться на екрані (рис. 3.10).

Крок 5. Збереження файла даних. Щоб зберегти створений файл, клацніть мишею на кнопці **Save Data File** (Зберегти файл

Data Base				
N	1	2	3	4
V.	VAR1	VAR2	VAR3	VAR4
1	47,000	35,000		1446000.
2	47,000	73,000		2768000.
3	47,000	111,000		3974000.
4	47,000	149,000		5147000.
5	47,000	187,000		6290000.
6	47,000	225,000		7537000.
7	47,000	263,000		8828000.
8	47,000	301,000		10260000.

Рис. 3.8. Таблиця з даними, що введені з клавіатури

Variable 3 ? | X

Name: ПЛОЩАДЬ MD code: -9999

Display Format

Column width: 8 Decimals: 0

Category: Representation:

Number	1000; -1000
Date	1 000; -1 000
Time	1000; (1000)
Scientific	1 000; (1 000)
Currency	
Percentage	

OK Cancel

<< >>

All Specs Text Values Values/Stats Graphs

Long name (label, link, or formula with Functions):

=v1*v2

Examples: Label: Gross income in 1991 Formulas: = v1 + v2 ; comment
 Link: @Excel[c:\file.xls]r2c2:r4c4 = (v1>0)*AGE + v3

Рис. 3.9. Обчислення значень змінної ПЛОЩА

Data: Reklama (47 n)				
Цена рекламы				
	1	2	3	4
	Длина	Ширина	Площадь	Цена
1	47	35	1645	1446
2	47	73	3431	2768
3	47	111	5217	3974
4	47	149	7003	5147
5	47	209	9823	6290
6	47	225	10575	7537
7	47	263	12361	8828
8	47	301	14147	10260

Рис. 3.10. Заповнена електронна таблиця з вихідними даними

даних) або наберіть на клавіатурі CTRL + S. Створений файл буде збережено і буде забезпечено доступ до нього. Далі ми будемо використовувати цей файл даних для аналізу.

Аналогічно до наведеного вище введемо дані щодо курсу валют, які представлені у додатку (табл. 1).

Після введення вихідних даних перевіримо дотримання передумов кореляційного аналізу. З цією метою у панелі меню **Statistics** оберемо пункт **Basic Statistic / Tables**. У діалозі, що з'явився, оберемо **Correlation Matrices** (див. рис. 3.11)

Далі слід задати змінні, за якими буде будуватися матриця. Використовуючи кнопку **ON variable List**, оберемо всі вхідні змінні, як показано на рис. 3.12.

Натиснувши кнопку **Summary**, отримаємо матрицю коефіцієнтів кореляції (рис. 3.13).

Проведемо порівняльний аналіз коефіцієнтів кореляції. Нас буде цікавити, у якій мірі кожна зі змінних X впливає на всі інші змінні X у вільному стані (тобто не зафіксовані).

Нагадаємо, що кореляція слугує мірою сили статистичного взаємозв'язку, та оцінюють його за абсолютним значенням. Кореляція, що наближається до 1 або -1, вказує на сильний взаємозв'язок, а кореляція, що наближається до 0, свідчить про відсутність взаємозв'язку. Абсолютне значення кореляції вказує

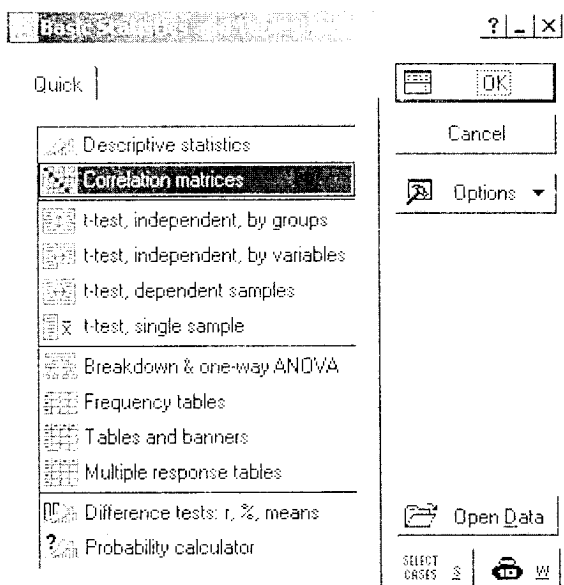


Рис. 3.11. Меню Correlation Matrices

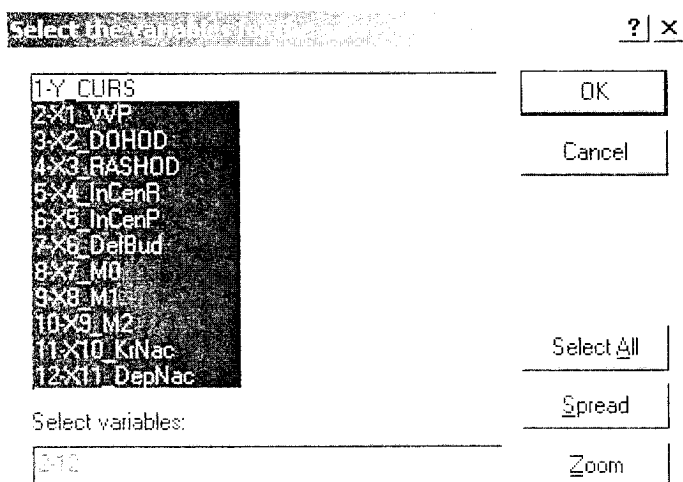


Рис. 3.12. Вибір вхідних змінних

Variable	X1_VWF_2_DONHO_3_RASHC_4_InCenF_5_InCenF_6_DefBud_7_M0_8_M1_9_M2_10_KrNa_11_DepNa										
X1_VVP	1,00	0,07	0,44	-0,11	-0,22	0,57	0,91	0,34	0,07	-0,57	-0,57
X2_DONHO	0,07	1,00	0,50	-0,15	-0,27	-0,58	0,53	0,96	0,35	0,71	-0,10
X3_RASHOD	0,44	0,50	1,00	-0,15	0,25	-0,63	0,67	0,90	-0,37	0,71	0,70
X4_InCenR	-0,11	-0,15	-0,15	1,00	0,51	0,03	-0,16	-0,18	-0,16	0,38	0,18
X5_InCenP	-0,22	-0,27	0,25	0,51	1,00	0,06	-0,12	0,27	0,24	0,50	0,50
X6_DefBud	0,57	-0,58	-0,53	0,03	0,06	1,00	-0,59	0,71	-0,73	0,44	0,49
X7_M0	0,91	0,96	0,90	-0,16	-0,12	-0,59	1,00	1,00	0,68	0,74	-0,23
X8_M1	0,34	0,96	0,97	-0,18	-0,27	-0,71	1,00	1,00	-0,73	-0,71	-0,73
X9_M2	0,07	0,35	0,37	-0,16	0,24	-0,73	0,71	1,00	1,00	-0,71	-0,73
X10_KrNa	-0,57	0,71	-0,73	0,38	0,50	0,44	-0,72	-0,73	0,71	1,00	0,98
X11_DepNa	-0,57	-0,10	-0,70	0,18	0,50	0,49	-0,75	-0,71	-0,70	0,98	1,00

Рис. 3.13. Матриця коефіцієнтів кореляції

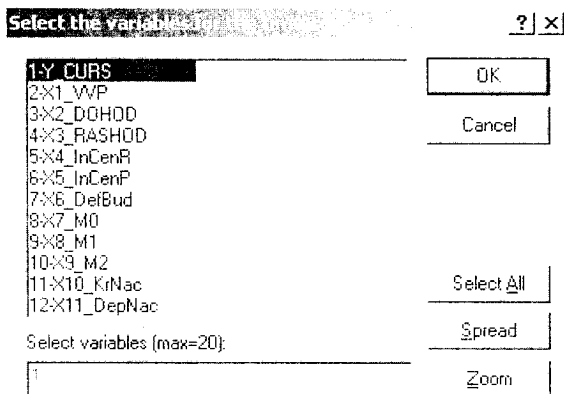
на силу взаємозв'язку, не визначаючи його напрямку. Отже, якщо не треба враховувати поправку на дію інших змінних, можна скористатися аналізом кореляцій. Якщо треба брати до уваги вплив інших змінних, тоді слід користуватися множинною регресією.

Матрицю коефіцієнтів кореляції, що наведена на рис. 3.13, можна використати для перевірки припущення про незалежність вхідних змінних, тобто вхідні змінні не повинні мати функціональний зв'язок.

Перевірка незалежності вимірювань

Перевірку незалежності змінних проведемо за допомогою автокореляційної функції. Ця можливість присутня в модулі **Time Series / Forecasting**. Для цього оберемо **Statistics → dvanced Linear / Nonlinear Models → Time Series / Forecasting**.

У діалоговому вікні, що з'явилося, натиснувши кнопку **Variables**, оберемо змінну, що досліджується. Наприклад, **Y1_CURS**:



Потім натиснемо кнопку **Arima & autocorrelation functions**. Далі перейдемо до вкладки **Autocorrelations**.

Single Series Arima

OK (Begin parameter estimation)

Lock Variable Long variable (series) name

L Y_CURS

Cancel

Options

Number of backups per variable (series): 1

Save variables

Quick | Advanced | Autocorrelations | Review series

Autocorrelations

☒ White noise standard errors

Partial autocorrelations

p-level for highlighting: 0.50

Number of lags: 40

У полі **Number of lags** задамо кількість лагів, що досліджуються. Натиснувши кнопку **Autocorrelations**, отримаємо графік автокореляційної функції (див. рис. 3.14). Автокореляційна функція показує ступінь залежності між вимірюваннями.

Такі дослідження слід проводити за кожною з змінних, що досліджуються.

Тепер перейдемо безпосередньо до проведення кореляційного аналізу. Для цього скористаємося модулем **Multiple Regression** у панелі меню **Statistics** (див. рис. 3.15).

У першу чергу необхідно обрати залежну та незалежну змінні. Для цього натиснемо кнопку **Variables**. У розділі залежних змінних (**Dependent var.**) оберемо Y1_CURS. У розділі незалежних (**Independent var.**) — решту (див. рис. 3.16).

Натиснемо кнопку **OK** і отримаємо результати регресійного аналізу. Значення отриманих змінних описані у розділі «Регре-

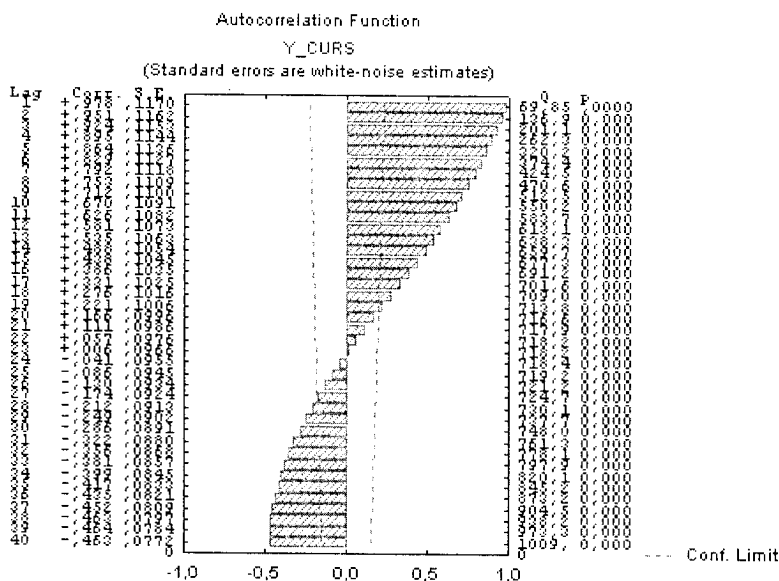


Рис. 3.14. Графік автокореляційної функції

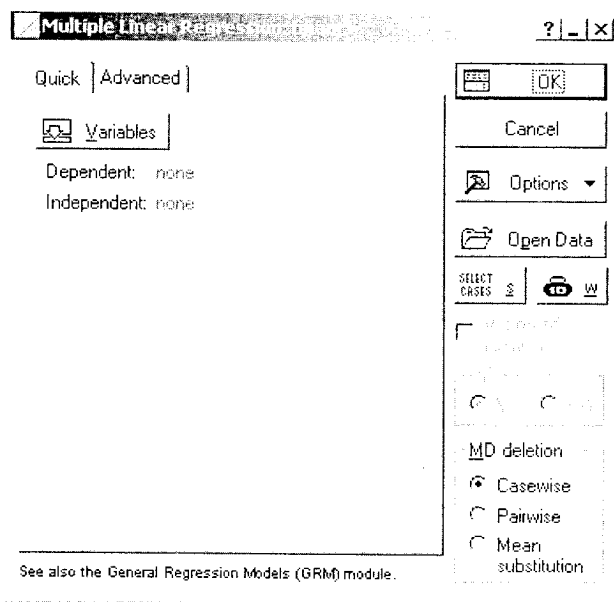


Рис. 3.15. Модуль Multiple Regression

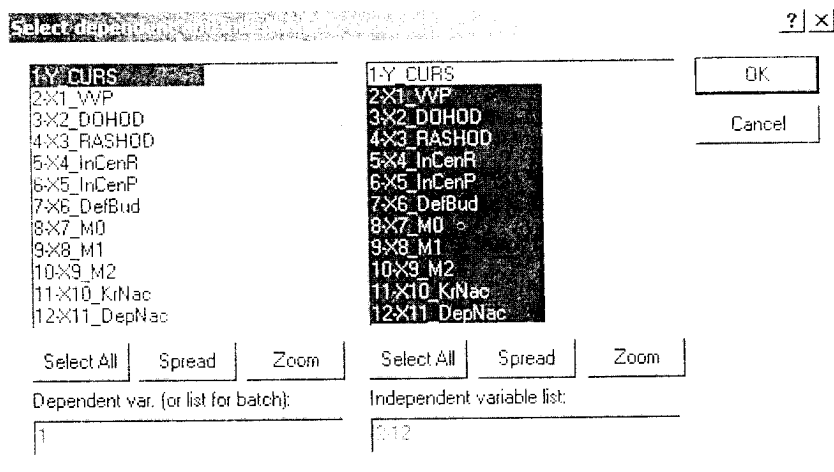


Рис. 3.16. Вибір залежних та незалежних змінних

сійний аналіз». Нас же цікавить таблиця результатів регресійного аналізу, яку можна отримати, натиснувши кнопку «Summary: Regression results».

Таблиця 3.10

Результати регресійного аналізу

N=70	Beta	Std. Err. of Beta	B	Std. Err. of B	t(58)	p-level
Intercept			-220,982	345,9274	-0,63881	0,525461
X1_VVP	0,09386	0,077727	0,003	0,0029	1,20761	0,232097
X2_DOHOD	0,23175	0,165663	0,018	0,0130	1,39894	0,167155
X3_RASHOD	-0,28228	0,192864	-0,024	0,0165	-1,46363	0,148694
X4_InCenR	-0,06829	0,035109	-5,986	3,0772	-1,94519	0,056605
X5_InCenP	0,12009	0,036794	10,094	3,2364	3,11671	0,003827
X6_DefBud	-0,05154	0,066973	-0,021	0,0269	-0,76958	0,444672
X7_M0	1,04198	0,327135	0,045	0,0141	3,18517	0,003329
X8_M1	-6,64092	0,737891	-0,175	0,0194	-9,00863	0,000000
X9_M2	6,38376	0,824997	0,106	0,0137	7,73602	0,000000
X10_KrNac	0,21535	0,124333	1,936	1,1175	1,73204	0,088583
X11_DepNac	-0,03242	0,168219	-5,827	1,9668	-2,97932	0,004214

Рівняння регресії (кореляції) будується за допомогою отриманих коефіцієнтів та має вигляд:

$$t_{Y1\_CURS} = 0,09t_{X1} + 0,23t_{X2} - 0,28t_{X3} - 0,07t_{X4} + 0,12t_{X5} - 0,05t_{X6} + \\ + 1,04t_{X7} - 6,64t_{X8} + 6,38t_{X9} + 0,22t_{X10} - 0,32t_{X11}.$$

Значущість стандартизованих коефіцієнтів Beta оцінимо згідно з табл. 3.10 за величиною p -level. Якщо значення p -level $< 0,05$, коефіцієнт вважається значущим. Виходячи з табл. 3.10, значущими виявилися коефіцієнти $X_5, X_7, X_8, X_9, X_{11}$.

Проведемо інтерпретацію стандартизованих коефіцієнтів регресії. Оскільки усі звичайні коефіцієнти регресії $b_1, \dots, b_k$ можна виразити у різних одиницях виміру, безпосереднє їх порівняння доволі ускладнено: невеликий коефіцієнт може насправді виявитися більш важливим, ніж великий.

Коефіцієнт регресії b_j вказує вплив зміни X на змінну Y , коли всі інші X -змінні залишаються незмінними. Коефіцієнт регресії b_j визначається в одиницях виміру Y на одну одиницю виміру X_j . Якщо, наприклад, Y являє собою курс валют, а X_1 — величина VVP у країні, то b_1 виражається у кількості курсу валют (обсяг) на одну одиницю VVP. Припустимо, що наступний коефіцієнт регресії, b_2 , виражається кількістю курсу валют (обсяг) на прибуток населення. Безпосереднє порівняння b_1 та b_2 не дозволить відповісти на запитання, який з цих двох факторів більше впливає на кількість курсу валют, тому що різні одиниці виміру порівнювати не можна.

Стандартизований коефіцієнт регресії, який обчислюється шляхом помноження коефіцієнта регресії b_j на S_{x_j} та ділення отриманого добутку на S_Y , являє собою очікуваний вимір Y (у стандартизованих одиницях S_Y), викликаний збільшенням X_j на одну відповідну стандартизовану одиницю (тобто S_{x_j}), коли усі інші змінні X залишаються незмінними. Абсолютні значення стандартизованих коефіцієнтів регресії можна порівнювати, отримуючи при цьому деяке уявлення про відносну важливість відповідних змінних. Кожен стандартизований коефіцієнт регресії вимірюється в одиницях стандартних відхилень Y на одне стандартне відхилення X_j .

Наведемо інтерпретацію одного з цих стандартизованих коефіцієнтів. Значення 0,23 свідчить про те, що збільшення курсу валют на одне стандартне відхилення (S_{x_2}) призведе до очікуваного збільшення курсу валют на 0,23 його стандартних відхилень.

Однак більш важлива та обставина, що ці стандартизовані коефіцієнти регресії тепер можна порівнювати між собою.

Найбільший за абсолютним значенням коефіцієнт свідчить про те, що ця змінна є найважливішою з решти.

Було б неправильно порівнювати коефіцієнти регресії безпосередньо, не стандартизувавши їх попередньо.

Абсолютні значення стандартизованих коефіцієнтів регресії можна порівнювати одне з іншим, що дозволяє отримати приблизне уявлення про важливість відповідних змінних. Слід підкреслити, що ці результати не є ідеальними, оскільки взаємозв'язки між змінними X можуть зробити принципово неможливим з'ясування того, яка з них насправді «відповідає» за поведінку змінної Y .

Множинна регресія робить поправку на інші змінні, а коефіцієнт кореляції — ні. Якщо потрібно враховувати вплив інших змінних, тоді слід користуватися множинною регресією, в іншому випадку можна скористатися аналізом кореляцій.

Множинна регресія дозволяє контролювати значення інших змінних; взаємозв'язки з Y , стандартизовані коефіцієнти регресії з множинної регресії можуть надати важливу додаткову інформацію.

Як можна оцінити та інтерпретувати значущість множинного коефіцієнту кореляції та довірчі інтервали для стандартизованих коефіцієнтів регресії показано у розділі 4.

У наступному прикладі оцінимо ступінь статистичного зв'язку між змінними X_1 (VVP), X_2 (DOHOD) та X_3 (RASHOD), що представлені у табл. 3.11, без використання ППП STATISTICA.

Припустимо, що розглянуті ознаки X_1 , X_2 та X_3 у генеральній вибірці розподілені за нормальним законом і являють собою вибірку цієї сукупності.

Таблиця 3.11

№	X_1	X_2	X_3	№	X_1	X_2	X_3
1	5060	2539	2581	9	10615	3422	3394
2	5140	3022	2824	10	4810	3787	3476
3	6488	2953	2953	11	7529	3311	3304
4	5062	3300	3067	12	12115	4521	4068
5	5730	3200	3051	13	5920	3283	3220
6	7075	3049	2919	14	6107	3462	3403
7	5155	3849	3485	15	6701	3938	3704
8	6740	3358	3839	16	7493	4144	3699

Закінчення табл. 3.11

№	X_1	X_2	X_3	№	X_1	X_2	X_3
17	6998	3885	3824	45	11844	5935	5492
18	5994	4140	3880	46	12851	5260	5676
19	8383	4844	4407	47	10622	7092	6577
20	7503	5005	4366	48	9934	4500	5322
21	10190	4655	4826	49	10887	5775	5530
22	6584	4398	4566	50	12477	6731	6202
23	8205	3811	4003	51	11868	7121	5967
24	13287	5199	4766	52	12705	6293	6393
25	6544	3210	3603	53	14359	7397	7040
26	6511	3796	3755	54	15211	7608	6972
27	7816	4848	4269	55	16564	8456	7875
28	7640	3923	3995	56	19283	7386	7878
29	8538	4523	4414	57	14096	7737	8457
30	8262	4929	4733	58	16662	7907	7884
31	8594	4448	4646	59	18906	10000	8257
32	12052	4536	4613	60	13616	6182	7002
33	7701	4936	4726	61	13720	7397	7053
34	9905	4819	4766	62	15069	8345	7867
35	11841	6116	5556	63	16179	8540	7728
36	8017	3359	3825	64	16104	8184	8231
37	7960	4091	4058	65	15641	9020	8045
38	9003	4839	4837	66	20896	9323	8959
39	9318	4950	4328	67	21388	10039	9335
40	10043	4541	4251	68	16684	9249	8632
41	9835	5189	4904	69	21130	10241	9671
42	12173	5204	5181	70	18121	10040	9416
43	11980	5483	5185	71	13379	12275	10340
44	13480	5833	5396				

Для отримання точкових оцінок глобальних середніх, дисперсій середніх квадратичних відхилень та парних коефіцієнтів кореляції результати проміжних обчислень через великий розкид значень попередньо стандартизуємо та для зручності помістимо у розрахункову табл. 3.12. З метою контролю обчислень дані розбиті на групи по 8 вимірювань. Для кожної «вісімки»

у розрахунковій таблиці наведено рядкові суми елементів кожного стовпчика. В останньому контрольному стовпчику наведено суми елементів відповідних рядків. Сума восьми елементів контрольного стовпчика має співпадати з сумою елементів підсумкового рядка.

Таблиця 3.12

X_1	X_2	X_3	X_1^2	X_2^2	X_3^2	X_1X_2	X_1X_3	X_2X_3	Конт- рольна сума
0,2366	0,2068	0,2496	0,0560	0,0428	0,0623	0,0489	0,0591	0,0516	1,0137
0,2403	0,2462	0,2731	0,0578	0,0606	0,0746	0,0592	0,0656	0,0672	1,1446
0,3033	0,2406	0,2856	0,0920	0,0579	0,0816	0,0730	0,0866	0,0687	1,2893
0,2367	0,2688	0,2966	0,0560	0,0723	0,0880	0,0636	0,0702	0,0797	1,2320
0,2679	0,2607	0,2951	0,0718	0,0680	0,0871	0,0698	0,0791	0,0769	1,2763
0,3308	0,2484	0,2823	0,1094	0,0617	0,0797	0,0822	0,0934	0,0701	1,3580
0,2410	0,3136	0,3370	0,0581	0,0983	0,1136	0,0756	0,0812	0,1057	1,4241
0,3151	0,2736	0,3713	0,0993	0,0748	0,1378	0,0862	0,1170	0,1016	1,5767
2,1718	2,0587	2,3906	0,6004	0,5364	0,7246	0,5585	0,6522	0,6216	10,3147
0,4963	0,2788	0,3282	0,2463	0,0777	0,1077	0,1384	0,1629	0,0915	1,9279
0,2249	0,3085	0,3362	0,0506	0,0952	0,1130	0,0694	0,0756	0,1037	1,3770
0,3520	0,2697	0,3195	0,1239	0,0728	0,1021	0,0950	0,1125	0,0862	1,5337
0,5664	0,3683	0,3934	0,3209	0,1357	0,1548	0,2086	0,2229	0,1449	2,5158
0,2768	0,2675	0,3114	0,0766	0,0715	0,0970	0,0740	0,0862	0,0833	1,3443
0,2855	0,2820	0,3291	0,0815	0,0795	0,1083	0,0805	0,0940	0,0928	1,4334
0,3133	0,3208	0,3582	0,0982	0,1029	0,1283	0,1005	0,1122	0,1149	1,6494
0,3503	0,3376	0,3577	0,1227	0,1140	0,1280	0,1183	0,1253	0,1208	1,7747
2,8656	2,4332	2,7338	1,1207	0,7493	0,9392	0,8847	0,9916	0,8381	13,5563
0,3272	0,3165	0,3698	0,1071	0,1002	0,1368	0,1036	0,1210	0,1170	1,6991
0,2803	0,3373	0,3752	0,0785	0,1138	0,1408	0,0945	0,1052	0,1266	1,6521
0,3919	0,3946	0,4262	0,1536	0,1557	0,1817	0,1547	0,1671	0,1682	2,1937
0,3508	0,4077	0,4222	0,1231	0,1663	0,1783	0,1430	0,1481	0,1722	2,1117
0,4764	0,3792	0,4667	0,2270	0,1438	0,2178	0,1807	0,2224	0,1770	2,4911
0,3078	0,3583	0,4116	0,0948	0,1284	0,1950	0,1103	0,1359	0,1582	1,9303
0,3836	0,3105	0,3871	0,1472	0,0964	0,1499	0,1191	0,1485	0,1202	1,8625
0,6212	0,4235	0,4609	0,3859	0,1794	0,2125	0,2631	0,2863	0,1952	3,0282
3,1393	2,9277	3,3499	1,3171	1,0839	1,4127	1,1690	1,3345	1,2346	16,9687
0,3060	0,2615	0,3485	0,0936	0,0684	0,1214	0,0800	0,1066	0,0911	1,4771

Продовження табл. 3.12

X_1	X_2	X_3	X_1^2	X_2^2	X_3^2	X_1X_2	X_1X_3	X_2X_3	Конт- рольна сума
0,3044	0,3092	0,3632	0,0927	0,0956	0,1319	0,0941	0,1106	0,1123	1,6140
0,3654	0,3949	0,4129	0,1335	0,1560	0,1705	0,1443	0,1509	0,1631	2,0915
0,3361	0,3499	0,4156	0,1130	0,1224	0,1727	0,1176	0,1397	0,1454	1,9124
0,3572	0,3196	0,3864	0,1276	0,1021	0,1493	0,1142	0,1380	0,1235	1,8178
0,3992	0,3685	0,4269	0,1594	0,1358	0,1822	0,1471	0,1704	0,1573	2,1467
0,3863	0,4015	0,4577	0,1492	0,1612	0,2095	0,1551	0,1768	0,1838	2,2813
0,4018	0,3624	0,4493	0,1615	0,1313	0,2019	0,1456	0,1805	0,1628	2,1971
2,8565	2,7676	3,2603	1,0304	0,9729	1,3394	0,9981	1,1735	1,1393	15,5380
0,5635	0,3695	0,4461	0,3175	0,1366	0,1990	0,2082	0,2514	0,1649	2,6567
0,3601	0,4021	0,4571	0,1296	0,1617	0,2089	0,1448	0,1646	0,1838	2,2126
0,4631	0,3926	0,4609	0,2145	0,1541	0,2125	0,1818	0,2135	0,1810	2,4739
0,5536	0,4982	0,5373	0,3065	0,2483	0,2887	0,2758	0,2975	0,2677	3,2737
0,3748	0,2736	0,3699	0,1405	0,0749	0,1368	0,1026	0,1387	0,1012	1,7131
0,3722	0,3333	0,3925	0,1385	0,1111	0,1540	0,1240	0,1461	0,1308	1,9024
0,4209	0,3942	0,4678	0,1772	0,1554	0,2188	0,1659	0,1969	0,1844	2,3816
0,4357	0,4033	0,4186	0,1898	0,1626	0,1752	0,1757	0,1824	0,1688	2,3119
3,5439	3,0669	3,5502	1,6142	1,2046	1,5940	1,3789	1,5909	1,3826	18,9261
0,4696	0,3699	0,4111	0,2205	0,1369	0,1690	0,1737	0,1930	0,1521	2,2958
0,4598	0,4227	0,4743	0,2115	0,1787	0,2249	0,1944	0,2181	0,2005	2,5849
0,5692	0,4240	0,5011	0,3239	0,1797	0,2511	0,2413	0,2852	0,2124	2,9878
0,5601	0,4467	0,5015	0,3137	0,1995	0,2515	0,2502	0,2809	0,2240	3,0280
0,6303	0,4752	0,5219	0,3972	0,2258	0,2723	0,2995	0,3289	0,2480	3,3991
0,5538	0,4835	0,5311	0,3067	0,2338	0,2821	0,2677	0,2941	0,2568	3,2096
0,6009	0,4285	0,5489	0,3610	0,1836	0,3013	0,2575	0,3298	0,2352	3,2468
0,4966	0,5778	0,6361	0,2466	0,3338	0,4046	0,2869	0,3159	0,3675	3,6658
4,3402	3,6283	4,1259	2,3812	1,6718	2,1568	1,9712	2,2460	1,8965	24,4179
0,4645	0,3666	0,5147	0,2157	0,1344	0,2649	0,1703	0,2391	0,1887	2,5588
0,5090	0,4705	0,5348	0,2591	0,2213	0,2860	0,2395	0,2722	0,2516	3,0441
0,5834	0,5484	0,5998	0,3403	0,3007	0,3598	0,3199	0,3499	0,3289	3,7310
0,5549	0,5801	0,5771	0,3079	0,3365	0,3330	0,3219	0,3202	0,3348	3,6665
0,5940	0,5127	0,6183	0,3529	0,2628	0,3823	0,3045	0,3673	0,3170	3,7117
0,6714	0,6026	0,6809	0,4507	0,3631	0,4636	0,4046	0,4571	0,4103	4,5042
0,7112	0,6198	0,6743	0,5058	0,3841	0,4546	0,4408	0,4795	0,4179	4,6881

Закінчення табл. 3.12

X_1	X_2	X_3	X_1^2	X_2^2	X_3^2	X_1X_2	X_1X_3	X_2X_3	Конт- рольна сума
0,7745	0,6889	0,7616	0,5998	0,4746	0,5800	0,5335	0,5898	0,5247	5,5273
4,8628	4,3895	4,9614	3,0322	2,4776	3,1242	2,7349	3,0752	2,7738	31,4317
0,9016	0,6017	0,7619	0,8128	0,3621	0,5805	0,5425	0,6869	0,4584	5,7084
0,6591	0,6303	0,8179	0,4344	0,3973	0,6689	0,4154	0,5390	0,5155	5,0778
0,7790	0,6442	0,7625	0,6069	0,4149	0,5814	0,5018	0,5940	0,4912	5,3758
0,8840	0,8147	0,7985	0,7814	0,6637	0,6377	0,7201	0,7059	0,6505	6,6565
0,6366	0,5036	0,6772	0,4053	0,2536	0,4586	0,3206	0,4311	0,3410	4,0277
0,6415	0,6026	0,6821	0,4115	0,3631	0,4653	0,3866	0,4376	0,4110	4,4013
0,7046	0,6798	0,7608	0,4964	0,4622	0,5789	0,4790	0,5360	0,5172	5,2149
0,7565	0,6957	0,7474	0,5722	0,4840	0,5586	0,5263	0,5654	0,5200	5,4260
5,9627	5,1726	6,0083	4,5209	3,4009	4,5298	3,8923	4,4959	3,9050	41,8884
0,7529	0,6667	0,7960	0,5669	0,4445	0,6337	0,5020	0,5994	0,5307	5,4929
0,7313	0,7348	0,7780	0,5348	0,5400	0,6054	0,5374	0,5690	0,5717	5,6024
0,9770	0,7595	0,8664	0,9545	0,5769	0,7507	0,7420	0,8465	0,6581	7,1317
1,0000	0,8178	0,9028	1,0000	0,6689	0,8151	0,8178	0,9028	0,7384	7,6636
0,7801	0,7535	0,8348	0,6085	0,5677	0,6969	0,5878	0,6512	0,6290	6,1095
0,9879	0,8343	0,9353	0,9760	0,6961	0,8748	0,8242	0,9240	0,7803	7,8330
0,8473	0,8179	0,9106	0,7178	0,6690	0,8293	0,6930	0,7715	0,7448	7,0013
0,6255	1,0000	1,0000	0,3913	1,0000	1,0000	0,6255	0,6255	1,0000	7,2679
6,7020	6,3846	7,0241	5,7499	5,1630	6,2058	5,3298	5,8900	5,6531	54,1022

Далі отримаємо таблицю підсумкових рядків (табл. 3.13)

Таблиця 3.13

Таблиця підсумкових рядків

X_1	X_2	X_3	X_1^2	X_2^2	X_3^2	X_1X_2	X_1X_3	X_2X_3	Конт- роль
2,1718	2,0587	2,3906	0,6004	0,5364	0,7246	0,5585	0,6522	0,6216	10,3147
2,8656	2,4332	2,7338	1,1207	0,7493	0,9392	0,8847	0,9916	0,8381	13,5563
3,1393	2,9277	3,3499	1,3171	1,0839	1,4127	1,1690	1,3345	1,2346	16,9687
2,8565	2,7676	3,2603	1,0304	0,9729	1,3394	0,9981	1,1735	1,1393	15,5380
3,5439	3,0669	3,5502	1,6142	1,2046	1,5940	1,3789	1,5909	1,3826	18,9261
4,3402	3,6283	4,1259	2,3812	1,6718	2,1568	1,9712	2,2460	1,8965	24,4179
4,8628	4,3895	4,9614	3,0322	2,4776	3,1242	2,7349	3,0752	2,7738	31,4317

Закінчення табл. 3.13

X_1	X_2	X_3	X_1^2	X_2^2	X_3^2	X_1X_2	X_1X_3	X_2X_3	Конт- роль
5,9627	5,1726	6,0083	4,5209	3,4009	4,5298	3,8923	4,4959	3,9050	41,8884
6,7020	6,3846	7,0241	5,7499	5,1630	6,2058	5,3298	5,8900	5,6531	54,1022
36,4448	32,829	37,4	21,367	17,2604	22,0265	18,917	21,4498	19,4446	227,144

В останньому рядку підсумкової таблиці подано суми значень у кожному стовпчику.

Для отримання точкових оцінок парних коефіцієнтів кореляції використовуємо формулу

$$R_{X_iX_j} = \frac{\sum X_iX_j - \frac{\sum X_i \sum X_j}{n}}{\sqrt{\left(\sum X_i^2 - \frac{(\sum X_i)^2}{n}\right) \left(\sum X_j^2 - \frac{(\sum X_j)^2}{n}\right)}}, \quad i \neq j = 1, 2, 3,$$

$$R_{X_1X_2} = \frac{\sum X_1X_2 - \frac{\sum X_1 \sum X_2}{n}}{\sqrt{\left(\sum X_1^2 - \frac{(\sum X_1)^2}{n}\right) \left(\sum X_2^2 - \frac{(\sum X_2)^2}{n}\right)}} = 0,88922,$$

$$R_{X_1X_3} = \frac{\sum X_1X_3 - \frac{\sum X_1 \sum X_3}{n}}{\sqrt{\left(\sum X_1^2 - \frac{(\sum X_1)^2}{n}\right) \left(\sum X_3^2 - \frac{(\sum X_3)^2}{n}\right)}} = 0,91428,$$

$$R_{X_2X_3} = \frac{\sum X_2X_3 - \frac{\sum X_2 \sum X_3}{n}}{\sqrt{\left(\sum X_2^2 - \frac{(\sum X_2)^2}{n}\right) \left(\sum X_3^2 - \frac{(\sum X_3)^2}{n}\right)}} = 0,97995.$$

Далі обчислимо значення квадратичних відхилень для кожної зі змінних за формулою:

$$S_{X_i}^2 = \overline{X_i^2} - (\overline{X_i})^2, \quad i = 1, n,$$

де

$$\overline{X_i^2} = \frac{1}{n} \sum_{i=1}^n X_i^2, \quad \overline{X_i} = \frac{1}{n} \sum_{i=1}^n X_i,$$

$$S_{X_1}^2 = \overline{X_1^2} - (\overline{X_1})^2 = 0,0405, \quad S_{X_2}^2 = \overline{X_2^2} - (\overline{X_2})^2 = 0,0318,$$

$$S_{X_3}^2 = \overline{X_3^2} - (\overline{X_3})^2 = 0,0360.$$

Тепер обчислимо точкові оцінки умовних середніх квадратичних відхилень (при одній фіксованій змінній). Для цього скористаємося формулою:

$$S_{X_i/X_j} = \sqrt{S_{X_i}^2 (1 - R_{X_i X_j}^2)}, \quad i \neq j = 1, 2, 3,$$

$$S_{X_1/X_2} = \sqrt{S_{X_1}^2 (1 - R_{X_1 X_2}^2)} = 0,0921,$$

$$S_{X_1/X_3} = \sqrt{S_{X_1}^2 (1 - R_{X_1 X_3}^2)} = 0,0816,$$

$$S_{X_2/X_1} = \sqrt{S_{X_2}^2 (1 - R_{X_1 X_2}^2)} = 0,0816,$$

$$S_{X_2/X_3} = \sqrt{S_{X_2}^2 (1 - R_{X_2 X_3}^2)} = 0,0355,$$

$$S_{X_3/X_1} = \sqrt{S_{X_3}^2 (1 - R_{X_1 X_3}^2)} = 0,0769,$$

$$S_{X_3/X_2} = \sqrt{S_{X_3}^2 (1 - R_{X_2 X_3}^2)} = 0,0378.$$

Отримаємо точкові оцінки окремих коефіцієнтів кореляції:

$$R_{X_i X_j / X_k} = \frac{R_{X_i X_j} - R_{X_i X_k} R_{X_j X_k}}{\sqrt{(1 - R_{X_i X_k}^2)(1 - R_{X_j X_k}^2)}}, \quad i \neq j \neq k = 1, 2, 3,$$

$$R_{X_1 X_2 / X_3} = \frac{R_{X_1 X_2} - R_{X_1 X_3} R_{X_2 X_3}}{\sqrt{(1 - R_{X_1 X_3}^2)(1 - R_{X_2 X_3}^2)}} = -0,0833,$$

$$R_{X_1 X_3 / X_2} = \frac{R_{X_1 X_3} - R_{X_1 X_2} R_{X_2 X_3}}{\sqrt{(1 - R_{X_1 X_2}^2)(1 - R_{X_2 X_3}^2)}} = 0,4705,$$

$$R_{X_2 X_3 / X_1} = \frac{R_{X_2 X_3} - R_{X_1 X_2} R_{X_1 X_3}}{\sqrt{(1 - R_{X_1 X_2}^2)(1 - R_{X_1 X_3}^2)}} = 0,9009.$$

Обчислимо точкові оцінки залишкових дисперсій (при двох фіксованих змінних):

$$S_{X_i/X_jX_k}^2 = S_{X_i/X_k}^2 (1 - R_{X_iX_j/X_k}^2), \quad i \neq j \neq k = 1, 2, 3,$$

$$S_{X_1/X_2X_3}^2 = S_{X_1/X_3}^2 (1 - R_{X_1X_2/X_3}^2) = 0,0066,$$

$$S_{X_2/X_1X_3}^2 = S_{X_2/X_3}^2 (1 - R_{X_1X_2/X_3}^2) = 0,0013,$$

$$S_{X_3/X_1X_2}^2 = S_{X_3/X_2}^2 (1 - R_{X_1X_3/X_2}^2) = 0,0011.$$

Отримаємо оцінки множинних коефіцієнтів детермінації та кореляції за формулою:

$$R_{X_i}^2 = 1 - \frac{S_{X_i/X_jX_k}^2}{S_{X_i}^2}, \quad i \neq j \neq k = 1, 2, 3,$$

$$R_{X_1}^2 = 1 - \frac{S_{X_1/X_2X_3}^2}{S_{X_1}^2} = 0,8370, \quad R_{X_1} = 0,9149,$$

$$R_{X_2}^2 = 1 - \frac{S_{X_2/X_1X_3}^2}{S_{X_2}^2} = 0,9606, \quad R_{X_2} = 0,9801,$$

$$R_{X_3}^2 = 1 - \frac{S_{X_3/X_1X_2}^2}{S_{X_3}^2} = 0,9691, \quad R_{X_3} = 0,9844.$$

Перевіримо з рівнем $\alpha = 0,05$ та відповідними степенями вільності значущість множинних коефіцієнтів кореляції. Обчислимо значення F -критерію, що спостерігається:

$$F_{\text{розр}}(R_{X_i}^2) = \frac{R_{X_i}^2/2}{(1 - R_{X_i}^2)/(n-3)}, \quad i = 1, 2, 3,$$

$$F_{\text{розр}}(R_{X_1}^2) = \frac{R_{X_1}^2/2}{(1 - R_{X_1}^2)/(n-3)} = 177,2115,$$

$$F_{\text{розр}}(R_{X_2}^2) = \frac{R_{X_2}^2/2}{(1 - R_{X_2}^2)/(n-3)} = 840,6062,$$

$$F_{\text{розр}}(R_{X_3}^2) = \frac{R_{X_3}^2/2}{(1 - R_{X_3}^2)/(n-3)} = 1081,6120.$$

Тепер за таблицею F -розподілу знаходимо критичне значення F -статистики для рівня значущості $\alpha = 0,05$, числа степенів вільності чисельника $\nu_1 = 2$ та знаменника $\nu_2 = n - 3 = 69$: $F_{\text{кр}}(0,05; 2; 69) = 3,13$.

Виходячи з того, що усі значення F -статистик, що спостерігаються, перевищують критичне значення, можна сказати, що гіпотеза про рівність нулю кожного коефіцієнта множинної кореляції глобальної вибірки відкидається з ймовірністю помилки $\alpha = 0,05$.

Перевіримо значущість коефіцієнтів кореляції з рівнем значущості $\alpha = 0,05$. За таблицею розподілу r -статистики для числа степенів вільності $\alpha = n - 3 = 69$ та $\alpha = 0,05$ знаходимо:

$$R_{\text{кр}}(0,05; 70) < R_{\text{кр}}(0,05; 69) < R_{\text{кр}}(0,05; 65)$$

$$\text{або } 0,1954 < R_{\text{кр}}(0,05; 69) < 0,2027.$$

Унаслідок того, що значення $|R_{X_1X_3/X_2}|$ та $|R_{X_2X_3/X_1}|$, що спостерігаються, більші за $R_{\text{кр}}(0,05; 65)$, то вони будуть більші й за $R_{\text{кр}}(0,05; 70)$. Отже, гіпотезу про рівність нулю глобальних окремих коефіцієнтів кореляції $R_{X_1X_3/X_2}$ та $R_{X_2X_3/X_1}$ можна відкинути з ймовірністю помилки $0,05$. Значення r -статистики для окремого коефіцієнта $|R_{X_1X_2/X_3}|$ менше за $R_{\text{кр}}(0,05; 70)$ та, відповідно, за $R_{\text{кр}}(0,05; 65)$. Тому гіпотезу $H_0: \rho_{X_1X_2/X_3} = 0$ відкинути не можна. Прийнемо цю гіпотезу, тобто вважатимемо, що глобальний окремий коефіцієнт кореляції між X_1 та X_3 дорівнює нулю, тобто не є значущим. Отже, значущими виявилися тільки два окремих коефіцієнти кореляції $\rho_{X_1X_3/X_2}$ та $\rho_{X_2X_3/X_1}$.

З вірогідністю $\gamma = 1 - \alpha = 0,95$ знайдемо інтервальні оцінки значущих окремих коефіцієнтів кореляції. Для $\rho_{X_1X_3/X_2}$ отримаємо за таблицею функцій $Z_R = \frac{1}{2} \ln \frac{1+R}{1-R}$ значення: $Z_{R_{X_1X_3/X_2}} = Z_{0,4705} = 0,5107$. Потім обчислимо точність інтервальної оцінки для ΔZ :

$$\Delta Z = t_{0,95} \frac{1}{\sqrt{n-4}}.$$

Значення $t_{0,95}$ знаходиться за таблицею Лапласа: $\Delta Z = 0,2021$. Інтервальна оцінка для ΔZ :

$$0,5107 - 0,2022 \leq \Delta Z \leq 0,5107 + 0,2022$$

або

$$0,3086 \leq \Delta Z \leq 0,7129.$$

Щоб отримати інтервальну оцінку для $p_{X_1 X_3 / X_2}$ за таблицями Z-перетворення Фішера, виконаємо зворотний розрахунок:

$$0,299 \leq p_{X_1 X_3 / X_2} \leq 0,612.$$

Для $p_{X_2 X_3 / X_1}$ отримаємо $Z_{R_{X_2 X_3 / X_1}} = Z_{0,9009} = 1,4770$. Тоді

$$1,4770 - 0,2022 \leq \Delta Z \leq 1,4770 + 0,2022$$

або $1,2649 \leq \Delta Z \leq 1,6792$.

Інтервальна оцінка для $p_{X_2 X_3 / X_1}$ складатиме

$$0,852 \leq p_{X_2 X_3 / X_1} \leq 0,932.$$

Означимо через p_X^2 множинний коефіцієнт детермінації з більшою статистикою. Йому відповідають обидва значущі часткові коефіцієнти детермінації $p_{X_1 X_3 / X_2}^2$ та $p_{X_2 X_3 / X_1}^2$, які мають й оцінки, більші за $p_{X_1 X_2 / X_3}^2$.

Оцінка відповідного рівня регресії набуде вигляду:

$$\overline{X_3 / X_1 X_2} - \overline{X_3} = b_{X_3 X_1 / X_2} (X_1 - \overline{X_1}) + b_{X_3 X_2 / X_1} (X_2 - \overline{X_2}).$$

Обчислимо коефіцієнти окремої множинної регресії:

$$b_{X_3 X_1 / X_2} = r_{X_1 X_3 / X_2} \frac{S_{X_3 / X_2}}{S_{X_1 / X_2}} = 0,1932,$$

$$b_{X_3 X_2 / X_1} = r_{X_2 X_3 / X_1} \frac{S_{X_3 / X_1}}{S_{X_2 / X_1}} = 0,8488.$$

Таким чином, отримуємо:

$$\overline{X_3 / X_1 X_2} - 0,5195 = 0,1932(X_1 - 0,5062) + 0,8488(X_2 - 0,4560).$$

Знайдемо з вірогідністю $\gamma = 1 - \alpha = 0,95$ інтервальні оцінки множинних коефіцієнтів регресії. За таблицею розподілу Стюдента знайдемо $t(\alpha, n - 3) = t(0,05, 69) = 1,667$ і обчислимо відносно β нерівності:

$$\left| \frac{(0,1932 - \beta_{X_3 X_1 / X_2}) \cdot 0,0378 \cdot \sqrt{69}}{0,0921 \cdot \sqrt{1 - 0,47051^2}} \right| \leq 1,667,$$

$$\left| \frac{(0,8488 - \beta_{X_3 X_2 / X_1}) \cdot 0,0816 \cdot \sqrt{69}}{0,0769 \cdot \sqrt{1 - 0,90091^2}} \right| \leq 1,667.$$

Обчислимо точність інтервальних оцінок:

$$\Delta b_{X_3 X_1 / X_2} = 1,667 \cdot \frac{0,0921 \cdot \sqrt{1 - 0,47051^2}}{0,0378 \cdot \sqrt{69}} = 0,4313,$$

$$\Delta b_{X_3 X_2 / X_1} = 1,667 \cdot \frac{0,0769 \cdot \sqrt{1 - 0,90091^2}}{0,0816 \cdot \sqrt{69}} = 0,0821.$$

Звідки

$$0,1932 - 0,4313 \leq \beta_{X_3 X_1 / X_2} \leq 0,1932 + 0,4313,$$

$$0,8488 - 0,0821 \leq \beta_{X_3 X_2 / X_1} \leq 0,8488 + 0,0821,$$

$$-0,2381 \leq \beta_{X_3 X_1 / X_2} \leq 0,6245, \quad 0,8027 \leq \beta_{X_3 X_2 / X_1} \leq 0,9669.$$

Отримаємо покращені точкові оцінки дослідження коефіцієнтів детермінації:

$$R_{X_1 X_3 / X_2}^2 = \frac{(n-2) \cdot R_{X_1 X_3 / X_2}^2 - 1}{n-3} = 0,2101,$$

$$R_{X_2 X_3 / X_1}^2 = \frac{(n-2) \cdot R_{X_2 X_3 / X_1}^2 - 1}{n-3} = 0,8089,$$

$$R_{X_3}^2 = \frac{(n-1) \cdot R_{X_3}^2 - 1}{n-2} = 0,9686.$$

3.4. Застосування кореляційного аналізу для оцінки показників ефективності роботи підприємств

З метою аналізу взаємозв'язку показників ефективності виробництва продукції (продуктивність праці, фондовіддача та матеріалоемність продукції) була відібрана група з 25 однотипних машинобудівних підприємств [8].

На основі річних звітів підприємств були отримані такі дані (табл. 3.14): x — виробіток валової продукції у незмінних цінах

на одного працюючого середньої реєстрової чисельності ППП (млн. грн); y — випуск валової продукції на одну гривню середньорічної вартості основних промислово-виробничних фондів (грн); z — матеріалоемність у вартісному вираженні: вартість матеріалів для валової продукції у незмінених цінах (%).

Таблиця 3.14

№ підприємства	x	y	z	№ підприємства	x	y	z
1	6,0	2,0	25	14	5,7	2,2	25
2	4,9	0,8	30	15	5,1	1,3	30
3	7,0	2,7	20	16	5,2	1,5	14
4	6,7	3,0	21	17	7,3	2,7	20
5	5,8	1,0	28	18	6,1	2,4	27
6	6,1	2,1	26	19	6,2	2,2	28
7	5,0	0,9	30	20	5,9	2,0	26
8	6,9	2,6	22	21	6,0	2,0	26
9	6,8	3,0	20	22	4,8	0,9	31
10	5,9	1,1	29	23	7,3	3,2	19
11	5,0	0,8	27	24	7,2	3,3	20
12	5,6	2,2	25	25	7,0	3,0	20
13	6,0	2,4	24				

Припустимо, що ознаки x , y та z , що розглядаються, у генеральній сукупності підкоряються нормальному закону розподілу та вказані дані являють собою вибірку з цієї сукупності.

Щоб отримати точкові оцінки генеральних середніх, дисперсій середніх квадратичних відхилень та парних коефіцієнтів кореляції, результати проміжних обчислень слід помістити у розрахункову табл. 3.15. Для полегшення контролю обчислення дані розбито на п'ятірки. Для кожної п'ятірки в підсумковому (виділеному) рядку наведені суми елементів кожного стовпчика. В останньому контрольному стовпчику наводяться суми елементів відповідних рядків. Сума п'яти елементів контрольного стовпчика має співпадати з сумою елементів підсумкового рядка.

Таблиця 3.15

x	y	z	x^2	y^2	z^2	xy	xz	yz	Контрольна сума
6,0	2,0	25	36,00	9,00	625	12,00	150,0	50,0	910,00
4,9	0,8	30	24,01	0,64	900	3,92	147,0	24,0	1135,27
7,0	2,7	20	49,00	7,29	400	18,90	140,0	54,0	698,89

Закінчення табл. 3.15

x	y	z	x^2	y^2	z^2	xy	xz	yz	Контроль- на сума
6,7	3,0	21	44,89	9,00	441	20,10	140,7	63,0	749,39
5,8	1,0	28	33,64	1,00	784	5,80	162,4	28,0	1049,64
30,4	9,5	124	187,54	21,93	3150	60,72	740,1	219,0	4543,19
6,1	2,1	26	37,21	4,41	676	12,81	158,6	54,6	977,83
5,0	0,9	30	25,00	0,81	900	4,50	150,0	27,0	1143,21
6,9	2,6	22	47,61	6,76	484	17,94	151,8	57,2	796,81
6,8	3,0	20	46,21	9,00	400	20,40	136,0	60,0	701,44
5,9	1,1	29	34,81	1,21	841	6,49	171,1	31,9	1122,51
30,7	9,7	127	190,87	22,19	3301	62,14	767,5	230,7	4741,80
5,0	0,8	27	25,00	0,64	729	4,00	135,0	21,6	948,04
5,6	2,2	25	31,36	4,84	625	12,32	140,0	55,0	301,32
6,0	2,4	24	36,00	5,76	576	14,40	144,0	57,6	866,16
5,7	2,2	25	32,49	4,84	625	12,54	142,5	55,0	905,27
5,1	1,3	30	26,01	1,69	900	6,63	153,0	39,0	1162,73
27,4	8,9	131	150,86	17,77	3455	49,89	714,5	228,2	4783,52
5,2	1,5	24	27,04	2,25	576	7,80	124,8	36,0	804,59
7,3	2,7	20	53,29	7,29	400	19,71	146,0	54,0	710,29
6,1	2,4	27	37,21	5,76	729	14,64	164,7	64,8	1051,61
6,2	2,2	28	38,44	4,84	784	13,64	173,6	61,6	1112,52
5,9	2,0	26	34,81	4,00	676	11,80	153,4	52,0	965,91
30,7	10,8	125	190,79	24,14	3165	67,59	762,5	268,4	4644,92
6,0	2,0	26	36,00	4,00	676	12,00	156,0	52,0	970,00
4,8	0,9	31	23,04	0,81	961	4,32	148,8	27,9	1202,57
7,3	3,2	19	53,29	10,24	361	23,36	138,7	60,8	676,89
7,2	3,3	20	51,84	10,89	400	23,76	144,0	66,0	726,99
7,0	3,0	20	43,00	9,00	400	21,00	140,0	60,0	709,00
32,3	12,4	116	213,17	34,94	2798	84,44	727,5	266,7	4285,45

Далі отримуємо таблицю підсумкових рядків (табл. 3.16).

Таблиця 3.16

1	2	3	4	5	6	7	8	9	10
30,4	9,5	124	187,54	21,93	3150	60,72	740,1	219,0	4543,19
30,7	9,7	127	190,87	22,19	3301	62,14	767,5	230,7	4741,80
27,4	8,9	131	150,86	17,77	3455	49,89	714,5	228,2	4783,52
30,7	10,8	125	190,79	24,14	3165	67,59	762,5	268,4	4644,92
32,3	12,4	116	213,17	34,94	2798	84,44	727,5	266,7	4285,45
151,5	51,3	623	933,23	120,97	15869	324,78	3712,1	1213,0	22798,88

З останнього рядка підсумкової таблиці отримуємо:

$$\sum x = 151,5, \quad \sum y = 51,3, \quad \sum z = 623;$$

$$\sum x^2 = 933,23, \quad \sum y^2 = 120,97, \quad \sum z^2 = 15869;$$

$$\sum xy = 324,78, \quad \sum xz = 3712,1, \quad \sum yz = 1213,0.$$

Отже,

$$\bar{x} = \frac{1}{25} \cdot 151,5 = 6,06, \quad \bar{x}^2 = \frac{1}{25} \cdot 933,23 = 37,3292;$$

$$\bar{y} = \frac{1}{25} \cdot 51,3 = 2,052, \quad \bar{y}^2 = \frac{1}{25} \cdot 120,97 = 4,8388;$$

$$\bar{z} = \frac{1}{25} \cdot 623 = 24,92, \quad \bar{z}^2 = \frac{1}{25} \cdot 15869 = 634,76;$$

$$S_x^2 = x^2 - (\bar{x})^2 = 37,3292 - 6,06^2 = 0,6056, \quad S_x = 0,778203;$$

$$S_y^2 = 4,8388 - 2,052^2 = 0,628096, \quad S_y = 0,792525;$$

$$S_z^2 = 634,76 - 24,92^2 = 13,7536, \quad S_z = 3,7085846.$$

Щоб обчислити точкові оцінки парних коефіцієнтів кореляції, використовуємо формулу:

$$r_{xy} = \frac{\sum xy - (\sum x \sum y)/n}{\left\{ \left[\sum x^2 - (\sum x)^2/n \right] \left[\sum y^2 - (\sum y)^2/n \right] \right\}^{1/2}}.$$

Отримуємо

$$r_{xy} = \frac{324,78 - 151,5 \cdot 51,3/25}{[(933,23 - 151,5^2/25)(120,97 - 51,3^2/25)]^{1/2}} = 0,9016362;$$

$$r_{xz} = \frac{3712,1 - 151,5 \cdot 623/25}{[15,14(15869 - 623^2/25)]^{1/2}} = \frac{-63,28}{\sqrt{15,14 \cdot 343,84}} = -0,8770519;$$

$$r_{yz} = \frac{1213,0 - 51,3 \cdot 623/25}{\sqrt{15,7024 \cdot 343,84}} = -0,8899999.$$

Обчислюємо точкові оцінки умовних середніх квадратичних відхилень (при фіксуванні однієї змінної):

$$S_{x/y} = \sqrt{S_x^2(1 - r_{xy}^2)} = \sqrt{0,6056(1 - 0,9016362^2)} = 0,3365691;$$

$$S_{y/x} = \sqrt{S_y^2(1 - r_{xy}^2)} = \sqrt{0,628096 \cdot 0,1870522} = 0,3427633;$$

$$S_{x/z} = 0,3738452, \quad S_{y/z} = 0,3613602;$$

$$S_{z/x} = 1,7815879, \quad S_{z/y} = 1,6909689.$$

Отримуємо точкові оцінки окремих коефіцієнтів кореляції:

$$r_{xy/z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}} = \frac{0,9016362 - 0,8770519 \cdot 0,8899999}{\sqrt{0,23078 \cdot 0,2079002}} = 0,5526811;$$

$$r_{xz/y} = \frac{-0,8770519 + 0,9016362 \cdot 0,8899999}{\sqrt{0,1870552 \cdot 0,2079002}} = -0,3782736;$$

$$r_{yz/x} = \frac{-0,8899999 + 0,9016362 \cdot 0,8770519}{\sqrt{0,1870552 \cdot 0,23078}} = -0,4775413.$$

Обчислюємо точкові оцінки залишкових дисперсій (при фіксуванні двох змінних):

$$S_{x/yz}^2 = S_{x/z}^2(1 - r_{xy/z}^2) = 0,3738452^2(1 - 0,5526811^2) = 0,0970695;$$

$$S_{y/xz}^2 = S_{y/z}^2(1 - r_{xy/z}^2) = 0,0906942, \quad S_{z/xy}^2 = S_{z/x}^2(1 - r_{yz/x}^2) = 2,450226.$$

Отримуємо оцінки множинних коефіцієнтів детермінації та кореляції:

$$r_x^2 = 1 - \frac{S_{x/yz}^2}{S_x^2} = 1 - \frac{0,0970695}{0,6056} = 0,8397136, \quad r_x = 0,9163588;$$

$$r_y^2 = 1 - \frac{S_{y/xz}^2}{S_y^2} = 1 - \frac{0,0906942}{0,628096} = 0,8556046, \quad r_y = 0,9249889;$$

$$r_z^2 = 1 - \frac{S_{z/xy}^2}{S_z^2} = 1 - \frac{2,450226}{13,7536} = 0,8218484, \quad r_z = 0,9065585.$$

Для контролю обчислювань корисно скористатися іншими формулами точкових оцінок множинних коефіцієнтів кореляції, наприклад,

$$r_x = \sqrt{\frac{r_{xy}^2 + r_{xz}^2 - 2r_{xy}r_{xz}r_{yz}}{1 - r_{yz}^2}} = \sqrt{\frac{0,1745766}{0,2079002}} = \sqrt{0,8397134} = 0,9163587.$$

Помилка в останньому розряді на одиницю є припустимою. Перевіримо з рівнем $\alpha = 0,05$ значущість множинних коефіцієнтів кореляції. Обчислимо значення F -критерію, що спостерігаються:

$$F_{\text{спост}}(r_x^2) = \frac{r_x^2/2}{(1 - r_x^2)/(n - 3)} = \frac{0,8397136 \cdot 22}{2 \cdot 0,1602864} = 57,627157;$$

$$F_{\text{спост}}(r_y^2) = \frac{0,8556046 \cdot 22}{2 \cdot 0,1443954} = 65,179712;$$

$$F_{\text{спост}}(r_z^2) = \frac{0,8218484 \cdot 22}{2 \cdot 0,1781516} = 50,745165.$$

Знаходимо тепер за таблицею F -розподілу критичне значення F -статистик для рівня значущості $\alpha = 0,05$, числа степенів вільності чисельника $v_1 = 2$ та знаменника $v_2 = n - 3 = 22$. Воно складає $F_{\text{кр}}(0,05; 2; 22) = 3,44$.

Оскільки всі значення F -статистик, що спостерігаються, перевищують критичне, то гіпотеза про рівність нулю кожного множинного коефіцієнта кореляції генеральної сукупності відкидається з ймовірністю помилки, яка дорівнює $0,05$. Отже, ці коефіцієнти значуще відрізняються від нуля.

Перевіримо значущість окремих коефіцієнтів кореляції, для яких $\alpha = 0,05$. За таблицею розподілу r -статистики для числа степенів вільності $v = n - 3 = 25 - 3 = 22$ та $\alpha = 0,05$ знаходимо:

$$r_{\text{кр}}(0,05; 25) < r_{\text{кр}}(0,05; 22) < r_{\text{кр}}(0,05; 20)$$

$$\text{або } 0,381 < r_{\text{кр}}(0,05; 22) < 0,423.$$

Оскільки значення $|r|$, $|r_{xy/z}|$ та $|r_{yz/x}|$, що спостерігаються, більші за $0,423$, то вони перевищуватимуть і $r_{\text{кр}}(0,05, 22)$. Отже, гіпотеза про рівність нулю генеральних окремих коефіцієнтів кореляції $r_{xy/z}$ та $r_{yz/x}$ відкидається з ймовірністю помилки $0,05$. Значення r -статистики, що спостерігається, для окремого коефіцієнта кореляції між x та y менше за $0,381$, тобто й менше,

ніж $r_{кр}(0,05, 22)$. Отже, гіпотеза $H_0: p_{xz/y} = 0$ не відкидається. Прийнемо цю гіпотезу, тобто будемо вважати, що генеральний окремий коефіцієнт кореляції між x та z дорівнює нулю, тобто не є значущим.

Значущими виявились лише два окремих коефіцієнти кореляції — $p_{xy/z}$ та $p_{yz/x}$.

Зауважимо, що при необхідності можна було б обчислити $r_{кр}(0,05; 22)$, наприклад, за допомогою лінійної інтерполяції:

$$r_{кр}(0,05; 22) = 0,381 + \frac{0,423 - 0,381}{25 - 20}(25 - 22) = 0,4062.$$

З надійністю $\gamma = 1 - \alpha = 0,95$ знайдемо інтервальні оцінки значущих окремих коефіцієнтів кореляції.

Для $p_{xy/z}$ отримуємо за таблицею функції $Z_r = \frac{1}{2} \ln \frac{1+r}{1-r}$ значення $Z_{0,55} = 0,6184$. Потім обчислюємо точність інтервальної оцінки для MZ :

$$\Delta Z = t_{0,95} \frac{1}{\sqrt{n-4}} = 1,96 \cdot \frac{1}{\sqrt{21}} = 0,4277$$

($t_{0,95}$ знаходиться за таблицею Лапласа). Інтервальною оцінкою для MZ є:

$$0,6184 - 0,4277 \leq MZ \leq 0,6184 + 0,4277$$

або $0,1907 \leq MZ \leq 1,0461$.

Щоб отримати інтервальну оцінку для $p_{xy/z}$ за таблицями Z_r , знайдемо числа, найближчі до 0,1907 або до 1,0461: $Z_{r_{\min}} = 0,1923$ та $Z_{r_{\max}} = 1,0454$. Переходячи до аргументу r , отримаємо

$$0,19 \leq p_{xy/z} \leq 0,78.$$

Для $p_{yz/x}$ маємо $Z_{-0,48} = -Z_{0,48} = -0,5230$. Тоді

$$-0,5230 - 0,4277 \leq MZ \leq -0,5230 + 0,4277$$

або $-0,9507 \leq MZ \leq 0,0953$.

Найближчими до чисел 0,0953 та 0,9507 будуть 0,1003 та 0,9505. Отже, інтервальною оцінкою для $p_{yz/x}$ є

$$-0,74 \leq p_{yz/x} \leq -0,10.$$

Для трьох ознак, що вивчаються, немає однобічного причинно-наслідкового зв'язку. Тому оберемо у якості моделі взаємозалежності евристично найбільш надійну в умовах цієї вибірки. Значущим множинним коефіцієнтом детермінації, який має найбільшу статистику, є p_y^2 . Йому відповідають обидва значущі часткові коефіцієнти детермінації $p_{xy/z}^2$ та $p_{yz/x}^2$ (які мають й оцінки більші, ніж оцінка $p_{xz/y}^2$).

Оцінка відповідного рівняння регресії має вигляд:

$$\overline{y/(x, z)} - \bar{y} = b_{yx/z}(x - \bar{x}) + b_{yz/x}(z - \bar{z}).$$

Обчислимо коефіцієнти множинної (часткової) регресії:

$$b_{yx/z} = r_{xy/z} \frac{S_{y/z}}{S_{x/z}} = 0,5526811 \cdot \frac{0,3613601}{0,3738452} = 0,5342236;$$

$$b_{yz/x} = r_{yz/x} \frac{S_{y/x}}{S_{z/x}} = -0,4775413 \cdot \frac{0,3427633}{1,7815879} = -0,0918751.$$

Таким чином, отримаємо

$$\overline{y/(x, z)} - 2,052 = 0,5342(x - 6,06) - 0,091875(z - 24,92).$$

Знайдемо з надійністю $\gamma = 1 - \alpha = 0,95$ інтервальні оцінки для множинних коефіцієнтів регресії. За таблицею розподілу Стюдента знаходимо $t(\alpha, n - 3) = t(0,05, 22) = 2,074$ та розв'язуємо відносно β нерівності

$$\left| \frac{(0,5342236 - \beta_{yx/z})0,3738452\sqrt{22}}{0,3613602\sqrt{1 - 0,5526811^2}} \right| \leq 2,074;$$

$$\left| \frac{(-0,0918751 - \beta_{yz/x})1,7815879\sqrt{22}}{0,3427633\sqrt{1 - 0,4775413^2}} \right| \leq 2,074.$$

Обчислимо точності інтервальних оцінок:

$$\Delta b_{yx/z} = 2,074 \cdot \frac{0,3613602\sqrt{1 - 0,5526811^2}}{0,3738452\sqrt{22}} = 0,3562008;$$

$$\Delta b_{yz/x} = 2,074 \cdot \frac{0,3427633\sqrt{1 - 0,4775413^2}}{1,7815879\sqrt{22}} = 0,0747445.$$

Звідки

$$0,5342236 - 0,3562008 \leq \beta_{yx/z} \leq 0,5342236 + 0,3562008;$$
$$-0,0918751 - 0,0747445 \leq \beta_{yz/x} \leq -0,0918751 + 0,0747445;$$

$$0,1780 \leq \beta_{yx/z} \leq 0,9804;$$

$$-0,0171306 \leq \beta_{yz/x} \leq -0,1666196.$$

Отримаємо найкращі точкові оцінки коефіцієнтів детермінації, що вивчаються:

$$\hat{p}_{xy/z}^2 = \frac{(n-2)r_{yx/z}^2 - 1}{n-3} = \frac{23 \cdot 0,5526811^2 - 1}{22} = 0,2738861;$$

$$\hat{p}_{yz/x}^2 = \frac{23 \cdot 0,4775413^2 - 1}{22} = 0,1929567;$$

$$\hat{p}_y^2 = \frac{(n-1)r_y^2 - 2}{n-3} = \frac{24 \cdot 0,8556046 - 2}{22} = 0,8424777.$$

Спираючись на отримані розрахунки, можна зробити такі висновки:

☑ доведений цільний взаємозв'язок кожного з показників ефективності роботи підприємства, що досліджується, з іншими показниками (всі множинні коефіцієнти детермінації значущі та перевищують 0,8);

☑ особливо цільний зв'язок існує між фондovіддачею та двома іншими показниками — продуктивністю праці та матеріалоемністю. Зміна фондovіддачі в середньому на 84,25 % зумовлена зміною продуктивності праці та матеріалоемності (зміна фондovіддачі у середньому на 15,75 % зумовлена впливом неконтрольованих факторів, ознак). При збільшенні продуктивності праці на 1 млн. грн фондovіддача збільшується в середньому на 0,55 грн на 1 грн основних виробничих фондів. При зменшенні матеріалоемності на 1 % фондovіддача збільшується в середньому на 0,48 %. Зазначені нормативи відносно стабільні за умови, що показники, які вивчаються, незначно відхиляються від своїх середніх рівнів (стабільність вказується довірчими інтервалами та ймовірністю 0,95). Взаємозалежність між матеріалоемністю та продуктивністю праці (без урахування фондovіддачі) не доведена (частковий коефіцієнт кореляції $r_{xz/y}$ не є значущим) за цих умов. Для більш надійної перевірки такої залежності необхідний великий обсяг вибірки.

Контрольні завдання і вправи

1. Побудуйте діаграми розсіювання, користуючись даними, наведеними в табл. 1.1.

2. Що являє собою коваріація між X та Y ?

3. Що легше інтерпретувати, коваріацію чи кореляцію? Чому?

4. Якщо більші значення X викликають появу більших значень Y , то якою, на вашу думку, має бути кореляція — додатною, від'ємною або нульовою? Чому?

5. Чи свідчить наявність сильної додатної кореляції про те, що більші значення X викликають появу більших значень Y ? Якщо ні, то які ще можливі варіанти?

6. Знайдіть та інтерпретуйте стандартну помилку оцінки коефіцієнта регресії, використовуючи дані табл. 1.1 (X_2, Y).

7. Чи є ця модель значущою? Про що це свідчить?

8. Знайдіть та інтерпретуйте стандартизовані коефіцієнти регресії, використовуючи дані табл. 1.1 (X_3, Y).

9. Використовуючи дані табл. 3.14, побудуйте стандартизоване рівняння регресії.

РОЗДІЛ 4

РЕГРЕСІЙНИЙ АНАЛІЗ І ЙОГО ВИКОРИСТАННЯ У МАТЕМАТИЧНОМУ ЗАБЕЗПЕЧЕННІ ІНФОРМАЦІЙНО-УПРАВЛЯЮЧИХ СИСТЕМ

4.1. Поняття регресії

Регресійний аналіз відноситься до статистичних методів аналізу та обробки даних і широко застосовується у багатьох предметних областях. Найбільшого поширення він набув у побудові математичних залежностей на основі експериментальних даних.

Історія розвитку регресійного аналізу починається з 1805 року, коли Андріаном Марі Лежандром було розроблено метод найменших квадратів. Поняття регресії було введено у 1885 році Франсісом Гальтоном у роботі «Природна спадковість», в якій були систематизовані дослідження з кореляційного та регресійного аналізу. Термін «регресія» (від лат. *regression* — відступ, повернення до будь-чого) пов'язаний лише зі специфікою одного з перших конкретних прикладів, у якому це поняття було використано Ф. Гальтоном. Він перший використав метод найменших квадратів для оброблення статистичних даних. Подальші дослідження були проведені К. Пірсоном, який 1897 року розробив коефіцієнт кореляції. У наукових працях Джорджа А. Юла (1897 та 1907 рр.) закладено основи теорії часткової кореляції та лінійної регресії для будь-якої кількості вхідних змінних. Запро-

понований Дж. А. Юлом метод знайшов широке застосування майже в усіх галузях людської діяльності.

Регресійний аналіз (РА) застосовується у різних предметних галузях, коли одна спостережувана змінна має очікуване значення, яке залежить від значень інших змінних. Регресійний зв'язок — це зв'язок між змінними, у якому очікуване значення однієї випадкової (залежної, прогнозованої, вихідної, регресійної) змінної співвідноситься з іншими змінними (що називаються незалежними, вхідними змінними, предикторами, регресорами, аргументами регресії). Аргументи регресії, або незалежні змінні, можуть бути як випадковими, так і не випадковими. У разі випадковості незалежних змінних регресія визначається як умовна відносно спостережуваних значень її випадкових аргументів.

З огляду на наявність великого розмаїття публікацій з регресійного аналізу, включення його у численні пакети з обробки даних, достатньо довгу історію, а також застосовність до більшості предметних галузей, можна припустити, що як метод він цілком розроблений і в науковому плані не є цікавим. Проте при глибокому вивченні методу, за умови аналізу припущень та передумов, на яких він ґрунтується, виникає велика кількість проблем, пов'язаних з коректністю використання статистичних критеріїв.

Формування основних напрямків розвитку регресійного аналізу сьогодні зумовлене труднощами, які виникають через порушення передумов та припущень регресійного аналізу, і пов'язане з розробленням модифікованих методів, які долали б ці труднощі. Більш широко розвивається аналіз залишків, удосконалюються обчислювальні аспекти РА, формалізуються процедури вибору структури моделей, набуло розвитку застосування методу для корельованих та неоднорідних спостережень. Розроблюються та вдосконалюються методи РА, що мають застосовуватися за наявності помилок у рівнях факторів, що задаються, корельованості незалежних змінних, залежності майбутніх значень від передісторії, а також порушень припущень про нормальність закону розподілів остач. Крім цього, розроблюється цілий клас методів, який отримав назву стійких, або робастних, оцінок. В ІУС регресійний аналіз застосовується для побудови статичних моделей та моделей прогнозування, управління, а також для дослідження впливу різних факторів на процес або об'єкт.

Мета написання розділу полягає у викладенні найновіших досягнень у регресійному аналізі. З огляду на те, що формальне

використання пакетів прикладних програм (ППП) для побудови математичних моделей часто призводить до помилкових результатів та помилкової інтерпретації, у розділі на інженерному рівні подано методику коректного використання ППП для побудови математичних моделей у різних предметних галузях. Виклад методу використання ППП для побудови математичних моделей за допомогою регресійного аналізу супроводжується численними прикладами.

4.2. Побудова математичних моделей в ІУС з використанням регресійного аналізу

Регресійний аналіз в ІУС використовується для побудови моделей статичних та динамічних об'єктів. *Статичними* будемо називати об'єкти, в яких відсутня «пам'ять». Реакція виходу об'єкта на вхідні фактори миттєва й не залежить від передісторії. *Динамічні* об'єкти характеризуються тим, що вихід об'єкта залежить від передісторії.

Статичні моделі встановлюють залежність між вхідними та вихідними змінними у стані об'єкта, що встановився. Статичні моделі застосовуються:

- ☑ для синтезу систем автоматичного управління, зокрема для вибору структури системи управління, оптимальних налаштувань регуляторів, а також при проектуванні нових та вдосконаленні діючих промислових об'єктів, установок тощо;

- ☑ у задачах оптимального управління у статичі;

- ☑ для більш глибокого розуміння процесів управління.

Задача знаходження статичної моделі (характеристики) зводиться до апроксимації функцій багатьох змінних на деяких класах функцій, які визначаються досліджуванним об'єктом управління. Слід зазначити, що для апроксимації статичних характеристик бажано використовувати клас лінійних залежностей. По суті, немає реальних об'єктів, які б мали лінійні статистичні характеристики, але значення цього класу велике, оскільки:

- ☑ існує багато об'єктів, що мають характеристики, близькі до лінійних;

- ☑ статичні характеристики можна розглядати не в цілому діапазоні функціонування об'єкта, а в деякій периферії цього режиму та в ній апроксимувати нелінійну статичну харак-

теристичку лінійної. Сьогодні широкого розповсюдження для побудови статичних залежностей набув регресійний та кореляційний аналіз.

При побудуванні моделей для динамічних об'єктів з використанням регресійного аналізу вхідні та вихідні змінні враховуються з передісторією. У цьому випадку має місце автокореляція між помилками моделі та оцінки моделі є неефективними.

4.2.1. Теоретичні основи прикладного регресійного аналізу

Регресійний аналіз застосовується для побудування математичних залежностей об'єктів, явищ за результатами експериментальних даних, отриманих на основі проведення активного або пасивного експериментів.

Припустимо, що математична залежність буде належати до певного класу функцій з декількома невідомими параметрами. У загальному випадку такі функції мають вигляд:

$$Y = F(X, B, E),$$

де $Y = \{y_1, y_2, \dots, y_n\}$ — вектор залежної (вихідної) змінної розмірністю $n \times 1$; X — матриця незалежних (вхідних) змінних розмірністю $n \times p$; B — вектор невідомих параметрів розмірністю $p \times 1$; E — вектор збуджень розмірністю $n \times 1$; P — кількість незалежних змінних; n — кількість експериментальних даних; F — клас функціональних залежностей.

Значення Y є випадковою величиною, значення X можуть розглядатися або як фіксовані, або як випадкові. При цьому очікуване значення однієї випадкової змінної співвідноситься зі значеннями інших спостережуваних випадкових змінних у вигляді умовної регресії.

Розглянемо залежність між випадковими величинами X та Y , яка подана у вигляді деякої таблиці спостережень значень X та Y . Переносячи табличні значення X та Y на площину XY , отримуємо поле кореляції (див. рис. 4.1).

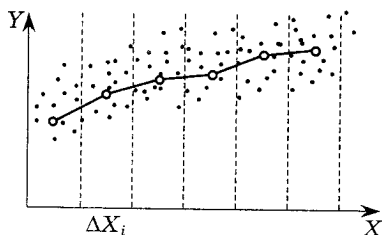


Рис. 4.1. Експериментальне рівняння регресії

Розіб'ємо діапазон зміни x на m рівних інтервалах Δx . Усі точки, що потрапили до інтервалу Δx_i , віднесемо до середини інтервалу x_i , у результаті отримаємо трансформоване поле кореляції.

Визначимо часткові середні арифметичні $\bar{y}_i$ для кожного значення $\bar{x}_i$:

$$\bar{y}_i = \frac{1}{n} \sum_{ik=1}^{n_{ik}} y_{ik},$$

де n_{ik} — число точок, що опинилися в інтервалі Δx_i , причому $\sum_{i=1}^m n_{ik} = N$, N — загальне число спостережень.

З'єднаємо послідовно точки з координатами $\bar{x}_i$ та $\bar{y}_i$ відрізкамі прямих. Отримана ламана лінія називається *емпіричною лінією регресії* y за x ; вона показує, як у середньому змінюється y зі зміною x . Граничне положення емпіричної лінії регресії, до якого вона прямує при необмеженому збільшенні числа спостережень та одночасному зменшенні Δx , називається *граничною теоретичною лінією регресії*. Її знаходження і є основною задачею регресійного аналізу. Зазначимо, що за лінією регресії неможливо точно обчислити значення y за x_i в одному досліді. Однак залежність $y = f(x)$ дозволяє знайти середнє значення y при багаторазовому повторенні дослідів при фіксованому значенні x_i .

У регресійному аналізі розглядається зв'язок між однією змінною, яка називається *залежною*, та декількома іншими, які називаються *незалежними*. Цей зв'язок подається у вигляді математичної моделі, тобто у вигляді функції регресії. Якщо функція лінійна відносно параметрів, але не обов'язково лінійна відносно незалежних змінних, то математична модель є лінійною. У протилежному випадку модель є нелінійною.

Статистичними проблемами обробки у регресійному аналізі є такі:

- ☒ отримання найкращих точкових та інтервальних оцінок невідомих параметрів регресійного аналізу;
- ☒ перевірка гіпотез щодо цих параметрів;
- ☒ перевірка адекватності;
- ☒ перевірка множини передбачуваних (імовірних) припущень.

Об'єкт, що досліджується, може бути представлений у вигляді, наведеному на рис. 4.2.

З урахуванням n дослідів модель можна записати у матричній формі

$$Y = XB + E,$$

де

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}, \quad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad B = \begin{bmatrix} b_1 \\ \vdots \\ b_p \end{bmatrix}, \quad E = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}.$$

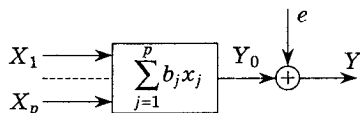


Рис. 4.2. Вигляд об'єкта дослідження

Незалежно від способу отримання вибірки є два способи, що називаються попередніми, для визначення існування та ступеня лінійної залежності між X та Y .

Перший спосіб полягає у графічному відображенні точок $(x_1, y_1), \dots, (x_n, y_n)$ на площині XY . Такий графік називається *діаграмою розсіювання*.

Другий спосіб — обчислення вибіркового коефіцієнта кореляції:

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

Для коректного використання регресійного аналізу застосовують такі припущення щодо властивостей регресійної помилки $E = \{e_i\}$ ($e_i = y_i - \hat{y}_i$, де $\hat{y}_i$ — значення залежної змінної, отримане шляхом підстановки X_{ij} у рівняння $y_i = \sum_{j=1}^p x_{ij} b_j$, $i = \overline{1, n}$, $j = \overline{1, p}$; n — кількість експериментальних даних; p — кількість незалежних змінних):

Пр. 1. У кожному досліді e_i має нормальний закон розподілу:

$$e_i \approx N(0), \quad i = 1, \dots, n.$$

Пр. 2. У кожному досліді математичне сподівання e_i дорівнює нулю:

$$M[e_i] = 0, \quad i = 1, \dots, n.$$

Пр. 3. В усіх дослідях дисперсія e_i постійна й однакова:

$$D[e_i] = \sigma_e^2, \quad i = 1, \dots, n.$$

Пр. 4. В усіх дослідях помилки e_i незалежні:

$$\text{cov}[e_i e_q] = 0, \quad i \neq q.$$

У векторній формі припущення Пр. 1–Пр. 4 для n дослідів можна подати у вигляді узагальненої нуль-гіпотези:

$$H_0: E \approx N_n(0, \hat{\sigma}_e^2, I_n)$$

де N_n — позначення нормального закону розподілу; I_n — n -вимірна одинична матриця.

Окрім припущень Пр. 1–Пр. 4, додатково вводять такі передумови:

Пер. 1 — матриця спостережень X має повний ранг, тобто $\text{rg}(x) = p$.

Пер. 2 — структура моделі адекватна істинній залежності.

Пер. 3 — значення випадкової помилки e_i не залежать від значень регресорів x_i .

Пер. 4 — помилки реєстрації Δx_{ij} регресорів зневажливо малі у порівнянні з випадковою помилкою e .

4.2.2. Принцип отримання файла вихідних даних при обробці методом РА

Для діючих об'єктів статичні та динамічні характеристики можуть бути отримані лише за експериментальними даними, одержаними в результаті проведення активного або пасивного експерименту. Обсяг експериментальних робіт істотно залежить від типу або класу об'єкта та мети дослідження, використання математичної моделі тощо.

Активний експеримент не завжди вдається реалізувати. Це пов'язано, передусім, із тим, що неможливо застосувати достатньо широкий діапазон варіювання керованих змінних через небезпеку зриву технологічного режиму та недотримання у вихідному продукті регламентованих значень.

У якості істотних змінних можуть опинитися контрольовані, але некеровані фактори, у цьому випадку активний експеримент складно реалізувати через неможливість знайти головний екстремум функції відгуку.

Шумове поле, разом з усіма некерованими змінними, може виявитися настільки великим, що математична модель, отримана за допомогою активного експерименту, буде неадекватною. У зв'язку з цим з'явилися припущення про сполучення активного та пасивного експериментів.

Випадкові завади у реальних умовах можуть накладатися навіть на керовані фактори, так що їх частину не вдається стабілізувати на необхідний час на рівнях, передбачених планом. Це призводить або до зростання похибок в усіх розрахунках, або до необхідності обробляти дані активного експерименту за методикою пасивного експерименту.

Розглянемо принцип організації пасивного експерименту. Для його проведення об'єкт слід оснастити необхідною контрольно-вимірювальною апаратурою, яка має лінійні статистичні характеристики, що забезпечують сталість амплітудно-частотних характеристик у діапазоні робочих частот об'єкта (від нуля до частоти зрізу $\omega_{\text{зріз}}$).

Аналогічні вимоги висуваються до датчиків, призначених для вимірювання вхідної та вихідної величин. Приклад безперервного вимірювання показаний на рис. 4.3, де результати пасивного експерименту отримані у вигляді записів вхідних та вихідних параметрів на папері.

У випадку використання міні- чи мікро-ЕОМ результати вимірювання датчиків через комутатори та аналогово-цифрові перетворювачі надходять у пам'ять ЕОМ через певні проміжки часу. Зауважимо, що результати безперервних вимірювань для обробки на ЕОМ необхідно подавати у дискретній формі, через певні інтервали часу. Слід зазначити, що вірогідність отриманих експериментальних даних суттєво залежить від організації проведення експериментальних робіт, яка має відповідати вимогам методу обробки.

Якщо обробка експериментальних даних ведеться кореляційними та регресійними методами, час між сусідніми вимірюваннями за кожною змінною $x_j(t)$ та $x_j(t + 1)$ необхідно обирати так, щоб кореляційні зв'язки були відсутні. Це означає, що інтервал часу Δt_ϵ періодичної реєстрації параметрів має бути більшим або дорівнювати максимальному значенню часу згасання кореляційної функції τ_z , який визначається за формулою $Rx_j(\tau_z) = 0,05 \cdot Rx_j(0)$. Графік визначення часу згасання кореляційної функції наведено на рис. 4.4.

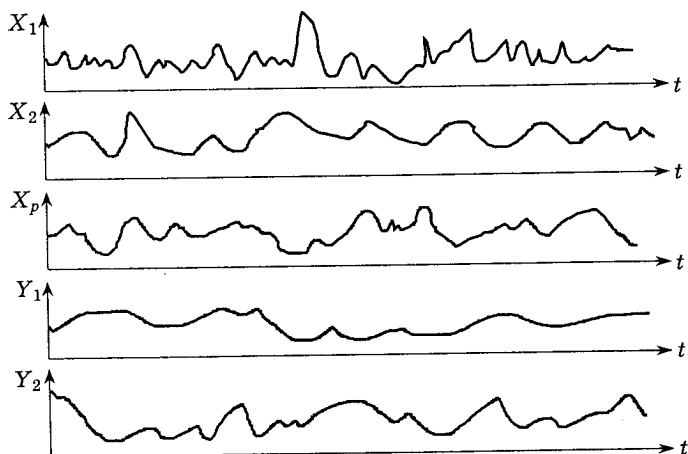
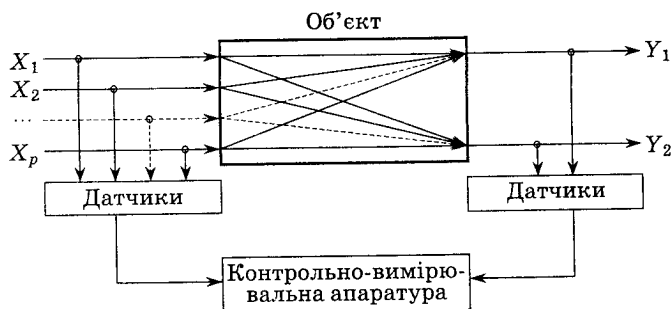


Рис. 4.3. Безперервна зміна входних та вихідних параметрів об'єкта

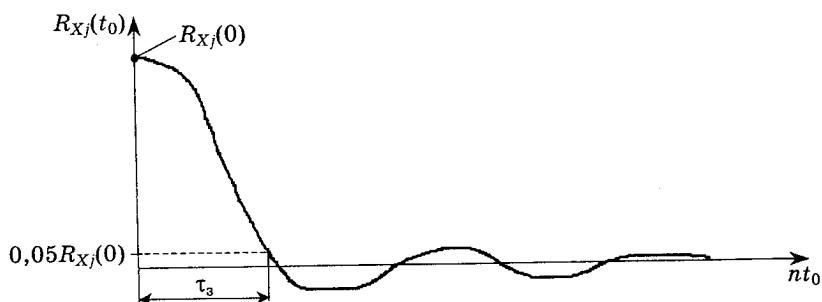


Рис. 4.4. Визначення часу згасання кореляційної функції

За відсутності такого матеріалу τ_z легко визначити приблизно за зв'язком часу згасання $Rx_j(t_0)$ з середнім числом нулів n_0 за одиницю часу $\tau_{zj} = \frac{2,0}{n_{0j}}$.

Середнє число нулів n_{0j} за одиницю часу визначається при запису безперервного процесу впродовж часу T_1 : $n_0 = \frac{N_1}{T_1}$, де N_1 — загальне число нулів на обраному відрізку реалізації; T_1 — відповідний час.

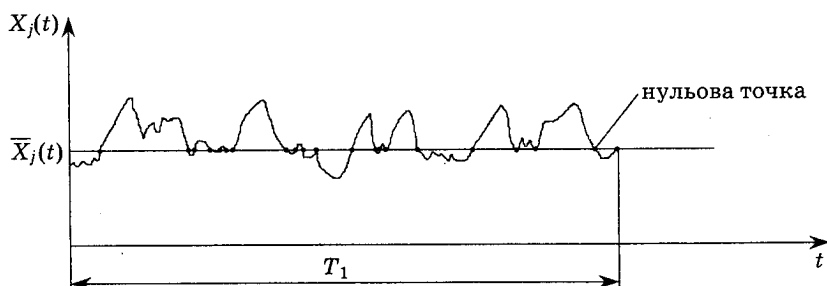


Рис. 4.5. Графік процесу для визначення n_0

Зазначимо, що результати не повинні піддаватися впливу змін технологічних параметрів, тому інтервал знімання даних необхідно обирати по можливості ближчим до нижньої межі, тобто $\Delta t_{0j} \geq \tau_{zj}$.

