МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ

А. М. Єріна

Статистичне моделювання та прогнозування

Навчальний посібник

Допущено Міністерством освіти і науки України

Київ 2001

ББК 65.В6
Є 71

Рецензенти

В. Л. Ревенко, д-р екон наук
(Міжнар наук -навч центр інформ технологій
і систем НАНУ та М-ва освіти і науки України)

О. П. Гусєва, канд екон наук, доц
(Одеський держ екон ун-т)

Гриф надано Міністерством освіти і науки України
Лист № 2/1147 від 10.07.2000

Єріна А. М.

Є 71 Статистичне моделювання та прогнозування: Навч. посібник. — К.: КНЕУ, 2001. — 170 с.

ISBN 966-574-209-4

У навчальному посібнику розглядаються методологічні принципи статистичного моделювання та прогнозування соціально-економічних явищ процесів, різні модифікації моделей динаміки, структури і взаємозв'язків, умови адаптації їх до специфіки об'єктів моделювання.

Аналітичні можливості та межі застосування конкретних моделей ілюструються на прикладах, різних за соціально-економічним змістом та інформаційною базою. Розрахунки виконано за технологіями статистичного аналізу та обробки даних, реалізованими в системі Statistica 5

Для студентів, аспірантів, викладачів, науковців і практиків — усіх, хто прагне опанувати методи поглибленого аналізу закономірностей, формування масових соціально-економічних явищ і передбачення їх розвитку в умовах невизначеності

ББК 65.В6

ISBN 966-574-209-4

ПЕРЕДМОВА

Статистичний аналіз даних стає невід'ємним атрибутом системи управління на усіх її рівнях — від невеликої фірми до національної економіки в цілому. Статистичні моделі використовують для діагностики стану об'єктів управління, при вивченні причинно-наслідкового механізму формування варіації та динаміки соціально-економічних явищ і процесів, у моніторингу економічної кон'юнктури, при прогнозуванні та прийнятті оптимальних управлінських рішень.

Оволодіння багатим арсеналом методів статистичної обробки даних з використанням комп'ютерних технологій є важливою складовою професійної підготовки економіста. Саме цій меті підпорядковано курс «Статистичне моделювання та прогнозування». Відповідно до програми курсу в навчальному посібнику розглядаються:

- методологічні принципи статистичного моделювання та прогнозування, перевірки гіпотез і верифікації прогнозів;
- моделі багатовимірних оцінок (рейтингів, латентних факторів) і моделі класифікацій;
- різноманітні моделі динаміки (трендові, сезонного ритму, повного циклу), комплексне їх використання при прогнозуванні;
- модифікації множинної регресії; адаптація основних засад регресійного аналізу до специфіки об'єктів моделювання та інформаційної бази;
- моделі багатофакторного прогнозування за даними взаємозв'язаних динамічних рядів;
- моделювання причинних комплексів системами рівнянь.

При викладенні навчального матеріалу наголошується на двох аспектах:

1) на аналітичних можливостях і межах застосування кожного типу моделей;

2) на використанні інтегрованої системи обробки даних *Statistica,* яка надає користувачеві унікальні можливості поглибленого аналізу статистичних закономірностей.

Логічна структура аналізу ілюструється на конкретних прикладах соціально-економічного змісту (за умовними даними). Для кожного типу моделей розглядаються принципи формування інформаційної бази, вибору процедур аналізу, інтерпретації результатів. Методологія обробки даних у системі *Statistica* ґрунтується на електронних таблицях типу *MS Excel.*

Акцентуючи увагу студентів на параметрах моделей, таблиці з результатами аналізу і графіки наводяться у стандартному вигляді англійською мовою. Специфікація включених у модель ознак і змістовна інтерпретація параметрів моделі розкривається в коментарях до таблиць і графіків.

Для ймовірнісної оцінки параметрів моделей у таблицях результатів пропонуються фактичні рівні істотності *p-level.* З метою самостійної перевірки гіпотез щодо окремих властивостей процесу чи адекватності моделі в цілому в додатках наведено фрагменти таблиць найпоширеніших статистичних критеріїв.

Посібник рекомендується для студентів, аспірантів, викладачів, науковців і практиків, діяльність яких пов'язана з обробкою та аналізом статистичної інформації.

Розділ 1

МЕТОДОЛОГІЧНІ ОСНОВИ СТАТИСТИЧНОГО МОДЕЛЮВАННЯ ТА ПРОГНОЗУВАННЯ

1.1. ЛОГІКА ПРИКЛАДНОГО СТАТИСТИЧНОГО МОДЕЛЮВАННЯ

Моделювання — один з ефективних засобів пізнання законів і закономірностей навколишнього світу. Суть моделювання полягає в заміні реального процесу певною конструкцією, яка відтворює основні, найістотніші риси процесу, абстрагуючись від вторинних, неістотних. Будь-яка конструкція — фізична чи математична — це спрощений, схематичний образ реальності. Мистецтво моделювання саме й полягає в тому, щоб знати, що, де, коли та як можна і треба спрощувати.

Особливого значення набувають моделі при вивченні закономірностей масових процесів, які недоступні прямому спостереженню і не піддаються експериментуванню. Передусім це стосується соціально-економічних явищ і процесів, закономірності яких формуються під впливом безлічі взаємопов'язаних факторів і за складністю переважають закони фізики, хімії чи біології.

За своєю природою соціально-економічні явища і процеси — стохастичні, ймовірнісні; невизначеність — їх внутрішня властивість. Вивчення цих процесів, передбачення перспектив їх подальшого розвитку, прийняття оптимальних управлінських рішень мають спиратися на такі моделі, які й в умовах невизначеності забезпечують сталість і надійність висновків. Такими є статистичні моделі. Вони належать до класу математичних, виражаються у формі рівнянь, функцій, алгоритмів; при їх розв'язуванні поєднуються логіко-алгебраїчні та ймовірнісні методи.

Формально статистична модель являє собою абстрактну схему відношень між величинами, що характеризують властивості реального процесу. Вибір же цих властивостей і розробка схем відношень між ними здійснюється неформальним шляхом. На основі апріорного аналізу природи процесу формулюються гіпотези щодо окремих його властивостей і закономірностей. Гіпотези перевіряються на фактичних даних.

Зв'язок між математичною схемою моделі і реальним процесом забезпечується поєднанням у моделі інформації двох типів:

1) апріорі логічно обґрунтованих гіпотез щодо природи та характеру властивостей процесу, співвідношень і взаємозв'язків між ними;

2) емпіричних даних, які характеризують ці властивості.

Модель встановлює відповідність між сукупністю фактів і гіпотезами, імітує механізм формування закономірностей. На моделях проводяться експерименти, результати яких поширюються на реальність. Основна вимога, що ставиться до моделі, — подібність, адекватність її реальному процесу.

Аби зрозуміти загальну логіку статистичного моделювання, умовно розкладемо його на етапи:

1) Характеристика мети та об'єкта моделювання.
2) Розвідувальний аналіз даних.
3) Математична формалізація моделі.
4) Оцінювання параметрів моделі.
5) Перевірка адекватності моделі.
6) Аналіз та інтерпретація результатів.

На першому етапі визначаються мета та об'єкт моделювання. *Мета* — це кінцеве призначення моделі. Скажімо, діагностика процесу, аналіз механізму його формування, тенденцій розвитку тощо. Залежно від мети дослідження один і той самий процес можна описати різними моделями.

Об'єктом моделювання виступає статистична сукупність, в якій реалізується закономірність. Формально будь-яку сукупність можна представити у вигляді впорядкованого набору даних з параметрами *n, m, T,* де *n* — кількість елементів сукупності (j=1, 2, ..., *n), m* — кількість зареєстрованих у j-го елемента ознак (i = 1, 2,..., m), *T* — календарний термін періоду з певними квантами часу (рік, квартал, місяць, доба тощо). Отже, інформаційна одиниця об'єкта моделювання — значення i-ї ознаки у j-го елемента сукупності у t-му періоді — x_{ijt}. Якщо сукупність вивчається в статиці, то інформація представляється матрицею $n \cdot m$, якщо в динаміці, то матрицею $T \cdot m$.

Характеристика об'єкта моделювання включає такі моменти:

* вибір одиничного елемента сукупності — носія характерних для закономірності рис;
* визначення просторових і часових меж об'єкта моделювання;
* формування ознакової множини моделі.

Вибір первинного елемента сукупності залежить від рівня об'єкта моделювання. Скажімо, продуктивність праці можна вивчати на рівні галузі, окремих підприємств, цехів і навіть окремих робітників. Очевидно, що у кожному випадку елемент сукупності

буде іншим. Межі об'єкта моделювання задаються обсягом сукупності *n* для статичних моделей і тривалістю періоду *T* — для динамічних.

При формуванні ознакової множини *X* вирішальну роль відіграють експертні оцінки значущості та інформативності окремих ознак, враховується можливість їх точного вимірювання, діапазон варіації, трудомісткість збирання інформації.

У статистичному моделюванні сукупність завжди розглядається як вибірка — класична чи гіпотетична. Класична вибірка — це частина реальної генеральної сукупності, відібрана для обстеження за принципами вибіркового методу. Гіпотетична генеральна сукупність оперує не кількістю елементів, а кількістю можливих наслідків функціонування об'єкта моделювання в одних і тих самих умовах. Отже, фактичні дані, навіть якщо вони є результатом суцільного обстеження сукупності, розглядаються як випадкові реалізації стохастичного, непередбачуваного процесу. Це дає підстави для ймовірнісного оцінювання результатів моделювання.

Завдання ймовірнісного оцінювання — встановити, наскільки виявлена закономірність позбавлена випадкових впливів, наскільки вона характерна для того комплексу умов, у яких функціонує об'єкт моделювання. Якісна своєрідність і неповторність статистичних сукупностей потребує інтерпретації цих оцінок щодо конкретних умов простору і часу. В окремих випадках імовірнісне оцінювання результатів суцільного спостереження недоречне, скажімо, при визначенні рейтингів окремих елементів сукупності. Проте мета конкретного дослідження не може відкинути правомірність використання таких оцінок.

Розвідувальний аналіз даних передбачає:

* статистичне описування об'єкта — визначення середніх, стандартних відхилень, інших характеристик розподілу;
* уніфікацію типів ознак, приведення їх до одного виду;
* тестування сукупності на однорідність, ідентифікацію аномальних спостережень;
* відтворення пропущених даних;
* оцінювання взаємозв'язків між ознаками.

Побудова моделі ґрунтується на основі певних правил та алгоритмів, які визначають порядок розрахунків і математичних дій, необхідних для обробки інформації. На етапі *математичної формалізації моделі* обґрунтовується алгебраїчна форма розрахунків, відношення між властивостями процесу описуються символами та знаками, порядок розрахунків — блок-схемами.

Оцінювання параметрів моделі — це етап комп'ютерної обробки даних. В 1.4 анонсується система *Statistica,* яка надає унікальні можливості експериментування, розвідки, графічного відображення і поглибленого аналізу даних, у якій сучасні методи статистичного моделювання та прогнозування реалізовані з використанням новітніх комп'ютерних технологій.

Перевірка адекватності моделі означає оцінювання ступеня відповідності параметрів моделі характеристикам об'єкта. На цьому етапі використовують різні процедури порівняння модельних висновків, перевірки статистичних гіпотез за допомогою статистичних критеріїв. Перевірка адекватності моделі має сенс лише щодо мети дослідження і не може бути абстрактною.

Заключний етап моделювання — *аналіз та інтерпретація результатів* — один із найскладніших і найвідповідальніших. Складність його полягає у тому, що для інтерпретації результатів не існує готових алгоритмів чи рецептів. Єдина спільна для всіх моделей вимога — інтерпретація має узгоджуватися з первинними гіпотезами. Основні висновки формулюються в змістовних термінах: зміст параметрів моделі, правильність перевірюваних гіпотез, оцінювання ступеня їх вірогідності.

Отже, можна сформулювати два принципи статистичного моделювання:

* підпорядкованість меті дослідження на всіх етапах моделювання;

* забезпечення адекватності моделі.

Слід пам'ятати, що єдино правильної, «ідеальної» моделі не існує. Ту ж саму закономірність можна описати різними моделями. Вибір того чи іншого типу моделі залежить від мети дослідження, специфіки процесу (явища), масштабу об'єкта моделювання, наявної інформації, технічного та програмного забезпечення.

1.2. СУТНІСТЬ І ВИДИ СТАТИСТИЧНИХ ПРОГНОЗІВ

Одна з найскладніших проблем системи управління — передбачити майбутнє і віднайти ефективні рішення в умовах невизначеності. Інструментом мінімізації невизначеності слугує *прогнозування,* а *прогнозом* називають науково обґрунтований висновок про майбутні події, про перспективи розвитку процесів, про можливі наслідки управлінських рішень.

За специфікою об'єктів прогнозування прогнози поділяють на науково-технічні, економічні, соціальні, військово-політичні тощо. Економічні прогнози, в свою чергу, класифікують за масштабністю об'єкта на глобальні (світові), макроекономічні, структурні (міжгалузеві та міжрегіональні), регіональні, галузеві, мікроекономічні.

Прогнозування передбачає систему наукових доведень, використання методів і прийомів з різним ступенем формалізації, узгодженість окремих висновків і оцінок щодо майбутнього розвитку процесу. В світовій практиці прикладного прогнозування використовують різні методи: статистичні (прогнозна екстраполяція), функціонально-ієрархічні (прогнозні сценарії), методи структурної аналогії, імітаційного моделювання, експертні оцінки. Кожен метод має свої особливості, позитивні якості й вади, свої межі використання.

При прогнозуванні соціально-економічних процесів перевага віддається статистичним методам, прогнозним результатом яких є очікувані у майбутньому значення характеристик процесу.

Очевидно, що майбутнє неможливо спостерігати, а очікуваний результат — виміряти, його можна лише передбачити за певних умов, скажімо, *«...якщо тенденція не зміниться, то...»* або *«...якщо станеться подія А, то...»* і т. ін. Якщо умови змінять-ся, то автоматично зміниться й результат прогнозування. Отже, статистичний прогноз, побудований за схемою *«...якщо, то...»,* завжди є умовним.

Іншою особливістю статистичного прогнозу є визначеність його в часі. Часовий горизонт прогнозу називають *періодом упередження.* За тривалістю цього періоду вирізняють прогнози: короткострокові (до 1 року), середньострокові (до 5 років) і довгострокові (від 5 до 20 років і більше). Тривалість періоду упередження залежить від специфіки об'єкта прогнозування, інтенсивності динаміки, тривалості дії виявлених закономірностей та тенденцій.

Прогнозний результат на період упередження можна представити одним числом (точковий прогноз) або інтервалом значень, до якого з певною ймовірністю належить прогнозна величина (інтервальний прогноз).

Статистичні прогнози ґрунтуються на гіпотезах про стабільність значень величини, що прогнозується; закону її розподілу; взаємозв'язків з іншими величинами тощо. Основний інструмент прогнозування — *екстраполяція.*

Суть прогнозної екстраполяції полягає в поширенні закономірностей, зв'язків і відношень, виявлених в *t-му* періоді, за його

межі. Залежно від гіпотез щодо механізму формування і подальшого розвитку процесу використовуються різні методи прогнозної екстраполяції, їх можна об'єднати в дві групи:

• екстраполяція закономірностей розвитку — тенденцій і коливань;

• екстраполяція причинно-наслідкового механізму формування процесу — багатофакторне прогнозування.

Ці методи різняться не процедурою розрахунків прогнозу, а способом описування об'єкта моделювання. Екстраполяція закономірностей розвитку ґрунтується на вивченні його передісторії, виявленні загальних і усталених тенденцій, траєкторій зміни в часі. Абстрагуючись від причин формування процесу, закономірності його розвитку розглядають як функцію часу. Інформаційною базою прогнозування слугують одномірні динамічні ряди.

При багатофакторному прогнозуванні процес розглядається як функція певної множини факторів, вплив яких аналізується одночасно або з деяким запізненням. Інформаційною базою виступає система взаємозв'язаних динамічних рядів. Оскільки фактори включаються в модель у явному вигляді, то особливого значення набуває апріорний, теоретичний аналіз структури взаємозв'язків.

Важливим етапом статистичного прогнозування є *верифікація прогнозів,* тобто оцінювання їх точності та обґрунтованості. На етапі верифікації використовують сукупність критеріїв, способів і процедур, які дають можливість оцінити якість прогнозу.

Найбільш поширене *ретроспективне оцінювання прогнозу,* тобто оцінювання прогнозу для минулого часу (ex-post прогноз). Процедура перевірки така. Динамічний ряд поділяється на дві частини: перша — для $t = 1, 2, 3, \ldots, p$ — називається ретроспекцією (передісторією), друга — для $t = p + 1, p + 2, p + 3, \ldots, p + (n - p)$ — прогнозним періодом.

За даними ретроспекції моделюється закономірність динаміки і на основі моделі розраховується прогноз Y_{p+v}, де V — період упередження. Ретроспекція послідовно змінюється, відповідно змінюється прогнозний період, що унаочнює рис. 1.1 (для $v = 1$).

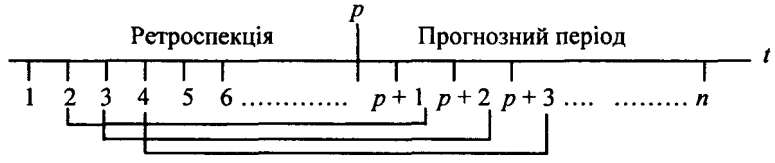


Рис. 1. 1. Схема ретроспективної перевірки точності прогнозу для V =1

Оскільки фактичні значення прогнозного періоду відомі, то можна визначити похибку прогнозу як різницю фактичного y_t і прогнозного Y_t рівнів: $e_t = y_t - Y_t$. Всього буде $n - p$ похибок. Узагальнюючою оцінкою точності прогнозу слугує *середня похибка:*

$$\textit{абсолютна } \bar{e} = \frac{\sum |e_t|}{n - p},$$

$$\textit{квадратична } s = \sqrt{\frac{\sum e_t^2}{n - p}}.$$

Для порівняння точності прогнозів, визначених за різними моделями, використовують *похибку апроксимації* (%):

$$\hat{A} = 100 \frac{\sum \left| \frac{e_t}{y_t} \right|}{n - p}$$

Якщо результат оцінювання точності прогнозу задовольняє визначені критерії точності, скажімо, 10 %, то прогнозна модель вважається прийнятною і рекомендується для практичного використання. Очевидно, що похибка прогнозу залежить від довжини ретроспекції та горизонту прогнозування. Оптимальним співвідношенням між ними вважається 3: 1.