Вимога щодо відсутності кореляційних зв'язків між вимірюваннями пов'язана з тим, що матриця $(X^T X)^{-1}$ має бути добре зумовленою, і статистична модель задовільно інтерпретуватиме реальний процес тільки в тому випадку, якщо при організації експериментальних робіт будуть враховані динамічні властивості об'єкта.

Урахування динаміки об'єкта зводиться до визначення часу зсуву Δt_z між моментами реєстрації вхідної та вихідної величин. Час зсуву можна визначити за авто- та взаємнокореляційними функціями: $\Delta t_z = t_x - t_{xy}$ (див. рис. 4.6).

Точність такого способу визначення часу зсуву Δt_z залежить від близькості досліджуваного випадкового процесу до стаціонарного, величини похибки розрахунку оцінок кореляційних функцій тощо.

Далі складається таблиця експериментальних даних.

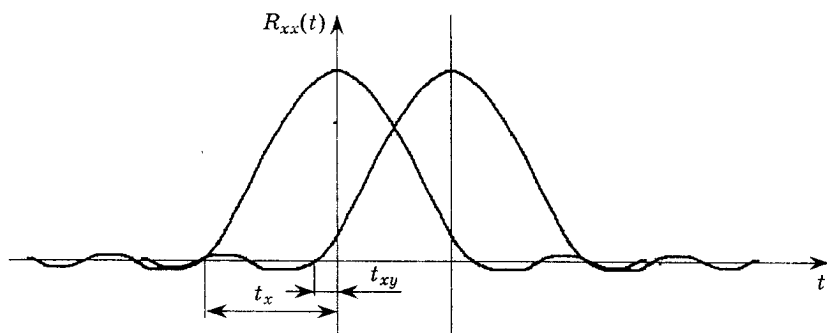


Рис. 4.6. Графічне визначення часу зсуву Δt_3

Таблиця 4.1

Таблиця експериментальних даних

Номер спостереження	x_1	x_2	...	x_p	y_1	y_2
1	x_{11}	x_{21}	...	x_{p1}	y_{11}	y_{21}
2	x_{12}	x_{22}	...	x_{p2}	y_{12}	y_{22}
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	$\vdots$	$\vdots$
N	x_{1N}	x_{2N}	...	x_{pN}	y_{1N}	y_{2N}

Значення $x_{11}, x_{21}, x_{31}, \dots, x_{pN}$ були отримані в момент часу t , а значення y_1 — у момент $(t + \Delta t_3)$, де Δt_3 — час динамічного зсуву.

Якщо обробка результатів вимірювань виконується шляхом регресійного аналізу, послідовність дій експериментатора з визначення Δt та T в умовах проведення пасивного експерименту для визначення статичних характеристик має бути такою:

1. За попереднім записом випадкового процесу визначити попередній крок квантування за формулою: $\Delta t = \frac{0,2}{n_0}$, де n_0 — середнє число перетинів (нулів) за одиницю часу.

2. Попередньо визначити час згасання кореляційної функції: $\tau_3 = \frac{2,0}{n_0}$. Провести попередню оцінку тривалості реалізації T_1 за формулою: $T \geq 10\tau_3$.

3. Урахувати динаміку об'єкта в експериментальних даних і скласти табл. 4.1.

Слід зазначити, що експериментальні роботи необхідно проводити за таких Δt і T , що відповідають умовам $\Delta t_j = \min_j |\Delta t_j|$ та $T = \max_j |T_j|$, де $j = \overline{1, p}$ — кількість вхідних змінних.

Після заповнення таблиці результати експерименту опрацьовують методами кореляційного та регресійного аналізу.

Зазначимо, що в цьому випадку необхідно перевірити всі передумови застосування цих методів. У результаті обробки експериментальних даних отримують статичні характеристики, за якими обирають близькі до лінійних режимів роботи об'єкта та значення можливих за технологічними умовами вхідних впливів.

Більш інформативні змінні отримують при проведенні активного експерименту.

Розглянемо деякі критерії оптимального планування експериментів.

1. *Критерій ортогональності плану.* Обирається план для оцінок параметрів моделі такого виду, для якого інформаційна матриця $\Phi = X^T X$ плану є діагональною. У цьому випадку матриця $(X^T X)^{-1}$ є діагональною, отже й коваріація оцінок дорівнює нулю. Це означає, що заміна нулем будь-якого коефіцієнта у рівнянні моделі не змінить значень усієї решти коефіцієнтів.

2. *Критерій рототабельності* потребує такого розташування експериментальних точок в області планування Ω_x , при якому дисперсія σ_y^2 оцінки значень залежної змінної у точці X залежить лише від відстані між цією точкою та центром плану.

Названі критерії забезпечують деякі корисні властивості оцінок коефіцієнтів, однак вони ніяк не пов'язані з вимогою максимальної точності побудування моделі.

3. *Критерій А-оптимальності.* Необхідно обрати план X , за якого слід матриці $C = (X^T X)^{-1}$ буде мінімальним (тобто сума діагональних елементів матриці C буде мінімальною). Критерій А-оптимальності потребує мінімуму середньої дисперсії оцінок.

4. *Критерій D-оптимальності.* Цей критерій потребує такого розташування точок в області Ω_x , за якого визначник матриці $C = (X^T X)^{-1}$ мінімальний (або визначник матриці $X^T X$ максимальний). D-оптимальний план мінімізує обсяг еліпсоїда розсіювання оцінок коефіцієнтів.

5. *Критерій G-оптимальності* потребує такого розташування точок в області Ω_x , за якого досягається найменша величина

максимальної дисперсії оцінки залежної змінної в області Ω_x . Критерій G -оптимальності вимагає максимальної точності оцінки залежної змінної.

Розглянемо принцип організації повного факторного експерименту (ПФЕ), який для більшості технологічних процесів є цілком задовільним.

Кількість дослідів становить $N = 2^P$, де P — кількість вхідних змінних. Кожна зі змінних варіюється на двох рівнях: $(+1, -1)$ або $(+, -)$.

Послідовність проведення експерименту є такою:

1. Стабілізується режим роботи об'єкта.
2. Визначається кількість вхідних змінних об'єкта (p) та їх рівень варіювання $(+, -)$.

3. Складається матриця планування повного факторного експерименту за таким алгоритмом:

3.1. Кожний q -й рядок матриці відповідає набору параметрів x_j , що варіюються.

3.2. Визначається значення x_0 , яке в усіх дослідженнях дорівнює $+$.

3.3. У першому досліді ($q = 1$) усі змінні, що варіюються, повинні знаходитися на нижньому рівні, тобто $x_1 = -1$, $x_2 = -1$, $x_3 = -1$.

3.4. Наступні рядки матриці планування обирають за таким правилом: при порядковому сортуванні варіантів частота зміни знаків змінних, що варіюються, повинна бути вдвічі меншою, ніж для попереднього рядка.

Наведемо матрицю планування для трифакторного експерименту:

Номер дослідів	x_0	x_1	x_2	x_3	x_1x_2	x_1x_3	x_2x_3	$x_1x_2x_3$	Кодове позначення
1	+	-	-	-	+	+	+	-	(1)
2	+	+	-	-	-	-	+	+	a
3	+	-	+	-	-	+	-	+	b
4	+	+	+	-	+	-	-	-	ab
5	+	-	-	+	+	-	-	+	c
6	+	+	-	+	-	+	-	-	ac
7	+	-	+	+	-	-	+	-	bc
8	+	+	+	+	+	+	+	+	abc

Реєструється вихідний параметр Y_q з урахуванням динаміки об'єкта. Необхідно у кожній точці дослідів провести K паралельних дослідів і результати спостережень $y_{q1}, \dots, y_{qk}$ усереднити за виразом

$$\bar{y}_q = \frac{1}{k} \sum_l^k y_{ql}, \quad (k=4 \dots 6).$$

Досліди з K серій необхідно рандомізувати з урахуванням даних табл. 4.2.

Таблиця 4.2

Рандомізація дослідів

Номер дослідів	Значення реалізації дослідів в серії							
1	1	5	4	2	3	8	5	3
2	6	3	7	5	7	5	1	1
3	5	7	5	3	1	1	7	5
4	8	2	1	6	4	4	4	4
5	2	4	8	1	2	2	8	7
6	7	1	3	7	6	3	3	8
7	4	8	6	8	8	7	6	2
8	3	6	2	4	5	6	2	6

Рандомізація здійснюється у такому порядку: для кожної серії дослідів з табл. 4.2 обирають будь-який зі стовпців чисел, які й визначають послідовність реалізації матриці планування експерименту в цій серії дослідів.

Результати експериментів опрацьовуються відомими методами обробки результатів активного експерименту.

Обчислювальна процедура регресійного аналізу

Нехай значення вихідної змінної Y залежать від значень змінних x_1, x_2, e , і ця залежність має вигляд:

$$y = b_1 x_1 + b_2 x_2 + e = y_0 + e.$$

За результатами спостережень над контрольованими змінними x_1, x_2 , у необхідно визначити невідомі коефіцієнти $b_j, j=1, 2, \dots$. У цьому випадку однозначність між X та Y порушується,

бо має місце ϵ , що призводить до необхідності збільшити число дослідів $n > p$ для отримання оцінок b_1, b_2 .

Для отриманих оцінок $\hat{b}_1, \hat{b}_2$ емпіричне рівняння приймає вигляд:

$$\hat{y} = \hat{b}_1 x_1 + \hat{b}_2 x_2, \quad (4.1)$$

Для визначення $\hat{b}_1, \hat{b}_2$ використовується МНК, згідно з яким необхідно мінімізувати суму квадратів різниць $\Delta y_i = y_i - \hat{y}_i$, $i = 1, \dots, n$, де y_i — значення вихідної змінної у i -му досліді; $\hat{y}_i$ — значення, отримане шляхом підстановки x_1, x_2 у (4.1).

Іншими словами, для визначення b необхідно мінімізувати

$$S = \sum_{i=1}^n (\Delta y_i)^2 = \sum_{i=1}^n (y_i - \hat{b}_1 x_{i1} - \hat{b}_2 x_{i2})^2 \rightarrow \min, \quad (4.2)$$

Функція (4.2) досягається за виконання умов

$$\frac{\partial S}{\partial \hat{b}_1} = 0, \quad \frac{\partial S}{\partial \hat{b}_2} = 0.$$

Продиференціюємо S за $\hat{b}_1, \hat{b}_2$ та прирівняємо похідні до нуля:

$$\begin{aligned} \frac{\partial S}{\partial \hat{b}_1} &= \frac{\partial}{\partial \hat{b}_1} \sum_{i=1}^n (y_i - \hat{b}_1 x_{i1} - \hat{b}_2 x_{i2})^2 = \sum_{i=1}^n 2(y_i - \hat{b}_1 x_{i1} - \hat{b}_2 x_{i2})(-x_{i1}) = 0; \\ \frac{\partial S}{\partial \hat{b}_2} &= \frac{\partial}{\partial \hat{b}_2} \sum_{i=1}^n (y_i - \hat{b}_1 x_{i1} - \hat{b}_2 x_{i2})^2 = \sum_{i=1}^n 2(y_i - \hat{b}_1 x_{i1} - \hat{b}_2 x_{i2})(-x_{i2}) = 0. \end{aligned}$$

Після перетворень отримуємо систему нормальних рівнянь:

$$\begin{cases} \hat{b}_1 \sum_{i=1}^n x_{i1}^2 + \hat{b}_2 \sum_{i=1}^n x_{i1} x_{i2} = \sum_{i=1}^n y_i x_{i1}; \\ \hat{b}_1 \sum_{i=1}^n x_{i1} x_{i2} + \hat{b}_2 \sum_{i=1}^n x_{i2}^2 = \sum_{i=1}^n y_i x_{i2}. \end{cases} \quad (4.3)$$

Розглянемо геометричну інтерпретацію терміну «нормальні рівняння». Спостереження Y_i можна інтерпретувати як компоненти вектора спостережень Y з основою на початку координат у просторі спостережень. Потрібно обрати такі b_j (з простору параметрів), які визначають оцінку $\hat{Y}_i$, що реалізують $\min S$. Водночас необхідно знайти такі b_j , які визначатимуть вектор $(Y - \hat{Y})$, який прямуватиме через кінець вектора Y перпенди-

кулярно до поверхні оцінок. Цей перпендикуляр i є вектором, який реалізує мінімум S (рис. 4.7).

Для розв'язання систем нормальних рівнянь скористаємося правилом Крамера:

$$V = \begin{vmatrix} \sum x_{i1}^2 & \sum x_{i1}x_{i2} \\ \sum x_{i1}x_{i2} & \sum x_{i2}^2 \end{vmatrix} = \begin{vmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{vmatrix},$$

$$b_1 = \frac{|v_1|}{|v|}, \quad b_2 = \frac{|v_2|}{|v|},$$

де V — інформаційна матриця; V_j ($j=1, 2$) — матриця V , у якій j -й стовпець замінено стовпцем правої частини системи (4.3).

Розв'язок існує, якщо $r_x(x)=2$, тобто матриця X має бути матрицею повного рангу. Коефіцієнти $\hat{b}_1, \hat{b}_2$ є випадковими величинами. Якщо у рівняння регресії включити нову змінну, то необхідно коефіцієнт $\hat{b}_j$ обчислити заново.

У матричній формі узагальнене рівняння регресії набуває вигляду:

$$Y = XB + E.$$

Згідно з МНК шукані оцінки B мінімізують суму квадратів відхилень $(y_i - \hat{y}_i)$, тобто

$$S = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = (Y - \hat{Y})^T (Y - \hat{Y}) = (Y - X\hat{B})^T (Y - X\hat{B}) \rightarrow \min, \quad (4.4)$$

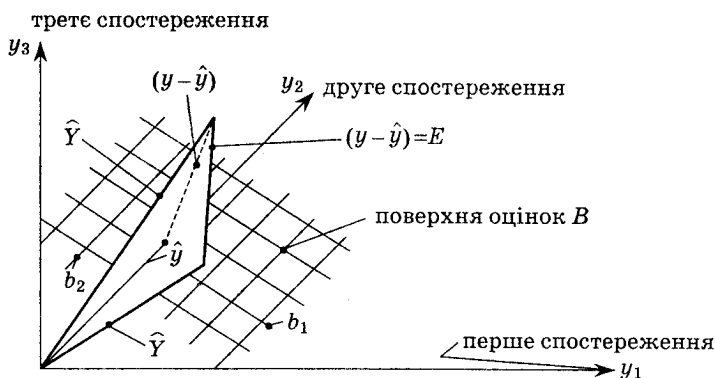


Рис. 4.7. Геометричне подання МНК

Тоді система рівнянь (4.9) набуває вигляду

$$b_1 \sum_{i=1}^n (x_{i1} - \bar{x}_1)^2 + \dots + b_p \sum_{i=1}^n (x_{i1} - \bar{x}_1)(x_{ip} - \bar{x}_p) = \sum_{i=1}^n (x_{i1} - \bar{x}_1)(y_i - \bar{y}),$$

.....

$$b_1 \sum_{i=1}^n (x_{i1} - \bar{x}_p)^2 (x_i - x_1) + \dots + b_p \sum_{i=1}^n (x_{ip} - \bar{x}_p) = \sum_{i=1}^n (x_{ip} - \bar{x}_p)(y_i - \bar{y}),$$

що за умови $x_0 = 1$ приводить до моделі

$$(y - \bar{y}) = b_1^0 (x_1 - \bar{x}_1) + \dots + b_p^0 (x_p - \bar{x}_p) + e,$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i, \quad \bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}, \quad j = \overline{1, p}. \quad (4.10)$$

Розв'язок рівняння (4.10) приводить до розв'язку рівняння регресії

$$\hat{y} = \bar{y} - \sum_{j=1}^p \hat{b}_j^0 \bar{x}_j + \hat{b}_1^0 x_1 + \dots + \hat{b}_p^0 x_p,$$

$$\text{де } \bar{y} - \sum_{j=1}^p \hat{b}_j^0 \bar{x}_j = b_0.$$

Таким чином, введення у рівняння регресії вільного члена приводить до центрування вихідних даних матриці спостережень (ХУ). Центрування зменшує кількість рівнянь регресії й полегшує обчислювальні процедури, оскільки при великому масштабі X та Y та великому n елементи $\sum_{i=1}^n x_{ij} y_i$, $\sum_{i=1}^n x_{ij}^2$ являють собою багаторозрядні числа, що призводить до втрати точності при округленні.

Розв'язання системи (4.10) полегшується, якщо попередньо провести стандартизацію вихідних даних, тобто привести всі дані до єдиного масштабу:

$$y_i^* = \frac{y_i - \bar{y}}{\hat{\sigma}_y}, \quad x_{ij}^* = \frac{x_{ij} - \bar{x}_j}{\hat{\sigma}_{x_j}}, \quad j = \overline{1, p},$$

$$\hat{\sigma}_y = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}, \quad \hat{\sigma}_{x_j} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}.$$

У процесі стандартизації змінних переходимо до кореляційного методу. Для стандартизованих змінних $y^*, x_1^*, \dots, x_k^*$ є справедливими рівності $\bar{y}^* = 0, \bar{x}^* = 0, \hat{\sigma}_y^2 = 1, \hat{\sigma}_{x_j}^2 = 1$ й система нормальних рівнянь має вигляд:

$$\begin{aligned}\hat{\beta}_1 \hat{r}_{11} + \hat{\beta}_2 \hat{r}_{12} + \dots + \hat{\beta}_p \hat{r}_{1p} &= \hat{r}_{1y}, \\ \hat{\beta}_1 \hat{r}_{21} + \hat{\beta}_2 \hat{r}_{22} + \dots + \hat{\beta}_p \hat{r}_{2p} &= \hat{r}_{2y}, \\ &\dots\dots\dots \\ \hat{\beta}_1 \hat{r}_{p1} + \hat{\beta}_2 \hat{r}_{p2} + \dots + \hat{\beta}_p \hat{r}_{pp} &= \hat{r}_{py},\end{aligned}$$

де $\hat{\beta}_j$ ($j = 1, \dots, p$) — вибіркові оцінки регресійного рівняння у стандартизованому вигляді.

Вибірковий коефіцієнт кореляції між j та g змінними визначається за формулою

$$r_{jq} = \frac{1}{n-1} \frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{iq} - \bar{x}_q)}{\bar{\sigma}_j \bar{\sigma}_q},$$

причому $j = q, r_{jq} = 1, -1 \leq r_{jq} \leq 1$.

Використання величин $\hat{r}_{jq}$ у якості коефіцієнтів системи нормальних рівнянь вельми корисне, оскільки всі числа, що брали участь у розв'язанні, знаходяться між +1 та -1. При цьому похибки, що виникають унаслідок округлення, мінімізуються. Емпіричне рівняння регресії для стандартизованих змінних $x_1, \dots, x_p$ не має вільного члена $\hat{y}^* = \hat{\beta}_1 x_1^* + \dots + \hat{\beta}_p x_p^*$ та є зручним для приблизної оцінки порівняльного впливу незалежних змінних $x_1, \dots, x_p$ на залежну змінну y .

Напрявленість ефекту, що викликаний зміною x , визначається знаком перед коефіцієнтом $\hat{\beta}_j$, а величина ефекту пропорційна до абсолютного значення коефіцієнта $\hat{\beta}_j$.

Зв'язок між оцінками у натуральному масштабі $\hat{b}_j$ та стандартному та $\hat{\beta}_j$ визначається співвідношенням

$$\hat{\beta}_j = \frac{\hat{\sigma}_{xj}}{\hat{\sigma}_y} \hat{b}_j, \quad j = \overline{1, p}.$$

Оскільки $\hat{\sigma}_y^2 > \hat{\sigma}_{xj}^2$, то й $\hat{b}_j > \hat{\beta}_j$.

Слід зазначити, що введені для зручності величини $\hat{r}_{jq}$, $\hat{r}_{jy}$ у регресійному аналізі не мають того ймовірнісного сенсу, який притаманний коефіцієнтам парної кореляції у кореляційному аналізі, коли дві випадкові величини, наприклад x_j та x_q , пов'язані спільним законом розподілу.

При дотриманні уведених припущень МНК оцінки регресійних коефіцієнтів $\hat{b}_j$ є незміщеними, переконливими та ефективними. Зазначимо, що регресійний аналіз не потребує, щоб усі вхідні змінні X_j були розподілені нормально, а потребує лише, щоб для будь-якого фіксованого набору X_{ij} вихідна змінна y була розподілена нормально. Це необхідно для коректної інтерпретації результатів з використанням розподілів Фішера та Стюдента.

Зауважимо, що незалежні змінні X у регресійному аналізі можуть бути випадковими й детермінованими. Якщо матриця X випадкова, велике значення має проведення попереднього статистичного аналізу, суть якого полягає в оцінюванні щільності статистичного зв'язку між незалежними змінними й залежною. Це оцінювання може бути визначено за допомогою кореляційного аналізу, який є потужним засобом для отримання апіорної інформації при застосуванні РА.

4.2.4. Довірчі інтервали для рівняння регресії

Для проведення поглибленого аналізу рівняння регресії перш за все необхідно впевнитись у тому, що вектор помилок E розподілений за нормальним законом. Для побудови довірчих інтервалів коефіцієнтів моделі, передбачених значень рівняння регресії, середнього значення використовуються стандартні статистичні розподіли F , t , X , що вимагають нормальності розподілів.

Визначення довірчого інтервалу для істинного значення y_i^0 рівняння регресії

Визначення довірчого інтервалу зводиться до знаходження інтервалу, в якому з імовірністю $P = 1 - \alpha$ міститься істинне значення y_i^0 , що відповідає деякому досліді $x_{ij} = x(x_{i1}, \dots, x_{ip})$ з матриці спостережень X . Інакше кажучи, є інтервал, у якому з заданою імовірністю знаходиться лінія регресії.

Підставляючи x_i в емпіричне рівняння регресії, отримаємо

$$\hat{y}_i = \hat{b}_1 x_{i1} + \hat{b}_2 x_{i2} + \dots + \hat{b}_p x_{ip}.$$

Різниця між $\hat{y}_i$ та y_i^0 пояснюється дією різних помилок. Зазначимо, що $\hat{y}_i$ має випадковий характер, оцінки $\hat{b}_i$ та $\hat{y}_i$ розподілені нормально з параметрами

$$M[\hat{y}_i] = M[X, \hat{B}] = X_i B = y_i^0,$$

$$D[\hat{y}_i] = D[x_i \hat{B}] = X_i D[\hat{B}] X_i^T = X_i (X^T X)^{-1} X_i^T \sigma_e^2.$$

Можна стверджувати, що $\lim_n D[\hat{y}_i] = 0$.

Інакше кажучи, y є обґрунтованою оцінкою істинного значення y , що відповідає досліді x , тобто при необмеженій кількості дослідів емпірична лінія регресії збігається з дійсною залежністю $y^0 = b_1 x_1 + \dots + b_p x_p$.

Складаючи дріб Стюдента, отримуємо:

$$n = \frac{\hat{y}_i - y_i^0}{\hat{\sigma}_{\text{зал}} \sqrt{x_i (X^T X)^{-1} X_i^T}} \sim t_{(n-p)}.$$

Якщо задати рівень значущості α та знайти табличне значення $t_{\alpha/2(n-p)}$, можна побудувати достовірний інтервал для $P = 1 - \alpha$ у вигляді

$$\hat{y}_i - t_{\alpha/2(n-p)} \hat{\sigma}_{\text{зал}} \sqrt{x_i (X^T X)^{-1} X_i^T} \leq y_i^0 \leq \hat{y}_i + t_{\alpha/2(n-p)} \hat{\sigma}_{\text{зал}} \sqrt{x_i (X^T X)^{-1} X_i^T}.$$

Щоб надати графічну ілюстрацію отриманого результату, звернемося до двовимірної моделі за умови $x_j = 1$.

Інтервальні оцінки регресійних коефіцієнтів

Для побудови інтервальних оцінок для параметрів регресійної моделі та перевірки гіпотез необхідно ввести припущення у вигляді розподілу E , оскільки ці процедури використовують стандартні статистичні розподіли, що ґрунтуються на нормальному законі. Це узгоджується з центральною граничною теоремою теорії ймовірностей:

$$E \sim N_n(0, \sigma e^2, I_n).$$

Інтервальною оцінкою параметра θ_0 будемо називати такий інтервал $L = \theta_2 - \theta_1$ на осі θ , який із заздалегідь обраною ймовірністю $P = 1 - \alpha$ покриває істинне значення параметра

$$P[\theta_2 \leq \theta_0 \leq \theta_1] = 1 - \alpha. \quad (4.11)$$

Побудувати інтервальну оцінку для параметра θ_0 — означає визначити за даними вибірки (або за точковою оцінкою $\hat{\theta}$) кінцеві значення θ_1 та θ_2 інтервалу, який відповідає умові (1). На практиці співвідношення означає, що у $(1 - \alpha) \cdot 100\%$, істинне значення θ_0 завжди перебуватиме в інтервалі $[\theta_2, \theta_1]$.

Для обраного рівня значущості α , що визначає довірчу ймовірність $P = 1 - \alpha$, можна знайти табличне значення $t_{\frac{\alpha}{2}(n-2)}$. Тоді можна записати співвідношення

$$P\left[-t_{\frac{\alpha}{2}(n-2)} \leq \frac{\hat{b}_j - b_j^*}{\hat{\sigma}_{\text{зал}} \sqrt{c_{jj}}} \leq t_{\frac{\alpha}{2}(n-2)}\right] = 1 - \alpha, \quad (4.12)$$

де $C = V^{-1} = \begin{vmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{vmatrix}$; V — матриця коефіцієнтів системи. Так, для двовимірного випадку система рівнянь має вигляд:

$$\hat{b}_1 \sum_{i=1}^n x_{i1}^2 + \hat{b}_2 \sum_{i=1}^n x_{i1} x_{i2} = \sum_{i=1}^n y_i x_{i1};$$

$$\hat{b}_1 \sum_{i=1}^n x_{i1} x_{i2} + \hat{b}_2 \sum_{i=1}^n x_{i2}^2 = \sum_{i=1}^n y_i x_{i2};$$

$$V = \begin{vmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{vmatrix}, \quad C_{jj} = \frac{|v_{jj}|}{|v|},$$

V_{jj} — отримується викресленням j -го рядка та j -го стовпця

$$|V| = \sum_{i=1}^n x_{i1}^2 \sum_{i=1}^n x_{i2}^2 - \left(\sum_{i=1}^n x_{i1} \sum_{i=1}^n x_{i2} \right)^2.$$

Зробивши перетворення у квадратних дужках, отримаємо вираз для довірчого інтервалу коефіцієнта b_1 , для довірчої ймовірності $P = 1 - \alpha$.

$$\hat{b}_1 - t_{\frac{\alpha}{2}(n-2)} \hat{\sigma}_{\text{зал}} \sqrt{c_{11}} \leq b_1 \leq \hat{b}_1 + t_{\frac{\alpha}{2}(n-2)} \hat{\sigma}_{\text{зал}} \sqrt{c_{11}}.$$

Для ширини довірчого інтервалу справедливо

$$L = 2t_{\frac{\alpha}{2}(n-2)} \hat{\sigma}_{\text{зал}} \sqrt{c_{11}} = 2t_{\frac{\alpha}{2}(n-2)} \hat{\sigma}_{\text{зал}} \sqrt{\frac{\sum x_2^2}{\sum x_1^2 \sum x_1^2 - (\sum x_1 x_2)^2}}.$$

Довірчий інтервал залежить від деяких факторів, а саме:

1. Довірчий інтервал є симетричним відносно виборчої оцінки $\hat{b}_j$.
2. Ширина довірчого інтервалу залежить від $\hat{\sigma}_{\text{зал}}$ та x_1, x_2 .
3. Ширина довірчого інтервалу мінімальна, якщо $\sum x_1 x_2 = 0$ (тобто вектори ортогональні).
4. Ширина довірчого інтервалу прямує до нескінченності, якщо вектор-стовпці x_1 та x_2 у матриці спостережень x колінеарні, тобто якщо $\sum x_1^2 \sum x_2^2 = (\sum x_1 x_2)^2$.
5. У загальному випадку в регресійних рівняннях довірчий інтервал для окремо взятого регресійного коефіцієнта b_j визначається виразом:

$$P = \left[\hat{b}_j - t_{\frac{\alpha}{2}(n-p)} \hat{\sigma}_{\text{зал}} \sqrt{c_{jj}} \leq b_j \leq \hat{b}_j + t_{\frac{\alpha}{2}(n-p)} \hat{\sigma}_{\text{зал}} \sqrt{c_{jj}} \right] = 1 - \alpha.$$

Використовуючи F -розподіл, отримаємо

$$F_{(\alpha, p, n-p)} = \frac{(B - \hat{B})X^T X(B - \hat{B})}{k \hat{\sigma}_{\text{зал}}^2}.$$

Для випадку двовимірної регресійної моделі довірча область у координатах площини b_1, b_2 набуде вигляду:

$$\begin{aligned} (b_1 - \hat{b}_1)^2 \sum_{i=1}^n x_{i1}^2 + 2(b_1 - \hat{b}_1)(b_2 - \hat{b}_2) \sum x_{i1} x_{i2} + (b_2 - \hat{b}_2)^2 \sum_{i=1}^n x_{i2}^2 = \\ = 2\hat{\sigma}_{\text{зал}}^2 F_{\alpha}(2, n-2). \end{aligned}$$

Це рівняння еліпса у координатах b_1, b_2 (див. рис. 4.8). Нахил головних осей еліпса щодо координатних осей, як і його товщина, залежать від ступеня корельованості змінних x_1, x_2 .

У випадку, коли $r_{12} = 0$, спільна довірча область розташовується паралельно до координатних осей, а її витягнутість визначається значеннями $\hat{\sigma}_1^2$ та $\hat{\sigma}_2^2$ (див. рис. 4.9).

Спільна довірча область у вигляді еліпса тим ширша, чим менше α .

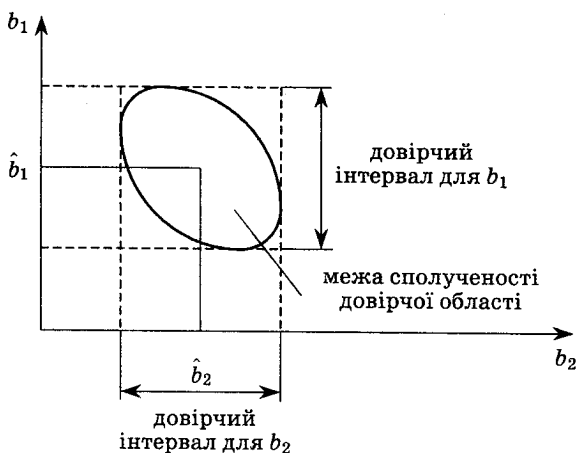


Рис. 4.8. Графічна інтерпретація довірчого інтервалу для двовимірної регресійної моделі

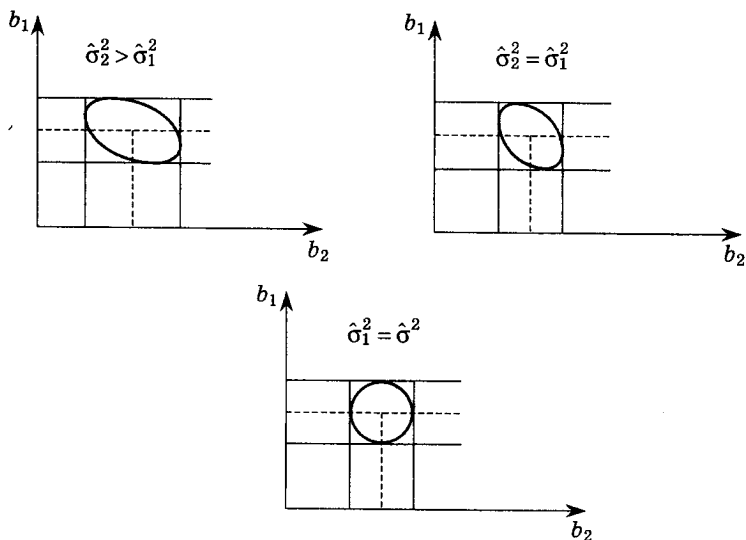


Рис. 4.9. Графічна інтерпретація довірчої області для $r_{12} = 0$

Визначимо довірчий інтервал для рівняння регресії. Нехай існує двовимірний модель виду $y = b_1 + b_2 x_2 + e$. Якщо $x \equiv 1$, отримаємо

$$y = b_1 + b_2 x_2 + e.$$

Тоді довірчий інтервал для y можна подати у вигляді:

$$y_i^0 = (\hat{b}_1 + \hat{b}_2 x_{i2}) \pm t_{\frac{\alpha}{2}(n-2)} \hat{\sigma}_{\text{зал}} \left[\frac{1}{n} + \frac{(x_{i2} - \bar{x}_2)^2}{\sum_{i=1}^n (x_{i2} - \bar{x}_2)^2} \right]^{\frac{1}{2}}, \quad (4.13)$$

де

$$\hat{\sigma}_{\text{зал}}^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{b}_1 - \hat{b}_2 x_{2i})^2.$$

Вираз (4.13) на площині з координатними осями — y та x_2 визначає дві гілки гіперболи, віссю якої є емпіричне рівняння регресії $\hat{y} = \hat{b}_1 + \hat{b}_2 x_2$.

Ці гілки є геометричним місцем точок довірчих інтервалів для y_j при зміні значення x_{j2} (див. рис. 4.10).

Довірчий інтервал має найменшу довжину, коли $x_{i2} = \bar{x}_2$, де $\bar{x}_2$ — середнє значення змінної x_2 .

Довжина довірчого інтервалу зростає, якщо збільшується значення $(x_{i2} - \bar{x}_2)$.

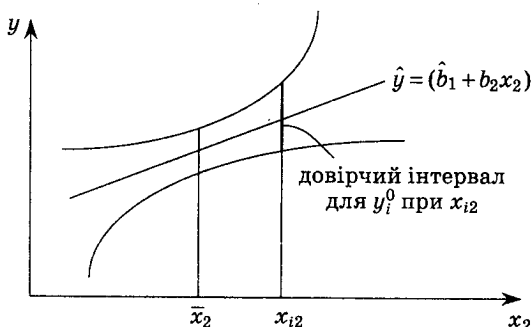


Рис. 4.10. Довірчий інтервал для y_i^0 при x_{i2}

4.2.5. Оцінка дисперсії помилки та ступінь невизначеності моделі

Визначимо ступінь невизначеності рівняння

$$\hat{y} = \hat{b}_1 x_1 + \hat{b}_2 x_2 \quad (4.14)$$

або наскільки у цілому є відмінними значення y_i , $i = 1, \dots, n$, що спостерігалися в експерименті, та значення $\hat{y}_i$, $i = 1, \dots, n$, отримані шляхом підстановки у рівняння (4.14) відповідних даних x_{i1} , x_{i2} .

Розглянемо дисперсію вихідної змінної

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

яка характеризує міру розкиду вибірових значень y біля вибіркового середнього $\bar{y}$.

Для наочності приймемо $x = 1$ та нанесемо пари значень $(x_{i2} y_i) \dots (x_{n2} y_n)$ у вигляді точок на координатну площину YX_2 . Позначимо $\bar{y}$ та лінію $\hat{y} = \hat{b}_1 + \hat{b}_2 x_2$ (див. рис. 4.11).

Для будь-якого значення y є справедливим вираз

$$(y_i - \bar{y}) = (y_i - \hat{y}_i) - (\bar{y} - \hat{y}_i),$$

або для точки x_{i2}

$$(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}). \quad (4.15)$$

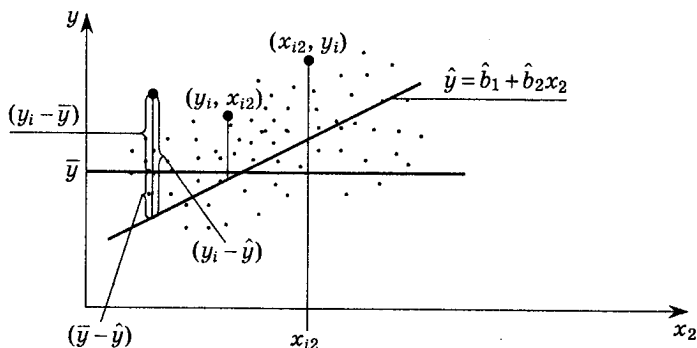


Рис. 4.11. Дисперсія вихідної змінної

Піднесемо до квадрата (4.15) та, підсумувавши остаточний результат, отримаємо

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y})^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2, \quad (4.16)$$

Співвідношення (4.16) можна інтерпретувати таким чином:

а) сума квадратів відхилень

$$\sum_{i=1}^n (y_i - \bar{y})^2$$

характеризує розкид значень $y_1, \dots, y_n$, що спостерігалися в експерименті, відносно $\bar{y}$;

б) сума квадратів відхилень

$$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

характеризує розкид відносно $\bar{y}$ значень $\hat{y}_1, \dots, \hat{y}_n$, отриманих шляхом підстановки у вираз

$$\hat{y} = \hat{b}_1 + \hat{b}_2 x_2$$

значень x , що відповідають експерименту. Тобто

$$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

характеризує ту частку загального розкиду $y_1, \dots, y_n$ відносно $\bar{y}$, яка зумовлена зміною значення x_2 , що спостерігалася в експерименті;

в) сума квадратів

$$\sum_{i=1}^n (y_i - \hat{y})^2$$

характеризує ту частку розкиду значень $y_1, \dots, y_n$ відносно $\bar{y}$, яка не пояснюється зміною значення x_2 , тобто таку, що зумовлена дією в часі експерименту помилки e .

Позначимо

$$\sum_{i=1}^n (y_i - \bar{y})^2 = S_{\text{заг}} = SS_t, \quad \sum_{i=1}^n (y_i - \hat{y})^2 = S_{\text{зал}} = SS_r,$$

$$\sum_{i=1}^n (\hat{y} - \bar{y})^2 = S_{\text{рег}} = SS_d,$$

де SS_r — відхилення регресії.

Тоді $S_{\text{заг}} = S_{\text{зал}} + S_{\text{рег}}$, $SS_t = SS_d + SS_r$.

Якщо $S_{\text{заг}} = S_{\text{рег}}$ та $S_{\text{зал}} = 0$, то всі значення y збігаються з $\bar{y}$, тобто випадкова помилка відсутня і $\sigma_e = 0$.

Отже, маємо функціональний зв'язок y з x . Якщо $S_{\text{заг}} = S_{\text{зал}}$, тобто $S_{\text{рег}} = 0$, то це означає, що між змінними y та x відсутній зв'язок й усі зміни y пов'язані з дією випадкової помилки e .

4.2.6. Адекватність регресійної моделі

Відмінність між дійсною залежністю $\psi(\cdot)$ і залежністю апроксимуючою називатимемо *адекватністю моделі*. Розглянемо вплив неадекватності моделі на властивості оцінок емпіричної моделі для одновимірного випадку.

Нехай є об'єкт з вхідною (x_2) і вихідною (y) змінними. Залежність між y та x_2 носить нелінійний характер $y^0 = \psi(x_2)$, водночас постулюється залежність між y та x_2 у вигляді лінійної функції $y^0 = b_1 + b_2 x_2$.

Регресійна модель має вигляд $y = b_1 + b_2 x_2 + e$. У результаті оброблення вхідних даних методом МНК отримаємо емпіричне рівняння регресії $\hat{y} = \hat{b}_1 + \hat{b}_2 x_2$. Зобразимо на координатній площині y, x_2 такі дані (див. рис. 4.12):

$$\begin{aligned} y_i - \hat{y}_i = \hat{e}_i &= [y_i - \psi(x_{i2})] + [\psi(x_{i2}) - y_i^0] + [y_i^0 - \hat{y}_i], \\ \hat{e}_i &= [\hat{y}_i - \psi(x_{i2})] + [\psi(x_{i2}) - y_i^0] + [y_i^0 - \hat{y}_i] = g_i + d_i. \end{aligned} \quad (4.17)$$

Величина d_i характеризує в кожній точці x_{i2} різницю між дійсною залежністю $y^0 = \psi(x_2)$ та постульованою моделлю $y^0 = b_1 + b_2 x_2$.

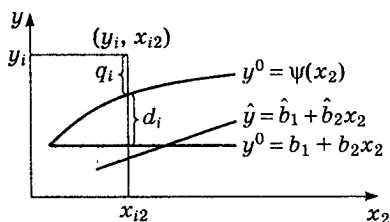


Рис. 4.12. Зображення даних на координатній площині

ня регресії задається незсуненою оцінкою не σ_e^2 , а величини $(\sigma_e^2 + \sum d_i^2)$, тобто

$$\hat{\sigma}_{\text{зал}}^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2} \approx \sigma_e^2 + \sum_{i=1}^n d_i^2. \quad (4.19)$$

Таким чином, у загальному випадку залишкова дисперсія $\hat{\sigma}_{\text{зал}}^2$ оцінює дисперсію помилки $\hat{\sigma}_e^2$ і неадекватність.

Якщо дисперсія помилки $\hat{\sigma}_e^2$ невідома, але помилки $e_1, e_2, \dots, e_n$ розподілені за законом $N(0, \hat{\sigma}_e^2)$, для перевірки гіпотези про те, що проста лінійна регресія Y за X буде відсутня, тобто гіпотези $H_0: \beta_1 = 0$ проти альтернативи $H_1: \beta_1 \neq 0$, досліджується F -відношення з таблиці дисперсійного аналізу для одновимірного випадку:

$$F_0 = \frac{MS_d}{MS_R} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 / \nu_D}{\sum_{i=1}^n (y_i - \hat{y}_i)^2 / \nu_R} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - b_0 - b_2 x_{i2}) / (n-2)}. \quad (4.20)$$

Якщо вірною є гіпотеза H_0 , то й F_0 має F -розподіл з $\nu_D = 1$ і $\nu_R = n - 2$ степенями вільності. P -значення є площа області під кривою щільності розподілу $F(\nu_D, \nu_R)$ праворуч від F_0 . Ми відхиляємо H_0 , якщо P менше, ніж рівень значущості α . Якщо H_0 приймається, то найкращою оцінкою Y при будь-якому $X = x$ буде середнє значення $\bar{y}$.

З іншого боку, існує відношення (4.20), яке можна використати для оцінювання адекватності, порівнюючи $F_{\text{розр}}$ та $F_{\text{табл}}$. Розрахункове $F_{\text{розр}}$ визначається за формулою (4.20). Значення $F_{\text{табл}}$ береться з таблиць з певним числом степенів вільності і для прийнятого рівня значущості α .

Математичне сподівання суми квадратів залишків дорівнює:

$$M \left[\hat{E}^T \hat{E} \right] = \frac{\sigma_{\text{зал}}^2}{n-2} = \frac{\sigma_e^2 + \sum_{i=1}^n d_i^2}{n-2}. \quad (4.18)$$

З (4.18) випливає, що у випадку неадекватності залишкова дисперсія емпіричного рівняння

Якщо розраховане значення $F_{\text{розр}}$ виявиться більшим за $F_{\text{табл}}$, то це означає, що залишкова дисперсія MSR статистично менша дисперсії MSD відносно $\bar{y}$, у цьому випадку отримане рівняння регресії можна вважати дієздатним.

Зрізані вибірки, як правило, не підкоряються нормальному закону розподілу. Тому використання (4.20) для оцінки адекватності є проблематичним. У зв'язку з цим в роботі [8] пропонується, щоб $F_{\text{розр}}$ перевищувало $F_{\text{табл}}$ у 5–6 разів.

Розглянемо питання оцінки адекватності, у випадку наявності дисперсії помилки σ_e^2 , яку можна оцінити експериментально.

Нехай є p -компонент вектора X , наприклад, $x_1, \dots, x_p$. Для кожного з цих x_j існує n_l спостережень $y_{j1}, y_{j2}, \dots, y_{jnk}$ змінної Y $j = 1, \dots, p, l = 1, \dots, k$.

$$X \left\{ \begin{array}{l} x_1 \left\{ \begin{array}{l} y_{11} \\ \hat{y}_1 y_{21} \\ \vdots \\ y_{n1} \end{array} \right\} \bar{y} \\ x_2 \left\{ \begin{array}{l} y_{12} \\ \hat{y}_2 y_{22} \\ \vdots \\ y_{n2} \end{array} \right\} \bar{y}, \\ x_3 \left\{ \begin{array}{l} y_{1k} \\ \hat{y}_k y_{2k} \\ \vdots \\ y_{nk} \end{array} \right\} \bar{y} \end{array} \right.$$

$$\bar{y}_1 = \frac{1}{k} \sum_{l=1}^k y_{1l}, \quad \bar{y}_2 = \frac{1}{k} \sum_{l=1}^k y_{2l}, \quad \bar{y}_p = \frac{1}{k} \sum_{l=1}^k y_{pl},$$

$$\bar{y} = \frac{1}{p} \sum_{j=1}^p \bar{y}_j.$$

Статистику критерію для перевірки гіпотези H_0 , якщо відома дисперсія помилки, подамо у вигляді:

$$F_0 = \frac{\hat{\sigma}_{\text{ост}}^2}{\hat{\sigma}_e^2} < F_T[\alpha, (n-2), k(n-1)].$$

Якщо $F_0 < F_{\text{табл}} (1 - \alpha)$, гіпотеза приймається. Таким чином, має місце $\sum_{i=1}^n d_i^2 \neq 0$ і модель є неадекватною.

І лише за умови адекватності подання моделлю дійсної залежності величини $d_i = 0$. Припустимо, що під час експерименту можна спостерігати і значення регресійної похибки e . Тоді матриця спостережень має вигляд (табл. 4.4).

Використовуючи $D[e_1] = D[e_2] = \dots = D[e_n] = \alpha e^2$, можна отримати вираз:

$$\hat{\sigma}_e^2 = \frac{\sum (e_i - \bar{e})^2}{n - 1}.$$

Таблиця 4.4

Матриця спостережень регресійної помилки e

x_2	y	E
x_{12}	y_1	e_1
x_{22}	y_2	e_2
x_{32}	y_3	e_3
$\vdots$	$\vdots$	$\vdots$
x_{n2}	y_n	e_n

Оцінкою $\hat{\sigma}_e^2$ і залишковою дисперсією $\hat{\sigma}_{\text{зал}}^2$ скористаємося для перевірки статистичної гіпотези $H_0: \sigma_{\text{зал}}^2 = \sigma_e^2$ проти альтернативної $H_1: \sigma_{\text{зал}}^2 > \sigma_e^2$.

Якщо нуль-гіпотеза вірна, то $d_i = 0$, $i = 1, \dots, n$ і модель адекватна. Якщо нуль-гіпотеза відхиляється $H_1: \sigma_{\text{зал}}^2 > \sigma_e^2$, то присутнє зміщення d і модель неадекватна.

Якщо відома оцінка $\hat{\sigma}_e^2$ дисперсії чистої помилки, процедура перевірки адекватності така:

1. Обчислюється залишкова дисперсія $\hat{\sigma}_{\text{зал}}^2$.

2. Складається співвідношення $f = \frac{\hat{\sigma}_{\text{зал}}^2}{\hat{\sigma}_e^2}$. Відомо, що у випадку дотримання умов нуль-гіпотези $H_0: \sigma_{\text{зал}}^2 = \sigma_e^2$ величина f має F -розподіл з $(n - k)$ та $(n_0 - 1)$ степенями вільності чисельника та знаменника.

3. Для рівня значущості α з таблиць F -розподілу знаходимо значення $F_{\text{табл}}(\alpha, n - p, n_0 - 1)$.

4. Порівнюючи значення f і $F_{\text{табл}}(\alpha, n - p, n_0 - 1)$, визначаємо адекватність. Якщо $f < F_{\text{табл}}$ — це означає, що з імовірністю $P = 1 - \alpha$ дані матриці спостережень (XY) не суперечать припущенню про адекватність. Якщо $f > F_{\text{табл}}$ — модель не адекватна.

Хоча F -критерій показав, що підгонка моделі в цілому є задовільною, все ж можуть мати місце вагомі розбіжності. Такого роду розбіжності можна помітити за допомогою аналізу залишків, тобто досліджувати набір відхилень $e_i = y_i - \hat{y}_i$, який може надати ключ до того, як покращити модель. Існує достатнє розмаїття критеріїв аналізу залишків. Але зазвичай залишки досліджують такі характерні ознаки:

- ☑ виявлення викидів;
- ☑ виявлення тренду в залишках;
- ☑ виявлення різкої зміни рівня процесу;
- ☑ виявлення змін у дисперсії похибки;
- ☑ перевірка нормальності закону розподілу.

Зазвичай для відкриття викидів різких зрушень та тренду в залишках використовують графіки залишків.

Для відкриття змін у дисперсії помилки використовують критерії Кохрена і Бартлета.

Для перевірки незалежності помилок використовують критерій Дарбіна — Ватсона. При порушенні передумов і припущень регресійного аналізу використовують різноманітні модифіковані методи.

Недолік множинної регресії полягає в тому, що необхідно правильно обрати структуру моделі, що не завжди є можливим. Більш надійними способами побудови статичних моделей є покроковий регресійний аналіз (із включенням або з видаленням змінних) і метод групового обчислення аргументів.

4.2.7. Покрокова лінійна регресія

У багатьох випадках при застосуванні регресійного аналізу необхідно обрати структуру моделі. Ця задача може бути зведена до вибору підмножини S_k з усіх можливих підмножин змінних вимірності K , $k = 1, p$, які відповідає найбільше значення множинного коефіцієнта кореляції. Для підмножини S_i за допомогою статистики $F = \frac{(SS'_R - SS_R)/m}{MS_R}$, де SS'_R — залишкова сума квадратів для змінних $x_{m+1}, \dots, x_p$, перевіряється гіпотеза про те, що додавання $(P - 1)$ змінних не поліпшує припущення Y . Якщо ця гіпотеза відкидається, то перевіряється гіпотеза про те, що додавання в підмножину S_2 $(P - 2)$ змінних, що залишилися, не поліпшує припущення Y . Таку перевірку виконують, доки для

деякого S_m , $1 \leq m \leq p$, не приймається гіпотеза про відсутність поліпшення припущення Y при додаванні $(P - 2)$ змінних, що залишилися. Якщо число незалежних змінних велике, такий підхід пов'язаний із значним обсягом обчислень. Так, для $P = 5$ існує 31 рівняння регресії, при $P = 10$ кількість рівнянь становитиме 1023. Якщо є P змінних, маємо $(2^P - 1)$ рівнянь регресії. Одним з рішень такої проблеми є покрокова регресія, коли змінні X_j включаються одна за одною до підмножини згідно з заданим критерієм. У процесі розрахунку деяка змінна може бути замінена іншою.

Стандартна покрокова процедура (F -метод) полягає у включенні та виключенні змінних за допомогою квадрата t -критерію, що має F -розподіл. Припустимо, що до набору C уже включено k змінних ($k = 0, 1, \dots, p - 1$). Тоді значення F для включення змінної X , що не входить в C , обчислюється за формулою

$$F_{yx.c} = r_{yx.c}^2 \frac{n - k - 2}{1 - r_{yx.c}^2}, \quad (4.21)$$

де $r_{yx.c}$ — частковий коефіцієнт кореляції, квадрат якого визначає частку залишкової дисперсії Y , «поясненої» додаванням змінної X до набору C . Величина $F_{yx.c}$ слугує статистикою критерію для перевірки гіпотези про те, що припущення Y значуще не поліпшиться при включенні X до набору C , тобто $H_0: r_{yx.c} = 0$. Якщо ця гіпотеза є вірною, то статистика розподілена за законом $F(1, n - k - 2)$. Аналогічно, величина F -виключення для будь-якої змінної X з C слугує статистикою критерію для перевірки гіпотези про те, що набір C_1 , що здобутий з набору C при видаленні X і містить $k_1 = k - 1$ змінних, передбачає Y так само добре, як і набір C . Інакше кажучи, виконується перевірка гіпотези $H_0: r_{yx.c1} = 0$, і статистикою критерію є величина видалення F

$$F_{yx.c1} = r_{yx.c1}^2 \frac{n - k_1 - 2}{1 - r_{yx.c1}^2}, \quad (4.22)$$

розподілена за законом $F(1, n - k_1 - 2)$, якщо H_0 є вірним. Правило зупину засновано на завданні припустимого мінімуму F -включення та припустимого мінімуму F -видалення, $F_{вк} > F_{уд}$.

Розглянемо далі стандартну процедуру покрокової регресії.

Крок 0. Обчислюються прості коефіцієнти кореляції r_{yx} та величина F -включення F_{yx_j} , $j = \overline{1 \dots p}$. Статистика критерію визна-

часться за виразом $F_{yx_j} = r_{yx_j}^2 \frac{n-2}{1-r_{yx_j}^2}$, отриманим із формули (4.21)

шляхом підстановки $k=0$ або як квадрат t -статистики.

Крок 1. Змінна x_{j1} , якій відповідає найбільше значення F -включення (або, що еквівалентно, найбільша величина квадрата коефіцієнта кореляції з Y), обирається як найкращий предиктор для Y . Обчислюються коефіцієнти регресійної моделі за методом найменших квадратів, будується таблиця дисперсійного аналізу та визначається множинний коефіцієнт кореляції $r_{y, x_j} = |r_{yx_j}|$. Величина F -видалення для x_{j1} в цьому випадку співпадає з величиною F -включення. Далі обчислюються коефіцієнти часткової кореляції $r_{yx_j x_{j1}}$ та значення F -включення

$$F_{yx_j x_{j1}} = r_{yx_j x_{j1}}^2 \frac{n-3}{1-r_{yx_j x_{j1}}^2}$$

для $j=1, \dots, p$, $j \neq j_i$, тобто для кожної змінної, що не ввійшла до рівняння регресії. Ця статистика має 1 та $(n-3)$ степенів вільності і використовується для перевірки гіпотези $H_0: \rho_{yx_i x_{ji}} = 0$, $j=1 \dots p$, $i \neq i_1$. Якщо всі обчислені значення F -включення менші за встановлений мінімум, то далі виконується крок S , в іншому випадку відбувається перехід до кроку 2. Статистика $F_{yx_j x_{j1}}$ необхідна для перевірки гіпотези $H_0: \rho_{yx_j x_{j1}} = 0$, $j \neq j_1$.

Крок 2. Змінна x_{j2} , яка має найбільше значення F -включення (або, що еквівалентно, найбільший квадрат часткового коефіцієнта кореляції з Y при фіксованому значенні x_{j1}), обирається як найкращий предиктор для Y за умови, що вже обрана змінна x_{j1} . Обчислюються МНК-оцінки коефіцієнтів, будується таблиця дисперсійного аналізу, обчислюється множинний коефіцієнт кореляції $r_{yx_{j1} x_{j2}}$ та значення F -видалення $F_{yx_{j1} x_{j2}}$ та $F_{yx_{j2} x_{j1}}$. Ці статистики мають 1 та $(n-3)$ степенів вільності та визначаються виразами:

$$F_{yx_{j1} x_{j2}} = \frac{r_{yx_{j1} x_{j2}}^2 (n-3)}{1-r_{yx_{j1} x_{j2}}^2} \quad \text{і} \quad F_{yx_{j2} x_{j1}} = \frac{r_{yx_{j2} x_{j1}}^2 (n-3)}{1-r_{yx_{j2} x_{j1}}^2}.$$

Їх використовують для перевірки гіпотези $H_0: \rho_{yx_{j2} x_{j1}} = 0$. Обчислюються частковий коефіцієнт кореляції $r_{yx_j x_{j1} x_{j2}}$ та значення F -включення

$$F_{yx_j x_{j1} x_{j2}} = r_{yx_j x_{j1} x_{j2}}^2 \frac{n-4}{1-r_{yx_j x_{j1} x_{j2}}^2}$$

для перевірки гіпотези $H_0: \rho_{yx_j x_{j_1} x_{j_2}} = 0$ з 1 та $(n - 4)$ степенями вільності при $j = 1, \dots, p, j \neq j_1, j \neq j_2$. Якщо всі значення F -включення менше встановленого мінімуму, то далі виконують крок S , в іншому випадку відбувається перехід до кроку 3.

Крок 3. Нехай L означає набір з l незалежних змінних, які включені у рівняння регресії. У цьому разі виконують такі дії:

1. Якщо будь-які зі значень F -виключення для змінних з L менше за відповідний мінімум, то змінну, якій відповідає найменше значення F -виключення, видаляють із набору та виконують пункт 2 кроку 3 із заміною l на $(l - 1)$.

2. Якщо для всіх змінних, що не входять до L , значення F -включення менше за встановлений мінімум, то виконують крок S . В іншому випадку до набору L додають змінну, якій відповідає максимальне значення F -включення, і l замінюють на $(l + 1)$.

3. Визначають коефіцієнти моделі за МНК, складають таблицю дисперсійного аналізу, обчислюють множинний коефіцієнт кореляції r_{yl} між Y та змінними з L та значення F -виключення $F_{y_{jz}x}(l - 1)$ для Y та змінної x_{jz} з L при заданій решті $(l - 1)$ змінних з L . Кожна з цих величин має 1 та $(n - l - 1)$ степенів вільності та використовується для перевірки гіпотези $H_0: \rho_{yx_j}(l - 1) = 0$. Визначають величину часткового коефіцієнта кореляції $r_{yx_j.l}$ та значення F -включення $F_{yx_j.l}$ між Y та кожною змінною x_j , що не входить до L , за даних змінних з L . Ця статистика має 1 та $(n - l - 2)$ степенів вільності та перевіряє гіпотезу $H_0: \rho_{yx_j.l} = 0$ для x_j , що не входять в $L, j = 1, \dots, p$. Кроки 4, 5 рекурентно повторюють крок 3.

Крок S виконується якщо:

☒ значення F -включення для всіх змінних, що не входять в L , менше за встановлений мінімум;

☒ для всіх змінних з L значення F -виключення більше встановленого мінімуму;

☒ число включених змінних дорівнює p .

Крок S . На запит користувача друкується таблиця результатів для кожного кроку, в якій виводиться номер кроку, номер включених та видалених змінних, значення F -включення та F -виключення та множинного коефіцієнта кореляції між Y та включеними змінними.

Розглянемо принцип зупину процедури включення та виключення у покроковій процедурі [8].

Найкраще значення набору предикторів H може бути визначене, коли всі значення F -включення, що обчислюються, стануть менші заданих значень F -включення та F -виключення. Розглядають декілька правил зупину при реалізації покрокових процедур [8, 14, 15, 31]. Стандартне правило зупину засноване на розрахунку F -включення та F -виключення та порівнянні їх з припустимими значеннями. Правило зупину, засноване на зміні R_2 , передбачає перевірку гіпотези $H_0: \rho_y h = \rho_y t l$ за допомогою статистики t

$$F = \frac{n-l-1}{l-h} \cdot \frac{r_{yl}^2 - r_{yh}^2}{1 - r_{yl}^2},$$

де l — набір вхідних змінних, що входять у модель; h — набір змінних, що входять у рівняння на проміжному кроці. Якщо гіпотеза є правильною, то величина F має F -розподіл з $(l-h)$ та $(n-l-1)$ степенями вільності. Цей критерій застосовується послідовно на кожному кроці, доки не буде отримане перше незначуще значення F . Третє правило зупину засновано на перевірці гіпотези про те, що при переході до наступного кроку значення безумовного середнього квадрата помилки (UMSE) не зменшується. В ППП важливо обрати значення F -включення та F -виключення. У деяких ППП за замовчуванням передбачається, що мінімум F -включення дорівнює 4,0. Для змінних, що видаляються, також обирається припустимий принцип F -виключення (ця величина має бути менше за мінімум F -включення). Значення мінімуму F -включення еквівалентно максимуму рівня значущості, тобто $\min F$ -включення дорівнює $F_{1-\alpha(1, v)}$ для деякого числа степенів вільності v . Звичайно, $v = n - p - 1$, а рекомендоване значення α складає 0,15, хоча можна встановити $\alpha = 0,05$. У роботі [31] показано, що якщо $v_R < 40$ більш можливо використовувати стандартне правило зупину з $\alpha = 0,15$ (тобто $\min F$ -включення приблизно дорівнює 2,5). На панелі вводу даних є такі додаткові поля:

- ☒ **Method: Forward** — прямий метод;
- ☒ **Backward** — зворотний метод;
- ☒ **Standart** — стандартна регресія;
- ☒ **F-enter** — F -статистика, у разі перевищення якої змінна вводить у модель (за замовчуванням — менше за 4);
- ☒ **F-remove** — F -статистика, нижче якої змінні будуть видалятися з моделі (за замовчуванням менше за 4). **F-remove** не повинна перевищувати **F-enter**;

☒ **Max Steps** — максимальне число кроків, виконуваних системою перед зупиненням процесу відбору (за замовчуванням — 500).

Після заповнення полів система починає підбирати модель на основі встановлених критеріїв.

4.2.8. Перевірка передумов та припущень регресійного аналізу

Регресійний аналіз є одним з найпоширеніших методів обробки результатів спостережень. Він слугує підґрунтям для багатьох розділів математичної статистики та методів обробки даних. Регресійний аналіз базується на низці припущень та передумов, порушення яких призводить до некоректного його використання та помилкової інтерпретації результатів.

Хоча F -критерій й показав, що підгонка моделі в цілому є задовільною, доцільно провести аналіз залишків для перевірки дотримання передумов та припущень.

У цьому випадку досліджується набір відхилень між експериментальними й передбаченими значеннями залежної змінної $Y_i - \hat{Y}_i = E_i$.

Перевірка передумов та припущень регресійного аналізу передбачає розв'язання комплексу задач, до якого входять:

- ☒ оцінка випадковості залежної змінної;
- ☒ оцінка стаціонарності та ергодичності залежних та незалежних змінних;
- ☒ перевірка достовірності гіпотези про нормальність розподілу помилок E ;
- ☒ виявлення викидів;
- ☒ перевірка сталості математичного сподівання та дисперсії помилок;
- ☒ оцінка корельованості залишків;
- ☒ виявлення мультиколінеарності.

Перевірка випадковості

Експериментально-статистичний підхід, що реалізований для лінійних систем методами множинного регресійного аналізу, базується на припущенні випадковості помилки в нормальній лінійній моделі $Y = XB + E$, де Y — $(n \times 1)$ -вектор спостережень залежної

змінної; X — $(n \times p)$ -матриця спостережень незалежних змінних; B — $(p \times 1)$ -вектор невідомих коефіцієнтів; E — $(n \times 1)$ -вектор помилок.

У наведеній моделі E та Y є випадковими величинами, що потребують частого підтвердження, особливо при застосуванні множинного регресійного аналізу.

Задачу перевірки випадковості можна розділити на дві підзадачі:

☑ перевірка випадковості власне величини Y ;

☑ перевірка випадковості вибірки, тобто припущення про відсутність істотного зміщення середньої величини у часі.

Перша підзадача розв'язується шляхом застосування *критерію серій*. З цією метою послідовність спостережень величини Y подають як послідовність нулів та одиниць, де одиницею позначено значення, які перевищують середнє або медіану, а нулем — значення, менші за медіану. Після позначення вектор спостережень перетворюється на послідовність серій $\{r_i\}$, де r_i — кількість елементів одного виду, що йдуть поспіль, i — номер серії. Доведено, що за $N > 10$ розподіл величини r наближається до нормального з характеристиками

$$\bar{r} = \frac{N+1}{2}, \quad S_r = \frac{1}{2}\sqrt{N-1}.$$

Тоді з імовірністю 0,954 теоретичне число серій r перебуватиме в межах

$$\frac{N+1}{2} - \sqrt{N-1} \leq r_T \leq \frac{N+1}{2} + \sqrt{N-1}.$$

Якщо фактичне значення $r_{\text{ф}}$ потрапляє у вказані межі, то Y можна вважати випадковою величиною.

Серією називається послідовність спостережених значень, перед якими й після яких розташовані спостережені значення іншої категорії. Якщо послідовність N спостережень являє собою незалежні спостережені значення однієї й тієї ж самої випадкової величини, тобто ймовірність знаків (+) та (−) не змінюється від одного спостереження до іншого, то вибірковий розподіл числа серій у послідовності є випадкова величина r із середнім значенням

$$\mu_r = \frac{2N_1N_2}{N} + 1 \quad (4.23)$$

та дисперсією

$$\sigma_r^2 = \frac{2N_1N_2(2N_1N_2 - N)}{N^2(N-1)}. \quad (4.24)$$

Тут N_1 — число спостережень зі знаком (+), N_2 — число спостережень зі знаком (-). За умови $N_1 = N_2 = N/2$ співвідношення (4.23) та (4.24) набувають вигляду:

$$\mu_r = \frac{N}{2} + 1, \quad \sigma_r^2 = \frac{N(N-2)}{4(N-1)}.$$

У [2] та у прикладі 1 наведена таблиця, яка містить дані про $100\alpha\%$ -ві точки функції розподілу для критерію серій. Гіпотезу можна перевірити при будь-якому заданому рівні значущості α шляхом співставлення числа серій з граничними значеннями $r_{n, 1-\alpha/2}$ та $r_{n, \alpha/2}$, де $n = N/2$. Якщо спостережене число серій виходить за межі цього інтервалу, гіпотезу слід відкинути при рівні значущості α .

Приклад 4.1

Нехай $N = 20$ складається з таких чисел:

5,5 5,7 6,8 5,4
 5,1 5,0 5,5 6,8
 5,7 6,5 4,9 5,8
 5,2 5,4 5,4 6,9
 4,8 5,8 5,9 5,5

Визначимо, чи є незалежні спостереження випадковими.

Розв'язання

Оберемо $\alpha = 0,05$. Медіаною для $N = 20$ буде величина $x = 5,6$. Складемо послідовність з (+) та (-).

-- + -- + -- + - +++ -- + - +++ --
 1 2 3 4 5 6 7 8 9 10 11 12 13

Число серій дорівнює 13. Область прийняття гіпотези визначається інтервалом $r_{10, 1-\alpha/2} < r \leq r_{10, \alpha/2}$. З таблиці, що містить дані про $100\alpha\%$ -ві точки функції розподілу для критерію серій, наведеної у [2], за $\alpha = 0,05$ знаходимо, що $r_{10, 1-\alpha/2} = r_{10, 0,975} = 6$

та $r_{10, \alpha/2} = r_{10, 0,025} = 15$. Гіпотеза приймається, оскільки $r = 13$ входить до інтервалу.

Для розв'язання другої підзадачі застосовують *метод послідовних різниць*.

Елементи досліджуваної вибірки $\{y_i\}$ розташовуються за порядком отримання спостережень й для них обчислюються вибіркові середнє та дисперсія

$$\bar{y} = \frac{1}{N} \sum_{i=1}^N y_i, \quad \hat{S}^2 = \frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{y})^2.$$

Визначають різниці Δy між сусідніми спостереженнями $\Delta y_1 = y_2 - y_1$, $\Delta y_2 = y_3 - y_2$, ..., $\Delta y_{N-1} = y_N - y_{N-1}$ та математичне сподівання квадрата різниці

$$M[\Delta y] = \frac{1}{N-1} \sum \Delta y_i^2 = 2\hat{c}^2,$$

де $\hat{c}^2$ — оцінка генеральної дисперсії.

Фактична величина критерію випадковості вибірки

$$t_\Phi = \hat{c}^2 / S^2.$$

Відомо, що за умови $N > 20$ та випадкової вибірки, t_Φ розподілена нормально. Виходячи з $I\%$ -го рівня значущості, маємо $F(t_T) = 0,99$ та $t_T = 2,58$. Тоді теоретичне значення критерію $t_T = 1 - \frac{2,58}{\sqrt{N+1}}$. За умови $t_\Phi < t_T$ для конкретного N гіпотеза випадковості відкидається.

Перевірка стаціонарності

Аналіз випадкових процесів може проводитися шляхом осереднення величин за ансамблем вибіркових реалізацій або за однією реалізацією. При аналізі за ансамблем процес $\{y(t)\}$ вважається *нестационарним*, математичне сподівання

$$m_y(t_1) = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N y_i(t_1)$$

та автокореляційна функція

$$R_y(t_1, t_1 + \tau) = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N y_i(t_1) y_i(t_1 + \tau)$$

змінюються з плином часу. Якщо $\mu_y(t_1)$ та $R_y(t_1, t_1 + \tau)$ не залежать від t_1 , то випадковий процес $y(t)$ називається *слабо стаціонарним*, або стаціонарним у широкому розумінні. Середнє значення такого процесу не залежить від часу, а його автокореляційна функція залежить лише від величини зсуву, тобто $m_y(t_1) = m_y$, $R_y(t_1, t_1 + \tau) = R_y(\tau)$.

Коли всі початкові та змішані моменти не залежать від часу, то процес $\{y(t)\}$ є *строго стаціонарним*, або стаціонарним у вузькому розумінні.

На практиці для аналізу частіше за все доступна одна реалізація $y(t)$. У цьому випадку здійснюють осереднення за часом та кажуть про *внутрішню стаціонарність*, під якою розуміють, що характеристики процесу

$$m_y(t_1, T) = \frac{1}{T} \int_{t_1}^{t_1+T} y(t) dt, \quad R_y(t_1, t_1 + \tau, T) = \frac{1}{T} \int_{t_1}^{t_1+T} y(t) y(t + \tau) dt$$

не змінюються значуще для різних відрізків часу. Аналогічно до зазначеного вище вводяться поняття слабкої та строгої внутрішньої стаціонарності.

Оскільки на практиці перевірка за ансамблем достатньо довгих вибірових реалізацій нездійсненна, то для використання тестів перевірки стаціонарності процесу приймається низка припущень:

- 1) перевірка полягає у дослідженні поведінки не ансамблю, а його окремих реалізацій; це означає, що доведення внутрішньої стаціонарності окремих реалізацій може слугувати доказом стаціонарності випадкового процесу, якому належить ця реалізація;

- 2) для більшості процесів достатньо перевірити слабку стаціонарність, оскільки, по-перше, для ефективного використання спектрального та кореляційного аналізу випадкових процесів достатньо виконання умови слабкої стаціонарності, а по-друге, для реальних процесів зазвичай слабка стаціонарність тягне за собою й строгу; якщо процес визначається нормальною щільністю, то це доведення здійснюється автоматично, оскільки всі моменти вищих порядків повністю визначаються середнім і автокореляційною функцією;

- 3) довжина досліджуваної реалізації має бути достатньо великою порівняно з випадковими коливаннями, що містяться в ній, тобто довжина реалізації має дозволяти відокремити довгоперіодичні складові (*тренд*) від флуктуаційних;

4) на практиці часто стаціонарність автокореляційної функції забезпечується стаціонарністю дисперсії.

Ураховуючи ці припущення, перевірку стаціонарності здійснюють, досліджуючи одну реалізацію $x(t)$. Для цього вона розділяється на N рівних інтервалів — таких, що її ділянки в межах кожного інтервалу можна вважати незалежними. Для всіх інтервалів обчислюють середні значення й середні значення квадратів, з яких складають дві послідовності, які перевіряють на наявність тренда:

$$\bar{x}_1, \bar{x}_2, \dots, \bar{x}_N;$$

$$\bar{x}_1^2, \bar{x}_2^2, \dots, \bar{x}_N^2.$$

Якщо відомий вибірковий розподіл, то для перевірки можна застосувати непараметричні критерії (t -критерій Стьюдента, χ^2 -критерій Пірсона, F -критерій Фішера), однак у звичайній ситуації перевірка стаціонарності здійснюється за високої невідомості відносно досліджуваного процесу. У цьому випадку доцільно використати непараметричні критерії, наприклад, критерій серій або тренда. Розглянемо їх.

Припустимо, що вибіркові середні та квадрати мають істинні значення m_x та s_x^2 . Якщо гіпотеза є правильною і зміни у вибіркових значеннях не містять тренда, то число серій відносно будь-якої величини, наприклад медіани, дорівнює їхньому очікуваному числу в послідовності незалежних випадкових значень. Якщо число серій значною мірою відрізняється від очікуваного, то гіпотезу стаціонарності слід відкинути.

Аналогічно до застосування критерію серій при перевірці випадковості, в цьому разі формують послідовність, що складається зі знаків (+) та (-). Якщо виміряна величина x_i перевищує середнє (або медіану) за вибірками, тобто $x_i > \bar{x}$, то ставиться знак «+», в протилежному випадку ($x_i < \bar{x}$) — знак «-». Вибірковий розподіл числа серій є випадковою величиною r із середнім значенням

$$m_r = \frac{2N_1N_2}{N} + 1$$

та дисперсією

$$s_r^2 = \frac{2N_1N_2(2N_1N_2 - N)}{N^2(N - 1)},$$

де N_1 та N_2 — число спостережень зі знаком «+» та «-» відповідно. У [2] наведені дані про розподіл величини r за $N_1 = N_2 = N/2$. Якщо розраховане число серій r_p перебуває в інтервалі $[r_{N/2, 1-\alpha/2}, r_{N/2, \alpha/2}]$, то й гіпотеза про стаціонарність приймається, в іншому випадку — відкидається.

Використання критерію тренда засновано на підрахунку числа випадків k , коли $x_i > x_j$ для $i < j$ у послідовності N спостережених значень величини x . Така нерівність називається *інверсією*, а число k визначається зі співвідношення

$$k = \sum_{i=1}^{N-1} k_i,$$

де

$$k = \sum_{j=i+1}^N k_{ij}, \quad k_{ij} = \begin{cases} 1, & \text{при } x_i > x_j, \\ 0, & \text{при } x_i < x_j. \end{cases}$$

Число інверсій є також випадковою величиною із середнім

$$m_k = \frac{N(N-1)}{4}$$

та дисперсію

$$s_k^2 = \frac{2N^3 + 3N^2 + 5N}{72} = \frac{N(2N+5)(N-1)}{72}.$$

Область прийняття гіпотези обмежена інтервалом $[k_{N, 1-\alpha/2}, k_{N, \alpha/2}]$.

Критерій тренда має більшу потужність при виявленні монотонного тренда, однак при виявленні коливального тренда його потужність невелика, в цьому випадку доцільніше використовувати критерій серій.

Непараметричні критерії мають низку особливостей:

☑ немає необхідності знати ширину смуги частот досліджуваних процесів;

☑ не треба точно знати час осереднення, що витрачений на обчислення середніх та квадратів;

☑ для перевірки не обов'язково, щоб досліджувані процеси були повністю випадковими. При вивченні процесів може мати місце випадок, коли незалежність від часу середніх та квадратів не є достатньою умовою для твердження про незалежність від часу та автокореляційної функції.

Більшість висновків регресійного аналізу основані на припущенні про нормальний розподіл помилок. Цей факт потребує підтвердження за допомогою існуючих критеріїв згоди емпіричного та теоретичного розподілу: Пірсона, Смирнова — Крамера — Мізеса, Колмогорова, Шапіро — Уїлкса [8, 9, 10].

Критерії згоди основані на розрахунку статистик, які є мірою розбіжності емпіричного та теоретичного розподілів.

Перевага критерію χ^2 полягає в тому, що він може застосовуватись для перевірки різних розподілів і навіть значень окремих ймовірностей. Недоліками цього критерію є великий обсяг вибірки та відносно велика трудомісткість обчислень, до того ж результат, отриманий за критерієм χ^2 , залежить от способу розбиття на інтервали. Як правило, оптимальне число інтервалів визначається зі співвідношення $k = 3,322 \lg n + 1$.

Розглянемо n незалежних випадкових величин $Z_1, Z_2, \dots, Z_n$ кожна з яких розподілена нормально з нульовим середнім та одиничною дисперсією. Визначимо нову випадкову величину

$$x_n^2 = Z_1^2 + Z_2^2 + \dots + Z_n^2.$$

Щільність розподілу дорівнює

$$P(x^2) = [2^{n/2} \Gamma(n/2)]^{-1} (x^2)^{[(n/2)-1]} e^{-x^2/2}, \quad x^2 \geq 0,$$

де $\Gamma(n/2)$ — гамма-функція.

Окремий вид критерію, який часто використовується для перевірки узгодженості щільності розподілу, отриманого за даними вибірки, з деякою теоретичною щільністю розподілу, носить назву *критерію згоди χ^2* .

Отримують статистику

$$X^2 = \sum_{i=1}^k \frac{(f_i - F_i)^2}{F_i},$$

де f_i — число спостережених значень в i -му розряді, яке називається *спостереженою частотою*; F_i — очікувана частота, коли істинна щільність розподілу — величина X — є $P_0(X)$.

Таблиця 4.5

Мінімальне число розрядів K для виборок обсягу N за $\alpha = 0,05$

N	200	400	600	800	1000	1500	2000
K	16	20	24	27	30	35	39

Область прийняття гіпотези визначається нерівністю $x^2 \leq x_{n,\alpha}^2$, де дані про функції $x_{n,\alpha}^2$ беруть з таблиці хі-квадрат-розподілу.

Якщо вибіркове значення суми x^2 більше за $x_{n,\alpha}^2$, гіпотеза $P(x) = P_0(x)$ відкидається. За критерієм згоди Колмогорова оцінюємо близькість фактичного розподілу до теоретичного шляхом знаходження величини D , тобто максимальної різниці накопичених окремих значень фактичного та теоретичного розподілів. Якщо випробування незалежні та їх число n зростає, то ймовірність нерівності $A\sqrt{n} \geq 1$, прямує до межі:

$$\lim_{n \rightarrow \infty} P(A\sqrt{n} \geq 1) = 1 - \sum_{k=-\infty}^{\infty} (-1)^{k_e - 2k^2 l}.$$

Перевага критерію Колмогорова полягає у меншій трудомісткості розрахунку та можливості будувати довірчі межі для оцінок розподілів. Більш висока чутливість виявляється в такому: якщо для прийняття рішення про розподіл за критерієм x^2 потрібно N_1 відліків, то для l -критерію — $N^{4/5}$ відліків за тієї ж самої якості результатів. Цей критерій може використовуватись і при дуже малих обсягах вибірки, за яких x^2 -критерій взагалі не можна застосувати.

Критерій w^2 — Крамера — Мізеса — Смирнова заснований не на абсолютній, а на середньоквадратичній розбіжності між емпіричним та теоретичним розподілами. Цей критерій подібний до критерію згоди Колмогорова, однак має більшу потужність, тобто може використовуватися для меншої кількості відліків, але з більшою трудомісткістю.

Недоліком усіх критеріїв згоди є те, що вони не дають прямої відповіді на запитання про застосовність до цього емпіричного розподілу теоретичного розподілу, а вказують лише на ймовірнісну оцінку розбіжностей між ними. Цей недолік усувається за використання критерію Колмогорова, який дає пряму відповідь на запитання щодо міри розбіжності між емпіричним та теоретичним розподілами.

У критерій Ястремського використовується критерій Пірсона. У загальному вигляді критерій Ястремського може бути записаний таким чином:

$$l = \frac{|c^2 - n|}{\sqrt{2n + 4Q}},$$

де n — кількість груп; Q — величина, що залежить від кількості груп, проте якщо кількість груп менше 20, вважають, що $Q = 0,6$.

Оскільки величина l розподілена нормально, то з імовірністю 0,997 за абсолютним значенням вона має не перевищувати 3. Отже, за умови $|l| < 3$ емпіричний розподіл підпорядковується теоретичному закону розподілу. За умови $|l| > 3$ емпіричний розподіл не вкладається у теоретичний розподіл.

Наведені критерії згоди вирішують питання лише про те, існує чи ні відмінність між емпіричним та гіпотетичним розподілом, однак не визначають ступеня невідповідності.

Для визначення виду закону розподілу застосовують подання емпіричного розподілу у вигляді гістограм або графіка функції розподілу, за якими попередньо оцінюють його тип.

Вибір типу теоретичного розподілу, зазвичай, здійснюється на основі візуального аналізу. Як правило, його вигляд визначають за вибірковими точковими оцінками числових характеристик. До них належать оцінки:

$$\checkmark \text{ середнього — } \hat{m}_x = \frac{1}{N} \sum_{i=1}^N x_i;$$

$$\checkmark \text{ дисперсії — } \hat{D}_x = \frac{1}{N-1} \sum_{i=1}^N (x_i - \hat{m}_x)^2 = \hat{S}_x^2;$$

$$\checkmark \text{ асиметрії — } \hat{A} = \hat{m}_3 / \hat{s}_x^3;$$

$$\checkmark \text{ ексцесу } \hat{E} = \hat{m}_4 / \hat{s}_x^4 - 3, \text{ де}$$

$$\hat{m}_3 = \sum_{i=1}^k p_i (x_i - \hat{m}_x)^3; \quad \hat{m}_4 = \sum_{i=1}^k p_i (x_i - \hat{m}_x)^4,$$

(p_i — частоти за інтервалами);

$$\checkmark \text{ медіани } \hat{M}_e = x_{Me} + \Delta x \frac{(N/2 - H)}{n_{Me}};$$

$$\checkmark \text{ моди } \hat{M}_o = x_{Mo} + \frac{\Delta x (p_1 - p_{Mo})}{p_1 - 2p_{Mo} + p_2},$$

де x_{Me} та x_{Mo} — початки медіанного та модального інтервалів, тобто тих, у які потрапили Me та Mo ; Δx — довжина інтервалу; n_{Me} — частота медіанного інтервалу; H — сума відліків до початку медіанного інтервалу; p_{Mo} — частість модального інтервалу; p_1 та p_2 — частоти сусідніх з модальним інтервалів.

Точкові оцінки дозволяють отримати уявлення про характер щільності розподілу ймовірностей, використовуються для порівнянь та обчислень. Однак вони є величинами випадковими, тому не дозволяють безпосередньо відповісти на запитання, яку помилку ми робимо, приймаючи замість точного значення параметра його оцінку. Про це дозволяють судити інтервальні оцінки, які ґрунтуються на виявленні меж довірчих інтервалів, у яких з заданою ймовірністю знаходяться невідомі значення параметрів.

Однією з умов можливості застосування регресійних методів є добра зумовленість даних.

Таким чином, на кожному етапі ретроспективного аналізу здійснюється не лише вивчення характеру процесу, але й виявлення порушень окремих теоретичних передумов основних математичних методів, що використовуються для ідентифікації моделей. Такий аналіз дозволяє досліднику прийняти рішення про вибір того чи іншого математичного апарату, методу або класу методів.

Перевірку гіпотези про існування нелінійного зв'язку між вхідними та вихідними величинами здійснюють за формулою $n_{yx}^2(t) = h_{yx}^2(t) - r_{yx}(t)$, де $h_{yx}(t)$ — нормована-дисперсійна функція; $r_{yx}(t)$ — нормована кореляційна функція.

Загальна кількісна оцінка ступеня лінійності визначається за формулами:

$$Q_r = \frac{|r|}{s_r}, \quad Q_h = \frac{|h|}{s_h},$$

$$\text{де } s_r = \frac{1-r^2}{\sqrt{N}}, \quad s_h = \frac{1-h^2}{\sqrt{N}}.$$

Якщо $Q_r \geq 2,56$, то підтверджується наявність лінійного зв'язку між вхідними та вихідними змінними, якщо $Q_h \geq 2,56$ — підтверджується наявність нелінійного зв'язку.

Перевірка гіпотези про відхилення коефіцієнта кореляції від дисперсійного відношення проводиться за F -критерієм:

$$F = \frac{N-k}{k-2} \cdot \frac{h^2 - r^2}{1 - h^2},$$

де $k = 1 + 3,322 \ln N$.

Зв'язок можна вважати лінійним, якщо $F \leq F_{n_1, n_2}$ при довірчій ймовірності p_g , де n_1 та n_2 — степені вільності: $n_1 = N - M$,

$n_2 = k - 2$. У цьому випадку оператор об'єкта слід шукати у класі лінійних.

У процесі аналізу залишків слід пам'ятати про те, що у міру того, як незалежні змінні вводяться в модель, залишки стають менш інформативними. По суті, кожний залишок являє собою деяке зважене середнє арифметичне помилок e ; чим більше несуттєвих незалежних змінних x_j додається до моделі, тим більш схожими стають залишки.

Розглянемо питання про виявлення викидів у вибірці. Викид серед залишків являє собою залишок, який значно перевищує за абсолютною величиною решту й відрізняється від середнього за залишками на три, чотири або навіть більше стандартних відхилень.

Для виявлення викидів необхідно побудувати графік залишків, визначених за формулою

$$\delta_{ei} = \frac{y_i - \hat{y}_i}{\left\{ \frac{1}{n - K - 1} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \right\}^{\frac{1}{2}}}.$$

Якщо $\delta_{ei} \geq 3$, ця точка характеризуватиме викид. Слід зазначити, що інколи викид може надати корисну інформацію; у цьому разі їх необхідно ретельно досліджувати, а не відкидати. Проте викиди мають бути виключені з розгляду, якщо вони викликані помилками у реєстрації даних, незадовільним станом апаратури тощо. Якщо є не одне аномальне вимірювання, то критерій δ_{ei} їх не виявляє, особливо якщо аналізується менш ніж 30 вимірювань. За наявності декількох підозрілих значень доцільно перейти до інших критеріїв.

Виявлення тренда

Виявлення тренда залишків можна провести за графіком (рис. 4.13), який дозволяє виявляти тренди розмахів та нелінійні тренди.

Наявність різкого зсуву рівня слід визначати за графіком δ_{ei} (рис. 4.14). За графіком визначається точка експерименту, за якої спостерігається різкий зсув. Для оцінки числового значення зсуву застосовують різницю між середнім вибірковим значенням залишків до й після викиду.

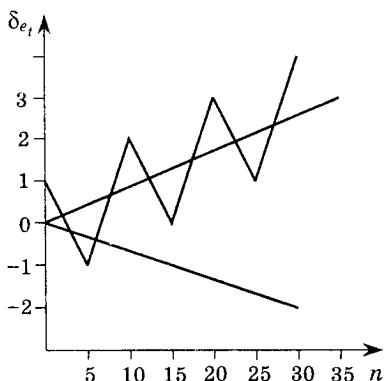


Рис. 4.13. Графік виявлення тренда

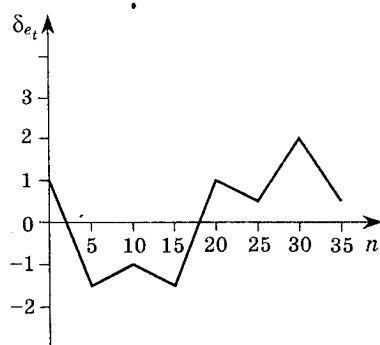


Рис. 4.14. Графік виявлення зсуву рівня

Виявлення змін у дисперсії помилки

Перевіряють гіпотезу $H_0: \delta_1^2 = \delta_2^2 = \dots = \delta_n^2$ про постійність дисперсії помилки у різних дослідках, тобто

$$\text{cov}[E] = Z_1 = \text{diag}\{\delta_1^2, \dots, \delta_n^2\}.$$

Непостійність $D[e]$, як і у випадку непостійності μ , розуміється як обов'язкова наявність залежності $D[e]$ від факторів $\hat{Y}$, $f(x_j)$, t .

Для виявлення непостійності дисперсії помилки можна використовувати розмахи залишків. Аналіз розмаху залишків змушує припустити, що дисперсія за двох найбільших значень x_t істотно відрізняється від дисперсії за решти значень цієї змінної.

У якості графічної процедури перевірки пропонується провести аналіз графіків залежностей типу

$$e_i^2 = f(i), \quad e_i^2 = f(x_{ij}), \quad e_i^2 = f(\hat{y}_i), \quad i = \overline{1, n}, \quad j = \overline{1, m_0},$$

де n — кількість експериментальних даних, m_0 — кількість груп по n_0 значень у кожній групі.

При постійності дисперсій помилки, тобто при дотриманні нуль-гіпотези $H_0: \delta_1^2 = \delta_2^2 = \dots = \delta_n^2 = \delta^2$, графіки залишків матимуть вигляд, як на рис. 4.15, в іншому разі — як на рис. 4.16.

Для перевірки правильності графічної інтерпретації, необхідно використовувати аналітичну тестову процедуру, зокрема критерії Кохрена або Бартлетта.