При оцінюванні та порівнянні точності прогнозів використовують також коефіцієнт розбіжності Г. Тейла, який дорівнює нулю за відсутності похибок прогнозу і не має верхньої межі:

$$V = \frac{\sqrt{\sum (y_t - Y_t)^2}}{\sqrt{\sum y_t^2}}.$$

Існуючі методи верифікації прогнозів у більшості своїй ґрунтуються на статистичних процедурах, які зводяться до побудови довірчих меж прогнозу, себто до побудови інтервальних прогнозів.

При прогнозуванні процесів, розвиток яких повністю або частково не піддається формалізації (наприклад, розвиток науки і техніки, соціально-економічні та політичні наслідки прийняття певних управлінських рішень), використовують методи експертних оцінок. Вони ґрунтуються на мобілізації професійного досвіду та інтуїції експертів, які добираються за принципом компетентності.

1. 3. МЕТОД ЕКСПЕРТНИХ ОЦІНОК

Характерною особливістю моделювання та прогнозування соціально-економічних процесів є багатоваріантність, тобто можливість використання різних методів, моделей, інформаційного забезпечення, критеріїв оцінювання адекватності моделі тощо. Вибір між конкуруючими варіантами базується на певній системі правил, що забезпечують надання обґрунтованих оцінок кожному варіанту.

Уважається, що експерт (лат. expertus — досвідчений) володіє цією системою правил і може порівняти варіанти, приписуючи кожному з них числа. Найчастіше перевага чи відносна значущість варіантів встановлюється за допомогою методів ранжування, попарних порівнянь або безпосереднього оцінювання.

При *ранжуванні* експерт повинен розмістити варіанти (фактори, моделі, об'єкти тощо) у порядку, який вважає раціональним, і приписати кожному з них числа натурального ряду — ранги 1, 2, ..., n. Кількість рангів дорівнює кількості варіантів. Якщо експерт надає двом і більше варіантам однакові ранги, то кожному з цих варіантів приписується середній ранг, обчислений з відповідних чисел натурального ряду.

При обґрунтуванні складних управлінських рішень в умовах невизначеності, при довгостроковому прогнозуванні розвитку науки, техніки, економіки використовують групові експертизи. Надійність групових оцінок залежить від узгодженості думок експертів, що потребує відповідної статистичної обробки інформації.

При *груповій експертизі* (n експертів) для кожного i-го варіанта визначається сума рангів ΣR_i, за якою упорядковуються варіанти. Скажімо, перший — найвищий — ранг надається варіанту, який набирає найменшу суму рангів, а останній — варіанту з найбільшою сумою рангів. Результати опитування експертів оформляються у вигляді матриці.

Наприклад, за даними ранжування трьох варіантів п'ятьма експертами (табл. 1.1), перший ранг надається варіанту А, для якого $\Sigma R_i = 6$, другий — варіанту В, третій — варіанту С. Слід зазначити, що ранги визначають лише місця варіантів поміж іншими, не враховуючи існуючих між ними відстаней.

Таблиця 1.1

Варіант	Експерт					Сума рангів	d	d^2
	1	2	3	4	5			
A	2	1	1	1	1	6	–4	16
B	1	2	3	2	2	10	0	0
C	3	3	2	3	3	14	4	16
Разом	X	X	X	X	X	30	X	32

Статистична обробка результатів ранжування передбачає оцінювання ступеня узгодженості думок експертів. Мірою узгодженості слугує *коефіцієнт конкордації* W, в основу розрахунку якого покладено відхилення d сум рангів за окремими варіантами ΣR_i від середньої суми рангів, яка становить $\frac{1}{2} n (m + 1)$. Коефіцієнт конкордації — це відношення суми квадратів названих відхилень $S = \Sigma d^2$ до максимально можливої суми квадратів відхилень $S_{max} = n^2 (m^3 - m) / 12$. Якщо ранги не повторюються, то

$$W = \frac{12S}{n^2\left(m^3 - m\right)},$$

де m — кількість варіантів;

n — кількість експертів.

При неузгодженості думок експертів $W = 0$. Чим вищий ступінь узгодженості, тим більше значення W наближається до 1. За даними табл. 1.1, середня сума рангів становить 30 : 3 = 10, сума квадратів відхилень $S = 32$, а коефіцієнт конкордації

$$W = (12 \cdot 32) / 5^2 (3^3 - 3) = 0{,}64,$$

що свідчить про певні розбіжності в оцінках експертів щодо значущості варіантів.

Перевірка істотності коефіцієнта конкордації W здійснюється за допомогою критерію χ^2 з $(m - 1)$ *числом ступенів вільності (свободи)*. Статистична характеристика критерію розраховується за формулою $\chi^2 = Wn(m - 1)$. Для наведеного прикладу $\chi^2 = 0{,}64 \times \times 5(3 - 1) = 6{,}4$, що перевищує критичне значення $\chi^2(2) = 5{,}99$ (див. додаток 2). Це дає підстави стверджувати з імовірністю 0,95, що значення $W = 0{,}64$ не випадкове і думки експертів узгоджені.

При *попарних порівняннях* експерти використовують дві оцінки: 0 або 1. Більш вагомому варіанту надається оцінка 1, менш вагомому — 0. Результати попарних порівнянь оформляються у вигляді матриці, елементами якої є кількості наданих переваг a_{ij}.

Діагональні елементи такої матриці представлені нулями. Одна із властивостей матриці $a_{ij} + a_{ji} = n$, де n — кількість експертів. За результатами опитування (табл. 1.1) матриця кількості переваг має такий вигляд (табл. 1.2):

Таблиця 1.2

Варіант	A	B	C	Разом	ω,
A	0	4	5	9	0,60
B	1	0	4	5	0,33
C	0	1	0	1	0,07
Разом	1	5	9	15	1,00

Відношення кількості наданих відповідному варіанту переваг до загальної суми елементів матриці характеризує його *вагомість*. За даними табл. 1.2, найвагомішим виявився варіант A, для якого $\omega = 9 : 15 = 0{,}60$.

Часто завданням експерта є не ранжування варіантів, а безпосереднє оцінювання рівнів певного явища чи окремих його властивостей, скажімо, якості продукції, конкурентоспроможності фірм тощо. У таких ситуаціях спершу визначається шкала (діапазон) оцінок, у межах якої експерт і оцінює явище (властивість) певним балом z_{ij}, де i — властивість, j — елемент сукупності.

Для певної множини m властивостей одного явища визначається середній бал $G_j = \Sigma z_{ij} / m$.

На таких методичних засадах ґрунтується більшість рейтингових систем. Так, всесвітньо відома рейтингова система CAMEL, якою користуються органи нагляду за банківською діяльністю, має п'ятибальну шкалу оцінок: від 1 (добре) до 5 (незадовільно). Для кожного банку оцінюється достатність капіталу, якість активів, ефективність менеджменту, прибутковість і ліквідність балансу. Середній бал G_j є рейтингом фінансового стану j-го банку. Від його значення залежить ступінь втручання органів банківського нагляду і комплекс заходів щодо усунення недоліків.

Якщо властивості z_i не рівновагомі, то рейтинг визначається як середня арифметична зважена $G_j = \Sigma z_{ij} \omega_i$, де ω_i — вага i-ї властивості. Саме так оцінюються комерційні, політичні ризики тощо. Наприклад, комерційний ризик, пов'язаний з інтернаціоналізацією банківської діяльності, оцінюється індексом Бері. Ознакова множина цього індексу включає 15 різновагомих показників, які характеризують політичну та економічну ситуацію в країні-партнерові. Зокрема, політична стабільність (вага 12 %), стан платіжного балансу (вага 6 %), темп економічного розвитку (вага 10 %), інші. Сума ваг становить 100 %.

Одним з популярних методів формування групової експертизи є метод Дельфи, назва якого походить від дельфійських мудреців, які славилися в давнину передбаченнями майбутнього. Основні принципи методу Дельфи: анонімність, регульованість зворотного зв'язку та узгодженість групової оцінки.

Автономне опитування експертів проводиться, як правило, в чотири тури. Кожного разу експерт виражає свою думку певною оцінкою в межах визначеної шкали. Результати опитування групи експертів упорядковуються; на основі упорядкованого ряду визначається медіана *Me* й квартилі оцінок — нижній Q_1 і верхній Q_3. Медіана розглядається як узагальнююча групова оцінка процесу; для характеристики варіації оцінок використовують інтерквартильний розмах $R = Q_3 - Q_1$.

Значення медіани і розмаху повідомляють усім експертам. Тим з них, чиї оцінки виявилися за межами діапазону $(Q_3 - Q_1)$, пропонують аргументувати свої висновки, аби ознайомити з ними решту експертів. Такий зворотний зв'язок відсікає «шуми», зменшує вплив індивідуальних і групових інтересів, не пов'язаних з проблемою.

Ітераційна процедура упорядкування та узагальнення експертних оцінок дає можливість зблизити точки зору експертів, що робить групові оцінки надійнішими за просте усереднення. Проте сама по собі процедура опитування не розв'язує всіх проблем точності прогнозів. Вирішальну роль відіграють компетентність експертів і досконалість програми опитування.

1.4. КОМП'ЮТЕРНІ ТЕХНОЛОГІЇ СТАТИСТИЧНОГО МОДЕЛЮВАННЯ

Програмне забезпечення статистичних досліджень досить розвинуте. Всесвітньо відомі статистичні пакети для комплексної обробки даних: *BMDP*, *SPSS*, *SAS*, *Statgraphics*. З 1995 р. світовим лідером на ринку статистичного програмного забезпечення визнається інтегрована система *Statistica* для *Windows* (версія 5.0). Багатофункціональна, графічно орієнтована на обробку масових даних система *Statistica* відповідає основним стандартам *Windows*. Передусім це стандарти користувацького інтерфейсу — MDI, використання буфера обміну, механізму динамічного зв'язку (DDE)

з іншими додатками; система підтримує всі операції, реалізовані за допомогою методу *Drag-and-Drop* — Перетягти та опустити, включаючи автозаповнення, інші.

Складніші процедури обробки даних у системі *Statistica* виконує спеціалізований модуль *Data Management* — Управління даними, а для обробки великих масивів даних або даних з довгими текстовими значеннями застосовують процедури *Megafile Manager Data* — Менеджера мегафайлів.

Система *Statistica* працює з чотирма типами документів. Це:

- електронна таблиця *Spreadsheet*, призначена для введення і перетворення первинних даних;
- електронна таблиця *Scrollsheet* — для виведення результатів аналізу;
- *графік* — для візуалізації результатів обробки та аналізу даних;
- *звіт* — файл у формі RTF (розширений текстовий формат), в якому зберігається текстова, числова і графічна інформація.

Усі статистичні процедури системи розбито на окремі модулі, кожен з яких об'єднує групу логічно зв'язаних між собою статистичних методів і в рамках конкретної моделі забезпечує повний і всебічний аналіз закономірностей. Наприклад, у модулі *Basic Statistics / Tables* — Основні статистики і таблиці пропонується широкий вибір методів розвідувального статистичного аналізу: характеристики варіації і форми розподілу, групування та класифікації, таблиці дисперсійного аналізу *Anova*, всі види коефіцієнтів щільності зв'язку, критерії для тестування нормальності розподілу, істотності зв'язку тощо.

Модуль *Multiple Regression* — Множинна регресія включає вичерпний набір засобів множинної лінійної і нелінійної регресії, багатофакторного прогнозування, аналіз залишків і викидів, тестування гіпотез регресійного аналізу.

Модуль *Time Series / Forecasting* — Часові ряди і прогнозування об'єднує процедури аналізу закономірностей динаміки — тенденцій розвитку і коливань. Модуль пропонує різні методи згладжування рядів, описування трендів, сезонної декомпозиції, авторегресійного аналізу, прогнозної екстраполяції.

Система *Statistica* включає модуль *Anova / Manova* — Дисперсійний аналіз, увесь арсенал методів багатовимірного аналізу (кластерний, дискримінантний, факторний аналіз, факторне шкалювання, канонічні кореляції).

Особливе місце посідає модуль *Sepath* — Моделювання взаємозв'язків системами структурних рівнянь. Зазначені модулі по-

кривають практично весь спектр сучасних методів статистичного моделювання.

Запуск модуля здійснюється через перемикач модулів — *Module Switcher*. У кожному модулі робота починається із Стартової панелі, де відкривається файл первинних даних, вибирається процедура обробки даних і визначаються відповідні їй параметри.

Стартова панель — основне діалогове вікно модуля. Структуру діалогу в усіх модулях уніфіковано, її можна подати схематично (рис. 1.2).

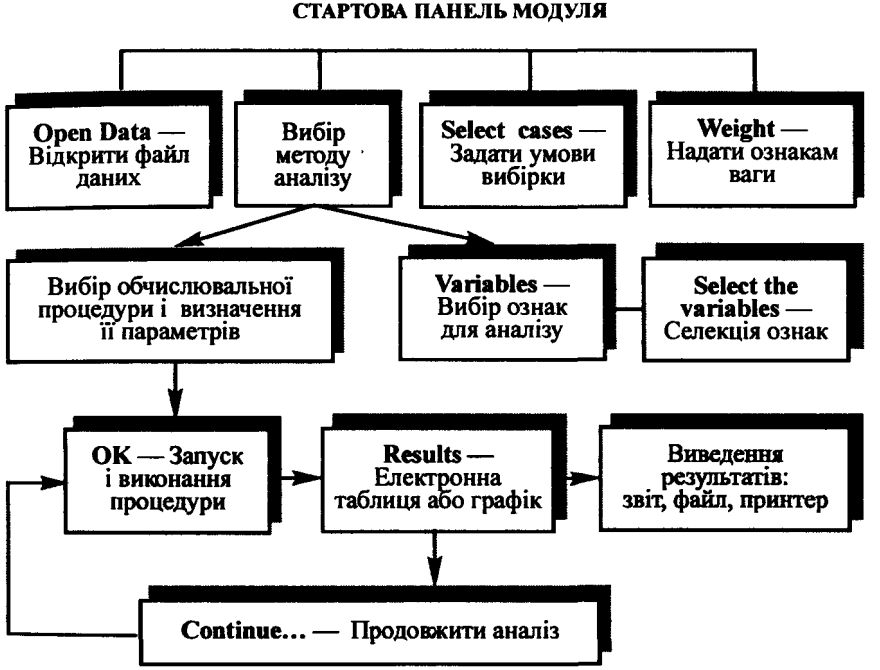


Рис. 1.2. Структура діалогу в модулі

У системі *Statistica* реалізовано принцип постійного логічного підказування. Якщо користувач не може визначитися щодо наступного кроку діалогу, через команду *Enter* система сама спрямує до відповідного діалогового вікна. Якщо виникають складнощі з вибором параметрів обчислювальної процедури, вони задаються системою «за умовчуванням».

Важливою характеристикою системи є наявність засобів всебічної графічної підтримки процесу обробки даних і візуалізації результатів аналізу. Графічні можливості й засоби системи уні-

кальні. Вона включає сотні різних типів користувацьких і спеціальних статистичних графіків, доступних у будь-якому модулі й на будь-якому етапі статистичної обробки даних. Інструменти компонування складної графічної інформації з текстовою і числовою інформацією розглядаються у кожному модулі.

Використання сучасних комп'ютерних технологій обробки даних, інтерактивний спосіб взаємодії з системою перетворюють статистичне моделювання та прогнозування в захоплююче дослідження закономірностей навколишнього світу.

Завдання для самоконтролю

1. З метою ретроспективного оцінювання точності прогнозу грошової маси М1 (млн. грн.) ряд динаміки поділено на: період ретроспекції (9 кварталів) і прогнозний період (3 квартали). Складено два варіанти прогнозу. Використовуючи стандартну похибку, оцініть точність прогнозу за кожним варіантом, зробіть висновок.

Прогнозний період	Прогнозний рівень за варіантом		Фактичний рівень
	1	2	
p+1	7162	7130	7158
p+2	7257	7252	7240
p+3	7352	7374	7365

2. За даними поквартальної динаміки грошової маси М3 (млрд. грн.) і грошового мультиплікатора складено прогнози на період упередження v =1, 2, 3. Використовуючи похибку апроксимації, порівняйте точність прогнозів, зробіть висновок.

Прогнозний період	Прогнозний рівень		Фактичний рівень	
	Грошова маса М3	Грошовий мультиплікатор	Грошова маса М3	Грошовий мультиплікатор
p+1	15,7	1,82	15,4	1,80
p+2	16,5	1,86	16,9	1,77
p+3	17,3	1,89	17,7	1,82

3. Групи респондентів здійснили ранжування дестабілізуючих факторів економіки:

Фактор	Ранги, надані		
	промисловцями	аграріями	гуманітаріями
Податки	1	1	1
Тіньова економіка	4	2	3
Законодавство	3	4	2
Державний борг	2	3	4
Вимоги Світового банку	5	5	5

Оцініть ступінь узгодженості думок респондентів, висновок зробіть з імовірністю 0,95.

4. Цільова установка проекту — максимальний прибуток. Можливі три стратегії досягнення цілі. За даними матриці переваг, наданих експертами кожній стратегії, ранжуйте їх за вагомістю. Зробіть висновки.

Стратегія	A	B	C
A	—	3	4
B	2	—	3
C	1	2	—

5. За наведеними даними оцініть ризик підприємця, який планує вийти на ринок з новим товаром. Фактори ризику оцінювались експертами в діапазоні від 0 до 10 балів, вагові коефіцієнти — від 0 до 100 %. Межа мінімального ризику — 2,5 бала.

Фактор ризику	Бал	Вага, %
Ємність ринку	2	20
Сталість попиту	5	20
Конкурентоспроможність товару	2	25
Фінансовий стан і кредитоспроможність	4	16
Якість роботи маркетингової служби	3	12
Імідж фірми	2	7

Розділ 2

ОПИСУВАННЯ ОБ'ЄКТА МОДЕЛЮВАННЯ

2.1. ФОРМУВАННЯ ІНФОРМАЦІЙНОЇ БАЗИ МОДЕЛІ

Масив первинних даних у системі *Statistica* організується та зберігається у вигляді електронної таблиці *Spreadsheet* з рядками *Cases* і стовпцями *Variables*. Тобто в рядках електронної таблиці розміщуються елементи статистичної сукупності ($j = 1, 2, \ldots, n$), по стовпцях — ознаки ($i = 1, 2, \ldots, m$). Клітинки таблиці призначені для числової або текстової інформації. В системі реалізовано так званий механізм «подвійного запису», за яким встановлюється еквівалент: число = текстове значення. Скажімо, для ознаки «стать»: 1 = чол., 2 = жін. У процесі опрацювання інформації можна переключатися з одного типу даних на інший.