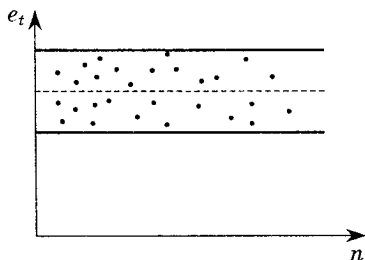


Рис. 4.15. Графік залишків
за умови $\delta = \text{const}$

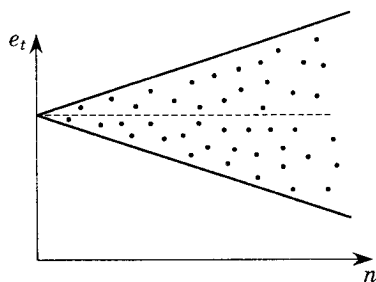


Рис. 4.16. Графік залишків
за умови $\delta \neq \text{const}$

При використанні *критерію Кохрена* треба n значень e_i розбити на m_0 груп (зазвичай, $m_0 = 3 \dots 5$) по n_0 значень у кожній групі. Далі для кожної j -групи слід обчислити значення s_j^2 за формулою:

$$s_j^2 = \frac{1}{n_0 - 1} \sum_{i=1}^{n_0} (\tilde{e}_{ij} - \hat{\mu}_j)^2,$$

де $\tilde{e}_{ij}$ — значення e_i , що потрапили до j -групи

$$\hat{\mu}_j = n_0^{-1} \sum_{i=1}^{n_0} e_{ij}.$$

На підставі отриманих значень s_j^2 треба обчислити значення статистики Кохрена

$$G_p = \frac{\max s_j^2}{\sum_{j=1}^m s_j^2}. \quad (4.25)$$

При дотриманні нуль-гіпотези величина (4.25) має розподіл Кохрена.

Якщо розрахункове значення G_p , обчислене за формулою (4.25), менше за $G_{\text{табл}}$ для відповідного рівня значущості та числа степенів вільності, то гіпотеза про постійність дисперсії не суперечить нашим даним.

У подальшому необхідно досліджувати залишки для підтвердження гіпотези про нормальність їх розподілу. Для цього часто використовують імовірнісний папір. Якщо його нема, здійснюють аналіз гістограм, побудованих на рівномірних інтервалах, з цією метою використовують критерій Пірсона.

Критерій Кохрена застосовується до виборок, що підкоряються нормальному закону розподілу помилок. Критерій доцільно використовувати за умови рівності чисельності всіх виборок $n_{01} = n_{02} = \dots = n_{0m}$.

Для виявлення різниці двох та більшого числа дисперсій широко використовується критерій Бартлетта, що полягає у порівнянні логарифма середньої дисперсії з сумою логарифмів окремих дисперсій. Формули, необхідні для прийняття цього критерію, ґрунтуються на гіпотезі $H_0: \delta_1^2 = \delta_2^2 = \dots = \delta_n^2 = \delta^2$ та припущенні, що вимірювані змінні розподілені за нормальним законом.

Якщо гіпотеза правильна, то об'єднана оцінка

$$S^2 = \frac{\sum_{i=1}^n v_i s_i^2}{\sum_{i=1}^n v_i} = \frac{1}{\sum_{i=1}^n p_i - n} \sum_{i=1}^n (p_i - 1) s_i^2$$

(P_i — число повторювань у вибірці) розподілена як s^2 із середнім значенням δ^2 та v степенями вільності, де $V = \sum_{i=1}^n v_i$.

Величина

$$\lambda = -\frac{1}{c} \sum_{i=1}^n v_i \ln \frac{s_i^2}{s^2}, \quad (4.26)$$

де $c = 1 + \frac{1}{3(n-1)} \left(\sum_{i=1}^n \frac{1}{v_i} - \frac{1}{\sum_{i=1}^n v_i} \right)$, приблизно підкоряється χ^2 -роз-

поділу з $(n-1)$ степенями вільності. Для великих значень v_i величина $c \approx 1$.

Для окремого випадку, коли v_i рівні, так що

$$\sum v_i = nv_i, \quad \chi^2 \approx \frac{1}{c} nv_i \ln S^2 - \frac{1}{n} \sum_{i=1}^n \ln S_i^2, \quad (4.27)$$

де $c = 1 + [n+1]/3nv_i$. Якщо χ^2 перевищує значення $\chi_{1-\alpha}^2$ для $(n-1)$ -го степеня вільності, то використовується гіпотеза, яка стверджує, що $\delta_1^2 = \delta_2^2 = \dots$, відкидається.

Перевірка постійності математичного сподівання регресійних помилок $H_0: M[e_i] = \mu = \text{const}$

Перевіряється гіпотеза $H_0: \mu = \text{const}$, а не гіпотеза $H_0: \mu = 0$, тому що $\mu = \text{const} \neq 0$, то постійне математичне сподівання регресійної помилки входить у вільний член емпіричного порівняння регресії та, отже, розрізнити випадки $\mu = 0$ та $\mu = \text{const} \neq 0$ не видається можливим.

Для візуальної перевірки можна використовувати графіки типу $e_i = f(i)$, $e_i = f(x_{ij})$, $e_i = f(\hat{y}_i)$.

У випадку виконання нуль-гіпотези графіки повинні мати вигляд, як на рис. 4.17.

Графіки на рис. 4.12б,в,г свідчать про лінійну або квадратичну залежність μ від помилки відповідного фактора.

Ефективність графічної процедури можна підвищити, якщо перейти до групування значень $\tilde{e}_i$. Усі $\tilde{e}_i$ розбиваються на m груп по n_0 значень у кожній групі. У подальшому можна використовувати й аналітичний критерій, суть якого полягає у наступному. Якщо $M[e]$ постійне, то різниця між відхиленнями $\tilde{e}_i$

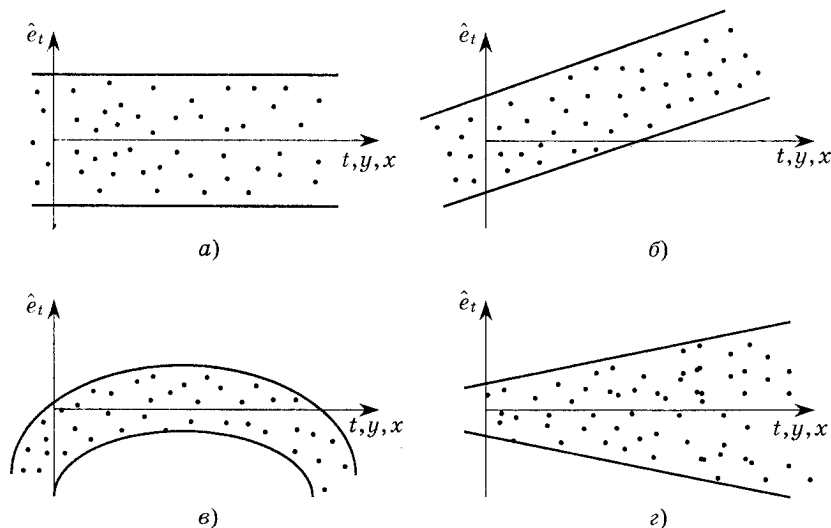


Рис. 4.17. Графічна процедура перевірки гіпотези постійності математичного сподівання

щодо загального середнього $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n \tilde{e}_i$ та відхиленнями групових

середніх $\hat{\mu}_j$ щодо $\hat{\mu}$ повинні бути статистично незалежні.

Значення S_*^2 , S_{**}^2 обчислюються за формулами

$$S_*^2 = \frac{\sum_j^m (n_0 - 1) \cdot S_j^2}{n - m}, \quad S_j^2 = \frac{1}{n_0 - 1} \sum_i^{n_0} (\tilde{e}_{ji} - \hat{\mu}_j)^2,$$

$$S_{**}^2 = \frac{1}{m - 1} \sum_{j=1}^m n_j (\hat{\mu}_j - \hat{\mu})^2,$$

де n_0 — число значень $\tilde{e}_i$ в j -групі.

Величина $F_p = \frac{S_{**}^2}{S_*^2}$ має F -розподіл із числом степенів вільності $(m - 1)$ та $(n - m)$ при дотриманні нуль-гіпотези.

Розрахувавши F_p та задавшись α за таблицями, знаходимо F_T . Якщо $F_p < F_T$, то дані не суперечать гіпотезі про постійність математичного сподівання регресійних помилок e_i .

Інтерпретація коефіцієнта Дарбіна — Ватсона

Кореляція між залишками оцінюється за *критерієм Дарбіна — Ватсона*. Розглянемо його.

Якщо залишки корельовані, оцінки дисперсій параметрів за методом найменших квадратів виявляються зміщеними та не-ефективними, то для встановлення наявності кореляції залишків використовується критерій Дарбіна — Ватсона. Цей критерій не є вельми строгим, тому що він вимагає відсутності статистичної залежності залишків від X . Обчислюється статистика D для серії з n вимірювань:

$$D = \frac{\sum_{i=2}^n (E_i - E_{i-1})^2}{\sum_{i=1}^n E_i^2},$$

де E_i — залишок $(y_i - \hat{y}_i)$ в момент часу i , а $E_i - E_{i-1}$ — права послідовна різниця.

На рис. 4.18 представлені розподіл D та області, в яких гіпотеза про серіальну кореляцію приймається або відкидається.

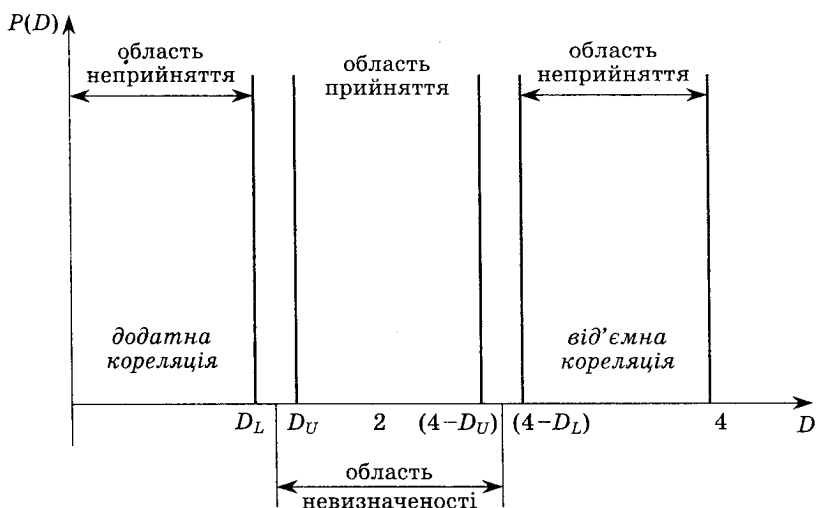


Рис. 4.18. Розподіл D для оцінки серіальної кореляції

Якщо розрахункове значення D виявляється менше, ніж D_L , або перевищує значення $4 - D_L$, то слід вважати, що серіальна кореляція має місце. Якщо величина D перебуває у межах інтервалу між D_U та $4 - d_U$, кореляція відсутня. В областях, позначених знаком питання, цей критерій не дозволяє прийняти певне рішення.

Таблиця 4.6

Статистика Дарбіна — Ватсона для серіальної кореляції

Обсяг вибірки, n	Імовірність на верхньому хвості розподілу	Значення D_L та D_U за Дарбіном та Ватсоном									
		$K - 1$		$K - 2$		$K - 3$		$K - 4$		$K - 5$	
		D_L	D_U	D_L	D_U	D_L	D_U	D_L	D_U	D_L	D_U
15	0,01	0,81	1,07	0,70	1,25	0,59	1,46	0,49	1,70	0,39	1,96
	0,025	0,95	1,23	0,83	1,10	0,71	1,61	0,59	1,84	0,48	2,09
	0,05	1,08	1,36	0,95	1,51	0,82	1,75	0,69	1,97	0,56	2,21
20	0,01	0,95	1,15	0,86	1,27	0,77	1,41	0,68	1,57	0,60	1,74
	0,025	1,08	1,28	0,99	1,41	0,89	1,55	0,79	1,70	0,70	1,87
	0,05	1,20	1,41	1,10	1,54	1,00	1,68	0,90	1,83	0,79	1,99
25	0,01	1,05	1,21	0,98	1,30	0,90	1,41	0,83	1,52	0,75	1,65
	0,025	1,18	1,34	1,10	1,43	1,02	1,54	0,94	1,65	0,86	1,77
	0,05	1,29	1,45	1,21	1,55	1,12	1,56	1,04	1,77	0,95	1,89

Закінчення табл. 4.6

Обсяг вибірки, n	Імовірність на верхньому хвості розподілу	Значення D_L та D_U за Дарбіном та Ватсоном									
		$K - 1$		$K - 2$		$K - 3$		$K - 4$		$K - 5$	
		D_L	D_U	D_L	D_U	D_L	D_U	D_L	D_U	D_L	D_U
30	0,01	1,13	1,26	1,07	1,34	1,01	1,42	0,94	1,51	0,88	1,61
	0,025	1,25	1,38	1,18	1,46	1,12	1,54	1,05	1,63	0,98	1,73
	0,05	1,35	1,49	1,28	1,57	1,21	1,65	1,14	1,74	1,07	1,83
40	0,01	1,25	1,34	1,20	1,40	1,15	1,46	1,10	1,52	1,05	1,58
	0,025	1,35	1,45	1,30	1,51	1,25	1,57	1,20	1,63	1,15	1,69
	0,05	1,44	1,54	1,39	1,60	1,34	1,66	1,29	1,72	1,23	1,79
50	0,01	1,32	1,40	1,28	1,45	1,21	1,49	1,20	1,51	1,16	1,59
	0,025	1,42	1,50	1,38	1,51	1,31	1,59	1,30	1,61	1,26	1,69
	0,05	1,50	1,59	1,46	1,63	1,42	1,67	1,38	1,72	1,34	1,77
60	0,01	1,38	1,45	1,35	1,48	1,32	1,52	1,28	1,56	1,25	1,60
	0,025	1,47	1,54	1,44	1,57	1,40	1,61	1,37	1,65	1,33	1,69
	0,05	1,55	1,62	1,51	1,65	1,48	1,69	1,44	1,73	1,41	1,77
80	0,01	1,47	1,52	1,44	1,54	1,42	1,57	1,39	1,60	1,36	1,52
	0,025	1,51	1,59	1,53	1,62	1,49	1,65	1,47	1,67	1,44	1,70
	0,05	1,61	1,66	1,59	1,69	1,56	1,72	1,53	1,74	1,51	1,77
100	0,01	1,52	1,56	1,50	1,58	1,48	1,60	1,46	1,63	1,44	1,65
	0,025	1,59	1,63	1,57	1,65	1,55	1,67	1,53	1,70	1,51	1,72
	0,05	1,65	1,69	1,63	1,72	1,61	1,74	1,59	1,76	1,57	1,78

Імовірність, зазначена у другому стовпчику таблиці, дорівнює площі під верхнім хвостом розподілу. K — число незалежних змінних, окрім постійного члену.

Мультиколінеарність змінних

Одне з основних припущень регресійного аналізу (РА) відноситься до матриці вихідних даних. Воно формулюється так: серед незалежних змінних не повинно бути лінійно залежних. Ця вимога необхідна для обчислення оцінки $B = (X^T X)^{-1} X^T Y$ методом найменших квадратів.

Мультиколінеарність приводить до:

☑ зниження точності, дисперсія оцінок збільшується, параметри моделі корельовані, що призводить до труднощів у інтерпретації моделі;

☑ оцінки коефіцієнтів стають чутливими до особливостей множини вибірових даних.

Причиною мультиколінеарності можуть служити:

- а) наявність автокореляції у ряді спостережень;
- б) кореляція між змінними;
- в) високий рівень перешкод.

Під мультиколінеарністю будемо розуміти сполученість незалежних змінних, це означає «майже лінійну залежність» векторів $x_1, \dots, x_k$, тобто існування чисел $v_1, \dots, v_k$ таких, що

$$v_1 x_1 + v_2 x_2 + \dots + v_k x_k \approx 0. \quad (4.28)$$

Коли рівність (4.28) має місце, говорять про строгу мультиколінеарність.

За наявності мультиколінеарності оцінки МНК стають додатними, тобто дисперсія оцінок буде вельми великою.

За наявності (4.28) матриця $(X'X)$ стає погано зумовленою, зокрема $|X'X| \approx 0$, тобто $\lambda_{\min}(X'X) = \lambda_1 \approx 0$.

Взагалі, мультиколінеарність — поняття досить багатогранне, й важко припустити міру, яка в усіх відношеннях була добра.

Розглянемо п'ять різних мір. Перші три з них — характеристики матриці плану $X'X$.

1. За наявності лінійної залежності (4.28) матриця $X'X$ стає погано зумовленою, тобто визначник близький до 0. Тоді визначник $|X'X| = \prod_{i=1}^k \lambda_i$ може виступати в якості міри мультиколінеарності.

2. Мінімальне характеристичне число матриці $(X'X)\lambda_{\min} = \lambda_1$. Чим менше λ_1 , тим сильніша мультиколінеарність.

3. Міра зумовленості матриці за Непманом — Голдстейном полягає у відношенні $\lambda_{\max}(X'X)/\lambda_{\min}(X'X)$. Експериментальним шляхом встановлено, що за $P = \lambda_{\max}/\lambda_{\min} \geq 50$ оцінки параметрів моделі, отримані за такими даними, стають зсунутими та неефективними.

Для розрахунку чисел зумовленості рекомендується застосовувати норму Фробеніуса

$$|X|_{\Phi} = \sqrt{S_p(X'X)} = \sqrt{S_p(XX')} = \sqrt{\sum_i \sigma_i^2(x)}.$$

Використання для оцінки зумовленості норми Фробеніуса має ту перевагу, що порядок числа зумовленості в цьому випадку показує, з якою точністю необхідно робити обчислення, щоб отримати стає рішення.

У якості числових характеристик зумовленості використовують числа Тюрінга, числа Тодда.

Вихід з глухого куту мультиколінеарності полягає, з одного боку, у пошуку нових даних, з іншого — у розробці спеціальних методів оцінювання.

Якщо в матриці $A = X^T X$ змінні нормалізовані, тобто

$$\sum_{i=1}^n X_{ji} = 0, \quad \sum_{i=1}^n X_{ij}^2 = 1, \quad j = \overline{1, m},$$

тоді A виявляється кореляційною матрицею R та її визначник дорівнює $\det R$. А за величиною R можна судити про наявність мультиколінеарності, тобто

☑ якщо в X стовпчики лінійно незалежні $\det R = 1$;

☑ якщо впорядкувати за величиною власні значення матриці R , тобто $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_m$, і розглянемо числа зумовленості

$$M = m \max_{i,j} |a_{ij}| \max_{i,j} |a_{ij}^{-1}|, \quad P = \frac{\lambda_m}{\lambda_1},$$

де a_{ij}^{-1} — елемент матриці A^{-1} .

Систему рівнянь та матриці з великими значеннями чисел зумовленості M та P називають погано зумовленими. В цьому випадку недіагональні елементи матриці R близькі до одиниці, тобто спостерігається щільний кореляційний зв'язок. Міра зумовленості $\nu(R)$ матриці R може вимірюватися через норми матриць R та R^{-1} , тобто

$$\nu(R) = |R| \cdot |R^{-1}|.$$

Візьмемо норми матриці R^{-1} величину $\text{tr}(R^{-1})$. Тоді

$$\text{tr}(R^{-1}) = \sum_{i=1}^k \lambda_i^{-1},$$

якщо $\lambda_1 \rightarrow 0$. Корисною інформацією, яку можна здобути з матриці R^{-1} , є величина $R_j = (1 - C_{jj}^{-1})^{1/2}$, де C_{jj} — діагональний елемент матриці R^{-1} . Величина R_j є аналогом коефіцієнту множинної кореляції між незалежною змінною x_j та рештою незалежних змінних. Якщо $C_{jj} = 1$, $R_j = 0$, то змінна x_j ортогональна решті змінних. Якщо C_{jj} , то $R = 1$, тоді x_j виявляється лінійною комбінацією з рештою змінних. Якщо модель стохастична, тобто $X^T X$ — стохастична матриця. Перевірку мультиколінеарності

можна здійснювати наступним способом за числом змінних k і числом спостережень n

$$x^2 = - \left[n - \frac{1}{6}(2k+5) - 1 \right] \ln \det R.$$

Передбачається, що для великого числа спостережень ця величина підкоряється x^2 -розподілу з $\gamma = 1/2m(m-1)$ степенями вільності.

Таким чином, якщо у дослідника є можливість обирати велику кількість незалежних змінних x , то він повинен обирати їх так, щоб виконувалася умова $X_j^T X_k = 0$, $j \neq k$, тобто стовпчики матриці X повинні бути ортогональними.

На практиці, коли мають справу з пасивним експериментом, тобто дослідник може лише спостерігати (реєструвати, виміряти) дані, але не формує їх сам, необхідно досліджувати матриці R , R^{-1} й власні значення кореляційної матриці. Це дозволяє лише виявити мультиколінеарність, але не усунути її.

Розглянемо деякі рекомендації щодо усунення мультиколінеарності.

Рекомендації щодо усунення мультиколінеарності

Найбільш простий спосіб усунення мультиколінеарності — виключення однієї змінної з пари змінних, коефіцієнт кореляції між якими більший за 0,8.

Найпростіші рекомендації з усунення мультиколінеарності зводяться до скорочення великої кількості пояснювальних змінних, що розглядаються, за рахунок тих з них, які лінійно зв'язані з вже включеними у модель. Виконання цих рекомендацій веде до побудови скороченої моделі, яка не завжди відповідає умові спостережуваності й керованості.

Щоб уникнути небажаних ефектів мультиколінеарності, зберігши при цьому весь набір пояснювальних змінних, що цікавлять нас, у [32] пропонується збільшити розміри вибірки шляхом отримання додаткової інформації. Зрозуміло, що не будь-яке довільне збільшення вибірки веде до послаблення ефектів мультиколінеарності. В [33] приводяться результати, що дозволяють визначити напрямок, у якому треба шукати додаткові дані, що забезпечують збільшення ефективності оцінок невідомих параметрів регресійної моделі. Однак, практичне застосування цих

результатів можливе, здебільшого, лише в тих випадках, коли додаткові дані можна отримати шляхом проведення активного експерименту, який не є характерним для функціонуючих технологічних процесів. Проводити спостереження необхідно у напрямку максимізації власних векторів, які відповідають малим (або нульовим) власним значенням.

Зауважимо, що *власним значенням* (або власним числом) *оператора* B називається таке число λ , яке для деякого ненульового вектора X має місце рівність $BX = X\lambda$.

Часто для усунення мультиколінеарності використовують прийоми, основані на попередньому перетворенні вихідних даних шляхом отримання відхилень від тренда. Однак, регресійна модель, отримана завдяки таким перетворенням, слабо піддається інтерпретації. Іноді використовують апіорну інформацію про існуючі між параметрами зв'язки у вигляді обмежень при обчисленні параметрів регресії, що оцінюються. За винятком найпростіших випадків, реалізація цих підходів досягається істотним ускладненням обчислювальної процедури знаходження оцінок.

Широкого розповсюдження отримали методи усунення мультиколінеарності, основані на заміні великої вихідної кількості пояснювальних змінних головними компонентами з подальшим відкиданням тих з них, які є незначущими у рівнянні регресії. Однак, як показано в [12], відкидання головних компонент, незначущих у рівнянні регресії при вихідних пояснюючих змінних не підвищує їх ефективності.

Близькими до розглянутого можна вважати методи, основані не на компонентному, а на факторному аналізі, причому аналогія простежується як за перевагами, так і за недоліками.

У випадках, коли перелік пояснювальних змінних регресійної моделі надто великий, в [32] рекомендується розділити їх на групи з високою кореляцією і в кожній групі побудувати узагальнюючі фактори у вигляді головної компоненти, які надалі використовуватимуться як нові змінні моделі, що будується.

4.3. Методи змішеного оцінювання

Останнім часом значного поширення набули дослідження з теорії змішеного оцінювання та її застосування для випадків, коли змінні, що включаються у модель, сильно корельовані.

Незміщені оцінки з мінімальною дисперсією, визначені за методом МНК з системи нормальних рівнянь

$$\hat{\beta} = (X'X)^{-1} X'Y,$$

достатньо добре вивчені. Розглянемо випадок, коли матриця $X'X$ відрізняється від одиничної. Про ненадійність оцінок при поганій зумовленості $X'X$ будемо судити за значенням власних чисел.

Нехай власні числа $X'X$ розташовані за порядком спадання $\lambda_{\max} = \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p = \lambda_{\min}$. Тоді середнє значення та дисперсію квадрата відстані оцінок від істинних параметрів можна подати через власні числа

$$E(L_1^2) = \sigma^2 \sum_{j=1}^p \lambda_j^{-1} \quad \text{та} \quad \text{VAR}(L_1^2) = 2\sigma^4 \sum_{j=1}^p \lambda_j^{-2},$$

де межами для математичного сподівання і дисперсії L_1^2 є $\sigma^2/\lambda_{\min}$ та $2\sigma^4/\lambda_{\min}^2$ відповідно. Таким чином, якщо $X'X$ має хоча б одне мале власне число, відстань між $\hat{\beta}$ та β матиме тенденцію до швидкого зростання.

Ця обставина стає вирішальною у процесі ідентифікації, базованій на регресійному аналізі, при застосуванні якого цілком можливе отримання невірних знаків оцінок параметрів. У цьому випадку для оцінок параметрів рівняння регресії доцільно використовувати апарат зміщеного оцінювання, основною ідеєю якого є відмова від вимоги незміщеності, тобто $E(\hat{\beta}^*) \neq \beta$, де $\hat{\beta}^*$ — будь-яка зміщена оцінка.

За рахунок відмови від вимоги незміщеності вдається істотно зменшити вплив корельованості входів, який обмежує можливість використання моделей, отриманих на основі даних пасивного експерименту. Крім того, за рахунок введення у розглядання фактора деформації зміщені оцінки набувають гнучкості, дозволяють адаптувати алгоритм обробки інформації до характеристик процесу. Ця особливість є важливою з огляду на те, що в адаптивних системах управління безперервними процесами ідентифікацію параметрів моделі доводиться вести частіше за все на ковзній базі при змінюваних статистичних характеристиках та корельованості входів.

Детальний виклад основних моментів зміщеного оцінювання (загальний підхід, рідж-оцінки, редуковані оцінки, оцінки Марквардта, оцінки Хокінса), а також деякі результати порівняльного аналізу цих оцінок методом Монте-Карло можна знайти у [33].

Для випадків, коли коваріаційні матриці недостатньо зумовлені, тобто проявляється один з небажаних ефектів мультиколінеарності, для обчислення оцінок коефіцієнтів регресії доцільно використовувати метод регуляції А. М. Тихонова [32]. Цей метод дозволяє з загальних позицій підійти до отримання відомих раніше процедур регуляризованого оцінювання, які за своєю суттю є методами зміщеного оцінювання. Через зміщеність отримуваних оцінок виникає необхідність у побудові такого процесу регуляризації, який не призводив би до зміни чи зникнення змістовного сенсу коефіцієнтів регресії. На жаль, результатів щодо цілеспрямованого добору регуляризуючих процесів, що забезпечують потрібну інтерпретацію остаточного варіанту моделі, поки що немає.

Мультиколінеарність веде до збільшення оцінок. За наявності значної мультиколінеарності оцінка МНК стає настільки незадовільною, що навіть знаки деяких координат часто не відповідають істинним.

Задачу оцінювання можна сформулювати таким чином: знайти оцінку параметрів регресії, яка була б сталою навіть за умови сильної залежності вхідних змінних.

Слід зазначити, що на сьогодні не існує критерію, який дозволяє за ступенем зумовленості вхідних даних однозначно вирішити питання щодо можливості використання того чи іншого методу оцінювання.

Вплив мультиколінеарності на властивості оцінок моделі, що отримані у вигляді $\hat{B} = (X'X)^{-1} \cdot X' \cdot Y$, легко продемонструвати, досліджуючи коваріаційну матрицю оцінок і відстань від $\hat{B}$ до B .

$$\text{VAR}(\hat{B}) = \sigma^2 (X'X)^{-1}. \quad (4.29)$$

Квадрат помилки оцінки коефіцієнтів визначається таким чином:

$$L_1^2 = (\hat{B} - B)'(\hat{B} - B), \quad (4.30)$$

$$E[L_1^2] = \sigma^2 \text{Trace}(X'X)^{-1}, \quad (4.31)$$

де L_1 — відстань від $\hat{B}$ до B .

Якщо помилка E нормально розподілена, то

$$\text{VAR}[L_1^2] = 2\sigma^4 \text{Trace}(X'X)^{-2}. \quad (4.32)$$

З огляду на (4.29), (4.31), (4.32) видно, що при зміні $X'X$ від одиничної до погано зумовленої матриці спостерігається невизначеність в оцінці $\hat{B}$.

При зменшенні власних значень матриці $X'X$ коефіцієнти моделі мають тенденцію до зростання. Модель багатовимірної регресії може бути подана в ортогональній формі з використанням сингулярного розкладу

$$X = P\Lambda^{1/2}Q'. \quad (4.33)$$

Матрицю $X'X$ можна подати у вигляді $Q\Lambda Q'$. Нехай вираз $Z = X\hat{Q}$ являє собою перетворення на головні компоненти, коли власні значення впорядковані так, що $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$.

У (4.33) матриця Q — це $(p \times p)$ -матриця, яка складається з власних ортонормованих векторів. Власні значення $X'X$ одержують за $\lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_p)$.

Нехай

$$\alpha = Q'B. \quad (4.34)$$

Якщо $Z = P\Lambda^{1/2}$, то матриця P надасть координати, які стосуються головних осей, нормалізованих так, що кожний стовпчик з P має одиничну довжину. Сингулярні значення $\sqrt{\lambda_i}$ дадуть довжини головних осей, довжина i -ї головної осі дорівнюватиме $2\sqrt{\lambda_i}$.

Тоді (4.34) можна записати так:

$$Y = Z\alpha + E. \quad (4.35)$$

Оцінюючи α за МНК, отримаємо

$$\hat{\alpha} = (Z'Z)^{-1}Z'Y = \lambda^{-1}Z'Y. \quad (4.36)$$

Узагальнена гребенева оцінка визначається за формулою:

$$\hat{\alpha}^*(K) = [A + K]^{-1}Z'Y, \quad (4.37)$$

де $K = \text{diag}[k_1, k_2, \dots, k_p]$, $k_i \geq 0$.

Якщо $K = kI_p$, $k \geq 0$, тоді (4.37) буде гребеневою оцінкою і

$$\hat{\alpha}^*(k) = [\lambda + kI_p]^{-1}Z'Y. \quad (4.38)$$

З (4.37) маємо

$$\hat{Z}'Y = \lambda\hat{\alpha}. \quad (4.39)$$

Використовуючи визначення Z і α , вираз (4.37) та (4.38) можна переписати у вигляді

$$\hat{B}^*(k) = [X'X + QKQ']^{-1} X'Y; \quad (4.40)$$

$$\bar{\bar{B}}^*(k) = [X'X + kI_p]^{-1} X'Y. \quad (4.41)$$

Доведено, що для того, щоб побудувати оцінки, які були б добрі у випадку мультиколінеарності, необхідно відмовитись від умови незміщеності. Засоби, які дозволяють оцінювати коефіцієнти моделі в умовах мультиколінеарності, отримали назву *зміщених*. У класі зміщених моделей можна побудувати методи, стійкі до мультиколінеарності. До таких оцінок передусім належать рідж-оцінки (від англ. *ridg* — гребінь, хребет), або гребеневі оцінки, запропоновані Хоерлом та Кеннордом, оцінки головних компонент з узагальненнями, стислі оцінки.

У загальному випадку задача оцінювання для зміщених оцінок формулюється таким чином: знайти оцінку параметрів регресії, яка була б сталою при сильній мультиколінеарності.

У класі зміщених оцінок не можна знайти оптимальні оцінки у розумінні мінімальної матриці середніх сум квадратів відхилень

$$L^2(b) = M(\hat{b}^* - b)'(\hat{b}^* - b).$$

Якщо розглядати клас лінійних оцінок, то *середня сума квадратів відхилень* (ССКВ) є обмеженою для усіх X .

Для отримання точкових оцінок з меншою середньоквадратичною помилкою використовується вираз

$$\hat{B} = (X'X + kI_p)^{-1} X'Y. \quad (4.42)$$

Оцінки, отримані з (4.42), і є *рідж-оцінками*, або *гребеневими оцінками*. У (4.42) $K = kI_p$ — деяка невід'ємна невизначена матриця $P \times P$. Додавання до матриці плану $X'X$ від'ємно визначеної матриці робить її кращою, а оцінки — більш сталими.

Додавання до діагональних елементів матриці $X'X$ деякого невід'ємного числа kI_p або діагональної матриці K приводить до однакового результату. Обчислювальна процедура гребеневого аналізу полягає у такому.

Позначимо

$$D = \begin{bmatrix} (X'X)_{11} & 0 & \dots & 0 \\ 0 & (X'X)_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & (X'X)_{pp} \end{bmatrix}, \quad Z = XD^{-1/2}, \quad \alpha = D^{1/2}.$$

Тоді модель $Y = XB + E$ зведеться до моделі

$$Y = Z + E\alpha, \quad (4.43)$$

Гребеневою оцінкою з матрицею $K = kI_p$ для моделі (4.43) є

$$\begin{aligned} \alpha(k) &= (Z'Z + kI)^{-1} Z'Y = [(D^{-1/2}X)'XD^{-1/2} + kI]^{-1} D^{-1/2} X'Y = \\ &= D^{-1/2} (X'X + K)^{-1} X'Y, \end{aligned}$$

$$K_{jj} = k(X'X)_{jj}.$$

Тому $\alpha(k) = D^{-1/2}B(k)$ являється рідж-оцінкою (4.42) вхідної моделі (4.43) з вибором K за правилом $K_{jj} = k(X'X)_{jj}$, $K_{jl} = 0$ ($j \neq l$), де $k \geq 0$.

У більш простій формі $\alpha(k) = (Z'Z + kI)^{-1} Z'Y$.

При $k = 0$ одержують оцінки МНК. При $k \neq 0$ одержують $\alpha(k) = [I + k(X'X)^{-1}] \alpha$, де α — оцінки МНК, одержувані з (4.43).

У випадках, коли матриця $X'X$ має малі власні значення, ймовірність близькості $\hat{b}$ до істинної оцінки b мала. Аналіз структури факторного простору за коефіцієнтами кореляцій матриці $X'X$ не завжди дозволяє пояснити чутливість результатів до конкретного набору даних. У зв'язку з цим оцінка чутливості результатів обчислень до статистичних характеристик вхідних даних може дати ключ до отримання найбільш ефективних параметрів, які забезпечують мінімальне відхилення $\hat{b}$ від b та залишкової суми квадратів (MSE). Для розв'язання цієї задачі використаємо значення $\hat{b}(k)$ та $MSE(k)$ для деякої множини K .

Якщо прийняти, що B є довірливою оцінкою b , залишкову суму квадратів запишемо у вигляді

$$\begin{aligned} MSE &= (Y - XB)'(Y - XB) = (Y - X\hat{b})'(Y - X\hat{b}) + (B - \hat{b})'X'X(B - \hat{b}) = \\ &= MSE_{\min} + MSE(B). \end{aligned}$$

Таким чином, маємо, що величина MSE дорівнює мінімальному значенню $MSE_{\min}$ плюс значення квадратичної форми в точці $(B - \hat{b})$. Чим гірше зумовлена матриця $X'X$, тим більше підстав очікувати, що вектор $\hat{b}$ буде більшим без істотного збільшення залишкової суми квадратів.

У зв'язку з цим при виборі параметрів K за критерієм мінімуму суми квадратів необхідно рухатися у напрямку, за яким довжина вектора регресії буде найменшою. У [6] ця задача сформульована таким чином: необхідно мінімізувати $B'B$ за умови

$$(B - \hat{b})'X'X(B - \hat{b}) = MSE_0, \quad (4.44)$$

або, зводячи (4.44) до задачі Лагранжа, мінімізувати

$$F = B'B + \left(\frac{1}{k}\right) [(B - \hat{b})'X'X(B - \hat{b}) - MSE_0],$$

де $\left(\frac{1}{k}\right)$ — множник Лагранжа. Далі

$$\frac{\partial F}{\partial B} = 2B + \frac{1}{k[2(X'X(B - 2(X'X)\hat{b}))]} = 0.$$

Це зводиться до $B = \hat{b}[X'X + KI]^{-1}X'Y$, де K обирається таким чином, щоб задовольнялося обмеження (4.44).

Оцінімо вплив параметра K на величину дисперсії і квадрат зміщення гребеневої оцінки від $\hat{b}$. Розглянемо вираз

$$\begin{aligned} E[L_1^2(K)] &= E[(\hat{b}^* - b)'(\hat{b} - b)] = E[(\hat{b}^* - b)'Z'Z(\hat{b}^* - b)] + \\ &+ (Zb - b)'(Zb - b) = \sigma^2 \sum_{i=1}^p \frac{\lambda_i}{(\lambda_i + k)^2} + \\ &+ K^2 b' \sum_{i=1}^p \frac{\lambda_i}{\lambda_i + k} (X'X + KI)^{-2} b = \gamma_1(k) + \gamma_2(k). \end{aligned}$$

Доданок $\gamma_1(k)$ дорівнює сумарній дисперсії оцінок параметрів $\hat{B}^* = Z\hat{B} = Z(X'X)^{-1}X'Y$.

Сумарна дисперсія зменшується зі збільшенням K . Доданок $\gamma_2(k)$ являє собою квадрат відстані від Zb до b . Вона дорівнюватиме 0 за умови $k=0$, оскільки $Z=I$. Інакше кажучи, $\gamma_2(k)$ можна розглядати як квадрат зміщення гребневих оцінок $(\hat{b}^*)$ від

звичайних (b) . Аналіз величини $E[L_1^2(k)]$ показує, що можуть існувати значення K , за яких середньоквадратична помилка для $\hat{b}^*$ виявиться меншою, ніж для розв'язку $\hat{b}$. Значення похідної кожної складової функції $E[L_1^2(K)]$ на початку координат представимо у вигляді:

$$\lim_{k \rightarrow 0} \left(\frac{d\gamma_1}{dK} \right) = -2\sigma^2 \sum_{j=1}^p \frac{\lambda}{\lambda_j^2}.$$

З огляду на це можна перейти в область $K > 0$, отримавши невелике зміщення, але при цьому істотно зменшити дисперсію і завдяки цьому зменшити середньоквадратичну помилку оцінки і прогнозу.

Таким чином, клас зміщених оцінок дозволяє обрати оцінку $\hat{b}^*$ з меншою середньоквадратичною помилкою порівняно зі звичайною оцінкою $\hat{b}$, але зміщеною. У роботах [6, 9, 10] показано, що не можна побудувати чітко визначену автоматичну процедуру для отримання точкової оцінки. Для будь-якого заданого набору даних можна обрати $\hat{b}^*$ краще, ніж $\hat{b}$.

Кожному набору даних, що обробляється за допомогою лінійної регресії, властиві внутрішні обмеження, які слід брати до уваги, обираючи спосіб обробки даних.

Клас алгоритмів зміщеного оцінювання можна умовно розділити на три групи:

- ☒ алгоритми стиснених оцінок;
- ☒ алгоритми оцінок дробового рангу;
- ☒ алгоритми гребневих оцінок.

Відповідно до наведеної вище класифікації, проведемо аналіз алгоритмів стиснених оцінок.

Типовим представником зміщених оцінок є клас стиснених оцінок, які мають вигляд $\hat{\beta}_l^* = l(X'X)^{-1}X'Y = l\hat{\beta}$, де $l \in [0, \infty]$. Якщо l — фіксований скаляр, то $\hat{\beta}$ — *детермінована стиснена оцінка*, якщо ж $l = f(\hat{\beta})$ — скалярна функція від $\hat{\beta}$, то $\hat{\beta}_l$ є *стохастичною стисненою оцінкою*.

Історично першими стисненими оцінками були оцінки *Стейна* [12]. Для них коефіцієнт стиснення не залежить від невідомих параметрів β , σ^2 . Недоліком стиснених оцінок Стейна є така обставина: за недіагональної матриці $X'X$ деякі β_i можуть змінювати знак більш ніж один раз, а також зростати за величиною. Розвитком оцінок Стейна є двопараметричні оцінки Стейна,

оцінки Склове [18], оцінки, запропоновані С. А. Гінсбергом [19], та деякі інші. У роботах Н. І. Андреева [20, 21] запропоновані стиснені оцінки середнього.

На жаль, визначення найбільш ефективної оцінки серед стиснених оцінок теоретично неможливо, найбільш ефективну оцінку для кожної конкретної задачі можна виділити лише шляхом моделювання.

Оцінки дробового рангу є розвитком оцінок узагальненого обігу, які також належать до зміщених оцінок. Слід зауважити, що інтерпретація оцінок узагальненого обернення як зміщених дозволяє отримати більш надійні результати. Оцінки узагальненого обернення визначаються як $\hat{\beta}_r^+ = A_r^+ X'Y$, де A_r^+ — узагальнена обернена матриця рангу r .

Для оцінок дробового рангу справедлива наступна умова припустимості. Для того, щоб середньоквадратична помилка оцінок дробового рангу була меншою, ніж варіація МНК-оцінок, необхідним є виконання умови

$$\sum_{j=r+1}^p \lambda_j^{-1} > \frac{1}{\sigma^2} (\hat{\beta}' \hat{\beta}).$$

Слід зазначити можливість комбінування оцінок дробового рангу з гребеновими $\hat{\beta}_{rk}^+ = (A_r^+ + kl)^{-1} X'Y$. Тут r обирається так, щоб виключити нульові власні значення, а k — за будь-яким із способів вибору k для гребенових оцінок.

Розглянемо клас методів гребенового оцінювання. Найбільш розповсюдженими та одночасно найбільш типовими алгоритмами групи зміщених оцінок є сімейство гребенових оцінок зі скалярною корекцією.

Окрім А. М. Тихонова, значних результатів у галузі побудови регуляризуючих алгоритмів домоглися Хоерл та Кеннард. Їх підхід було названо гребеновою (*ridge*-) регресією. Назва «гребенева регресія» пов'язана з тим, що довірчі еліпси при різних рівнях значущості геометрично можуть утворювати гребінь. Метод гребенової регресії можна й треба називати регуляризованим. Саме Хоерл та Кеннард надалі займалися питаннями коректного застосування регресійного аналізу в умовах мультиколінеарності і назвали розроблений *метод гребеновою регресією* [33].

Ці оцінки були вперше описані у вигляді

$$\hat{\beta}_k = (X'X + kl)^{-1} X'Y = WX'Y.$$

Природним узагальненням гребневих оцінок є оцінки вигляду $\hat{\beta}_c = (X'X + C)^{-1} X'Y$, де C — симетрична додатно визначена матриця.

Поряд з цим, Хоерл та Кеннард запропонували багатопараметричне сімейство узагальнених гребневих оцінок, що відрізняються формуванням матриці деформації $K = \text{diag}\{k_j\}_{j=1}^m$ у виразі $\beta(k) = (X'X + K)^{-1} X'Y$, та ітераційну процедуру оцінювання параметра K_j , виходячи з графічного зображення гребеневого сліду коефіцієнтів.

Теорія зміщених оцінок виявилась дуже плідним напрямком розв'язання деяких задач математичної статистики. Проте до цього часу предметом досліджень являється вибір параметрів деформації.

Слід зауважити, що розмаїття процедур визначення параметра деформації k_j породило й велику кількість методів зміщеного оцінювання. Покажемо, що тип алгоритму, його ефективність та галузь використання здебільшого залежать від вибору фактора деформації.

Дослідимо проблему вибору параметра k_j у загальній гребеневій регресії. Для розв'язання цієї задачі розглянемо проблему статистичної мінімізації критеріїв $E(L_1^2)$ та $E(L_2^2)$. Для деякого фіксованого $K > 0$ гребенева оцінка має вигляд:

$$b^* = (X'X + kI_p)^{-1} X'Y. \quad (4.45)$$

Зауважимо, що оцінка, отримана за (4.45), зводить матрицю $X'X$ до діагонального вигляду за допомогою деякого ортогонального перетворення P , тобто $P'(X'X)P = \lambda$, де λ — діагональна матриця власних значень матриці $X'X$, а стовпчики матриці P є відповідні до них власні вектори.

Уведемо позначення $X^* = XP$ та $b = P\alpha$. Лінійну модель будемо подавати у вигляді $Y = X^*\alpha + E$. Зауважимо, що $X^{*'}X^* = \lambda$. Тоді статистичну оцінку параметра α за методом МНК можна подати рівнянням

$$\hat{\alpha} = \Lambda^{-1} X^{*'} y, \quad (4.46)$$

а узагальнену гребеневу статистичну оцінку задати виразом

$$\alpha^* = (\Lambda + k)^{-1} X^{*'} y, \quad (4.47)$$

де k — діагональна матриця з невід'ємними діагональними елементами k_j .

У цьому випадку для того, щоб отримати узагальнену гребеневу статистичну оцінку для b , припустимо

$$b^* = P\alpha^*. \quad (4.48)$$

Розглянемо два критерії для визначення параметра α^* в узагальненій гребеневій регресії:

$$E(L_1^2) = E(\tilde{\alpha} - \alpha)'(\tilde{\alpha} - \alpha), \quad E(L_2^2) = E(\tilde{\alpha} - \alpha)' \Lambda (\tilde{\alpha} - \alpha),$$

де $\tilde{\alpha}$ є деякою оцінкою параметра α .

У [12, 32, 33, 34] показано, що існує таке $k > 0$, для якого виконується співвідношення $E(b^* - b)'B(b^* - b) \leq E(\hat{b} - b)'B(\hat{b} - b)$. Його результати легко поширюються на узагальнену гребеневу регресію: існує деяка матриця $k > 0$, для якої

$$E(\alpha^* - \alpha)' \Lambda (\alpha^* - \alpha) \leq E(\hat{\alpha} - \alpha)' \Lambda (\hat{\alpha} - \alpha).$$

Легко встановити, що для узагальненої гребеневої регресії обидва критерії $E(L_1^2)$ та $E(L_2^2)$ мінімізуються за умови

$$k_j = \frac{\sigma^2}{\alpha_j^2}. \quad (4.49)$$

Проблема полягає в тому, що необхідно оцінити параметри k_j за наявними даними, оскільки параметри α_j та σ^2 не відомі.

Для отримання початкових значень k_j у (4.49) застосовувалися статистичні оцінки для σ^2 та α_j . Ці значення надалі були використані у (4.47) для корекції оцінених α_j , а потім у (4.49) з метою отримання наступного набору параметрів k_j . Можна показати, що ця процедура допускає розв'язання, точно виражене у замкненій формі через початкові значення

$$e_{i(0)} = \frac{\hat{\sigma}^2}{\lambda_i \hat{\alpha}_j^2}. \quad (4.50)$$

У якості альтернативи, замість оцінювання параметрів k_j з виразу (4.49), побудуємо статистичну оцінку критерію, а потім мінімізуємо її, використовуючи методи оптимізації.

Для розв'язання цієї задачі розглянемо процедуру отримання $\min \hat{L}_1^2$ та $\hat{L}_2^2$. Перепишемо величину $\hat{L}_1^2$ у вигляді

$$\hat{L}_1^2 = E(\alpha^* - \alpha)'(\alpha^* - \alpha) = \sum_{i=1}^p \frac{\lambda_i \sigma^2 + k_j^2 \alpha_j^2}{(\lambda_j + k_j)^2},$$

$$E(\alpha^* - \hat{\alpha})'(\alpha^* - \hat{\alpha}) = \sum_{j=1}^p k_j^2 \left(\frac{\sigma^2}{\lambda_j + \alpha_j^2} \right) / (\lambda_j + k_j)^2,$$

причому

$$E(\alpha^* - \alpha)'(\alpha^* - \alpha) = E(\alpha^* - \hat{\alpha})'(\alpha^* - \hat{\alpha}) + \sigma^2 \sum_{j=1}^p \frac{\lambda_j - k_j}{\lambda_j(\lambda_j + k_j)}. \quad (4.51)$$

Отже,

$$\hat{L}_1^2 = (\alpha^* - \hat{\alpha})'(\alpha^* - \hat{\alpha}) + \hat{\sigma}^2 \text{Tr}\{(\Lambda - k)\Lambda^{-1}(\Lambda + k)^{-1}\} \quad (4.52)$$

є незміщеною статистичною оцінкою критерію $E(L_1^2)$ для узга-
льненого гребеня. Таким самим чином отримуємо

$$E(\alpha^* - \alpha)' \Lambda (\alpha^* - \alpha) = E(\alpha^* - \hat{\alpha})' \Lambda (\alpha^* - \hat{\alpha}) + \\ + \sigma^2 \text{Tr}\{(\Lambda - k)(\Lambda + k)^{-1}\}. \quad (4.53)$$

Отже,

$$\hat{L}_2^2 = E(\alpha^* - \hat{\alpha})' \Lambda (\alpha^* - \hat{\alpha}) + \hat{\sigma}^2 \text{Tr}\{(\Lambda - k)(\Lambda + k)^{-1}\} \quad (4.54)$$

є незміщеною статистичною оцінкою критерію $E(L_2^2)$.

Припустимо, що $v_j = \lambda_j / (\lambda_i + k_i)$ за умови $\alpha_j^* = \hat{\alpha}_j v_j$, а компо-
нента M_i виразу (4.52) може бути записана як

$$M_j = \hat{\alpha}_i^2 (v_j - 1)^2 + \left(\frac{\hat{\sigma}^2}{\lambda_i} \right) (2v_j - 1),$$

де $\hat{L}_1^2 = \sum_{j=1}^p M_j$ та $\hat{L}_2^2 = \sum_{j=1}^p \lambda_j M_j$.

Шляхом подальших алгебраїчних перетворень M_j можна по-
дати у вигляді

$$M_j = \hat{\alpha}_j^2 \{(v_j - 1 + e_{j(0)})^2 - e_{j(0)}^2\}, \quad (4.55)$$

де $e_{j(0)}$ задається виразом (4.50). Тепер множник у виразі (4.55)
 $\hat{\alpha}_j^2 e_{j(0)}$ є постійним у процесі мінімізації, і його можна не брати
до уваги. Задача зводиться до мінімізації виразу

$$\tilde{M}_j = \hat{\alpha}_j^2 \{(v_j - 1 + e_{j(0)}^2)\}, \quad (4.56)$$

для $j = 1, 2, \dots, p$, де v_j знаходиться у межах $0 \leq v_j \leq 1$, причому
 $v_j = 0$ відповідає нескінченному k_j .

Легко знайти розв'язок цієї задачі:

$$v_j = \begin{cases} 1 - e_{j(0)}, & e_{j(0)} \leq 1, \\ 0, & e_{j(0)} \geq 1, \end{cases}$$

що відповідає такому:

$$\alpha_j^* = \begin{cases} \hat{\alpha}_j(1 - e_{i(0)}), & e_{j(0)} \leq 1, \\ 0, & e_{j(0)} \geq 1. \end{cases} \quad (4.57)$$

Тепер, оскільки всі λ_j додатні, величина

$$\sum_{j=1}^p \lambda \widetilde{M}_j \quad (4.58)$$

буде досягати свого мінімуму за значення α_j^* , що задається виразом (4.57).

Таким чином, оскільки $\hat{L}_2^2$ та величина (4.58) відрізняються лише на константу, то обидві величини $\hat{L}_1^2$ та $\hat{L}_2^2$, тобто (4.52) та (4.54), досягають глобального мінімуму за одних й тих самих значень α_j^* . Ця властивість, ніби то, узгоджується з тим фактом, що обидва критерії $E(\hat{L}_1^2)$ та $E(\hat{L}_2^2)$ мінімізуються значенням (4.49).

Розглянемо алгоритм узагальненої гребеневої статистичної оцінки обмеженнями на оцінки параметрів лінійної моделі. Для отримання гребневих статистичних оцінок використовуємо метод квадратичного програмування.

Для статистичної оцінки $\hat{L}_1^2$ виразимо цю задачу з обмеженнями таким чином: необхідно мінімізувати $\hat{L}_1^2$ за обмежень

$$cb^* \geq k \text{ і } 0 \leq v \leq 1, \quad (4.59)$$

де b — p -вектор з елементами b_i ; c — діагональна $(p \times p)$ -матриця з елементами c_i , такими що

$$c_i = \begin{cases} 1, & \text{якщо } b_i^* \geq b_i, \\ -1, & \text{якщо } b_i^* \leq b_i, \\ 0, & \text{якщо } b_i^* \text{ не обмежено,} \end{cases} \quad (4.60)$$

v — p -вектор з компонентами v_i , що задаються співвідношеннями

$$v_j = \frac{\lambda_j}{\lambda_j + k_j}; \quad I — p\text{-вектор з одиниць.}$$

Припустимо, що на практиці часто матиме місце випадок $b = 0$, який відповідає наявності апріорної інформації про знаки, а не про величини параметрів b_i .

Обмеження $cb^* \geq b$ еквівалентне

$$c\alpha^* \geq b. \quad (4.61)$$

Якщо тепер позначимо через $\hat{A}$ діагональну $(p \times p)$ -матрицю з діагональними елементами $\hat{\alpha}_i$, то вираз (4.61) можна переписати так

$$c\hat{A}v \geq b, \quad (4.62)$$

оскільки $\alpha_i^* = \hat{\alpha}_i v_i$. Виконаємо ще одне додаткове спрощення позначень, припускаючи, що t — деякий p -вектор з компонентами t_i , такий що

$$t_i = I - e_{i(0)}. \quad (4.63)$$

Тоді, як було показано вище, мінімізація критерію $\hat{L}_1^2$ еквівалентна мінімізації виразу

$$\sum_{i=1}^p \tilde{M}_i = [\hat{A}(v-t)][\hat{A}(v-t)], \quad (4.64)$$

де M_i подано виразом (4.56).

Розглянемо деякі підходи до вибору параметра k . Подамо матрицю других моментів помилок у вигляді

$$M(B) = E(\hat{\beta} - \beta)'(\hat{\beta} - \beta). \quad (4.68)$$

Доведено, що завжди можна знайти $k > 0$, для якого $M(\hat{\beta}) - M(\hat{\beta}_k)$ є додатно визначеною матрицею. Для цього достатньо, щоб k задовольняло умові

$$K < \frac{2\sigma^2}{\beta'\beta}. \quad (4.69)$$

Якщо $\text{rank}(X) < p$ і вектор $c = Q\alpha$, то умова (4.64) гарантує для допускаючої оцінку функції $\hat{\beta}_k$ меншу у порівнянні з $\hat{\beta}$ середньоквадратичну помилку, тобто кращі прогнозовані властивості.

Якщо ж обирати k , виходячи з умови мінімізації гребеневого сліду, то його значення повинно задовольняти обмеженню $k < \rho\sigma^2/(\beta'\beta)$.

На жаль, у методі гребеневої оцінки істинні значення елементів матриці K залежать від невідомих параметрів α , σ^2 . Однак оцінку k_j і, отже, оцінку всієї матриці K можна отримати методом ітерацій. У початковому наближенні можна прийняти оцінки α та k , отримані звичайним методом найменших квадратів. Потім

виконують ітераційні обчислення оцінок α та k , доки не буде досягнута збіжність результатів.

Цей метод має додаткове ускладнення: замість величин k_j , по суті, наявні деякі стохастичні змінні $\hat{k}_j$. Оскільки використання оцінок для величин k_j є коректнішим за використання фактичних значень k_j , то твердження про те, що результуючі оцінки матимуть найменшу середньоквадратичну похибку, не буде обов'язково справедливим.

У роботі [34] показано, що задовільні результати при виборі K_j для кращої оцінки $\hat{b}^*$ можна отримати за умови $K_j = K^*$ для всіх j . Далі необхідно скористатися регуляризованим слідом для вибору єдиного значення $\hat{b}$.

За деяких значень K система стабілізується і набуває характерних властивостей ортогональної системи. Коефіцієнти приймають напрямки й величину відповідної вхідної матриці $X'X$.

Інший підхід полягає у тому, що вхідну інформацію приводять до канонічного вигляду за допомогою перетворення $X = X^*P$ і $\lambda = PB$. Визначаються оцінки оптимальних K_j за допомогою $\hat{K}_{\text{юпт}} = \sigma^2 / \alpha_j^{*2}$, що використовуються для отримання $\hat{b}$. Використання $\hat{K}_{\text{юпт}}$ скорочує довжину вектора регресії, що оцінюється. Тоді доцільно оцінити ще раз $\hat{k}_{\text{юпт}}$ за допомогою $\hat{\alpha}_j^*$. Таку процедуру продовжують, доки у значеннях $(\hat{\alpha}^*)'(\hat{\alpha}^*)$ та $\hat{k}_{\text{юпт}} = \hat{\sigma}^2 / \hat{\alpha}_j^{*2}$ не з'явиться сталість. Регуляризований слід може містити додаткову інформацію про методологію вибору $\hat{b}^*$.

Середньоквадратична помилка оцінки зменшується зі збільшенням ширини спектра власних значень вхідної матриці.

На цей час для кожної з трьох груп зміщених алгоритмів існує деяке розмаїття методів вибору параметра деформації, які в принципі мають різну ефективність, що залежить як від властивостей, так і від характеристик об'єкта.

4.4. Робастні методи оцінювання

У цьому підрозділі розглядаються проблеми, породжені порушенням припущення про нормальність у регресійному аналізі, а також підходи та засоби їх вирішення.

Нормальний розподіл посідає в математичній статистиці (зокрема, у регресійному аналізі) особливе місце. Ця обставина

зумовлена *центральною граничною теоремою*, за якою нормальний розподіл за певних умов слугує граничним розподілом суми більшого числа випадкових величин, якщо число спостережень прямує до нескінченності. Нормальний закон розподілу спостережень має низку властивостей, завдяки яким отримана переважна більшість результатів теоретичної математичної статистики, широко застосовуваних і в методах регресійного аналізу. За нормального закону розподілу спостережень МНК-оцінки параметрів регресійної моделі збігаються з оцінками *методу максимуму правдоподібності* (ММП) та мають певні оптимальні властивості. Загальноприйняті та широко розповсюджені критерії для перевірки гіпотез у регресійному аналізі теж засновані на припущенні про нормальність спостережень.

Власне підхід до визначення нормального закону розподілу спостережень через центральну граничну теорему показує, що нормальний розподіл — одна з теоретичних моделей, за допомогою якої описується (приблизно) поведінка певного класу випадкових явищ. Для значної частини реальних явищ та процесів існують умови, що забезпечують опис механізму спостережуваного випадкового явища моделлю нормального розподілу. Підкреслимо, що в цьому разі фактичний закон розподілу реальних випадкових величин, що спостерігаються, завжди відрізняється від точного (теоретичного) нормального розподілу. Треба також зазначити, що крім нормального, є ще й інші теоретичні розподіли, які цілком задовільно описують випадкові механізми низки реальних явищ.

Нехай $N(\mu, \sigma)$ позначає нормальний закон розподілу спостережень з математичним сподіванням μ та квадратичним відхиленням σ , з функцією розподілу

$$\Phi(y, \mu, \sigma) = \int_{-\infty}^z \varphi(t) dt, \quad z = \frac{y - \mu}{\sigma},$$

де $\varphi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$ — щільність розподілу $N(0, 1)$.

Характерна особливість цього розподілу полягає в тому, що переважна частина спостережень зосереджена в межах інтервалу $(\mu - 3\sigma, \mu + 3\sigma)$ з центром μ . Спостереження потрапляє в цей інтервал з імовірністю 0,9973. Тому значення, що опинилися за

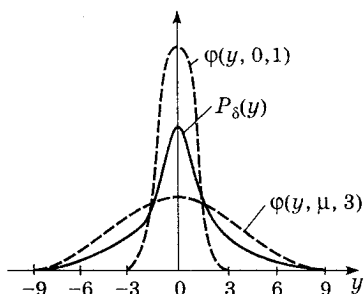


Рис. 4.19. Щільності розподілу основних спостережень, грубих помилок та «забрудненого розподілу»

межами інтервалу, зазвичай розглядають як грубі помилки («викиди»). При нормальному законі ймовірність появи в окремому спостереженні грубої помилки дорівнює 0,0027, тобто в цьому випадку поява такої помилки — надзвичайно рідкісна подія.

З рис. 4.19 видно, що за межами цього інтервалу крива щільності дуже близька до осі абсцис. Частини кривої, що знаходяться поза інтервалом, називаються *хвостами розподілу*, а оскільки основна маса щільності нормального розподілу зосереджена у 3-сигмовому інтервалі, кажуть, що у цього розподілу короткі хвости.

В значній частині реальних спостережень розподіли мають довші хвости, ніж за нормального розподілу, через що грубі помилки є не такими рідкісними подіями. Модель одного з таких законів розподілу спостережень запропонував Тьюкі. У цій моделі розподіл спостережень подається як суміш двох нормальних розподілів із щільністю

$$P_\delta(y) = (1 - \delta)\varphi(y, \mu, \sigma) + \delta\varphi(y, \mu, 3\sigma),$$

де $\varphi(y, \mu, \sigma)$ — щільність нормального розподілу основних спостережень, $\varphi(y, \mu, 3\sigma)$ — нормальний закон розподілу грубих помилок, але з потроєною квадратичною помилкою; δ — ймовірність появи грубих помилок (їхня інтенсивність). У спеціальній літературі модель $P_\delta(y)$ інтерпретують як «забруднення» нормального розподілу $N(\mu, \sigma)$ грубими помилками з розподілом $N(\mu, 3\sigma)$, а величину δ розглядають як ступінь «забруднення» нормальної сукупності грубими помилками. Доки інтенсивність грубих помилок мала — $\delta < 0,05$, «забруднений» розподіл приблизно нормальний; коли $\delta > 0,1$ — в розподілі з'являються довгі хвости і воно починає істотно відрізнятися від нормального. На рис. 4.19 пунктирною лінією позначені площини розподілу основних спостережень $\varphi(y, \mu, \sigma)$ та грубих помилок $\varphi(y, \mu, 3\sigma)$, а неперервною лінією — щільність «забрудненого» розподілу $P_\delta(y)$ при інтенсивності $\delta = 0,5$ та параметрах $\mu = 0$, та $\sigma = 1$.

Видно, що в результаті «забруднення» щільність розподілу спостережень $P_\delta(y)$ опиняється нижче та зі значно більш довгими хвостами, ніж нормальне.

У більш загальному випадку для грубих помилок у моделі $P_\delta(y)$ можна взяти довільний розподіл, що відрізняється від нормального.

При роботі з «забрудненими» розподілами слід мати на увазі, що виявлення грубих помилок та їх виключення з моделі у множинній регресії стає нелегкою, а іноді й неможливою справою. Це пов'язано із складністю розрізнення ефектів грубої помилки та неадекватності власне моделі. Грубі помилки фізично виникають у результаті відмови приладу, через випадковість, що призвела до неконтрольованої зміни умов експерименту, тощо.

Числовими характеристиками, які показують ступінь відмінності розподілу цієї випадкової величини у від нормального закону, є коефіцієнт асиметрії γ_1 та ексцес γ_2 , що визначаються таким чином: $\gamma_1 = \sigma^{-3}E[(y - \mu)^3]$, $\gamma_2 = \sigma^{-4}E[(y - \mu)^4] - 3$. Для нормального закону $\gamma_1 = \gamma_2 = 0$. За умови $\gamma_1 \neq 0$ розподіл асиметричний, до того ж, коли $\gamma_1 > 0$, кажуть, що він «скошений» праворуч, а коли $\gamma_1 < 0$ — ліворуч. Додатне значення ексцесу показує, що розподіл має більш довгі хвости, ніж нормальний. Експоненціальний розподіл, наприклад, має ексцес $\gamma_2 = 6$, розподіл екстремальних значень дорівнює $\gamma_2 = -2,4$, а за $\gamma_2 < 0$ розподіл характеризують хвости, більш короткі, ніж у нормального. Такий рівномірний розподіл с $\gamma_2 = -1,2$. Хоча асиметрія та ексцес не визначають розподіл однозначно, це зазвичай найважливіші показники, за допомогою яких можна кількісно оцінити ефект відхилення від нормальності розподілу спостережень.

Обговоримо коротко й проблеми, пов'язані з перевіркою припущення про нормальність розподілу спостережень. Перевіряти гіпотези про нормальність розподілу можна за стандартними критеріями згоди Колмогорова — Смирнова або χ^2 , але коректне й ефективне використання цих критеріїв можливе тільки при великій кількості спостережень — понад 100. Крім того, ці критерії вимагають апріорного знання параметрів μ та σ теоретичного розподілу (для критерію Колмогорова — Смирнова) або їх оцінок (для χ^2). Таких недоліків позбавлений спеціальний критерій для перевірки нормальності, запропонований Шапіро та Уїлксом і відомий як W -критерій. Дослідження показали, що цей критерій за ефективністю перевищує загальні критерії згоди.

Він, наприклад, дозволяє з великою вірогідністю виявляти помірні відхилення від нормальності вже при 20 спостереженнях, а значні відхилення — навіть при $N = 10$.

Із зазначеного вище випливає, що для виявлення помірних відхилень від нормальності треба мати у наявності щонайменше 20 отриманих за однакових умов спостережень випадкової величини, що розглядається. У термінах регресійного аналізу це означає проведення відповідного числа повторних дослідів у цій точці факторного простору. У переважній кількості випадків це неможливо через обмеження у часі, нестаціонарність експериментальних умов, з економічних міркувань та ін. Як правило, не вдається забезпечити навіть грубої перевірки нормальності — це одна з головних причин того, що припущення про нормальність зазвичай апріорно допускається без жодної перевірки.

Інша важлива причина апріорного, а часто й беззастережного, прийняття в регресійному аналізі припущення про нормальність — наявність достатньо розробленого математичного апарату оцінювання та перевірки гіпотез для цього випадку. Якщо є такий апарат, природно виникає питання: чи не можна поширити його додаток на випадок, коли умови нормальності не виконуються? Вже підкреслювалось, що нормальний розподіл — лише теоретична модель (апроксимація) деяких реальних явищ і на практиці спостереження розподіляються в кращому випадку лише приблизно нормально.

Зазначені обставини породжують проблему чутливості методів оцінювання та перевірки гіпотез, розроблених у припущенні про нормальність, до порушень цього припущення. Статистичні методи, нечутливі до порушень припущень, на яких вони засновані, називаються *сталими* (або *робастними*). Поняття «сталі методи» було введено Боксом. Сталість має важливе значення при практичному застосуванні методів, оскільки введені припущення — завжди певною мірою ідеалізують реальні умови, а обмежений обсяг вибірок ускладнює або робить зовсім неможливо його перевірку.

Методи оцінювання називають сталими до порушення припущення про нормальність спостережень, якщо основні властивості оцінок — незміщеність, обґрунтованість та ефективність — виявляються нечутливими до порушення цього припущення. Сталість за критерієм ефективності щільно пов'язана зі сталістю до

грубих помилок, що виражається у нечутливості значень оцінок до грубих помилок у спостереженнях. Оцінки, які мають високу ефективність як для нормального розподілу, так і для розподілу з довгими хвостами (що є ознакою появи грубих помилок), є також сталими до грубих помилок. Однак зворотне твердження не завжди правильне: сталість до грубих помилок не обов'язково призводить до сталості за критерієм ефективності.

З проведених до сьогодні досліджень сталості відомо, що МНК-оцінки мають низьку ефективність при розподілах з довгими хвостами (тобто таких, що відрізняються від нормального), що супроводжується й чутливістю до грубих помилок. Це спричинює необхідність розробки сталих методів оцінювання, які мають високу ефективність у широкому класі розподілів та нечутливих до грубих помилок.

Критерії перевірки гіпотез, що засновані на нормальному законі, називають *сталими до порушення нормальності*, якщо ймовірності прийняття неправильних рішень (помилки I та II роду) не чутливі до такого порушення. Нагадаємо, що помилка I роду має місце у випадку, коли гіпотеза, яку перевіряють, неправильно відкидається, а насправді вона вірна. Помилка II роду означає, що гіпотеза, яку перевіряють, приймається, хоча вона насправді невірна.

Сталість критеріїв до порушення нормальності — предмет численних досліджень. Більшість отриманих результатів стосуються сталості стандартних критеріїв для перевірки гіпотез щодо математичних сподівань та дисперсій однієї або декількох сукупностей до порушення нормальності розподілу спостережень. Ці результати показують, що критерії для математичних сподівань (t - та F -критерії) порівняно сталі до порушення нормальності, тоді як критерії для дисперсій (Бартлетта, Фішера) вельми чутливі до такого порушення. Також виявилось, що чутливість критеріїв у регресійному аналізі залежить ще й від плану експерименту.

Порушення нормальності у регресійному аналізі породжує дві групи проблем. Перш за все, це проблеми, пов'язані з дослідженням сталості стандартних методів регресії, заснованих на припущенні про нормальність, до порушення нормальності. Мета — виявлення умов, за яких стандартні методи є сталими й можуть використовуватися коректно та ефективно при інших розподілах. Друга група проблем виникає, якщо відповідні

методи регресійного аналізу не стали до порушення нормальності. Для вирішення проблем такого роду існують два підходи:

☑ модифікація існуючих методів з урахуванням відхилення від нормальності;

☑ розробка вільних від розподілу методів, які засновані не на припущенні про нормальність і мають високу ефективність для широкого класу розподілів.

У сучасній теорії сталого оцінювання розробляються три основних класи оцінок параметрів регресійних моделей, які стали до порушення припущення про нормальність спостережень.

Один з них — так звані *М-оцінки*. Їх позначають b_M і отримують, розв'язуючи задачу оптимізації [35, 36]:

$$\sum_{u=1}^n \rho(y_u - \sum_{i=1}^k \beta_i f_{iu}) = \sum_{u=1}^n \rho(e_u) = \min_{\beta_1, \dots, \beta_k}, \quad (4.69)$$

де $e_u = y_u - \sum \beta_i f_i$ — залишки, а $\rho(z)$ — придатна обрана функція. Властивості оцінок b_M , що отримані з (4.69), залежать від виду функції $\rho(z)$. Для забезпечення незміщеності, обґрунтованості та високої ефективності *М-оцінок* ця функція повинна задовольняти певним вимогам. Найзагальніша вимога — функція $\rho(z)$ має зростати повільніше за функцію z^2 , на якій базується МНК. При такому виборі $\rho(z)$ функція (4.69) та отримані *М-оцінки* виявляються менш чутливими до грубих помилок, ніж оцінки МНК.

Ще один клас сталих оцінок параметрів множинної регресії будується на основі впорядкованих співвідношень між залишками, e_u та відповідними спостереженнями y . Цей клас відомий як *L-оцінки*; будемо позначати їх b_L . Головне у побудуванні таких оцінок — отримання варіаційного ряду залишків, що являє собою впорядкований за зростанням ряд величин $e_u = y_u - \sum \beta_i f_{iu}$:

$$e(1) \leq e(2) \leq \dots \leq e(N), \quad (4.70)$$

де $e(i)$ — i -та порядкова статистика. Частіше за все *L-оцінки* параметрів регресійної моделі знаходять шляхом розв'язання мінімізаційної задачі:

$$\sum_{i=1}^N \lambda(i) e_{(i)}^2 = \min_{\beta_1, \dots, \beta_k}, \quad (4.71)$$

де $\lambda(i)$ — обрана придатна вагова функція; її вигляд визначає властивості оцінок b_L . Очевидно, що задача (4.71) є модифікацією МНК, яка полягає у введенні вагових коефіцієнтів $\lambda(i)$. Сталість цих оцінок до грубих помилок та їхня більш висока ефективність порівняно з МНК-оцінками досягаються відповідним вибором функції $\lambda(i)$. Зазвичай функція $\lambda(i)$ обирається так, що певна частина залишків, розташованих на краях варіаційного ряду (4.70), може розглядатися як результати грубих помилок та отримує менші значення порівняно з середніми членами ряду.

Третій клас сталих оцінок параметрів множинної регресії — так звані R -оцінки. Такі методи засновані на використанні рангів залишків $e_u = y_u - \sum \beta_i f_{iu}$. Ранг R_u величини e_u представляє її порядковий номер у варіаційному ряді (4.70). Він визначається рівністю $e_{(R_u)} = e_u$. Зазвичай R -оцінки, які позначаються b_R , отримують, розв'язуючи мінімізаційну задачу:

$$\sum_{u=1}^N a(R_u) e_u = \sum_{u=1}^N a(R_u) (y_u - \sum \beta_i f_{iu}) e_u = \min_{\beta_1, \dots, \beta_k}, \quad (4.72)$$

де $a(R_u)$ — обрана придатна вагова функція, що визначає властивості оцінок. Вагова функція обирається так, що залишки на краях варіаційного ряду набувають менших значень, ніж у середині, а це приводить до сталості оцінок щодо грубих помилок. Щоб гарантувати властивості незміщеності, обґрунтованості та ефективності R -оцінок для широкого класу розподілів, до функції $a(R_u)$ висувають певні вимоги.

Зазначимо, що введені вище сталі оцінки параметрів лінійної регресійної моделі самі виявилися нелінійними функціями спостережень. Тому ефективні оцінки можна знайти тільки у класі нелінійних оцінок.

4.5. Приклад побудови регресійної моделі та її інтерпретація за даними таблиці значень показників виробничо-господарської діяльності підприємства з використанням ППП «STATISTICA»

Пакет прикладних програм «STATISTICA» являє собою графічну статистичну систему, призначену для розв'язання задач математичної та прикладної статистики. Він може бути успішно

використаний для ідентифікації статичних характеристик технологічних процесів засобами регресійного та кореляційного аналізів, гребеневими та робасними засобами.

ППП дозволяє заздалегідь обирати структуру моделі на підставі статистичного аналізу вхідних даних. Припустимо, що об'єктом дослідження є технологічний процес, описуваний статичною залежністю $Y = F(X)$, де X — вектор входу та Y — вектор виходу, F — оператор моделі, який характеризує зв'язок між входом та виходом об'єкта.

При дослідженні процесу спостерігачеві доступні значення вхідних $X = \{x_j\}$ та вихідних $Y = \{y_i\}$ змінних, які записуються у вигляді таблиці вхідних даних; $j = 1, p$ — кількість вхідних змінних; $i = 1, n$ — кількість вимірювань.

Унаслідок наявності помилок вимірювання, які є в експерименті, та впливу різноманітних неврахованих факторів, Y є випадковим процесом з певними статистичними характеристиками.

Кожне значення y_t в t -й момент часу є реакцією на одночасний вплив x_{tj} , $j = 1 \dots p$, у той же самий момент часу. Якщо об'єкт характеризується інерцією (затримкою), де y_t є реакцією на $x_{tj} - \Delta t$, де $\Delta t = t_x - t_y$ (t_x — час початку дії незалежної змінної X , t_y — час початку реакції Y на X).

Будь-яка модель відображає лише деякі особливості об'єкта та ніколи не буває його точною копією. Отже, не можна говорити про «істинну» модель в повному розумінні слова. В окремих випадках під «істинним» значенням розуміють умовне математичне сподівання Y при заданих значеннях X , іншими словами, $M[Y] = M\{y/x\} = XB + M[E]$, де $M[\dots]$ — символ математичного сподівання; E — вектор помилок; $B = \{b_l\}$, $l = 1, m$ — вектор параметрів коефіцієнтів моделі.

Структура виразу може бути описана різноманітними функціями, але у випадку, що розглядається, обов'язково лінійними відносно коефіцієнтів, b_l , $l = 1, m$.

Створимо файл даних, який характеризуватиме виробничо-господарську діяльність машинобудівного підприємства (за прикладом із розділу 3). Вихідними даними вважатимемо показники, наведені у табл. 4.7:

- Y1 — продуктивність праці;
- Y2 — індекс зниження собівартості продукції;
- Y3 — рентабельність;
- X4 — трудомісткість одиниці продукції;

Таблиця 4.7

Вихідні дані

№ підприємства	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Y9	Y10	Y11	Y12	Y13	Y14	Y15	Y16	Y17
1	9,26	204,20	13,26	0,23	0,78	0,40	1,37	1,23	0,23	1,45	26006	167,69	47750	6,40			
2	9,38	209,60	10,16	0,24	0,75	0,26	1,49	1,04	0,39	1,30	23935	186,10	50391	7,80			
3	12,11	222,60	13,72	0,19	0,68	0,40	1,44	1,80	0,43	1,37	22589	220,45	43149	9,76			
4	10,81	236,70	12,85	0,17	0,70	0,50	1,42	0,43	0,18	1,65	21220	169,30	41089	7,90			
5	9,35	62,00	10,63	0,23	0,62	0,40	1,35	0,88	0,15	1,91	7394	39,53	14257	5,35			
6	9,87	53,10	12,10	0,43	0,76	0,19	1,39	0,57	0,34	1,68	11586	40,41	22661	9,90			
7	8,17	172,10	25,83	0,31	0,73	0,25	1,16	1,72	0,38	1,94	26609	102,96	52509	4,50			
8	9,12	56,50	23,39	0,26	0,71	0,44	1,27	1,70	0,09	1,89	7801	37,02	14903	4,88			
9	5,88	52,60	14,68	0,49	0,69	0,17	1,16	0,84	0,14	1,94	11587	45,74	25387	3,46			
10	6,30	46,60	10,05	0,36	0,73	0,39	1,25	0,60	0,21	2,06	9475	40,07	16821	3,60			
11	6,22	53,20	13,99	0,37	0,68	0,33	1,13	0,82	0,42	1,96	10811	45,44	19459	3,56			
12	5,49	30,10	9,68	0,43	0,74	0,25	1,10	0,84	0,05	1,02	6371	41,08	12973	5,65			
13	6,50	146,40	10,03	0,35	0,66	0,32	1,15	0,67	0,29	1,85	26761	136,14	50907	4,28			
14	6,61	18,10	9,13	0,38	0,72	0,02	1,23	1,04	0,48	0,88	4210	42,39	6920	8,85			
15	4,32	13,60	5,37	0,42	0,68	0,06	1,39	0,66	0,41	0,62	3557	37,39	5736	8,52			
16	7,37	89,80	9,86	0,30	0,77	0,15	1,38	0,86	0,62	1,09	14148	101,78	26705	7,19			
17	7,02	62,50	12,62	0,32	0,78	0,08	1,35	0,79	0,56	1,60	9872	47,55	20068	4,82			
18	8,25	46,30	5,02	0,25	0,78	0,20	1,42	0,34	1,76	1,53	5975	32,61	11487	5,46			
19	8,15	103,50	21,18	0,31	0,81	0,20	1,37	1,60	1,31	1,40	16662	103,25	32029	6,20			
20	8,72	73,30	25,17	0,26	0,79	0,30	1,41	1,46	0,45	2,22	9166	38,95	18946	4,25			
21	6,64	76,60	19,40	0,37	0,77	0,24	1,35	1,27	0,50	1,32	15118	81,32	28025	5,38			
22	8,10	73,01	21,0	0,29	0,78	0,10	1,48	1,58	0,77	1,48	11429	67,26	20968	5,88		8,64	15,06
23	5,52	32,30	6,57	0,34	0,72	0,11	1,24	0,68	1,20	0,68	6462	59,92	11049	9,27		97,20	20,09
24	9,37	199,60	14,19	0,23	0,79	0,47	1,40	0,86	0,21	2,30	24628	107,34	45933	4,36		80,28	14,76
25	13,17	598,10	15,81	0,17	0,77	0,53	1,45	1,98	0,25	1,37	49727	512,60	99400	10,31		51,48	10,08
26	6,67	71,20	5,23	0,29	0,80	0,34	1,40	0,33	0,15	1,51	11470	53,81	20719	4,69		105,12	14,76
27	5,68	90,80	7,99	0,41	0,71	0,20	1,28	0,45	0,66	1,43	19448	80,83	36813	4,16		128,52	10,44

Закінчення табл. 4.7

№ підпри- ємства	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Y9	Y10	Y11	Y12	Y13	Y14	Y15	Y16	Y17
28	5,22	82,10	17,50	0,41	0,79	0,24	1,33	0,74	0,74	1,82	18963	59,42	33956	3,13	94,68	14,76	15,41
29	10,02	76,20	17,16	0,22	0,76	0,54	1,22	0,03	0,32	2,62	9185	36,96	17016	4,02	85,32	20,52	19,35
30	8,16	119,50	14,54	0,29	0,78	0,40	1,28	0,99			17478	91,43	34873	5,23	76,32	14,40	16,83
31	3,78	21,90	6,24	0,51	0,62	0,20	1,47	0,24			6265	17,16	11237	2,74	153,00	24,84	30,53
32	6,48	48,40	12,08	0,36	0,75	0,64	1,27	0,57			8810	27,29	17306	3,10	107,64	11,16	17,98
33	10,44	173,50	9,49	0,23	0,71	0,42	1,51	1,22			17659	184,33	39250	10,44	90,72	6,48	22,09
34	7,65	74,10	9,28	0,26	0,74	0,27	1,46	0,68			10342	58,42	19074	5,65	82,44	9,72	18,29
35	8,77	68,60	11,42	0,27	0,65	0,37	1,27	1,00			8901	59,40	18452	6,67	79,92	3,24	26,05
36	7,00	60,80	10,31	0,29	0,66	0,38	1,43	0,81			8402	49,63	17500	5,91	120,96	6,48	26,20
37	11,06	355,60	8,65	0,01	0,84	0,35	1,50	1,27			32625	391,27	7888	11,99	84,60	5,40	17,26
38	9,02	264,80	10,94	0,02	0,74	0,42	1,35	1,14			31160	258,62	58947	8,30	85,32	6,12	18,83
39	13,28	526,60	9,87	0,18	0,75	0,32	1,41	1,89			46461	75,66	94697	1,63	101,52	8,64	19,70
40	9,27	118,60	6,14	0,25	0,75	0,33	1,47	0,67			13833	123,68	29626	8,94	107,64	11,88	16,87
41	6,70	37,10	12,93	0,31	0,79	0,29	1,35	0,96			6391	37,21	11688	5,82	85,32	7,92	14,63
42	6,69	57,70	9,78	0,38	0,62	0,30	1,40	0,67			11115	53,37	21955	4,80	131,76	10,08	22,17
43	9,42	51,60	13,22	0,24	0,70	0,56	1,20	0,98			6555	32,87	12243	5,01	116,64	18,72	22,62
44	7,24	64,70	17,29	0,31	0,66	0,42	1,15	1,16			11085	45,63	20193	4,12	138,24	13,68	26,44
45	5,39	48,30	7,11	0,42	0,69	0,26	1,09	0,54			9484	48,41	20122	5,10	156,96	16,56	22,26
46	5,61	15,00	22,49	0,51	0,71	0,16	1,26	1,23			3967	13,58	7612	3,49	137,52	14,76	19,13
47	5,59	87,50	12,14	0,31	0,73	0,45	1,36	0,78			15283	63,99	27404	4,19	135,72	7,92	18,28
48	6,57	108,40	15,25	0,37	0,65	0,31	1,15	1,16			20874	104,55	39648	5,01	155,52	18,36	28,23
49	6,54	267,30	31,34	0,16	0,82	0,08	1,87	4,44			19418	222,11	43799	11,44	48,60	8,28	12,39
50	4,23	34,20	11,56	0,18	0,80	0,68	1,17	1,06			3851	25,76	6235	7,67	42,84	14,04	11,64
51	5,22	26,80	30,14	0,43	0,83	0,03	1,61	2,13			6388	29,52	11524	4,66	142,20	16,92	8,62
52	18,00	43,6	19,71	0,40	0,70	0,02	1,34	1,21			9756	41,99	17309	4,30	145,80	11,16	20,10
53	11,03	72,02	23,56	0,31	0,74	0,22	1,22	2,20			11795	78,11	22225	6,62	120,52	14,76	19,41

- X5 — питома вага працівників у складі підприємства;
- X6 — питома вага покупних виробів;
- X7 — коефіцієнт змінності обладнання;
- X8 — премії та винагороди на одного працівника;
- X9 — питома вага втрат від браку;
- X10 — фондовіддача;
- X11 — середньорічна чисельність підприємства;
- X12 — середньорічна вартість загального виробничого фонду;
- X13 — середньорічний фонд заробітної платні підприємства;
- X14 — фондоозброєність праці;
- X15 — оборотність нормованих обігових коштів;
- X16 — оборотність ненормованих обігових коштів;
- X17 — невиробничі витрати.

Статистичний аналіз експериментальних даних з використанням ППП «STATISTICA» включає в себе побудову двовимірних діаграм розсіювання, обчислення одновимірних статистик, побудову гістограм, графіків функцій розподілу, побудову довірчих інтервалів для математичного сподівання та дисперсії, обчислення кореляційної функції. Перевірку гіпотез щодо стаціонарності та випадковості вибірки даних, нажаль, у ППП «STATISTICA» не передбачено.

Вид моделі обирається, виходячи з вигляду діаграм розсіювання вхідних та вихідних параметрів. Нелінійні моделі приводяться до лінійних шляхом логарифмічного перетворення або заміни змінної.

Розглянемо приклад, візьмемо набір даних, які описують процес виробництва. Залежність між параметрами подамо у вигляді:

$$Y = F(X_1, X_2, \dots, B, E),$$

де Y — продуктивність праці, яка є випадковою величиною; значення X можуть розглядатися як випадкові або як фіксовані. Якщо X — випадкова величина, очікуване значення вихідної змінної Y розглядається як умовне математичне сподівання $E\{Y/X\}$. У цьому разі необхідно перевірити гіпотезу про випадковість вхідних та вихідних змінних.

Теоретичну частину перевірки гіпотези про випадковість було наведено у розділі 4.4.

Таблиця 4.8

Відсоткові точки розподілу серій

$n = N/2$						
	0,99	0,975	0,95	0,05	0,025	0,01
5	2	2	3	8	9	9
6	2	3	3	10	10	1»
7	3	3	4	11	12	12
8	4	4	5	12	13	13
9	4	5	6	13	14	15
10	5	6	6	15	15	16
11	6	7	7	16	16	17
12	7	7	8	17	18	18
13	7	8	9	18	19	20
14	8	9	10	19	20	21
15	9	10	11	20	21	22
16	10	11	11	22	22	23
18	11	12	13	24	25	26
20	13	14	15	26	27	28
25	17	18	19	32	33	34
30	21	22	24	37	39	40
35	25	27	28	43	44	46
40	30	31	33	48	50	51
45	34	36	37	54	55	57
50	38	40	42	59	61	63
55	43	45	46	65	66	68
60	47	49	51	70	72	74
65	52	54	56	75	77	79
70	56	58	60	81	83	85
75	61	63	65	86	88	90
80	65	68	70	91	93	96
85	70	72	74	97	99	101
90	74	77	79	102	104	107
95	79	82	84	107	109	112
100	84	86	88	113	115	117

Перевіримо гіпотезу про випадковість параметрів Y_1 , X_4 , X_8 , X_{12} . Дані наведено у табл. 4.9.

Таблиця 4.9

	Y_1	X_4	X_8	X_{12}	$Y_1 - M$	$X_4 - M$	$X_8 - M$	$X_{12} - M$
1	9,26	0,23	1,23	167,69	1,28962	-0,07245	0,158113	75,81642
2	9,38	0,24	1,04	186,1	1,40962	-0,06245	-0,03189	94,22642
3	12,11	0,19	1,8	220,45	4,13962	-0,11245	0,728113	128,5764
4	10,81	0,17	0,43	169,3	2,83962	-0,13245	-0,64189	77,42642
5	9,35	0,23	0,88	39,53	1,37962	-0,07245	-0,19189	-52,8436
6	9,87	0,43	0,57	40,41	1,89962	0,127547	-0,50189	-51,4636
7	8,17	0,31	1,72	102,96	0,19962	0,007547	0,648113	11,08642
8	9,12	0,26	1,7	37,02	1,14962	-0,04245	0,628113	-54,8536
9	5,88	0,49	0,84	45,74	-2,09038	0,187547	-0,23189	-46,1336
10	6,3	0,36	0,6	40,07	-1,67038	0,057547	-0,47189	-51,8036
11	6,22	0,37	0,82	45,44	-1,75038	0,067547	-0,25189	-46,4336
12	5,49	0,43	0,84	41,08	-2,48038	0,127547	-0,23189	-50,7936
13	6,5	0,35	0,67	136,14	-1,47038	0,047547	-0,40189	44,26642
14	6,61	0,38	1,04	42,39	-1,36038	0,077547	-0,03189	-49,4836
15	4,32	0,42	0,66	37,39	-3,65038	0,117547	-0,41189	-54,4836
16	7,37	0,3	0,86	101,78	-0,60038	-0,00245	-0,21189	9,906415
17	7,02	0,32	0,79	47,55	-0,95038	0,017547	-0,28189	-44,3236
18	8,25	0,25	0,34	32,61	0,27962	-0,05245	-0,73189	-59,2636
19	8,15	0,31	1,6	103,25	0,17962	0,007547	0,528113	11,37642
20	8,72	0,26	1,46	38,95	0,74962	-0,04245	0,388113	-52,9236
21	6,64	0,37	1,27	81,32	-1,33038	0,067547	0,198113	-10,5536
22	8,1	0,29	1,58	67,26	0,12962	-0,01245	0,508113	-24,6136
23	5,52	0,34	0,68	59,92	-2,45038	0,037547	-0,39189	-31,9536
24	9,37	0,23	0,86	107,34	1,39962	-0,07245	-0,21189	15,46642
25	13,17	0,17	1,98	512,6	5,19962	-0,13245	0,908113	420,7264
26	6,67	0,29	0,33	53,81	-1,30038	-0,01245	-0,74189	-38,0636
27	5,68	0,41	0,45	80,83	-2,29038	0,107547	-0,62189	-11,0436
28	5,22	0,41	0,74	59,42	-2,75038	0,107547	-0,33189	-32,4536
29	10,02	0,22	0,03	36,96	2,04962	-0,08245	-1,04189	-54,9136
30	8,16	0,29	0,99	91,43	0,18962	-0,01245	-0,08189	-0,44358
31	3,78	0,51	0,24	17,16	-4,19038	0,207547	-0,83189	-74,7136
32	6,48	0,36	0,57	27,29	-1,49038	0,057547	-0,50189	-64,5836
33	10,44	0,23	1,22	184,33	2,46962	-0,07245	0,148113	92,45642

Закінчення табл. 4.9

	Y1	X4	X8	X12	Y1 - M	X4 - M	X8 - M	X12 - M
34	7,65	0,26	0,68	58,42	-0,32038	-0,04245	-0,39189	-33,4536
35	8,77	0,27	1	59,4	0,79962	-0,03245	-0,07189	-32,4736
36	7	0,29	0,81	49,63	-0,97038	-0,01245	-0,26189	-42,2436
37	11,06	0,01	1,27	391,27	3,08962	-0,29245	0,198113	299,3964
38	9,02	0,02	1,14	258,62	1,04962	-0,28245	0,068113	166,7464
39	13,28	0,18	1,89	75,66	5,30962	-0,12245	0,818113	-16,2136
40	9,27	0,25	0,67	123,68	1,29962	-0,05245	-0,40189	31,80642
41	6,7	0,31	0,96	37,21	-1,27038	0,007547	-0,11189	-54,6636
42	6,69	0,38	0,67	53,37	-1,28038	0,077547	-0,40189	-38,5036
43	9,42	0,24	0,98	32,87	1,44962	-0,06245	-0,09189	-59,0036
44	7,24	0,31	1,16	45,63	-0,73038	0,007547	0,088113	-46,2436
45	5,39	0,42	0,54	48,41	-2,58038	0,117547	-0,53189	-43,4636
46	5,61	0,51	1,23	13,58	-2,36038	0,207547	0,158113	-78,2936
47	5,59	0,31	0,78	63,99	-2,38038	0,007547	-0,29189	-27,8836
48	6,57	0,37	1,16	104,55	-1,40038	0,067547	0,088113	12,67642
49	6,54	0,16	4,44	222,11	-1,43038	-0,14245	3,368113	130,2364
50	4,23	0,18	1,06	25,76	-3,74038	-0,12245	-0,01189	-66,1136
51	5,22	0,43	2,13	29,52	-2,75038	0,127547	1,058113	-62,3536
52	18	0,4	1,21	41,99	10,02962	0,097547	0,138113	-49,8836
53	11,03	0,31	2,2	78,11	3,05962	0,007547	1,128113	-13,7636

Розрахункові значення медіан, середніх, стандартного відхилення подані в табл. 4.10.