Уведення первинних даних у таблицю здійснюється різними способами, а саме:

- безпосередньо з клавіатури;
- перетворенням існуючого масиву даних за допомогою певних операцій (ранжування, стандартизації), математичних або статистичних функцій (ln, sin тощо);
- імпортуванням даних з інших *Windows* додатків, наприклад з *Excel*.

При формуванні нового файла через команду *New Data* автоматично створюється таблиця розміром 10 · 10: по стовпцях — 10 ознак (VAR1, VAR2 ...) і по рядках — 10 спостережень. Залежно від обсягу наявної інформації стандартну структуру електронної таблиці можна змінити, додаючи або вилучаючи певні ознаки чи спостереження.

Скажімо, для конкретної моделі масив первинних даних містить 15 спостережень і чотири ознаки. Отже, необхідно із стандартної електронної таблиці вилучити шість ознак і додати п'ять спостережень. За допомогою кнопок *Vars* і *Cases* на панелі інструментів вибираються відповідні команди:

1) *Delete Variables* — Вилучити ознаки в діапазоні VAR5 — VAR10;

2) *Add Cases* — Додати спостереження (вказати їх кількість — 5 і номер спостереження, за яким їх треба розмістити, — 10).

Необхідним етапом формування інформаційної бази є специфікація ознак, тобто визначення основних параметрів кожної з них: імені, формату, коду для пропущених даних, формули чи зв'язку DDE.

Найпростіший спосіб надання специфікації ознакам — через команду *Current Specs* — Поточні специфікації (кнопка VARS). Основні з них:

Name — ім'я, не більше восьми символів. У противному разі необхідно використати мітку *Label* у полі *Long Name*. Скажімо, *Name*: ВВП; *Label*: Валовий внутрішній продукт 1999 р. у поточних цінах, млрд. грн.;

MD code — код, який приписується пропущеним даним. Можна вилучити дані з розрахунків, замінити їх середніми значеннями або інтерполювати. За умовчування *MD code* становить – 9999.

Decimals — кількість розрядів після коми.

Display Format — спосіб представлення ознак різних типів: числа, дати, час, науковий формат, грошовий формат, проценти.

Якщо значення *i*-ї ознаки необхідно розрахувати, то в текстовому полі *Long Name* задається формула, за якою ведеться розрахунок. Формула починається із знака «=». При потребі у формулі використовуються математичні чи статистичні функції — *Functions*. У цьому ж текстовому полі можна встановити зв'язок (link) з іншими *Windows* додатками за допомогою механізму DDE. Наприклад, *Link*: @ Excel | c:\ file.xls!r2c2:r4c4.

Еквівалент між числовим і текстовим значеннями ознак встановлюється через команду *Text Values*. За допомогою команди *All Specs* можна переглянути й відредагувати специфікації усіх ознак.

Інші параметри електронної таблиці встановлюються за допомогою функціональних кнопок на панелі інструментів. Так, імена спостережень (рядків) — через кнопку *Cases*, а заголовок таблиці — через кнопку *Workbook* — Робочі книги. Після того, як усі параметри таблиці задано, можна вводити дані.

Процедуру створення файла первинних даних розглянемо на прикладі 15 цукрових заводів, які характеризуються чотирма ознаками. Специфікацію ознак наведено в табл. 2.1.

Таблиця 2.1

Variables	Name	Long Name
VAR1	Якість	Цукристість буряка, %
VAR2	Втрати	Втрати сировини при транспортуванні та зберіганні, %
VAR3	Патока	Вміст цукру в патоці, %
VAR4	Ефект	Вихід цукру з 1 т цукрового буряка, %

Файл первинних даних по сукупності цукрових заводів RM.sta 4v*15c наведено в табл. 2.2.

Таблиця 2.2

	VAR1 Якість	VAR2 Втрати	VAR3 Патока	VAR4 Ефект
1	15,1	0,99	2,5	9,78
2	15,41	1,06	2,68	9,13
3	15,22	0,98	2,19	10,46
4	15,16	0,95	2,06	10,69
5	15,43	1	2,05	10,58
6	15,41	1	2,06	10,84
7	15,15	0,97	2,34	10,87
8	16,06	0,9	2,24	12,24
9	15,95	0,92	2,27	11,94
10	15,59	0,95	2,13	11,26
11	15,52	0,93	2,26	11,01
12	15,33	0,97	2	11,88
13	15,48	0,91	2,2	11,53
14	15,18	0,98	2,23	11,03
15	15,17	0,98	2,18	10,37

Спеціалізований модуль *Data Management* має значно ширші можливості доступу до команд з формування структури таблиць і специфікації первинних даних. У діалоговому вікні *Greate New Data File* можна одразу задати кількість спостережень і кількість ознак, їх імена, формат і заголовок файла. У файл первинних даних спеціального формату можна конвертувати кореляційну матрицю (див. 8.1).

Для роботи з даними в системі *Statistica* реалізовано всі стандартні операції методу *Drag-and-Drop*, зокрема:

- копіювання, переміщення, вставка;
- автозаповнення блоків з регулярною структурою;
- стандартизація даних;
- транспонування;
- зсування значень ознаки на певний лаг;
- ранжування даних;
- перекодування значень ознаки, перехід від однієї шкали вимірювання до іншої.

Іноді необхідно проаналізувати не всю сукупність, а певну її частину. В системі *Statistica* реалізовано процедуру вибору підмножини спостережень. Умови вибору визначаються стандартним діалогом *Case Selection Conditions* із меню *Options, Select*. Якщо підмножини формуються для подальшого аналізу, вибирається опція *Include if*, якщо ж підмножина виключається із подальшого аналізу — опція *Exclude if*. Так, за умови *Include if*: V4 ≤ 11 будуть аналізуватися лише ті спостереження, в яких значення V4 не перевищує 11. Можна задати складніші умови вибору, зокрема за допомогою логічних операторів.

2.2. РОЗВІДУВАЛЬНИЙ АНАЛІЗ ДАНИХ

Статистична обробка даних у будь-якому модулі генерує велику кількість вихідної інформації у вигляді електронних таблиць *Scrollsheet* і графіків. Таблиці *Scrollsheet* підтримують усі стандартні операції з виділеними блоками значень (копіювання, переміщення, вставка, екстраполяція, стандартизація даних тощо). Характерною їх особливістю є наявність у лівому верхньому куті кнопки *Continue...*, призначеної для продовження аналізу. Це може бути чергова таблиця або повернення в діалогове вікно тієї статистичної процедури, за якою ведеться аналіз. Таблицю *Scrollsheet* можна відредагувати, зберегти як файл результатів (з розширенням *scr*) або конвертувати в первинні дані (файл з розширенням *sta*), можна експортувати в інші *Windows* додатки.

Як приклад розглянемо порядок створення таблиці *Scrollsheet* у модулі *Basic Statistics / Tables* — Основні статистики і таблиці, який об'єднує методи розвідувального аналізу даних. На стартовій панелі модуля відкриваємо файл первинних даних, наприклад RM.sta (по сукупності цукрових заводів). На першому етапі аналізу даних використаємо процедури *Descriptive Statistics* — Описові статистики. Вибір ознак для аналізу здійснюється у вікні *Select the variables for the analysis.* Вибираємо одразу всі чотири ознаки: VAR1 — VAR4. З метою всебічного аналізу розподілу сукупності за цими ознаками скористаємося опцією *More Statistics* — Розширений набір описових статистик. З-поміж них виберемо: *Mean* — середню величину, *Median* — медіану, *Lower and Upper Quartiles* — нижній і верхній квартилі, *Standard Deviation* — стандартне відхилення, *Skewness* — коефіцієнт асиметрії та *Kurtosis* — коефіцієнт ексцесу.

Після команди на виконання процедури аналізу система створює електронну таблицю *Scrollsheet* з результатами розрахунку (табл. 2.3).

Таблиця 2.3

Descriptive Statistics (RM.sta)							
Continue...	Mean	Median	Lower Quartile	Upper Quartile	Std. Dev.	Skewness	Kurtosis
VAR1	15,37	15,33	15,16	15,52	0,331	0,603	0,755
VAR2	0,97	0,97	0,93	0,99	0,041	0,384	0,817
VAR3	2,23	2,2	2,06	2,27	0,179	1,275	1,970
VAR4	10,91	10,87	10,46	11,53	0,820	–0,398	0,391

Як показують дані, розподіл сукупності цукрових заводів характеризується невисокою варіацією ознак (відношення стандартного відхилення до середньої не перевищує 10 %) і помітною асиметрією, особливо за рівнем втрат цукру при переробці сировини (VAR3); для виходу цукру з 1 т сировини характерна лівостороння асиметрія.

Продовжуючи розвідувальний аналіз даних, з метою оцінювання взаємозв'язку між ознаками, виберемо на стартовій панелі модуля процедуру *Correlation matrices* — Кореляційні матриці. У вікні *Pearson Product-Moment Correlation* вибираємо тип матриці — *One variable list (square matrix)* — Один список ознак (квадратна матриця), а у вікні вибору ознак натиснемо кнопку *Select all* — Усі ознаки. За командою на виконання процедури створюється нова таблиця, елементами якої є парні коефіцієнти кореляції (табл. 2.4).

Таблиця 2.4

Variable	VAR1	VAR2	VAR3	VAR4
VAR1	1	– 0,265	0,238	0,377
VAR2	– 0,265	1	0,492	– 0,726
VAR3	0,238	0,492	1	– 0,573
VAR4	0,378	– 0,726	– 0,573	1

Для візуалізації взаємозв'язків необхідно вибрати тип графіка і вказати ознаки. Наприклад, зв'язок між цукристістю буряка (VAR1) і виходом цукру з 1 т сировини (VAR 4) представимо у вигляді двовимірної діаграми розсіювання. Команда *Continue...* повертає до вікна процедури *Correlation matrices*. Вибираємо опцію *2D scatterplot*, і відповідний графік кореляційного поля з'являється на екрані (рис. 2.1).

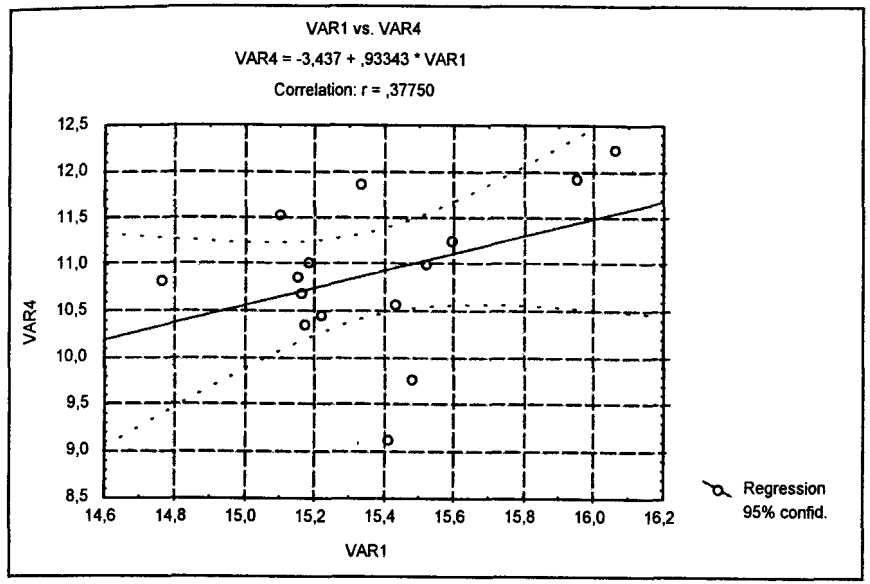


Рис. 2.1. Кореляційне поле

Отже, за результатами розвідувального аналізу даних на екрані монітора послідовно виведено три документи:

- таблиця із статистичними характеристиками розподілу сукупності;
- кореляційна матриця;
- діаграма розсіювання (кореляційне поле).

Ці документи мають стандартні заголовки і для змістовної інтерпретації результатів аналізу потребують відповідного редагування. Специфікація даних таблиці *Scrollsheet* задається через команду контекстового меню: для стовпців — *Column Specs*, для рядків — *Row Name*, назва таблиці — командою *Titles*. Графіки редагуються за опціями редактора *Graphic Text Editor*.

Відредаговану числову, текстову та графічну інформацію можна експортувати в інші *Windows* додатки, на принтер, у текстовий файл або у спеціальний файл, який називається звітом. Параметри виведення інформації вказуються в діалоговому вікні *Page / Output Setup* — Параметри сторінки / Виведення. Для числової і текстової інформації вибирається команда *Text / Scrollsheets / Spreadsheets*, для графічної — *Graphs* (з подальшим уточненням формату та інших параметрів графіка). В рамці *Output Header* можна вказати заголовок, дату, час тощо.

Звіт — це один з документів системи *Statistica*, який створюється за опцією *Windows* (рамка *Output*) у розширеному текстовому форматі (*RTF*). Файл *RTF* можна редагувати безпосередньо в *Statistica* або у будь-якому іншому текстовому процесорі, наприклад у *MS Word*. Система має блок опцій автоматичної підготовки звіту — *Auto-report*.

2.3. БАГАТОВИМІРНЕ РАНЖУВАННЯ

При описуванні об'єкта моделювання, окрім узагальнюючих характеристик по сукупності в цілому, важливо упорядкувати одиниці сукупності за певними властивостями (якостями, цінностями), визначити належність кожної з них до певного типу. Якщо властивість характеризується однією ознакою, то упорядкування одиниць сукупності здійснюється заміною значень цієї ознаки відповідними рангами. У системі *Statistica* це процедура *Rank Variables* меню Vars. У діалоговому вікні *Rank Order Values* вибираються ознаки, схема упорядкування (за зростанням чи зменшенням значень), умови обробки зв'язаних рангів, тип рангу: регулярний (від 1 до n) чи фракційний (від 0 до 1).

Оскільки властивості соціально-економічних явищ характеризуються, як правило, множиною ознак ($m \geq 2$), то при упорядкуванні одиниць сукупності виникає необхідність агрегування усіх ознак множини x_i в одну *інтегральну оцінку* G_j. Така оцінка геометрично інтерпретується як точка у багатовимірному просторі, координати якої вказують на масштаб або позицію j-ї одиниці. Алгебраїчно значення ознак для j-ї одиниці сукупності представляються вектором $x_j = | x_1, x_2, \ldots, x_m |$, а агрегування їх означає перетворення вектора в скаляр.

Агрегування ознак ґрунтується на так званій теорії «адитивної цінності», згідно з якою цінність цілого дорівнює сумі цінностей його складових. Такий підхід реалізовано при визначенні рейтингів на основі експертних оцінок, представлених рангами або балами (див. 1.3). Якщо ознаки множини X мають різні одиниці вимірювання, то адитивне агрегування потребує приведення їх до однієї основи, тобто попередньої *стандартизації*. Вектор первинних значень ознак $x_j = | x_1, x_2, \ldots, x_m |$ замінюється вектором стандартизованих значень $z_j = | z_1, z_2, \ldots, z_m |$.

Найчастіше інтегральна оцінка G_j визначається як середня арифметична стандартизованих значень ознак z_{ij}. Для j-ї одиниці сукупності

$$G_j = \frac{1}{m}\sum_{1}^{m} z_{ij} .$$

Якщо ознаки множини різновагомі, то кожній з них надається певна вага ω_i, тобто інтегральна оцінка має форму середньої арифметичної зваженої:

$$G_j = \sum_{1}^{m} z_{ij}\omega_i , \text{ де } \Sigma\, \omega_i = 1.$$

Конструювання інтегральної оцінки передбачає чотири етапи:

- формування ознакової множини;
- вибір способу стандартизації показників;
- обґрунтування функції вагових коефіцієнтів;
- визначення процедури агрегування показників.

На етапі формування ознакової множини X вирішальну роль відіграє апріорний якісний аналіз суті явища. Так, для характеристики демографічної ситуації використовують такі показники, як: очікувана тривалість життя, сумарний коефіцієнт плідності, коефіцієнт дитячої смертності, демографічне навантаження працездатного населення, валовий міграційний рух тощо. Незважаючи на значущість показників середньодушового доходу чи забезпеченості населення житлом, до ознакової множини демографічної ситуації вони не включаються, оскільки за своєю суттю є характеристиками життєвого рівня населення. Щодо вагових коефіцієнтів, то вибір їх також ґрунтується на теоретичному аналізі суті явища і в кожному конкретному дослідженні ω_i визначається експертно-статистичним методом (див. 1.3).

При формуванні ознакового простору важливо забезпечити інформаційну односпрямованість показників x_i. Демографічна ситуація за інших рівних умов буде тим краща, чим більша тривалість життя і менша дитяча смертність. Тобто тривалість життя і дитяча смертність інформаційно різноспрямовані, і це необхідно враховувати при агрегуванні їх в одну оцінку. З метою забезпечення інформаційної односпрямованості показників їх поділяють на *стимулятори* та *дестимулятори*. Зв'язок між оцінкою G і показником-стимулятором x_{st} прямий, між оцінкою G і показником-дестимулятором x_{dst} — обернений. При агрегуванні дестимулятори перетворюються на стимулятори, наприклад, $x_{st} = 1 - x_{dst}$ або $x_{st} = 1/x_{dst}$.

На практиці застосовують різні способи стандартизації. Усі вони ґрунтуються на порівнянні емпіричних значень показника x_{ij} з певною величиною a. Такою величиною може бути максимальне x_{max}, мінімальне x_{min}, середнє $\bar{x}$ чи еталонне x_0 значення показника. Результат порівняння можна представити відношенням $\frac{x_{ij}}{a}$ або відхиленням $\frac{x_{ij}-a}{q}$, де q — одиниця стандартизації.

Наприклад, визначимо рейтинги інвестиційної привабливості компаній — постачальників електронного обладнання. Ознаковий простір представляють (%): x_1 — рентабельність виробництва, x_2 — ліквідність активів, x_3 — частка видатків на наукові дослідження. Найпростіший спосіб стандартизації — відношення $\frac{x_{ij}}{a}$, а оскільки всі зазначені показники є стимуляторами, то доцільно взяти $a = x_{min}$, а отже, $z_{ij} = \frac{x_{ij}}{x_{min}}$.