Таблиця 4.10

	Mean	Median	Minimum	Std.Dev.
Y1	7,97038	7,37000	3,78000	2,61022
X4	0,30245	0,31000	0,01000	0,10548
X8	1,07189	0,96000	0,03000	0,67598
X12	91,87358	58,42000	13,58000	92,36779

Для отримання числа серій були використані середні значення. Так, для Y1 за $n = 53/2 \approx 26$ кількість серій дорівнює 19. Використовуючи табл. 4.8, отримуємо: $r_{53/2, 1-0,05/2} < r_{53} \leq r_{53/2, 0,05/2}$; $18 < 19 \leq 33$. Оскільки розрахункове значення 19 опинилося в інтервалі між табличними значеннями, то можна вважати, що дані стаціонарні.

Для X_4 , X_8 , X_{12} розрахункові значення r_{53} відповідно 22, 21, 20. Отже, для вхідних змінних X_4 , X_8 , X_{12} гіпотеза про випадковість дотримується.

Для перевірки гіпотези про випадковий характер вихідних даних використовується метод послідовних різниць. Елементи досліджуваної вибірки $\{y_i\}$ розташовуються за порядком отримання спостережень, для них обчислюються вибіркові середнє та дисперсія:

$$\bar{y} = \frac{1}{N} \sum_{i=1}^N y_i; \quad \hat{S}^2 = \frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{y})^2.$$

Визначають різниці Δy між сусідніми спостереженнями $\Delta y_1 = y_2 - y_1$, $\Delta y_2 = y_3 - y_2$, ..., $\Delta y_{N-1} = y_N - y_{N-1}$ та математичне сподівання квадрата різниці

$$M[\Delta y] = \frac{1}{N-1} \sum \Delta y_i^2 = 2\hat{c}^2,$$

де $\hat{c}^2$ — оцінка генеральної дисперсії.

Фактична величина критерію випадковості вибірки

$$t_\Phi = \frac{\hat{c}^2}{S^2}.$$

Відомо, що при $N > 20$ та випадковій вибірці t_Φ розподілена нормально. Виходячи з 1 %-го рівня значущості, маємо $F(t_T) = 0,99$ та $t_T = 2,58$. Тоді теоретичне значення критерію

$$t_T = 1 - \frac{2,58}{\sqrt{N+1}}.$$

За умови $t_\Phi < t_T$ для конкретного N гіпотеза випадковості відкидається.

Перевірка гіпотези про наявність викидів та визначення структури статистичного зв'язку

Двовимірні діаграми розсіювання зображують пари значень (x_1, y_1) , (x_2, y_2) , ..., (x_n, y_n) та призначені для оцінки виду залежності між X та Y (лінійна, нелінійна), а також візуального визначення викидів та грубих помилок. Формалізовані засоби оцінки нелінійності та лінійності викладені у [4]. Алгоритми оцінки викидів та грубих помилок представлені у [4, 8, 10]. Дослідник

за двовимірними діаграмами розсіювання візуально визначає вид залежності, наявність викидів і в подальшому оцінює їх за формалізованими процедурами. Подана на діаграмі розсіювання залежність Y від X може бути:

- ☒ лінійною;
- ☒ поліноміальною;
- ☒ експоненціальною;
- ☒ логарифмічною або ін.

ППП «STATISTICA» дозволяє обрану користувачем залежність подати графічно у вигляді графіків таких типів:

- ☒ Regular;
- ☒ Multiple;
- ☒ Double-Y та ін.

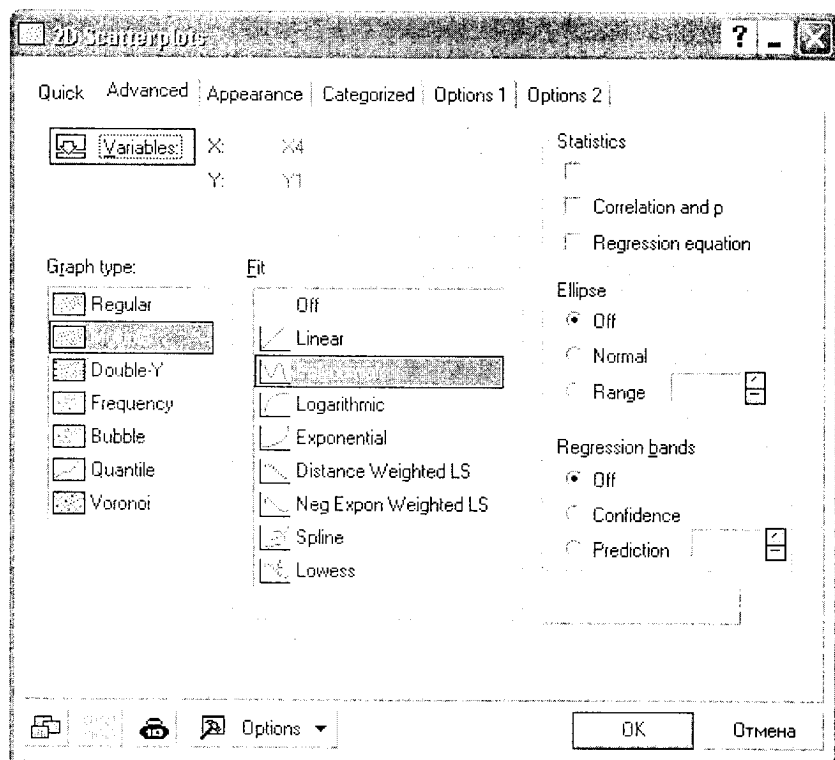
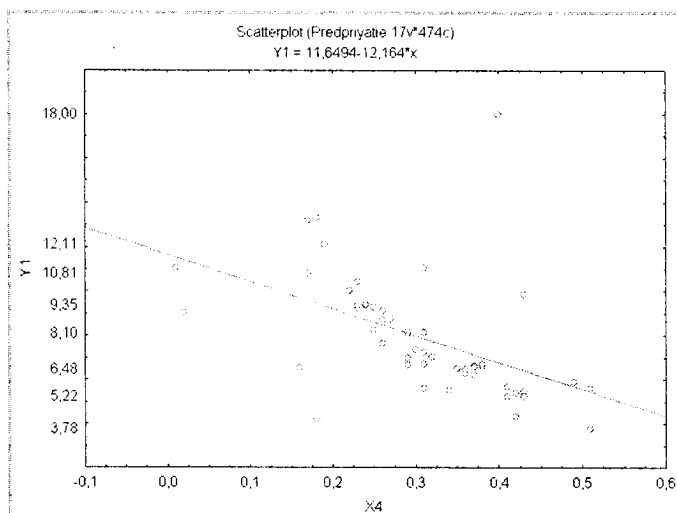
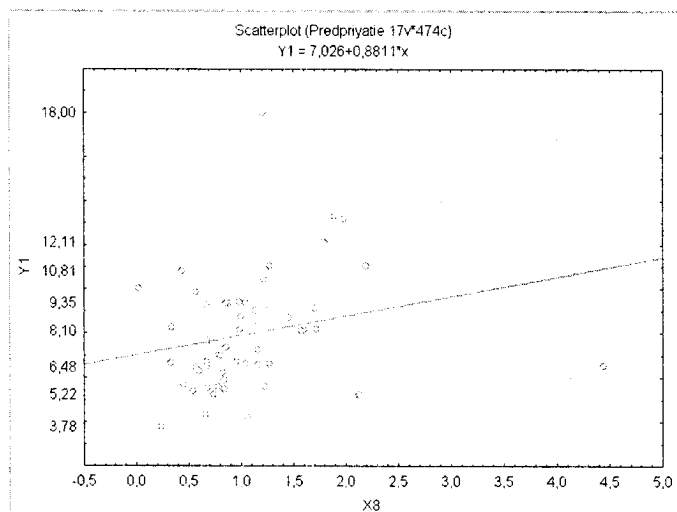


Рис. 4.20. Вікно вибору типу графіка

Розглянемо залежність Y_1 від X_4 , X_8 , X_{12} за діаграмами розсіювання (див. рис. 4.21, 4.22). Виходячи з аналізу графіків можна прийняти гіпотезу про лінійну структуру моделі.



a)



б)

Рис. 4.21. Діаграми розсіювання змінних X_4 (a) та X_8 (б)

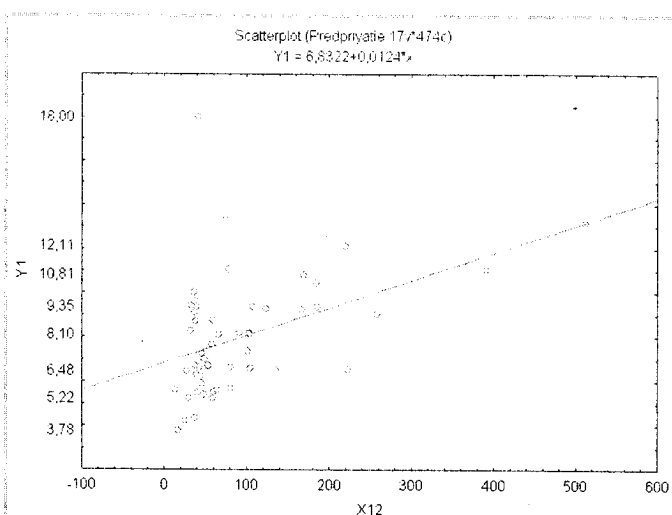


Рис. 4.22. Діаграма розсіювання для змінної X12

Аналіз графіків показує, що для змінної X4 спостерігається викид з координатами (0,4, 18), для змінної X8 — (1,3, 18) та (4,4, 6,48), для змінної X12 — (40, 18).

Користувачеві необхідно провести ретельний аналіз методу отримання вихідних даних. Слід зауважити, що апроксимація вихідних даних різними функціями «лежить на плечах користувача» та має бути зроблена для всіх вхідних та вихідних змінних. Ця інформація дозволяє коректно підійти до вибору структури багатовимірної моделі.

Перевіримо гіпотезу про стаціонарність залежної та незалежної змінних. Перевірку будемо проводити у широкому розумінні, тобто перевіримо незалежність перших двох моментів математичного сподівання та автокореляційних функцій від початку відліку. Для аналізу візьмемо змінні Y1, X4, X8, X12. Щоб перевірити їх на стаціонарність, необхідно розбити дані на декілька частин та перевірити на кожному інтервалі постійність перших двох моментів (математичного сподівання та дисперсії). Якщо ряд стаціонарний, вони не повинні залежати від початку відліку.

У табл. 4.11 наведені вибірки, що складаються з 53 значень змінних (Y1, X4, X8, X12), 44 значень змінних — від 10 до 53 (Y1(10), X4(10), X8(10), x12(10)) та 29 значень змінних — від 25 до 53 (Y1(25), X4(25), X8(25), x12(25)).

Таблица 4.11

	Y1	X4	X8	X12	Y1(10)	X4(10)	X8(10)	x12(10)	Y1(25)	X4(25)	X8(25)	x12(25)
1	9,26	0,23	1,23	167,69	6,3	0,36	0,6	40,07	13,17	0,17	1,98	512,6
2	9,38	0,24	1,04	186,1	6,22	0,37	0,82	45,44	6,67	0,29	0,33	53,81
3	12,11	0,19	1,8	220,45	5,49	0,43	0,84	41,08	5,68	0,41	0,45	80,83
4	10,81	0,17	0,43	169,3	6,5	0,35	0,67	136,14	5,22	0,41	0,74	59,42
5	9,35	0,23	0,88	39,53	6,61	0,38	1,04	42,39	10,02	0,22	0,03	36,96
6	9,87	0,43	0,57	40,41	4,32	0,42	0,66	37,39	8,16	0,29	0,99	91,43
7	8,17	0,31	1,72	102,96	7,37	0,3	0,86	101,78	3,78	0,51	0,24	17,16
8	9,12	0,26	1,7	37,02	7,02	0,32	0,79	47,55	6,48	0,36	0,57	27,29
9	5,88	0,49	0,84	45,74	8,25	0,25	0,34	32,61	10,44	0,23	1,22	184,33
10	6,3	0,36	0,6	40,07	8,15	0,31	1,6	103,25	7,65	0,26	0,68	58,42
11	6,22	0,37	0,82	45,44	8,72	0,26	1,46	38,95	8,77	0,27	1	59,4
12	5,49	0,43	0,84	41,08	6,64	0,37	1,27	81,32	7	0,29	0,81	49,63
13	6,5	0,35	0,67	136,14	8,1	0,29	1,58	67,26	11,06	0,01	1,27	391,27
14	6,61	0,38	1,04	42,39	5,52	0,34	0,68	59,92	9,02	0,02	1,14	258,62
15	4,32	0,42	0,66	37,39	9,37	0,23	0,86	107,34	13,28	0,18	1,89	75,66
16	7,37	0,3	0,86	101,78	13,17	0,17	1,98	512,6	9,27	0,25	0,67	123,68
17	7,02	0,32	0,79	47,55	6,67	0,29	0,33	53,81	6,7	0,31	0,96	37,21
18	8,25	0,25	0,34	32,61	5,68	0,41	0,45	80,83	6,69	0,38	0,67	53,37
19	8,15	0,31	1,6	103,25	5,22	0,41	0,74	59,42	9,42	0,24	0,98	32,87

Продовження табл. 4.11

	Y1	X4	X8	X12	Y1(10)	X4(10)	X8(10)	x12(10)	Y1(25)	X4(25)	X8(25)	x12(25)
20	8,72	0,26	1,46	38,95	10,02	0,22	0,03	36,96	7,24	0,31	1,16	45,63
21	6,64	0,37	1,27	81,32	8,16	0,29	0,99	91,43	5,39	0,42	0,54	48,41
22	8,1	0,29	1,58	67,26	3,78	0,51	0,24	17,16	5,61	0,51	1,23	13,58
23	5,52	0,34	0,68	59,92	6,48	0,36	0,57	27,29	5,59	0,31	0,78	63,99
24	9,37	0,23	0,86	107,34	10,44	0,23	1,22	184,33	6,57	0,37	1,16	104,55
25	13,17	0,17	1,98	512,6	7,65	0,26	0,68	58,42	6,54	0,16	4,44	222,11
26	6,67	0,29	0,33	53,81	8,77	0,27	1	59,4	4,23	0,18	1,06	25,76
27	5,68	0,41	0,45	80,83	7	0,29	0,81	49,63	5,22	0,43	2,13	29,52
28	5,22	0,41	0,74	59,42	11,06	0,01	1,27	391,27	18	0,4	1,21	41,99
29	10,02	0,22	0,03	36,96	9,02	0,02	1,14	258,62	11,03	0,31	2,2	78,11
30	8,16	0,29	0,99	91,43	13,28	0,18	1,89	75,66				
31	3,78	0,51	0,24	17,16	9,27	0,25	0,67	123,68				
32	6,48	0,36	0,57	27,29	6,7	0,31	0,96	37,21				
33	10,44	0,23	1,22	184,33	6,69	0,38	0,67	53,37				
34	7,65	0,26	0,68	58,42	9,42	0,24	0,98	32,87				
35	8,77	0,27	1	59,4	7,24	0,31	1,16	45,63				
36	7	0,29	0,81	49,63	5,39	0,42	0,54	48,41				

	Y1	X4	X8	X12	Y1(10)	X4(10)	X8(10)	x12(10)	Y1(25)	X4(25)	X8(25)	x12(25)
37	11,06	0,01	1,27	391,27	5,61	0,51	1,23	13,58				
38	9,02	0,02	1,14	258,62	5,59	0,31	0,78	63,99				
39	13,28	0,18	1,89	75,66	6,57	0,37	1,16	104,55				
40	9,27	0,25	0,67	123,68	6,54	0,16	4,44	222,11				
41	6,7	0,31	0,96	37,21	4,23	0,18	1,06	25,76				
42	6,69	0,38	0,67	53,37	5,22	0,43	2,13	29,52				
43	9,42	0,24	0,98	32,87	18	0,4	1,21	41,99				
44	7,24	0,31	1,16	45,63	11,03	0,31	2,2	78,11				
45	5,39	0,42	0,54	48,41								
46	5,61	0,51	1,23	13,58								
47	5,59	0,31	0,78	63,99								
48	6,57	0,37	1,16	104,55								
49	6,54	0,16	4,44	222,11								
50	4,23	0,18	1,06	25,76								
51	5,22	0,43	2,13	29,52								
52	18	0,4	1,21	41,99								
53	11,03	0,31	2,2	78,11								

Знайдемо математичне сподівання та дисперсію для цих виборок (результати наведені у табл. 4.12).

Таблиця 4.12

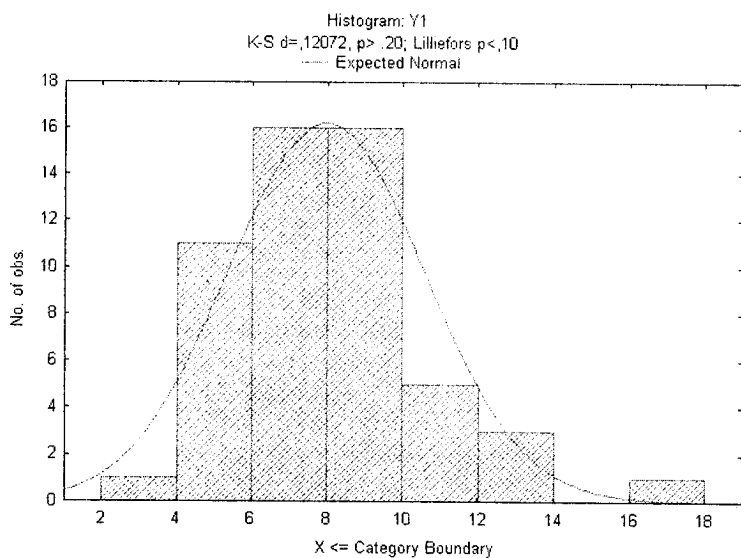
Variable	Descriptive Statistics (stacionarnost)				
	Valid N	Mean	Minimum	Maximum	Std. Dev.
Y1	53	7,97038	3,78000	18,0000	2,6102
X4	53	0,30245	0,01000	0,5100	0,1055
X8	53	1,07189	0,03000	4,4400	0,6760
X12	53	91,87358	13,58000	512,6000	92,3678
Y1 (10)	44	7,69273	3,78000	18,0000	2,6885
X4 (10)	44	0,30636	0,01000	0,5100	0,1057
X8 (10)	44	1,05909	0,03000	4,4400	0,7092
x12 (10)	44	87,72955	13,58000	512,6000	95,8556
Y1 (25)	29	8,06552	3,78000	18,0000	3,1220
X4 (25)	29	0,29310	0,01000	0,5100	0,1218
x8 (25)	29	1,12172	0,03000	4,4400	0,8304
x12 (25)	29	99,22793	13,58000	512,6000	114,7987

Проаналізуємо результати. Для Y1 значення математичного сподівання (**Mean**) дорівнює 7,97(53), 7,69(44), 8,06(25), а значення стандартного відхилення — 2,61, 2,68, 3,12. Оскільки ці значення практично збігаються, можна зробити висновок: *гіпотеза про стаціонарність підтверджується, тобто перші два моменти не залежать від початку відліку*.

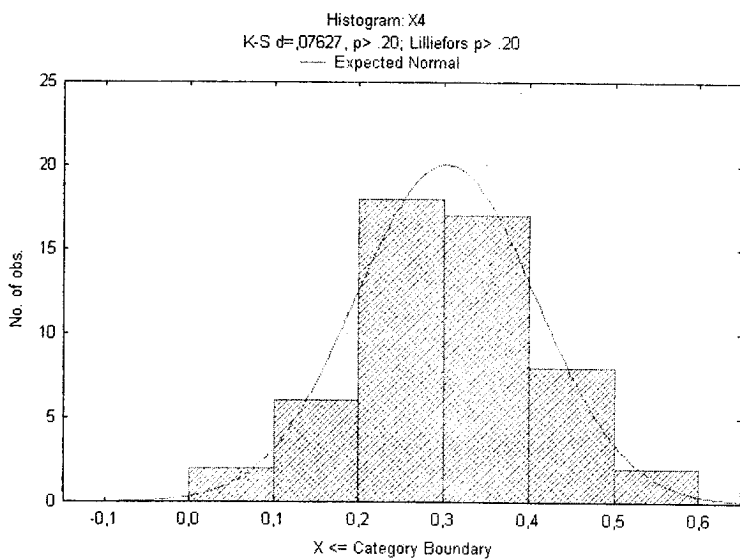
Наступним етапом є оцінка закону розподілу вхідних та вихідних змінних. З цією метою застосовуються графічні та аналітичні процедури. У якості графічних процедур використаємо гістограми та пряму нормального розподілу.

Побудуємо гістограму для Y1 та X4. Для цього увійдемо до модуля **Descriptive statistics** (Описові статистики). Після вибору змінних виберемо опцію **Normality** та побудуємо гістограми для Y1 та X4 (див. рис. 4.23). Проаналізуємо їх.

Візуальний аналіз гістограм показує, що змінні підкоряються нормальному закону розподілу. Цю гіпотезу перевіримо, аналізуючи графік нормального закону розподілу.



a)



б)

Рис. 4.23. Гістограми змінних Y1 (a) та X4 (б)

Вікно вибору модуля перевірки гіпотези про нормальність розподілу має вигляд, наведений на рис. 4.24.

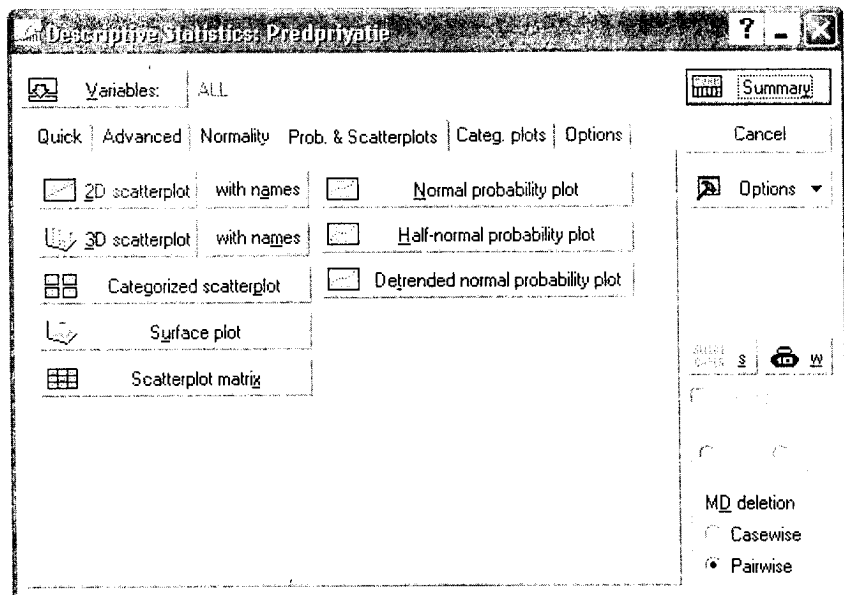


Рис. 4.24. Вікно вибору модуля перевірки гіпотези про нормальність розподілу

Для перевірки гіпотези увійдемо до опції «**Prob.&scatter-plots**» та оберемо «**Normal probability plot**», що представлені на рис. 4.25.

Аналіз гістограм показує, що практично вся площа вихідних даних лежить під кривою нормального розподілу. Для Y1 має місце «важкий хвіст» та скісність праворуч (значення $skewness = 1,2967$). Значення **Kurtosis** (3,02) показує гостроверхівність у порівнянні з нормальним законом розподілу.

Одновимірні статистики дозволяють дослідникові отримати інформацію про статистичні характеристики досліджуваного процесу та формально обрати спосіб побудови залежності $y = F(x)$.

Процедура **Normal Probability Plot** використовується для візуального порівняння відповідності функції розподілу експериментальних даних до функції нормального розподілу. Для перевірки цієї гіпотези користуються тим, що вибіркова функція

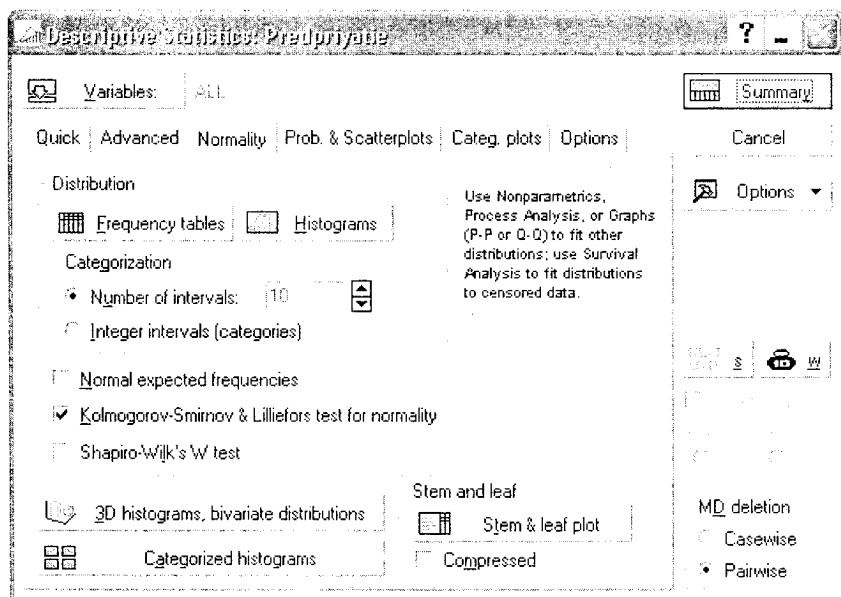


Рис. 4.25. Вікно модуля «Описові статистики»

розподілу $F_n(x)$ за великих обсягів вибірки рівномірно близька до теоретичної функції розподілу. Надалі переходять до варіаційного ряду.

Функція $F_n(x)$ — кусково-стала, яка в кожній з точок x здійснює стрибок, що дорівнює $1/n$, причому за умови $x < x(1)$ $F_n(x) = 0$, а якщо $x > x(n)$ — $F_n(x) = 1$. Для перевірки нормальності вибірки можна застосувати функцію Φ^{-1} до середини цих стрибків (значення функції береться з таблиці квантилів функції Лапласа). У результаті отримуємо точки $x_{(u)}$, $\Phi^{(-1)}((2i - 1)/2n)$ на площині (x, Z) . У залежності від того, наскільки добре ці точки «накладаються» на пряму лінію, можна судити про нормальність розподілу вибірки. Графіки зображені на рис. 4.26.

Аналіз графіків показує, що змінні Y1 та X4 розподілені за нормальним законом. Однак графічні процедури вимагають від користувача при прийнятті рішення про вид закону розподілу високої кваліфікації. Для більш точного оцінювання закону розподілу необхідно використовувати аналітичні процедури, до яких слід віднести описові статистики та аналітичні процедури Шапіро — Уїлкса, Колмогорова — Смирнова та Ліліфорса.

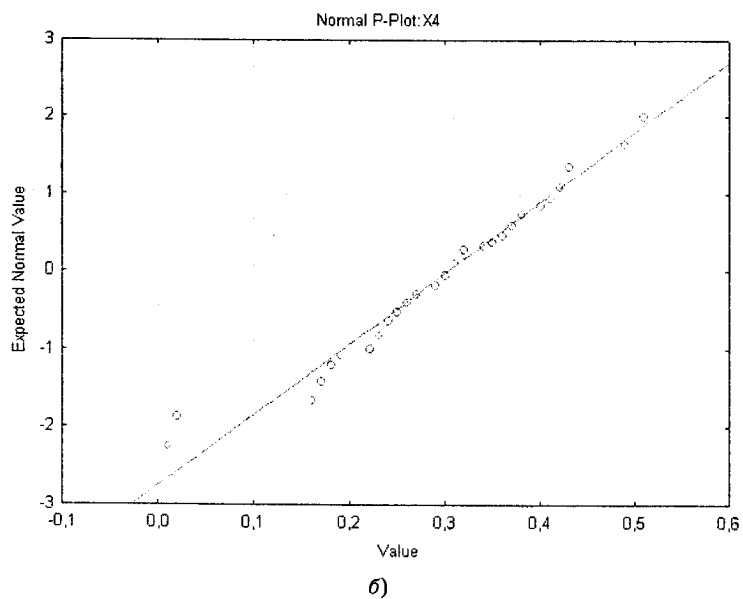
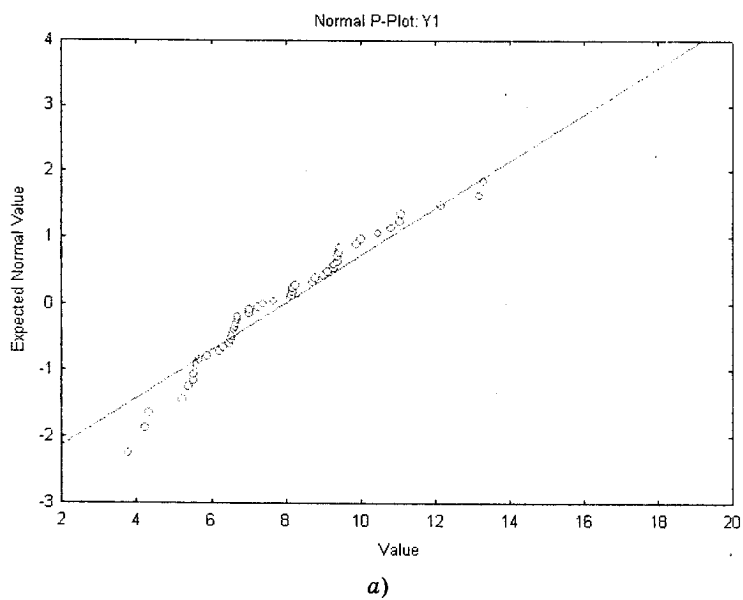


Рис. 4.26. Графіки нормальності розподілу вибірки Y (a) та $X4$ (б)

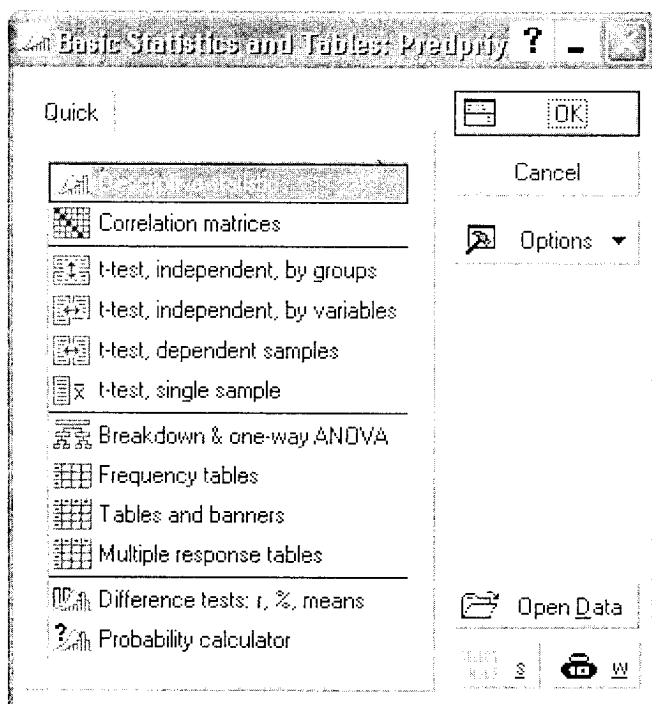


Рис. 4.27. Вікно вибору аналітичної процедури перевірки гіпотези про нормальність розподілу (**Frequency Tables**)

Обчислення одномірних статистик необхідне для оцінки мір положення (**Average** — середнє, **Median** — медіана, **Mode** — мода, **Geometric mean** — геометричне середнє), мір розсіяння (**Variance** — дисперсія, **Standard deviation** — стандартне відхилення, **Standard error** — стандартна помилка, **Minimum**, **Maximum**, **Range** — розмах, **Lower quartile** — нижній кuartиль, **Upper quartile** — верхній кuartиль, **Interquartile range** — міжкuartильний діапазон), статистику опису форми (**Skewness** — асиметрія, яка характеризує скісність розподілу: якщо асиметрія більша за 0, то крива розподілу скісна праворуч від 0, якщо менша за 0 — ліворуч; **Kurtosis** — ексцес, який характеризує гостроверхівність

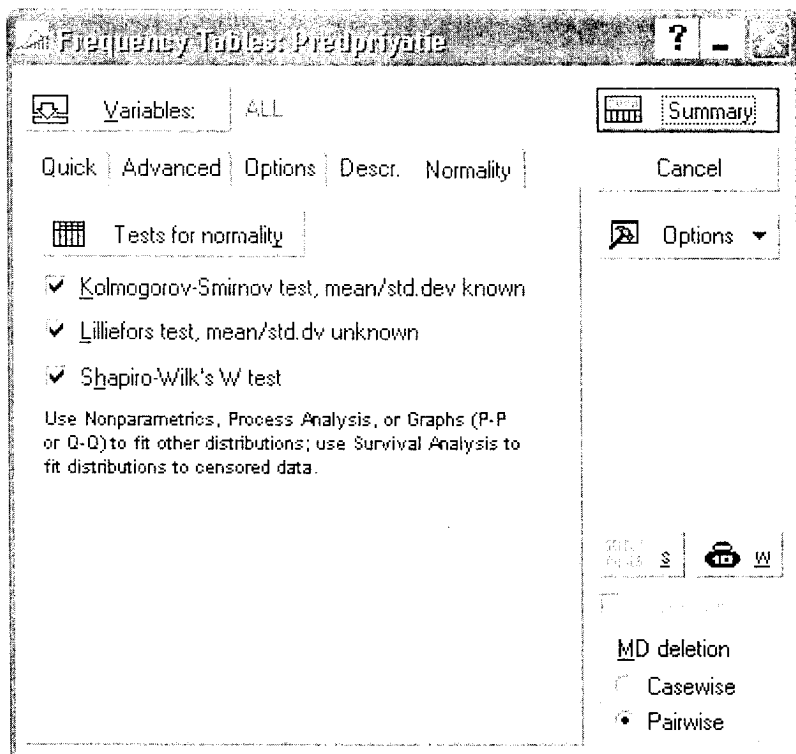


Рис. 4.28. Вікно модуля Frequency Tables

розподілу у порівнянні з нормальним розподілом $N(0, 1)$, якщо ексцес менший за 0, то крива має більш плоску верхівку у порівнянні з кривою $N(0, 1)$.

Standardized skewness та **Standardized kurtosis** — відповідно стандартизовані асиметрія та ексцес (поділені на стандартні відхилення), що характеризують близькість розподілу до нормального. Якщо ці коефіцієнти знаходяться у діапазоні $[-2, 2]$, то можна очікувати, що розподіл близький до нормального. Якщо щільність розподілу симетрична, то коефіцієнт асиметрії дорівнює 0. Для нормального розподілу коефіцієнт ексцесу дорівнює 0. Якщо розподіл сконцентрований навколо середнього більше, ніж нормальний, то він перевищує 0. Результати розрахунку одномірних статистик подані у табл. 4.13.

Таблиця 4.13

Результати розрахунку оновимірних статистик для залежної змінної
та незалежних змінних

	Valid N	Mean	Variance	Std.Dev.	Standard	Skewness	Std.Err.	Kurtosis	Std.Err.
Y1	53	7,97	7	2,61	0,359	1,296765	0,327446	3,02645	0,644420
Y2	53	113,19	13674	116,94	16,063	2,523834	0,327446	7,25800	0,644420
Y3	53	13,70	40	6,30	0,865	1,011741	0,327446	0,57388	0,644420
X4	53	0,30	0	0,11	0,014	-0,400756	0,327446	0,68504	0,644420
X5	53	0,73	0	0,06	0,008	-0,275835	0,327446	-0,59911	0,644420
X6	53	0,30	0	0,16	0,021	0,187033	0,327446	-0,22209	0,644420
X7	53	1,34	0	0,14	0,019	0,782345	0,327446	2,46385	0,644420
X8	53	1,07	0	0,68	0,093	2,544068	0,327446	10,92844	0,644420
X9	29	0,47	0	0,39	0,072	1,878882	0,433547	3,87634	0,845239
X10	29	1,58	0	0,47	0,087	-0,092746	0,433547	0,07446	0,845239
X11	53	14707,79	98151212	9907,13	1360,849	1,668189	0,327446	3,28661	0,644420
X12	53	91,87	8532	92,37	12,688	2,748663	0,327446	9,03254	0,644420
X13	53	27528,17	378330743	19450,73	2671,763	1,871513	0,327446	4,50595	0,644420
X14	53	5,97	6	2,38	0,327	0,774911	0,327446	-0,09246	0,644420
X15	32	105,07	1005	31,70	5,604	-0,082640	0,414457	-0,80070	0,809371
X16	32	12,01	24	4,85	0,858	0,538782	0,414457	0,17335	0,809371
X17	32	19,31	24	4,86	0,858	0,248969	0,414457	0,22291	0,809371

У ППП «STATISTICA» для розрахунку **Standardized skewness** та **Standardized kurtosis** необхідно поділити значення **Skewness** та **Kurtosis** на відповідні їм значення стандартного відхилення (**Std.Dev.**) й цими значеннями доповнити таблицю.

Аналіз таблиці показує, що значення **Standardized skewness** та **Standardized kurtosis** знаходяться у діапазоні $[-2, 2]$. Однак коректне використання цього критерію пов'язане з асимптотичною вихідних даних. Реальні дані, як правило, мають зрізану вибірку. У цьому випадку доцільно перейти до більш точних критеріїв оцінки нормальності закону розподілу, до яких слід віднести аналітичні критерії Шапіро — Уїлкса, Колмогорова — Смирнова та Ліліфорта.

Результати їх використання представлені на рис. 4.29.

Variable	Tests of Normality (Predpriyatie)					
	N	max D	K-S p	Lilliefors p	W	p
Y1	53	0,120724	p > .20	p < ,10	0,917544	0,001366
Y2	53	0,255183	p < ,01	p < ,01	0,701041	0,000000
Y3	53	0,136239	p > .20	p < ,05	0,916142	0,001204
X4	53	0,076271	p > .20	p > .20	0,970940	0,221553
X5	53	0,083721	p > .20	p > .20	0,975129	0,332215
X6	53	0,052057	p > .20	p > .20	0,984941	0,738861
X7	53	0,096809	p > .20	p > .20	0,939328	0,009586
X8	53	0,177186	p < ,10	p < ,01	0,803938	0,000001
X9	29	0,195349	p < ,20	p < ,01	0,811980	0,000138
X10	29	0,100038	p > .20	p > .20	0,983220	0,911330
X11	53	0,200530	p < ,05	p < ,01	0,844454	0,000007
X12	53	0,225963	p < ,01	p < ,01	0,681074	0,000000
X13	53	0,183701	p < ,10	p < ,01	0,824770	0,000002
X14	53	0,151393	p < ,20	p < ,01	0,935568	0,006739
X15	32	0,108378	p > .20	p > .20	0,962596	0,322968
X16	32	0,101186	p > .20	p > .20	0,969719	0,491687
X17	32	0,122796	p > .20	p > .20	0,981547	0,843130

Рис. 4.29. Результати розрахунку модуля **Frequency Tables**

Умови віднесення вихідних даних до нормального закону розподілу:

☑ для критерію Колмогорова — Смирнова: якщо $p > 0,05$, то ця змінна відповідає нормальному закону розподілу;

☑ для критерію Лїліфорса: якщо $p > 0,05$, то ця змінна відповідає нормальному закону розподілу;

☑ для критерію Шапіро — Уїлкса: якщо значення більше за 0,96 (див. табличне значення).

Розбіжності у результатах оцінки нормальності закону розподілу пов'язані з чутливістю критеріїв до обсягу вибірки. Однак у більшості випадків результати співпадають.

Зазначимо, що використовувати критерій χ^2 слід обережно, бо в основу статистики Пірсона є її асимптотичний характер.

Нормальний розподіл грає величезну роль при побудові статистичних моделей. Багато критеріїв оцінки адекватності, перевірки статистичних гіпотез та низка статистичних висновків ґрунтуються на нормальному законі розподілу. У зв'язку з цим при розв'язанні функціональних задач в ІУС важливо вміти користуватися засобами оцінок нормального розподілу.

Отже, попередній аналіз вихідних даних показує, що регресійний аналіз можливо використовувати для побудови лінійної моделі за вихідними даними. При виборі структури моделі необхідно враховувати не лише лінійні, але й ступеневі, й члени взаємодії, тобто представити модель у вигляді полінома Колмогорова — Габора

$$y = \sum_{j=1}^p b_j x_j + \sum_{l=p+1}^{2p} b_l x_l^2 + \sum_{z=2p+1}^k b_z x_z x_{z+1},$$

де $j = [1, p]$, $l = [p + 1, 2p]$, $z = [2p + 1, k]$, $k = C_p^2$ (число сполучень); b_j — коефіцієнти моделі; x_j — незалежні змінні.

Це все дозволяє обрати структуру моделі з членами взаємодії та різними степенями. Необхідно на кожному кроці перевіряти інформаційну матрицю $X^T X$ на мультиколінеарність.

Побудуємо модель виду $Y1 = F(X4, X5, X6)$ за допомогою регресійного аналізу у ППП «STATISTICA». До уваги візьмемо такі показники:

Y1 — продуктивність праці;

X4 — трудомісткість одиниці продукції;

X5 — питома вага працівників у складі підприємства;

X6 — питома вага покупних виробів.

Вихідні дані наведені у табл. 4.14.

Таблиця 4.14

	Y1_	X4_	X5_	X6_		Y1_	X4_	X5_	X6_
1	9,26	0,23	0,78	0,40	27	5,68	0,41	0,71	0,20
2	9,38	0,24	0,75	0,26	28	5,22	0,41	0,79	0,24
3	12,11	0,19	0,68	0,40	29	10,02	0,22	0,76	0,54
4	10,81	0,17	0,70	0,50	30	8,16	0,29	0,78	0,40
5	9,35	0,23	0,62	0,40	31	3,78	0,51	0,62	0,20
6	9,87	0,43	0,76	0,19	32	6,48	0,36	0,75	0,64
7	8,17	0,31	0,73	0,25	33	10,44	0,23	0,71	0,42
8	9,12	0,26	0,71	0,44	34	7,65	0,26	0,74	0,27
9	5,88	0,49	0,69	0,17	35	8,77	0,27	0,65	0,37
10	6,30	0,36	0,73	0,39	36	7,00	0,29	0,66	0,38
11	6,22	0,37	0,68	0,33	37	11,06	0,01	0,84	0,35
12	5,49	0,43	0,74	0,25	38	9,02	0,02	0,74	0,42
13	6,50	0,35	0,66	0,32	39	13,28	0,18	0,75	0,32
14	6,61	0,38	0,72	0,02	40	9,27	0,25	0,75	0,33
15	4,32	0,42	0,68	0,06	41	6,70	0,31	0,79	0,29
16	7,37	0,30	0,77	0,15	42	6,69	0,38	0,62	0,30
17	7,02	0,32	0,78	0,08	43	9,42	0,24	0,70	0,56
18	8,25	0,25	0,78	0,20	44	7,24	0,31	0,66	0,42
19	8,15	0,31	0,81	0,20	45	5,39	0,42	0,69	0,26
20	8,72	0,26	0,79	0,30	46	5,61	0,51	0,71	0,16
21	6,64	0,37	0,77	0,24	47	5,59	0,31	0,73	0,45
22	8,10	0,29	0,78	0,10	48	6,57	0,37	0,65	0,31
23	5,52	0,34	0,72	0,11	49	6,54	0,16	0,82	0,08
24	9,37	0,23	0,79	0,47	50	4,23	0,18	0,80	0,68
25	13,17	0,17	0,77	0,53	51	5,22	0,43	0,83	0,03
26	6,67	0,29	0,80	0,34	52	18,00	0,40	0,70	0,02
					53	11,03	0,31	0,74	0,02

Для побудови багатовимірної моделі регресії необхідно використовувати модуль **Multiple Regression** ППП «STATISTICA». Після вибору модуля необхідно обрати файл даних та змінні, які увійдуть у модель регресії, в діалоговому вікні **Multiple Linear Regression**.

Припустимо, що вибірка характеризується лінійним взаємозв'язком з випадковістю, яка представлена моделлю множинної

лінійної регресії, у відповідності до якої спостережуване значення Y визначається взаємозв'язком у генеральній сукупності плюс нормально розподілена випадкова помилка. Передбачається також, що ці випадкові помилки для різних спостережень (елементарних одиниць наших даних) не залежать одна від одної.

Коефіцієнти регресії для генеральної сукупності показують середній вплив кожної з X -змінних на Y (у цій генеральній сукупності) за умови, що решта X -змінних залишиться незмінними.

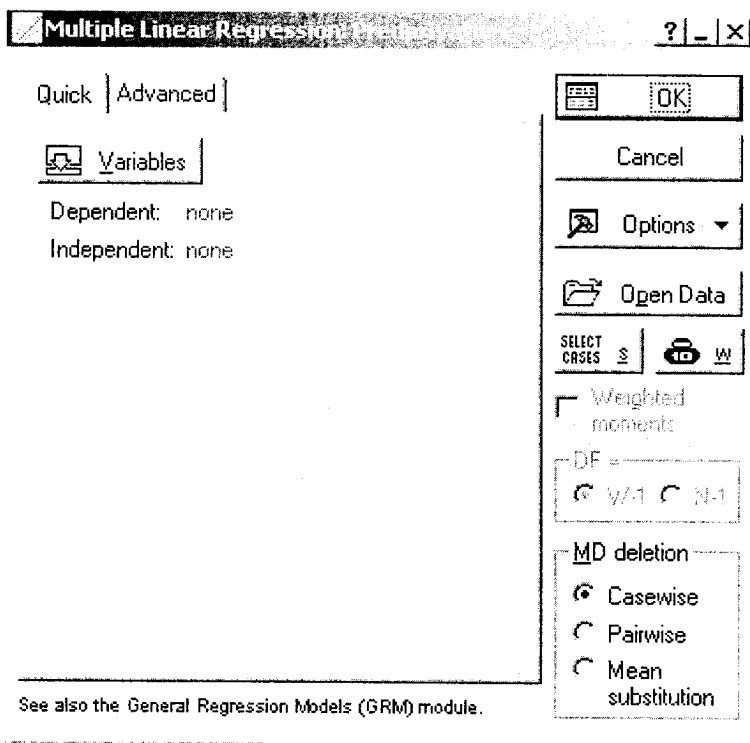


Рис. 4.30. Вікно вибору змінних для побудови моделі регресії

Після підтвердження вибору змінних на екрані з'являється таблиця з результатами множинної регресії (рис. 4.31).

У цій таблиці відображені такі параметри:

☒ **Dep. Var.** — ім'я залежної змінної (у нашому випадку — продуктивність);

Multiple Regression Results: Predpriyatie

?

×

Multiple Regression Results

Dependent: Производительн

Multiple R = ,52087901

F = 6,081439

R² = ,27131390

df = 3,49

No. of cases: 53

adjusted R² = ,32670047

p = ,001330

Standard error of estimate: 2,295365062

Intercept: 18,680923545 Std. Error: 5,475228 t(49) = 3,3507 p = ,0016

X4 -Трудоемк. beta=-,62 X5 - уд. вад beta=-,15 X6 - удельный beta=-,17

(significant betas are highlighted)

Alpha for highlighting effects: .05

OK

Quick | Advanced | Residuals/assumptions/prediction

Summary: Regression results

Cancel

Options

Рис. 4.31. Таблица результатів побудови моделі з використанням багатовимірною регресійного аналізу

- ☒ **No.of Cases** — кількість спостережень, що використовуються при побудові моделі;
- ☒ **R** — множинний коефіцієнт кореляції;
- ☒ **F** — критерій Фішера (його можна подивитись у таблицях, *див.* додаток; модель вважається доброю, якщо критерій більший за табличний у 5–6 разів);

☒ **R²** — коефіцієнт детермінації. Вказує, який відсоток варіації Y пояснюється впливом усіх X-змінних, включених у модель. У нашому прикладі коефіцієнт детермінації $R^2 = 0,22$, або 22 %, вказує на те, що незалежні змінні X (X4 — трудомісткість одиниці продукції; X5 — питома вага робітників у складі підприємства; X6 — питома вага покупних виробів) пояснюють 22,0 % продуктивності праці. При цьому 78 % залишаються непоясненими та пов'язуються з впливом інших факторів. Бажано, щоб значення R^2 було якомога більшим (великі значення R^2 свідчать

про те, що досліджуваний взаємозв'язок є достатньо сильним). В ідеальному випадку $R^2 = 100\%$; це можливо лише за умови, що всі помилки прогнозування дорівнюють 0;

☑ $df = n - p - 1$ — ступені вільності, де p — кількість вхідних змінних;

☑ **adjusted R^2** — скоректований коефіцієнт детермінації, з урахуванням кількості вимірювань, що обчислюється за формулою **adjusted $R^2 = 1 - (1 - R^2) \cdot (n/(n - p))$** .

☑ p — оцінка значущості моделі або рівень значущості (оцінка значущості F -критерію);

☑ **Standard error of estimate** — стандартна помилка в оцінці, яка характеризує міру розсіювання величин, що спостерігаються, відносно лінії регресії. Стандартна помилка оцінки $S_e = 2,29$ вказує залишок варіації продуктивності праці. Після того, як використали X -змінні у рівнянні регресії для прогнозування.

Якщо нічого не відомо про X -змінні, доцільно використовувати у якості оптимальної приблизної оцінки середнє значення продуктивності ($Y = 7,97$), і ми помилятимемося приблизно на $S_Y = 5,26$. Але якщо відомі такі характеристики, як трудомісткість одиниці продукції, питома вага робітників у складі підприємства, питома вага покупних виробів, то для прогнозування тарифів можна скористатися рівнянням регресії; в цьому випадку ми помилимося приблизно на $S_e = 2,29$. Таке зменшення помилки прогнозування (з 5,26 до 2,29) і є однією з переваг використання регресійного аналізу. Стандартна помилка в оцінці продуктивності праці $S_e = 2,29$ свідчить про те, що фактична продуктивність праці, як правило, відхиляється від прогнозованої не більше ніж на 2,29 (мова йде про стандартне відхилення);

☑ **Intercept** — вільний член;

☑ **Std Error** — стандартна помилка в оцінці вільного члена у рівнянні регресії;

☑ t — критерій Стьюдента (при ступенях вільності 49);

☑ p — значущість коефіцієнтів (коефіцієнт значущий, якщо p менше 0,05).

На підставі табличних даних попередньо можна зробити висновки про доцільність використання цієї моделі з практичною метою. У ППП «STATISTICA» червоним кольором вказаний коефіцієнт, який є значущим.

На закладці **Quick** оберемо опцію **Підсумковий результат регресії** — **Summary: Regression results** (див. рис. 4.32).

		Regression Summary for Dependent Variable: Производительность труда (Predpriyatie)					
		R= .52037801 R²= .27131390 Adjusted R²= .22670047 F(3,49)=6.0814 p<.00133 Std Error of estimate: 2.2954					
N=53		Beta	Std. Err. of Beta	B	Std. Err. of B	t(49)	p-level
Intercept				18,6809	5,57423	3,34970	0,00129
X4 -Трудоемк. ед. продукции		-0,154650	0,154650	-15,4650	15,4650	-1,00000	0,31691
X5 - уд. вес раб в сост ППП		-0,150262	0,134864	-7,0669	6,341912	-1,11432	0,270577
X6 - удельный вес покупок изделий		-0,167827	0,146260	-2,8226	2,459664	-1,14745	0,256766

Рис. 4.32. Результати регресійного аналізу

У цій таблиці подано таку інформацію:

☑ **Beta** — коефіцієнти рівняння регресії в стандартизованому вигляді. Стандартизовані коефіцієнти призначені для оцінки щільності зв'язку між вхідними й вихідними змінними;

☑ **Std Err of Beta** — дані цього стовпця потрібні для побудови довірчих інтервалів рівняння регресії та коефіцієнтів. Стандартний коефіцієнт (**Beta**) показує ступінь впливу, змінюючись від -1 до 1 (чим ближче до 1, тим більший ступінь впливу); знак при коефіцієнті показує напрямок дії;

☑ **B** — нормальні коефіцієнти та їх стандартна помилка;

☑ **t(49)** — оцінки значущості за критерієм Стюдента (бажано мати також табличне значення та порівнювати його з розрахунковим);

☑ **p-level** — на підставі даних у цьому стовпці можна судити про значущість коефіцієнтів (якщо $p < 0,05$ — коефіцієнт є значущим).

Рівняння регресії має такий вигляд:

$$Y_1 = 18,6809 - 15,46 * X_4 - 7,066 * X_5 - 2,82 * X_6.$$

У нашому прикладі коефіцієнти 7,066 та 2,82 при змінних $*X_5$ та $*X_6$ не є значущими.

Перейдемо на закладку **Advanced** та виберемо опцію **ANOVA** (рис. 4.33).

Analysis of Variance; DV: Производительность труда (Predpriyatie)					
Effect	Sums of Squares	df	Mean Squares	F	p-level
Regress.	96,1239	3	32,04128	6,081439	0,001301
Residual	258,1663	49	5,26870		
Total	354,2902				

Рис. 4.33. Дисперсійна таблиця

На підставі цієї таблиці можна судити про значущість дисперсії відносно регресії та дисперсії залишків; їх співвідношення показує коефіцієнт F . Sums of Squares — сума квадратів. Бажаючи F порівняти з табличним значенням для відповідного рівня значущості α та відповідних ступенів вільності ($\alpha = 0,05$, ступені вільності 3 та 49).

Далі оберемо опцію Covariances of Coefficients (рис. 4.34).

Variable	Covariances of Regression Coefficients E, DV: Производительность труда (Predpriyatie)		
	X4 -Трудоёмк. ед. произ.	X5 - уд. вес раб в сост	X6 - удельный вес покупных изд
X4 -Трудоёмк. ед. продукции	14,68317	10,27701	5,178372
X5 - уд. вес раб в сост ППП	10,27701	40,21985	4,400050
X6 - удельный вес покупных изделий	5,17837	4,40005	6,050931

Рис. 4.34. Коваріаційна матриця

Коваріаційна матриця показує наявність лінійного статистичного зв'язку між вхідними змінними.

Переходимо до опції Current sweep matrix (рис. 4.35).

Variable	Current Status of Sweep Matrix; DV: Производительность труда (Predpriyatie)			
	X4 -Трудоёмк.	X5 - уд. вес раб	X6 - удельный	Производительность труда
X4 -Трудоёмк. ед. продукции	-1,61243	-0,59388	-0,83669	-0,624820
X5 - уд. вес раб в сост ППП	-0,59388	-1,22307	-0,37412	-0,150282
X6 - удельный вес покупных изделий	-0,83669	-0,37412	-1,43849	-0,167827
Производительность труда	-0,62482	-0,15028	-0,16783	0,728686

Рис. 4.35. Матриця коефіцієнтів

Аналіз таблиці показує, що між X_5 та X_4 існує достатньо сильний зв'язок. Між X_5 та X_6 — також сильний зв'язок, який може призвести до поганої зумовленості інформаційної матриці $X^T X$.

Перейдемо до наступної опції — Partial correlation (див. рис. 4.36).

Variable	Variables currently in the Equation; DV: Производительность труда (Predpriyatie)						
	Beta in	Partial Cor.	Semipart Cor.	Tolerance	R-square	t(49)	p-level
X4 -Трудоёмк. ед. продукции	-0,62482	0,499400	-0,430056	0,620482	0,373815	-4,05499	0,000131
X5 - уд. вес раб в сост ППП	-0,150282	-0,167209	-0,135889	0,817616	0,182384	-1,11432	0,270577
X6 - удельный вес покупных изделий	-0,167827	-0,161763	-0,139929	0,696172	0,304828	-1,14745	0,256766

Рис. 4.36. Частковий коефіцієнт кореляції

Частковий коефіцієнт кореляції показує лінійний зв'язок між змінними без урахування впливу проміжних обчислень. Так,

наприклад, X_4 пов'язаний з Y без урахування впливу X_5 та X_6 , значення яких дорівнюють $-0,157$ та $-0,161$ відповідно. До того ж, X_5 пов'язаний з Y без урахування впливу X_4 та X_6 , значення яких дорівнюють $-0,499$ та $-0,161$ відповідно.

У стовпчику **Beta in** наведені стандартизовані коефіцієнти рівняння регресії. У **Semipart Cor.** наведені значення коефіцієнтів кореляції з урахуванням помилок. Стовпчик **Tolerance** показує оцінку чистого впливу вхідної змінної на Y . Стовпчик **R-square** містить значення коефіцієнта детермінації, який показує внесок дисперсії кожної змінної у загальну дисперсію Y . Стовпчик **t(49)** містить значення коефіцієнтів Стюдента, який показує значущість. Стовпчик **p-level** — оцінка значущості з урахуванням $\alpha = 0,05$.

Далі приступаємо до аналізу залишків. Для цього переходимо до закладки «**Residuals/assumptions/predictions**» (рис. 4.37).

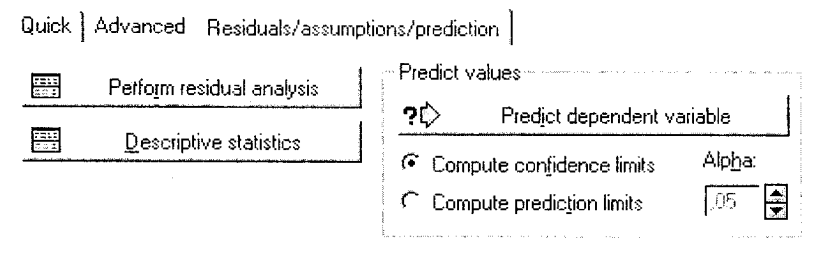


Рис. 4.37. Закладка **Residuals/assumptions/predictions**

Обравши опцію «**Perform residual analysis**», виконуємо перевірку передумови про нормальність розподілу залишків. На рис. 4.38 — вікно аналізу залишків.

Оберемо опцію «**Summary: Residuals & predicted**» отримуємо таблицю вигляду (див. рис. 4.39)

У цій таблиці відображені такі параметри:

- ☒ **Observed Value** — вихідні значення вихідної змінної Y ;
- ☒ **Predicted Value** — передбачені значення Y за моделлю;
- ☒ **Residual** — залишки $Y - Y_m$;
- ☒ **Standard Pred. v.** — стандартне значення передбачення;
- ☒ **Standard Residual** — стандартне значення залишків;
- ☒ **Std. Err. Pred. Val.** — стандартна помилка передбачення;
- ☒ **Mahalanobis Distance** — значення відстані Махаланобіса.

Residual Analysis: Predpriyatje ? | _ | X

Dependent: Производительность труда Multiple R : ,52087801 F = 6,081439
 R²: ,27131390 df = 8,49
 No. of cases: 53 adjusted R²: ,22670047 p = ,001330
 Standard error of estimate: 3,225365062
 Intercept: 18,580322545 Std. Error: 5,575228 t(49) = 3,3507 p < ,0016

Quick | Advanced | Residuals | Predicted | Scatterplots | Probability plots | Outliers | Save | Summary | Cancel | Options ▾

Summary: Residuals & predicted
 Normal plot of residuals

Рис. 4.38. Аналіз залишків

Case No.	Predicted & Residual Values (Predpriyatje)								
	Dependent variable: Производительность труда								
	Observed Value	Predicted Value	Residual	Standard Pred. v.	Standard Residual	Std. Err. Pred. Val	Mahalanobis Distance	Deleted Residual	Cook's Distance
1	9,26000	8,48353	0,77647	0,37743	0,33828	0,468355	1,18383	0,81020	0,001297
2	9,38000	8,93609	0,44391	0,71029	0,19340	0,421124	0,76920	0,45938	0,000337
3	12,11000	9,60869	2,30131	1,35209	1,00259	0,649466	3,18193	2,50159	0,023773
4	10,81000	9,69432	1,11568	1,26797	0,48606	0,633666	2,98183	1,20772	0,005275
5	9,35000	9,61424	-0,26424	1,20907	-0,11512	0,866412	6,42769	-0,30815	0,000642
6	9,87000	6,12531	3,74469	-1,35706	1,63141	0,581110	2,35172	4,00114	0,046687
7	8,17000	8,02335	0,14666	0,03896	0,06389	0,337110	0,14048	0,14989	0,000023
8	9,12000	8,40147	0,71853	0,31707	0,31304	0,442305	0,94969	0,74624	0,000981
9	5,88000	5,74876	0,13125	-1,63402	0,05718	0,651078	3,20262	0,14273	0,000078
10	6,30000	6,85511	-0,55511	-0,82029	-0,24184	0,492200	1,40989	-0,58186	0,000739
11	6,22000	7,22319	-1,00319	-0,54956	-0,43705	0,468100	1,18147	-1,04673	0,002162
12	5,49000	6,09729	-0,60729	-1,37766	-0,26457	0,549674	2,00089	-0,64424	0,001129
13	6,50000	7,70199	-1,20199	-0,19740	-0,52366	0,526971	1,75964	-1,26887	0,004027

Рис. 4.39. Залишки та передбачення

За аналізом даних таблиці можна визначити значення точок експерименту, які істотно впливають на адекватність моделі.

Виберемо опцію «Normal plot of residual» (рис. 4.40). За графіком оцінимо нормальність розподілу залишків. При нормальному розподілі залишки повинні наближатися до логарифмічної прямої нормального закону розподілу. У нашому випадку має місце незначне відхилення від нормального закону розподілу.

Для більш детального аналізу виберемо закладку «Advanced» (див. рис. 4.41):

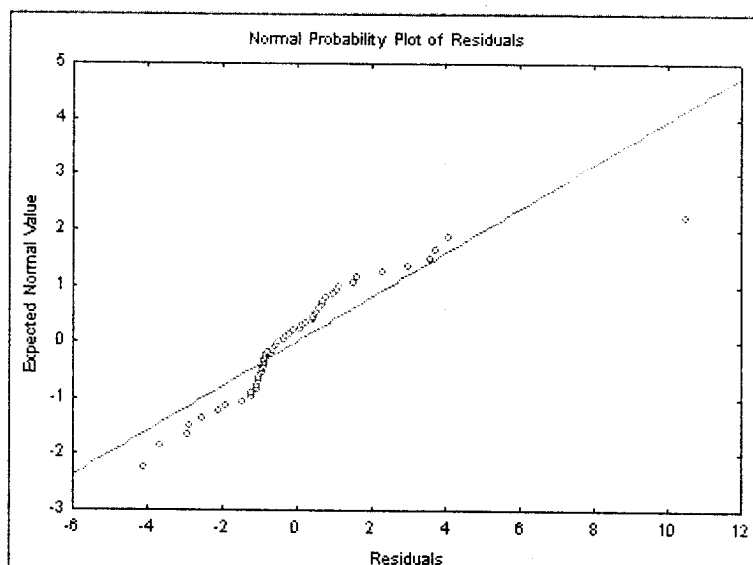


Рис. 4.40. Графік розподілу залишків

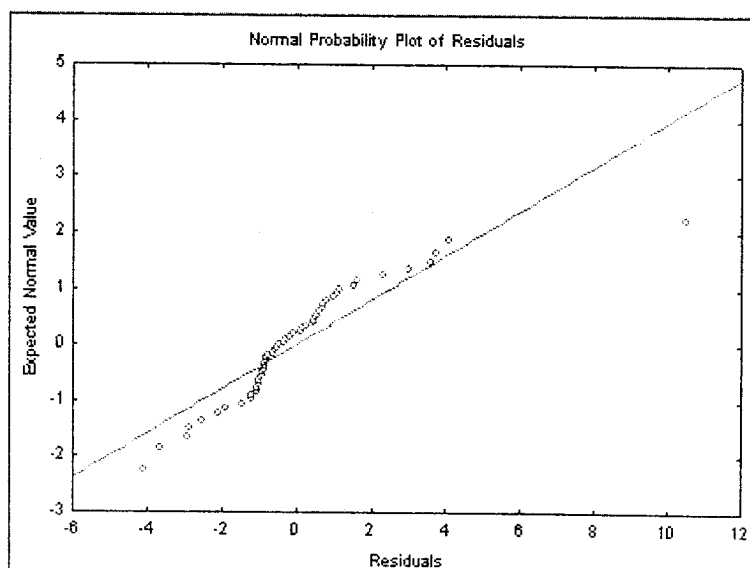


Рис. 4.41. Закладка Advanced

Опція **Summary: Residuals & predicted** була описана вище (співпадає з такою самою опцією на закладці **Quick**).

Обравши опцію **Descriptive statistic**, отримуємо вікно такого вигляду (рис. 4.42):

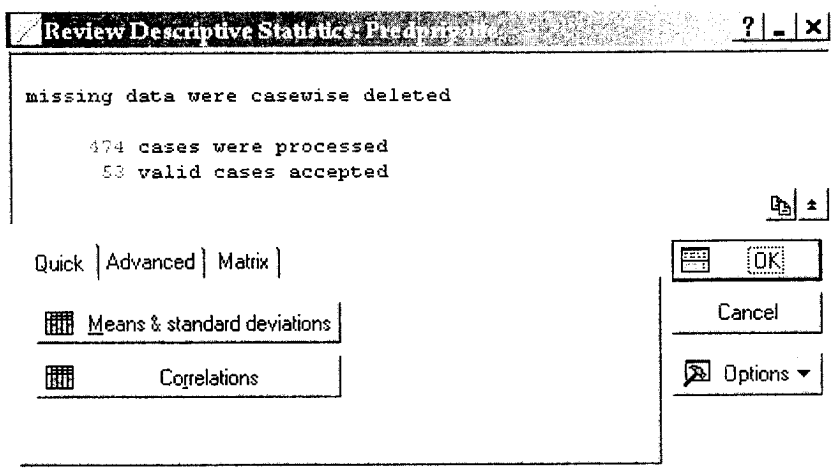


Рис. 4.42. Вікно Descriptive statistic

Обравши опцію **Means & standart deviations**, отримуємо таку таблицю (рис. 4.43):

Variable	Means and Standard Deviations (Predpriyat)		
	Means	Std.Dev.	N
X4 -Трудоемк. ед. продукции	0,302453	0,105483	53
X5 - уд. вес раб в сост ППП	0,733585	0,055508	53
X6 - удельный вес покупных изделий	0,301132	0,155200	53
Производительность труда	7,970377	2,610225	53

Рис. 4.43. Математичне сподівання та стандартне відхилення

На підставі аналізу даних таблиці можна стверджувати, що вибірка має однорідний характер і цілком імовірно, що дисперсія вихідної змінної не описується впливом X4, X5, X6, оскільки є помилки в їх стандартних відхиленнях.

Оберемо опцію **Correlation** (рис. 4.44):

Variable	Correlations (Predpriyatie)			
	X4 -Трудоемк. ед	X5 - уд. вес раб в	X6 - удельный	Производительность труда
X4 -Трудоемк. ед. продукции	1,000000	-0,334244	-0,494718	-0,491562
X5 - уд. вес раб в сост ППП	-0,334244	1,000000	-0,065663	0,069560
X6 - удельный вес покупных изделий	-0,494718	-0,065663	1,000000	0,151151
Производительность труда	-0,491562	0,069560	0,151151	1,000000

Рис. 4.44. Кореляційна матриця

Аналіз кореляційної матриці був представлений вище.

Повернемося до вікна аналізу залишків (натиснувши на кнопку **Cancel**). Переходимо до опції **Durbin-Watson statistic** (рис. 4.45).

	Durbin-Watson d (Predpriyatie) and serial correlation of residuals		
	Durbin- Watson d	Serial Corr.	
Estimate	1,579915	0,198413	

Рис. 4.45. Критерій Дарбіна — Ватсона

Критерій Дарбіна — Ватсона оцінює статистичний зв'язок між залишками (див. розділ 4.2.8).

Інтерпретація коефіцієнта Дарбіна — Ватсона

Якщо залишки корельовані, оцінки дисперсії параметрів за методом найменших квадратів виявляються зміщеними й не-ефективними, то для встановлення наявності кореляції залишків використовується критерій Дарбіна — Ватсона. Цей критерій є вельми жорстким, оскільки він вимагає відсутності статистичної залежності залишків від x .

У нашому випадку за умови, що $N = 53$, $\alpha = 0,05$ та $K = 3$ за табл. 4.1 знаходимо $D_l = 1,42$ та $D_u = 1,67$.

Розрахункове значення критерію Дарбіна — Ватсона дорівнює $d = 1,579$ і потрапляє в інтервал невизначеності. Таким чином, на підставі критерію Дарбіна — Ватсона можна судити про порушення передумови щодо відсутності кореляції між зміними.

Переходимо до закладки **Residuals**, на якій проводять графічний аналіз залишків на відповідність нормальному закону розподілу. Оберемо опцію **Histogram of Residuals** (рис. 4.46).

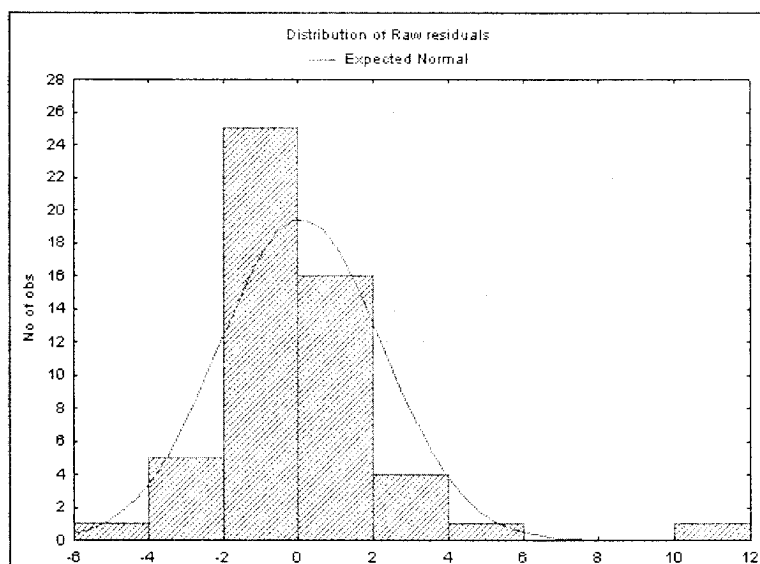


Рис. 4.46. Гістограма розподілу залишків

Аналіз гістограми розподілу залишків показує, що має місце асиметрія ліворуч, тобто скісність ліворуч та наявність «важких хвостів» праворуч.

Увійшовши в опцію **Casewise plot of residuals**, отримаємо таблицю з графіком відхилень залишків від нульового значення з розмахом 3σ (рис. 4.47).

Raw Residuals				Raw Residual (Predprnyatie)						
Dependent variable: Производительность труда				Observed	Predicted	Residual	Standard	Standard	Std.Err.	Distance
Case	-3s	0	+3s	Value	Value	Value	Residual	Pred.Val	Distance	Deleted
1	.	.	*	9,26000	8,48353	0,77647	0,37743	0,33828	0,468355	1,16383
2	.	.	*	9,38000	8,93609	0,44391	0,71029	0,19340	0,421124	0,76920
3	.	.	*	12,11000	9,60969	2,50131	1,35209	1,00259	0,649466	3,16193
4	.	.	*	10,81000	9,69432	1,11568	1,26797	0,48606	0,633666	2,98183
5	.	.	*	9,35000	9,61424	-0,26424	1,20907	-0,11512	0,866412	6,42769
6	.	.	*	9,87000	6,12531	3,74469	-1,35706	1,63141	0,581110	2,35172
7	.	.	*	8,17000	8,02335	0,14665	0,03696	0,06389	0,337110	0,14048
8	.	.	*	9,12000	8,40147	0,71853	0,31707	0,31304	0,442305	0,94969
9	.	.	*	5,88000	5,74876	0,13125	-1,63402	0,05718	0,651078	3,20262
10	.	.	*	6,30000	6,85511	-0,55511	-0,82029	-0,24184	0,492200	1,40989
11	.	.	*	6,22000	7,22319	-1,00319	-0,54956	-0,43705	0,468100	1,18147
12	.	.	*	5,48000	6,09729	-0,60729	-1,37766	-0,26457	0,549674	2,00089
13	.	.	*	6,50000	7,70199	-1,20199	-0,19740	-0,52366	0,526971	1,75964
14	.	.	*	6,61000	7,68090	-1,05090	-0,22762	-0,45784	0,677390	3,55430
15	.	.	*	4,32000	7,21221	-2,89222	-0,55763	-1,26002	0,678449	3,56177
16	.	.	*	7,37000	8,17754	-0,80754	0,15237	-0,35181	0,494583	1,43310
17	.	.	*	7,07000	7,98522	-0,91522	0,01827	-0,42487	0,610488	2,69773

Рис. 4.47. Відхилення залишків від нульового значення

У кінці таблиці (рис. 4.48) наведені мінімальне, максимальне середнє значення залишків та медіана. Середнє значення дорівнює нулю, тобто дотримується передумова, що математичне сподівання дорівнює нулю. Медіана не дорівнює середньому значенню, оскільки є зсув ліворуч. Щільність розподілу залишків не є симетричною.

Minimum	*		-	3.78000	5.32641	-0.09495	-1.94665	-1.78401	0.337110	0.14048	-4.9834
Maximum	*		*	18.00000	11.95668	10.50699	2.93195	4.57748	1.060209	10.11273	11.5951
Mean	*	*	*	7.97038	7.97038	0.00000	0.00000	0.00000	0.606016	2.94340	-0.0303
Median	*	*	*	7.37000	8.02533	-0.48882	0.04042	-0.21296	0.575766	2.29071	-0.5133

Рис. 4.48. Мінімальне, максимальне, середнє значення залишків і медіана

Увійдемо в опцію **Residuals vs. Independent var.** та отримаємо діаграму розсіювання (рис. 4.49).

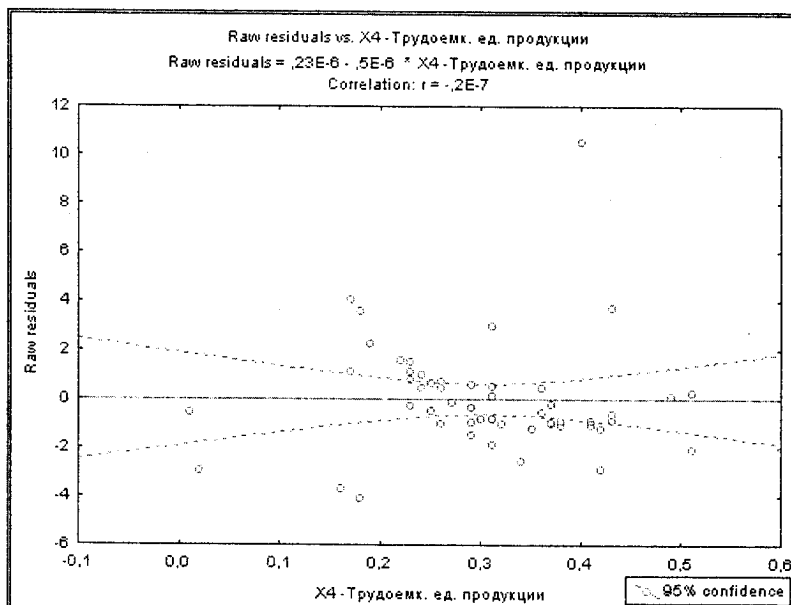


Рис. 4.49. Діаграма розсіювання залишків

Це діаграма розсіювання залишків у залежності від дії незалежної змінної. При достатньо доброму впливі змінної всі значення повинні попасти у довірчий інтервал.

Графік показує діаграму розсіювання залишків при зміні незалежної змінної X_4 . Наведений довірчий інтервал для залишків змінної X_4 . Значення, які вийшли за довірчий інтервал, потребують додаткового аналізу. Аналогічні графіки необхідно побудувати для змінних X_5 , X_6 .

Перейдемо до опції **Histogram of Observed** (рис. 4.50).

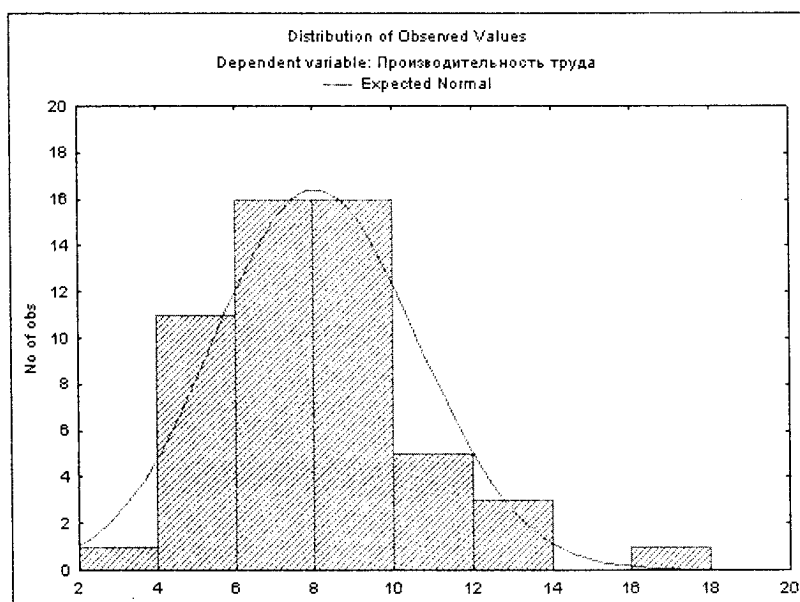


Рис. 4.50. Гістограма

Цей графік призначений для графічної інтерпретації залежної змінної про дотримання припущення нормальності закону розподілу. Аналіз цих даних був проведений вище.

Перейдемо до закладки **Predicted**. Тут можна вивчити вплив незалежних змінних на передбачене значення залежної змінної. Якщо вплив є значущим, то всі значення на графіку повинні лежати у межах довірчого інтервалу.

Отримаємо гістограму передбачених значень за моделлю (див. рис. 4.51).

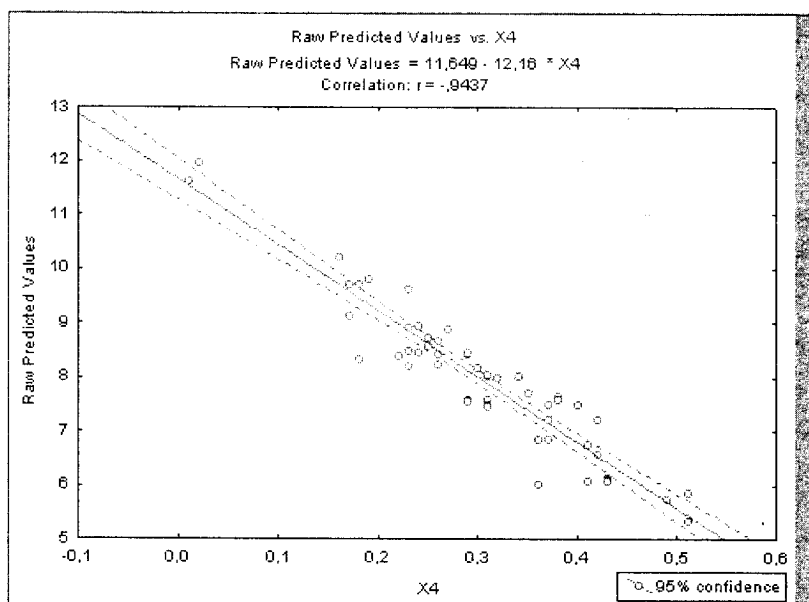


Рис. 4.51. Гістограма передбачених значень

Аналогічні графіки треба побудувати для змінних X5, X6.

Residual Analysis: Predpriyatt ? | _ | x

Dependent: *Продуктивність* Multiple R : ,52087201 F = 6,081439
 R²: ,27131390 df = 3,49
 No. of cases: 33 adjusted R²: ,22670047 p = ,001330
 Standard error of estimate: 2,235065063
 Intercept: 18,680923545 Std. Error: 5,575228 t(49) = 3,3507 p < ,0016

Quick | Advanced | Residuals | Predicted | Scatterplots | Probability plots | Outliers | Save | Summary

☒ Predicted vs. residuals ☒ Observed vs. squared residuals
☒ Predicted vs. squared residuals ☒ Residuals vs. deleted residuals
☒ Predicted vs. observed ☒ Bivariate correlation
☒ Observed vs. residuals ☒ Partial residual plot

Cancel
 Options ▾

Рис. 4.52. Закладка Scatterplots

Увійдемо в опцію **Predicted vs. residuals** (рис. 4.53).

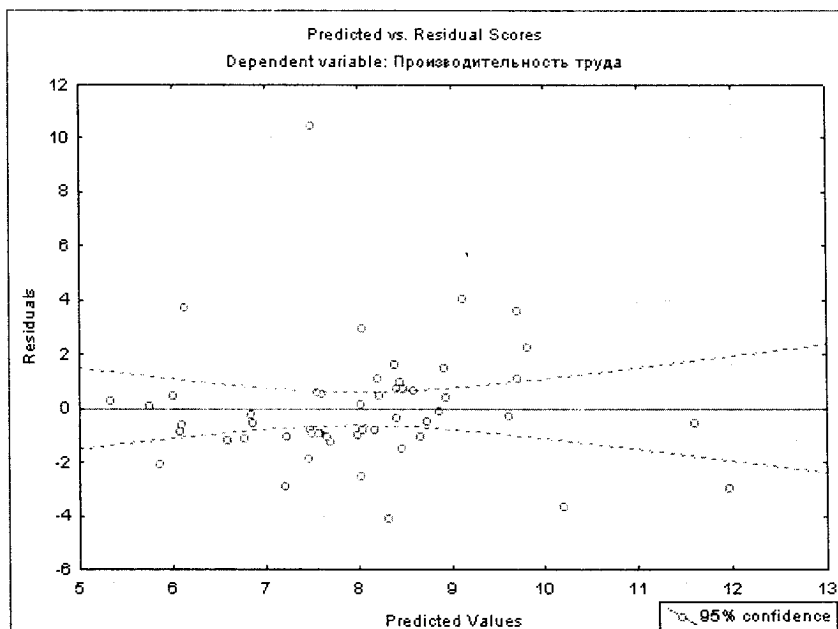


Рис. 4.53. Графік залежності залишків від передбачених значень

Графік показує вплив передбачених значень на залишки. При достатньо добрій апроксимації моделлю вихідних даних залишки повинні лежати в області довірчого інтервалу з імовірністю 95 %. У нашому випадку вплив передбачених значень на залишки має достатньо широкий розкид, на підставі якого можна судити про сумнівну адекватність побудованої моделі.

Перейдемо до опції **Predicted vs. squared residuals** (див. рис. 4.54)

Зі збільшенням діапазону зміни залишків значно збільшується довірчий інтервал, ширина якого значно більша за реальні дані. Це добре видно з графіка.

Увійдемо в опцію **Predicted vs. Observed** (див. рис. 4.55).

Цей графік показує відповідність значень, що спостерігаються, та передбачених. Якщо модель достатньо адекватна, то спостережані та передбачені значення повинні лежати на одній прямій або в області довірчої імовірності. Вихід точок за межі

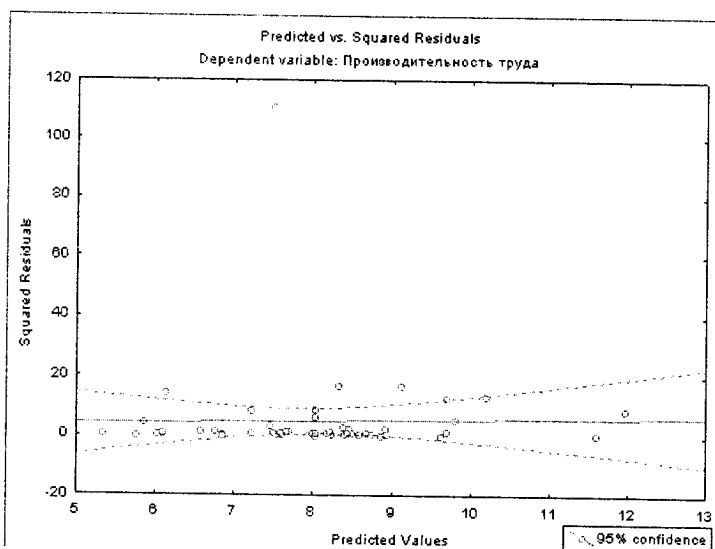


Рис. 4.54. Графік, що відображає вплив передбачених значень від квадрата залишків

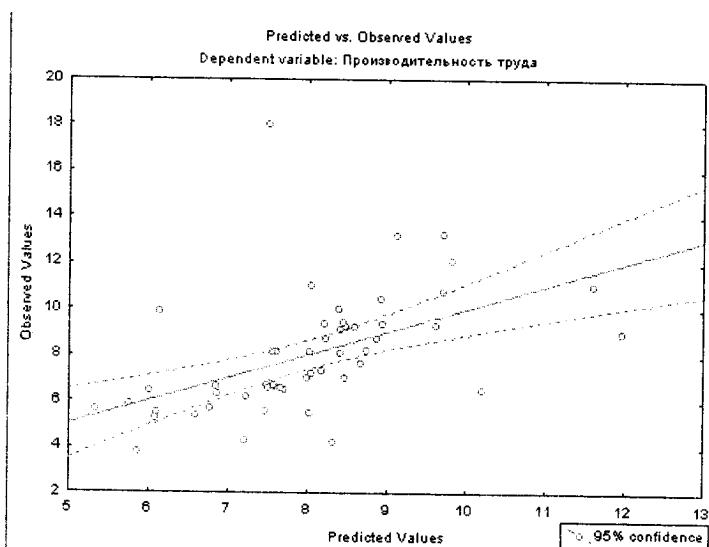


Рис. 4.55. Графік відповідності значень, що спостерігаються, та передбачених

довірчої імовірності говорить про недостатньо адекватну математичну модель. У нашому випадку більше 30 % точок лежать за межами довірчої області. В цьому випадку досліднику необхідно провести глибокий аналіз вихідних даних, можна змінити структуру моделі та включити у модель найбільш інформативні змінні.