Розраховані для трьох компаній рейтинги (табл. 2.5) показують, що з-поміж них для інвесторів найбільш привабливою є Motorola.

Таблиця 2.5

Компанія	x_{1j}	x_{2j}	x_{3j}	Z_{1j}	Z_{2j}	Z_{3j}	G_j
Motorola	21	31	34	1,3125	1,1481	1,2143	1,225
Nec	16	32	28	1	1,1852	1	1,0617
Hitachi	19	27	32	1,1875	1	1,1428	1,1101

Якщо існують стандарти, нормативи чи будь-які інші еталонні значення ознак x_{i0}, то, агрегуючи відношення $\frac{x_{ij}}{x_{i0}}$, можна оцінити ступінь відхилення від «еталона». Аби значення інтегральної оцінки G_j змінювалося в інтервалі від 0 до 1, розрахунок ведеться за формулою, в якій агрегуються і додатні, і від'ємні відхилення:

$$G_j = \frac{1}{m}\sum_{1}^{m}\left|\frac{x_{ij}}{x_{i0}} - 1\right|,$$

Залежно від конкретної мети дослідження можна агрегувати лише додатні або лише від'ємні відхилення. Іноді усереднюють не модулі, а квадрати відхилень, використовуючи середню квадратичну.

Порівняльний аналіз у межах сукупності, в якій кожний показник має типовий середній рівень, здійснюється на основі агрегування відношень x_{ij} до середнього рівня $\bar{x}$:

$$G_j = \frac{1}{m}\sum_{1}^{m}\frac{x_{ij}}{\bar{x}}.$$

Очевидно, що при $G_j > 1$ рівень розвитку явища в j-ї одиниці вищий за середній по сукупності, а при $G_j < 1$ — нижчий. Таку узагальнюючу оцінку називають *багатовимірною середньою*, за її значеннями здійснюють типологію одиниць сукупності, скажімо, автопідприємств за рівнем ефективності використання парку машин, агропідприємств — за рівнем забезпеченості ресурсами тощо.

Якщо ознаки множини різновагомі, то багатовимірна середня розраховується як арифметична зважена

$$G_j = \sum_{1}^{m}\frac{x_{ij}}{\bar{x}}\omega_i, \text{ де } \omega_i \text{ — вага } i\text{-ї ознаки, } \sum_{1}^{m}\omega_i = 1.$$

Скажімо, в агрогосподарствах регіону на 100 га сільськогосподарських угідь припадає в середньому: 13,5 працездатних, 0,85 ум. трактора і 23,8 корови. Середня оцінка якості ґрунтів — 51 бал. Оцінюючи ресурсний потенціал господарств, показникам якості ґрунтів і забезпеченості тракторами надається вага 0,3, показникам забезпеченості трудовими ресурсами і щільності поголів'я корів — 0,2. Якщо у j-му господарстві значення цих показників відповідно 13,8; 0,8; 29,9 і 62,7, то багатовимірна середня забезпеченості ресурсами становить

$$G_j = \frac{13,8}{13,5}0,2 + \frac{0,8}{0,85}0,3 + \frac{29,9}{23,8}0,2 + \frac{62,7}{51,0}0,3 = 1,106.$$

Тобто ступінь забезпеченості ресурсами вищий за середній по регіону.

Аналогічного змісту інтегральну оцінку можна обчислити і на основі часток

$$d_{ij} = \frac{x_{ij}}{\sum_{1}^{n}x_{ij}},$$

де $\sum_{1}^{n}x_{ij}$ — загальний обсяг значень i-ї ознаки по сукупності в цілому;

x_{ij} — обсяг j-ї складової за цією ознакою. Очевидно, що $\sum_{1}^{n} d_{ij} = 1$, або 100%. Формула інтегральної оцінки має такий вигляд:

$$G_j = \frac{n}{m}\sum_{1}^{m} d_{ij}.$$

Наприклад, на j-й регіон (j = 1, 2, 3) припадає: 25 % експорту продукції чорної металургії, 48 % — продукції машинобудування, 24 % — продукції хімічної промисловості та 32 % — продукції інших галузей промисловості. Рейтинг експортного потенціалу промисловості цього регіону становить

$$G_j = \frac{3}{4}(0{,}25 + 0{,}48 + 0{,}24 + 0{,}32) = 0{,}9675,$$

тобто нижчий за середній рівень по трьох регіонах.

У соціально-економічних дослідженнях широко використовують інтегральні оцінки, розраховані на основі відхилень $(x_{ij} - a)$, стандартизованих варіаційним розмахом $(x_{max} - x_{min})$. При цьому для стимуляторів $a = x_{min}$, для дестимуляторів $a = x_{max}$.

$$z_{ij} = \frac{x_{ij} - x_{min}}{x_{max} - x_{min}};\ z_{ij} = \frac{x_{max} - x_{ij}}{x_{max} - x_{min}}.$$

Тобто z_{ij} показує відносну позицію j-ї одиниці сукупності в діапазоні варіації за i-ю ознакою. При високих значеннях i-ї ознаки z_{ij} наближається до 1, при низьких — до 0. Таку саму властивість має й інтегральна оцінка $G_j = \frac{1}{m}\sum_{1}^{m} z_{ij}$. Чим вищий рівень розвитку властивості, тим далі від нуля відхиляється значення G_j.

Класичним прикладом такого типу інтегральної оцінки є індекс людського розвитку за методикою Програми розвитку ООН. Ознакову множину цього індексу представляють: x_1 — очікувана тривалість життя, x_2 — досягнутий рівень освіти, x_3 — реальний ВВП на душу населення. Одиниця стандартизації — теоретично можливий варіаційний розмах: для тривалості життя (років) — (85 – 25), для рівня освіти (%) (100 – 0), для ВВП на душу населення до 1997 р., (дол. США) — (5120 – 100). Якщо фактичний середньодушовий дохід перевищував 5120 дол. США, то величина перевищення дисконтувалася за певною методикою. Визначимо індекс людського розвитку для країни, де очікувана тривалість життя — 69,4 року, рівень освіти — 87 %, ВВП на душу населення — 5010 дол. США:

$$G_j = \frac{1}{3}\left[\frac{69{,}4 - 25}{85 - 25} + 0{,}87 + \frac{5010 - 100}{5120 - 100}\right] = 0{,}863.$$

Використання теоретично можливого варіаційного розмаху дає змогу провести порівняльний аналіз як у просторі, так і в часі. Якщо аналіз динаміки не передбачається, то за одиницю стандартизації можна взяти фактичний варіаційний розмах.

Відносну позицію j-ї одиниці сукупності у багатовимірному просторі характеризує також таксономічний показник рівня розвитку, розрахунок якого спирається на традиційний спосіб стандартизації відхилень від середньої:

$$\text{для стимуляторів } z_{ij} = \frac{x_{ij} - \bar{x}_i}{\sigma_i},$$

$$\text{для дестимуляторів } z_{ij} = \frac{\bar{x}_i - x_{ij}}{\sigma_i}.$$

Така стандартизація дає змогу позбутися одиниць вимірювання, але водночас відбувається вирівнювання середніх і дисперсій: для кожної ознаки $\bar{z} = 0$, дисперсія $\sigma_z^2 = 1$, а діапазон варіації z_{ij} згідно з правилом «трьох сигм» — від –3 до +3.

При розрахунку інтегральної оцінки використовують стандартний діапазон варіації для всіх ознак на одному і тому самому рівні. Скажімо, на рівні двох стандартних відхилень (від –2 до +2). Відстань між верхньою (+2) і нижньою (–2) точками діапазону у багатовимірному просторі становить $|C| = 2\, z_0 \sqrt{m}$, де z_0 — точка, взята за базу порівняння. Якщо $z_0 = -2$, то для п'яти ознак $|C| = 2(-2)\sqrt{5} = 8{,}94$.

Позиція j-ї одиниці відносно бази порівняння z_0 визначається як Евклідова відстань

$$G_{j0} = \left[\sum_{1}^{m}(z_{ij} - z_0)^2\right]^{1/2},$$

а відношення відстані G_{j0} до стандартного діапазону варіації $|C|$ називають *таксономічним показником рівня розвитку*:

$$G_j = \frac{C_{j0}}{|C|}.$$

Значення його коливаються в межах від 0 до 1. Чим вищий рівень розвитку явища, тим більше значення G_j. Якщо координати

умовного об'єкта визначити на рівні $z_0 = +2$ (по верхній межі діапазону варіації), то таку інтерпретацію має відхилення $(1 - G_j)$.