Перейдемо до опції **Observed vs. residuals** (рис. 4.56):

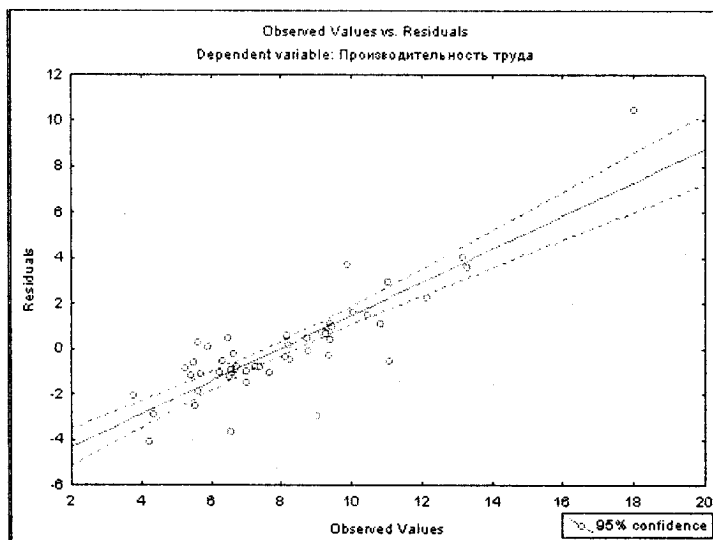


Рис. 4.56. Графік впливу діапазону зміни залежної змінної на зміну залишків

Цей графік показує вплив діапазону зміни залежної змінної на зміну залишків. Оскільки регресійна модель показує передбачення в середньому, то нульові залишки повинні лежати в середній точці значень залежної змінної. Відхилення від середнього значення призводить до збільшення довірчого інтервалу. Цей графік може показати, за якої зміни залежної змінної залишки є неприпустимими.

Перейдемо до опції **Observed vs. Squared Residuals** (див. рис. 4.57).

Аналіз наведеного графіка показує наявність експериментальних точок, для яких залишки виходять за межі допустимих значень.

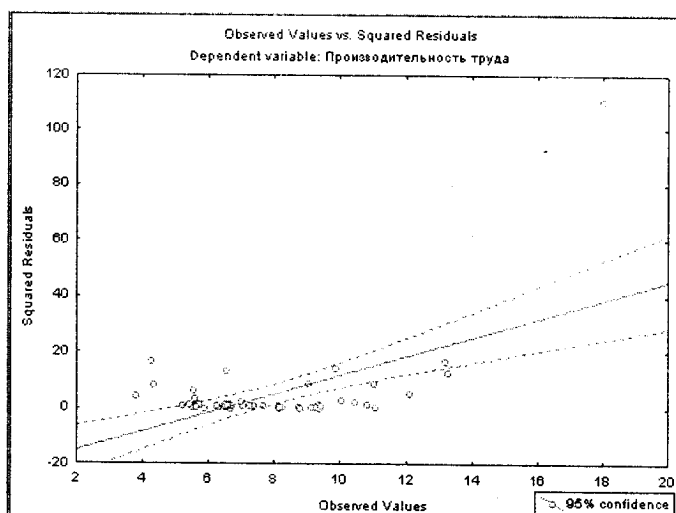


Рис. 4.57. Графік впливу діапазону зміни залежної змінної на зміну квадрата залишків

Перейдемо до опції **Residuals vs. deleted residuals** (див. рис. 4.58).

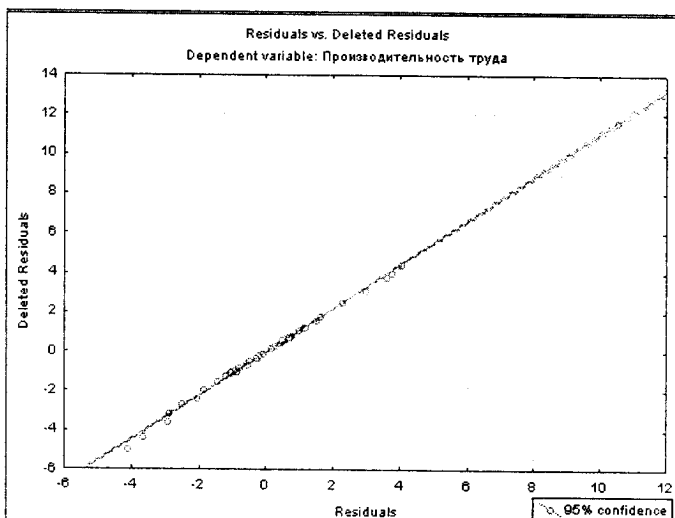


Рис. 4.58. Графік залишків

Переходимо до опції **Bivariate correlation** (рис. 4.59).

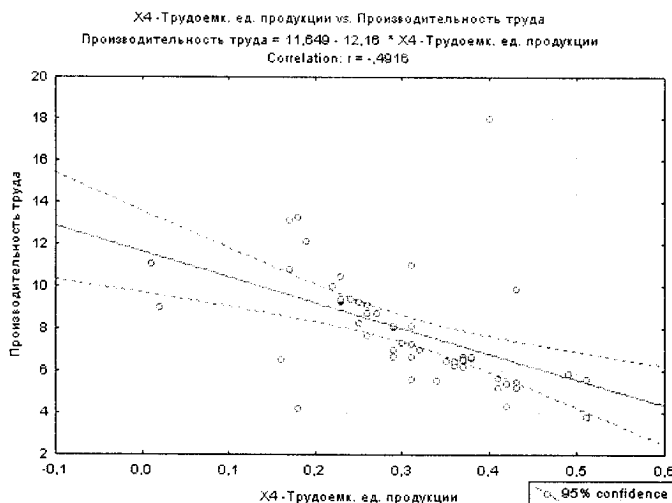


Рис. 4.59. Графік лінійної залежності $Y1 = F(X4)$

Цей графік показує лінійну модель $Y1 = F(X4)$. Більше 20 % точок виходять за межі довірчого інтервалу. Можна судити про недостатню адекватність моделі. Аналогічні моделі та графіки необхідно побудувати для змінних $X5$, $X6$.

Перейдемо до опції **Partial residual plot** (рис. 4.60).

Графік показує додавання залишків до моделі в залежності від зміни $X4$. Тобто відображує чистий вплив змінної $X4$ на передбачені значення залежної змінної $Y1$. Аналогічні графіки необхідно побудувати для змінних $X5$, $X6$.

Перейдемо на закладку **Probability plots** (див. рис. 4.61).

Увійдемо до опції **Normal plot of residuals** (див. рис. 4.62).

Ця опція перевіряє гіпотезу про нормальність розподілу залишків. Якщо залишки розподілені за нормальним законом, то всі дані повинні наближатися до на прямої лінії.

Опції **Half-normal plot** та **Detrended normal plot** використовуються у випадку зрізаної вибірки.

Отже, незважаючи на те, що у відповідності до таблиць регресійного та дисперсійного аналізу можна прийняти гіпотезу про адекватність моделі (задовільне значення F -критерію,

множинного коефіцієнта кореляції, коефіцієнта детермінації), глибокий аналіз доводить некоректність використання цього методу для побудови математичної моделі. Більш того, не дотримуються передумови нормальності закону розподілу залишків, постійності дисперсії та має місце мультиколінеарність. Аналіз гістограми ви-

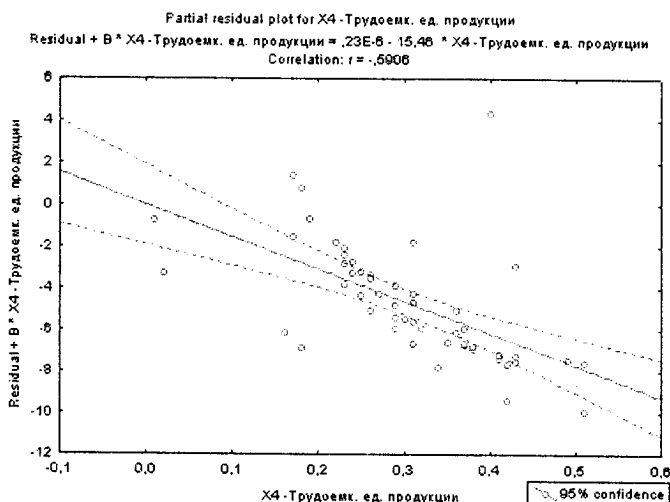


Рис. 4.60. Графік додавання залишків

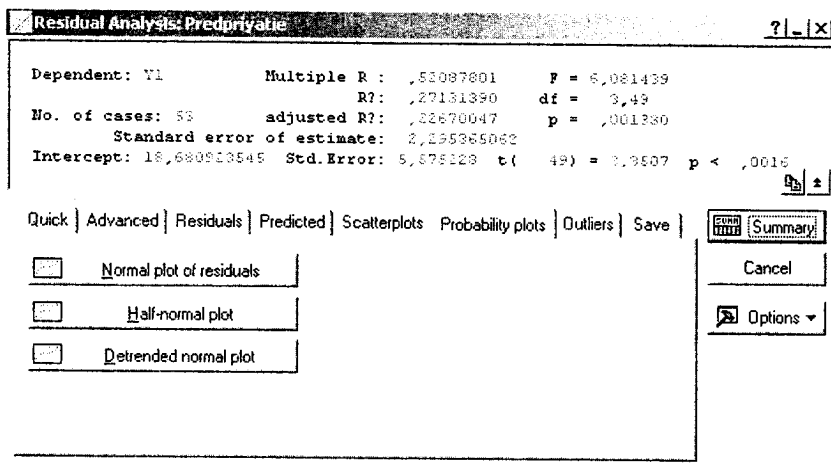


Рис. 4.61. Закладка Probability plots

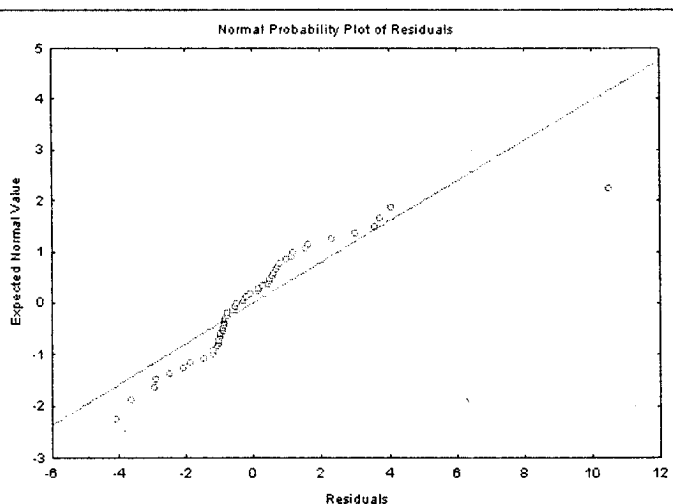


Рис. 4.62. Графік розподілу залишків

хідних даних показав наявність «важких хвостів». У цьому випадку для побудови математичної моделі $Y_1 = F(X_4, X_5, X_6)$ доцільно використовувати робасні або гребеневі методи оцінювання.

Слід мати на увазі, що за наявності кореляції між помилками можна робити припущення про наявність авторегресії першого порядку. Для цього випадку запропоновані методи оцінювання з параметром авторегресії. У роботі [4] наведені фрагменти вибору ефективної процедури оцінювання в залежності від використання матриці перетворення та параметра авторегресії.

Зауважимо, що вхідні та вихідні змінні можуть вимірюватись з випадковими помилками. В цьому випадку необхідно використовувати методи оцінювання, що залежать від статистичних властивостей помилок та характеру істинних значень змінних.

4.6. Приклад побудови моделі курсу валют за фундаментальними факторами з використанням ППП «STATISTICA»

Аналіз фундаментальних факторів проведемо на прикладі встановленого Національним банком України офіційного курсу гривні щодо долара США, який діяв з січня 1996 року до грудня

2001 року (представлено середнє значення за період). Дані було наведено у третьому розділі.

Визначимо, чи є незалежні змінні **VVP**, **DOHOD**, **RASHOD**, **INCENP**, **INCENR**, **DEFBUD** та залежна змінна **CURS** випадковими. Виконаємо перевірку для рівня значущості $\alpha = 0,05$.

Результати розрахунків представлені у таблиці 4.15.

Таблиця 4.15

Таблиця перевірки гіпотези випадковості

Змінні	Кількість серій	$\alpha = 0,975$	$\alpha = 0,025$
CURS	2	27	44
VVP	17	27	44
DOHOD	14	27	44
RASHOD	10	27	44
INCENP	28	27	44
INCENR	27	27	44
DEFBUD	8	27	44

Аналіз таблиці показує, що змінні **VVP**, **DOHOD**, **RASHOD**, **DEFBUD**, **CURS** мають тренд. Змінні **INCENR**, **INCENP** є випадковими.

Структура моделі може бути представлена різними функціями, але у випадку, що розглядається, функції обов'язково мають бути лінійними відносно коефіцієнтів b_j , $j = 1, p$.

При виборі структури моделі необхідно враховувати не лише лінійні, але й ступеневі, й члени взаємодії, тобто подати модель у вигляді полінома Колмогорова — Габора:

$$y = \sum_{j=1}^p b_j x_j + \sum_{l=p+1}^{2p} b_l x_l^2 + \sum_{z=2p+1}^k b_z x_z x_{z+1},$$

де $j = [1, p]$, $l = [p + 1, 2p]$, $z = [2p + 1, k]$; $k = C_p^2$ (число сполучень); b_j — коефіцієнти моделі; y , x_j — залежна й незалежна змінні відповідно.

Обирати попередню структуру моделі можна за двовимірними діаграмами розсіювання. Двовимірні діаграми розсіювання зображають пари значень (x_1, y_1) , (x_2, y_2) , ..., (x_n, y_n) і призначені для оцінки виду залежності між X та Y (лінійна, нелінійна), а також для візуального визначення викидів і грубих помилок.

На наступному етапі ми перевіримо гіпотезу про те, що спостереження розподілені за нормальним законом. Нагадаємо, що в РА незалежні змінні X_j ($j = 1, 2, \dots, p$) розглядаються як не випадкові величини, незалежно від істинного закону розподілу X_j . Вимога щодо нормальності закону розподілу Y необхідна лише для перевірки значущості рівняння регресії та його параметрів β_j , а також для інтервального оцінювання β_j . Для отримання точкових оцінок β_j ($j = 1, 2, \dots, p$) ця умова не потрібна. Зауважимо, що перевірку можна здійснювати на підставі аналізу графічних та аналітичних процедур. Для з'ясування питання про вид закону розподілу використовують подання емпіричного розподілу у вигляді гістограм або графіка функції розподілу, за якими попередньо оцінюють тип розподілу.

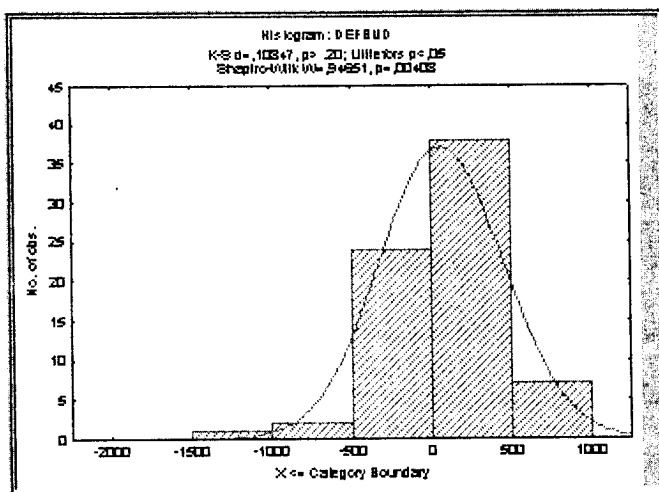
Вибір типу теоретичного розподілу зазвичай здійснюється на основі візуального аналізу. Як правило про його вид судять за вибірковими точковими оцінками числових характеристик. Фактичний розподіл відрізняється від теоретичного через вплив випадкових факторів, який згладжується збільшенням обсягу досліджуваної сукупності. Велике значення має зіставлення фактичних кривих розподілу з теоретичними.

На рис. 4.63 наведені гістограми змінних **VVP**, **DEFBUD**. Візуально проаналізувавши їх, можна прийняти гіпотезу про нормальний закон розподілу. Гістограми для змінних **DOHOD**, **CURS**, **RASHOD**, **INCENP**, **INCENR** показують, що прийняти гіпотезу про нормальний закон розподілу неможна.

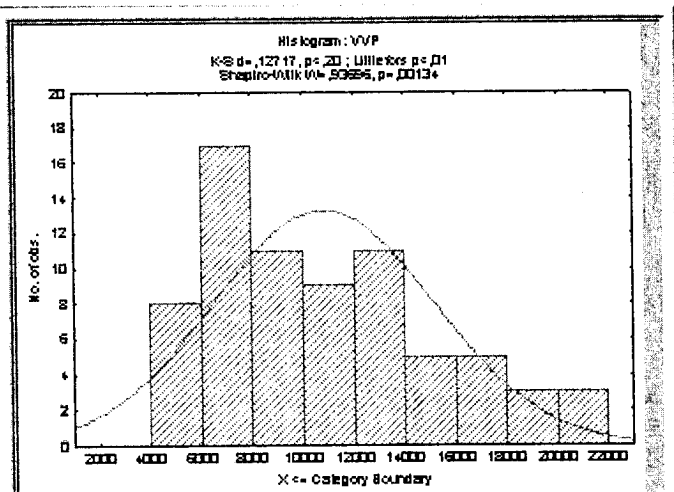
Подальшу перевірку гіпотези проведемо за графіками **Normal probability plots**.

На рис. 4.64, 4.65 наведені графіки розташування даних **RASHOD** та **DEFBUD** на прямій, що відповідає ідеальному нормальному закону розподілу.

Гіпотеза про рівність емпіричного закону розподілу до нормального закону приймається у випадку, якщо вихідні дані лежать на прямій ідеального закону розподілу. Зауважимо, що S -подібне розташування даних на графіках свідчить про необхідність якимось чином перетворити змінні (наприклад, логарифмічне перетворення часто використовується для того, щоб вирівняти кінець розподілу). На цих графіках можна візуально виявити викиди. Аналіз даних **VVP**, **DOHOD**, **RASHOD**, **INCENP**, **INCENR**, **DEFBUD** показує, що лише для змінних **VVP**, **DEFBUD** можна прийняти гіпотезу про нормальний закон розподілу.



a)



б)

Рис. 4.63. Гістограми змінних VVP (а), DEF BUD (б)

Найбільш вірогідний результат перевірки гіпотези про нормальний розподіл залежних та незалежних змінних отримаємо за допомогою існуючих критеріїв згоди емпіричного й теоретичного

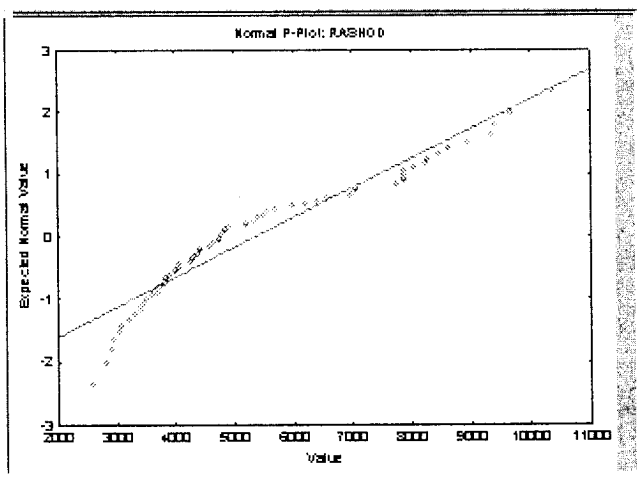


Рис. 4.64. Графік перевірки нормального закону розподілу

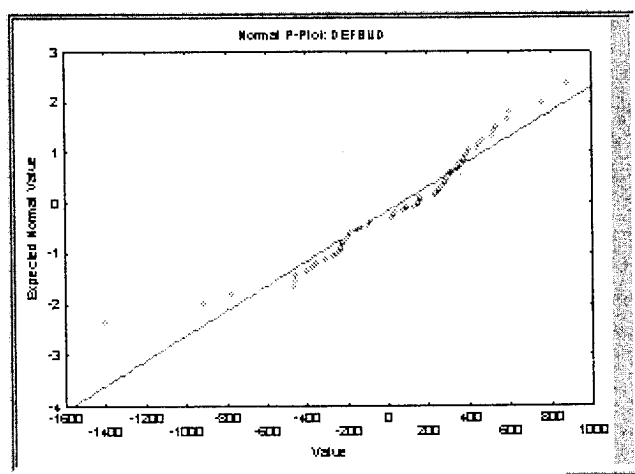


Рис. 4.65. Нормальний імовірнісний графік для змінних RASHOD та DEFBUD

розподілу: Пірсона, Крамера — Мізеса, Колмогорова — Смирнова, Шапіро — Уїлкса.

Для змінних CURS-BALANS отримані результати представлені на рис. 4.66.

Variable	Tests of Normality (Valuta1)					
	N	max D	K-S p	Lilliefors p	W	p
CURS	72	0,244179	p < ,01	p < ,01	0,781675	0,000000
VVP	72	0,127166	p < ,20	p < ,01	0,936964	0,001337
DOHOD	72	0,171818	p < ,05	p < ,01	0,908108	0,000068
RASHOD	72	0,162972	p < ,05	p < ,01	0,913644	0,000113
BEZRAB	72	0,498914	p < ,01	p < ,01	0,453462	0,000000
INCENP	72	0,234741	p < ,01	p < ,01	0,825043	0,000000
INCENR	72	0,178423	p < ,05	p < ,01	0,845436	0,000000
DEFBUD	72	0,108474	p > ,20	p < ,05	0,946513	0,004052
BALANS	22	0,137354	p > ,20	p > ,20	0,934906	0,155223

Рис. 4.66. Вікно результатів перевірки гіпотези нормальності розподілу змінних CURS-BALANS за критеріями згоди Колмогорова — Смирнова (K-S), Лілліфорса, Шаніро — Уїлкса (W)

За таблицею результатів приймаємо гіпотезу про нормальний розподіл лише для змінних VVP, -BALANS та DEFBUD. Побудуємо кореляційну матрицю.

Матрицю коефіцієнтів кореляції, наведену на рис. 4.67, можна використати для перевірки припущення про незалежність вхідних змінних та відсутність мультиколінеарності.

Мультиколінеарність легко помітити за величиною коефіцієнтів кореляції. Коли якийсь з коефіцієнтів кореляції великий (понад 0,85), можна вважати, що між відповідними змінними

Variable	X1_VVP	X2_DOHOD	X3_RASHOD	X4_InCenP	X5_InCenR	X6_DefBud	X7_M0	X8_M1	X9_M2	X10_KrNa	X11_DepNa
X1_VVP	1,00	0,33	0,54	-0,11	-0,35	-0,73	0,91	0,91	0,91	0,77	-0,53
X2_DOHOD	0,33	1,00	0,58	-0,15	-0,27	-0,75	0,96	0,96	0,95	0,71	-0,70
X3_RASHOD	0,54	0,58	1,00	-0,15	-0,26	-0,63	0,97	0,97	0,97	0,71	-0,70
X4_InCenR	-0,11	-0,15	-0,15	1,00	0,57	0,03	-0,16	-0,16	-0,16	0,78	0,85
X5_InCenP	-0,35	-0,27	-0,26	0,57	1,00	0,06	-0,33	-0,37	-0,24	0,60	0,60
X6_DefBud	-0,73	-0,75	-0,63	0,03	0,06	1,00	0,63	-0,71	0,73	0,44	0,43
X7_M0	0,91	0,96	0,97	-0,16	-0,33	0,63	1,00	1,00	0,99	0,73	0,73
X8_M1	0,91	0,96	0,97	-0,16	-0,37	-0,71	1,00	1,00	1,00	0,71	-0,71
X9_M2	0,91	0,95	0,97	-0,16	-0,24	0,73	0,99	1,00	1,00	0,71	0,70
X10_KrNa	-0,57	-0,71	-0,73	0,78	0,60	0,44	-0,73	-0,73	-0,71	1,00	0,95
X11_DepNa	-0,53	-0,70	-0,70	0,85	0,60	0,43	0,73	-0,71	-0,70	0,95	1,00

Рис. 4.67. Кореляційна матриця вихідної інформації

існує лінійний зв'язок і існує імовірність появи мультиколінеарності.

Проведемо порівняльний аналіз коефіцієнтів кореляції, що представлені на рис. 4.67.

Змінні X_1 та X_2 , X_2 та X_3 , X_2 та X_4 і багато інших пар мають значення коефіцієнтів кореляції понад 0,85. Дослідимо мультиколінеарність за числами зумовленості. Більш строго зумовленість матриці визначають різні співвідношення між її власними числами. Цими кількісними характеристиками є числа зумовленості, які обчислюються з використанням різних матричних норм. Використаємо найпростішу характеристику вигляду:

$$P \text{ — число, } r(A) = \frac{\max |l_j|}{\min |l_j|}, \quad j = \overline{1, M},$$

де l_i — власні числа матриці A , 11

У нашому випадку $P = 500$, тобто має місце мультиколінеарність (див. рис. 4.68).

Спираючись на отримані дані, можна з впевненістю говорити про наявність залежностей між вихідними даними, що порушує основні передумови регресійного аналізу. З огляду на це використовувати у подальшому результати регресійного аналізу є недоцільним.

На наступному етапі перевіримо гіпотезу про незалежність вимірювань за допомогою автокореляційної функції. Такі дослідження слід проводити за кожною з досліджуваних змінних.

Отже, попередній аналіз вихідних даних свідчить про наявність істотного лінійного зв'язку між змінними X_1 , X_2 , X_3 , що є причиною порушення вихідної передумови методу РА про повний ранг інформаційної матриці. Можна зробити висновок про некоректність застосування класичного методу РА для розв'язання задачі побудови статистичної моделі прогнозування курсу валют за фундаментальними факторами. У цьому випадку в силу специфіки вихідної інформації пропонується використовувати для побудови моделі прогнозування методи зміщеного оцінювання.

Використовуючи метод гребеневого оцінювання, отримаємо модель прогнозування вигляду $Y = 133,17 + 0,0415X_2 - 0,1805X_6$ або $Y = 0,54X_2 - 0,44X_6$. Аналіз множинного коефіцієнта кореляції

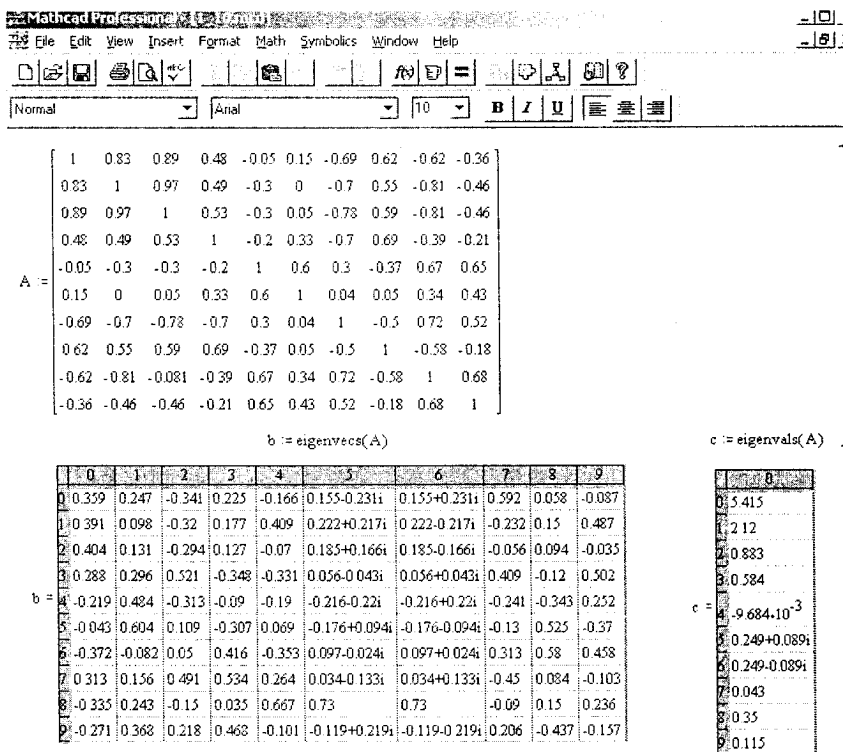


Рис. 4.68. Розрахункові значення власних чисел і векторів вихідної інформації

$R = 0,90086971$, коефіцієнта детермінації $R_0 = 0,81156624$, F — відношення $F(2,68) = 146,43$ показує, що адекватність моделі прогнозування є цілком задовільною.

Регресійний аналіз широко застосовується при розв'язанні багатьох завдань в інформаційно-управляючих системах. Його ефективне використання пов'язане з дотриманням певних передумов і припущень, однак на практиці майже завжди мають місце їх порушення, що призводить до необхідності вживати радикальні заходи для коректного використання регресійного аналізу. Застосування його в умовах лінійної залежності між вхідними векторами, порушення припущення про нормальність закону розподілу, за наявності неоднорідності й корельованості спостережень, помилок у незалежних змінних спричинює певні

труднощі. З огляду на ці обставини стала нагальною потреба у розробці класу проблемно-орієнтованих методів синтезу математичних моделей, нечутливих до порушення передумов. Так, для одержання оптимальних оцінок параметрів моделі при порушенні передумови про повний ранг інформаційної матриці (тобто наявності лінійної залежності між вхідними змінними) розроблений спеціальний клас методів — *методи зміщеного оцінювання*. Для побудови статистичних математичних моделей при порушенні припущень про нормальність залишків використовують робастні методи. Отже, якщо припущення РА порушені, необхідно в першу чергу звернути увагу на організацію одержання вихідних даних. Залежно від специфіки функціонування об'єкта обирають активний або пасивний експеримент. Вихідні дані подають у вигляді таблиці залежних і незалежних змінних, з урахуванням динамічних характеристик. Проводять перетворення даних з попередньою обробкою. Зазначимо, що після перетворення всіх або деяких змінних результати обробки можуть потребувати нової інтерпретації. Зокрема, стандартна помилка оцінки S_e має різну інтерпретацію залежно від того, чи виконувалося перетворення Y .

Перевірка гіпотези про стаціонарність, випадковість, оцінка динамічних характеристик, аналіз законів розподілу вихідних даних дозволяють формалізувати вибір структури моделі.

Аналізуючи дані, можна виявити деякі потенційні проблеми, пов'язані з нелінійністю, нерівною мінливістю й наявністю значень, що різко відхиляються. Незважаючи на те, що гістограми деяких змінних можуть бути сильно скошеними, а деякі діаграми розсіювання — нелінійними, модель лінійної множинної регресії й у цих випадках може бути цілком застосовна. У випадку, коли дані представляють часові ряди, можна застосовувати множинний регресійний аналіз, використовуючи для кожної змінної замість вихідних значень процентні зміни значення цієї змінної між різними періодами часу.

Якщо незалежні змінні є випадковими, будуюмо регресію як умовне математичне сподівання. Перевіряємо передумови й припущення регресійного аналізу. Тут можуть з'явитися проблеми, властиві для множинної регресії.

Якщо математичне сподівання помилки не дорівнює нулю й коваріаційна матриця помилок не діагональна, рекомендуємо використати узагальнений метод найменших квадратів.

Поява нових методів обробки інформації привела до необхідності розробки інтелектуальних систем, що дозволяють в умовах порушення передумов вибрати найбільш ефективний метод побудови моделей, який відповідає специфічним особливостям предметної області.

Контрольні запитання

1. Для чого застосовується регресійний аналіз?
2. Які проблеми називаються статистичними проблемами обробки в регресійному аналізі?
3. Напишіть формулу, за якою обчислюється попередній крок квантування.
4. Як обчислюється вибірковий коефіцієнт кореляції?
5. За якою формулою визначається час згасання кореляційної функції?
6. Які критерії планування експериментів ви знаєте?
7. Що таке довірчий інтервал?
8. Від чого залежить вибір довірчого інтервалу?
9. За якою формулою обчислюється дисперсія помилки?
10. Коли довірчий інтервал має найменшу довжину?
11. За яких умов довжина довірчого інтервалу збільшується?
12. Що називається адекватністю моделі?
13. Напишіть вираз для регресійної моделі.
14. За якою формулою обчислюється математичне сподівання суми квадратів залишків?
15. У чому складається процедура перевірки адекватності моделі?
16. Що таке покрокова регресія?
17. За допомогою алгоритму покрокового регресійного аналізу складіть рівняння регресії з максимальним числом значущих коефіцієнтів регресії за даними табл. 4.7.

РОЗДІЛ 5

ДИСПЕРСІЙНИЙ АНАЛІЗ

5.1. Теоретичні основи дисперсійного аналізу

Ефективність розв'язання функціональних задач в ІУС залежить від якості оцінок стану об'єктів управління. Стан багатьох об'єктів дослідження визначається вхідними та вихідними змінними, або факторами, що не мають кількісного опису. Такими факторами можуть бути керовані, некеровані, контрольовані та неконтрольовані змінні в організаційно-економічних системах та системах управління технологічними процесами, які з будь-яких причин не можуть бути виміряні кількісно. Наприклад, необхідно встановити, чи однакові середні розміри деталей, що виготовляються на різних верстатах, тобто оцінити, чи значущий вплив фактора індивідуальності верстата на процес обробки.

Використовуючи радіодеталі з декількох партій треба визначити чи істотно відрізняються параметри деталей від різних партій. Налаштовуючи радіоапаратуру візуальними та акустичними методами декількома операторами треба оцінити вплив на результат налаштування трьох факторів: апаратури, методу й оператора. Якість налаштування є відкликом. У хімічній промисловості треба оцінити вплив різних каталізаторів на швидкість проходження реакції, вплив різних добавок на якість та обсяг продукції.

Таким чином, дисперсійний аналіз (ДА) використовується у задачах визначення впливу тих чи інших факторів на вихідний параметр або відкликом.

За наявності більше двох виборок дисперсійний аналіз використовується з метою з'ясування наявності систематичних або невипадкових розбіжностей між додавками та визначає, чи існує взаємодія впливів додавок за їх комбінації. Дисперсійний аналіз набув достатньо широкого застосування у медицині, соціології та інших предметних галузях.

Загальна задача дисперсійного аналізу включає декілька факторів, кожен з яких має не менш двох рівнів. В основі кожної задачі лежить план експерименту, який зручно подавати у вигляді таблиці (див. табл. 5.1).

Визначимо постановку задачі. Дано:

1) вихідний параметр y або відклик може залежати від p незалежних факторів $x_k = (x_1, x_2, \dots, x_p)$, що не мають кількісного опису;

2) кожний фактор x_k може варіюватися на u_i рівнях, кількість яких для кожного x_k може бути різною $i = 1, 2, q, j, \dots$;

3) повний факторний експеримент складається з $N = u_1, u_2, \dots, u_i, \dots, u_n$ серій незалежних спостережень;

4) кожна y_{jg} серія містить $l = (1, 2, \dots, m)$ спостережень y_{jgl} або дублюючих дослідів, які для кожної вибірки y_{jg} можуть відрізнятися.

Завдання: визначити, наскільки істотним на фоні випадкових похибок є вплив того чи іншого фактора x_k або комбінації впливу таких факторів на вихідний параметр y , порівняти x_k з іншими факторами та виділити найбільш суттєві.

Припущення:

☑ величина y — є нормально розподіленою випадковою величиною;

☑ величина y — є стаціонарною випадковою величиною;

☑ дисперсія одиничного спостереження σ^2 , що зумовлена випадковими помилками, постійна в усіх дослідях та не залежить від $x_k = (x_1, x_2, \dots, x_p)$.

Кожне з цих припущень необхідно перевірити за результатами спостережень.

У деяких випадках припущення про нормальність розподілів у при дії факторів $x_k = (x_1, x_2, \dots, x_p)$ не дотримується. За таких умов доцільно використовувати непараметричні методи аналізу, найбільш розробленим з них є ранговий метод. Це може призвести до різних постановок задач дисперсійного аналізу, однак принцип розв'язання залишається незмінним.

Розглянемо декілька підходів до розв'язання задачі дисперсійного аналізу. Нехай фактор буде варіюватися на трьох рівнях. Якщо експериментатора цікавлять вибірки, що відповідають саме цим рівням цього фактора, тоді дисперсійний аналіз буде інтерпретуватися за детермінованою моделлю. Зокрема, при повторенні експерименту розглядатимуться випадкові вибірки з тих же самих рівнів. Фактор відноситься до випадкової моделі, якщо вибірки, що відповідають різним рівням фактора, вибираються випадково з генеральної сукупності. Тому при повторенні експерименту вибірки скоріш за все будуть випадковими.

Отже, техніка проведення дисперсійного аналізу доволі різноманітна.

При порушенні припущення про нормальність розподілів y доцільно використовувати математичні перетворення, наприклад логарифмування. Для перевірки гіпотези про нормальність розподілів y використовуємо методику, наведену у розділах 3 та 4.

З постановки задачі та вказаних припущень очевидно, що чим більший вплив фактора X на Y , тим більшою є розбіжність між середніми арифметичними $\bar{y}_j$ серій паралельних спостережень, зроблених при різних сполученнях рівнів варіювання досліджуваних факторів.

При двох серіях спостережень порівняння середніх та перевірка нуль-гіпотези проводиться за допомогою t -критерію Стьюдента. За наявності більше двох серій незалежних спостережень (виборок) для розрізнення середніх застосовується F -критерій. У цьому випадку необхідно зіставити велике довільне число середніх та на підставі їх аналізу зробити висновок про значущість впливу того чи іншого фактора.

Для оцінки впливу x_k на вихідний параметр y необхідно порівняти вплив кожного фактора та встановити деякий показник цього впливу.

Основна ідея дисперсійного аналізу полягає у розкладенні загальної дисперсії на складові, що залежать від випадкових причин, від кожного з факторів, що розглядаються, та їх взаємодії, а також в оцінці статистичної значущості дисперсій останніх з урахуванням помилки відтворюваності досліду.

Вивчати вплив факторів за їх дисперсіями зручно, оскільки це найпростіша міра розсіювання. Дослідження факторів за їх дисперсіями і називається *дисперсійним аналізом*.

Існує величезне розмаїття технік дисперсійного аналізу, зумовлене великою кількістю способів взаємозв'язку між вибірками. Види аналізу відрізняються один від одного планом, тобто способом збирання даних.

Для застосування дисперсійного аналізу необхідно, щоб дані являли собою набір виборок з однією й тією ж базовою одиницею вимірювання, були впорядковані у відповідності до деякої структури або деяким чином. Необхідно також забезпечити, щоб дані задовольняли таким основним припущенням:

☑ кожна вибірка являє собою випадкову вибірку з генеральної сукупності, для якої буде виконуватись узагальнення;

☑ кожна генеральна сукупність підкоряється нормальному розподілу;

☑ стандартні відхилення цих генеральних сукупностей рівні між собою.

У багатовимірному наборі даних одна або декілька змінних (основні вимірювані величини) є кількісними, а решта змінних є якісними. Якісні змінні у сукупності визначають, яким чином кількісні спостереження групуються у вибірки.

5.2. Однофакторний дисперсійний аналіз

Розглянемо ідею дисперсійного аналізу на прикладі вивчення впливу одного фактора X на відклик Y на фоні випадкових похибок (σ_0^2). Подамо результати експерименту у вигляді табл. 5.1, яка має $(u \times m)$ спостережень y_{jl} , де j — порядковий номер рівня варіювання фактора; l — порядковий номер дублюючого досліду в серії на кожному j -му рівні, $l = \overline{1, m}$.

Таблиця 5.1

Результати експерименту

№ рівня фактора X	№ дублюючого досліду						$\bar{y}_j = \frac{1}{m} \sum_{l=1}^m y_{jl}$
	1	2	...	l	...	m	
1	y_{11}	y_{12}	...	y_{1l}	...	y_{1m}	$\bar{y}_1$
2	y_{21}	y_{22}	...	y_{2l}	...	y_{2m}	$\bar{y}_2$
...	...	...	...	...	...	...	...
j	y_{j1}	y_{j2}	...	y_{jl}	...	y_{jm}	$\bar{y}_j$

Закінчення табл. 5.1

№ рівня фактора X	№ дублюючого досліджу						$\bar{y}_j = \frac{1}{m} \sum_{l=1}^m y_{jl}$
	1	2	...	l	...	m	
...	...	...	...	...	...	...	...
u	y_{u1}	y_{u2}	...	y_{ul}	...	y_{um}	$\bar{y}_u$
							$\bar{y} = \frac{1}{u \cdot m} \sum_{j=1}^u \sum_{l=1}^m y_{jl} = \frac{1}{u} \sum_{j=1}^u \bar{y}_j$

Розглянемо випадок, коли $m_1 = m_2 = \dots = m$.

Розсіювання між стовпчиками визначатиметься помилкою відтворюваності, розсіювання між рядками — додатковою дією фактора, що вивчається.

Нехай за відсутності помилок досліджу при варіюванні фактора X на різних рівнях отримані істинні значення $y_1, \dots, y_j, \dots, y_u$, вихідного параметра. Тоді у якості показника впливу фактора X приймемо величину, що називається *дисперсією фактора X* , тобто:

$$\sigma_{jx}^2 = \frac{1}{m-1} \sum_{l=1}^m (y_{jl} - \bar{y}_j)^2, \quad (5.1)$$

$$\text{де } \bar{y}_j = \frac{1}{m} \sum_{l=1}^m y_{jl}.$$

У результаті спостережень при варіюванні фактора X на U рівнях отримаємо $y_1, y_2, \dots, y_u$, розсіювання яких можна характеризувати вибірковою дисперсією

$$\sigma_l^2 = \frac{1}{u-1} \sum_{j=1}^u (y_{lj} - \bar{y}_l)^2, \quad (5.2)$$

$$\text{де } \bar{y}_l = \frac{1}{u} \sum_{j=1}^u y_{lj}.$$

Величину MS_R називатимемо *внутрішньовибірковою дисперсією* (у межах кожної вибірки) або *дисперсією відтворюваності* — σ_0^2 .

Вибіркова дисперсія розсіювання «всередині серії» або залишкова оцінка дисперсії відтворюваності MS_R знаходиться як середнє з вибірових дисперсій за кожною серією окремо

$$MS_R = \frac{1}{u} \sum_{j=1}^u \frac{1}{m-1} \sum_{l=1}^m (y_{jl} - \bar{y}_j)^2 = \frac{SS_R}{u(m-1)} \approx \sigma_0^2. \quad (5.3)$$

Чим більша величина σ_0^2 , тим складніше оцінити вплив фактора X на Y .

Коли дисперсія відтворюваності σ_0^2 не відома, схема дисперсійного аналізу має дозволити знайти її оцінку поряд з оцінками дисперсій факторів, що вивчаються. З цією метою необхідно планувати проведення серій дублюючих дослідів при кожному з усіх можливих сполучень рівнів факторів, що вивчаються.

Вибіркова дисперсія розсіювання між середніми серій слугить незміщеною оцінкою дисперсії $\frac{MS_B}{m}$, з якою нормально розподілені незалежні одне від одного середні серій:

$$MS_B(\bar{y}) = \frac{1}{u-1} \sum_{j=1}^u (\bar{y}_j - \bar{y})^2 = \frac{SS_B}{m(u-1)} \approx \frac{\sigma_x^2}{m} \quad (5.4)$$

з числом степенів вільності $v_x = u - 1$. Величину (5.4) називатимемо *міжвибірковою дисперсією* («між вибірками»).

Загальна дисперсія визначається з виразу

$$\sigma^2 = \frac{1}{um-1} \sum_{j=1}^u \sum_{l=1}^m (y_{jl} - \bar{y})^2 \approx MS_R.$$

Якщо відмінність σ^2 від σ_0^2 (дисперсія відтворюваності) незначуща, то розкид спостережень, який вона характеризує, пов'язаний лише з випадковими причинами і вплив фактора є незначущим. Якщо ж відмінність σ^2 від σ_0^2 значуща, то підвищений розкид спостережень спричинюється не лише випадковими причинами, але ще й впливом фактора X , який тепер треба визнати значущим.

Оскільки в останньому випадку поєднується вплив двох незалежних факторів — випадкових причин (з дисперсією σ_0^2) та фактора X (з дисперсією σ_X^2), що призводить до загального розсіювання спостережень, то загальна дисперсія буде сумою ($\sigma_0^2 + \sigma_X^2$), її оцінка $\sigma^2 \approx \sigma_0^2 + \sigma_X^2$, а дисперсія фактора визначатиметься виразом: $\sigma_X^2 \approx \sigma^2 - \sigma_0^2$.

Якщо розбіжність спостережуваних у досліді значень для різних рівнів пояснюється дією випадковості, тоді всі дані належать

одному й тому самому розподілу. У цьому випадку перевіряються гіпотези для однофакторного дисперсійного аналізу.

Нульова гіпотеза для F -тесту в однофакторному дисперсійному аналізі припускає, що U вибірки мають одне й те саме середнє значення. Альтернативна гіпотеза припускає, що не всі середні значення рівні між собою, тобто принаймні у двох сукупностей середні розрізняються.

Нульова гіпотеза: $H_0: \bar{y}_1 = \bar{y}_2 = \dots = \bar{y}_u$ (усі середні між собою рівні).

Альтернативна гіпотеза: $H_1: \bar{y}_1 \neq \bar{y}_u$ (принаймні для однієї пар генеральних сукупностей (не всі середні між собою рівні)).

Оскільки припускається, що стандартні відхилення всіх сукупностей однакові, нульова гіпотеза фактично зводиться до того, що вибіркові сукупності мають однакові розподіли. Альтернативна гіпотеза стверджує, що існують розбіжності без урахування кількості вибірових сукупностей, які фактично відрізняються між собою.

Розсіювання окремих спостережень відносно загального середнього значення зумовлене дією випадкових причин та впливом фактора X . Дія фактора випадковості виявляється у розсіюванні спостережень серій дублюючих дослідів на кожному рівні X навколо середнього арифметичного своєї серії. Вплив же фактора X (з дисперсією σ_x^2) спричиняє підвищене розсіювання середніх значень у серіях відносно загального середнього.

Розкладемо суму квадратів відхилень спостережень від загального середнього на дві складові суми, одна з яких буде характеризувати вплив фактора випадковості, а інша — фактора змінюваності X . Суму квадратів відхилень спостережень від загального середнього значення подамо у вигляді

$$\begin{aligned} S &= \sum_{j=1}^u \sum_{l=1}^m (y_{jl} - \bar{y})^2 = \sum_{j=1}^u \sum_{l=1}^m (y_{jl} - \bar{y}_j + \bar{y}_j - \bar{y})^2 = \sum_{j=1}^u \sum_{l=1}^m (y_{jl} - \bar{y}_j)^2 + \\ &+ 2 \sum_{j=1}^u \sum_{l=1}^m (y_{jl} - \bar{y}_j)(\bar{y}_j - \bar{y}) + \sum_{j=1}^u \sum_{l=1}^m (\bar{y}_j - \bar{y})^2 = \sum_{j=1}^u \sum_{l=1}^m (y_{jl} - \bar{y}_j)^2 + \\ &+ m \sum_{j=1}^u (\bar{y}_j - \bar{y})^2 = SS_R + SS_B, \quad \sum_{l=1}^m (y_{jl} - \bar{y}_j) = 0. \end{aligned}$$

Таким чином, сума квадратів відхилень, що характеризує розсіювання спостережень у результаті дії обох факторів — як

випадковості (з дисперсією MS_R), так і досліджуваного X (з дисперсією MS_B), подається у вигляді:

$$S = \sum_{j=1}^u \sum_{l=1}^m (y_{jl} - \bar{y})^2 = SS_B + SS_R.$$

Сума квадратів відхилень «усередині серії», що характеризує залишкове розсіювання випадкових похибок у межах дослідів, тобто їх відтворюваність (з дисперсією MS_R):

$$SS_R = \sum_{j=1}^u \sum_{l=1}^m (y_{jl} - \bar{y}_j)^2.$$

Сума квадратів відхилень «між серіями»

$$SS_B = m \sum_{j=1}^u (\bar{y}_j - \bar{y})^2.$$

Сума $\frac{SS_B}{m}$ характеризує розсіювання середніх $\bar{y}_j$ серій за рахунок випадкових причин (з дисперсією для $\frac{MS_R}{m}$ середніх серій) та досліджуваного фактора (з дисперсією MS_B).

Припустимо, що вплив X на Y відсутній, тобто нуль-гіпотеза про однорідність $\bar{y}_1, \bar{y}_2, \dots, \bar{y}_j, \dots, \bar{y}_u$ є правильною. Тоді всі серії паралельних спостережень можна розглядати як випадкові вибірки з однієї й тієї ж нормальної сукупності й, отже: S_0 та S_X незалежні одне від одного. За відсутності впливу фактора X вибіркові оцінки MS_R, MS_B однорідні, оскільки вони є оцінками однієї й тієї ж генеральної дисперсії $MS_T = \sigma^2$.

Припустимо тепер, що вплив фактора X на вихідний параметр, являється істотним, тобто нуль-гіпотеза про однорідність $\bar{y}_1, \bar{y}_2, \dots, \bar{y}_j, \dots, \bar{y}_u$ хибна. Тоді серії паралельних спостережень можна розглядати як випадкові вибірки з однією й тією ж дисперсією σ_0^2 та різними центрами розподілу $C_1, \dots, C_j, \dots, C_u$ і, отже:

1) вибіркова дисперсія σ^2 характеризує вплив як фактора випадковості, так і фактора X , тобто $\sigma^2 = \sigma_0^2 + \sigma_X^2$;

2) оскільки SS_R не змінюється при заміні y_{jl} на $(y_{jl} - C_j)$, то вибіркова дисперсія MS_R також не змінюється і все ще є незміщеною оцінкою для σ^2 , тобто $MS_R \approx \sigma^2$;

3) оскільки сума SS_B ураховує не лише випадкові, але й систематичні розбіжності між середніми серій і збільшується за рахунок впливу фактора X , то дисперсія $MS_B[\bar{y}_j]$ при цьому також збільшується і вже не є оцінкою тільки для $\frac{\sigma_0^2}{m}$, тобто

$$MS_B[\bar{y}_j] = \frac{\sigma_0^2}{m} + \sigma_x^2, \text{ звідки } \sigma^2 + m\sigma_x^2 = MS_R;$$

4) незалежність SS_R та SS_B одне від одного зберігається.

Таким чином, для дисперсії фактора X можна надати дві приблизні оцінки:

$$\checkmark \quad \sigma_x^2 \approx \sigma^2 - \sigma_0^2;$$

$$\checkmark \quad \sigma_x^2 \approx \frac{1}{m}(S_x^2 - \sigma_0^2) \approx \frac{1}{m}(MS_B - MS_R).$$

Перша оцінка менш точна через похибки величин MS_B та MS_R . Точність другої оцінки вища, оскільки дисперсії входять до неї, розділені на m . З цього випливає, що за наявності впливу фактора X вибіркові оцінки дисперсій є неоднорідними. Порівнюючи ці вибіркові дисперсії, можна прийняти рішення про справедливість першого або другого припущення щодо значущості впливу фактора X (з дисперсією σ_x^2) на вихідний параметр.

Оцінка впливу фактора

Для того, щоб вплив фактора X був значущим ($\sigma_x^2 > 0$), необхідно й достатньо, щоб дисперсія S_x^2 значуще відрізнялась від S_0^2 . Перевірку нуль-гіпотези про однорідність цих вибіркових оцінок можна здійснити за допомогою F -критерію:

$$F = \frac{MS_B}{MS_R}.$$

F -статистика для однофакторного дисперсійного аналізу являє собою співвідношення значень варіації з двох джерел: міжгрупову варіацію ділять на внутрішньогрупову. Можна вважати, що F -статистика показує, у скільки разів вибіркові середні більш змінювані у порівнянні з тим, що можна було б очікувати, якщо б розбіжність була випадковою. Для виконання F -тесту необхідно обчислити F -статистику та порівняти отримане значення з F табличним.

F -таблиця містить критичні значення для розподілу F -статистики за умови справедливості нульової гіпотези, тобто при істинній

нульовій гіпотезі значення F -статистики перевищує значення з F -таблиці в контрольованому відсотку випадків (наприклад, 5 %). Щоб знайти критичне значення у таблиці, необхідно використовувати кількість степенів вільності для визначення рядка F -таблиці та відповідний рівень перевірки (наприклад, 5 %) для визначення необхідного стовпчика.

Якщо обчислене за результатами спостережень дисперсійне відношення F перевищує табличне $F_{\alpha}(v_X, v_0)$ за розподілом Фішера для обраного рівня значущості α за відповідних степенів вільності v_X та v_0 , то вплив фактора X визнається значущим ($\sigma_x^2 > 0$). У цьому випадку необхідно відкинути нульову гіпотезу H_0 і прийняти альтернативну гіпотезу H_1 . За таких умов середні значення виборок значуще відрізняються одне від одного і розходження у значеннях вибірових середніх, що спостерігається, не можна прийнятно пояснити лише випадковістю.

Якщо значення F -статистики менше критичного значення з F -таблиці ($F \leq F_{\alpha}(v_X, v_0)$), то необхідно прийняти нульову гіпотезу H_0 і не приймати альтернативну гіпотезу H_1 . У такому випадку середні виборок незначуще відрізняються одне від одного, а розходження у значеннях вибірових середніх можна прийнятно пояснити тільки випадковістю. Результат не є статистично значущим.

У дисперсійному аналізі перевіряють нуль-гіпотезу $M[S]_B = M[S_R] = MS_T$ за альтернативи $MS_B > MS_R$, тому користуються однобічним F -критерієм. При цьому $\alpha = 0,05$.

Слід мати на увазі, що дисперсійний аналіз дозволяє визначити вплив фактора лише в цілому, не даючи кількісної оцінки цього впливу. Так, наприклад, при зміні діапазону варіювання фактора, що вивчається, оцінка впливу останнього буде змінюватись. F -критерій для перевірки гіпотези $H_0: \bar{y}_1 = \bar{y}_2 = \dots = \bar{y}_u$ є достатньо точним за рівності $\bar{y}_1, \bar{y}_2, \dots, \bar{y}_m$.

Випадок нерівночисельних розрахунків

Вище було розглянуто випадок лише рівночисельних серій спостережень на всіх рівнях фактора X . Ця обставина не є суттєвою для теорії дисперсійного аналізу, а тому для різного числа m_j паралельних спостережень на різних j -х рівнях схема проведення та основні способи аналізу залишаються такими самими. Зазнають змін лише деякі вирази:

☑ загальне число спостережень $M = \sum_{j=1}^u m_j$;

☑ підсумки спостережень за серіями $Y_j = \sum_{l=1}^{m_j} y_{jl}$;

☑ середні серій $\bar{y}_j = \frac{1}{m_j} \sum_{l=1}^{m_j} y_{jl} = \frac{y_j}{m_j}$;

☑ загальне середнє $\bar{y} = \frac{1}{M} \sum_{j=1}^u \sum_{l=1}^{m_j} y_{jl} = \frac{1}{M} \sum_{j=1}^u Y_j$;

☑ співвідношення для сум:

$$S = \sum_{j=1}^u \sum_{l=1}^{m_j} (y_{jl} - \bar{y})^2 = \sum_{j=1}^u \sum_{l=1}^{m_j} (y_{jl} - \bar{y}_j)^2 + \sum_{j=1}^u m_j (\bar{y}_j - \bar{y})^2 = SS_R + SS_B.$$

5.3. Приклад використання однофакторного дисперсійного аналізу

Для прикладу оцінимо якості покупних виробів (компонентів) для виробництва електронної апаратури, що придбані у постачальників № 1, 2, 3. Набір даних запозичений із роботи [2] і наведений у табл. 5.2. Оцінка якості вимірюється за шкалою від 0 до 100 і показує ступінь відповідності технічних специфікацій компонента вимогам фірми. Чим вища оцінка, тим краще. Оцінка 75 і вище достатня для того, щоб компонент можна було застосовувати. У середньому якість продукції постачальника № 3 має найвищу оцінку якості (87,7), наступним є постачальник № 1 (82,1) і, нарешті, постачальник № 2 (80,7).

Зауважимо, що найвища якість продукції постачальника № 3 може бути зумовлена випадковим добром компонентів. Чи збереже цей постачальник провідні позиції, якщо дослідити вибірки більшого розміру? Дисперсійний аналіз дає відповідь на це питання, не потребуючи великих витрат на отримання додаткових даних.

У нашому прикладі кількість виборок дорівнює $U = 3$. Сумарна кількість вимірювань

$$M = m_1 + m_2 + m_3 = 18 + 21 + 19 = 58.$$

Таблиця 5.2

Оцінка якості продукції постачальників № 1, № 2, № 3

Показник	Постачальник		
	№ 1	№ 2	№ 3
Оцінка якості продукції	75	94	90
	.	.	.
	.	.	.
	.	.	.
	73	90	83
Середнє	$\bar{y}_1 = 82,055556$	$\bar{y}_2 = 80,666667$	$\bar{y}_3 = 87,684211$
Стандартне відхилення	$S_1 = 7,124706$	$S_2 = 7,598245$	$S_3 = 5,228688$
Розмір вибірки	$m_1 = 18$	$m_2 = 21$	$m_3 = 19$

Визначимо загальне середнє:

$$Y = m_1 \bar{y}_1 + m_2 \bar{y}_2 + \frac{m_3 \bar{y}_3}{M} = 18 \cdot 82,055 + 21 \cdot 80,66 + \frac{19 \cdot 87,68}{58} =$$

$$= \frac{4837}{58} = 83,39.$$

Оцінимо міжгрупову дисперсію

$$M[S]_B = \frac{18(82,055 - 83,39)^2 + 21(80,66 - 83,39)^2 + 19(87,68 - 83,39)^2}{3 - 1} =$$

$$= \frac{538,163}{2} = 269,08.$$

Міжгрупова дисперсія показує, наскільки є відмінними вибіркві середні. Вона дорівнює нулю, якщо середні є рівними, і тим більша, чим сильніше розрізняються середні. Міжгрупова дисперсія представляє собою основу міри варіації вибірквих середніх.

Формулу для обчислення міжгрупової варіації можна розглядати як результат заміни усіх значень в усіх вибірках відповідними вибірквими середніми, об'єднання результатів цих заміи в один великий набір даних, обчислення звичайного вибірквого стандартного відхилення, піднесення його у квадрат для отримання дисперсії та помноження на масштабний коефіцієнт $(m - l)/(u - l)$.

Внутрішньогрупова дисперсія

$$M[S_R] = \frac{(m_1 - 1)(S_1)^2 + (m_2 - 1)(S_2)^2 + (m_3 - 1)(S_3)^2}{M - u} = \\ = \frac{(18 - 1)(7,12)^2 + (21 - 1)(7,59)^2 + (19 - 1)(5,228)^2}{58 - 3} = 45,63.$$

Внутрішньогрупова дисперсія вимірює, наскільки є неоднорідною кожна вибірка. Усі вибірки мають однакову внутрішньогрупову дисперсію, тобто лише одне значення внутрішньогрупової дисперсії. Це значення повинно дорівнювати нулю, якщо кожна вибірка складається зі своїх багаторазово повторених середніх, і воно буде тим більше, чим сильніше розрізняються числа у вибірці. Квадратний корінь від значення внутрішньогрупової дисперсії дає оцінку стандартних відхилень для генеральних сукупностей.

Число степенів вільності ($M - u$) при визначенні внутрішньогрупової дисперсії $M[S_R]$ відображає той факт, що вимірюють дисперсію усіх M значень даних відносно їх вибірових середніх, але при цьому віднімають u степенів вільності, оскільки відомі оцінки u різних вибірових середніх.

Число степенів вільності ($u - 1$) при визначенні міжгрупової дисперсії $M[S]_B$ відображає той факт, що вимірюють варіацію u середніх. Один степінь вільності при цьому віднімають (як і при звичайному стандартному відхиленні), оскільки для загального середнього оцінка відома.

Визначимо F -статистику у вигляді:

$$F = \frac{269,08}{45,63} = 5,89.$$

Далі визначаємо табличне значення F -статистики за F -таблицею, яка містить критичні значення для розподілу F -статистики за умови справедливості нульової гіпотези, тобто за істинної нульової гіпотези значення F -статистики перевищує значення з F -таблиці у контрольованому відсотку випадків (наприклад, 5 %). Щоб знайти критичне значення у таблиці, необхідно використовувати кількість степенів вільності для визначення рядка F -таблиці та відповідний рівень перевірки (наприклад, 5 %) для визначення необхідного стовпця. Критичне значення з F -таблиці для перевірки на звичайному рівні (5 %) знаходиться між 3,316

та 3,150 (для числа степенів вільності всередині груп, рівних 30 та 60 відповідно, оскільки шукане значення для числа степенів вільності 55 у таблиці відсутнє). Наприклад, на рівні 1 % критичне значення знаходиться між 5,390 та 4,977. Приблизне значення статистики можна отримати шляхом інтерполяції з використанням цих двох значень кількості степенів вільності, в той час як обчислення на комп'ютері дасть точне значення. Оскільки 5,89 перевищує будь-яке число, що знаходиться між 3,316 та 3,150, то мають місце істотні відмінності між вибірковими середніми.

Отже, постачальники значуще розрізняються з точки зору рівня якості компонентів ($p < 0,05$), що постачаються ними.

Зауважимо, що F -тест встановлює лише наявність або відсутність відмінностей. Однак часто потрібно визначити вибіркові середні, які істотно відрізняються від інших. Існує низка різних методів рішення цієї проблеми. Представлений тут метод перевірки найменшої значущої відмінності оснований на t -тесті для середньої різниці між парами виборок [2].

Якщо F -тест незначущий, не треба перевіряти відмінності окремих виборок між собою. F -тест вже встановив відсутність значущих відмінностей. Якщо F -тест значущий, то можна перевіряти відмінності для окремих виборок, щоб встановити, які саме пари виборок розрізняються між собою.

При проведенні t -тесту для прийняття рішення про відмінність двох конкретних виборок необхідно визначити різницю середніх цих виборок, оцінити стандартну помилку різниці середніх, оцінити довірчий інтервал різниці середніх для генеральних сукупностей, щоб визначити, чи містить цей інтервал задане значення 0 (що означає відсутність відмінностей).

Стандартну помилку обчислюють таким чином.

Стандартна помилка =

$$= \sqrt{(\text{внутрішньовибіркова змінність}) \left(\frac{1}{n_i} + \frac{1}{n_j} \right)},$$

де m_i та m_j — розміри двох виборок, які порівнюються. Величина стандартної помилки залежить від розмірів виборок.

Оцінимо наявність значущих відмінностей серед оцінок якості продукції між парами постачальників 1 і 2, 1 і 3 та 2 і 3.

Доцільність такого порівняння зумовлена наявністю значущих відмінностей, виявлену в результаті F -тестування.

Для порівняння постачальників 1 і 2 визначимо приблизне значення t з таблиці t -розподілу для $M - u = 58 - 3 = 55$ степенів вільності, яке дорівнює 1,960.

Різниця середніх $(\bar{y}_1, \bar{y}_2) = 80,667 - 82,056 = -1,389$.

$$\mu_{\sigma 0} = \sqrt{45,63 \left(\frac{1}{21} + \frac{1}{18} \right)} = 2,170.$$

Визначимо 95 %-й довірчий інтервал для різниці середніх генеральних сукупностей: $-1,389 - 1,960 \cdot 2,170 \approx -5,64$; $-1,389 + 1,960 \cdot 2,170 \approx 2,86$.

Значення t -статистики дорівнює:

$$t = \frac{1,389}{2,170} = -0,640.$$

Виходячи з 95 %-го довірчого інтервалу (від $-5,64$ до $2,86$) видно, що він включає в себе нуль. У цьому випадку можна прийняти нульову гіпотезу про відсутність відмінностей у генеральній сукупності між середніми оцінками якості виробів у постачальників 1 і 2. Цю гіпотезу підтверджує й t -тест, з якого видно, що значення t -статистики ($-0,640$) менше за абсолютним значенням, ніж табличне значення 1,960. Отже, постачальник 2 має нижчу оцінку якості виробів, ніж постачальник 1.

Виконаємо обчислення для порівняння середніх оцінок якості між постачальниками 1 і 3. Різниця середніх складає $y_3 - \bar{y}_1 = 87,684 - 82,056 = 5,628$.

$$\text{Стандартна помилка} = \sqrt{(45,63) \left(\frac{1}{19} + \frac{1}{18} \right)} = 2,222.$$

95 %-й довірчий інтервал для різниці середніх генеральних сукупностей знаходиться між

$$5,628 - 1,960 \cdot 2,222 = 1,27 \quad \text{та} \quad 5,628 + 1,960 \cdot 2,222 = 9,98,$$

$$t = \frac{5,628}{2,222} = 2,533.$$

Виходячи з 95 %-го довірчого інтервалу (від 1,27 до 9,98) видно, що він не включає в себе нуль. У цьому випадку можна прийняти альтернативну гіпотезу про наявність відмінностей у генеральній сукупності між середніми оцінками якості виробів у постачальників 1 і 3. Цю гіпотезу підтверджує й t -тест, з якого видно, що значення t -статистики (2,533) більше за абсолютним значенням, ніж табличне значення 1,960. Отже, постачальник 3 має більш високу оцінку якості, ніж постачальник 1.

Виконаємо обчислення для порівняння середніх оцінок якості виробів між постачальниками 2 і 3. Різниця середніх складає $y_3 - \bar{y}_2 = 87,684 - 80,667 = 7,017$.

$$\text{Стандартна помилка} = \sqrt{(45,63) \left(\frac{1}{19} + \frac{1}{21} \right)} = 2,139.$$

95 %-й довірчий інтервал для різниці середніх генеральних сукупностей знаходиться між

$$7,017 - 1,960 \cdot 2,139 = 2,82 \quad \text{та} \quad 7,017 + 1,960 \cdot 2,139 = 11,21,$$

$$t = \frac{7,017}{2,139} = 3,281.$$

Виходячи з 95 %-го довірчого інтервалу (від 2,82 до 11,21) видно, що він не включає в себе нуль. У цьому випадку можна прийняти альтернативну гіпотезу про наявність відмінностей у генеральній сукупності між середніми оцінками якості виробів постачальників 2 і 3. Цю гіпотезу підтверджує t -тест, з якого видно, що значення t -статистики (2,139) більше за абсолютною величиною, ніж табличне значення 1,960. Отже, постачальник 3 має більш високу оцінку якості, ніж постачальник 2.

На підставі результатів дисперсійного аналізу якості продукції трьох постачальників можливо зробити такі висновки.

☑ Існують значущі відмінності якості продукції постачальників. F -тест свідчить, що середні значення оцінок якості продукції цих трьох постачальників неоднакові.

☑ Постачальник 3 має істотно більш високу якість виробів у порівнянні з кожним з двох інших постачальників (на підставі тесту найменшої значущої відмінності).

Два інших постачальника, 1 і 2, не мають значущих відмінностей з точки зору середнього рівня якості своєї продукції.

Після виконання однофакторного аналізу необхідно перевірити коректність застосування використаного методу дисперсійного аналізу, тобто перевірити чи є величина у нормально розподіленою випадковою величиною, оцінити постійність в усіх дослідках дисперсії одиничного спостереження.

Якщо припущення про нормальність залишків викликає сумнів, необхідно використовувати рангові або знакові процедури аналізу даних.

5.4. Рангові методи однофакторного дисперсійного аналізу

Якщо нічого не відомо про розподіл спостережень, то безпосередньо використовувати для перевірки нульової гіпотези кількісні значення спостережень x_{jl} стає складно. У цьому випадку простіше за все спиратися у висновках лише на співвідношення «більше-менше» між спостереженнями, оскільки вони не залежать від розподілу спостережень. При цьому вся інформація з табл. 5.2, яку ми використовуємо, міститься у тих рангах, які отримують числа x_{jl} при впорядкованості усієї їх сукупності. Відповідні критерії для перевірки нульової гіпотези називаються ранговими, вони придатні для будь-яких неперервних розподілів спостережень. Тут конкретні числові значення величин x_{jl} взагалі є умовністю, а змістовний сенс мають лише відношення «більше-менше» між ними.

Коли серед чисел x_{jl} немає однакових, процедура призначення рангів досить проста. За наявності збігів надійність методу знижується.

Побудуємо таблицю рангів за даними табл. 5.2. Для цього впорядкуємо величини x_{jl} (все одно як — від більшого до меншого, або від меншого до більшого). Позначимо через r_{jl} ранг числа x_{jl} в усій сукупності. Результати подамо у табл. 5.3.

Якщо спостережуване у досліді розташування рангів, наведене у табл. 5.3, є випадковим і всі реалізації табл. 5.3 рівноймовірні, тоді можна прийняти статистичну гіпотезу H_0 про однорідність усіх k представлених виборок. В основі статистичного критерію, за допомогою якого можна було б судити про справедливість висунутої гіпотези, лежить рангова статистика r_{ij} . Основна вимога до цієї статистики така: її розподіл за

дотримання гіпотези H_0 має помітно відрізнятись від її розподілу за альтернативних гіпотез.

Таблиця 5.3

Ранги результатів вимірювань

Обробки	1	2	...	k
Ранги результатів вимірювань	r_{11}	r_{12}	...	r_{1k}
	r_{21}	r_{22}	...	r_{2k}
	$\vdots$	$\vdots$		$\vdots$
	r_{n1}	r_{n2}	...	r_{nk}

Розглянемо непараметричні критерії перевірки однорідності Краскала — Уолліса (довільні альтернативи) і Джонкхієра (альтернативи з упорядкуванням) [8].

Якщо не можна сказати будь-що певне про альтернативи до H_0 , можна скористатися для її перевірки вільним від розподілу критерієм Краскала — Уолліса. Для цього спостереження x_{ij} замінюються їх рангами r_{ij} , впорядковуючи всю сукупність $\|x_{ij}\|$ за зростанням (для визначеності). Потім для кожної обробки j (тобто для кожного стовпця вихідної таблиці) обчислюються

$$R_j = \sum_{i=1}^{n_i} r_{ij} \quad \text{та} \quad R_j = \frac{1}{n_j} \sum_{i=1}^{n_i} r_{ij},$$

де R_{ij} — це середній ранг, розрахований за стовпцем. Якщо між стовпчиками немає систематичних відмінностей, середні ранги R_{ij} , $j = 1, \dots, k$ не повинні значно відрізнятись від середнього рангу, розрахованого за всією сукупністю $\|r_{ij}\|$. Ясно, що останній до-

рівнює $\frac{(N+1)}{2}$, тому величини $\left(R_1 - \frac{N+1}{2}\right)^2, \dots, \left(R_k - \frac{N+1}{2}\right)^2$ при

H_0 у сукупності мають бути невеликими. Складаючи загальну характеристику, доцільно врахувати відмінності у числі спостережень для різних обробок та взяти за міру відступу від чистої випадковості величину

$$H = \frac{12}{N(N+1)} \sum_{j=1}^k n_j \left(R_j - \frac{N+1}{2}\right)^2. \quad (5.5)$$

Ця величина називається статистикою Краскала — Уолліса. Множник $\frac{12}{[N(N+1)]}$ потрібен для стабілізації її розподілу за великої кількості спостережень. Інша форма для обчислення H :

$$H = \frac{12}{N(N+1)} \sum_{j=1}^k \frac{R_j^2}{n_j} - 3(N+1). \quad (5.6)$$

Таблиці та асимптотика

Невеликі таблиці розподілу статистики H для гіпотези H_0 можна знайти у збірках статистичних таблиць. За великих обсягів $n_i, \dots, n_k$, які знаходяться за межами таблиць, випадкова величина H (для гіпотези H_0) приблизно розподілена як χ^2 з $(k-1)$ степенями вільності. При використанні цього наближення ми відкидаємо H_0 (на рівні значущості α), якщо $H_{\text{спос}} > \chi_{1-\alpha}^2$, де $\chi_{1-\alpha}^2$ — квантиль рівня $(1-\alpha)$ розподілу χ^2 з $(k-1)$ степенями вільності.

Значення, що збігаються

Якщо у табл. 5.2 є значення, що збігаються, при ранжуванні та переході до табл. 5.3 треба використовувати середні ранги. Якщо збігів багато, рекомендовано використовувати модифіковану форму статистики H' :

$$H' = \frac{H}{1 - \left(\sum_{j=1}^g T_j / [N^3 - N] \right)}, \quad (5.7)$$

де g — число груп спостережень, що збігаються; $T_j = (t_j^3 - t_j)$, t_j — число співпадаючих спостережень у групі з номером j . Детальніші відомості з цього приводу можна знайти, наприклад, у [2, 35, 40].

Якщо наявні вибірки впорядковані за зростанням впливу фактора, тоді доцільно використовувати критерій Джонкхієра. Статистика Джонкхієра детально викладена у роботі [8]. Розглянемо застосування непараметричних критеріїв на прикладі, взятому з роботи [8].

Для з'ясування впливу грошового стимулювання на продуктивність праці шести однорідним групам по п'ять чоловік були

запропоновані задачі однакової складності. Задачі пропонувалися кожному випробуваному незалежно від усіх інших. Групи відрізняються між собою величиною грошової винагороди за задачу, що розв'язується. Необхідно перевірити гіпотезу про відсутність впливу грошової винагороди на число розв'язаних задач. У табл. 5.4 наведено число розв'язаних задач членами кожної групи. Величина винагороди у таблиці збільшується від групи з меншим індексом до групи з більшим індексом.

Таблиця 5.4

Число розв'язаних задач членами кожної групи

Група 1	Група 2	Група 3	Група 4	Група 5	Група 6
10	8	12	12	24	19
11	10	17	15	16	18
9	16	14	16	22	27
13	13	9	16	18	25
7	12	16	19	20	24

На підставі аналізу таблиці можна зробити висновок, що фактор грошової винагороди має тенденцію до зростання. Нехай величини x_{ij} не мають нормального (гауссівського) закону розподілу, тоді для розв'язання задачі використовуємо непараметричні критерії Краскела — Уолліса та Джонкхієра. Як альтернативу до нульової гіпотези оберемо припущення, що чим вищий рівень стимулювання, тим вища продуктивність.

Застосуємо критерій Краскела — Уолліса для розв'язання цієї задачі. Результати взяті з роботи [2]. У зв'язку з наявністю у табл. 5.4 збігів, використовуємо середні ранги. Так, значення $x_{ij} = 10$ зустрічається у табл. 5.4 двічі й при впорядкуванні x_{ij} воно «поділяє п'яте та шосте місця». Тому середній ранг $x_{ij} = 10$ дорівнює 5,5. У результаті ранжування отримаємо табл. 5.5. У двох нижніх рядках наведені суми рангів R_j та середні ранги $R_j = R_j/n_j$ за стовпцями.

Для обчислення статистики Краскела — Уолліса H зручніше скористатися формулою (5.6). У нашому випадку загальне число спостережень $N = 30$, число спостережень при заданому значенні фактора $n_j = 5$, $j = 1, \dots, 6$. Підставляючи ці значення, маємо:

$$H = \frac{17682}{155} - 93 = 21,077.$$

Таблиця 5.5

Таблиця рангів спостережень

Група 1	Група 2	Група 3	Група 4	Група 5	Група 6
5,5	2	9	9	27,5	23,5
7	5,5	20	14	17	21,5
3,5	17	13	17	26	30
11,5	11,5	3,5	17	21,5	29
1	9	17	23,5	25	27,5
$R_1 = 28,5$	$R_2 = 45$	$R_3 = 62,5$	$R_4 = 80,5$	$R_5 = 117$	$R_6 = 131,5$
$R_1 = 5,7$	$R_2 = 9$	$R_3 = 12,5$	$R_4 = 16,1$	$R_5 = 23,4$	$R_6 = 26,3$

Величина H асимптотично має розподіл χ^2 з числом степенів вільності, що у даному випадку дорівнює 5. За таблицею розподілу χ^2 знаходимо, що мінімальний рівень значущості α трохи більший 0,001. Зауважимо, що цей висновок є приблизним у зв'язку з тим, що у табл. 5.4 було певне число збігів спостережень x_{ij} . Для урахування впливу зв'язків можна скористатися статистикою H' (5.7). У нашому випадку маємо вісім груп спостережень, що збігаються:

9, 9; 10, 10; 12, 12; 13, 13; 16, 16, 16, 16, 16; 18, 18; 19, 19; 24, 24.

Відповідно:

$$T_1 = 2^3 - 2 = 6, \quad T_2 = 2^3 - 2 = 6, \quad T_3 = 3^3 - 3 = 24,$$

$$T_4 = 6, \quad T_5 = 5^3 - 5 = 120, \quad T_6 = 6, \quad T_7 = 6, \quad T_8 = 6.$$

Знаменник дробу у виразі для H' дорівнює

$$1 - \sum_{j=1}^8 \frac{T_j}{(30^3 - 30)} = 1 - \frac{6}{899},$$

а саме значення H' приблизно дорівнює 21,2186.

Оскільки скоректоване значення H' статистики Краскела — Уолліса несуттєво відрізняється від значення H , можна відкинути гіпотезу на мініальному рівні значущості близько 0,001.

Застосування критерію Джонкхієра дає більш відчутний результат, оскільки має місце збільшення матеріального стимулювання зі зростанням індексу груп. За допомогою таблиць стандартного нормального розподілу знаходимо, що обчислене значення відповідає мініальному рівню значущості $\alpha \approx 3 \cdot 10^{-7}$.

З використанням критерію Краскела — Уолліса гіпотеза про однорідність відкидається на рівні значущості не менше $1 \cdot 10^{-3}$. Критерій Джонкхієра знизив рівень значущості майже на чотири порядки [2].

5.5. Однофакторний дисперсійний аналіз у пакетах прикладних програм

У пакетах прикладних програм (ППП) доволі широко відображені стандартні методи дисперсійного аналізу. Зауважимо, що у зарубіжній літературі дисперсійний аналіз представлений виразом *Analysis of variance* (аналіз варіацій), або *ANOVA*. Більшість програм однофакторного дисперсійного аналізу формують узагальнену таблицю, представлену нижче.

Таблиця 5.6

Таблиця однофакторного дисперсійного аналізу

Джерело дисперсії	Сума квадратів	Степені вільності	Середній квадрат	F-відношення	p
Між рівнями (групами)	$SS_B = \sum_{j=1}^u m_j (\bar{y}_j - \bar{y}_{..})^2$	$v_B = u - 1$	$MS_B = \frac{SS_B}{v_B}$	$F = \frac{MS_B}{MS_R}$	
Всередині рівнів (груп)	$SS_R = \sum_{j=1}^u \sum_{l=1}^{m_j} (y_{jl} - \bar{y}_{..})^2$	$v_R = m - 1$	$MS_R = \frac{SS_R}{v_R} = s^2$		
Повна	$SS_T = \sum_{j=1}^u \sum_{l=1}^{m_j} (y_{jl} - \bar{y}_{..})^2$	$v_T = m - 1$			

У таблиці зазвичай приводять суму квадратів (*SS*), число степенів вільності (*v*) та середній квадрат (*MS*) для кожного з джерел дисперсії — міжрівневого (або міжгрупового) та внутрішньорівневого (внутрішньогрупового). Останню величину називають ще *залишковою сумою квадратів* (або *сумою квадратів помилок*). Іноді щедрукуються ще повна сума квадратів і число степенів вільності. Для перевірки H_0 ми обчислимо *F*-відношення, тобто відношення середнього міжгрупового квадрата до середнього внутрішньогруповому квадрата.

Кожну гіпотезу перевіряють за допомогою F -тестування, в процесі якого порівнюють значення середнього квадрата для цього джерела варіації (яке є великим, якщо це джерело варіації призводить до відмінностей у значеннях кількісних показників) із середнім квадратом помилки, з'ясовуючи при цьому, наскільки сильніший вплив цього джерела варіації у порівнянні з чистою випадковістю. Щоб визначити, чи значущий вплив цього джерела варіації (скажемо, 1-го), порівнюють обчислене F -значення з табличним F -значенням, що узятो з числом степенів вільності для цього джерела у чисельнику та числом степенів вільності для помилки у знаменнику. Або ж слід просто подивитися на p -значення (p_i) й прийняти рішення про те, що результат значущий, якщо це значення достатньо мале, наприклад $p < 0,05$.

P -значенням служить площа праворуч від величини F під кривою площини розподілу $F(I - 1, n - I)$.

Як приклад, розглянемо розв'язання задачі дослідження впливу i -тої дієти ($i = 1, 2, 3, 4$) на вміст азоту, що видихається. Дієти містять різну кількість білків [1]. Припустимо, зафіксовано чотири дієти: $D1$ — безбілкова, $D2$ — 23 % білків, $D3$ — 32 % білків, $D4$ — 67 % білків. Припустимо також, що для i -тої дієти, $i = 1, \dots, 4$, експериментатор випадково відібрав 9 випробуваних. Такий експеримент адекватно описується моделлю однофакторного дисперсійного аналізу з фіксованими ефектами. У табл. 5.7 вказана кількість азоту, що видихається, для всіх $n = 36$ об'єктів,

Таблиця 5.7

Набір даних

	$D1$	$D2$	$D3$	$D4$
	4,079	4,368	4,169	4,928
	4,859	5,668	5,709	5,608
	3,540	3,752	4,416	4,940
	5,047	5,848	5,666	5,291
	3,298	3,802	4,123	4,674
	4,679	4,844	5,059	5,038
	2,870	3,578	4,403	4,905
	4,648	5,393	4,496	5,208
	3,847	4,374	4,688	4,806
Середні	4,096	4,625	4,747	5,044

а також для кожної дієти. Зауважимо, що кількість азоту, який видихається, у середньому збільшується при зростанні споживання білків. Цю задачу будемо розв'язувати за допомогою ППП «STATISTICA» (Release: 6).

У пакеті доволі широко відображені стандартні методи факторного аналізу. Доступ до частини з них здійснюється з пунктів **General ANOVA/MANOVA** та **Quick specs dialog**. Перш за все необхідно перевірити гіпотезу про наявність нормального закону розподілу для залежної змінної (АЗОТ). Перевірку здійснимо за критеріями, наведеними у розділах 3 і 4. Результати перевірки показали, що для нашого прикладу можна використовувати однофакторний дисперсійний аналіз.

Наступним етапом є підготовка даних. Звернемо увагу на особливості вихідних даних, що наведені у табл. 5.7. Є одна залежна змінна (АЗОТ), яка вимірюється кількісно, й один фактор (ДІЄТА), що варіюється на чотирьох рівнях (чотири дієти) і не має кількісного опису. У таблицю даних має бути введений вектор, який об'єднує всі значення змінної відклику (АЗОТ), що є в наявності, для кожного рівня варіювання (ДІЄТА). При цьому порядок розташування спостережень у цьому векторі для кожного рівня, взагалі кажучи, не є суттєвим. У нашому випадку дані, що відповідають різним дієтам, були записані в окремі змінні у вигляді послідовності чисел

111111111222222222333333333444444444.

Кожному числу відповідає значення залежної змінної. Після введення змінних увійдемо до опції **One-way ANOVA**. У якості залежної змінної виберемо АЗОТ, незалежним фактором буде ДІЄТА. Установивши прапорець **All effects/Graphs**, отримаємо таблицю, в якій представлені розрахункове значення $F = 3,211$ та величина $p = 0,036$. На підставі табличних значень робимо висновок, що на виділення азоту з довірчою ймовірністю 95 % впливають дієти, оскільки розрахункове значення $p = 0,036 < p = 0,05$. Далі, натискаючи кнопку **ОК**, отримаємо графік залежності середніх значень азоту від типу дієти та довірчі інтервали для цих середніх. На графіку добре видно, що кількість азоту, який видихається, залежить від типу дієти. Більш того, зі збільшенням номеру дієти кількість азоту монотонно збільшується. Увійшовши у меню **Symmary** та натиснувши кнопку **Cell statistics**, отри-

маємо оцінки середнього, стандартного відхилення, стандартної помилки, довірчих інтервалів для середнього за кожною дієтою. У меню **Comps** можна провести порівняльний аналіз середніх значень залежної змінної від типу дієт. Використовуючи опції **Resids, Matrix, Means** можна провести більш глибокий аналіз.

Перераховані вище процедури доволі слабо торкалися питання про правомірність застосування дисперсійного аналізу до аналізованих даних. Для більш детального розгляду цього питання користувачеві необхідно зберегти отримані залишки (процедура **Save residuals**) для подальшого аналізу в інших розділах пакетів. Зокрема, до них надалі можуть бути застосовані критерій χ^2 та критерії Колмогорова — Смирнова, Шапіро — Уїлкса, Лїлліфорса для перевірки узгодження з нормальним розподїлом. Також можна скористатися візуальним методом перевірки нормальності за допомогою графіка залишків на нормальному ймовїрнісному паперї або критеріями асиметрії та ексцесу. Для нашого прикладу складно серйозно говорити про можливість достовірної перевірки нормальності вибірки, яка складається з 36 елементів. Зауважимо лише, що певні сумніви у нормальності розподїлу залишків викликає вже гістограма залишків та їхній графік на нормальному ймовїрнісному паперї.

У наступному прикладї покажемо можливість розв'язання задачі визначення залежності величини зношеності шин від типу автомобіля й типу шин за допомогою одновимїрного дисперсійного аналізу, що наведений у ППП «STATGRAPHICS PLUS» Вихідні данї взятї з роботи [2] і наведенї у табл. 5.8.

Таблиця 5.8

Вихідні данї

№	Авто-мобіль	Шини	Зношеність	№	Авто-мобіль	Шини	Зношеність
1	1	1	1,85	9	3	1	1,67
2	1	2	2,12	10	3	2	2,18
3	1	3	1,83	11	3	3	2,24
4	1	4	1,83	12	3	4	1,65
5	2	1	1,73	13	4	1	2,03
6	2	2	2,28	14	4	2	2,52
7	2	3	2,27	15	4	3	1,89
8	2	4	1,93	16	4	4	2,01

У ППП «STATGRAPHICS PLUS» представлені такі модулі дисперсійного аналізу: **Factor Mean Plot, One-Way Anova, Multifaktor Anova, Variance Components**.

Дослідимо вплив типу шин на їхнє зношування. Для цього виберемо модуль **One-Way Anova** й заповнимо вікно однофакторного дисперсійного аналізу **ANOVA TABLE**. Натиснемо **OK** і отримаємо таблицю дисперсійного аналізу, а також графік розсіювання та графік з відгалуженнями.

Таблиця 5.9

Базова таблиця дисперсійного аналізу

Source of variation	Sum of Squares	d. f.	Mean square	F-ratio	P-value
Between groups	0,529	3	0,176	5,30	0,0147
Within groups	0,399	12	0,033		
Total (correct)	0,928	15			

Дамо переклад термінів, що фігурують у цій таблиці: **Source of Variation** — джерело варіації; **Between groups** — між групами; **Within groups** — всередині груп; **Total (correct)** — разом (скоректоване значення); **Sum of Squares** — сума квадратів; **d. f.** — степені вільності; **Mean square** — середні квадрати; **F-ratio** — F -відношення; **P-value** — рівень значущості.

Тепер пояснимо значення величин, що містяться у табл. 5.9. У рядку **Between groups** виводяться характеристики, пов'язані з дією аналізованого фактора: сума квадратів між групами (0,529), відповідне число степенів вільності (**d. f.**) та частку цих величин. У рядку **Within groups** виводяться сума квадратів усередині груп (0,399), відповідне число степенів вільності та частка цих величин. У рядку **Total (correct)** сума квадратів відхилень спостережень від їхнього середнього значення і число степенів вільності розподілу цієї величини при виконанні нульової гіпотези. У стовпчику **F-ratio** виводиться значення F -статистики, а у стовпчику **P-value** — її рівень значущості. Зазвичай, якщо ця величина близька до нуля, є підстава відкинути нульову гіпотезу.

Для даних нашого прикладу з наведеної таблиці дисперсійного аналізу (табл. 5.9) можна зробити висновок, що нульова гіпотеза про відсутність ефектів обробки повинна бути відкинута, оскільки ймовірність отримання вказаного або більшого значення F -відношення (рівень значущості F -статистики) при нульовій

гіпотезі практично дорівнює нулю. На підставі табличних значень робимо висновок, що зношення шин з довірчою ймовірністю 95 % впливають типи шин, оскільки розрахункове значення $p = 0,0147$ менше $p = 0,05$.

Для продовження поглибленого аналізу слід увійти до меню **Tabular Options**. При цьому на екрані з'явиться меню процедур аналізу у вигляді:

- ☒ **Analysis Summery**;
- ☒ **Summery statistics** — основні статистичні характеристики;
- ☒ **Table of Means** — таблиця середніх;
- ☒ **Multiple Range Tests** — множинні порівняння;
- ☒ **Variance Check** — тести дисперсій;
- ☒ **Kruskal-Wallis Test** — тест постійності розподілів.

Далі можна у будь-якому порядку виконати всі або деякі з процедур цього меню. Вони у різній формі надають інформацію щодо двох груп питань: оцінювання ефектів обробки й аналізу адекватності застосування методу дисперсійного аналізу.

Аналогічну інформацію можна отримати у модулі **Graphical Options**, де подані такі графіки:

- ☒ **Scatterplo** — графік розсіювання;
- ☒ **Means Plot** — графік середніх;
- ☒ **Box-and-Whiser Plot** — графік прямокутників;
- ☒ **Residuals versus Predicted**;
- ☒ **Residuals versus Row Number**;
- ☒ **Analysis of Means (FNOM) Plot**.

Отримати оцінки ефектів обробки й побудувати для них довірчі інтервали дозволяє процедура **Table of Means**. Процедура **Means plot** реалізує графічне подання даних таблиці, що видається процедурою **Means table**. Процедура **Scatterplo** дозволяє за графіком розсіювання визначати форму залежностей і знаходити «викиди». Графік прямокутників **Box-and-Whiser Plot** дозволяє для кожного рівня обробки отримати у компактній формі інформацію про медіани, верхні та нижні квартилі розмахів та низку інших параметрів. Ці процедури надають корисні відомості про симетрію закону розподілу даних на кожному рівні обробки, про можливі грубі викиди у спостереженнях, а також наочно порівнюють положення основних мас даних для різних рівнів фактора.

Процедура **Residual plots** пропонує побудування графіка залишків в одній з трьох можливих форм: у залежності від рівня

фактора, від передбачених значень або просто в залежності від індексу спостереження, у векторі введення даних. Кожна з цих форм підкреслює свій аспект у можливих причинах порушення однорідності розподілу залишків.

З точки зору обґрунтованості застосування у цілому методу дисперсійного аналізу, вельми важливою є процедура **Varian-**
ce check. Вона охоплює результати трьох статистичних крите-
ріїв — Кокрена, Бартлетта та Хартлі, — для порівняння розкидів
спостережень на різних рівнях фактора. Для даних прикладу ця
процедура виводить на екран такі результати:

Tests for Homogeneity of Variances

Cochran's test:	0,39	$p = 0,756$
Bartlett's test:	1,05	$p = 0,9$
Hartley's test:	2,189	

При цьому за першими двома критеріями (Кокрена та Бартлетта) немає підстав відкидати нульову гіпотезу цих критеріїв про те, що дані на різних рівнях обробки мають одну й ту саму дисперсію. Однак зауважимо, що критерії Кокрена та Бартлетта вельми чутливі до відхилення від припущення, що нормовані оцінки дисперсії для кожного рівня обробки підкоряються розподілу χ^2 (з відповідним числом степенів вільності). Ця вимога виконується, якщо дані для кожного рівня обробки належать до сімейства нормальних розподілів. Для інших сімейств розподілів ця вимога частіше за все не дотримується.

Критерій Краскела — Уолліса перевіряє гіпотезу, чи мають порівнювані вибірки один і той самий розподіл, і ґрунтується на рангах. У нашому прикладі критерій Краскела — Уолліса є значущим ($p = 0,04$). Таким чином, характеристики виборок типу шин значуще відрізняються одна від одної.

5.6. Двофакторний дисперсійний аналіз

Вивчається вплив двох одночасно діючих факторів X_1 та X_2 . Подамо результати експерименту у вигляді табл. 5.10 ($U_1 \times U_2 \times m$) спостережень y_{jgl} , де j — порядковий номер y рівня варіювання

фактора X_1 , $j = \overline{1, U_1}$; g — порядковий номер y рівня варіювання фактора X_2 , $g = \overline{1, U_2}$; l — порядковий номер дублюючого дослідів у серії при кожному ig сполученні рівнів двох факторів $l = \overline{1, m_{ig}}$.

Для спрощення розрахунків розглянемо спочатку випадок рівночисельних спостережень за всіх можливих сполучень рівнів, тобто $m_{11} = m_{12} = m_{13} = \dots = m_{ig} = \dots = m_{U_1 U_2} = m$.

Таблиця 5.10

Результати експерименту

№ j фактору X_1	№ j фактору X_2						y_j
	1	2	...	g	...	u_2	
1	y_{111}	y_{121}		y_{1g1}		$y_{1 u_2 1}$	$\bar{y}_1$
	y_{112}	y_{122}		y_{1g2}		$y_{1 u_2 2}$	
	...	...		...		...	
	y_{11m1}	y_{12m}		y_{1gm}		$y_{1 u_2 m}$	
2	y_{211}						$\bar{y}_2$
	y_{212}						
	...						
	y_{21m}						
...							...
j	y_{j11}						$\bar{y}_j$
	y_{j12}						
	...						
	y_{j1m}						
...							...
u_1	$y_{u1 11}$					$y_{u1 u_2 1}$	$\bar{y}_{u1}$
	$y_{u1 12}$					$y_{u1 u_2 2}$	
	$y_{u1 1m}$					$y_{u1 u_2 m}$	
$\bar{y}_g$	$\bar{y}_1$	$\bar{y}_2$	...	$\bar{y}_g$	...	$\bar{y}_{u2}$	$\bar{y}$

У таблиці середні значення по вертикалі й горизонталі оцінюються за такими формулами:

$$\bar{y}_g = \frac{1}{u_1 m} \sum_{j=1}^{u_1} \sum_{l=1}^m y_{gjl},$$

$$\bar{y}_j = \frac{1}{m u_2} \sum_{g=1}^{u_2} \sum_{l=1}^m y_{jgl},$$

$$\bar{y} = \frac{1}{u_2 u_1 m} \sum_{j=1}^{u_1} \sum_{g=1}^{u_2} y_{jgl}.$$

Обчислимо середні арифметичні y_{jg} серій з m повторюваних спостережень для кожного сполучення рівнів j та g факторів X_1 та X_2 :

$$\begin{aligned} \bar{y}_{jg} &= \frac{1}{m} \sum_{l=1}^m y_{jgl}, \\ \bar{y} &= \frac{1}{u_1 u_2 m} \sum_{j=1}^{u_1} \sum_{g=1}^{u_2} \sum_{l=1}^m y_{jgl} = \frac{1}{u_1 u_2} \sum_{j=1}^{u_1} \sum_{g=1}^{u_2} \bar{y}_{jg} = \frac{1}{u_1} \sum_{j=1}^{u_1} \bar{y}_j = \frac{1}{u_2} \sum_{g=1}^{u_2} \bar{y}_g. \end{aligned} \quad (5.8)$$

Розсіювання спостережень у кожній серії зумовлене дією лише випадкових причин з дисперсією δ^2 . Розсіювання ж власне середніх серій за всіма можливими сполученнями рівнів X_1 та X_2 навколо загального середнього, окрім фактора випадковості, спричинюється фактором взаємодії X_{12} (з дисперсією δ_{12}^2). Окрім цих факторів, на розсіювання середніх за рядками впливає лише один фактор X_1 (з дисперсією δ_1^2), а на розсіювання середніх за стовпцями — лише один фактор X_2 з дисперсією δ_2^2 , оскільки всі рівні іншого фактора в кожному з цих випадків осереднені.

У відповідності до основної ідеї дисперсійного аналізу розкладемо суму S квадратів відхилень спостережень від загального середнього на компоненти, які відповідають зазначеним факторам:

$$\begin{aligned} S &= \sum_{j=1}^{u_1} \sum_{g=1}^{u_2} \sum_{l=1}^m (y_{jgl} - \bar{y})^2 = \sum_{j=1}^{u_1} \sum_{g=1}^{u_2} \sum_{l=1}^m (y_{jgl} - \bar{y}_{jg})^2 + u_2 m \sum_{j=1}^{u_1} (\bar{y}_j - \bar{y})^2 + \\ &+ u_1 m \sum_{g=1}^{u_2} (\bar{y}_g - \bar{y})^2 + m \sum_{j=1}^{u_1} \sum_{g=1}^{u_2} (\bar{y}_{jg} - \bar{y}_j - \bar{y}_g + \bar{y})^2 = S_0 + S_1 + S_2 + S_{21}, \end{aligned}$$

де S — загальна сума квадратів, що характеризує розсіювання окремих спостережень y_{jgl} у загальній сукупності за рахунок впливу всіх факторів; S_0 — сума квадратів відхилень «всередині серій». Сума характеризує розсіювання окремих спостережень за рахунок впливу фактора випадковості; S_1 — сума квадратів відхилень «між рядками». Сума $\frac{S_1}{u_2 m}$ характеризує розсіювання

середніх $\bar{y}_j$ за рядками у результаті дії фактора випадковості з дисперсією $\frac{\delta^2}{u_2 m}$, фактора X_1 з дисперсією δ_1^2 та фактора взаємодії з дисперсією середнього рядка $\frac{\delta_{12}}{u_2}$; S_2 — сума квадратів відхилень «між рядками». Сума $\frac{S_2}{u_1 m}$ характеризує розсіювання середніх $\bar{y}_g$ за стовпцями в результаті дії фактора випадковості з дисперсією середнього стовпця $\frac{\delta^2}{u_1 m}$, фактора X_2 (з дисперсією δ_2^2 та фактора взаємодії з дисперсією середнього стовпчику $\frac{\delta_{12}^2}{u_1}$); S_{12} — сума квадратів відхилень «між серіями». Сума S_{12} характеризує розсіювання середніх $\bar{y}_{jg}$ серій у результаті дії фактора випадковості з дисперсією середнього $\frac{\delta^2}{m}$ та фактора взаємодії з дисперсією δ_{12}^2 .

Суми квадратів S, S_0, S_1, S_2, S_{12} , поділені на відповідне число степенів вільності v, v_0, v_1, v_2, v_{12} , дають незміщені оцінки дисперсії відтворюваності δ^2 . Оцінимо такі вибіркові дисперсії:

☑ вибіркова загальна дисперсія за всіма $u_1 u_2 m$ спостереженнями:

$$S^2 = \frac{S}{u_1 u_2 m - 1} \approx \delta^2 + \delta_1^2 + \delta_2^2 + \delta_{12}^2,$$

з числом степенів вільності $v = u_1 u_2 m - 1$;

☑ вибіркова дисперсія розсіювання «всередині серій», або залишкова оцінка, є середньозваженою дисперсією з усіх серій спостережень:

$$S_0^2 = \frac{S}{u_1 u_2 m - 1} \approx \delta^2$$

з числом степенів вільності $v_0 = u_1 u_2 (m - 1)$;

☑ вибіркова дисперсія розсіювання «між рядками»:

$$S_1^2 = \frac{S_{B1}}{u_1 - 1} = \delta^2 + u_2 m \delta_1^2 + m \delta_{12}$$

з числом степенів вільності $v_1 = u_1 - 1$;

☑ вибіркова дисперсія розсіювання «між стовпчиками»:

$$S_2^2 = \frac{S_{B2}}{u_2 - 1} \approx \delta^2 + u_1 m \delta_2^2 + m \delta_{12}^2$$

з числом степенів вільності $v_2 = u_2 - 1$;

☑ вибіркова дисперсія розсіювання «між серіями»:

$$S_{12}^2 = \frac{S_{B12}}{u_1 u_2 - u_1 - u_2 + 1} \approx \delta^2 + m \delta_{12}^2$$

з числом степенів вільності $v_{12} = (u_1 - 1)(u_2 - 1)$.

Число степенів вільності перевіряється за відношенням $v = v_0 + v_1 + v_2 + v_{12}$.

Аналіз значущості впливу факторів x_1 , x_2 та їх взаємодії x_{12} проводиться за критерієм Фішера при обраному рівні значущості α у такому порядку:

1) Вплив факторів x_1 та x_2 відповідно з дисперсіями

$$\delta_1^2 = \frac{1}{u_2 m} (S_1^2 - S_{12}^2), \quad \delta_2^2 = \frac{1}{u_1 m} (S_2^2 - S_{12}^2)$$

визнається значущим ($\delta_1^2 > 0$, $\delta_2^2 > 0$), якщо виявиться значущою відповідно відмінність S_1^2 від S_{12}^2 і S_2^2 від S_{12}^2 та виконуватиметься умова:

$$F_1 = \frac{S_1^2}{S_{12}^2} > F_\alpha(v_1, v_{21}), \quad F_2 = \frac{S_2^2}{S_{12}^2} > F_\alpha(v_2, v_{21}).$$

Якщо одне з цих дисперсійних відношень, тобто вплив відповідного фактора, незначущий ($\delta_1^2 = 0$ або $\delta_2^2 = 0$), то для дисперсії ($\delta^2 + m \delta_{12}^2$) отримаємо дві оцінки S_1^2 та S_{12}^2 або S_2^2 та S_{12}^2 відповідно, які можна об'єднати у вільну оцінку $S_{12}^2 = \delta^2 + m \delta_{12}^2$ з більшим числом степенів вільності $\bar{v}_{12} = u_2(u_1 - 1)$ або $\bar{v}_{12} = u_1(u_2 - 1)$.

Якщо обидва фактори x_1 та x_2 незначущі ($\delta_1^2 = 0$, $\delta_2^2 = 0$), то оцінки для дисперсії $\delta^2 + m \delta_{12}^2$ можна об'єднати в оцінку $S_{12}^2 = \delta^2 + m \delta_{12}^2$ з більшим числом степенів вільності $\bar{v}_{12} = u_1 u_2 - 1$.

2) Вплив взаємодії x_{12} з дисперсією:

$$\delta_{12}^2 = \frac{1}{m} (S_{12}^2 - S_0^2),$$

визнається значущим ($\delta_{12}^2 > 0$), якщо відмінність S_{12}^2 та S_0^2 виявиться значущою, тобто якщо критерій:

$$F = \frac{S_{12}^2}{S_0^2} > F_\alpha(v_{12}, v_0).$$

В іншому випадку $\delta_{12}^2 = 0$ й обидві оцінки S_{12}^2 та S_0^2 для δ^2 можна об'єднати в одну $\bar{S}^2 \approx \delta^2$ з більшим числом степенів вільності $\bar{v}_0 = u_1 u_2 m - u_1 - u_2 + 1$.

Якщо x_1 , x_2 та x_{12} значущі, то дисперсію відтворюваності можна оцінити вибірковою загальною дисперсією S^2 .

Взаємодію факторів можна оцінити графічно. Якщо на одній осі відобразити один фактор, а на іншій — середнє значення оцінок функції відклику, то можна провести лінію для кожного рівня варіювання другого фактора. Наявність різного кута нахилу ліній вказуватиме на взаємодію факторів.

За наявності трьох і більше факторів, що визначають вибірку, дисперсійний аналіз теж дозволяє вивчити головний ефект кожного фактора (визначити наявність його впливу) та взаємодію факторів (визначити зв'язок факторів між собою). Тут ми маємо справу з більшою порівняно з двофакторним дисперсійним аналізом кількістю видів взаємодії, кожен з яких вивчається окремо.

За наявності більш однієї кількісної залежної змінної для вивчення відмінності залежних змінних між вибірками можна використовувати багатовимірний дисперсійний аналіз.

Висновки

Загальну схему проведення дисперсійного аналізу подамо у вигляді послідовності розв'язання задач:

- ☒ вибір або розробка методики отримання вихідних даних та проведення їх аналізу;
- ☒ вибір методу проведення дисперсійного аналізу;
- ☒ підготовка даних для обраного методу;
- ☒ аналіз результатів з проведенням поглибленого аналізу.

Слід зауважити, що в пакетах прикладних програм модуль **ANOVA** не завжди розкриває повністю аналіз, який потрібний для глибокого розуміння суті взаємозв'язків між залежними та незалежними змінними в досліджуваному об'єкті. Так,

наприклад, у ППП «STATISTICA» для проведення поглибленого аналізу доцільно також використовувати модулі **Basic Statistics, Nonparametrics**.

Для отримання вихідних виборок бажано провести активний експеримент.

На першому етапі проводиться попередній аналіз даних, який дозволить порівняти одразу декілька розподілів, оцінити структуру даних та перевірити початкові умови, необхідні для дисперсійного аналізу, а саме нормальність розподілів та рівність варіації.

Дані для однофакторного дисперсійного аналізу являють собою набір незалежних одновимірних виборок, елементи яких виміряні в однакових одиницях.

Попередній аналіз даних дозволить коректно обрати метод дисперсійного аналізу для об'єкта, що досліджується.

Існує багато методів дисперсійного аналізу, включаючи двофакторний план та плани більш високого порядку. Так, однофакторний дисперсійний аналіз досліджує ефект дії одного якісного фактора на одну змінну відклику у припущенні, що дані належать до сімейства нормальних розподілів з однаковою дисперсією. Багатофакторний дисперсійний аналіз досліджує ефект дії одного або декількох якісних факторів на одну змінну відклику у припущенні, що дані належать до сімейства нормальних розподілів з однаковою дисперсією.

Ранговий однофакторний аналіз Краскела — Уолліса досліджує ефект дії одного фактора класифікації для збалансованого або незбалансованого однофакторного плану. Ранговий двофакторний аналіз Фрідмана досліджує ефект дії двох факторів класифікації для збалансованого двофакторного плану. Аналіз групових планів досліджує ефект дії одного або декількох якісних факторів на одну змінну відклику, якщо дані повністю згруповані або ієрархічні, а число спостережень дорівнює всім комбінаціям рівнів факторів. Багато інших типів планів експерименту в ППП можна дослідити за допомогою процедури множинної регресії.

Слід звернути особливу увагу на той факт, що перевірити припущення про нормальність закону розподілу для малих виборок практично неможливо. Це призводить до необхідності використовувати непараметричні методи дисперсійного аналізу, які менш ефективні порівняно з класичними методами.

Контрольні запитання та вправи

1. Поясніть сенс дисперсійного аналізу.
2. Назвіть та прокоментуйте два джерела варіації в однофакторному дисперсійному аналізі.
3. Чи мають бути однаковими розміри виборок у однофакторному дисперсійному аналізі?
4. Що характеризує сума квадратів відхилень «між серіями»?
5. Які приблизні оцінки можна дати для дисперсії фактора X ?
6. Як оцінити дисперсії у двофакторному дисперсійному аналізі?
7. За яким критерієм проводиться аналіз значущості впливу факторів x_1 , x_2 та їх взаємодії x_{12} за обраного рівня значущості α ?
8. Чому не слід використовувати t -тест для незалежних виборок замість однофакторного дисперсійного аналізу?
9. Які умови повинні виконуватись для коректного застосування класичного дисперсійного аналізу?

Завдання 1

У чотирьох кішок народилося по три кошеняти, для яких були складені три раціони харчування. Протягом трьох тижнів за кожним з раціонів харчувалося одне кошеня, випадково обране з трьох. Отримані збільшення у вазі наведені в табл. 5.11.

Таблиця 5.11

Кішка	Раціон		
	1	2	3
1	9,1	7,4	5,2
2	13,8	13,0	11,4
3	8,8	9,5	4,2
4	13,0	11,9	10,7

Перевірте гіпотезу: чи впливав конкретний раціон на збільшення ваги кошенят?

Завдання 2

Для виявлення зв'язку між дихальною функцією і звичкою до паління результати легеневих проб у групі співробітників установи були зіставлені з режимом паління. В одній з таких проб вимірювався обсяг (у літрах) повітря, що видихається через 1 с після форсованого вдиху. Результати обстеження наведено в табл. 5.12.

Перевірте за допомогою однофакторного дисперсійного аналізу, чи свідчать ці дані про зв'язок між палінням і дихальною функцією?

Використовуючи однофакторний дисперсійний аналіз, перевірте гіпотезу про відсутність такої залежності.

Таблиця 5.12

Легенева проба	Нікотинова залежність	
	не палять	палять
Ненормальна	2	16
Нормальна	64	83

Завдання 3

Нижче наведено дані (див. табл. 5.13) про ступінь кмітливості школярів, які мають атлетичну та неатлетичну статуру.

Таблиця 5.13

Статура школяра	Ступінь кмітливості	
	високий	низький
Атлетична	581	567
Неатлетична	209	351

Чи можна сказати, що існує зв'язок між кмітливістю школярів та їхньою статуєю? Перевірте за допомогою однофакторного дисперсійного аналізу.

Завдання 4

Проводилось обстеження з метою виявлення наявності зв'язку між віком тих, хто одружується вперше, та рівнем їхніх доходів. Результати обстеження наведено у табл. 5.14.

Таблиця 5.14

Рівень доходів	Вік молодят, років		
	до 18	18–21	старші 21
Низький	45	25	15
Середній	35	60	25
Високий	10	28	24

Чи свідчать ці дані про наявність зв'язку між віком першого одруження та рівнем доходів молодих? Розрахуйте суму квадратів відхилень «всередині серій», суму квадратів відхилень «між серіями».

РОЗДІЛ 6

КЛАСТЕРНИЙ АНАЛІЗ

6.1. Галузі застосування

Кластерний аналіз (від англ. *cluster* — гроно, пучок, скупчення, група елементів, що характеризуються будь-якою загальною властивістю) призначений для того, щоб поділити велику кількість об'єктів на задане або невідоме число класів на підставі деякого математичного критерію якості. Кластерний аналіз (КА) широко використовується для розв'язання багатьох функціональних задач в ІУС, що розробляються для різних предметних галузей. Наприклад, в організаційно-економічних ІУС кластерний аналіз знаходить застосування при розв'язанні задач перспективного та поточного планування, обліку, контролю і аналізу діяльності підприємства, прогнозування його розвитку, управління основним та допоміжним виробництвами, матеріально-технічним постачанням та збутом продукції, а також задач управління технічною підготовкою виробництва, транспортними процесами, трудовими ресурсами, фінансовими та грошовими коштами тощо. В АСУТП кластерний аналіз застосовують для первинної обробки інформації, побудови динамічних та статичних моделей, синтезу систем управління.

У медичній галузі КА (зокрема, метод таксономії) покладено в основу класифікації захворювань, їх симптомів та способів лікування. У психіатрії правильна діагностика кластерів симптомів деяких захворювань, таких як параноя, шизофренія тощо,

є вирішальною для успішної терапії. В археології за допомогою кластерного аналізу дослідники намагаються встановити таксономії кам'яних знарядь, поховальних об'єктів тощо. Ефективним виявилось застосування кластерного аналізу в маркетингових дослідженнях. Взагалі, в предметних галузях, де необхідно класифікувати велику кількість об'єктів за групами, кластерний аналіз виявляється вельми корисним та ефективним.

6.2. Постановка задачі

Кластерний аналіз є одним із методів багатовимірної класифікації і характеризується відсутністю апріорної інформації про розподіл генеральної сукупності вектора ознак. Методи КА дозволяють поділити велику кількість об'єктів на групи «схожих» об'єктів на підставі деякого критерію якості класифікації.

Кластерний аналіз ґрунтується на таких вимогах:

- ☒ всередині груп об'єкти повинні бути щільно пов'язані між собою;
- ☒ об'єкти різних груп мають перебувати на відстані одне від одного;
- ☒ за інших рівних умов розподіл об'єктів за групами має бути рівномірним [18].

Методи кластерного аналізу використовуються у більшості випадків, коли немає будь-яких апріорних гіпотез щодо класів. Слід розуміти, що кластерний аналіз визначає «найбільш можливе значуще рішення», тому перевірка статистичної значущості насправді тут непридатна — навіть у випадках, коли відомі P -рівні (як, наприклад, у методі K середніх).

Задачу знаходження кластерів сформулюємо таким чином: нехай досліджується сукупність n об'єктів, кожен з яких характеризується за K виміряними на ньому ознаками.

Вихідні дані подаються у вигляді матриці

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix},$$

складеної з n вимірювань (об'єктів) та K ознак. Потрібно поділити цю сукупність на однорідні групи (класи) у відповідності з вибраними критеріями близькості.

Розглянемо принцип подання даних у кластерному аналізі. Числові значення, що входять у матрицю X , можуть відповідати трьом типам змінних [18]:

- ☒ кількісним,
- ☒ ранговим,
- ☒ якісним.

Кількісні змінні мають властивість впорядкованості, над ними можна проводити математичні операції. Значення *рангових змінних* також впорядковані, їх можна пронумерувати натуральними числами. Однак використання цих чисел в арифметичних операціях буде некоректним. *Якісні змінні* набувають двох (дихотомні) або більше значень. Цим значенням можна поставити у відповідність деякі числа, що відображатимуть будь-яку впорядкованість значень якісної змінної.

Бажано, щоб таблиця вихідних даних відповідала одному типу змінних. В іншому випадку різні типи змінних намагаються звести до якогось одного типу змінних. Наприклад, усі змінні можна звести до дихотомних, використовуючи подану нижче процедуру. Кількісні змінні переводять у рангові, розбиваючи області значень кількісної змінної на інтервали, які потім нумеруються числами натурального ряду. Рангові змінні автоматично стають якісними, якщо не враховувати впорядкованості їхніх значень. Що стосується якісних змінних, то кожному з можливих їх значень доводиться ставити у відповідність дихотомну змінну (вона дорівнюватиме 1, якщо якісна змінна прийняла це значення, і 0 — в іншому випадку).

Зазначимо, що форма запису вихідних даних, їх зведення до одного типу, можливість використання тільки частини даних відіграють певну роль в оцінці практичної ефективності обчислювального комплексу, призначеного для розв'язання задач класифікації.

Матриця X не є єдиним способом подання вихідних даних у задачах кластерного аналізу. Іноді вихідна інформація задана у вигляді квадратної матриці:

$$R = (\rho_{ij}), \quad i, j = 1, 2, \dots, k,$$

елемент ρ_{ij} якої визначає ступінь близькості i -го та j -го об'єктів.

6.3. Короткий аналіз методів кластеризації

Алгоритми кластерного аналізу відрізняються великим розмаїттям. Це можуть бути, наприклад, алгоритми повного перебору сполучень об'єктів або генерації випадкових груп. Разом з тим, більшість таких алгоритмів складається з двох етапів. На першому етапі задається початкове (штучна або довільна) розбиття великої кількості об'єктів на класи і визначається математичний критерій якості автоматичної класифікації. На другому етапі об'єкти переносяться з класу в клас, доки значення критерію не перестане поліпшуватися.

Розмаїття алгоритмів кластерного аналізу зумовлено також безліччю різних критеріїв, що виражають ті чи інші аспекти якості автоматичного групування. Найпростіший критерій якості безпосередньо базується на величині відстані між кластерами. Однак такий критерій не враховує «населеність» кластерів — відносну щільність розподілу об'єктів усередині виділюваних групувань. Тому інші критерії ґрунтуються на обчисленні середніх відстаней між об'єктами всередині кластерів. Але найчастіше застосовуються критерії у вигляді відносин показників «населеності» кластерів до відстані між ними. Це, наприклад, може бути відношення суми міжкласових відстаней до суми внутрішньокласових (між об'єктами) відстаней або відношення загальної дисперсії даних до суми внутрішньокласових дисперсій і дисперсії центрів кластерів.

Різноманітні процедури кластерного аналізу входять до складу практично всіх сучасних пакетів прикладних програм для статистичної обробки багатовимірних даних. Значного поширення набули алгоритми ієрархічного групування об'єктів і ознак. Ці алгоритми призначені для отримання наочного уявлення про стратифікаційну структуру всієї досліджуваної сукупності об'єктів. Вони засновані на послідовному об'єднанні кластерів (*агломеративні процедури*, від слова «агломерат» — скупчення) або на послідовному розбитті (*дивізивні процедури*, від англ. *division* — розподіл, поділ).

Найбільшу популярність мають агломеративні процедури. В агломеративних, або об'єднавчих, методах відбувається послідовне об'єднання найбільш близьких об'єктів у один кластер. Процес такого послідовного об'єднання можна показати на гра-

фіку у вигляді *дендрограми*, або *дерева об'єднання*. Таке подання дозволяє наочно зобразити кластеризацію агломеративними алгоритмами.

Розглянемо відповідну послідовність операцій. На першому кроці в таких процедурах всі об'єкти вважаються окремими кластерами. Потім на кожному наступному кроці два найближчих кластери об'єднуються в один. Кожне об'єднання зменшує число кластерів на одиницю, отже зрештою всі об'єкти поєднуються в один кластер. Найбільш прийнятне розбиття обирає найчастіше сам дослідник, якому надається дендрограма, що відображає результати групування об'єктів на всіх кроках алгоритму. Водночас можна використовувати також і математичні критерії якості об'єднання.

Напрявленість алгоритмів кластерного аналізу на певну структуру групувань може призводити до неоптимальних чи навіть неправильних результатів, якщо гіпотеза про тип групувань є хибною. За відмінності реальних і гіпотетичних розподілів зазначені алгоритми часто «нав'язують» даним не властиву їм структуру і дезорієнтують дослідника. Тому експериментатор, враховуючи даний факт, в умовах апріорної невизначеності звертається до застосування «батареї» алгоритмів кластерного аналізу і надає перевагу якому-небудь висновку на підставі комплексної оцінки сукупності результатів роботи алгоритмів [20].

6.4. Міра близькості між об'єктами або ознаками

Більшість алгоритмів кластерного аналізу або цілком входять з матриці близькості, або вимагають обчислення окремих її елементів, тому, якщо дані представлені у вигляді матриці X , то першим етапом розв'язання задачі пошуку кластерів буде вибір способу обчислення відстаней або близькості між об'єктами чи ознаками (у цьому відношенні розбіжність між об'єктами й ознаками є істотною).

Відносно просто визначається близькість між ознаками. Як правило, кластерний аналіз ознак переслідують ті ж самі цілі, що й факторний аналіз — виділення груп пов'язаних між собою ознак, що відображають певну сторону досліджуваних об'єктів. У цьому випадку мірами близькості служать різні статистичні коефіцієнти зв'язку.

Якщо ознаки кількісні, то можна використовувати оцінки звичайних парних вибірових коефіцієнтів кореляції r_{ij} , де $i, j = 1, 2, \dots, k$. Однак коефіцієнт кореляції вимірює тільки лінійний статистичний зв'язок, тому якщо зв'язок нелінійний, то слід використовувати кореляційне відношення або виконати відповідне перетворення шкали ознак. Не слід забувати, що кореляційний аналіз припускає наявність стаціонарної вибірки. В економіці й інших предметних галузях змінні представлені у вигляді часових рядів, для яких характерна наявність тренда, сезонної, циклічної складових. У цьому випадку ми маємо нестационарні послідовності і використання класичної кореляційної теорії для оцінки щільності зв'язку між ознаками може призвести до помилкової класифікації. Інтерес становить використання модернізованих коефіцієнтів кореляції, що дозволяють вимірювати щільність статистичного зв'язку в часі, а також залежно від амплітуди їхніх відхилень щодо певної критичної величини [28]. Існують також різні коефіцієнти зв'язку, визначені для рангових, якісних і дихотомних змінних.

Найбільш складним і найменш формалізованим у задачі класифікації є визначення поняття однорідності об'єктів. У загальному випадку однорідність об'єктів задається шляхом введення правила обчислень відстані $\rho(X_i, X_j)$ між будь-якою парою досліджуваних об'єктів $(X_1, X_2, \dots, X_n)$ або завданням деякої функції $r(X_i, X_j)$, що характеризує ступінь близькості i -го і j -го об'єктів. Якщо задана функція $\rho(X_i, X_j)$, то близькі з точки зору цієї метрики об'єкти вважаються однорідними, належними до одного класу. При цьому необхідно зіставляти $\rho(X_i, X_j)$ з деяким граничним значенням, що визначається у кожному конкретному випадку по-своєму.

Аналогічно використовується й міра близькості $r(X_i, X_j)$, при завданні якої потрібно пам'ятати про необхідність виконання умов симетрії $r(X_i, X_j) = r(X_j, X_i)$; максимальної подібності об'єкта із самим собою $r(X_i, X_i) = \max r(X_i, X_j)$, при $1 \leq j \leq n$, і монотонного спадання $r(X_i, X_j)$ по $\rho(X_i, X_j)$, тобто з нерівності $\rho(X_k, X_l) \geq \rho(X_i, X_j)$ має впливати нерівність $r(X_k, X_l) \geq r(X_i, X_j)$.

Розглянемо деякі статистичні критерії оцінки неоднорідностей. Зауважимо, що застосування критеріїв Бартлетта і Кокрена [15] для виявлення неоднорідностей у часових рядах дуже проблематичні, тому що вони ґрунтуються на наявності повторних вимірів.

За відсутності повторних досвідів використовуємо модифіковані критерії перевірки неоднорідності вибірки типу M -критерію Бартлетта, критерій Харрісона і Мак-Кейба, параметричний критерій Голдфілда і Квандта.

Модифікований M -критерій Бартлетта ґрунтується на припущенні про нормальність часового ряду. У цьому випадку перевіряється нуль-гіпотеза H_0 : всі значення часового ряду X_i , $i = 1, n$, розподілені як $\delta^2 X^2$, проти альтернативи H_0 : $\hat{X}_i - \hat{\delta}_x^2 X^2$, де не всі $\hat{\delta}_x^2$ однакові. Для перевірки гіпотези H_0 вихідну вибірку розбивають на три групи і для кожної обчислюють середній квадрат за формулами

$$S_1^2 = \sum_{j1=1}^{N1} \frac{\hat{X}_{j1}^2}{N1}; \quad S_2^2 = \sum_{j2=N+1}^{N2} \frac{\hat{X}_{j2}^2}{N2}; \quad S_3^2 = \sum_{j3}^{N3} \frac{\hat{X}_{j3}^2}{N3}; \quad j = 1, 2, 3,$$

де $N1 + N2 + N3 = N$.

Для S_1^2 , S_2^2 , S_3^2 обчислюємо статистику:

$$\mu = \frac{\sum_{j=1}^3 3 \ln S_j^2 / S^2}{1 + \left(\frac{1}{3} (N_j - 1) \left(\sum_{j=1}^P \frac{1}{V_j} - \frac{1}{V} \right) \right)},$$

де $V_j = 2$, $S^2 = N \sum_{i=1}^N \hat{x}_i^2$, $V = N_j - 1$.

При нуль-гіпотезі μ має розподіл приблизно X^2 з двома степенями вільності. Задаючись рівнем значущості з таблиці X^2 розподілу, вибираємо $X_T^2(\alpha, 2)$. Якщо $\mu < X_T^2(\alpha, 2)$, то H_0 виконується і дисперсія у часовому ряді однорідна, у протилежному випадку гіпотеза H_0 не виконується.

M -критерій Бартлетта не вимагає неоднорідності даних, але дуже чутливий до порушення нормальності.

Наявність ексцесу й асиметрії у вихідних даних свідчить про відхилення від нормального розподілу, що, безумовно, позначається на кінцевому результаті.

Аналогічно для оцінки неоднорідності часових рядів можна запропонувати використовувати граничний критерій Харрісона та Мак-Кейба. У припущенні про нормальність перевіряється нульова гіпотеза H_0 : $\delta^2 J$ проти альтернативної H_1 : $\text{diag}(\delta_1^2, \delta_2^2, \dots, \delta_N^2)$.

Для перевірки гіпотези обчислюємо $C = X^T A X / X^T X$, де A — діагональна $(N \times N)$ -матриця з q елементами 0 $< q < N$, що дорівнюють 1, і $(N - q)$ елементами, що дорівнюють 0.

Знаходимо значення C_L і C_U за такими виразами:

$$C_L = \left(1 + \frac{(N - q)F_\alpha(N - q, q)}{q} \right)^{-1}, \quad C_U = \left(1 + \frac{(N - q)F_\alpha(N - q, q)}{q} \right)^{-1},$$

де F_α — табличне значення розподілу Фішера для обраного рівня значущості α і зазначеної в дужках степені вільності. Якщо $C < C_L$, H_0 відкидається на користь H_1 .

Складність використання цього критерію полягає в чутливості критерію до порушення нормальності і вибору q , а також до розташування 0 і 1 на діагоналі A .

Застосування параметричного критерію Голдфільда і Квандта для оцінки неоднорідності часових рядів ускладнено необхідністю впорядкування спостережень за зростанням X_i , викидання ряду спостережень у середині ряду і побудови рівнянь регресії для перших і останніх спостережень. Критерій передбачає знаходження залишкових сум квадратів S_1 і S_2 відповідно для малих і великих значень X_i . Далі визначається значення $Q = S_2/S_1$. Якщо Q більше F -значення для обраного рівня значущості α , гіпотеза про однорідність дисперсії відкидається.

Моделювання методом Монте-Карло показало, що потужність критерію залежить від співвідношення між стандартним відхиленням фактора і його середнім, від числа значень, що опинилися всередині, і зменшується для значень X_i , розташованих близько до найбільшого середнього. Більшість розглянутих критеріїв потребує нормальності розподілу.

У багатьох випадках доцільним являється використання непараметричних критеріїв, заснованих на кількості піків у вихідних даних. Ідея використання непараметричних критеріїв на базі пікових характеристик призвела до необхідності розробки серії непараметричних процедур, позбавлених недоліків, пов'язаних з кореляціями і неоднорідностями часових рядів. Для визначення наявності автокореляції між вимірюваннями існує безліч критеріїв, таких як: d -статистика Дарбіна — Уотсона, метод Тейла, критерії, засновані на рекурентних залишках, тощо. Багато критеріїв чутливі до порушення нормальності розподілу даних.

Найбільш простими для застосування на практиці є критерії серій і критерії Дарбіна — Уотсона. Більшість динамічних рядів, які характеризують об'єкти технічних систем, економічні процеси, біржові ціни є нестационарними, однак їхні властивості можуть бути однорідні. Поводження таких рядів у різні часи відносно різних рівнів флуктуації багато в чому можуть бути схожою.

Вибір метрики або міри близькості являється вузловим моментом дослідження, від якого в основному залежить остаточний варіант розбиття об'єктів на класи за даного алгоритму розбиття. У кожному конкретному випадку цей вибір має здійснюватися у певний спосіб, у залежності від мети дослідження, фізичної і статистичної природи вектора спостережень X , апіорних відомостей про характер імовірнісного розподілу X .

Існує чотири стандартних критерії, яким має задовольняти міра подібності, щоб бути метрикою:

1) *Симетрія*. Дано два об'єкти — x та y ; відстань між ними задовольняє умові $d(x, y) = d(y, x) \geq 0$.

2) *Нерівність трикутника*. Дано три об'єкти — x, y, z ; відстані між ними задовольняють умові $d(x, y) \leq (d(x, z) + d(y, z))$.

3) *Розрізнення нетотожних об'єктів*. Дано два об'єкти — x та y ; якщо $d(x, z) \neq 0$, то $x \neq y$.

4) *Нерозрізненість ідентичних об'єктів*. Для двох ідентичних об'єктів x та x' $d(x, x') = 0$, тобто відстань між цими об'єктами дорівнює нулю.

Іншою важливою величиною у кластерному аналізі є відстань між цілими групами об'єктів.

Наочний приклад найбільш розповсюджених відстаней і мір близькості, що характеризують взаємне розташування окремих груп об'єктів, представлений у роботі [3, 15] (див. рис. 6.1).

Припустимо, що w_i — i -та група (клас, кластер) об'єктів, N_i — число об'єктів, які утворюють групу w_i , вектор μ_i — середнє арифметичне об'єктів, що входять у w_i (іншими словами, μ_i — «центр ваги» i -ї групи), а $q(w_l, w_m)$ — відстань між групами w_l та w_m .

Відстань між найближчими сусідами — це відстань між найближчими об'єктами кластерів:

$$q_{\min}(w_l, w_m) = \min d(x_i, x_j).$$

$$x_i^* w_l, x_j^* w_m$$

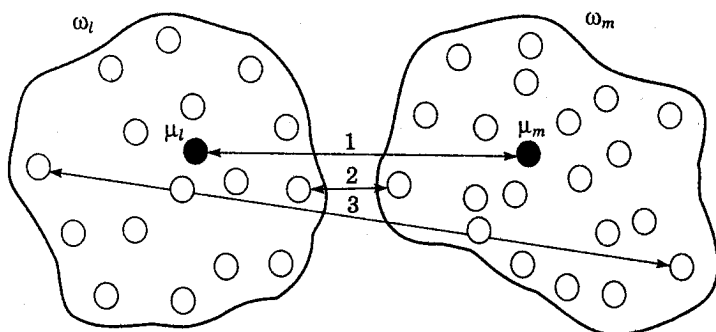


Рис. 6.1. Способи визначення відстані між кластерами ω_l і ω_m :
 1 — за центрами ваги; 2 — за найближчими об'єктами;
 3 — за найвіддаленішими об'єктами

Відстань між далекими сусідами — відстань між найвіддаленішими об'єктами кластерів:

$$q_{\max}(\omega_l, \omega_m) = \max_{x_i^* \omega_l, x_j^* \omega_m} d(x_i, x_j).$$

Відстань між «центрами ваги» дорівнює відстані між центральними точками кластерів.

Вибір тієї чи іншої міри відстані між кластерами впливає на вигляд виділюваних алгоритмами кластерного аналізу геометричних групувань об'єктів у просторі ознак. Так, алгоритми, засновані на відстані до найближчого сусіда, добре працюють у випадку групувань, що мають складну (наприклад, ланцюгову) структуру. Відстань до далекого сусіда застосовується, коли шукані групування утворюють у просторі ознак кулясті хмари. Проміжне місце займають алгоритми, що використовують відстані між «центрами ваги», відстані між середніми значеннями, які найкраще працюють у випадку групувань еліпсоїдної форми.

Розглянемо найбільш часто використовувані відстані і міри близькості в задачах кластерного аналізу.

Відстань Махаланобіса (загальний вигляд)

У випадку залежних компонент $x_1, x_2, \dots, x_n$ вектора спостережень X і їх різної значущості в рішенні питання класифікації зазвичай використовують узагальнену (зважену) відстань Махаланобіса, що задається формулою:

$$\rho_0(X_i, X_j) = \sqrt{(X_i - X_j)^T \Lambda^{-1} \Lambda (X_i - X_j)}$$

де Σ — коваріаційна матриця генеральної сукупності, з якої витягаються спостереження; Λ — деяка симетрична невід'ємно-визначена матриця «вагових» коефіцієнтів, яка найчастіше вибирається діагональною.

Три види відстаней, наведені нижче, є окремими випадками метрики ρ_0 .

Звичайна евклідова відстань

Визначається за формулою:

$$\rho_E(X_i, X_j) = \sqrt{\sum_{l=1}^k (x_{il} - x_{jl})^2},$$

де x_{il}, x_{jl} — величина l -тої компоненти у i -го (j -го) об'єкта ($l = 1, 2, \dots, k; i, j = 1, 2, \dots, n$).

Використання цієї відстані виправдане за таких умов:

☑ спостереження беруться з генеральних сукупностей, що мають багатовимірний нормальний розподіл з коваріаційною матрицею виду $\sigma^2 E_k$, тобто компоненти X взаємно незалежні і мають однакову дисперсію;

☑ компоненти вектора спостережень X однорідні за фізичним сенсом й однаково важливі для класифікації;

☑ простір ознак збігається з геометричним простором.

Природно, що з геометричної точки зору і змістовної інтерпретації евклідова відстань може виявитися безглуздою, якщо її ознаки мають різні одиниці вимірювання. Для приведення ознак до однакових одиниць звертаються до нормування кожної ознаки шляхом ділення центрованої величини на середнє квадратичне відхилення і переходять від матриці X до нормованої матриці з елементами

$$x_{il}^H = \frac{x_{il} - \bar{x}_l}{s_l},$$

де x_{il} — значення l -тої ознаки у i -го об'єкта; $\bar{x}_l$ — середнє арифметичне значення l -тої ознаки; $s_l = \sqrt{\frac{1}{n} \sum_i (x_{il} - \bar{x}_l)^2}$ — середнє квадратичне відхилення l -тої ознаки.

Однак ця операція може призвести до небажаних наслідків. Якщо кластери добре розділені за одною ознакою і не розділені за іншою, то після нормування дискримінувальні можливості першої ознаки будуть зменшені через збільшення «шумового» ефекту другої. Для додання ваги більш віддаленим об'єктам використовується квадрат евклідової відстані.

«Зважена» евклідова відстань

Визначається за формулою

$$\rho_{BE}(X_i, X_j) = \sqrt{\sum_{l=1}^k w_l (x_{il} - x_{jl})^2}$$

і застосовується у випадках, коли кожній компоненті вектора спостережень X вдається приписати деяку «вагу» w_l , пропорційну до ступеня важливості ознаки в задачі класифікації. Зазвичай приймають $0 \leq w_l \leq 1$, де $l = 1, 2, \dots, k$.

Визначення «ваги», як правило, пов'язане з додатковими дослідженнями, наприклад, з організацією опитування експертів і обробкою їхніх думок. Визначення w_l тільки за даними вибірки може призвести до помилкових висновків.

Хеммінгова відстань

Визначається за формулою

$$\rho_{BE}(X_i, X_j) = \sum_{l=1}^k |x_{il} - x_{jl}|$$

і використовується як міра розбіжності об'єктів, що задаються дихотомічними ознаками. Хеммінгова відстань дорівнює числу розбіжностей значень відповідних ознак у розглянутих i -му та j -му об'єктах.

У деяких задачах класифікації як міру близькості об'єктів можна використовувати деякі фізично змістовні параметри, що так чи інакше характеризують взаємовідносини між об'єктами. Наприклад, задачу класифікації галузей народного господарства з метою агрегування розв'язують на основі матриці міжгалузевого балансу.

У цій задачі об'єктом класифікації є галузь народного господарства, а матриця міжгалузевого балансу подана елементами

S_{ij} , що характеризують суму річних поставок i -тої галузі в j -ту в грошовому вираженні. Як міру близькості $\{r\}$ приймають симетризовану нормовану матрицю міжгалузевого балансу. З метою нормування грошове вираження поставок i -тої галузі j -ту заміняють часткою цих поставок стосовно всіх поставок i -тої галузі. Симетризацію нормованої матриці міжгалузевого балансу можна проводити, усереднивши близькість взаємних поставок між i -тою та j -тою галузями таким чином, що в цьому випадку $r_{ij} = r_{ji}$.

Як правило, розв'язання задач класифікації багатовимірних даних передбачає як попередній етап дослідження реалізацію методів, які дозволяють вибрати з компонент $x_1, x_2, \dots, x_k$ векторів x , що спостерігаються, порівняно невелике число найбільш істотних інформативних ознак, тобто зменшити розмірність простору, що спостерігається. З цією метою кожен з компонент $x_1, x_2, \dots, x_k$ розглядають як об'єкт, що підлягає класифікації. Після розбиття на невелике число однорідних у деякому сенсі груп для подальшого дослідження залишають по одному представнику від кожної групи. Водночас передбачається, що ознаки, які потрапили в одну групу, у певному сенсі пов'язані одна з одною і несуть інформацію про якусь одну властивість об'єкта.

У якості близькості між окремими ознаками зазвичай використовують різні характеристики ступеня їхньої корельованості, в першу чергу коефіцієнти кореляції. У ряді задач застосовуються й інші відстані (метрики). Вибір метрики визначається структурою ознакового простору і метою класифікації. До цього часу формалізувати цей етап задачі класифікації не виявляється можливим.

Якщо два об'єкти розрізняють за однією координатою (деяким виміром) використовується $\rho_c(x_i, x_j) = \max |x_i, x_j|$.

Для збільшення розбіжності між об'єктами використовується степенева відстань, що обчислюється за формулою:

$$\rho_p(X_i, X_j) = \left(\sum_{l=1}^k |x_{il} - x_{jl}|^p \right)^{\frac{1}{p}},$$

де r і p — параметри, які визначає користувач. Параметр p відповідає за поступове зважування різниць за окремими координатами; параметр r відповідає за прогресивне зважування великих відстаней між об'єктами. Якщо обидва параметри — r та p — дорівнюють 2, то ця відстань збігається з відстанню Евкліда.

6.5. Відстань між кластерами

Якщо об'єкти являють собою окремі кластери, відстані між ними визначаються обраною мірою. Однак коли зв'язуються разом кілька об'єктів, виникає питання: як слід визначити відстані між кластерами? Іншими словами, необхідно встановити правило об'єднання чи зв'язку для двох кластерів. Тут є різні можливості: зв'язати два кластери разом, коли будь-які два об'єкти у двох кластерах ближчі один до одного, ніж відповідні відстані зв'язку. У цьому випадку використовується «правило ближнього сусіда» для визначення відстані між кластерами; цей метод називається *методом одиночного зв'язку*. Це правило буде «волокнисті» кластери тобто кластери, «зчеплені» окремими елементами, які випадково опинились ближче інших одне до одного. Як альтернативу можна використовувати «сусідів» у кластерах, які знаходяться далі один від одного за всі інші пари об'єктів. Цей метод називається *методом повного зв'язку*. Існує безліч інших методів об'єднання кластерів, подібних до тих, що були розглянуті.

Нехай S_i — i -та група (клас, кластер), що складається з n_i об'єктів; $\bar{x}_i$ — середнє арифметичне векторних спостережень S_i , групи, тобто «центр ваги» i -тої групи; $\rho(S_i, S_m)$ — відстань між групами S_i і S_m .

Розглянемо найчастіше використовувані на практиці міри близькості і відстань між класами.

Одиночний зв'язок (метод найближчого сусіда). Як було зазначено вище, у цьому методі відстань між двома кластерами визначається відстанню між двома найбільш близькими об'єктами (найближчими сусідами) у різних кластерах. За цим правилом об'єкти, у певному сенсі, «нанизують» для формування кластерів, отже результуючі кластери мають тенденцію бути представленими довгими «ланцюжками».

$$\rho_{\min}(S_i, S_m) = \min_{x_i \in S_i, x_j \in S_m} \rho(x_i, x_j).$$

У методі одиночного зв'язку на першому кроці поєднуються два об'єкти, що мають між собою максимальну міру подібності. На наступному кроці до них приєднується об'єкт із максимальною мірою подібності з одним з об'єктів кластера. У такий спосіб процес продовжується далі. Отже, для включення об'єкта

в кластер потрібна максимальна подібність лише з одним членом кластера. Звідси і назва методу одиночного зв'язку: потрібен тільки один зв'язок, щоб приєднати об'єкт до кластера; зв'язок нового елемента з кластером визначається лише за одним з елементів кластера. Недоліком цього методу є утворення занадто великих, «довгастих» кластерів.

Повний зв'язок (метод найбільш віддалених сусідів). У цьому методі відстань між кластерами визначають найбільшою відстанню між двома будь-якими об'єктами в різних кластерах (тобто «найбільш віддаленими сусідами»). Тут міра подібності між об'єктом — кандидатом на включення в кластер і всіма членами кластера не може бути менше деякого граничного значення.

Цей метод зазвичай працює дуже добре, коли об'єкти входять насправді з реально різних «гаїв». Якщо ж кластер має подовжену форму або за природою є «ланцюговим», то цей метод не є придатним:

$$\rho_{\max}(S_l, S_m) = \max_{x_i \in S_l, x_j \in S_m} \rho(x_i, x_j).$$

Незважене попарне середнє. У цьому методі відстані між двома різними кластерами обчислюються як середня відстань між усіма парами об'єктів у них. Метод ефективний, коли об'єкти в дійсності формують різні «гаї», однак він працює однаково добре у випадках подовжених («ланцюгового типу») кластерів.

Зважене попарне середнє. Метод ідентичний до методу незваженого попарного середнього, за винятком того, що при обчисленні розмір відповідних кластерів (тобто число об'єктів, що міститься в них) використовується як ваговий коефіцієнт. Тому пропонований метод має застосовуватися (скоріш навіть, ніж попередній), коли передбачаються нерівні розміри кластерів.

У незваженому центроїдному методі визначається відстань між центрами ваги.

Зважений центроїдний метод (медіана) ідентичний до попереднього, за винятком того, що при обчисленнях використовуються ваги для врахування різниці між розмірами кластерів (тобто числами об'єктів у них). Тому за наявності (або за припущення про наявність) значних відмінностей у розмірах кластерів перевагу слід надати цьому методу.

Метод Варда відрізняється від інших, тому що для оцінки відстаней між кластерами він використовує методи дисперсійного

аналізу. Метод мінімізує суму квадратів (SS) для будь-яких двох (гіпотетичних) кластерів, що можуть бути сформовані на кожному кроці. У цілому метод виявляється дуже ефективним, однак він прагне створювати кластери малого розміру.

Ідея ще одного агломеративного методу — *методу Уорда* — полягає в тому, щоб проводити об'єднання, яке дає мінімальне збільшення внутрішньогрупової суми квадратів відхилень. Помічено, що метод Уорда призводить до утворення кластерів приблизно рівних розмірів і таких, що мають форму гіперсфер.

Ітеративний *метод групування k-середніх (k-means clustering)* працює безпосередньо з об'єктами, а не з матрицею подібності. У методі *k-середніх* об'єкт належить до того класу, відстань до якого мінімальна. Відстань розуміється як евклідова відстань, тобто об'єкти розглядаються як точки евклідового простору. Відстань між об'єктом і класом є відстань між об'єктом і центром класу.

Метод *k-середніх* передбачає такий алгоритм дії:

1) задається деяке розбиття даних на кластери (число кластерів визначається користувачем); обчислюються центри ваги кластерів;

2) відбувається переміщення точок: кожна точка поміщається у найближчий до неї кластер;

3) обчислюються центри ваги нових кластерів;

4) кроки 2, 3 повторюються, доки не буде знайдена стабільна конфігурація (тобто кластери припинять змінюватися) або число ітерацій не перевищить задане користувачем. Підсумкова конфігурація і є шуканою.

Академіком А. М. Колмогоровим була запропонована «узагальнена відстань» між класами, що як окремі випадки містять усі розглянуті вище види відстаней. «Узагальнена відстань» заснована на понятті так званого «узагальненого середнього», а точніше — степеневого середнього, і визначається за формулою:

$$\rho_{об}(S_l, S_m) = \left[\frac{1}{n_l n_m} \sum_{x_i \in S_l} \sum_{x_j \in S_m} \rho^r(x_i, x_j) \right]^{\frac{1}{r}}.$$

Можна показати, що за умови $r \rightarrow \infty$ $\rho_{об}(S_l, S_m) = \rho_{\max}(S_l, S_m)$, за умови $r \rightarrow -\infty$ $\rho_{об}(S_l, S_m) = \rho_{\min}(S_l, S_m)$, за умови $r = 1$ $\rho_{об}(S_l, S_m) = \rho(S_l, S_m)$.

З формули (7.9) випливає, що якщо $S(m, q) = S_m \cup S_q$ — група елементів, отримана шляхом об'єднання кластерів S_m і S_q , то узагальнена відстань між кластерами S_l і $S(m, q)$ визначається за формулою

$$\rho_{06}(S_l, S_{(m,q)}) = \left\{ \frac{n_m [\rho_{06}(S_l, S_m)]^r + n_q [\rho_{06}(S_l, S_q)]^r}{n_m + n_q} \right\}^{\frac{1}{r}}. \quad (7.10)$$

Відстань між групами елементів особливо важлива в так званих агломеративних ієрархічних кластерах-процедурах, оскільки принцип роботи таких алгоритмів полягає в послідовному об'єднанні спочатку найближчих елементів, а потім і цілих груп, усе більш і більш віддалених один від одного елементів.

При цьому відстань між класами S_l і $S(m, q)$, що є об'єднанням двох інших класів S_m і S_q , можна визначити за формулою:

$$\rho_{l,(m,q)} = \rho(S_l, S_m) = \alpha \rho_{lm} + \beta \rho_{lq} + \gamma \rho_{mq} + \delta |\rho_{lm} - \rho_{lq}|, \quad (6.1)$$

де $\rho_{lm} = \rho(S_l, S_m)$; $\rho_{lq} = \rho(S_l, S_q)$; $\rho_{mq} = \rho(S_m, S_q)$ — відстані між класами S_l , S_m і S_q ; α , β , γ і δ — числові коефіцієнти, значення яких визначають специфіку процедури, її алгоритм.

Наприклад, при $\alpha = \beta = -\delta = 1/2$ і $\gamma = 0$ приходимо до відстані, побудованої за принципом «найближчого сусіда». При $\alpha = \beta = \delta = 1/2$ і $\gamma = 0$ відстань між класами визначається за принципом «далекого сусіда», як відстань між двома найвіддаленішими елементами цих класів. І нарешті, при

$$\alpha = \frac{n_m}{n_m + n_q}, \quad \beta = \frac{n_q}{n_m + n_q}, \quad \gamma = \delta = 0$$

співвідношення (6.1) призводить до відстані ρ між класами, обчисленої як середнє з відстаней між усіма парами елементів, один із яких береться з одного класу, а інший — з іншого класу.

6.6. Функціонали якості розбиття

Оскільки існує велика кількість різних способів розбиття заданої сукупності елементів на класи, тому є актуальним порівняльний аналіз якості цих способів. З цією метою вводиться

поняття функціонала якості розбиття $Q(S)$, визначеного на множині всіх можливих розбиттів [41, 42].

Найкращім виявляється таке розбиття S^* , за якого досягається екстремум обраного функціонала якості. Слід зазначити, що вибір того чи іншого функціонала якості розбиття, як правило, спирається на емпіричні міркування.

Розглянемо деякі найбільш розповсюджені функціонали якості розбиття. Нехай дослідженням обрана метрика ρ у просторі X і $S = (S_1, S_2, \dots, S_p)$ деяка фіксоване розбиття спостережень $X_1, X_2, \dots, X_n$ на задане число p класів $S_1, S_2, \dots, S_p$.

Існують такі характеристики функціонала якості:

☑ сума внутрішньокласових дисперсій

$$Q_1(S) = \sum_{l=1}^p \sum_{x_i \in S_l} \rho^2(x_i, \bar{x}_l);$$

☑ сума попарних внутрішньокласових відстаней між елементами

$$Q_2(S) = \sum_{l=1}^p \sum_{x_i, x_j \in S_l} \rho^2(x_i, x_j) \quad \text{чи} \quad Q'_2(S) = \sum_{l=1}^p \frac{1}{n_l} \sum_{x_i, x_j \in S_l} \rho^2(x_i, x_j).$$

$Q_1(S)$ і $Q_2(S)$ широко використовуються в задачах кластерного аналізу для порівняння якості процедур розбиття;

☑ узагальнена внутрішньокласова дисперсія

$$Q_3(S) = \det \left(\sum_{l=1}^p n_l W_l \right),$$

де $\det A$ — визначник матриці A ; W_l — вибіркова коваріаційна матриця класу S_l , елементи якої визначаються за формулою

$$w_{qm}(l) = \frac{1}{n_l} \sum_{x_i \in S_l} (x_{iq} - \bar{x}_q)(x_{im} - \bar{x}_m), \quad q, m = 1, 2, \dots, k,$$

де x_{iq} — q -а компонента багатовимірного спостереження x_i ; $\bar{x}_q$ — середнє значення q -ї компоненти, обчислене за спостереженнями l -го класу.

Якість розбиття можна визначати за іншим видом узагальненої дисперсії, у якій операція підсумовування W_l замінена операцією множення

$$Q_4(S) = \prod_{l=1}^p (\det W_l)^{n_l}.$$

Зазначимо, що функціонали $Q_3(S)$ і $Q_4(S)$ зазвичай використовують при вирішенні питання: чи не зосереджені спостереження, розбиті на класи, у просторі, розмір якого менший за k .

Розглянемо принцип роботи деяких ієрархічних кластер-процедур. Ієрархічні (деревоподібні) процедури є найбільш поширеними алгоритмами кластерного аналізу за їх реалізацією на ЕОМ. Вони поділяються на *агломеративні* та *ітеративні дивізивні*.

Більшість цих алгоритмів виходить із матриці відстаней (подібності).

До недоліків ієрархічних процедур слід віднести складність обчислювальної реалізації. Алгоритми вимагають на кожному кроці матриці обчислення відстаней, а отже, великої ємності машинної пам'яті і багато часу. У зв'язку з цим реалізація таких алгоритмів при кількості спостережень понад декілька сотень недоцільна, а в ряді випадків й неможлива. Наочно це видно з прикладу, наведеного в роботі [3].

На першому кроці кожне спостереження X_i ($i = 1, 2, \dots, n$) розглядається як окремий кластер. Надалі на кожному кроці роботи алгоритму відбувається об'єднання двох найближчих кластерів і, з урахуванням прийнятої відстані, за формулою перераховується матриця відстаней, розмірність якої знижується на одиницю. Робота алгоритму закінчується, коли всі спостереження об'єднані в один клас. Результати класифікації подаються у вигляді дендрограми.

Приклад 6.1

Проведемо класифікацію об'єктів ($n = 6$) за двома ознаками:

№	1	2	3	4	5	6
X_{i1}	5	6	5	10	11	10
X_{i2}	10	12	13	9	9	7

Розв'язання

Скористаємося агломеративним ієрархічним алгоритмом класифікації. У якості відстані між об'єктами візьмемо звичайну евклідову відстань. Тоді відповідно до формули (6.1) відстань між першим і другим об'єктами становитиме

$$\rho_{12} = \sqrt{(5-6)^2 + (10-12)^2} = 2,24,$$

а між першим і третім об'єктами —

$$\rho_{13} = \sqrt{(5-5)^2 + (10-13)^2} = 3.$$

Очевидно, що $\rho_{11} = 0$.

Аналогічно знаходимо відстані між шістьма об'єктами і будемо матрицю відстаней

$$R_1 = \{\rho(x_i, x_j)\} = \begin{pmatrix} 0 & 2,24 & 3 & 5,10 & 6,08 & 5,83 \\ 2,24 & 0 & 1,41 & 5 & 5,38 & 6,40 \\ 3 & 1,41 & 0 & 6,40 & 7,21 & 7,81 \\ 5,10 & 5 & 6,40 & 0 & 1 & 2 \\ 6,08 & 5,83 & 7,21 & 1 & 0 & 2,24 \\ 5,83 & 6,40 & 7,81 & 2 & 2,24 & 0 \end{pmatrix}.$$

З матриці відстаней випливає, що четвертий і п'ятий об'єкти найбільш близькі $\rho_{4,5} = 1,00$ і тому поєднуються в один кластер. Після об'єднання об'єктів маємо п'ять кластерів:

Номер кластера	1	2	3	4	5
Склад кластера	(1)	(2)	(3)	(4,5)	(6)

Відстань між кластерами визначимо за принципом «найближчого сусіда», скориставшись формулою перерахування (6.1). Так, відстань між об'єктом S_i і кластером $S_{(4,5)}$ становитиме

$$\begin{aligned} \rho_{1,(4,5)} = \rho(S_1, S_{(4,5)}) &= \frac{1}{2}\rho_{14} + \frac{1}{2}\rho_{15}|\rho_{14} - \rho_{15}| = \\ &= \frac{1}{2}[5,10 + 6,08] - \frac{1}{2}[5,10 - 6,08] = 5,10. \end{aligned}$$

Таким чином, відстань $\rho_{1,(4,5)}$ дорівнює відстані від об'єкта 1 до найближчого до нього об'єкта, що входить у кластер $S_{(4,5)}$, тобто $\rho_{1,(4,5)} = \rho_{1,4} = 5,10$. Тоді матриця відстаней

$$R_2 = \begin{pmatrix} 0 & 2,24 & 3 & 5,10 & 5,83 \\ 2,24 & 0 & 1,41 & 5 & 6,40 \\ 3 & 1,41 & 0 & 6,40 & 7,81 \\ 5,10 & 5 & 6,40 & 0 & 2 \\ 6,08 & 5,83 & 7,21 & 1 & 2,24 \\ 5,83 & 6,40 & 7,81 & 2 & 0 \end{pmatrix}.$$

Об'єднаємо другий і третій об'єкти, що мають найменшу відстань $\rho_{23} = 1,41$. Після об'єднання об'єктів маємо чотири кластери: $S_{(1)}$, $S_{(2,3)}$, $S_{(4,5)}$, $S_{(6)}$.

Знову знайдемо матрицю відстаней. Для того, щоб визначити відстань до кластера $S_{(2,3)}$, скористаємося матрицею відстаней R_2 . Наприклад, відстань між кластерами $S_{(4,5)}$ та $S_{(2,3)}$ дорівнює

$$\begin{aligned}\rho_{(4,5),(2,3)} &= \frac{1}{2}\rho_{(4,5),2} + \frac{1}{2}\rho_{(4,5),3} - \frac{1}{2}|\rho_{(4,5),2} - \rho_{(4,5),3}| = \\ &= \frac{5}{2} + \frac{6,40}{2} - \frac{1,40}{2} = 5.\end{aligned}$$

Провівши аналогічні розрахунки, отримаємо

$$R_3 = \begin{pmatrix} 0 & 2,24 & 5,10 & 5,83 \\ 2,24 & 0 & 5 & 6,40 \\ 5,10 & 5 & 0 & 2 \\ 5,83 & 6,40 & 2 & 0 \end{pmatrix}.$$

Об'єднаємо кластери $S_{(4,5)}$ і $S_{(6)}$, відстань між якими, відповідно до матриці R_3 , найменша $\rho_{(4,5),6} = 62$. У результаті отримаємо три кластери $S_{(1)}$, $S_{(2,3)}$ та $S_{(4,5,6)}$.

Матриця відстаней матиме вигляд:

$$R_4 = \begin{pmatrix} 0 & 2,24 & 5,10 \\ 2,24 & 0 & 5 \\ 5,10 & 5 & 0 \end{pmatrix}.$$

Об'єднаємо тепер кластери S_1 і $S_{(2,3)}$, відстань між якими $\rho_{1,(2,3)} = 2,24$. У результаті отримаємо два кластери: $S_{(1,2,3)}$ та $S_{(4,5,6)}$, відстань між якими, знайдена за принципом «найближчого сусіда», становить $\rho_{(1,2,3),(4,5,6)} = 5$.

Результати ієрархічної класифікації об'єктів представлені на рис. 6.2 у вигляді дендрограми, де наводяться відстані між поєднуваними на даному етапі кластерами (об'єктами). У прикладі перевагу слід надати передостанньому етапу класифікації, коли всі об'єкти об'єднані в два кластери $S_{(1,2,3)}$ і $S_{(4,5,6)}$.

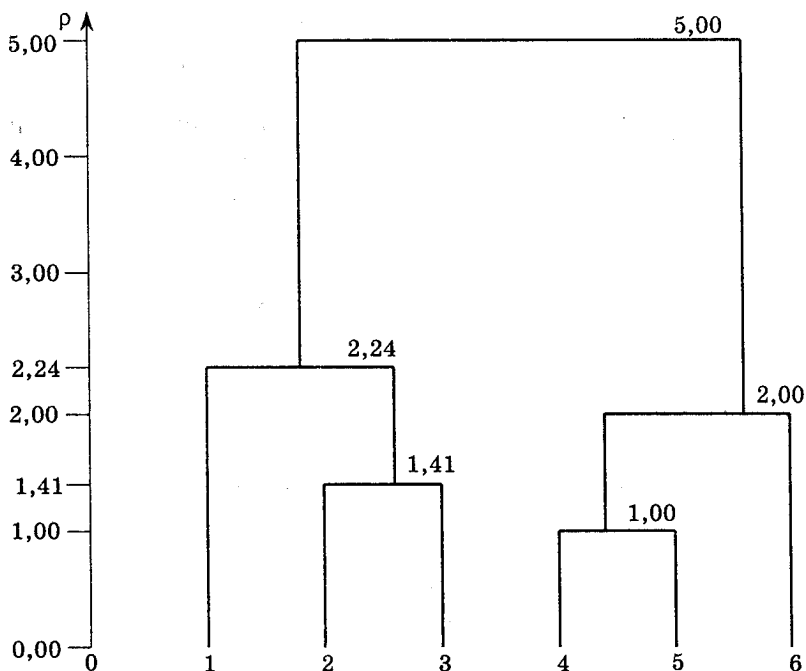


Рис. 6.2. Дендрограма

6.7. Використання ППП для розв'язання задачі кластеризації

Для розв'язання задачі кластеризації використовуємо ППП «STATISTICA» і «STATGRAPHICS Plus».

У ППП «STATISTICA» реалізовано декілька мір подібності і правил об'єднання кластерів. Для кожної конкретної задачі користувачеві на підставі глибокого вивчення об'єктів кластеризації і професійної підготовки необхідно самому приймати рішення про вибір міри подібності і правила об'єднання кластерів. Іноді доцільно використовувати декілька мір подібності та правил об'єднання кластерів, і за результатом обрати найбільш прийнятний варіант. Зазначимо, що в ППП вихідні дані перед початком класифікації зазвичай стандартизуються.

У ППП «STATISTICA» доступні такі міри подібності об'єктів: евклідова метрика, квадрат евклідової метрики, манхеттенська відстань, чи «відстань міських кварталів», метрика Чебишева,

метрика Мінковського, пірсонівський коефіцієнт кореляції, коефіцієнт співзв'язності.

У ППП «STATISTICA» реалізовані такі методи кластеризації: агломеративні, по'єднувальні (групи дерев), методи двобічного по'єднання, а також метод k -середніх.

У зазначеному ППП можна вибрати такі правила ієрархічного об'єднання кластерів:

☒ **Single linkage** — метод одиночного зв'язку;

☒ **Complete linkage** — метод повного зв'язку;

☒ **Unweighted pair group average** — незважений метод «середнього зв'язку»;

☒ **Weighted pair group average** — зважений метод «середнього зв'язку»;

☒ **Weighted centroid pair group (median)** — зважений центроїдний метод;

☒ **Ward method** — метод Уорда.

Ці алгоритми розрізняються правилами об'єднання об'єктів у кластери.

Як приклад застосування кластерного аналізу розглянемо класифікацію і розпізнавання ознак у діагностиці захворювань. Для виконання цієї роботи скористаємося ППП «STATISTICA 6.0».

Вихідною інформацією слугують дані клініки, у яких зареєстровані результати обстеження 103 осіб із установленим діагнозом: група 1 — гангренозний апендицит (28 спостережень), група 2 — флегмонозний апендицит (25 спостережень), група 3 — катаральний апендицит (26 спостережень) і група 4 — не підтверджений діагноз (24 спостереження). Вихідними ознаками служили 8 симптомів, охарактеризованих у табл. 6.1.

Таблица 6.1

Дані клінічного обстеження

№ групи	x1	x2	x3	x4	x5	x6	x7	x8
1	2	3	1	2	1	2	2	2
1	2	2	2	2	1	2	0	2
1	2	3	1	3	1	2	2	2
1	2	2	3	1	1	0	2	2
1	2	3	2	2	1	2	2	0
1	2	3	1	3	0	0	2	2
1	2	2	2	2	1	2	0	2

Продовження табл. 6.1

№ групи	x1	x2	x3	x4	x5	x6	x7	x8
1	2	2	1	3	1	2	2	2
1	1	2	2	3	1	2	2	2
1	2	3	2	2	1	2	2	2
1	2	3	1	3	1	2	2	0
1	2	3	2	2	1	2	2	2
1	2	2	1	3	0	2	0	2
1	2	3	2	2	1	0	2	2
1	2	4	2	2	1	2	2	2
1	2	2	1	3	1	2	2	2
1	2	3	3	2	1	2	0	2
1	1	1	2	2	0	2	2	2
1	2	3	2	3	1	2	2	2
1	2	1	1	3	1	0	2	2
1	2	3	3	2	1	2	2	2
1	2	3	2	3	1	2	2	0
1	2	2	1	2	1	2	0	2
1	2	3	2	2	0	2	2	2
1	2	3	1	2	1	2	2	2
1	2	3	2	3	1	2	2	2
1	2	3	1	3	1	2	2	2
1	2	3	1	2	1	2	2	2
2	2	3	1	2	1	2	2	0
2	1	4	2	1	0	2	0	2
2	2	3	1	3	1	0	2	2
2	1	4	2	2	1	2	2	2
2	2	4	1	2	0	2	2	2
2	2	4	2	2	1	2	0	0
2	1	2	1	2	1	2	2	2
2	2	4	2	3	0	0	2	2
2	1	3	1	1	1	2	0	2
2	2	4	1	2	1	2	2	0
2	2	4	1	3	0	2	2	2
2	1	2	1	2	1	0	0	2

Продовження табл. 6.1

№ групи	x1	x2	x3	x4	x5	x6	x7	x8
2	2	3	1	3	1	2	2	0
2	1	4	1	1	1	2	2	2
2	2	4	1	2	0	2	0	2
2	2	3	1	2	1	0	2	0
2	1	4	2	2	1	2	2	2
2	2	4	1	3	0	2	2	2
2	2	3	1	2	1	2	0	0
2	1	4	2	1	1	0	2	2
2	2	3	1	2	0	2	2	2
2	2	4	1	2	1	2	2	2
2	2	4	2	2	1	2	2	2
2	2	4	2	3	1	0	2	2
2	1	3	2	2	1	2	2	2
3	1	3	1	2	1	0	2	2
3	2	4	1	1	0	2	0	0
3	2	3	1	2	1	0	2	2
3	2	4	2	2	1	2	0	0
3	1	2	1	1	0	0	2	2
3	2	3	1	3	1	2	2	0
3	2	4	1	2	1	2	2	2
3	2	1	1	1	1	2	2	0
3	1	4	1	2	0	0	0	2
3	2	1	2	2	1	2	2	0
3	2	3	1	1	1	2	0	2
3	2	4	1	2	1	0	0	0
3	1	3	1	1	0	2	2	0
3	2	4	1	2	1	0	2	2
3	2	3	2	2	1	2	2	2
3	1	4	1	1	0	0	2	0
3	2	3	1	2	1	2	2	0
3	2	4	2	2	1	2	0	2
3	2	3	1	3	0	0	2	2
3	2	4	1	2	1	0	0	0

Закінчення табл. 6.1

№ групи	x1	x2	x3	x4	x5	x6	x7	x8
3	1	3	1	1	1	2	2	0
3	2	3	1	2	0	2	2	2
3	2	4	1	2	1	2	2	2
3	1	4	2	1	1	2	2	0
3	2	3	1	2	1	2	2	0
3	1	4	1	2	1	2	2	0
4	1	2	1	1	0	0	0	0
4	1	1	2	1	0	0	0	0
4	1	3	1	1	1	0	0	0
4	2	1	1	2	0	0	0	0
4	1	2	1	1	0	0	0	0
4	1	1	1	1	0	0	0	0
4	1	2	1	1	0	0	0	0
4	1	1	2	1	1	0	0	0
4	1	2	1	2	0	0	0	0
4	2	1	1	1	0	0	0	0
4	1	2	1	2	1	0	0	0
4	1	2	1	2	1	0	0	0
4	1	1	1	2	0	0	0	0
4	1	1	2	1	0	0	2	0
4	1	4	1	1	0	0	0	0
4	1	3	1	1	0	0	0	0
4	2	1	1	2	1	0	0	0
4	1	4	1	1	0	0	0	0
4	1	2	1	1	0	0	0	0
4	1	1	1	2	1	0	0	0
4	1	2	1	1	0	0	0	0
4	1	1	2	1	0	0	0	0
4	2	1	1	1	0	0	2	0
4	1	2	1	1	0	0	0	0

Уведемо представлені дані в електронну таблицю ППП «STATISTICA». Збережемо їх у файлі з ім'ям **appendix**.

У ППП обираємо **Statistics → Multivariate Exploratory Techniques → Cluster Analysis**. Отримуємо вікно діалогу вибору використовуваного аналізу (рис. 6.3).

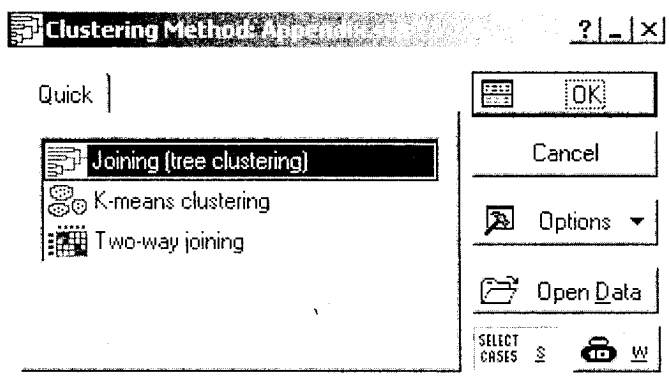


Рис. 6.3. Вікно вибору використовуваного аналізу

Розглянемо метод поєднання **joining (tree clustering)**, для чого виберемо його і натиснемо кнопку **OK**. У вікні, що з'явився, оберемо пункт **Variables** (рис. 6.4), де зазначимо змінні (рис. 6.5), на підставі яких будемо проводити кластеризацію.

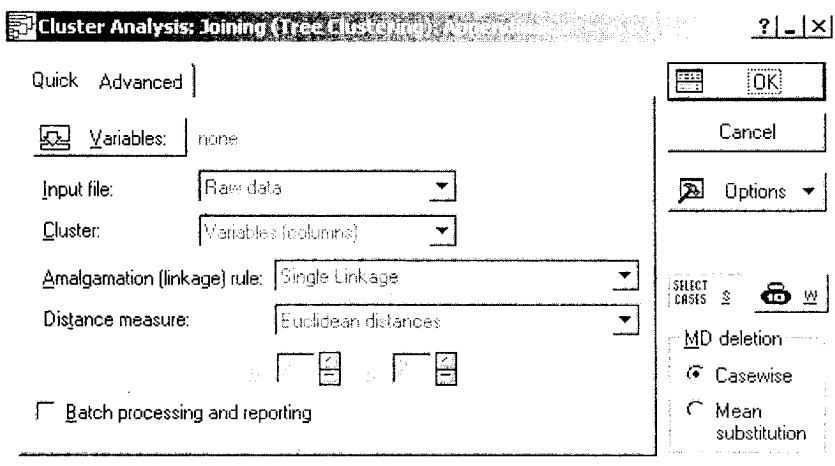


Рис. 6.4. Головне вікно класифікації методом **joining (tree clustering)**

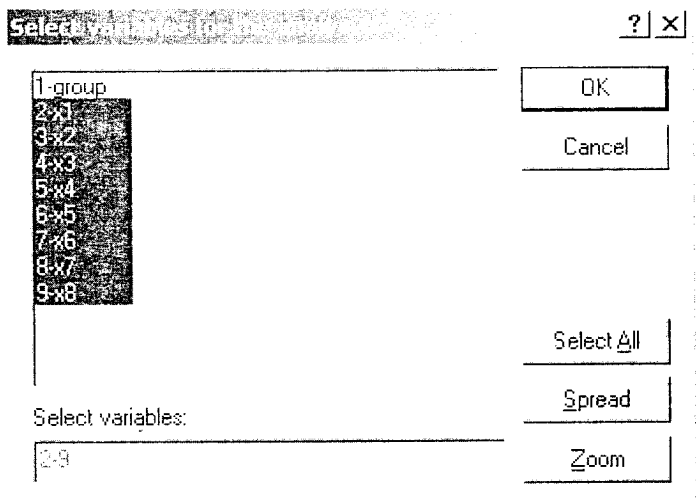


Рис. 6.5. Вікно вибору змінних для проведення кластеризації

Серед пропонованих ППП «STATISTICA» правил ієрархічного об'єднання кластерів — **Amalgamation (linkage) rule** — метод одиночного зв'язку **Single linkage**. Після натискання кнопки **OK** з'явиться вікно результатів кластеризації (рис. 6.6).

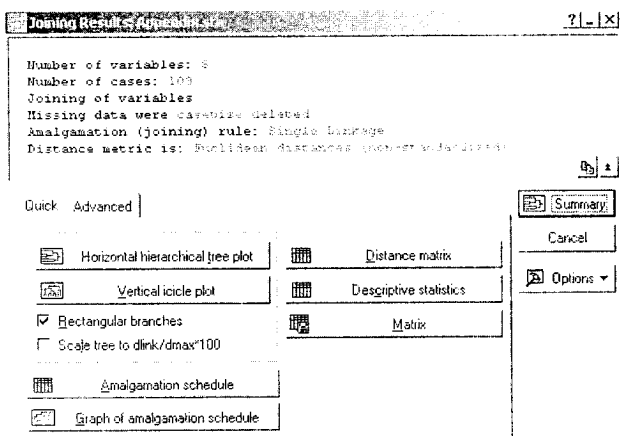


Рис. 6.6. Вікно результатів кластеризації методом joining (tree clustering)

Кнопка **Horizontal hierarchical plot** дозволяє отримати горизонтальну деревоподібну діаграму (рис. 6.7), кнопка **Vertical icicle plot** — вертикальну деревоподібну діаграму (рис. 6.8).

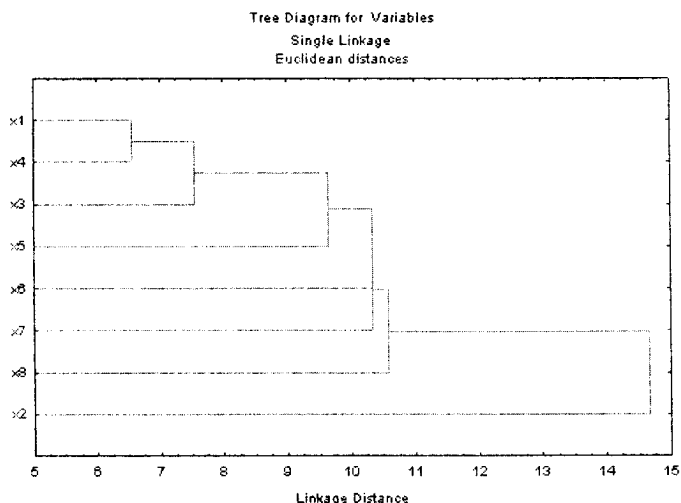


Рис. 6.7. Горизонтальна деревоподібна діаграма

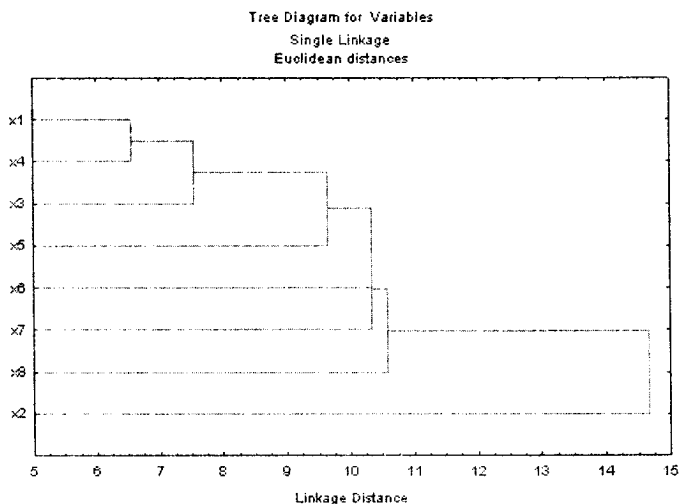


Рис. 6.8. Вертикальна деревоподібна діаграма

З рис. 6.7 і 6.8 видно, що деревоподібні діаграми починаються з кожного об'єкта в класі (вихідні змінні). Потім поступово (дуже малими кроками) «послаблюється» критерій про те, які об'єкти є унікальними, а які ні. Іншими словами, знижується межа, що відноситься до рішення про об'єднання двох чи більше об'єктів в один кластер. У результаті разом пов'язуються усе більше і більше число об'єктів і агрегуються усе більше кластерів, які складаються з елементів, що все сильніше розрізняються. На останньому кроці всі об'єкти поєднуються.

На таких діаграмах осі являють собою відстань об'єднання. Так, для кожного вузла в графі (місце формування нового кластера) можна побачити величину відстані, для якої відповідні елементи пов'язуються в новий єдиний кластер.

Дані діаграми, в результаті успішного аналізу методом об'єднання, дають можливість виявити кластери (галузі) і інтерпретувати їх.

Кнопка **Amalgamation schedule** дозволяє відобразити послідовність об'єднання змінних у кластери із зазначенням відстані об'єднання (рис. 6.9).

linkage distance	Amalgamation Schedule (Appendix.sta)							
	Single Linkage							
	Euclidean distances							
	Obj. No. 1	Obj. No. 2	Obj. No. 3	Obj. No. 4	Obj. No. 5	Obj. No. 6	Obj. No. 7	Obj. No. 8
6,557438	x1	x4						
7,549834	x1	x4	x3					
9,643651	x1	x4	x3	x5				
10,34408	x1	x4	x3	x5	x6			
10,34408	x1	x4	x3	x5	x6	x7		
10,58300	x1	x4	x3	x5	x6	x7	x8	
14,66288	x1	x4	x3	x5	x6	x7	x8	x2

Рис. 6.9. Послідовність об'єднання змінних у кластери

Кнопка **Graph of amalgamation schedule** дозволяє графічно відобразити моменти об'єднання змінних у кластери (див. рис. 6.10).

Кнопка **Distance matrix** виводить на екран матрицю відстаней між центрами кластерів (див. рис. 6.11).

Кнопка **Descriptive statistics** виводить на екран матрицю статистики — середніх і стандартних відхилень (див. рис. 6.12).

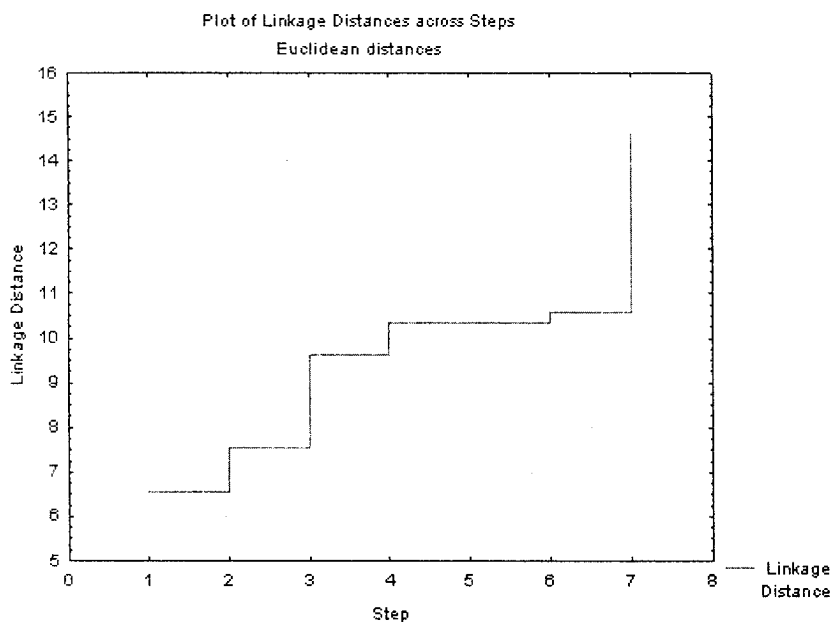


Рис. 6.10. Графік об'єднання змінних у кластери

Variable	Euclidean distances (Appendix.sta)							
	x1	x2	x3	x4	x5	x6	x7	x8
x1	0,0	15,9	7,5	6,6	11,4	10,3	10,3	11,3
x2	15,9	0,0	18,6	14,7	24,3	20,3	20,2	21,3
x3	7,5	18,6	0,0	10,4	9,6	10,6	10,8	10,6
x4	6,6	14,7	10,4	0,0	14,6	12,7	11,7	12,7
x5	11,4	24,3	9,6	14,6	0,0	10,7	11,3	11,3
x6	10,3	20,3	10,6	12,7	10,7	0,0	11,1	11,7
x7	10,3	20,2	10,8	11,7	11,3	11,1	0,0	11,1
x8	11,3	21,3	10,6	12,7	11,3	11,7	11,1	0,0

Рис. 6.11. Матриця відстаней між класами

Кнопка **Matrix** виводить на екран повну статистику кластеризації (див. рис. 6.13).

Variable	Means and Standard Deviations (Appendix.sta)	
	Mean	Std.Dev.
x1	1,621359	0,487420
x2	2,805825	1,019979
x3	1,359223	0,539687
x4	1,902913	0,693309
x5	0,650486	0,479148
x6	1,165049	0,991108
x7	1,223301	0,979516
x8	1,087379	1,001046

Рис. 6.12. Матриця статистики

	Appendix.sta							
	1 x1	2 x2	3 x3	4 x4	5 x5	6 x6	7 x7	8 x8
x1	0,00000	15,93738	7,54983	6,55744	11,40175	10,34408	10,34408	11,26943
x2	15,93738	0,00000	18,62794	14,66288	24,26992	20,32240	20,22375	21,26380
x3	7,54983	18,62794	0,00000	10,39230	9,64365	10,58301	10,77033	10,58301
x4	6,55744	14,66288	10,39230	0,00000	14,59452	12,72792	11,74734	12,72792
x5	11,40175	24,26992	9,64365	14,59452	0,00000	10,72381	11,26943	11,26943
x6	10,34408	20,32240	10,58301	12,72792	10,72381	0,00000	11,13553	11,66190
x7	10,34408	20,22375	10,77033	11,74734	11,26943	11,13553	0,00000	11,13553
x8	11,26943	21,26380	10,58301	12,72792	11,26943	11,66190	11,13553	0,00000
Means	1,62136	2,80583	1,35922	1,90291	0,65049	1,16505	1,22330	1,08738
Std.Dev.	0,48742	1,01998	0,53969	0,69331	0,47915	0,99111	0,97952	1,00105
No.Cases	103,00000							
Matrix	3,00000							

Рис. 6.13. Повна статистика кластеризації

Для отримання більш точної картини розбиття на кластери, проведемо кластеризацію вихідної вибірки за допомогою методу *k*-середніх — ***k*-means clustering**. Для цього повернемося до діалогу вибору методу кластеризації (рис. 6.3), виберемо метод ***k*-means clustering** і натиснемо кнопку **ОК**. При цьому з'явиться у вікні головне вікно класифікації методом ***k*-means clustering** (рис. 6.14).

Аналогічно до попереднього методу вкажемо змінні, на підставі яких будемо проводити кластеризацію (рис. 6.15).

Після підтвердження вибору змінних і натискання кнопки **ОК** з'явиться вікно результатів кластеризації (рис. 6.16).

Кнопка **Summary: Cluster means & Euclidean distance** дозволяє вивести таблиці, у першій з яких зазначені середні для

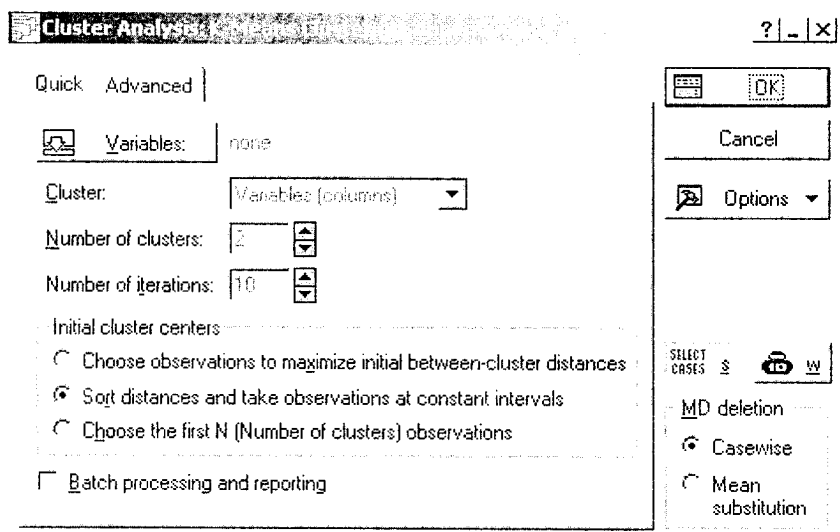


Рис. 6.14. Головне вікно класифікації методом k -means clustering

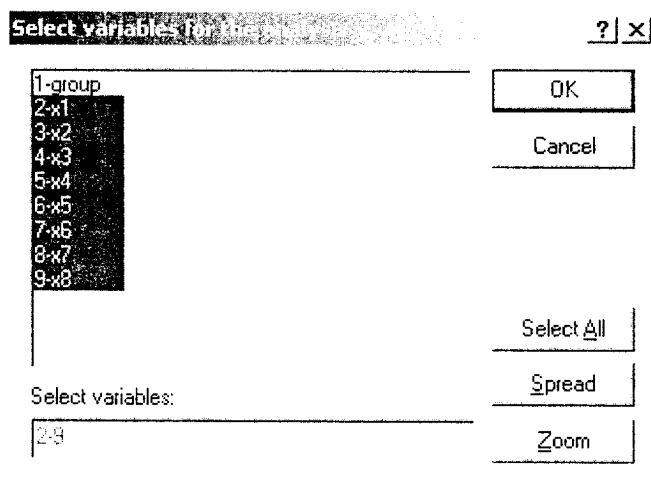


Рис. 6.15. Вікно вибору змінних для проведення кластеризації

кожного кластера (усереднення проводиться всередині кластера) (рис. 6.17), у другій зазначені евклідови відстані і квадрати евклідових відстаней між кластерами (рис. 6.18).

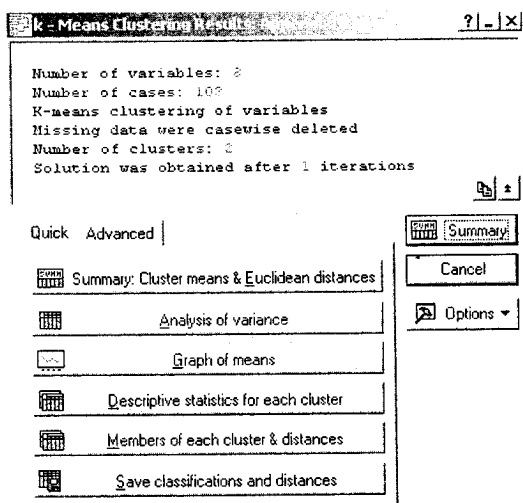


Рис. 6.16. Вікно результатів кластеризації методом *k*-means clustering

Case ID	Cluster Means (Appendix sta)	
	Cluster No. 1	Cluster No. 2
C_1	3,000000	1,714286
C_2	2,000000	1,571429
C_3	3,000000	1,857143
C_4	2,000000	1,571429
C_5	3,000000	1,571429
C_6	3,000000	1,428571
C_7	2,000000	1,571429
C_8	2,000000	1,857143
C_9	2,000000	1,857143
C_10	3,000000	1,857143
C_11	3,000000	1,571429
C_12	3,000000	1,857143
C_13	2,000000	1,428571
C_14	3,000000	1,571429
C_15	4,000000	1,857143
C_16	2,000000	1,857143
C_17	3,000000	1,714286
C_18	1,000000	1,571429

Рис. 6.17. Середні відстані між кластерами

Cluster Number	Euclidean Distances between Clusters (Appendix.sta)				
	Distances below diagonal				
	Squared distances above diagonal				
	No. 1	No. 2			
No. 1	0,000000	3,215178			
No. 2	1,793092	0,000000			

Рис. 6.18. Евклідова відстань між кластерами

Кнопка **Analysis of variation** дозволяє переглянути таблицю дисперсійного аналізу (рис. 6.19).

Case ID	Analysis of Variance (Appendix.sta)					
	Between SS	df	Within SS	df	F	signif. p
C_1	1,44643	1	1,428571	6	6,07500	0,048804
C_2	0,16071	1	3,714286	6	0,25962	0,628577
C_3	1,14286	1	2,857143	6	2,40000	0,172308
C_4	0,16071	1	5,714286	6	0,16875	0,695493
C_5	1,78571	1	3,714286	6	2,88462	0,140346
C_6	2,16071	1	7,714286	6	1,68056	0,242476
C_7	0,16071	1	3,714286	6	0,25962	0,628577
C_8	0,01786	1	2,857143	6	0,03750	0,852840
C_9	0,01786	1	2,857143	6	0,03750	0,852840
C_10	1,14286	1	0,857143	6	8,00000	0,030020
C_11	1,78571	1	5,714286	6	1,87500	0,219944
C_12	1,14286	1	0,857143	6	8,00000	0,030020
C_13	0,28571	1	7,714286	6	0,22222	0,654005
C_14	1,78571	1	3,714286	6	2,88462	0,140346
C_15	4,01786	1	0,857143	6	28,12500	0,001824
C_16	0,01786	1	2,857143	6	0,03750	0,852840
C_17	1,44643	1	5,428572	6	1,59868	0,252984
C_18	0,28571	1	3,714286	6	0,46154	0,522236
C_19	0,87500	1	2,000000	6	2,62500	0,156320
C_20	0,28571	1	5,714286	6	0,30000	0,603645
C_21	0,87500	1	2,000000	6	2,62500	0,156320

Рис. 6.19. Таблиця дисперсійного аналізу

Кнопка **Graph of means** дозволяє подивитися середні значення для кожного кластера на лінійному графіку (рис. 6.20).

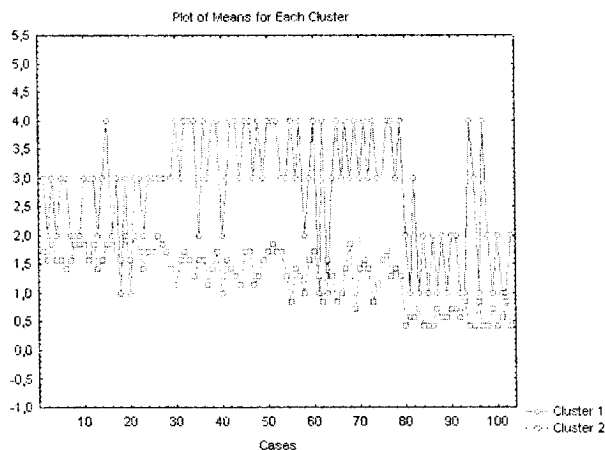


Рис. 6.20. Графік середніх значень кластерів

Кнопка «**Descriptive Statistics for each clusters**» відкриває електронну таблицю з описовими статистиками для кожного кластера: середнє, стандартне відхилення і дисперсія (рис. 6.21 і 6.22).

Descriptive Statistics for Cluster 1 (Appendix.sta)				
Cluster contains 1 variables				
Case ID	Mean	Standard Deviation	Variance	
C_1	3,000000	0,00	0,00	
C_2	2,000000	0,00	0,00	
C_3	3,000000	0,00	0,00	
C_4	2,000000	0,00	0,00	
C_5	3,000000	0,00	0,00	
C_6	3,000000	0,00	0,00	
C_7	2,000000	0,00	0,00	
C_8	2,000000	0,00	0,00	
C_9	2,000000	0,00	0,00	
C_10	3,000000	0,00	0,00	
C_11	3,000000	0,00	0,00	
C_12	3,000000	0,00	0,00	
C_13	2,000000	0,00	0,00	
C_14	3,000000	0,00	0,00	
C_15	4,000000	0,00	0,00	
C_16	2,000000	0,00	0,00	
C_17	3,000000	0,00	0,00	

Рис. 6.21. Описова статистика для першого кластера

Descriptive Statistics for Cluster 2 (Appendix.sta) Cluster contains 7 variables			
Case ID	Mean	Standard Deviation	Variance
C_1	1,714286	0,487950	0,238095
C_2	1,571429	0,786796	0,619048
C_3	1,857143	0,690066	0,476191
C_4	1,571429	0,975900	0,952381
C_5	1,571429	0,786796	0,619048
C_6	1,428571	1,133893	1,285714
C_7	1,571429	0,786796	0,619048
C_8	1,857143	0,690066	0,476191
C_9	1,857143	0,690066	0,476191
C_10	1,857143	0,377964	0,142857
C_11	1,571429	0,975900	0,952381
C_12	1,857143	0,377964	0,142857
C_13	1,428571	1,133893	1,285714
C_14	1,571429	0,786796	0,619048
C_15	1,857143	0,377964	0,142857
C_16	1,857143	0,690066	0,476191

Рис. 6.22. Описова статистика для другого кластера

Кнопка **Member of each cluster & distances** дозволяє переглянути розподіл даних за кластерами (рис. 6.23 і 6.24).

Members of Cluster Number 1 (Appendix.sta) and Distances from Respective Cluster Center Cluster contains 1 variables	
Variable	x2
Distance	0,00

Рис. 6.23. Розподіл даних для першого кластера

Members of Cluster Number 2 (Appendix.sta) and Distances from Respective Cluster Center Cluster contains 7 variables							
Variable	Variable	Variable	Variable	Variable	Variable	Variable	Variable
x1	x3	x4	x5	x6	x7	x8	
Distance	0,537295	0,571082	0,801475	0,783094	0,738427	0,719399	0,767891

Рис. 6.24. Розподіл даних для першого кластера

Кнопка **Save classifications and distances** дозволяє переглянути підсумкові дані класифікації (рис. 6.25).

	Appendix.sta		
	1	2	3
	VARIABLE	CLUSTER	DISTANCE
x1	2	2	0,54
x2	3	1	0,00
x3	4	2	0,57
x4	5	2	0,80
x5	6	2	0,78
x6	7	2	0,74
x7	8	2	0,72
x8	9	2	0,77

Рис. 6.25. Підсумкові дані класифікації

У ППП «STATGRAPHICS Plus for Windows» реалізовано сім видів ієрархічних агломеративних процедур і одна неієрархічна процедура кластерного аналізу типу k -середніх. Для прикладу розглянемо задачу про ринок цінних паперів, зокрема проблему оцінки різних фондів, що оперують цими паперами (дані наведені в посібнику зі застосування ППП «STATGRAPHICS Plus for Windows» [20].

У розглянутому прикладі буде досліджено стан 16 відомих інвестиційних фондів. Змінними є такі показники (більшість з них подано в умовних одиницях):

- ☒ прибутковість за п'ятирічний період — змінна **Five_Yr**;
- ☒ ризик — змінна **Risk**;
- ☒ відсоток прибутку (**performance**) (для кожного року) — **Perf90, Perf91, Perf92, Perf93, Perf94**;
- ☒ видаткова частина — змінна **Expense**;
- ☒ податковий рейтинг — змінна **Tax**.

Нижче наводиться таблиця з вихідними даними по досліджуваних фондах (табл. 6.2). У першому стовпчику зазначено найменування фонду, а в останньому — рекомендації експертів щодо операцій з цінними паперами цих фондів. У прикладі приведена лише послідовність розв'язання задачі кластеризації. Таблиці і рисунки не наводяться. Користувачеві пропонується вивчити методи кластеризації разом із ППП «STATGRAPHICS Plus».

Уведемо наведені дані в електронну таблицю STATGRAPHICS і збережемо їх у файлі з ім'ям **growth**. Виберемо **Special | Multivariate Methods | Cluster Analysis**. Система відобразить вікно діа-

Таблиця 6.2

Дані про інвестиційні фонди

Fund	Five_Yr	Risk	Perf90	Perf91	Perf92	Perf93	Perf94	Exprence	Tax	Recom.
F. Chip	16476	2	10	25	6	55	4	1.22	89	Buy
F. Contra	15476	2	-1	21	16	55	4	1.03	90	Buy
F. Destiny	14757	3	4	26	15	39	-3	0.7	69	Buy
Vista A	15145	4	-1	20	13	71	-6	1.49	96	Hold
Berger 100	15596	5	-7	21	9	89	-6	1.7	95	Hold
Gab. Assett	13640	1	0	22	15	18	-6	1.33	87	Buy
Neub. Focus	14081	3	1	16	21	25	-6	0.85	75	Buy
F. Magellan	13827	3	-2	25	7	41	-5	0.96	73	Buy
Janus	13187	2	-1	11	7	43	-1	0.91	85	Sell
L. Mason Value	13029	4	1	12	11	35	-17	1.82	92	Hold
Gabelli Growth	12301	3	-3	11	4	34	-2	1.41	80	Buy
Franklin Growth	11793	2	3	7	3	27	2	0.77	90	Sell
Janus 20	12441	4	-7	3	2	69	1	1.02	95	Sell
AARP Capital	11728	4	-10	16	5	41	-16	0.97	68	Sell
Kemper Growth A	11386	4	-6	2	-2	67	4	1.09	86	Sell
20th Cent. Growth	11258	4	-8	15	-4	32	0	1	60	Buy

логу для введення даних у кластерний аналіз. Двічі клацнемо лівою кнопкою миші на змінних **Exprence**, **Five_Yr**, **Perf90**, **Perf91**, **Perf92**, **Perf93**, **Perf94**, **Risk** і **Tax** для залучення їх в аналізі.

Виведемо характеристику **Fund** у поле **Point Labels** і залишимо поле даних **Select** порожнім. Натиснемо **OK**. Система видасть вікно з первинним зведенням кластерного аналізу.

Оскільки в нашому випадку бажано, щоб кластерний алгоритм добре працював з невеликою кількістю спостережень і був націлений на виділення кластерів із приблизно рівним числом членів, зупинимо свій вибір на методі Варда (**Wards method**).

Клацнемо правою кнопкою миші — на екрані з'являється вікно діалогу для вибору параметрів кластерного аналізу.

Установимо прапорець **Wards**, а всі інші залишимо у вихідному положенні. Натиснемо **OK**. На екрані відобразиться зведення кластерного аналізу для обраного методу. Натиснемо кнопку для завдання графічних опцій (третя ліворуч у верхньому ряді вікна аналізу). Система надасть спеціальне вікно діалогу.

Виберемо відображення у вигляді дендрограми (**Dendrogram**) і натиснемо кнопку **ОК**. Система додасть до табличного вікна графічне вікно.

Двічі клацнемо на дендрограмі для максимального розкриття цього вікна.

Дендрограма відображає ієрархічну структуру групування інвестиційних фондів. На ній чітко видні як мінімум три групування: одне закінчується на фонді **Gabelli Growth**, друге закінчується на фонді **Legg Mason Value** і третє, досить щільне групування, — на фонді **20th Century Growth**. Звідси випливає, що для більш детального розгляду групувань слід задати їхню кількість рівною 3.

Двічі клацнемо на рисунку для мінімізації розмірів вікна.

Клацнемо правою кнопкою миші на вікні зведення кластерного аналізу — з'явиться вікно діалогу для завдання параметрів дослідження, що проводиться.

Змінимо кількість кластерів (**Number of Clusters**) з 1 до 3.

Натиснемо кнопку **ОК**. Відповідно до введених змін будуть зроблені табличні перетворення. У зведенні кластерного аналізу насамперед вказуються: імена змінних, що беруть участь в аналізі, кількість повних зразків (спостережень без пропусків), використаний метод кластерного аналізу і прийнята метрика. Потім у зведенні описуються: число кластерів, кількість об'єктів у кожному кластері (населеність) і відповідний відсоток населеності. Крім того, у нижній частині зведення наводиться додаткова важлива інформація.

Наприклад, за координатами центроїдів можна судити про те, які змінні відіграють найбільш важливу роль у кожному кластері. Зокрема, у першому кластері видно, що витрати були розумними: незважаючи на низькі доходи у 1990 році, помітно, що в інших роках стан фондів першого кластера постійно поліпшувався. Також у першому кластері індицується найнижчий рейтинг ризику серед усіх кластерів, а податкові збори були теж досить невисокими.

Змінні, що представляють другий кластер, свідчать про те, що тут мали місце найбільші витрати, хоча за п'ятирічний період доходи залишалися найвищими. Оцінка ризику і податкові збори виявляються максимальними серед усіх кластерів.

Про третій кластер можна сказати, що він займає друге місце по витратах відносно до доходів за п'ятирічний період. Оцінка

ризикі була найвища, однак податкові збори істотно нижче, ніж у першого кластера.

Натиснемо кнопку табличних опцій (друга ліворуч у верхньому ряді). Система надасть відповідне вікно діалогу.

Встановимо **Membership Table** (таблиця приналежності спостережень), потім натиснемо кнопку **OK**.

Двічі клацнемо лівою кнопкою миші на таблиці населеності для максимального розкриття вікна.

У цій таблиці описані обрані параметри кластерного аналізу і потім дається повний список усіх спостережень, їхні імена і номери кластерів, у які входять зазначені спостереження.

Створення двовимірної діаграми розсіювання

Натиснемо кнопку графічних опцій (третю ліворуч у верхній частині вікна аналізу). З'явиться вікно діалогу для завдання відповідних параметрів. Встановимо прапорець **2D Scatterplot** (двовимірна діаграма розсіювання). Натиснемо кнопку **OK**, в результаті чого, система відобразить ще одне графічне вікно. Двічі клацнемо лівою кнопкою миші на вікні дендрограми, щоб розгорнути його.

На дендрограмі видні три дерева. За вертикальною віссю відкладена відстань для кожного кроку роботи агломеративного ієрархічного алгоритму кластеризації. На горизонтальній осі показані спостереження, скомбіновані у відповідності до проведеного аналізу. Дендрограма дозволяє побачити виразну картину трьох групувань й імена спостережень (інвестиційних фондів), що увійшли у виділені кластери.

Двічі клацнемо на дендрограмі й тим самим знову мінімізуємо її. Двічі клацнемо лівою кнопкою миші на двовимірній діаграмі розсіювання.

Діаграма показує, як групуються досліджувані спостереження на площині двох змінних **Expence** і **Five_Yr**. Кожен кластер поданий на діаграмі власним символом, а якби це було в кольоровому виконанні, то і кольором. З графіка випливає, що 1-й кластер має низькі відносні витрати; видно, як розподілені доходи фондів за п'ятирічний період. У кластері 2 спостерігаються найвищі витрати, але і максимальні п'ятирічні доходи. У кластері 3 низькі витрати супроводжуються й невисокими п'ятирічними доходами.

Для того щоб відобразити інші діаграми розсіювання, досить клацнути правою кнопкою миші і, отримавши в розпорядження відповідне вікно діалогу, вибрати цікавлячі пари змінних.

В основі кластерного аналізу лежить припущення, що всередині класів об'єкти мають бути пов'язані між собою. Вибір того чи іншого критерію оцінки щільності зв'язку залежить від багатьох факторів. Пакети прикладних програм пропонують декілька мір подібності. Природно, виникає питання: яку з них вибрати? Однозначної відповіді не існує. Однак на підставі статистичного аналізу даних і глибокого вивчення предметної галузі можна зробити попередній вибір критерію оцінки щільності зв'язку.

Ще одне природне запитання: яке застосувати правило об'єднання кластерів? Це також вирішується неформально, а після вивчення наявного матеріалу, залучення теоретичних результатів і випробування різних варіантів дій.

Можна запропонувати таку послідовність розв'язання задачі розбиття множини даних на класи:

- ☒ оцінити однорідність виборок;
- ☒ у випадку недотримання гіпотез про однорідність провести трансформацію вихідних даних;
- ☒ провести кластеризацію з використанням усіх критеріїв і вибрати найбільш кращий.

Контрольні запитання та завдання

1. Поясніть сутність кластерного аналізу.
2. За якими ознаками можна визначити однорідність виборок?
3. Які критерії використовуються в кластерному аналізі для визначення однорідностей?
4. Яким вимогам повинна задовольняти міра подібності, для того щоб бути метрикою?
5. Які найбільш розповсюджені способи визначення відстані між кластерами?
6. Поясніть принцип роботи агломеративних кластер-процедур.
7. Як можна обчислити відстань між об'єктом та кластером?

Завдання 1

За ієрархічним агломеративним алгоритмом провести класифікацію $n = 17$ марок автомобілів, що розрізняються ціною, витратою пального, часом розгону, необхідним для того, щоб досягти швидкості 70 миль на годину. Розв'язати задачу методом кластеризації — з'єднан-

ня. Таблиця вихідних даних наведена у ППП «STATISTICA» у файлі вихідних даних **EXAMPLES**.

Завдання 2

Розв'язати задачу 1 методом кластеризації — K -середніх.

Завдання 3

Розв'язати задачу 1, використовуючи метод двобічного з'єднання вихідних даних.

Завдання 4

За агломеративним алгоритмом провести класифікацію $n = 6$ регіонів за рівнем медичного обслуговування населення, який характеризується показниками: x_1 — число лікарів на 10 тис. мешканців та x_2 — число лікарняних ліжок на 10 тис. мешканців (див. табл. 6.3).

Таблиця 6.3

№ регіону	1	2	3	4	5	6
x_1	34,8	31,2	32,1	35,7	30,2	34,2
x_2	126	112	123	128	115	123

За відстань між об'єктами прийняти евклідову відстань s , а відстань між кластерами вимірювати за принципом: а) «найближчого сусіда»; б) «далекого сусіда».

Завдання 5

Розв'язати задачу 4 за умови, що відстань між кластерами вимірюється за принципом «центра ваги» груп.

РОЗДІЛ 7

ВИКОРИСТАННЯ ДИСКРИМІНАНТНОГО АНАЛІЗУ ДЛЯ ОБРОБКИ ТЕХНІКО-ЕКОНОМІЧНОЇ ІНФОРМАЦІЇ В ІНФОРМАЦІЙНО-УПРАВЛЯЮЧИХ СИСТЕМАХ

7.1. Постановка задачі

7.1.1. Припущення дискримінантного аналізу

Дискримінантний аналіз (ДА) є розділом багатовимірною статистичного аналізу, який дозволяє вивчати розбіжність між групами об'єктів за декількома змінними одночасно. Дискримінантний аналіз — це загальний термін, що відноситься до декількох щільно зв'язаних між собою статистичних процедур. Ці процедури можна поділити на методи інтерпретації міжгрупових розбіжностей — дискримінації і методи класифікації спостережень за групами. При інтерпретації потрібно відповісти на питання: чи можливо, використовуючи даний набір змінних, відрізнити одну групу від іншої, наскільки добре ці змінні допомагають провести дискримінацію і які з них найбільш інформативні? Дискримінантний аналіз знаходить широке застосування в різних предметних галузях. Наприклад, у соціології, медицині, економіці, техніці, а також при розв'язанні багатьох функціональних задач в інформаційно-управляючих системах. У загальному випадку задача дискримінації формулюється таким чином.

Припустимо, що ми маємо сукупність об'єктів (спостережень), розбиту на декілька класів W_i , $i = 1, 2, \dots, k$ (тобто щодо кожного об'єкта можна сказати, до якого класу він належить). Нехай для кожного об'єкта є виміри $X = (x_1, x_2, \dots, x_p)$ декількох кількісних характеристик. Необхідно знайти спосіб, за яким на підставі цих характеристик можна встановити клас, до якого належить об'єкт. Це дозволить визначати групи, до яких належать нові об'єкти з тієї ж сукупності. Характеристики, застосовувані для того, щоб відрізнити один клас від іншого, називаються *дискримінантними змінними* і вимірюються за *інтервальною шкалою* або за *шкалою відносин*.

Наприклад, досліджуваними об'єктами можуть бути люди — здорові чи хворі, а характеристиками — результати медичних аналізів. Якщо ми навчимося за цими характеристиками дізнатися, здоровий пацієнт чи має якесь захворювання, це дозволить значно підвищити ефективність медичних обстежень.

Для розв'язання цієї задачі необхідно побудувати функції вимірюваних характеристик, значення яких і пояснюють розбиття об'єктів на групи. Бажано, щоб таких функцій (дискримінуючих ознак) було небагато — у цьому випадку результати аналізу легше змістовно витлумачити.

Особливу роль завдяки своїй простоті відіграє лінійний дискримінантний аналіз, у якому класифікаційні ознаки обираються як лінійні функції від первинних ознак. Правило дискримінації визначається на підставі розв'язання оптимізаційної задачі. Точність розв'язання залежить від повноти апріорної інформації. При цьому треба брати до уваги втрати від неправильної класифікації. У випадку поділу декількох нормальних сукупностей лінійний дискримінантний аналіз виявляє статистичні властивості.

Правило дискримінації можна реалізувати за допомогою параметричних і непараметричних методів. Найбільш повно вивчені параметричні методи. Розподіл векторів ознак у кожній сукупності має бути нормальним, однак часто параметри цих розподілів невідомі. У цьому разі невідомі параметри розподілів замінюють їхніми оптимальними оцінками.

Непараметричні методи дискримінації не потребують знань про функціональний вид розподілів. У практичному застосуванні ці методи є найбільш цінними, оскільки дозволяють розв'язувати задачу дискримінації на підставі незначної апріорної інформації.

Застосування дискримінантного аналізу буде коректним, якщо дотримуються припущення, покладені в його основу. Якщо експериментальні дані для деякої функціональної задачі не відповідатимуть цим припущенням, то статистичні висновки не будуть точним відображенням реальності.

Розглянемо сутність припущень дискримінантного аналізу. По-перше, жодна змінна не може бути лінійною комбінацією інших змінних. Лінійна комбінація — це сума однієї або більше змінних із сталими вагами. Таким чином, не можна користуватися сумою змінних або їх середнім арифметичним значенням спільно із самими змінними. Відповідно неприпустиме використання змінних, коефіцієнт кореляції яких дорівнює 1. Змінна, що є лінійною комбінацією інших змінних, не несе будь-якої нової інформації, окрім тієї, яка міститься в компонентах суми, тому вона є зайвою.

Друге припущення, що приймається в багатьох випадках, полягає в тому, що коваріаційні матриці для генеральних сукупностей (*генеральні коваріаційні матриці*) однакові для різних класів. Часто використовуваний формі ДА притаманні лінійні дискримінантні функції, що відповідають просто лінійній комбінації дискримінантних змінних. Цей метод є найпростішим, оскільки припущення про однакові коваріаційні матриці у класах спрощує формули обчислення дискримінантних функцій, а також полегшує перевірку гіпотез про статистичну значущість.

Наступне припущення стосується того, що закон розподілу для кожного класу є багатовимірним нормальним законом, тобто кожна змінна має нормальний розподіл за фіксованих інших змінних. Це припущення дозволяє отримати точні значення ймовірності належності до цього класу і критерію значущості. При порушенні припущення про нормальність розподілу значення ймовірності точно обчислити вже не можна, але відповідні оцінки можуть бути корисні за дотримання певних обмежень.

У дискримінантному аналізі класи розглядаються як значення деякої класифікуючої змінної, що є вимірною за шкалою найменувань (кожному класу надається своє позначення). Залежно від зв'язку класифікуючої змінної та дискримінантних змінних задачу дискримінантного аналізу можна розв'язати за допомогою множинної регресії або множинного дисперсійного аналізу.

Якщо в конкретній ситуації класифікуючі змінні можна вважати залежними від дискримінантних змінних, то задача яв-

ляється аналогічною до задачі множинної регресії. Основна відмінність полягає в тому, що в дискримінантному аналізі залежна змінна вимірюється за шкалою найменувань (класів). Але коли передбачається, що значення дискримінантних змінних залежать від класів, дискримінантний аналіз є узагальненням багатовимірного дисперсійного аналізу. Це є типовим для задач, у яких належність змінних до деякого класу викликає зміни одночасно в декількох змінних.

7.1.2. Концептуальна модель дискримінантного аналізу

Розглянемо принцип роботи концептуальної моделі функціонування дискримінантного аналізу, поданої на рис. 7.1.

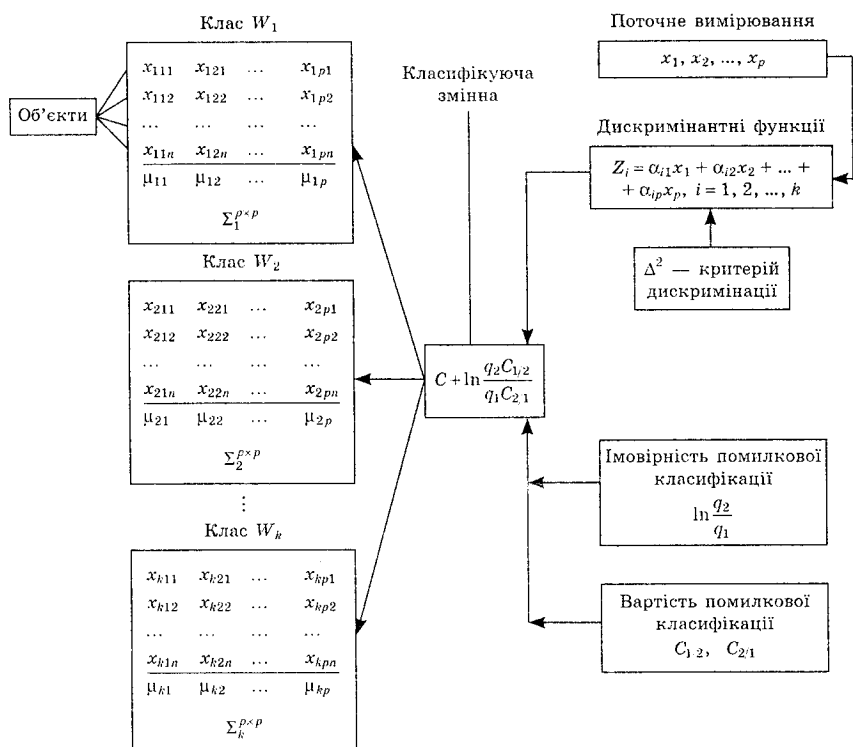


Рис 7.1. Концептуальна модель дискримінантного аналізу

Тут може бути декілька постановок задач. Для векторів X_i можуть бути задані параметри μ_{ij} і $\Sigma_i^{p \times p}$. З іншого боку, є вектори спостережень X_{ij} , але середні значення μ_{ij} і матриці коваріацій $\Sigma_i^{p \times p}$ невідомі. Відомі вартості помилкової класифікації $C_{i,j}$ та апріорні ймовірності $q_1, q_2, \dots, q_k$ віднесення X_i до того чи іншого класу. Мають місце довільні розподіли класів.

Процедура класифікації матиме такий вигляд. Є поточне вимірювання вектора $X = (x_1, x_2, \dots, x_p)$, значення якого підставляється у дискримінантні функції $Z_i = \alpha_{i1}x_1 + \alpha_{i2}x_2 + \dots + \alpha_{ip}x_p$. Величини $\alpha_{ij}^{b(1)}$ визначаються у відповідності до критерію дискримінації (7.1). Класифікована змінна оцінюється з урахуванням ймовірності помилкової класифікації. Залежно від значень дискримінантних функцій поточне спостереження відносять до того чи іншого класу.

Визначимо основні принципи, які покладено в основу обчислення дискримінантних функцій, та методи визначення їх кількості.

Канонічна дискримінантна функція Z є лінійною комбінацією дискримінантних змінних X і відповідає певним умовам. Вона має таке математичне подання:

$$f_i(x) = Z_i = \alpha_{i1}x_1 + \alpha_{i2}x_2 + \dots + \alpha_{ip}x_p, \quad (7.1)$$

де α_{ij} — коефіцієнти, які забезпечують виконання необхідних умов.

Коефіцієнти α_{ij} для першої функції обираються таким чином, щоб її середні значення для різних класів якомога більше відрізнялися одне від одного. Коефіцієнти другої функції обирають так само, тобто відповідні середні значення мають максимально відрізнятися за класами. При цьому накладається додаткова умова: значення другої функції мають бути не корельовані зі значеннями першої; значення третьої функції мають бути не корельовані зі значеннями перших двох функцій і т. д. Максимальне число дискримінантних функцій, яке можна отримати описаним способом, дорівнює числу класів без одиниці або числу змінних, залежно від того, яка з цих величин менша.

Розглянемо геометричну інтерпретацію дискримінантних функцій. Нехай дискримінантні змінні — осі p -вимірного евклідового простору. Кожен об'єкт (спостереження) є точкою цього простору з координатами, що являють собою спостережувані значення кожної змінної. Якщо класи відрізняються один від одного

досліджуваними змінними, їх можна подати як скупчення точок у деяких площинах простору, що розглядається. Оскільки класи можуть частково перекриватися, відповідні їм «території» не збігаються. Для визначення положення класу можна обчислити його центроїд.

Центроїд класу є уявлюваною точкою, координати якої — це середні значення змінних у цьому класі. Центроїд можна використовувати для вивчення розбіжностей між класами, оскільки він займає положення типових спостережень відповідного класу. Розгляд окремих змінних не дозволяє проводити багатовимірний аналіз — число змінних може бути великим, тому сукупну інформацію складно систематизувати. Виявляється, що для розрізнення відносного положення центроїдів не потрібна занадто велика розмірність. Як правило, досить обмежитися розмірністю, на одиницю меншою за число класів.

Оскільки центроїди задають простір, то у наявності є відповідне необмежене число точок, де можна помістити систему координат. Найбільш зручна точка, у якій кожна вісь має нульове значення, — це «головний» центроїд. Він займає положення, обумовлене середніми значеннями сукупності об'єктів за кожною з осей. Щодо цього центра існує нескінченна множина орієнтацій осей за умови, що вони належать простору, «натягнутому» на центроїди. Якщо спрямувати одну з цих осей під кутом, для якого середні значення класів розділяються в більшому ступені, ніж для будь-якого іншого напрямку, то отримаємо вісь, яка, як нам здається, повинна бути особливо важливою. Припускаючи, що є два і більше класи, можна орієнтувати другу вісь таким чином, щоб був забезпечений максимальний поділ класів, але при додатковому обмеженні — друга вісь ортогональна першій (і належить простору, що розглядається).

Аналогічно проводяться інші осі. Розташування осей за таким принципом приводить до критерію для канонічних дискримінантних функцій. Співвідношення (7.1) задає математичне перетворення p -вимірного простору дискримінантних змінних у q -вимірний простір канонічних дискримінантних функцій (де q — максимальне число функцій). Кожній осі відповідає своє співвідношення вигляду (7.1). Для цього спостереження значення Z_i інтерпретується як координата об'єкта в просторі канонічних дискримінантних функцій. Виключення з наведеного правила складають випадки, коли один чи декілька центроїдів не визначають нового напрямку.

У випадку, коли число дискримінантних змінних p менше за число класів, максимальне число функцій q дорівнює p . При цьому вже не відбувається перетворення з простору з більшою розмірністю в простір з меншою розмірністю. Слід лише виконати заміну координат, яка задовольняє деякому критерію.

На рис. 7.2 зображені об'єкти, що належать двом різним множинам W_1 і W_2 . Кожен об'єкт характеризується в цьому випадку двома змінними x_1 і x_2 . Якщо розглядати проекції об'єктів (точок) на кожну вісь, то ці множини перетинаються, тобто за кожною змінною окремо деякі об'єкти обох множин мають подібні характеристики. Щоб якнайкраще розділити розглянуті множини, потрібно побудувати відповідну лінійну комбінацію змінних x_1 і x_2 . Для двовимірного простору ця задача зводиться до визначення нової системи координат. Причому нові осі L і C повинні бути розташовані таким чином, щоб проекції об'єктів, які належать різним множинам, на вісь L , були максимально розділені. Вісь C перпендикулярна до осі L і розділяє дві «хмари» точок найкращим чином, тобто множини виявилися по різні боки від цієї прямої. При цьому ймовірність помилки класифікації повинна бути мінімальною. Сформульовані умови слід взяти до уваги при визначенні коефіцієнтів α_1 і α_2 у (7.1).

Позначимо x_{ij} — середнє значення j -тої ознаки у об'єктів i -тої множини (класу). Тоді для множини M_1 середнє значення функції $f_1(x)$ дорівнюватиме:

$$\bar{f}_1(x) = a_1 \bar{x}_{11} + a_2 \bar{x}_{12},$$

а для множини M_2 середнє значення функції $f_2(x)$ буде дорівнювати:

$$\bar{f}_2(x) = a_1 \bar{x}_{21} + a_2 \bar{x}_{22}.$$

Геометрична інтерпретація цих функцій являє собою дві паралельні прямі, що проходять через центри класів (множин) (див. рис. 7.3).

Дискримінантна функція може бути як лінійною, так і нелінійною. Вибір її виду залежить від геометричного розташування поділюваних класів у просторі дискримінантних змінних. Для спрощення викладень надалі розглядається лінійна дискримінантна функція.

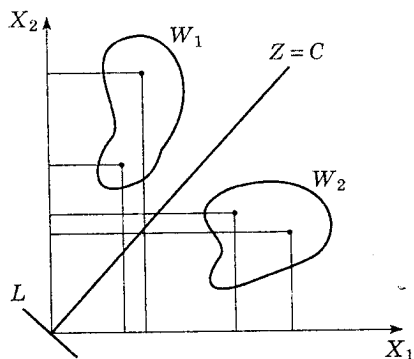


Рис. 7.2. Розділення дискримінантною функцією Z

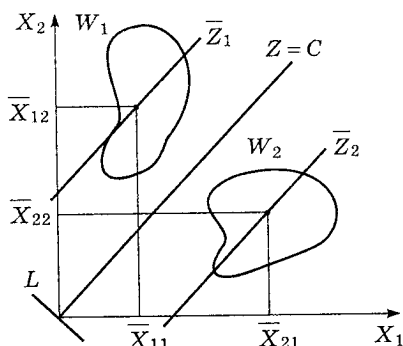


Рис. 7.3. Центри поділюваних множин і константа дискримінації

7.2. Обчислення дискримінантних функцій і їх інтерпретація

Для більш глибокого розуміння суті дискримінантних функцій, розглянемо процедуру класифікації для двох класів [8].

Нехай процедура класифікації для випадку p безперервних змінних припускає, що спостереження належать одному з двох класів, що мають багатовимірні нормальні розподіли. Спостереження $x_1, x_2, \dots, x_p$ записуються у вигляді вектора $X = (x_1, x_2, \dots, x_p)$. Передбачається, що W_1 має розподіл $N(\mu_1^{p \times 1}, \Sigma_1^{p \times p})$, а W_2 — розподіл $N(\mu_2, \Sigma_2^{p \times p})$, де $\mu_i = (\mu_{i1}, \mu_{i2}, \dots, \mu_{ip})$, $i = 1, 2$; μ_1, μ_2 — вектори середніх; Σ — матриця коваріацій. Інше спрощувальне припущення полягає в тому, що $\Sigma_1 = \Sigma_2 = S_{vj}$, $v = 1, \dots, p$; $j = 1, \dots, p$. Отже, вектор X має розподіл $N(\mu_1, \Sigma)$ чи $N(\mu_2, \Sigma)$.

Нехай задані вектори середніх μ_1 і μ_2 і матриці коваріацій $\Sigma_1 = \Sigma_2 = S_{vj}$. Дискримінантну функцію задамо у вигляді

$$Z_i = \alpha_{i1}x_1 + \alpha_{i2}x_2 + \dots + \alpha_{ip}x_p, \quad (7.2)$$

де $\alpha_1, \dots, \alpha_p$ — деякі сталі величини. Віднесемо X до W_1 , якщо $Z \geq C$, і до W_2 , якщо $Z < C$ (C — стала величина).

Тоді задача буде зведена до визначення значень $\alpha_1, \dots, \alpha_p$ і C , що мінімізують імовірність помилкової класифікації. Спочатку будемо вибирати ці сталі на інтуїтивному рівні, а потім покажемо, що отримана процедура класифікації є оптимальною.

Якщо спостереження X надійшло з W_1 , то величина Z має нормальний розподіл із середнім значенням дискримінантних функцій

$$\xi_1 = \sum_{j=1}^p \alpha_j \mu_{1j} \quad (7.3)$$

і дисперсією

$$\sigma_Z^2 = \sum_{v=1}^p \sum_{j=1}^p \alpha_v S_{vj} \alpha_j. \quad (7.4)$$

Аналогічно для X з W_2 величина Z має нормальний розподіл із середнім значенням дискримінантних функцій:

$$\xi_2 = \sum_{j=1}^p \alpha_j \mu_{2j} \quad (7.5)$$

і з тією ж дисперсією σ_Z^2 . Має сенс вибрати такі $\alpha_1, \dots, \alpha_p$, за яких ξ_1 і ξ_2 були б якомога більше віддалені один від одного відносно σ_Z^2 . Для цього введемо відстань між об'єктом і кожним із центроїдів класів, або *відстань Махаланобіса*:

$$\Delta^2 = \frac{(\xi_1 - \xi_2)^2}{\sigma_Z^2}, \quad (7.6)$$

щоб потім віднести об'єкт до найближчого класу. Співвідношення (7.6) припускає, що класи мають однакові коваріаційні матриці. Таким чином, потрібно знайти коефіцієнти $\alpha_1, \dots, \alpha_p$, що максимізують Δ^2 . У роботі [8] показано, що такі α_i є розв'язком системи лінійних рівнянь

$$\begin{array}{cccccc} \alpha_1 S_{11} & \alpha_2 S_{12} & \dots & \alpha_p S_{1p} & = & \mu_{11} - \mu_{21}; \\ \alpha_1 S_{21} & \alpha_2 S_{22} & \dots & \alpha_p S_{2p} & = & \mu_{12} - \mu_{22}; \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \alpha_1 S_{p1} & \alpha_2 S_{p2} & \dots & \alpha_p S_{pp} & = & \mu_{1p} - \mu_{2p}. \end{array} \quad (7.7)$$

Після підстановки отриманих α_i у (7.2) кожному об'єкту $x_1, x_2, \dots, x_p$ ставиться у відповідність значення дискримінантної функції Z .

У деяких випадках коефіцієнти α_j , $j = 1, 2, \dots, p$ можуть використовуватися при класифікації, у цьому разі є доцільним їх нормування. Для нормованих коефіцієнтів точка, де проєкції всіх

дискримінантних функцій нульові, збігається з головним центроїдом, де дискримінантні змінні набувають середніх значень.

Для нормованих коефіцієнтів дискримінантні значення вимірюються в одиницях стандартного квадратного відхилення. Спосіб приведення змінних до стандартної форми залежить від того, чи нормалізуються вихідні значення спостережень.

Для обчислення дискримінантних функцій скористаємося не-стандартизованими коефіцієнтами.

Розглянемо евристичну процедуру класифікації, у якій дискримінантні функції мають не стандартизовані коефіцієнти.

Для визначення сталої C слід розглянути рис. 7.4, на якому показані два розподіли величини Z відносно сталої C . Якщо вектор X належить W_2 , але $Z = \sum_{i=1}^p \alpha_i x_i \geq C$, то X відноситься до W_1 і, таким чином, буде зроблена помилка. Імовірність $\text{Pr}(1|2)$ цієї помилки зазначена на рис. 7.4. На цьому рисунку показана також імовірність $\text{Pr}(2|1)$ помилкової класифікації об'єкта з популяції W_1 , як приналежного до популяції W_2 . Природно шукати таку сталу C , щоб сума ймовірностей $\text{Pr}(1|2) + \text{Pr}(2|1)$ була мінімальною. Цього можна досягти вибором сталої C , рівновіддаленої від середніх значень, тобто

$$C = \frac{\xi_1 + \xi_2}{2}. \quad (7.8)$$

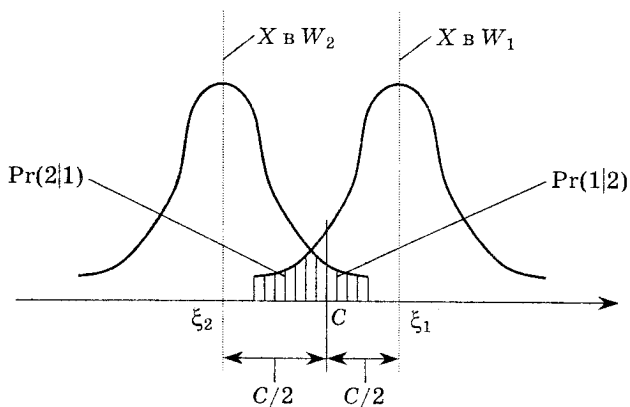


Рис. 7.4. Розподіл величини Z , коли X належить W_1 і W_2 відповідно; W_i має розподіл $N(\mu_i, \Sigma)$, $i = 1, 2$

Отже, евристична процедура класифікації полягає в обчисленні $\alpha_1, \dots, \alpha_p$, що задовольняють системі рівнянь (7.7), оцінок ξ_1 і ξ_2 , згідно з (7.3) і (7.5), а потім сталої C з формули (7.7). Для кожного вектора спостережень обчислюється значення (7.2) дискримінантної функції Z і X відноситься до популяції W_1 , якщо виконується умова $Z \geq C$, інакше — до W_2 .

Розглянемо приклад використання дискримінантного аналізу для проведення багатовимірної класифікації об'єктів. При цьому в якості навчальних будемо використовувати спочатку дві вибірки, що належать двом класам, а потім узагальнимо алгоритм класифікації на випадок k класів.

Нехай наявні дані щодо двох груп промислових підприємств машинобудівного комплексу: X_1 — фондівіддача основних виробничих фондів, грн; X_2 — витрати на гривну виробленої продукції, коп.; X_3 — витрати сировини і матеріалів на одну гривну продукції, коп.

Таблиця 7.1

	Номер підприємства	X_1	X_2	X_3
1-а група	1	0,50	94,0	8,50
	2	0,67	75,4	8,79
	3	0,68	85,2	9,10
	4	0,55	98,8	8,47
2-а група	5	1,52	81,5	4,95
	6	1,20	93,8	6,95
	7	1,46	86,5	4,70

Необхідно провести класифікацію чотирьох нових підприємств, що мають такі значення вихідних змінних:

✓ 1-е підприємство — $x_1 = 1,07$, $x_2 = 93,5$, $x_3 = 5,30$;

✓ 2-е підприємство — $x_1 = 0,99$, $x_2 = 84,0$, $x_3 = 4,85$;

✓ 3-е підприємство — $x_1 = 0,70$, $x_2 = 76,8$, $x_3 = 3,50$;

✓ 4-е підприємство — $x_1 = 1,24$, $x_2 = 88,0$, $x_3 = 4,95$.

Для зручності запишемо значення вихідних змінних для кожної групи підприємств у вигляді матриць X_1 і X_2 :

$$X_1 = \begin{pmatrix} 0,50 & 94,0 & 8,50 \\ 0,67 & 75,4 & 8,79 \\ 0,68 & 85,2 & 9,10 \\ 0,55 & 98,8 & 8,47 \end{pmatrix}, \quad X_2 = \begin{pmatrix} 1,52 & 81,5 & 4,95 \\ 1,20 & 93,8 & 6,95 \\ 1,46 & 86,5 & 4,70 \end{pmatrix}.$$

Обчислимо середнє значення кожної змінної в окремих групах для визначення положення центрів цих груп:

$$\boxed{\text{I група}} \quad \bar{x}_{11}=0,60, \quad \bar{x}_{21}=88,4, \quad \bar{x}_{31}=8,72;$$

$$\boxed{\text{II група}} \quad \bar{x}_{12}=1,39, \quad \bar{x}_{22}=87,3, \quad \bar{x}_{32}=5,53.$$

Дискримінантна функція Z у даному випадку має вигляд:

$$Z = \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3. \quad (7.9)$$

Коефіцієнти α_1, α_2 і α_3 обчислюються за формулою:

$$A = S_*^{-1}(\bar{X}_1 - \bar{X}_2),$$

де $\bar{X}_1, \bar{X}_2$ — вектори середніх значень у першій і другій групах; A — вектор коефіцієнтів; S_* — матриця, обернена до спільної коваріаційної матриці.

Для визначення спільної коваріаційної матриці S_* потрібно розрахувати матриці S_1 і S_2 . Кожен елемент цих матриць являє собою різницю між відповідним значенням вихідної змінної x_{ij} і середнім значенням цієї змінної в даній групі $\bar{x}_{ik}$ (k — номер групи):

$$S_1 = \begin{pmatrix} 0,0238 & -2,2460 & 0,0698 \\ -2,2460 & 318,76 & -5,958 \\ 0,0698 & -5,958 & 0,2602 \end{pmatrix},$$

$$S_2 = \begin{pmatrix} 0,0579 & -2,0450 & -0,4033 \\ -2,0450 & 76,530 & 13,258 \\ -0,4033 & 13,258 & 3,0417 \end{pmatrix}$$

Тоді спільна коваріаційна матриця S_* дорівнюватиме:

$$S_* = \frac{1}{n_1 + n_2 - 2}(S_1 + S_2),$$

де n_1, n_2 — число об'єктів 1-ї і 2-ї групи;

$$\begin{aligned} S_* &= \frac{1}{4+3-2} \begin{pmatrix} 0,0817 & -4,291 & -0,3335 \\ -4,291 & 395,290 & 7,300 \\ -0,3335 & 7,300 & 3,3019 \end{pmatrix} = \\ &= \begin{pmatrix} 0,01634 & -0,8582 & -0,0667 \\ -0,8582 & 79,058 & 1,460 \\ -0,0667 & 1,460 & 0,6604 \end{pmatrix}. \end{aligned}$$

Обернена матриця S_*^{-1} дорівнюватиме:

$$S_*^{-1} = \begin{pmatrix} 339,970 & -3,190 & 27,290 \\ -3,190 & 0,043 & -0,227 \\ 27,290 & -0,227 & 8,380 \end{pmatrix}.$$

Звідси знаходимо вектор коефіцієнтів дискримінантної функції за формулою:

$$\begin{aligned} A = S_*^{-1}(\bar{X}_1 - \bar{X}_2) &= \begin{pmatrix} 339,970 & -3,190 & 27,290 \\ -3,190 & 0,043 & -0,227 \\ 27,290 & -0,227 & 8,380 \end{pmatrix} \begin{pmatrix} -0,79 \\ 1,10 \\ 3,19 \end{pmatrix} = \\ &= \begin{pmatrix} -185,03 \\ 1,84 \\ 4,92 \end{pmatrix}, \end{aligned}$$

тобто $a_1 = -185,03$, $a_2 = 1,84$, $a_3 = 4,92$.

Підставимо отримані значення коефіцієнтів у формулу (7.9) і розрахуємо значення дискримінантної функції для кожного об'єкта:

☑ для першої множини

$$Z_{11} = 0,5 \times (-185,03) + 94 \times 1,84 + 8,5 \times 4,92 = 122,265;$$

$$Z_{12} = 0,67 \times (-185,03) + 75,4 \times 1,84 + 8,79 \times 4,92 = 58,012;$$

$$Z_{13} = 0,68 \times (-185,03) + 85,2 \times 1,84 + 9,1 \times 4,92 = 75,71;$$

$$Z_{14} = 0,55 \times (-185,03) + 98,8 \times 1,84 + 8,47 \times 4,92 = 121,6979;$$

$$Z_{1\text{порів}} = 94,4238;$$

☑ для другої множини

$$Z_{21} = 1,52 - 185,03 + 81,5 \times 1,84 + 4,95 \times 4,92 = -106,9316;$$

$$Z_{22} = 1,20 - 185,03 + 93,8 \times 1,84 + 6,95 \times 4,92 = -15,25;$$

$$Z_{23} = 1,46 - 185,03 + 86,5 \times 1,84 + 4,7 \times 4,92 = -87,85;$$

$$Z_{2\text{порів}} = -70,0138.$$

Після отримання константи дискримінації можна перевірити правильність розподілу об'єктів у вже існуючих двох класах, а також здійснити класифікацію нових об'єктів. Розглянемо, наприклад, об'єкти з номерами 1, 2, 3, 4. Для того, щоб віднести

ці об'єкти до однієї з двох множин, обчислимо для них значення дискримінантних функцій (за трьома змінними):

$$Z_1 = -185,03 \times 1,07 + 93,5 \times 1,84 + 5,30 \times 4,92 = 0,1339;$$

$$Z_2 = -185,03 \times 0,99 + 84,0 \times 1,84 + 4,85 \times 4,92 = -4,7577;$$

$$Z_3 = 0,70 \times (-185,03) + 76,8 \times 1,84 + 3,50 \times 4,92 = 29,011;$$

$$Z_4 = 1,24 \times (-185,03) + 88,0 \times 1,84 + 4,95 \times 4,92 = -43,1632.$$

Таким чином, об'єкти 1, 2 і 4 належать до другого класу, а об'єкт 3 належить до першого класу, оскільки $Z_1 < C$, $Z_2 < C$, $Z_4 < C$, $Z_3 > C$.

7.3. Байєсівська процедура класифікації

Байєсівська процедура класифікації досить повно викладена в роботі [8]. Зауважимо, що відстань Махаланобіса Δ^2 має ті ж властивості, що і статистика χ^2 з p -степенями вільності. Таким чином, відстань вимірюється в « χ^2 -одинацях». Якщо припустити, що кожен клас є частиною генеральної сукупності з багатовимірним нормальним розподілом, то більшість об'єктів групуватиметься поблизу центроїда і їхня щільність спадатиме в міру віддалення від нього. Знаючи відстань від центроїда, можна сказати, яка частина класу знаходиться ближче, а яка — далі від нього. Отже, можна оцінити ймовірність того, що об'єкт, який знаходиться на деякій відстані від центроїда, належить до того чи іншого класу. Оскільки наші відстані вимірюються в χ^2 -одинацях, то спробуємо знайти значущість отримання цієї ймовірності. Позначимо через $\Pr(X|G_k)$ ймовірність того, що об'єкт, який знаходиться далеко від центроїда, дійсно належить до класу K .

Відносячи об'єкт до найближчого класу у відповідності зі значенням Δ^2 , ми неявно приписуємо його до того класу, щодо якого він має найбільшу ймовірність приналежності. За наявності даних щодо таких ймовірностей про об'єкт можна стверджувати, що він є «найближчим» до якогось конкретного класу. Насправді об'єкт може з великими ймовірностями належати більш ніж до одного класу або не належати жодному з них. Розглянемо ситуацію з низькою розбіжністю і високим перекриттям класів. У цьому випадку об'єкт, близький до центроїду класу 1, буде

з більшою імовірністю «належати» до класу 2, оскільки він також «близький» до цього класу. Інша важлива ситуація: об'єкт знаходиться на великій відстані від усіх класів, інакше кажучи — всі ймовірності малі. Рішення щодо віднесення цього об'єкта до найближчого класу може виявитися позбавленим сенсу, оскільки він мало схожий на будь-який об'єкт з цього класу.

Наведемо тепер більш строгі розв'язання задачі класифікації, засновані на теоремі Байєса. Визначимо спочатку апіорну ймовірність q_i , як ймовірність того, що об'єкт належить до класу W_i , $i = 1, 2$. Передбачається, що сума апіорних ймовірностей ($q_1 + q_2$) дорівнює 1.

Ймовірність $\Pr(X|W_i)$ оцінює частку об'єктів у цьому класі, які знаходяться далі від центроїда, ніж X .

Визначимо умовну ймовірність $\Pr(X|W_i)$ отримання деякого вектора спостережень X , якщо відомо, що об'єкт належить до популяції W_i , $i = 1, 2$. Позначимо також через $\Pr(W_i|X)$ умовну ймовірність того, що об'єкт належить до популяції W_i при цьому векторі спостережень X . Величини $\Pr(W_1|X)$ і $\Pr(W_2|X)$ називаються *апостеріорними ймовірностями*. Різниця між апіорними й апостеріорними ймовірностями полягає в тому, що до отримання вектора спостережень цього об'єкта відомо, що він належить до сукупності W_i з ймовірністю q_i . Після аналізу вектора спостережень цього об'єкта можна сказати, що він належить до популяції W_i з ймовірністю $\Pr(W_i|X)$.

Дотепер під час обговорення класифікації передбачалося, що всі класи рівноправні, але на практиці це не завжди так. Розглянемо, наприклад, випадок двох класів, коли 90 % об'єктів генеральної сукупності міститься в класі 1. Ще до обчислення зрозуміло, що з дуже великою ймовірністю будь-який заданий об'єкт належить до класу 1. Отже, він буде віднесений до класу 2 тільки за наявності дуже переконливих свідчень на користь такого рішення. У цьому можна пересвідчитися, обчислюючи апостеріорні ймовірності з урахуванням апіорних знань про ймовірну приналежність до класу.

Інша ситуація, за якої є бажаним використання апостеріорних ймовірностей, виникає, коли «вартість» неправильної класифікації істотно змінюється від класу до класу. Якби ці витрати від неправильної класифікації могли бути виражені у вигляді відношення, то їх слід було б використовувати у той самий спосіб, що й апіорні ймовірності.

Для підвищення точності припущення і для зменшення «вартості» помилок бажано включити апіорні ймовірності у класифікуючі функції. Для простих класифікуючих функцій це можна зробити за допомогою додавання натурального логарифма від апіорної ймовірності до сталого члена. Також можна модифікувати відстань Δ^2 , двічі віднімаючи натуральний логарифм від апіорної ймовірності. Ця зміна у відстані математично ідентична до множення $\Pr(X|G_k)$ на апіорну ймовірність для цього класу. Якщо класи дуже розрізняються, то залучення апіорних ймовірностей навряд чи вплине на результат, оскільки поблизу межі між класами буде знаходитися дуже мало об'єктів. Таким чином, апіорні ймовірності будуть найбільш впливати, коли класи перекриваються і, отже, багато об'єктів з більшою ймовірністю можуть належати до декількох класів. Звичайно, в основу рішення про використання апіорних ймовірностей мають бути покладені теоретичні передумови. Якщо ж таких передумов немає, то краще цього не робити. Слід також пам'ятати, що апіорні ймовірності обчислені на основі генеральної сукупності та будуть відрізнятися від обчислених на основі вибірки.

Сформулюємо *теорему Байеса для задачі класифікації*.

В уведених вище позначеннях рівняння

$$\Pr(W_i | X) = \frac{q_i \Pr(X | W_i)}{q_1 \Pr(X | W_1) + q_2 \Pr(X | W_2)}, \quad i = 1, 2. \quad (7.10)$$

є справедливим для будь-якого розподілу величини X .

Якщо X має багатовимірний нормальний розподіл $N(\mu_1, \Sigma)$ чи $N(\mu_2, \Sigma)$, то $\Pr(X|W_1)$ і $\Pr(X|W_2)$ можна замінити відповідно на щільності розподілів $f_1(x)$ і $f_2(x)$. У результаті отримаємо

$$\Pr(W_i | X) = \frac{q_i f_i(x)}{q_1 f_1(x) + q_2 f_2(x)}, \quad i = 1, 2. \quad (7.11)$$

Байесівська процедура класифікації полягає у віднесенні вектора спостережень X до W_1 , якщо $\Pr(X|W_1) \geq \Pr(X|W_2)$, і до W_2 , якщо $\Pr(X|W_1) < \Pr(X|W_2)$. Підставляючи в ці нерівності значення апостеріорних ймовірностей з (7.11), отримуємо таке правило класифікації: X відноситься до популяції W_1 , якщо

$$\frac{q_1 f_1(x)}{q_2 f_2(x)} \geq 1, \quad (7.12)$$

і до W_2 , якщо

$$\frac{q_1 f_1(x)}{q_2 f_2(x)} < 1. \quad (7.13)$$

Така процедура мінімізує очікувану ймовірність помилкової класифікації [8]

$$q_1 \Pr(2|1) + q_2 \Pr(1|2). \quad (7.14)$$

Зауважимо, що ця величина є ймовірністю того, що об'єкт, який належить до популяції W_1 , помилково класифікується як приналежний до W_2 , чи навпаки, об'єкт з W_2 помилково відноситься до W_1 .

Алгебраїчними перетвореннями нерівності (7.12) можна показати, що байесівська процедура еквівалентна віднесенню X до W_1 , якщо

$$\sum_{j=1}^p \alpha_j x_j \geq \frac{\xi_1 + \xi_2}{2} + \ln \left(\frac{q_2}{q_1} \right) \quad (7.15)$$

і до W_2 , якщо

$$\sum_{j=1}^p \alpha_j x_j < \frac{\xi_1 + \xi_2}{2} + \ln \left(\frac{q_2}{q_1} \right). \quad (7.16)$$

Сталі α_j є розв'язками системи рівнянь (7.6), а ξ_1 і ξ_2 задаються відповідно виразами (7.4) і (7.5). Зауважимо, що при $q_1 = q_2 = 1/2$ байесівська процедура аналогічна евристичній процедурі класифікації у відповідності до (7.2).

У випадку відомих багатовимірних нормальних розподілів популяцій апостеріорна ймовірність того, що об'єкт належить до W_1 , має вигляд

$$\Pr(W_1 | X) = \frac{1}{1 + \frac{q_2}{q_1} \exp \left\{ -z + \frac{\xi_1 + \xi_2}{2} \right\}},$$

де Z задається рівністю (7.1), а ξ_1 і ξ_2 — формулами (7.3) і (7.5) відповідно. Для апостеріорної ймовірності виконується рівність $\Pr(W_2 | X) = 1 - \Pr(W_1 | X)$. При використанні оцінок параметрів можна замінити ξ_1 на $\bar{Z}_i$, $i = 1, 2$.

Подальше вдосконалення байесівської процедури пов'язане з включенням у неї вартості помилкової класифікації. Для цього введемо величину $C(2|1)$ — вартість втрати через віднесення об'єкта з W_1 до популяції W_2 . Аналогічно, $C(1|2)$ являє собою вартість втрати через віднесення об'єкта з W_2 до W_1 .

Узагальнена процедура класифікації Байеса полягає у віднесенні X до W_1 , якщо

$$\sum_{j=1}^p \alpha_j x_j \geq \frac{\xi_1 + \xi_2}{2} + \ln \frac{q_2 C(1|2)}{q_1 C(2|1)}, \quad (7.17)$$

і до W_2 , якщо

$$\sum_{j=1}^p \alpha_j x_j < \frac{\xi_1 + \xi_2}{2} + \ln \frac{q_2 C(1|2)}{q_1 C(2|1)}. \quad (7.18)$$

Така процедура мінімізує очікувану вартість помилкової класифікації

$$q_1 C(2|1) \Pr(2|1) + q_2 C(1|2) \Pr(1|2). \quad (7.19)$$

Ця процедура зводиться до байесівської процедури, коли вартості однакові, і до евристичної, коли при цьому $q_1 = q_2 = 1/2$.

Для узагальненої байесівської процедури ймовірності помилкової класифікації мають вигляд

$$\Pr(2|1) = \Phi \left(\frac{K - \frac{1}{2} \Delta^2}{\Delta} \right) \quad (7.20)$$

і

$$\Pr(1|2) = \Phi \left(\frac{-K - \frac{1}{2} \Delta^2}{\Delta} \right), \quad (7.21)$$

де

$$K = \ln \frac{q_2 C(1|2)}{q_1 C(2|1)}, \quad (7.22)$$

а Δ^2 задається рівністю (7.6). Зауважимо, що за умови $C(1|2) = C(2|1)$ і $q_1 = q_2 = 1/2$,

$$\Pr(2|1) = \Pr(1|2) = \Phi\left(-\frac{\Delta}{2}\right). \quad (7.23)$$

Можна показати, що набір $\alpha_1, \alpha_2, \dots, \alpha_p$, отриманий множенням розв'язку системи (7.7) на будь-яке додатне число, також максимізує Δ^2 . Якщо $\alpha_1, \alpha_2, \dots, \alpha_p$ помножені на додатну сталу, то величина $K = \ln \frac{q_2 C(1|2)}{q_1 C(2|1)}$, що зустрічається в процедурі класифікації, збільшується на ту ж сталу.

Величини $\Pr(2|1)$ і $\Pr(1|2)$ є спадними функціями від Δ^2 , тобто імовірності помилкової класифікації спадають, коли відстань між двома популяціями зростає.

У випадку, коли параметри розподілів відомі, значення імовірностей помилкової класифікації $\Pr(2|1)$ і $\Pr(1|2)$ визначаються за формулами (7.20–7.21). У випадку, коли параметри розподілів невідомі, у формулах (7.20–7.21) можна замінити Δ^2 на D^2 . Слід, однак, зауважити, що такі оцінки будуть зміщеними, тобто дійсна ймовірність помилкової класифікації буде в середньому більша, ніж така оцінка. Перевагою методу є простота отримання таких оцінок [8].

Розв'язання системи (7.7) можна записати в матричних позначеннях: $\alpha = \Sigma^{-1}(\mu_1 - \mu_2)$, де $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_p)$.

Підстановка у (7.6) дозволяє записати відстань Махаланобіса у вигляді $\Delta^2 = (\mu_1 - \mu_2) \Sigma^{-1} (\mu_1 - \mu_2)$.

Якщо X належить до однієї з двох відомих популяцій з довільними функціями щільності $f_1(x)$ і $f_2(x)$ відповідно, то узагальнена байєсівська процедура зводиться до віднесення X до W_1 , якщо $\frac{q_1 C(2|1) f_1(x)}{q_2 C(1|2) f_2(x)} \geq 1$, і до W_2 у протилежному випадку.

Приклад 7.1 [8]

Приймальна комісія коледжу має на меті віднести кожного абітурієнта або до популяції W_1 — студентів, що успішно закінчать коледж, або до популяції W_2 — тих, хто не закінчить коледж. Рішення комісії складається на основі оцінок абітурієнта за p вступними іспитами.

Розв'язання

Припустимо, що приймальна комісія враховує оцінки, отримані абітурієнтами на $p = 2$ іспитах. Нехай $X = (x_1, x_2)$ — вектор

оцінок абітурієнта. З досвіду попередніх років відомо, що $\mu_1 = (60, 57)$, $\mu_2 = (42, 39)$ і $\sum \begin{bmatrix} 100 & 70 \\ 70 & 100 \end{bmatrix}$. Нехай $q_1 = 1/3$, $q_2 = 2/3$ і приймемо для ілюстрації, що $C(1|2) = 2000$ і $C(2|1) = 3000$ дол., де C — «вартість» помилкової класифікації. Підставляючи ці значення в систему рівнянь (7.7), отримуємо $100\alpha_1 + 70\alpha_2 = 18$; $70\alpha_1 + 100\alpha_2 = 18$, звідки $\alpha_1 = \alpha_2 = 54/510$. Отже, дискримінантна функція має вигляд $Z = (54/510)(x_1 + x_2)$. Згідно з (7.4), $\xi_1 = (54/510)(60 + 57) = 12,39$, а за формулою (7.5) маємо $\xi_2 = 8,58$. Отже, (7.8) і (7.22) відповідно дають $(12,39 + 8,58)/2 = 10,49$ і $K = \ln(4/3) = 0,288$. Отже, узагальнена байесівська процедура відносить X до популяції W_1 , якщо $(54/510)(x_1 + x_2) \geq 10,49 + 0,288$, тобто $x_1 + x_2 \geq 101,79$. Якщо ж $x_1 + x_2 < 101,79$, то вектор відноситься до W_2 . Величина σ^2 (7.4) дорівнює 3,81 і відстань Махаланобіса Δ^2 (7.6) також дорівнює 3,81. Потім за формулами (7.20–7.21) можна отримати ймовірності помилкової класифікації: $\Pr(2|1) = \Phi(-0,83) = 0,203$; $\Pr(1|2) = \Phi(-1,12) = 0,131$. Отже, абітурієнт приймається, якщо лінійна комбінація його оцінок більше чи дорівнює 101,79 і не приймається в іншому випадку. За такої процедури 20,3 % потенційно добрих студентів відкидається комісією і приймається 13,1 % потенційно поганих.

7.4. Класифікація у випадку двох багатовимірних нормальних популяцій за невідомих параметрів

Нехай є об'єкт, якому відповідає вектор спостережень $X = (x_1, x_2, \dots, x_p)$. Потрібно віднести його на основі цих спостережень до класу W_1 з розподілом $N(\mu_1^{p \times 1}, \Sigma^{p \times p})$ або до W_2 : $N(\mu_2^{p \times 1}, \Sigma^{p \times p})$ [8]. Передбачається, що відомі апіорні ймовірності і вартості помилкової класифікації, але середні μ_1, μ_2 і матриця коваріацій Σ невідомі. Якщо $x_{11}, \dots, x_{1n_1}$ і $x_{21}, \dots, x_{2n_2}$ — незалежні випадкові вибірки з класів W_1 і W_2 відповідно, то можна оцінити μ_1 вибіркоким вектором середніх $\bar{x}_i = (x_{i1}, \dots, x_{ip})$, $i = 1, 2$, а Σ — об'єднаною вибірковою коваріаційною матрицею $S = (S_{vj})$, $j = 1, \dots, p$; $v = 1, \dots, p$. Оскільки наведені вище оцінки обґрунтовані, узагальнена процедура байесівської класифікації на основі оцінок параметрів полягає в такому: спочатку розв'язується система рівнянь (7.7) із заміною μ_{ij} на $\bar{x}_{ji}$, $i = 1, 2$, $j = 1, \dots, p$, і заміною S_{vj} на $\hat{S}_{vj}$, $v = 1, \dots, p$. Потім отримані оцінки коефіцієнтів

$\alpha_1, \dots, \alpha_p$ (позначимо їх $a_1, \dots, a_p$) використовуються для визначення значення дискримінантної функції Z_{il} (7.2) для кожного вектора спостережень x_{il} , $l = 1, \dots, n_i$. Далі ξ_i , задані формулами (7.3) і (7.5), оцінюються величинами

$$\bar{Z}_i = \frac{1}{n_i} \sum_{l=1}^{n_i} Z_{il}, \quad (7.24)$$

а σ^2 , задані (7.3), — величиною

$$\hat{\sigma}_z^2 = \sum_{j=1}^p \sum_{v=1}^p \alpha_i S_{iv} \alpha_v. \quad (7.25)$$

Таким чином, узагальнена байєсівська процедура оцінювання полягає у віднесенні $X = (x_1, x_2, \dots, x_p)$ до W_1 , якщо

$$Z = \sum_{j=1}^p \alpha_j x_j \geq \frac{\bar{Z}_1 + \bar{Z}_2}{2} + \ln \frac{q_2 C(1|2)}{q_1 C(2|1)} \quad (7.26)$$

і до W_2 у протилежному випадку.

Вибіркова відстань Махаланобіса

$$D^2 = \frac{(\bar{Z}_1 - \bar{Z}_2)^2}{\sigma_z^2} \quad (7.27)$$

є оцінкою для Δ^2 (7.6). Отже, алгоритм роботи дискримінантного аналізу реалізується в такий спосіб:

- ☒ визначаються коефіцієнти дискримінантної функції $\alpha_1, \dots, \alpha_p$;
- ☒ оцінюються значення дискримінантної функції Z_{il} для кожного вектора спостережень x_{il} , $i = 1, 2$; $l = 1, \dots, n_i$;
- ☒ визначаються вибіркові середні $\bar{Z}_1$ і $\bar{Z}_2$;
- ☒ розраховується вибіркова відстань Махаланобіса D^2 ;
- ☒ реалізується процедура класифікації у відповідності до виразу (7.26).

Часто буває складно визначити апіорні ймовірності q_1 та q_2 . Якщо випадкові вибірки обсягу n_1 і n_2 беруться відповідно з класів W_1 і W_2 , то q_1 і q_2 можна оцінити величинами

$$\hat{q}_1 = \frac{n_1}{n_1 + n_2} \quad \text{і} \quad \hat{q}_2 = \frac{n_2}{n_1 + n_2}.$$

У результаті роботи програми дискримінантного аналізу зазвичай виводиться значення F -статистики

$$F = \frac{n_1 + n_2 - p - 1}{p(n_1 + n_2 - 2)} \frac{n_1 n_2}{n_1 + n_2}, \quad (7.28)$$

яке можна використовувати для перевірки гіпотези $H_0: \Delta^2 = 0$. Числа степенів вільності F дорівнюють p та $(n_1 + n_2 - p - 1)$.

7.5. Покроковий дискримінантний аналіз

Дослідники часто зустрічаються з ситуаціями, коли в їхньому розпорядженні виявляється декілька можливих дискримінантних змінних, а вони не впевнені, чи всі з цих змінних корисні й необхідні. Подібні ситуації часто виникають, коли важко привести точний список дискримінантних змінних. У результаті збираються дані про всі змінні, котрі, як передбачається, являються добрими дискримінаторами, чи ж дослідження носить попередній характер і фахівці намагаються знайти корисні дискримінантні змінні.

У цих ситуаціях одна чи більше змінних можуть виявитися поганими дискримінаторами, оскільки середні класи слабо розрізняються за цими змінними. Крім того, дві чи більше змінних можуть нести однакову інформацію, хоча кожна є добрим дискримінатором. Якщо деякі з них зайняті в аналізі, решта виявляються зайвими і не роблять жодного внеску в аналіз (хоча самі по собі вони можуть бути добрими дискримінаторами), тому що в них замало нової інформації. Якщо немає переконливих теоретичних міркувань на користь збереження таких «надлишкових» змінних, їх рекомендується виключати, оскільки вони тільки ускладнюють аналіз і можуть навіть збільшити число неправильних класифікацій. Для оцінки внеску окремої змінної в значення дискримінантної функції доцільно користатися стандартизованими коефіцієнтами дискримінантної функції.

Стандартизовані коефіцієнти застосовують у випадках, коли потрібно визначити, яка з використовуваних змінних вносить найбільший вклад у величину дискримінантної функції. Крім визначення внеску кожної вихідної змінної в дискримінантну функцію, можна проаналізувати і ступінь кореляційної залежності між ними.

Для оцінки щільності зв'язку між окремими змінними і дискримінантними функціями служать коефіцієнти кореляції.

Один зі способів виключення непотрібних змінних полягає у використанні процедури послідовного відбору найбільш корисних дискримінантних змінних. Пряма процедура послідовного відбору починається з вибору змінної, що забезпечує найкраще одновимірне розрізнення. Потім аналізуються пари, утворені відбіраною й однією з змінних, що залишилися, після чого знаходиться пара, що дає найкраще розрізнення, з якої й відбирається змінна. Далі процедура переходить до утворення трійок з перших двох і кожної зі змінних, що залишилися. Найкраща трійка визначає третю змінну. На кожному кроці цієї процедури відбирається змінна, яка у сполученні з відібраними раніше дає найкраще розрізнення. Процес продовжується доти, доки не будуть розглянуті всі можливі змінні чи доки змінні, що залишилися, не припинять поліпшувати розрізнення.

Процедура послідовного відбору може працювати й у зворотному напрямку; тобто коли всі змінні спочатку вважаються «вхідними» у систему, а потім на кожному кроці відкидається одна, найгірша. Прямий і зворотний відбори можуть сполучатися, але частіше застосовується пряма процедура. Якщо яка-небудь змінна більше не дає значущого внеску в процес розрізнення, то вона відкидається, але на наступному кроці може бути знову відібрана. Усунення раніше відбіраної змінної відбувається тому, що вона в значній мірі містить ту ж дискримінантну інформацію, що й інші змінні, відібрані на попередніх кроках. На той час, коли ця змінна відбиралася, вона вносила істотний вклад у процес розрізнення. Однак змінні, відібрані на наступних кроках, у сполученні з однією чи декількома, відібраними раніше, дублюють цей внесок, отже змінна стає надлишковою й видаляється.

Процедури послідовного відбору породжують оптимальну множину дискримінантних змінних, яка може не бути максимальною (найкращою) комбінацією. Щоб отримати максимальний розв'язок, потрібно перевірити всі можливі сполучення (пари, трійки тощо). Така перевірка може виявитися дорогою і потребувати великих часових витрат. Процедура послідовного відбору є логічним і ефективним способом пошуку кращої комбінації, але немає гарантії, що її кінцевий продукт дійсно перевершує всі інші.

Послідовність, у якій відбираються змінні, не обов'язково відповідає їх відносній значущості. Внаслідок корельованості

(що розділяє дискримінантні можливості) навіть добрі дискримінатори можуть пізно потрапити чи взагалі не потрапити у дослідність, оскільки їхній внесок у розрізнення може виявитися менше внеску інших змінних.

Процедури послідовного відбору мають використовувати деяку міру якості розрізнення як критерій відбору. Існують різні критерії відбору змінних, що дозволяють ефективно розділяти класи, вибір яких залежить від статистичних характеристик вихідної інформації. Розглянемо деякі з них.

Одним з таких критеріїв являється Λ -статистика Уїлкса та часткове F -відношення.

Λ -статистика Уїлкса враховує як розбіжність між класами, так однорідність кожного класу. Під однорідністю слід розуміти ступінь скупчення об'єктів навколо центроїда їхнього класу. Тому змінна, яка збільшує однорідність не змінюючи поділ центроїдів, при відборі може мати перевагу щодо змінної, яка збільшує поділ без зміни однорідності.

Оскільки Λ -статистика Уїлкса являється «оберненою» статистикою, ми будемо відбирати ту змінну, для якої на цьому кроці вона приймає найменше значення. Можливо перетворити Λ -статистику на повну F -статистику для перевірки розбіжностей між класами. Якщо таке перетворення відбувається, то вибір робиться за найбільшим значенням. Замість повного F -відношення використовується часткове F -відношення, яке обчислюється так само, як і значення F -включення. Використання усіх трьох статистик приведе до однакового результату [8].

Міра, що вимірює поділ центроїдів класів і не чіпає однорідності всередині класів, отримала назву *V-статистика Рао*. Ця узагальнена міра відстані припустима при будь-якій кількості класів. Застосовуючи відстань Махаланобіса, V -статистика вимірює відстані від кожного центроїда класу до головного центроїда з вагами, що дорівнюють розміру відповідного класу. Отже, V -статистика не забезпечує максимального поділу між усіма парами класів.

Коли розглядається велике число об'єктів, V -статистика має вибірковий розподіл, що приблизно збігається з розподілом χ^2 з відповідними степенями вільності.

Крім того, зміна V -статистики, викликана додаванням (чи видаленням) змінних, також має розподіл χ^2 з числом степенів вільності, помножений на число змінних, доданих (вилучених) на цьому кроці. Ця властивість використовується при перевірці

статистичної значущості зміни загального поділу. Якщо зміна не являється значимою, змінну можна не включати. При додаванні змінних зміна V -статистики може виявитися негативною, що означає погіршення поділу центроїді [42].

Розглянуті статистики використовують квадрат відстані Махаланобіса між центроїдами двох класів.

Міжгрупова F -статистика використовується, коли необхідно збільшити розбіжність між парами, що містять великі групи, і враховує вибіркові розміри класів. Відстані для малих класів отримують менші ваги, ніж відстані для великих класів.

Критерій, призначений для мінімізації залишкової дисперсії між класами, сприяє формуванню рівномірного поділу класів. Цей критерій відрізняється від перших двох, у яких два класи можуть залишатися близькими один одному, а значне поліпшення поділу отримане для інших класів або за рахунок збільшення внутрішньогрупової однорідності. Він відрізняється від третього й четвертого критеріїв, у яких основна увага звертається лише на найщільнішу пару.

Тест толерантності може забезпечити необхідну точність обчислень. Толерантність ще не відібраної змінної дорівнює одиниці мінус квадрат множинної кореляції між цією змінною й усіма вже відібраними змінними, коли кореляції визначаються за внутрішньогруповою кореляційною матрицею. Якщо змінна, що перевіряється, є лінійною комбінацією (чи приблизно дорівнює лінійній комбінації) однієї чи декількох відібраних змінних, то її толерантність дорівнює нулю (чи близька до нуля). Змінна з малою толерантністю (наприклад, менше 0,001) може призвести до помилки в обчисленні матриці, оберненої до W , через швидке накопичення помилок округлення. Змінну, яка є лінійною комбінацією відібраних змінних, небажано використовувати, тому що вона не дає ніякої нової інформації, крім обчислювальних проблем.

У пакетах прикладних програм часто використовується статистика F -включення, яка являє собою часткову F -статистику, що оцінює поліпшення розрізнення від використання змінної, що розглядається, у порівнянні з розрізненням, досягнутим за допомогою інших, уже відібраних змінних.

Статистика F -видалення також є окремою F -статистикою. Однак вона оцінює значущість погіршення розрізнення після видалення змінної зі списку вже відібраних змінних. Ця процедура

проводиться на початку кожного кроку, щоб перевірити, чи є яка-небудь змінна, що вже не вносить досить великого вкладу в розрізнення, оскільки відібрані пізніше змінні дублюють її внесок.

На заключному кроці статистика F -видалення може бути використана для ранжування дискримінантних можливостей відібраних змінних. Змінна з найбільшим значенням статистики F -віддалення дає найбільший внесок у розрізнення, досягнуте завдяки іншим змінним. Змінна, що має другу за величиною статистику F -видалення, є другою за значущістю тощо. Це ранжування не обов'язково збігається з тим, яке можна було б отримати за допомогою одновимірної F -статистики, тому що вона вимірює повну дискримінантну здатність змінної без урахування дублювання її іншими змінними.

Розглянемо роботу *алгоритму покрокового дискримінантного аналізу*. Нехай необхідно віднести p -вимірний вектор спостережень $X = (x_1, x_2, \dots, x_p)$ до однієї з k популяцій W_i , об'єкти яких мають багатовимірні нормальні розподіли $N(\mu_i^{p \times 1}, \Sigma^{p \times p})$, де $\mu_i = (\mu_{i1}, \dots, \mu_{ip})$, $i = 1, \dots, k$. Оскільки X служить реалізацією випадкового вектора $X = (x_1, \dots, x_p)$, для поділу k популяцій можна використовувати всі змінні $x_1, \dots, x_p$. Однак на практиці часто потрібно виявити таку підмножину цих змінних, за якою можна побудувати найкращий поділ k популяцій. Для відбору змінних використовується F -статистика, що ґрунтується на критерії однофакторного дисперсійного аналізу.

По суті логіка покрокового аналізу така: на початку визначається змінна, для якої середні значення в k популяціях найбільш різні. Для кожної змінної розходження вимірюється за допомогою F -статистики однофакторного дисперсійного аналізу і вибирає (чи включається) та змінна, якій відповідає найбільше значення F . На кожному кроці процедури розглядається умовний розподіл кожної змінної, не включеної в підмножину, при заданих включених змінних.

З числа не включених змінних визначається змінна, для якої середні значення умовних розподілів у k популяціях найбільш різні. Це розходження вимірюється за допомогою F -статистики однофакторного дисперсійного аналізу. Процес завершується, коли жодна з змінних, що залишилися не вносить значимого внеску в поділ k популяцій. Як і в покроковій регресії показник вибирає припустимий мінімум F -включення, що відповідає максимальному рівню α (стандартне значення дорівнює 4,0) і мінімум

F -видалення (стандартне значення дорівнює 3,9), причому мінімум F -видалення має бути менше, ніж мінімум F -включення.

Розглянемо покрокову процедуру детальніше, для цього нехай $x_{i1}^{p \times 1}, \dots, x_{in_i}^{p \times 1}$ — випадкова вибірка з W_i , $i = l, \dots, k$. Тоді, використовуючи позначення і визначення цього розділу, можна описати покрокову процедуру в такий спосіб.

Крок 0. Для кожної змінної x_j , $j = l, \dots, p$, аналогічно F -статистики для перевірки гіпотези $H_0: \mu_{1j} = \dots = \mu_{kj}$ в однофакторному дисперсійному аналізі обчислюється статистика F -включення з $(k - l)$ і $(n - k)$ степенями вільності. Якщо всі значення F -включення менше прийнятого мінімуму, то вважається, що жодна змінна не вносить вагомого внеску в поділ популяцій.

Крок 1. Змінна x_{j1} , який відповідає найбільше значення F -включення, вважається першою. Для кожної популяції W_i , $i = l, \dots, k$, оцінюється коефіцієнт і визначається стала лінійної дискримінантної функції. Крім того, обчислюється таблиця результатів класифікації, U -статистика і її F -апроксимація. Обчислюється також значення F -видалення з $(k - 1)$ і $(n - k)$ степенями вільності для змінної x_{j1} , яке дорівнює значенню F -включення. Потім знаходиться значення F -включення з $(k - 1)$ і $(n - k - 1)$ степенями вільності для кожної з змінних, не включених у шукану підмножину. Таким чином, перевіряється гіпотеза $H_0: \mu_{1j1} = \dots = \mu_{kj1}$, де μ_{ij1} — середнє умовного розподілу x_j у популяції W_i при фіксованому значенні змінної x_{j1} $i = l, \dots, k$; $j = l, \dots, p$; $j \neq j_1$. Якщо всі значення F -включення менше його мінімуму, то виконується крок S , у протилежному випадку процедура переходить до виконання другого кроку.

Крок 2. Вибирається змінна x_{j2} , для якої значення F -включення є максимальним. Обчислюються оцінки двох коефіцієнтів і сталі дискримінантних функцій для кожної популяції W_i , $i = l, \dots, k$. Визначається таблиця результатів класифікації, U -статистика і її F -апроксимація. Крім того, для x_{j1} і x_{j2} шукаються значення статистик F -видалення з $(k - 1)$ і $(n - k - 1)$ степенями вільності. Таким чином, перевіряються відповідно гіпотези $H_0: \mu_{1j1j2} = \dots = \mu_{kj1j2}$ і $\mu_{1j2j1} = \dots = \mu_{kj2j1}$. Потім для кожної не включеної змінної обчислюється статистика F -включення з $(k - 1)$ та $(n - k - 2)$ степенями вільності. З її допомогою перевіряється гіпотеза $H_0: \mu_{1jj1j2} = \dots = \mu_{kjj1j2}$, де μ_{ijj1j2} — середнє умовного розподілу x_j при заданих x_{j1} і x_{j2} у популяції W_i , $i = l, \dots, k$, $j = l, \dots, p$, $j \neq j_1, j_2$. Якщо всі значення F -включення менше встановленого

мінімуму, то виконується крок S , в протилежному випадку — крок 3.

Крок 3. а) Позначимо символом L множину з l змінних, попередньо відібраних для побудови процедури класифікації. Якщо значення F -видалення менше прийнятого мінімуму хоча б для однієї змінної з L , то змінна, для якої це значення мінімальне, виключається з L і виконується (б), у якому l замінюється на $(l - 1)$. Якщо тепер деякі значення F -включення для змінних, не приналежних L , більше прийнятого для цієї статистики мінімуму, то змінна, для якої це значення максимальне, включається в L і l замінюється на $(l + 1)$.

б) Для кожної популяції W_i , $i = 1, \dots, k$, оцінюються l коефіцієнтів дискримінантної функції й обчислюється стала. Визначається таблиця результатів класифікації, значення U -статистики і її F -апроксимація. Крім того, для кожної змінної з L обчислюється значення F -видалення і відповідні степені вільності. Тим самим перевіряється гіпотеза $H_0: \mu_{1s(l-1)} = \dots = \mu_{ks(l-1)}$ для кожної змінної x_l з L при фіксованих значеннях інших $(l - 1)$ змінних у L . Символ $\mu_{is(l-1)}$ позначає середнє умовного розподілу x_s у W_i при фіксованих значеннях інших змінних у L . І нарешті, обчислюється значення статистики F -включення і відповідних степенів вільності для кожної змінної, не включеної в L . Таким чином, перевіряється гіпотеза $H_0: \mu_{1j(l)} = \dots = \mu_{kj(l)}$, де $\mu_{ij(l)}$ позначає середнє значення умовного розподілу змінної x_j у W_i при фіксованих значеннях усіх змінних з L , $i = 1, \dots, k$; $j = 1, \dots, p$; x_j не належить до L .

Кроки 4, 5, ... Рекурентно здійснюється повторення кроку 3, якщо F -включення стають менше заданого мінімуму для всіх змінних, не включених у L , чи коли всі змінні виявляються включеними у шукану підмножину і значення F -видалення стає менше заданого мінімуму, виконується крок S . У деяких програмах рекурентна процедура зупиняється також і в тому випадку, коли l дорівнює $\min(n_i)$, $i = 1, \dots, k$.

Крок S . На цьому кроці для кожного вектора x_{im} , $m = 1, \dots, n_i$, $i = 1, \dots, k$, проводиться обчислення апостеріорних імовірностей його приналежності до популяції $W_1, \dots, W_k$. На підставі цих імовірностей кожен об'єкт класифікується як приналежний одній з популяцій і складається таблиця результатів класифікації.

Зазвичай, якщо користувач вимагає, можна вивести на друк таблицю, у якій відображаються дії процедури на кожному її

кроці. Крім того, на кожному кроці виводяться такі необхідні величини, як номер кроку, включені та видалені змінні, значення статистик F -включення і F -видалення, U -статистика і її F -апроксимація.

7.6. Оцінка застосовності дискримінантного аналізу при порушенні припущень

На практиці вихідні дані часто не погоджують з припущеннями ДА. Важче за все задовольнити вимогам про нормальність багатовимірною розподілу дискримінантних змінних і рівняння коваріаційних матриць класів. Припущення про нормальність багатовимірною розподілу важливе для перевірки значущості, де зіставляються статистики, обчислені за вибірковими даними, з теоретичним імовірнісним розподілом для цієї статистики. Можна обчислити теоретичний розподіл, зробивши деякі зручні математичні припущення (наприклад, такі, як вимога, щоб генеральна сукупність мала багатовимірний нормальний розподіл). Якщо цікавляча нас генеральна сукупність не задовольняє цій вимозі, істинний вибірковий розподіл статистики буде відрізнятися від розподілу, отриманого теоретично. Розходження між цими двома розподілами можуть бути дуже малими або дуже великими в залежності від ступеня порушення припущень. У сучасних роботах показано, що дискримінантний аналіз не дуже чутливий до невеликих порушень припущення про нормальність. Це призводить лише до деяких втрат в ефективності й точності.

Припущення про нормальність розподілу відіграє важливу роль у класифікації, що ґрунтується на використанні імовірності приналежності до класу. Ці імовірності обчислюються виходячи з розподілу χ^2 , що виправдано лише, коли дискримінантні змінні мають багатовимірний нормальний розподіл. Якщо це припущення не виконується, обчислені імовірності будуть неточними. Може виявитися, наприклад, що імовірності для деяких груп будуть перебільшені, у той час як імовірності для інших груп — недооцінені. Отже, ця процедура не буде оптимальною в розумінні зменшення числа неправильних класифікацій.

Якщо коваріаційні матриці класів не рівні, бажано встановити причини викривлення дискримінантних функцій і рівнянь класифікації. Одне джерело помилок пов'язане з обчисленням

внутрішньогрупової коваріаційної матриці (чи інших, що мають відношення до матриці W). Внутрішньогрупова коваріаційна матриця служить оцінкою загальної коваріаційної матриці класів для генеральної сукупності, утвореної вибірками з декількох класів. Якщо матриці для всієї генеральної сукупності не рівні, матрицю W усе ще можна обчислити, але вона вже не буде сприяти спрощенню різних формул. Отже, канонічні дискримінантні функції не дадуть максимального поділу класів і імовірності приналежності до класів будуть викривлені. Пропонується використовувати коваріаційні матриці окремих класів для обчислення ймовірності приналежності до класу (так звана «квадратична дискримінація»).

ДА може бути проведений і коли припущення про нормальність багатовимірного розподілу і рівність коваріаційних матриць класів не виконуються. Задача при цьому полягає в інтерпретації результатів. Що вони означають? І яка кількість помилок вважається припустимою? У деяких підручниках пропонуються можливі процедури, але вони призводять лише до мінімальних поліпшень, оскільки вихідні відхилення не були великими. Звичайно, нам важко довідатися, скільки помилок було зроблено у зв'язку з конкретними порушеннями припущень. Однак тут можуть виявитися корисними деякі статистики, що не залежать від цих припущень.

При визначенні значущості і мінімального числа канонічних дискримінантних функцій ми не покладаємося на Λ -статистику Уїлкса чи пов'язаний з нею тест значущості, заснований на χ^2 -розподілі. Замість цього ми можемо розглянути канонічну кореляцію і відносний відсотковий вміст. Якщо будь-яка з цих величин виявиться невеликою, функція буде для нас малоціковою, навіть якщо вона статистично значима. Тести значущості становлять найбільший інтерес у випадку малих вибірок. Таким чином, маючи з ними справу, ми повинні з великою увагою поставитися до задоволення припущень. Однак у випадку великих вибірок ми можемо обійтися без тестів значущості або використовувати їх «консервативно», коли наші дані порушують припущення.

При класифікації точність передбачення найбільш важлива для об'єктів, розташованих поблизу межі. Якщо деякий об'єкт з імовірністю 0,90 належить до класу 1 і тільки з імовірністю 0,10 — до класу 2, то нам нема чого турбуватися про невеликі неточності, що виникають через порушення припущення. Хоча

визначена імовірність приналежності до класу може бути невісною, наше рішення приписати об'єкт до класу 1 буде правильним, якщо помилка в обчисленні ймовірностей не буде великою. З іншого боку, якщо об'єкт має імовірності 0,51 для класу 1 і 0,49 для класу 2, ми повинні бути дуже обережні, приймаючи рішення. Тут невелика помилка через порушення припущень може привести до неправильної класифікації.

На ефективність і інтерпретацію результатів дискримінантного аналізу можуть вплинути сильно корельовані змінні, велика кількість відсутніх даних, змінна з нульовим стандартним відхиленням всередині одного чи декількох класів, великі розходження в розмірах класів і викиди.

У випадку, коли не виконується припущення про рівність коваріаційних матриць, тобто кожна популяція W_i має розподіл $N(\mu_i, \Sigma_i)$ для $i = 1, \dots, k$, можна отримати квадратичну дискримінантну функцію.

При необхідності можна проводити розбиття множини об'єктів на k класів (при $k > 2$). У цьому випадку потрібно розрахувати k дискримінантних функцій, оскільки класи будуть відокремлюватися одне від одного індивідуальними поділяючими поверхнями.

Зміна числа змінних сильно впливає на результат ДА. Щоб судити про доцільність включення (видалення) дискримінантної змінної, зазвичай використовують спеціальні статистичні критерії, що дозволяють оцінити значимість погіршення чи поліпшення розбиття після включення (видалення) кожної з відібраних змінних. Математична модель зазвичай припускає, що змінні вимірюються за інтервальною шкалою і мають багатовимірний нормальний розподіл. Ми обмежилися обговоренням лінійного дискримінантного аналізу, який зазвичай вимагає рівності коваріаційних матриць класів. Маючи це на увазі, дослідник може широко використовувати підпрограми дискримінантного аналізу в стандартних пакетах комп'ютерних програм.

7.7. Використання ППП «STATISTICA» для розв'язання задач класифікації

Знайомство з можливостями проведення ДА в системі STATISTICA найкраще почати з розбору апробованого прикладу. Таким прикладом є класичний приклад Фішера — аналіз квіток

ірису. Завдання полягає в тому, щоб за результатами вимірювання довжини і ширини чашолистків і пелюсток квіток ірису віднести квітку до одного з трьох типів: **SETOSA**, **VERSICOL**, **VIRGINIC**. Дані для цього прикладу (результати вимірювань 150 квіток ірису, по 50 кожного типу) містяться у файлі **Irisdat.sta**. Аналіз виконуватимемо послідовно по кроках.

Крок 1. У ППП вибираємо **Statistics → Multivariate Exploratory Techniques → Discriminant Analysis**. Отримуємо вікно діалогу вибору аналізу, що використовується (рис. 7.5).

Крок 2. Після натискання кнопки **Open Data** (Відкрити дані) на екрані з'явиться файл даних **Irisdat.sta** з каталогу **Example**: Кнопка **Open Data** дозволяє відкрити файл даних для аналізу.

За допомогою кнопки **Variables** (Змінні), розташованої у лівому верхньому куті панелі, відкривається вікно **Select one grouping variable and independent variable list** (Вибрати одну групову змінну і список незалежних змінних). Групова змінна не повинна входити до списку незалежних змінних.

Кнопка **Codes for grouping variable** (Коди для групової змінної) відкриває вікно, де можна задати коди для назви груп, до яких належить об'єкт.

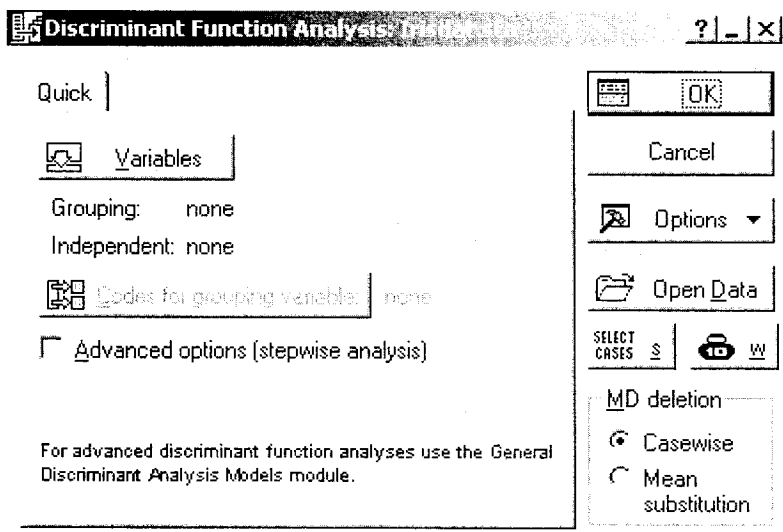


Рис. 7.5. Стартова панель модуля Дискримінантний аналіз

Опція **Missing data (Пропущені дані)** дозволяє задати спосіб обробки пропущених даних. Можливо два способи: підстановка середнього і видалення пропущених спостережень.

Кнопка **CANCEL** закриває стартову панель, кнопка **OK** відкриває діалогове вікно **Model Definition (Визначення моделі)**.

Кнопка **Advanced options (sterwise analysis)** призначена для вибору покрокового методу.

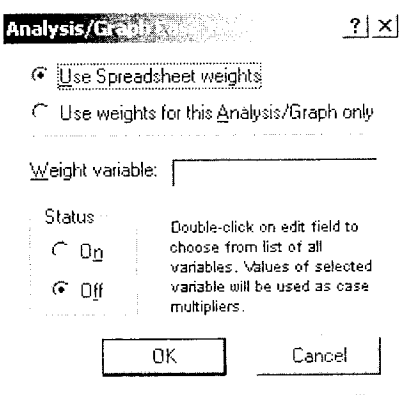


Рис. 7.6. Вікно визначення вагової змінної

Дві кнопки у правому нижньому куті вікна мають таке функціональне призначення. Кнопка **SELECT CASES (Вибрати спостереження)** відкриває діалогове вікно, у якому задаються умови вибору спостережень. Кнопка **W** відкриває діалогове вікно **Define weight (Визначити вагу)** (див. рис. 7.6).

У цьому вікні можна вибрати вагову змінну, яка дозволяє визначити (задати) ваги для різних спостережень (ваги будуть задаватися пропорційно значенням цієї змінної).

Після того як змінні для аналізу обрані і проведені при необхідності додаткові установки, натискається кнопка **OK** у правому верхньому куті і відкривається вікно побудови моделі.

Data: Irisdat.sta (5v by 150c)					
Fisher (1936) iris data: length & width of sepals and petals, 3 types of Iris					
	1	2	3	4	5
	SEPALLEN	SEPALWID	PETALLEN	PETALWID	IRISTYPE
1	5.1	3.3	1.4	0.2	SETOSA
2	6.4	2.8	5.6	2.2	VIRGINIC
3	6.5	2.8	4.6	1.5	VERSCOL
4	6.7	3.1	5.6	2.4	VIRGINIC
5	6.3	2.8	5.1	1.5	VIRGINIC
6	4.6	3.4	1.4	0.3	SETOSA
7	6.9	3.1	5.1	2.3	VIRGINIC
8	6.2	2.2	4.5	1.5	VERSCOL
9	5.9	3.2	4.8	1.8	VERSCOL
10	4.6	3.6	1	0.2	SETOSA
11	6.1	3	4.6	1.4	VERSCOL
12	6	2.7	5.1	1.6	VERSCOL
13	6.5	3	5.2	2	VIRGINIC

Рис. 7.7. Файл даних Iris.sta

Крок 3. Натиснемо кнопку **Variables (Змінні)** на стартовій панелі модуля дискримінантного аналізу (рис. 7.5) і виберемо змінні для аналізу.

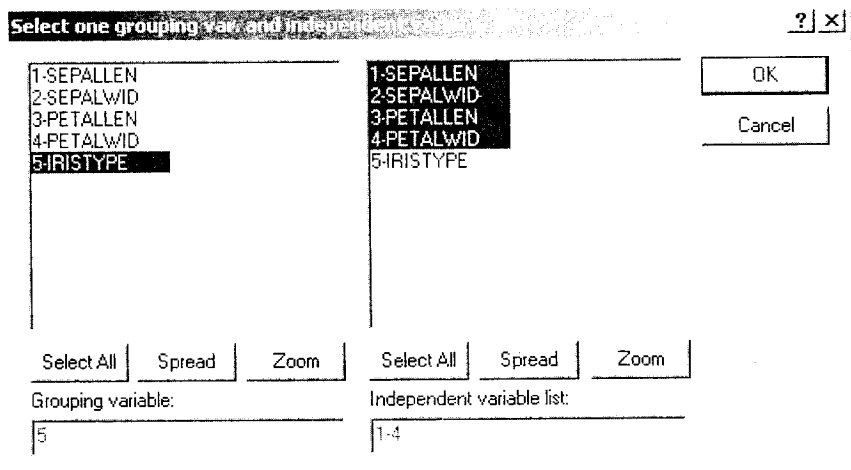


Рис. 7.8. Панель вибору змінних

У якості **Grouping variable** (Групової змінної) виберемо змінну **IRISTYPE** — ТИП ІРИСУ (5).

У якості **Independent variables** (Незалежних змінних) виберемо змінні **SEPALLEN** — ДОВЖИНА ЧАШОЛИСТКА, **SEPALWID** — ШИРИНА ЧАШОЛИСТКА, **PETALLEN** — ДОВЖИНА МАТОЧКИ, **PETALWID** — ШИРИНА МАТОЧКИ.

Крок 4. На стартовій панелі модуля дискримінантного аналізу відзначимо кнопку **Advanced optins (Stervise analysis)** і після натискання **OK** отримаємо меню, у якому виберемо метод дискримінантного аналізу. У нашому випадку виберемо покроковий метод (**Forward sterwise**) (див. рис. 7.9).

На цій панелі можна вибрати такі методи: **Standard** — стандартний; **Forward stepwise** — покрокові включення; **Backward stepwise** — покрокові виключення.

Кнопка **Variables (Змінні)** дозволяє вибрати множину незалежних змінних. В опції **Tolerance** (Толерантність) задається нижня межа толерантності (змінні з толерантністю менше цього значення в модель не включаються).

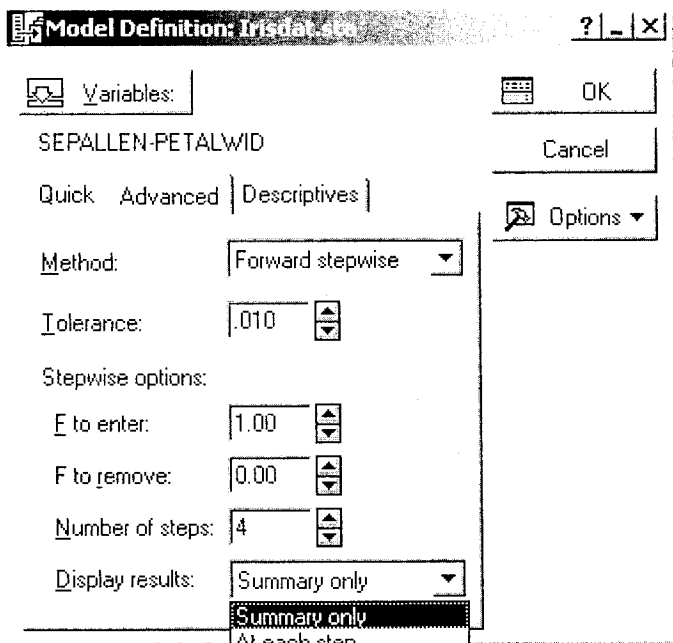


Рис. 7.9. Панель вибору методу дискримінантного аналізу

За визначенням, толерантність дорівнює 1 мінус квадрат множинної кореляції цієї змінної з іншими змінними в моделі (у покрокових методах моделі розглядаються на кожному кроці і кореляція обчислюється за включеними у модель змінними). змінні з малою толерантністю можуть призвести до помилок при обчисленні оберненої матриці. Очевидно, що якщо толерантність мала, то змінна несе малу додаткову інформацію і включення її в модель недоцільно. Задавши в опції **Tolerance** (Толерантність) мале число **epsilon**, ви вилучите з моделі змінні з толерантністю менше **epsilon**.

Для покрокових методів у вікні є спеціальна група опцій: **Stepwise Discriminant Function Analysis** (Покроковий дискримінантний аналіз), де задаються:

☒ **F to enter** — значення статистики F для включення змінної в модель;

☒ **F to remove** — значення статистики F для виключення змінної з моделі;

- ☒ **Number of steps** — число кроків;
- ☒ **Display of results** — відображення результатів на екрані (**Summary only** — тільки підсумкові, **At each step** — на кожному кроці).

Кнопка **Descriptives** відкриває діалогове вікно, у якому можна переглянути описові статистики для групованих даних і різні графіки (гістограми, графіки на нормальному імовірнісному папері тощо).

Це діалогове вікно використовується для перевірки основних припущень дискримінантного аналізу.

Діалогове вікно **Review Descriptive Statistics** (Описові статистики) (рис. 7.10) дозволяє отримати:

- ☒ об'єднані внутрішньогрупові коваріації і кореляції (**Pooled within-groups covariances & correlations**);
- ☒ повні коваріації і кореляції (**Total covariances & correlations**);
- ☒ середні значення для кожної змінної (**Means & number of cases**);
- ☒ стандартні відхилення змінних у кожній групі (**Standart deviations**);

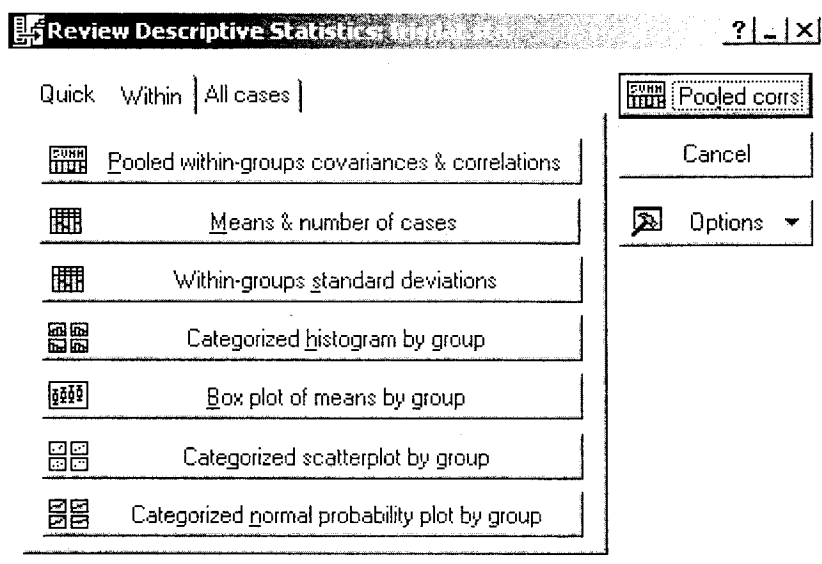


Рис. 7.10. Описові статистики

- ☒ «категоризовані» гістограми по групах для кожної змінної (**Categorized histogram (by group)**);
- ☒ діаграми розмаху по групах — «категоризовану» діаграму розсіювання (по групах) (**Box & whisker plot (by group)**);
- ☒ для двох будь-яких змінних (**Categorized scatterplot (by group)**);
- ☒ «категоризований» нормальний графік для будь-якої змінної по групах (**Categorized normal probability plot (by group)**).

Зробивши необхідні установки, натисніть у діалоговому вікні **Визначення моделі** кнопку **ОК** у правому верхньому куті. Система зробить обчислення, і вікно результатів з'явиться на екрані.

Крок 5. Після натискання кнопок і **ОК** на панелі вибору методу дискримінантного аналізу (рис. 7.10) відкриється діалогове вікно **Analysis Results**.

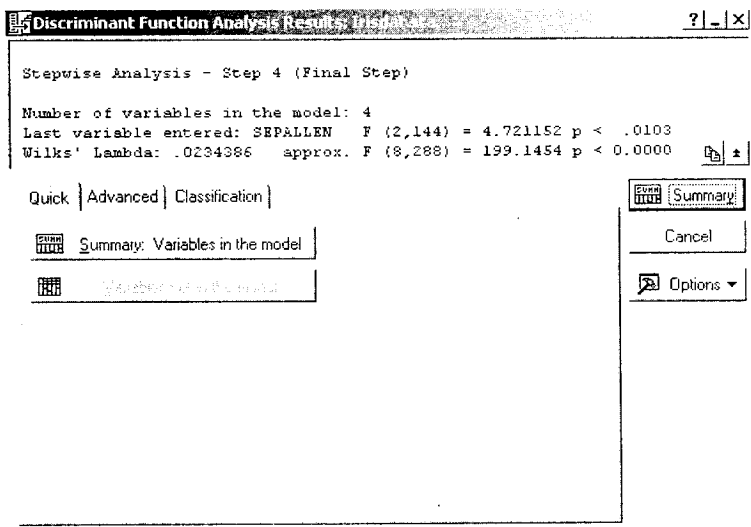


Рис. 7.11. Вікно аналізу результатів

Вікно **Discriminant Function Analysis Results** (Результати дискримінантного функціонального аналізу). Складається з двох частин: верхньої — інформаційної, і нижньої, де містяться функціональні кнопки, що дозволяють всебічно переглянути результати аналізу. В інформаційній частині міститься інформація про метод аналізу змінних, включених у модель, значення стандарт-

них статистик. У вікні наведені число кроків (4) і число змінних (4), що беруть участь у моделі, значення коефіцієнта Уїлкса (Wilks lambda = 0,0234); наближене значення F -статистики, пов'язаної з коефіцієнтом Уїлкса (approx. $F(8,288) = 199,1454$); рівень значущості p для $F = 199,1454$; остання включена змінна (Last variable entered) SEPALLEN; відповідне значення статистики F -критерію $F(2,144) = -4,72$, рівень значущості $p < 0,0103$.

Значення статистики коефіцієнта Уїлкса знаходяться в інтервалі $[0, 1]$, і якщо вони лежать близько 0, то це свідчить про добру дискримінацію. Значення статистики Уїлкса, що лежать близько до 1, свідчать про погану дискримінацію. Іншими словами, якщо значення коефіцієнта Уїлкса близькі до 0, то потужність дискримінації (потужність = 1 — імовірність помилки) близька до 1, якщо ж коефіцієнт Уїлкса близький до 1, то потужність близька до 0.

Якщо у вікні **Визначити модель** (див. рис. 7.9) в опції **Display results** (Відображення результатів на екрані) задати **At each step** (На кожному кроці), то результати будуть відображені на кожному кроці аналізу. При цьому у вікні результатів з'явиться кнопка **Next** (Наступний), натиснувши на яку, ви відкриєте вікно результатів на наступному кроці аналізу (див. рис. 7.12).

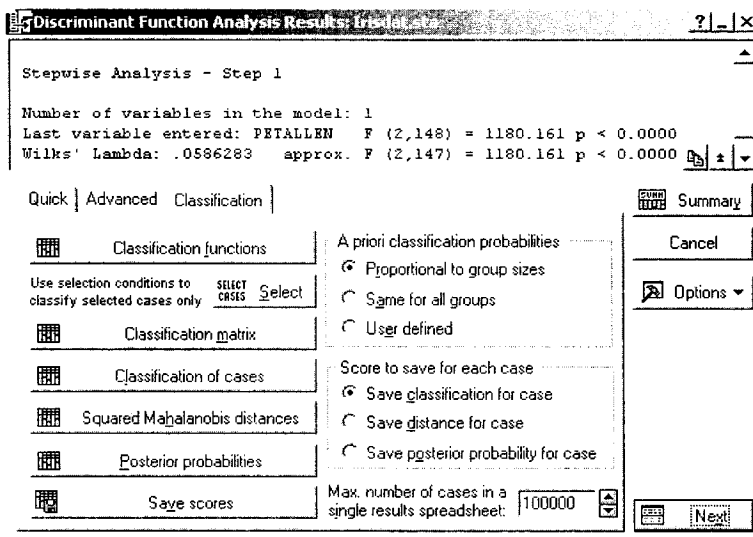


Рис. 7.12. Вікно результатів на кроці 1

При застосуванні покрокових методів (включення або виключення) на кожному кроці аналізу виникають вікна результатів. У цих вікнах доступні всі кнопки, що і в підсумковому вікні результатів. Натиснувши кнопку **Advanced**, задамо такі функціональні кнопки:

- ☒ **Summary: Variables in the model;**
- ☒ **Variables not in the model;**
- ☒ **Distances between groups;**
- ☒ **of stepwise analysis;**
- ☒ **Perform canonical analysis & graphs;**
- ☒ **Sterwise analysis summary.**

Натиснувши кнопку **Classification**, задамо такі функціональні кнопки:

- ☒ **Classification functions;**
- ☒ **Classification matrix;**
- ☒ **Classification of cases** — класифіковані спостереження;
- ☒ **Squared Mahalanobis distances** — квадрати відстаней Махаланобіса;
- ☒ **Posterior probabilities** — апостеріорні ймовірності;
- ☒ **Save scores** — зберегти.

Крок 6. Натиснувши кнопки **Quick** і **Summary** (див. рис. 7.11), запусимо обчислювальну процедуру. Проведемо дослідження результатів дискримінантного аналізу (рис. 7.13).

Variable	Wilks' Lambda	Partial Lambda	F-remove	p-level
SEPALLEN	0.024976	0.938464	4.72115	0.0103
SEPALWID	0.030580	0.766480	21.93593	0.0000
PETALLEN	0.036025	0.669206	35.59018	0.0000
PETALWID	0.031546	0.743001	24.90433	0.0000

Рис. 7.13. Вікно результатів дискримінантного аналізу даних з файла **Iris.sta**

Значення статистики Уїлкса **Wilks lambda** і p знаходяться близько до нуля, що свідчить про досить добру дискримінацію (див. рис. 7.13).

Крок 7. Натиснувши кнопку **Advanced** і ввійшовши в опцію **Perform canonical analysis** можна провести поглиблений аналіз, тобто перевірити передумови дискримінантного аналізу, провести аналіз гістограм для кожного класу тощо. На рис. 7.14 представлений канонічний аналіз.

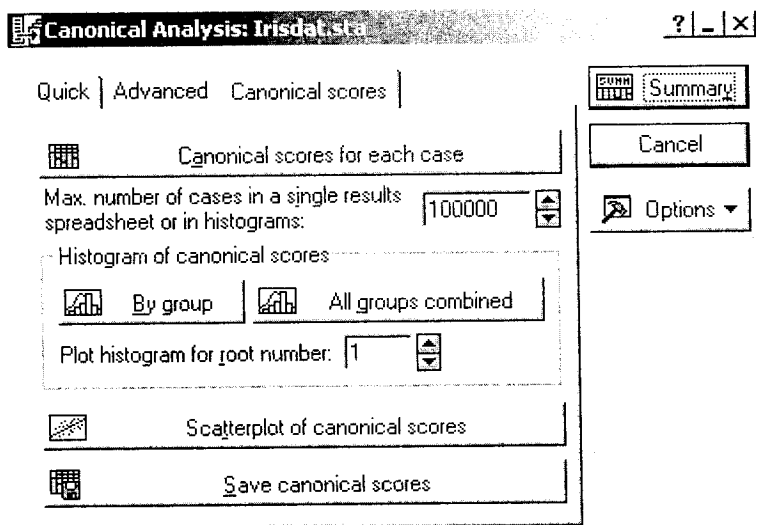


Рис. 7.14. Канонічний аналіз

Тут можна переглянути поділ груп на графіку. Для цього слід натиснути кнопку **Scatterplot of canonical scores** (Діаграма розсіювання канонічних значень). На екрані з'явиться графік поділ трьох класів ірису, наведений на рис. 7.15

Крок 8. На цьому кроці отримаємо дискримінантні функції і використаємо їх для віднесення поточного вимірювання квітки до того чи іншого класу. Корисною функцією видається оцінка статистичного зв'язку між параметрами квітки. Для проведення статистичного аналізу ввійдемо в опцію **Model Definition**, далі натиснувши послідовно кнопки **Descriptives** — **Review descriptive statistics** — **Total covariances** отримаємо кореляційну матрицю (див. рис. 7.16). Аналіз кореляційної матриці показує,

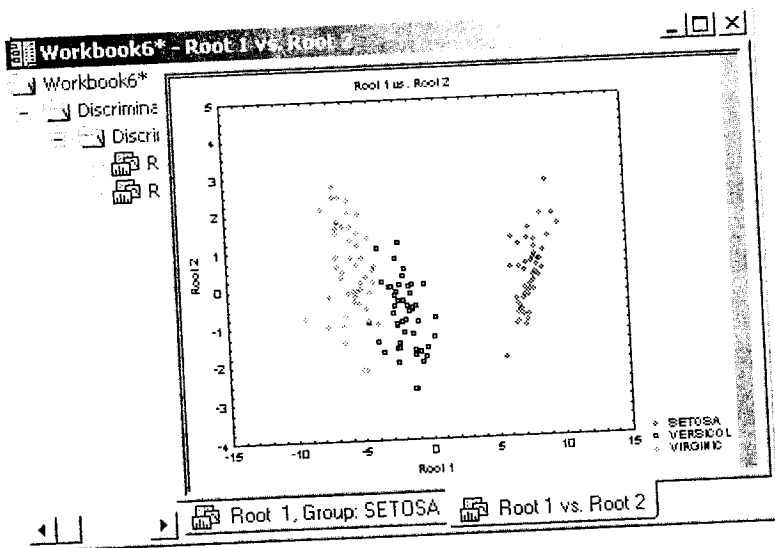


Рис. 7.15. Поділ трьох типів ірису

що між довжиною чашолистка і довжиною маточки, довжиною чашолистка і шириною маточки, довжиною маточки і шириною

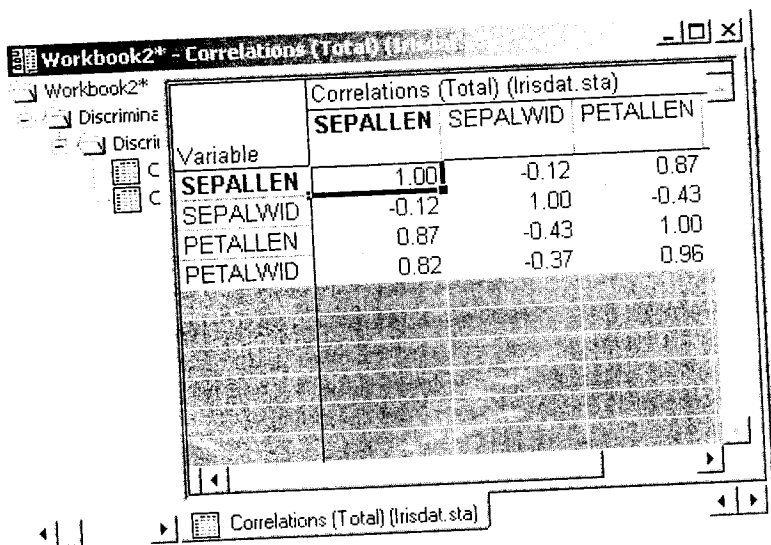


Рис. 7.16. Матриця коефіцієнтів кореляції

маточки мається досить високий статистичний зв'язок. Коефіцієнти кореляції відповідно дорівнюють 0,87, 0,82, 0,96. Має місце лінійний статистичний зв'язок між незалежними змінними. При використанні регресійного аналізу для побудови дискримінантних функцій необхідно в цьому випадку перевірити передумови і припущення методу й оцінити ступінь мультиколінеарності. Кореляційна матриця являє собою таблицю, що містить коефіцієнти кореляції для кожної пари змінних з багатовимірної сукупності даних. Чим вище коефіцієнт кореляції між двома X -змінними, тим більше мультиколінеарність. Це пояснюється тим, що висока кореляція (близька до 1 чи -1) вказує на сильний зв'язок і свідчить про те, що ці дві X -змінні мають лінійну залежність.

Наслідки мультиколінеарності можуть бути пов'язані з труднощами проведення статистичних тестів для окремих коефіцієнтів регресії внаслідок збільшення стандартних помилок. У результаті значуща змінна може бути оцінена як незначуща. Виникають обчислювальні труднощі, пов'язані з «несталістю обчислень». У цьому випадку програма може намагатися виконати ділення на нуль і видати безглузді і невірні результати.

Невелика чи середня мультиколінеарність зазвичай не представляє проблеми. Дуже сильна мультиколінеарність завжди буде представляти проблему і може призводити до серйозних помилок (обчислювальні наслідки).

Основний статистичний результат мультиколінеарності полягає у рості стандартних помилок деяких чи всіх коефіцієнтів регресії (S_{bi}). Натиснувши кнопку **Classification** і ввійшовши в опцію **Classification functions (Функції класифікації)** (рис. 7.17), побудуємо дискримінантні функції класифікації.

У якості **Незалежних змінних — Independent variables** — виберемо змінні: **SEPALLEN** — X_1 (ДОВЖИНА ЧАШОЛИСТКА), **SEPALWD** — X_2 (ШИРИНА ЧАШОЛИСТКА), **PETALLEN** — X_3 (ДОВЖИНА МАТОЧКИ), **PETALWD** — X_4 (ШИРИНА МАТОЧКИ).

У якості залежних змінних приймаємо: **SETOSA-Y1**, **VER-SICOL-Y2**, **VIRGINIC-Y3**. Нову квітку зі значеннями: X_1 , X_2 , X_3 , X_4 будемо відносити до того чи іншого типу ірису за трьома значеннями: за максимальним значенням дискримінантних функцій, мінімальним значенням квадрата відстані Махалано-біса і максимальною апостеріорною імовірністю.

Variable	SETOSA p=.33333	VERSICOL p=.33333	VIRGINIC p=.33333
PETALLEN	-16.4306	5.2115	12.767
SEPALWID	23.5879	7.0725	3.685
PETALWID	-17.3984	6.4342	21.079
SEPALLEN	23.5442	15.6982	12.446
Constant	-86.3085	-72.8526	-104.368

Classification Functions; grouping: IRISTYPE (Irisdat.sta)

Рис. 7.17. Функції класифікації, побудовані покроковим методом

Побудуємо дискримінантні функції. За допомогою цих функцій можна обчислити класифікаційні значення (мітки) для квіток, що знову спостерігаються, за формулами:

$$Y_1 = -86,31 - 16,43X_1 + 23,69X_2 - 17,4X_3 + 23,54X_4;$$

$$Y_2 = -72,85 + 5,21X_1 + 7,07X_2 - 6,43X_3 + 15,70X_4;$$

$$Y_3 = -104,37 + 12,76X_1 + 3,69X_2 - 21,08X_3 + 12,5X_4.$$

Віднесемо новий об'єкт до певного класу за допомогою відстані Махаланобіса.

Після натискання кнопки **Squared Mahalanobis distance** (Квадрат відстані Махаланобіса) отримаємо таблицю з квадратами відстані Махаланобіса від точок (спостережень) до центрів груп. Випадок відноситься до групи, до якої відстань Махаланобіса мінімальна (див. рис. 7.18).

На наступному етапі віднесемо кожну квітку до того чи іншого типу за допомогою апостеріорних імовірностей. До аналізу необхідно задати для кожного спостереження (у цьому прикладі квітки) імовірність, з якою вона належить до визначеного класу. Після того як аналіз виконаний, можна перерахувати ці імовірності й отримати апостеріорні імовірності класифікації.

Workbook1* - Squared Mahalanobis Distances from Group Centroids (Iris)

Squared Mahalanobis Distances from Group Centroids (Iris)

Incorrect classifications are marked with *

Case	Observed Classif.	SETOSA p=.33333	VERSICOL p=.33333	VIRGIN p=.33333
1	SETOSA	0.2419	90.6602	181.55
2	VIRGINIC	208.5713	27.3188	1.89
3	VERSICOL	105.2663	2.2329	13.07
4	VIRGINIC	207.9180	31.7492	4.45
* 5	VIRGINIC	133.0668	5.2529	7.23
6	SETOSA	1.3337	84.0118	170.05
7	VIRGINIC	173.1838	26.5620	11.04
8	VERSICOL	131.6617	8.4307	14.75
* 9	VERSICOL	130.8624	8.6697	6.50

Squared Mahalanobis Distances from Group Centroids (Iris)

Рис. 7.18. Відстані Махаланобіса для даних з файла Iris.sta

Натиснувши кнопку **Classification** і ввійшовши в опцію **Posterior probabilities** (Апостеріорні ймовірності), отримаємо таблицю з апостеріорними ймовірностями приналежності об'єкта до визначеного класу.

Workbook1* - Posterior Probabilities (Irisdat.sta)

Posterior Probabilities (Irisdat.sta)

Incorrect classifications are marked with *

Case	Observed Classif.	SETOSA p=.33333	VERSICOL p=.33333	VIRGIN p=.33333
1	SETOSA	1.000000	0.000000	0.000000
2	VIRGINIC	0.000000	0.000003	0.999997
3	VERSICOL	0.000000	0.995590	0.004410
4	VIRGINIC	0.000000	0.000001	0.999999
* 5	VIRGINIC	0.000000	0.729388	0.270612
6	SETOSA	1.000000	0.000000	0.000000
7	VIRGINIC	0.000000	0.000428	0.999572
8	VERSICOL	0.000000	0.969573	0.030427
* 9	VERSICOL	0.000000	0.253228	0.746772

Рис. 7.19. Таблиця апостеріорних ймовірностей

Інтерпретація цієї таблиці дуже проста. У першому стовпчику зазначений тип ірису для кожного спостереження. У другому, третьому, четвертому стовпчиках дані апостеріорні імовірності віднесення кожної квітки до певного типу. Квітка відноситься до групи з максимальною апостеріорною імовірністю. Знаком * позначаються неправильно класифіковані при використанні цього правила спостереження.

Крок 9. Класифікація нових спостережень. Не закриваючи діалогу **Результати дискримінантного аналізу**, додайте в таблицю вихідних даних новий випадок, наприклад, так:

151	5.3	3.1	2.7	0.5

Рис. 7.20. Нове спостереження в даних Iris.sla

Для того щоб зрозуміти, до якого класу відноситься цей об'єкт, натисніть кнопку **Posterior probabilities (Апостеріорні ймовірності)**. Після цього ви побачите ту ж таблицю з апостеріорними ймовірностями, до якої буде доданий рядок:

151	---	0.999	0.001	0.000

Рис. 7.21. Класифікація нового спостереження

Отже, нове спостереження з імовірністю 0,999 можна віднести до типу **SETOSA**.

Натисніть кнопку **Squared Mahalanobis distance (Квадрат відстані Махаланобіса)**, і ви побачите таблицю з квадратами відстані Махаланобіса. В останньому рядку таблиці ми бачимо відстані нового спостереження до групових центрів (див. рис. 7.22)

Знову відстань від нового спостереження до центра груп мінімальна для групи **SETOSA**. Отже, з високим ступенем імовірності нова квітка — це ірис типу **SETOSA**.

Крок 10. Побудова класифікаційної функції іншим методом. Завжди має сенс спробувати різні способи побудови класифікаційної функції.

151	---	16.2	31.2	97.2

Рис. 7.22. Відстань Махаланобіса від нового спостереження до центрів груп

Закрийте діалогове вікно результатів дискримінантного аналізу, натиснувши кнопку **CANCEL**. Ви повернетеся у вікно **Визначення моделі**. У рядку **Method (Метод)** виберіть стандартний метод і натисніть кнопку **OK**.

У вікні результатів за допомогою кнопки **Classification functions (Функції класифікації)** переглянете класифікуючі функції:

Squared Mahalanobis Distances from Group C				
Incorrect classifications are marked with *				
Case	Observed Classif.	SETOSA p=.33333	VERSICOL p=.33333	VIRGINIC p=.33333
	SETOSA	0.2419	90.6602	181.5587
	VIRGINIC	208.5713	27.3188	1.8944
	VERSICOL	105.2663	2.2329	13.0720
	VIRGINIC	207.9180	31.7492	4.4506
*	VIRGINIC	133.0668	5.2529	7.2359
	SETOSA	1.3337	84.0118	170.0569
	VIRGINIC	173.1838	26.5620	11.0484
	VERSICOL	131.6617	8.4307	14.7647
*	VERSICOL	130.8624	8.6697	6.5068

 Squared Mahalanobis Distances from Group Centroids (Irisdat.sta)

Рис. 7.23. Функції класифікації, побудовані за стандартним методом

Видно, що функції класифікації, побудовані двома методами, збігаються.

Крок 11. Використання кнопок **Select Cases (Вибрати спостереження)**. Ці кнопки знаходяться в двох вікнах модуля **Дискримінантний аналіз**: у стартовій панелі й у вікні результатів дискримінантного аналізу. Використовуючи цю кнопку в стартовій панелі, вибираються спостереження, за якими будуть побудовані дискримінантні функції і функції класифікації. У вікні результатів за допомогою цієї кнопки можна подивитися, як, використовуючи побудовані функції, класифікуються нові спостереження, тобто спостереження, які не використовувалися для побудови дискримінантних функцій і функцій класифікації. Використання цих кнопок важливий момент роботи в модулі.

Приклад використання цих кнопок дається нижче. Робочим файлом як і раніше являється файл **Irisdat.sta**, і змінні вибрані як у прикладі, розібраному в попередньому розділі.

У стартовій панелі модуля натисніть кнопку **Select Cases (Вибрати спостереження)**. Вікно умови вибору спостережень з'явиться на екрані:

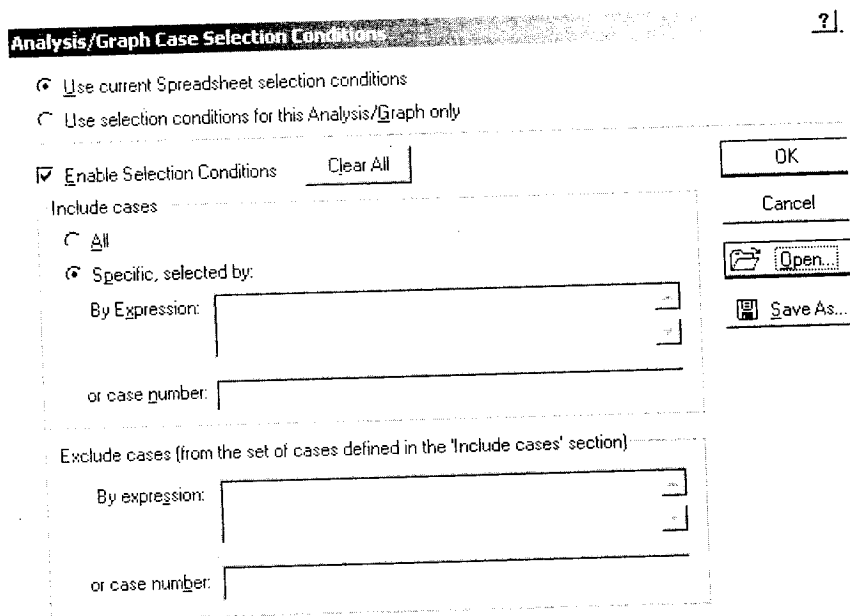


Рис. 7.24. Вікно умови вибору спостережень

Опції **Include cases (Включити спостереження)**, **Exclude cases (Виключити спостереження)** дозволяють, відповідно, включити чи виключити з аналізу спостереження, задані у вікні. Зробіть установки в цьому вікні, як показано на рисунку в прикладі. Натисніть кнопку **OK** і поверніться на стартову панель.

Висновки

Дискримінантний аналіз так само, як і кластерний аналіз, відноситься до методів багатовимірної класифікації, але при цьому базується на інших передумовах. Основна відмінність

полягає в тому, що в ході дискримінантного аналізу нові класи не утворюються, а формулюється правило, за яким нові одиниці сукупності відносяться до однієї з вже існуючих множин (класів). Підставою для віднесення кожної одиниці сукупності до визначеної множини служить величина дискримінантної функції, розрахована за відповідним значенням дискримінантних змінних.

Основними проблемами дискримінантного аналізу являються, по-перше, визначення набору дискримінантних змінних, по-друге, вибір виду дискримінантної функції. Існують різні критерії послідовного відбору змінних, що дозволяють отримати найкраще розрізнення множин. Можна також скористатися алгоритмом покрокового дискримінантного аналізу, який у літературі докладно описаний. Після уточнення оптимального набору дискримінантних змінних досліднику треба буде вирішити питання про вибір типу дискримінантної функції, тобто вибрати вид поділяючої поверхні. Найчастіше на практиці застосовують лінійний дискримінантний аналіз. У цьому випадку дискримінантна функція являє собою або пряму, або площину (гіперплощину).

Лінійна дискримінантна функція не завжди підходить у якості опису поділяючої поверхні між множинами. Наприклад, у тих випадках, коли множини, що розрізняються, не являються опуклими, правомірно припустити, що дискримінантна функція, що призводить до найменших помилок класифікації, не може бути лінійною.

Якщо множини, використовувані як навчальні вибірки, близько розташовані одна до одної, то зростає ймовірність помилкової класифікації нових об'єктів, особливо в тих випадках, коли об'єкт, що класифікується, сильно віддалений від центрів обох множин. Складається ситуація, за якої розпізнавання об'єкта утруднене. Одним з можливих виходів у такому випадку являється перегляд набору дискримінантних змінних.

У якості дискримінантних змінних можуть виступати не тільки вихідні (ті, що спостерігаються) ознаки, але й головні компоненти чи головні фактори.

У роботах [1, 2, 3, 7, 8, 9] показано, що дискримінантний аналіз являється досить стійким методом, що допускає деякі відхилення від цих припущень. Крім того, не всі припущення дискримінантного аналізу вимагають виконання на сто відсотків.

Знаючи статистичні характеристики і вимоги, що висуваються моделлю, можна визначити, коли вони були порушені, коли слід застосувати коригувальні міри і коли методика не відповідає цілям цього дослідження.

Контрольні запитання і вправи

1. Поясніть сутність дискримінантного аналізу.
2. Сформулюйте припущення дискримінантного аналізу
3. Як визначається кількість дискримінантних функцій?
4. У чому полягає сутність байєсівської процедури класифікації?
5. Як визначаються апостеріорні імовірності?
6. Поясніть принцип класифікації при невідомих параметрах.
7. Які критерії відбору змінних використовуються в покроковій процедурі дискримінації?
8. Поясніть принцип покрокового методу дискримінації.
9. Яким чином можна перевірити припущення дискримінантного аналізу?
10. У чому полягає необхідність використання коефіцієнта толерантності.

Завдання 1

У табл. 7.2 наведено групи машинобудівних підприємств з високим та низьким рівнями організації управління виробництвом [2]. Характеризуючи діяльність підприємств показниками рентабельності і продуктивності праці за допомогою дискримінантного аналізу, потрібно класифікувати три останніх підприємства.

Таблиця 7.2

№ підприємства	Показник Група підприємств	Рентабельність, %	Продуктивність праці, млн грн/люд.
1	Високий рівень, X	23,4	9,1
2		19,1	6,6
3		17,5	5,2
4		17,2	10,1
5	Низький рівень, Y	5,4	4,3
6		6,6	5,5
7		8,0	5,7
8		9,7	5,5
9		9,1	6,6
10	Підлягають дискримінації, Z	9,9	7,4
11		14,2	9,4
12		12,9	6,7

Завдання 2

При оцінці ефективності діяльності підприємств приладобудування були отримані два класи підприємств А і В (табл. 7.3). Використовуючи їх як навчальні вибірки провести дискримінацію трьох останніх підприємств за показниками фондовіддачі і матеріалоемності.

Таблиця 7.3

№ підприємства	Показник	Фондовіддача	Матеріалоемність
	Підприємства		
1	Клас А, Х	1,47	3,90
2		1,52	6,18
3		1,38	13,80
4		1,27	8,18
5	Клас У, У	0,59	9,08
6		0,57	2,10
7		0,46	6,5
8		0,53	13,3
9	Підлягають дискримінації, Z	0,92	13,80
10		1,21	4,80
11		0,63	12,33

Завдання 3

У ППП STATISTICA за допомогою кнопки **Variables** виберіть три змінні у файлі даних **iridat.sta** — ДЛЛЕПЕСТ, ШИРЧАШЕЛ, ШИРЛЕПЕС — і виконайте такі дії:

1. Запустіть обчислювальну процедуру, що реалізує покроковий метод включення, і проведіть інтерпретацію результатів.

2. Визначте, чи зміниться результат у порівнянні з результатом, отриманим у розділі «Використання ППП «STATISTICA» для розв'язання задач класифікації».

3. Поекспериментуйте з різними методами і різними наборами незалежних змінних.

ЛИТЕРАТУРА

1. Айвазян С. А., Мхитарян В. С. Прикладная статистика и основы эконометрики: Учебник.— М.: ЮНИТИ, 1998.
2. Сигел Э. Практическая бизнес-статистика.— М.: Изд. дом «Вильямс», 2002.— 1056 с.
3. Дубров А. М., Мхитарян В. С., Трошин Л. И. Многомерные статистические методы.— М.: Финансы и статистика, 2000.— 352 с.
4. Химмельблау Д. Анализ процессов статистическими методами.— М.: Мир, 1973.— 948 с.
5. Общая теория статистики: статистическая методология в изучении коммерческой деятельности / Под ред. А. А. Спирина, О. Э. Башиной.— М.: Финансы и статистика. 1995.— 296 с.
6. Рябушкин Т. В. и др. Статистические методы и анализ социально-экономических процессов.— М.: Наука, 1990.— 296 с.
7. Клейнен Дж. Статистические методы в имитационном моделировании.— М.: Статистика, 1978.— 221 с.
8. Афифи А., Эйзен С. Статистический анализ: Подход с использованием ЭВМ / Пер. с англ.— М.: Мир, 1982.— 488 с.
9. Тюрин Ю. Н., Макаров А. А. Статистический анализ данных на компьютере.— М.: Инфра-М, 1998.— 528 с.
10. Бендат Дж., Пирсол А. Измерение и анализ случайных процессов.— М.: Мир, 1974.— 448 с.
11. Рао Р. Линейные статистические методы и их приложения.— М.: Мир, 1967.— 548 с.
12. Демиденко Е. З. Линейная и нелинейная регрессия.— М.: Наука, 1987.— 248 с.

13. Алберт А. Регрессия, псевдорегрессия и рекуррентное оценивание.— М.: Наука, 1977.— 224 с.

14. Себер Дж. Линейный регрессионный анализ.— М.: Мир, 1980.— 456 с.

15. Вучков И., Бояджиева Л., Солаков Е. Прикладной линейный регрессионный анализ: Пер. с болгар.— М.: Финансы и статистика, 1987.— 230 с.

16. Канторович Г. Г. Анализ временных рядов // Экономический журнал Высшей школы экономики.— 2002.— Т. 6.— №1.— С. 85–116.— №2.— С. 251–273.— №3.— С. 379–401.

17. Zadeh L. A. Fuzzu Setz, information and Control, 1965, vol. 8, p. 338–353.

18. Факторный, дискриминантный и кластерный анализ.— М.: Финансы и статистика, 1989.

19. Knowledge Discovery Through Data Mining: What Is Knowledge Discovery.— Tandem Computers Inc., 1996.

20. Дюк В., Самойленко А. Data Mining: Учебный курс.— СПб: Питер, 2001.— 368 с.

21. Jiamei Han, Micheline Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers.— San Francisco, 2003.— 550 p.

22. Горбань А. Н., Дунин-Барковский В. Л. и др. Нейроинформатика.— Новосибирск: Наука, 1998.

23. Осовский С. Нейронные сети для обработки информации: Пер. с польского И. Д. Рудинского.— М.: Финансы и статистика, 2004.— 344 с.

24. Бэстэнс Д. Э., Ван ден Берг В. М., Вуд Д. Нейронные сети и финансовые рынки: принятие решений в торговых операциях.— М.: ТВП, 1997.— 236 с.

25. Рутковская Д., Пилиньский М., Рутковский Л. Нейронные сети, генетические алгоритмы и нечеткие системы: Пер. с польск. И. Д. Рудинского.— М.: Горячая линия-Телеком, 2004.— 452 с.

26. Grossberg S. 1974. Classical and instrumental learning by neural networks // Progress in theoretical biology.— New York: Academic Press.— Vol. 3, pp. 51–141.

27. Kohonen T. 1984. Self-organization and associative memory // Series in Information Sciences, vol. 8. Berlin: Springer verlag.

28. Лукашин Ю. П. Адаптивные методы краткосрочного прогнозирования временных рядов.— М.: Финансы и статистика, 2003.— 416 с.

29. *Боровиков В. П.* Прогнозирование в системе STATISTICA в среде Windows. Основы теории и интенсивная практика на компьютере.— М.: Финансы и статистика, 2000.— 384 с.
30. *Боровиков В. П.* STATISTICA: искусство анализа данных на компьютере: Для профессионалов.— СПб.: Питер, 2001.— 656 с.
31. *Дрейнер Н., Смит Г.* Прикладной регрессионный анализ: Пер. с англ.— М.: Финансы и статистика, 1986.— 260 с.
32. *Тихонов А. Н., Арсенин В. Я.* Методы решения некорректных задач.— М.: Наука, 1974.— 223 с.
33. *Hoerl A. E., Kennard R. W.* Ridge regression: applications to nonorthogonal problems // *Technometrics*.— 1970.— Vol. 12.— № 1.— P. 69–82.
34. *Obenchain R. L.* Influential observations in ridge regression // *The Annals of Statistics*.— 1997.— Vol. 3.— № 2.— P. 123–131.
35. *Шефе Г.* Дисперсионный анализ.— М.: Наука, 1980.— 512 с.
36. *Ивахненко А. Г., Степашко В. С.* Помехоустойчивость в моделировании.— К.: Наукова думка, 1985.— 216 с.
37. *Романенко В. Д., Игнетенко Б. В.* Адаптивное управление технологическими процессами на базе микро-ЭВМ.— К.: Вища школа, 1990.— 334 с.
38. Проектирование систем автоматизации технологических процессов: Справ. пособие / Под ред. А. С. Ключева.— М.: Энергоатомиздат, 1990.— 464 с.
39. Наладка средств автоматизации и автоматических систем регулирования: Справ. пособие / Под ред. А. С. Ключева.— М.: Атомэнергоиздат, 1989.— 368 с.
40. *Боровиков В. П.* Программа STATISTICA для студентов и инженеров.— М.: Компьютер Пресс, 2001.— 301 с.
41. *STATGRAPHICS PLUS for Windows. User Manuals*, Manugistics Inc., USA, 1995.
42. Классификация и кластер / Под ред. Дж. Вэн Райзин.— М.: Мир, 1980.
43. *Ester M., Kriegel H.-P., Xu X.* Knowledge discovery in large spatial databases: Focusing techniques for efficient class identification, SSD'95.
44. *Knorr E., Ng R.* Algorithms for mining distance-based outliers in large datasets, VLDB'98.
45. *Wang W., Yang X., Muntz R.* STING: A Statistical Information grid Approach to Spatial Data Mining, VLDB'97.

ДОДАТОК

МАТЕМАТИКО-СТАТИСТИЧНІ ТАБЛИЦІ

Таблиця 1

Нормальний закон розподілу. Значення функції $\Phi(t) = P(|T| \leq t_{\text{табл}})$

Цілі та десяткові частки t	Соті частки t									
	0	1	2	3	4	5	6	7	8	9
0,0	0	0,008	0,016	0,0239	0,0319	0,0399	0,0478	0,0558	0,0638	0,0717
0,1	797	876	955	1034	1113	1192	1271	1350	1428	1507
0,2	1585	1663	1741	1819	1897	1974	2051	2128	2205	2282
0,3	2358	2434	2510	2586	2661	2737	2812	2886	2960	3035
0,4	3108	3182	3255	3328	3401	3473	3545	3616	3688	3759
0,5	3829	3899	3969	4039	4108	4177	4245	4313	4381	4448
0,6	4515	4581	4647	4713	4778	4843	4907	4971	5035	5098
0,7	5161	5223	5285	5346	5407	5467	5527	5587	5646	5705
0,8	5763	5821	5878	5935	5991	6047	6102	6157	6211	6265
1,0	0,6827	0,6875	0,6923	0,697	0,7017	0,7063	0,7109	0,7154	0,7199	0,7243
1,1	7287	7330	7373	7415	7457	7499	7540	7580	7620	7660
1,2	7699	7737	7775	7813	7850	7887	7923	7959	7994	8029
1,3	8064	8098	8132	8165	8198	8230	8262	8293	8324	8355
1,4	8385	8415	8444	8473	8501	8529	8557	8584	8611	8638

Продовження табл. 1

Цілі та десят- кові частки t	Соті частки t									
	0	1	2	3	4	5	6	7	8	9
0,9	6319	6372	6424	6476	6528	6579	6629	6679	6729	6778
1,5	8664	8690	8715	8740	8764	8789	8812	8836	8859	8882
1,6	8904	8926	8948	8969	8990	9011	9031	9051	9070	9090
1,7	9109	9127	9146	9164	9181	9199	9216	9233	9249	9265
1,8	9281	9297	9312	9327	9342	9357	9371	9385	9399	9412
1,9	9426	9439	9451	9464	9476	9488	9500	9512	9523	9534
2,0	0,9545	0,9556	0,9566	0,9576	0,9586	0,9596	0,9606	0,9616	0,9625	0,9634
2,1	9643	9651	9660	9668	9676	9684	9692	9700	9707	9715
2,2	9722	9729	9736	9743	9749	9756	9762	9768	9774	9780
2,3	9786	9791	9797	9802	9807	9812	9817	9822	9827	9832
2,4	9836	9841	9845	9849	9853	9857	9861	9865	9869	9872
2,5	9876	9879	9883	9886	9889	9892	9895	9898	9901	9904
2,6	9907	9910	9912	9915	9917	9920	9922	9924	9926	9928
2,7	9931	9933	9935	9937	9939	9940	9942	9944	9946	9947

Закінчення табл. 1

Цілі та десяти- кові частки t	Соті частки t									
	0	1	2	3	4	5	6	7	8	9
2,8	9949	9951	9952	9953	9955	9956	9958	9959	9960	9961
2,9	9963	9964	9965	9966	9967	9968	9969	9970	9971	9972
3,0	0,9973	0,9974	0,9975	0,9976	0,9976	0,9977	0,9978	0,9979	0,9979	0,998
3,1	9981	9981	9982	9983	9983	9984	9984	9985	9985	9986
3,5	9995	9996	9996	9996	9996	9996	9996	9996	9997	9997
3,6	9997	9997	9997	9997	9997	9997	9997	9998	9998	9998
3,7	9998	9998	9998	9998	9998	9998	9998	9998	9998	9998
3,8	9999	9999	9999	9999	9999	9999	9999	9999	9999	9999
3,9	9999	9999	9999	9999	9999	9999	9999	9999	9999	9999
4,0	0,999936	9999	9999	9999	9999	9999	9999	9999	9999	9999
4,5	0,999994	---	---	---	---	---	---	---	---	---
5,0	0,99999994	---	---	---	---	---	---	---	---	---

Таблиця 2

Розподіл Стюдента (t -розподіл)

ν	Імовірність $\alpha = St(t) = P(T > t_{\text{табл}})$										
	0,9	0,8	0,7	0,6	0,5	0,4	0,3	0,2	0,1	0,05	0,01
1	0,158	0,325	0,510	0,727	1,000	1,376	1,963	3,078	6,314	12,706	63,657
2	0,142	0,289	0,445	0,617	0,816	1,061	1,386	1,886	2,290	4,303	9,925
											31,598

v	Імовірність $a = St(t) = P(T > t_{\text{таб}})$												
	0,9	0,8	0,7	0,6	0,5	0,4	0,3	0,2	0,1	0,05	0,02	0,01	0,001
3	0,137	0,277	0,424	0,584	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	12,941
4	0,134	0,271	0,414	0,569	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,132	0,267	0,408	0,559	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,043	6,859
6	0,131	0,265	0,404	0,553	0,718	0,906	1,134	1,440	1,953	2,447	3,143	3,707	5,959
7	0,130	0,263	0,402	0,549	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499	5,405
8	0,130	0,262	0,399	0,546	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,129	0,261	0,398	0,543	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,129	0,260	0,327	0,542	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	4,583
11	0,129	0,260	0,396	0,540	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,128	0,259	0,395	0,539	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,128	0,259	0,394	0,538	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	0,128	0,258	0,393	0,537	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977	4,140
15	0,128	0,258	0,393	0,536	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947	4,073
16	0,128	0,258	0,392	0,535	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921	4,015
17	0,128	0,257	0,392	0,534	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898	3,965
18	0,127	0,257	0,392	0,534	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878	3,922
19	0,127	0,257	0,391	0,533	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861	3,833

Імовірність $a = St(t) = P(T > t_{\text{табл}})$													
v	0,9	0,8	0,7	0,6	0,5	0,4	0,3	0,2	0,1	0,05	0,02	0,01	0,001
20	0,127	0,257	0,391	0,533	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845	3,850
21	0,127	0,257	0,391	0,532	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831	3,819
22	0,127	0,256	0,390	0,532	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819	3,792
23	0,127	0,256	0,390	0,532	0,685	0,868	1,060	1,319	1,714	2,069	2,500	2,807	3,767
24	0,127	0,256	0,390	0,531	0,685	0,857	1,059	1,318	1,711	2,064	2,402	2,797	3,745
25	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787	3,725
26	0,127	0,256	0,390	0,531	0,684	0,856	0,058	0,315	1,706	2,056	2,479	2,779	3,707
27	0,127	0,256	0,389	0,531	0,684	0,855	0,057	0,314	1,703	2,052	2,473	2,771	3,690
28	0,127	0,256	0,389	0,530	0,683	0,855	0,056	0,313	1,701	2,048	2,467	2,763	3,674
29	0,127	0,256	0,389	0,530	0,683	0,854	0,055	0,311	1,699	2,045	2,462	2,756	3,659
30	0,127	0,256	0,389	0,530	0,683	0,854	0,055	0,310	1,697	2,042	2,457	2,750	3,646
40	0,126	0,255	0,388	0,529	0,681	0,851	0,050	0,303	1,684	2,021	2,423	2,704	3,551
60	0,126	0,254	0,387	0,527	0,679	0,848	0,046	0,296	1,671	2,000	2,390	2,660	3,460
120	0,126	0,254	0,386	0,526	0,677	0,845	0,041	0,289	1,658	1,980	2,358	2,617	3,373
∞	0,126	0,253	0,385	0,524	0,674	0,842	0,036	0,282	1,645	1,960	2,326	2,576	3,291

Таблиця 3
Розподіл Пірсона (χ^2 -розподіл). Значення $\chi^2_{\text{табл}}$ для імовірностей $P(\chi^2 > \chi^2_{\text{табл}})$

v	Імовірність											
	0,999	0,995	0,99	0,98	0,98	0,98	0,95	0,90	0,80	0,75	0,70	0,50
1	0,05157	0,04393	0,03157	0,03628	0,03982	0,03982	0,00393	0,0158	0,0642	0,102	0,148	0,455
2	0,002	0,01	0,0201	0,0404	0,0506	0,0506	0,103	0,211	0,446	0,575	0,713	1,386
3	0,0243	0,0717	0,115	0,185	0,216	0,216	0,352	0,584	1,005	1,213	1,424	2,366
4	0,0908	0,207	0,297	0,429	0,484	0,484	0,711	1,064	1,649	1,923	2,195	3,357
5	0,210	0,412	0,554	0,752	0,831	0,831	1,145	1,610	2,343	2,675	3,000	4,351
6	0,381	0,676	0,872	1,134	1,237	1,237	1,635	2,204	3,070	3,455	3,828	5,348
7	0,598	0,989	1,239	1,564	1,690	1,690	2,167	2,833	3,822	4,255	4,671	6,346
8	0,857	1,344	1,646	2,032	2,180	2,180	2,733	3,490	4,594	5,071	5,527	7,344
9	1,152	1,735	2,088	2,532	2,700	2,700	3,325	4,168	5,380	5,899	6,393	8,343
10	1,479	2,156	2,558	3,059	3,247	3,247	3,240	4,865	6,179	6,737	7,267	9,342
11	1,834	2,603	3,053	3,609	3,816	3,816	4,575	5,578	6,989	7,584	8,148	10,341
12	2,214	3,074	3,571	4,178	4,404	4,404	5,226	6,304	7,807	8,438	9,034	11,340
13	2,617	3,565	4,107	4,765	5,009	5,009	5,892	7,042	8,634	9,299	9,926	12,340
14	3,041	4,075	4,660	5,368	5,629	5,629	6,571	7,790	9,467	10,165	10,821	13,339
15	3,483	4,601	5,229	5,985	6,262	6,262	7,261	8,547	10,307	11,036	11,721	14,339
16	3,942	5,142	5,812	6,614	6,908	6,908	7,962	9,312	11,152	11,912	12,624	15,338

Продовження табл. 3

v	Імовірність												
	0,999	0,995	0,99	0,98	0,98	0,95	0,90	0,80	0,75	0,70	0,50		
17	4,416	5,697	6,408	7,255	7,564	8,672	10,085	12,002	12,892	13,531	16,338		
18	4,905	6,265	7,015	7,906	8,231	9,390	10,865	12,857	13,675	14,440	17,338		
19	5,407	6,844	7,633	8,567	8,907	10,117	11,651	13,716	14,562	15,352	18,338		
20	5,921	7,434	8,260	9,237	9,591	10,871	12,443	14,578	15,452	16,266	19,337		
21	6,447	8,034	8,897	9,915	10,283	11,591	13,240	15,445	16,344	17,182	20,337		
22	6,983	8,643	9,542	10,600	10,982	12,338	14,041	16,314	17,240	18,101	21,337		
23	7,529	9,260	10,196	11,293	11,688	13,091	14,848	17,187	18,137	19,021	22,337		
24	8,035	9,886	10,856	11,992	12,401	13,848	15,659	18,062	19,037	19,943	23,337		
25	8,649	10,520	11,524	12,697	13,120	14,611	16,173	18,940	19,939	20,887	24,337		
26	9,222	11,160	12,198	13,409	13,844	15,379	17,292	19,820	20,843	21,792	25,336		
27	9,803	11,808	12,879	14,125	14,573	16,151	18,114	20,703	21,749	22,719	26,136		
28	10,391	12,461	13,565	14,847	15,308	16,928	18,937	21,588	22,657	23,617	27,336		
29	10,986	13,121	14,256	15,574	16,047	17,708	19,768	22,475	23,567	24,577	28,336		
30	11,588	13,787	14,953	16,306	16,791	18,493	20,599	23,364	24,478	25,508	29,336		

v	Імовірність									
	0,30	0,25	0,20	0,10	0,05	0,025	0,02	0,01	0,005	0,001
1	1,074	1,323	1,642	2,706	3,841	5,024	5,412	6,635	7,879	10,827
2	2,408	2,773	3,219	4,605	5,991	7,378	7,824	9,210	10,597	13,815
3	3,665	4,108	4,642	6,251	7,815	9,348	9,837	11,345	12,838	16,268
4	4,878	5,385	5,989	7,779	9,488	11,143	11,668	13,277	14,860	18,465
5	6,064	6,626	7,289	9,236	11,070	12,839	13,388	15,086	16,750	20,517
6	7,231	7,841	8,558	10,645	12,592	14,449	15,033	16,812	18,548	22,457
7	8,383	9,037	9,803	12,017	14,067	16,013	16,622	18,475	20,278	24,322
8	9,524	10,219	11,030	13,362	15,507	17,535	18,168	20,090	21,955	26,125
9	10,656	11,389	12,242	14,684	16,919	19,023	19,679	21,666	23,589	27,877
10	11,781	12,549	13,412	15,987	18,307	20,483	21,161	23,209	25,188	29,588
11	12,899	13,701	14,631	17,275	19,675	21,920	22,618	24,725	26,757	31,264
12	14,011	14,845	15,812	18,549	21,026	23,337	24,054	26,217	28,300	32,909
13	15,119	15,984	16,985	19,812	22,362	24,736	25,472	27,688	29,819	34,528
14	16,222	17,117	18,151	21,064	23,685	26,119	26,873	29,141	31,319	36,123
15	17,322	18,245	19,311	22,307	24,996	27,488	28,259	30,578	32,801	37,697
16	18,418	19,369	20,465	23,542	26,296	28,845	29,633	32,000	34,267	39,252

Закінчення табл. 3

v	Імовірність									
	0,30	0,25	0,20	0,10	0,05	0,025	0,02	0,01	0,005	0,001
17	19,511	20,489	21,615	24,769	27,587	30,191	30,995	33,409	35,718	40,790
18	20,601	21,605	22,760	25,989	28,869	31,526	32,346	34,805	37,156	42,312
19	21,689	22,718	23,900	27,204	30,144	32,852	33,687	36,191	38,582	43,820
20	22,775	23,828	25,038	28,412	31,410	34,170	35,020	37,566	39,997	45,315
21	23,858	24,935	26,171	29,615	32,671	35,479	36,343	38,932	41,401	46,797
22	24,939	26,039	27,301	30,813	33,924	36,781	37,659	40,289	42,796	48,268
23	26,018	27,141	28,429	32,007	35,172	38,076	38,968	41,638	44,181	49,728
24	27,096	28,241	29,553	33,196	36,415	39,364	40,270	42,980	45,558	51,170
25	28,172	29,339	30,675	34,382	37,652	40,046	41,566	44,314	46,928	52,620
26	29,246	30,434	31,795	35,563	38,885	41,923	42,856	45,642	48,290	54,052
27	30,319	31,528	32,912	36,741	40,113	43,194	44,140	46,963	49,645	55,476
28	31,391	32,620	34,027	37,916	41,337	44,461	45,419	48,278	50,993	56,893
29	32,461	33,711	35,139	39,087	42,557	45,722	46,693	49,588	52,336	58,302
30	33,530	34,800	36,250	40,256	43,773	46,979	47,962	50,892	53,672	59,703

Таблиця 4

Розподіл Фішера — Снедекора (F -розподіл)

Значення $F_{\text{табл}}$ за умови $P (F > F_{\text{табл}})$. Перше значення відповідає імовірності 0,05, друге — імовірності 0,01 і третє — імовірності 0,001; v_1 — число степенів вільності чисельника, v_2 — число степенів вільності знаменника

$v_1 \backslash v_2$	1	2	3	4	5	6	8	12	24	∞	t
1	161,4 4052 406523	199,5 4999 500016	215,7 5403 536700	224,6 5625 562527	230,2 5764 576449	234,0 5859 585953	238,9 5981 598149	243,9 6106 610598	249,0 6234 623432	253,3 6366 636535	12,71 63,66 636,2
2	18,51 98,49 998,46	19,00 99,01 999,00	19,16 00,17 999,20	19,25 99,25 999,20	19,30 99,30 999,20	19,33 99,33 999,20	19,37 99,36 999,40	19,41 99,42 999,60	19,45 99,46 999,40	19,50 99,50 999,40	4,30 9,92 31,00
3	10,13 34,12 67,47	9,55 30,81 148,51	9,28 29,46 141,10	9,12 28,71 137,10	9,01 28,24 134,60	8,94 27,91 132,90	8,84 27,49 130,60	8,74 27,05 128,30	8,64 26,60 125,90	8,53 26,12 123,50	3,18 5,84 12,94
4	7,71 21,20 74,13	6,94 18,00 61,24	6,59 16,69 56,18	6,39 15,98 53,43	6,26 15,52 51,71	6,16 15,21 50,52	6,04 14,80 49,00	5,91 14,37 47,41	5,77 13,93 45,77	5,63 13,46 44,05	2,78 4,60 8,61
5	6,61 16,26 47,04	5,79 13,27 36,61	5,41 12,06 33,20	5,19 11,39 31,09	5,05 10,97 20,75	4,95 10,67 28,83	4,82 10,27 27,64	4,68 9,89 26,42	4,53 9,47 25,14	4,36 9,02 23,78	2,57 4,03 6,86
6	5,99 13,74 35,51	5,14 10,92 26,99	4,76 9,78 23,70	4,53 9,15 21,90	4,39 8,75 20,81	4,28 8,47 20,03	4,15 8,10 19,03	4,00 7,72 17,99	3,84 7,31 16,89	3,67 6,88 15,75	2,45 3,71 5,96

Продовження табл. 4

$\frac{v_1}{v_2}$	1	2	3	4	5	6	8	12	24	∞	t
7	5,59 12,25 29,22	4,74 9,55 21,69	4,35 8,45 18,77	4,12 7,85 17,19	3,97 7,46 16,21	3,87 7,19 15,52	3,73 6,84 14,63	3,57 6,47 13,71	3,41 6,07 12,73	3,23 5,65 11,70	2,36 3,50 5,40
8	5,32 11,26 25,42	4,46 8,65 18,49	4,07 7,59 15,83	3,84 7,10 14,39	3,69 6,63 13,49	3,58 6,37 12,86	3,44 6,03 12,04	3,28 5,67 11,19	3,12 5,28 10,30	2,99 4,86 9,35	2,31 3,36 5,04
9	5,12 10,56 22,86	4,26 8,02 16,39	3,86 6,99 13,90	3,63 6,42 12,56	3,48 6,06 11,71	3,37 5,80 11,13	3,23 5,47 10,37	3,07 5,11 9,57	2,90 4,73 8,72	2,71 4,31 7,81	2,26 3,25 4,78
10	4,96 10,04 21,04	4,10 7,56 14,91	3,71 6,55 12,55	3,48 5,99 1,28	3,33 5,64 10,48	3,22 5,39 9,92	3,07 5,06 9,20	2,91 4,71 8,45	2,74 4,33 7,64	2,54 3,91 6,77	2,23 3,17 4,59
11	4,84 9,65 19,69	3,98 7,20 13,81	3,59 6,22 11,56	3,36 5,67 10,35	3,20 5,32 9,58	3,09 5,07 9,05	2,95 4,74 8,35	2,79 4,49 7,62	2,61 4,02 6,85	2,40 3,60 6,00	2,20 3,11 4,49
12	4,75 9,33 18,64	3,88 6,93 12,98	3,49 5,95 10,81	3,26 5,41 9,63	3,11 5,06 8,89	3,00 4,82 8,38	2,85 4,50 7,71	2,69 4,16 7,00	2,5 3,78 6,25	2,30 3,36 5,42	2,18 3,06 4,32
13	4,67 9,07 17,81	3,80 6,70 12,31	3,41 5,74 10,21	3,18 5,20 9,07	3,02 4,86 8,35	2,92 4,62 7,86	2,77 4,30 7,21	2,60 3,96 6,52	2,42 3,59 5,78	2,21 3,16 4,97	2,16 3,01 4,12
14	4,60 8,86 17,14	3,74 6,51 11,78	3,34 5,56 9,73	3,11 5,03 8,62	2,96 4,69 7,92	2,85 4,46 7,44	2,70 4,14 6,8	2,53 3,80 6,13	2,35 3,43 5,41	2,13 3,00 4,60	2,14 2,98 4,14

Продовження табл. 4

$\frac{v_1}{v_2}$	1	2	3	4	5	6	8	12	24	∞	t
15	4,45	3,68	3,29	3,06	2,90	2,79	2,64	2,48	2,29	2,07	2,13
	8,68	6,36	5,42	4,89	4,56	4,32	4,00	3,67	3,29	2,87	2,95
	16,59	11,34	9,34	8,25	7,57	7,09	6,47	5,81	5,10	4,31	4,07
16	4,41	3,63	3,24	3,01	2,85	2,74	2,59	2,42	2,24	2,01	2,12
	8,53	6,23	5,29	4,77	4,44	4,20	3,89	3,55	3,18	2,75	2,92
	16,12	10,97	9,01	7,94	7,27	6,80	6,20	5,55	4,85	4,06	4,02
17	4,45	3,59	3,20	2,96	2,81	2,70	2,55	2,38	2,19	1,96	2,11
	8,40	6,11	5,18	4,67	4,34	4,10	3,79	3,45	3,08	2,65	2,90
	15,72	10,66	8,73	7,68	7,02	6,56	5,96	5,32	4,63	3,85	3,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,51	2,34	2,15	1,92	2,10
	8,28	6,01	5,09	4,58	4,25	4,01	3,71	3,37	3,01	2,57	2,88
	15,38	10,39	8,49	7,46	6,81	6,35	5,76	5,13	4,45	3,67	3,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,48	2,31	2,11	1,88	2,09
	8,18	5,93	5,01	4,50	4,17	3,94	3,63	3,30	2,92	2,49	2,86
	15,08	10,16	8,28	7,26	6,61	6,18	5,59	4,97	4,29	3,52	3,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,28	2,08	1,84	2,09
	8,10	5,85	4,94	4,43	4,10	3,87	3,56	3,23	2,86	2,42	2,84
	14,82	9,95	8,10	7,10	6,46	6,02	5,44	4,82	4,15	3,38	3,85
21	4,32	3,47	3,07	2,84	2,68	2,57	2,42	2,25	2,05	1,82	2,08
	8,02	5,78	4,87	4,37	4,04	3,81	3,51	3,17	2,80	2,36	2,83
	14,62	9,77	7,94	6,95	6,32	5,88	5,31	4,70	4,03	3,26	3,82
22	4,30	3,44	3,05	2,82	2,66	2,55	2,40	2,23	2,03	1,78	2,07
	7,94	5,72	4,82	4,31	3,99	3,75	3,45	3,12	2,75	2,30	2,82
	14,38	9,61	7,80	6,81	6,19	5,76	5,19	4,58	3,92	3,15	3,79

Продовження табл. 4

$\frac{v_1}{v_2}$	1	2	3	4	5	6	8	12	24	∞	t
23	4,28 7,88 14,19	3,42 5,66 9,46	3,03 4,76 7,67	2,80 4,26 6,70	2,64 3,94 6,08	2,53 3,71 5,56	2,38 3,41 5,09	2,20 3,07 4,48	2,00 2,70 3,82	1,76 2,26 3,05	2,07 2,81 3,77
24	4,26 7,82 14,03	3,40 5,61 9,34	3,01 4,72 7,55	2,78 4,22 6,59	2,62 3,90 5,98	2,51 3,67 5,55	2,36 3,36 4,99	2,18 3,03 4,39	1,98 2,66 3,74	1,73 2,21 2,97	2,06 2,80 3,75
25	4,24 7,77 13,88	3,38 5,57 9,22	2,99 4,68 7,45	2,76 4,18 6,49	2,60 3,86 5,89	2,49 3,63 5,46	2,34 3,32 4,91	2,16 2,99 4,31	1,96 2,62 3,66	1,71 2,17 2,89	2,06 2,79 3,72
26	4,22 7,72 13,74	3,37 5,53 9,12	2,9 4,64 7,36	2,74 4,14 6,41	2,59 3,82 5,80	2,47 3,59 5,38	2,32 3,29 4,83	2,15 2,96 4,24	1,95 2,58 3,59	1,69 2,13 2,82	2,06 2,78 3,71
27	4,21 7,68 13,61	3,35 5,49 9,02	2,96 4,60 7,27	2,73 4,11 6,33	2,57 3,78 5,73	2,46 3,56 5,31	2,30 3,26 4,76	2,13 2,93 4,17	1,93 2,55 3,52	1,67 2,10 2,76	2,05 2,77 3,69
28	4,19 7,64 13,50	3,34 5,54 8,93	2,95 4,57 7,18	2,71 4,07 6,25	2,56 3,75 5,66	2,44 3,53 5,24	2,29 3,23 4,69	2,12 2,90 4,11	1,91 2,52 3,46	1,65 2,06 2,70	2,05 2,76 3,67
29	4,18 7,60 13,39	3,33 5,42 8,85	2,93 4,54 7,12	2,70 4,04 6,19	2,54 3,73 5,59	2,43 3,50 5,18	2,28 3,20 4,65	2,10 2,87 4,05	1,90 2,49 3,41	1,64 2,03 2,64	2,05 2,76 3,66
30	4,17 7,56 13,29	3,32 5,39 8,77	2,92 4,51 7,05	2,69 4,02 6,12	2,53 3,70 5,53	2,42 3,47 5,12	2,27 3,17 4,58	2,09 2,84 4,00	1,89 2,47 3,36	1,62 2,01 2,59	2,04 2,75 3,64

Закінчення табл. 4

$\frac{v_1}{v_2}$	1	2	3	4	5	6	8	12	24	∞	t
60	4,00 7,08 11,97	3,15 4,98 7,76	2,76 4,13 6,17	2,52 3,65 5,31	2,37 3,34 4,76	2,25 3,12 4,37	2,10 2,82 3,87	1,92 2,50 3,31	1,70 2,12 2,76	1,39 1,60 1,90	2,00 2,66 3,36
∞	3,84 6,64 10,83	2,99 4,60 6,91	2,60 3,78 5,42	2,37 3,32 4,62	2,21 3,02 4,10	2,09 2,80 3,74	1,94 2,51 3,27	1,75 2,18 2,74	1,52 1,79 2,13	1,03 1,04 1,05	1,96 2,58 3,29

Таблиця 5

Значення щільності $f(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$ для нормованого нормального закону розподілу, $F(-t) = f(t)$

Цілі та десяткові частки t	Соті частки t									
	0	1	2	3	4	5	6	7	8	9
0	0,3989	0,3989	0,3989	0,3988	0,3986	0,3984	0,3982	0,398	0,3977	0,3973
0,1	3970	3965	3961	3956	3951	3945	3939	3932	3925	3918
0,2	3910	3902	3894	3885	3876	3867	3857	3847	3836	3825
0,3	3814	3802	3790	3778	3765	3752	3739	3726	3712	3697
0,4	3683	3668	3653	3637	3621	3605	3589	3572	3555	3538
0,5	3525	3503	3485	3467	3448	3429	3410	3391	3372	3352
0,6	3332	3312	3292	3271	3251	3230	3209	3187	3166	3144
0,7	3123	3101	3079	3056	3034	3011	2989	2966	2943	2920

Продовження табл. 5

Цілі та десяти- кові частки t	Соті частки t									
	0	1	2	3	4	5	6	7	8	9
0,8	2897	2874	2850	2827	2803	2780	2756	2732	2709	2685
0,9	2661	2631	2613	2589	2565	2541	2516	2492	2468	2444
1,0	0,242	0,2396	0,2371	0,2347	0,2323	0,2299	0,2275	0,2251	0,2227	0,2203
1,1	2179	2155	2131	2107	2083	2059	2036	3012	1989	1965
1,2	1942	1919	1895	1872	1849	1826	1804	1781	1758	1736
1,3	1714	1691	1669	1647	1626	1604	1582	1561	1539	1518
1,4	1497	1476	1456	1435	1415	1394	1374	1354	1334	1315
1,5	1295	1276	1257	1238	1219	1200	1182	1163	1145	1127
1,6	1109	1092	1074	1057	1040	1023	1006	9989	9973	9957
1,7	0940	0925	0909	0893	0878	0863	0848	0833	0818	0804
1,8	0790	0775	0762	0748	0734	0721	0707	0694	0681	0669
1,9	0656	0644	0632	0620	0608	0596	0584	0573	0562	0551
2,0	0,0540	0,0529	0,0519	0,0508	0,0498	0,0488	0,478	0,0568	0,0459	0,0449
2,1	0440	0431	0422	0413	0404	0396	0387	0379	0371	0363
2,2	0355	0347	0339	0332	0325	0317	0310	0303	0297	0290
2,3	0283	0277	0270	0264	0258	0252	0246	0241	0235	0229
2,4	0224	0219	0213	0208	0203	0198	0194	0189	0184	0180

Цілі та десяти- кові частки t	Соті частки t									
	0	1	2	3	4	5	6	7	8	9
2,5	0175	0171	0167	0163	0158	0154	0151	0147	0143	0139
2,6	0136	0132	0129	0126	0122	0119	0116	0113	0110	0107
2,7	0104	0101	0099	0096	0093	0091	0088	0086	0084	0081
2,8	0079	0077	0075	0073	0071	0069	0067	0065	0063	0061
2,9	0060	0058	0056	0055	0053	0051	0050	0048	0047	0046
3,0	0,0044	0,0043	0,0042	0,0040	0,0039	0,0038	0,0037	0,0036	0,0035	0,0034
3,1	0033	0032	0031	0030	0029	0028	0027	0026	0025	0025
3,2	0024	0023	0022	0022	0021	0020	0020	0019	0018	0018
3,3	0017	0017	0016	0016	0015	0015	0014	0014	0013	0013
3,4	0012	0012	0012	0011	0011	0010	0010	0010	0009	0009
3,5	0009	0008	0008	0008	0008	0007	0007	0007	0007	0006
3,6	0006	0006	0006	0005	0005	0005	0005	0005	0005	0004
3,7	0004	0004	0004	0004	0004	0004	0003	0003	0003	0003
3,8	0003	0003	0003	0003	0003	0002	0002	0002	0002	0002
3,9	0002	0002	0002	0002	0002	0002	0002	0002	0001	0001
4,0	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001

Таблиця 6

Значення функція Пуассона $P(X = m) = \frac{\lambda^m}{m!} e^{-\lambda}$

$\lambda \backslash m$	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0
0	0,9048	0,8187	0,7408	0,6703	0,6065	0,5488	0,4966	0,4493	0,4066	0,3679
1	0,0905	0,1637	0,2223	0,2681	0,3033	0,3293	0,3476	0,3595	0,3659	0,3679
2	0,0045	0,0164	0,0333	0,0536	0,0758	0,0988	0,1216	0,1438	0,1547	0,1839
3	0,0002	0,0011	0,0033	0,0072	0,012	0,0198	0,0284	0,0383	0,0494	0,0613
4	0,0000	0,0001	0,0003	0,0007	0,0016	0,0030	0,0050	0,0077	0,0111	0,0153
5	0,0000	0,0000	0,0000	0,0001	0,0002	0,0003	0,0007	0,0012	0,0020	0,0031
6	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0003	0,0005
7	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
$\lambda \backslash m$	2,0	3,0	4,0	5,0	6,0	7,0	8,0	9,0	10,0	
0	0,1353	0,0498	0,0183	0,0067	0,0025	0,0009	0,0003	0,0001	0,0001	
1	0,2707	0,1494	0,0733	0,0337	0,0149	0,0064	0,0027	0,0011	0,0005	
2	0,2707	0,2240	0,1465	0,0842	0,0446	0,0223	0,0107	0,0050	0,0023	
3	0,1805	0,2240	0,1954	0,1404	0,892	0,0521	0,0286	0,0150	0,0076	
4	0,0902	0,1681	0,1954	0,1755	0,1339	0,0912	0,0572	0,0337	0,0189	

Закінчення табл. 6

$\lambda \backslash m$	2,0	3,0	4,0	5,0	6,0	7,0	8,0	9,0	10,0
5	0,0361	0,1008	0,1563	0,1755	0,1606	0,1277	0,0916	0,0607	0,0378
6	0,0120	0,0504	0,1042	0,1462	0,1606	0,1490	0,1221	0,0911	0,0631
7	0,0034	0,0216	0,0595	0,1045	0,1377	0,1490	0,1396	0,1171	0,0901
8	0,0009	0,0081	0,0298	0,0653	0,1033	0,1304	0,1396	0,1318	0,1126
9	0,0002	0,0027	0,0132	0,363	0,0689	0,1014	0,1241	0,1318	0,1251
10	0,0000	0,0008	0,0053	0,0181	0,0413	0,0710	0,0993	0,1186	0,1251
11	0,0000	0,0002	0,0019	0,0082	0,0225	0,0452	0,0722	0,970	0,1137
12	0,0000	0,0001	0,0006	0,0034	0,0113	0,0264	0,0481	0,728	0,0948
13	0,0000	0,0000	0,0002	0,0013	0,0052	0,0142	0,0296	0,0504	0,0729
14	0,0000	0,0000	0,0001	0,0005	0,0022	0,0071	0,0169	0,0324	0,0521
15	0,0000	0,0000	0,0000	0,0002	0,0009	0,0033	0,0090	0,0194	0,0347
16	0,0000	0,0000	0,0000	0,0000	0,0003	0,0015	0,0045	0,0109	0,0217
17	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006	0,0021	0,0058	0,0128
18	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0009	0,0029	0,0071
19	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0014	0,0037

Навчальне видання

ГУРЖІЙ Андрій Миколайович
ДУДАР Зоя Володимирівна
ЛЕВИКІН Віктор Макарович
ШАМША Борис Володимирович

МАТЕМАТИЧНЕ ЗАБЕЗПЕЧЕННЯ ІНФОРМАЦІЙНО-КЕРУЮЧИХ СИСТЕМ

Підручник
для вищих навчальних закладів

Редактор — *Л. В. Ситнік*
Комп'ютерна верстка — *Т. Є. Сергієнко*
Художнє оформлення — *А. В. Пивоваров*

Видано за рахунок державних коштів
Продаж заборонено

Підписано до друку 22.12.2005. Формат 60 × 90 ¹/₁₆. Друк офсетний.
Папір офсетний. Гарнітура SchoolBookC. Умов. друк. арк. 28,0.
Обл.-вид. арк. 31,66. Тираж 2250 прим. Зам. №

Видавництво «Компанія СМІТ»
61166, м. Харків, просп. Леніна, 14
Тел.: 8(057)717-54-94
Факс: 8(057)702-13-07
E-mail: book@smit.kharkov.ua
<http://www.smit-book.com>

Свідцтво про внесення суб'єкта видавничої справи до державного реєстру видавців,
виготівників і розповсюджувачів видавничої продукції ДК № 435 від 26.04.2001

Друк — ТОВ «СІМ»
61012, м. Харків, пров. Лопанський, 3-а
Тел.: 8(057)719-19-30
E-mail: tipa_graf@pisem.net