Визначимо таксономічний показник розвитку країн за такими ознаками: x_1 — ВВП на душу населення, тис. дол. США; x_2 — зовнішній борг, % до ВВП; x_3 — ступінь самозабезпеченості енергоресурсами, %. У табл. 2.6 наведено абсолютні та стандартизовані значення цих ознак. Оскільки x_2 — дестимулятор, то при розрахунку C_{j0} стандартизоване значення z_{2j} помножується на (–1). Наприклад, для першої країни Евклідова відстань становить

$$C_{10} = \left[(0{,}524+2)^2 + (1{,}259+2)^2 + (0{,}481+2)^2\right]^{1/2} = 4{,}811,$$

а таксономічний показник

$$G_j = \frac{4{,}811}{8{,}94} = 0{,}538.$$

Таблиця 2.6

Країна, j	Первинні значення ознак, x_{ij}			Стандартизовані значення ознак, z_{ij}			C_{j0}	G_j
	x_{1j}	x_{2j}	x_{3j}	z_{1j}	z_{2j}	z_{3j}		
1	5,8	14	48	0,524	–1,259	0,481	4,811	0,538
2	4,7	22	55	–0,543	0,140	1,182	3,963	0,443
3	3,9	28	29	–1,319	1,189	–1,422	1,206	0,135

У системі *Statistica* розрахунок інтегральних (багатовимірних) оцінок за будь-якою з розглянутих методик можна здійснити в модулі *Data Management* у такій послідовності:

Формування файла первинних даних обсягом $n \cdot m$ плюс додатковий стовпець *Newvar* для інтегральної оцінки → Стандартизація масиву первинних даних за опціями *Standardize Columns* або *Current Specs* → Розрахунок інтегральної оцінки за алгоритмом, вказаним у *Long Name* ознаки *Newvar*

Основне призначення інтегральних оцінок — ранжування, типологія об'єктів. Проте, як і будь-який інший статистичний показник, G_j має певний соціально-економічний зміст, варіація його значень підпорядкована певним законам розподілу, а отже, правомірним є використання таких оцінок при вивченні закономірностей розподілу, взаємозв'язку й тенденцій розвитку.

Завдання для самоконтролю

1. Для аналізу фінансового стану банків самостійно визначте сукупність банків (не менше 25) і по кожному з них випишіть інформацію щодо розміру активів, капіталу, зобов'язань і прибутку. По кожному банку визначте: прибутковість активів і капіталу, співвідношення капіталу і зобов'язань. У системі *Statistica* сформуйте файл первинних даних, здійсніть специфікацію показників.

2. Використовуючи процедури модуля *Basic Statistics and Tables*, здійсніть розвідувальний аналіз даних завд. 1: а) для показників прибутковості активів і капіталу визначте всі характеристики розподілу (*Statistics*); б) складіть ряд розподілу банків за рівнем прибутковості активів (*Frequency tables*), здійсніть частотний аналіз розподілу; в) складіть комбінаційне групування банків за рівнем прибутковості активів і прибутковості капіталу (*Tables and banners*); г) для візуалізації розподілу банків за усіма показниками скористайтеся графіками.

3. За процедурою *Correlation matrices* оцініть взаємозв'язки між показниками прибутковості та співвідношенням капіталу і зобов'язань (завд. 1), подайте їх у матричному вигляді та графічно. Зробіть висновки.

4. За наведеними даними визначте рейтинги банків за достатністю капіталу. Класифікуйте показники на стимулятори та дестимулятори.

Показник	Банк			
	А	Б	В	Г
H_1 —відношення зобов'язань до капіталу	16	14	11	15
H_3 — достатність капіталу	12	7	10	13
H_4 — ліквідність балансу	2,24	1,16	0,74	1,23
H_6 — ліквідність активів	0,18	0,21	0,32	0,24

Нормативи показників: H_1 — не більше 8; H_3 — не менше 0,5; H_4 — не більше 0,7; H_6 — не менше 0,5.

5. Визначте рейтинги регіонів за рівнем розвитку інформаційних комунікацій:

Показник на 1000 чол	Регіон			У середньому по країні
	А	Б	В	
Кількість телевізорів	310	340	250	330
Кількість радіоприймачів	220	300	230	240
Кількість телефонів	105	150	100	120

Обґрунтуйте вибір узагальнюючого показника.

6. Географічна структура зовнішньоекономічних зв'язків країни А з іншими країнами характеризується такими даними (%):

Показник	Разом	У тому числі з країною				
		Б	В	Г	Д	Е
Експорт	100	28	12	36	15	9
Імпорт	100	39	17	24	8	12

Оцініть ступінь активності зовнішньоекономічної діяльності країни А з іншими країнами, зробіть висновки.

7. Визначте рейтинги країн за рівнем науково-технічного розвитку. Обґрунтуйте вибір узагальнюючого показника.

Країна	Частка витрат на НДДКР* у ВВП, %	Патенти у країні, тис шт	Обсяг експорту ліцензій, млн грош од
А	2,4	15	720
Б	2,7	28	575
В	2,2	16	426
Г	2,5	32	682

* На науково-дослідну діяльність і конструкторську роботу.

8. Визначте рейтинги країн за рівнем економічного розвитку. Обґрунтуйте вибір узагальнюючої оцінки, класифікуйте показники на стимулятори і дестимулятори.

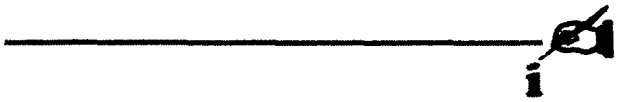
Країна	ВВП на 1 кг енерговитрат	Норма інвестицій, %	Рівень безробіття, %	Державний борг, % до ВВП
А	4,2	25	14	28
Б	1,6	30	10	78
В	0,9	28	15	62
Г	1,4	20	13	56
Д	0,8	32	9	45

9. За допомогою таксономічного показника оцініть екологічну ситуацію в регіонах. У таблиці наведено стандартизовані значення показників (у розрахунку на 1 жителя): x_1 — споживання свіжої води; x_2 — обсяг скидання забруднених стічних вод у природні водоймища; x_3 — викиди шкідливих речовин в атмосферне повітря стаціонарними джерелами забруднення.

Показник	Регіон				
	А	Б	В	Г	Д
z_1	–0,05	0,37	0,51	–1,30	–0,56
z_2	–0,08	–0,13	–0,38	–0,91	0,04
z_3	0,45	0,22	–0,72	–0,12	1,72

Класифікуйте показники на стимулятори і дестимулятори, зробіть висновки.

10. Яку з-поміж розглянутих інтегральних оцінок можна застосувати для аналізу динаміки складних соціально-економічних явищ? Відповідь обґрунтуйте.



Розділ 3

МОДЕЛІ КЛАСИФІКАЦІЇ

3.1. ОДНОРІДНІСТЬ І ТИПОЛОГІЯ

Однією з умов статистичного моделювання є однорідність сукупності. Лише в однорідній сукупності виявлені закономірності є сталими і їх можна застосувати до усіх одиниць сукупності.

Поняття *однорідності* пов'язують з наявністю в усіх одиниць сукупності таких спільних властивостей і рис, які визначають їх одноякісність, належність до одного й того ж типу. Оцінювання ступеня однорідності здійснюється за допомогою критеріїв математичної статистики, більшість з яких орієнтовано на аналіз форми одновершинних розподілів.

Однорідними вважаються сукупності, яким властивий симетричний, нормальний розподіл. Звісно, в соціально-економічних явищах нормальний розподіл у чистому вигляді не зустрічається. Але він близький до інших одновершинних розподілів, його часто використовують як перше наближення при моделюванні. Деякі одновершинні розподіли приводяться до нормального виду перетворенням значень ознак, скажімо, заміною їх логарифмами. Лог-нормальною кривою можна описати низку асиметричних розподілів, передусім з правосторонньою асиметрією.

Основні властивості нормального розподілу:

- крива розподілу симетрична відносно максимальної ординати, яка відповідає значенню середньої арифметичної $\bar{x}$;
- у межах $\bar{x} \pm \sigma$ міститься 68,3 % усіх частот ряду розподілу, в межах $\bar{x} \pm 2\sigma$ — 95,4 % частот, у межах $\bar{x} \pm 3\sigma$ — 99,7 % частот;
- співвідношення стандартного відхилення σ і середнього модуля відхилень $\bar{l}$ становить $\frac{\bar{l}}{\sigma} = \sqrt{\frac{2}{\pi}} = 0{,}8$ або $\frac{\sigma}{\bar{l}} = 1{,}25$. Значення його залежить від наявності в сукупності нетипових, аномальних спостережень і може слугувати індикатором її «засміченості»;
- третій центральний момент розподілу $m_3 = 0$, четвертий $m_4 = 3\,m_2^2$, звідси коефіцієнт асиметрії $a_3 = m_3 / \sigma^3 = 0$ і коефіцієнт ексцесу $a_4 = m_4 / m_2^2 = 3$.

Завдяки цим властивостям нормальна крива застосовується як *стандарт* і відіграє значну роль при використанні методів вибір-

кового, регресійного, факторного аналізу. Оцінка ступеня наближеності до цього стандарту ґрунтується на порівнянні емпіричних f_j і теоретичних $\hat{f}_j$ частот розподілу, де j — номер інтервалу. Теоретичні частоти визначають за формулою:

$$\hat{f}_j = n\left|F_{x_i} - F_{x_{i-1}}\right|,$$

де: $F_x = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{z^2}{2}} dz$ — інтегральна функція нормального розподілу (додаток 1);

$z = \frac{x - \bar{x}}{\sigma}$ — стандартизоване значення ознаки x;

$n = \Sigma f_j$ — обсяг сукупності.

З-поміж критеріїв, що застосовуються для оцінювання узгодженості емпіричного розподілу з нормальним, найпоширенішими є критерії Пірсона χ^2 та Колмогорова-Смирнова d.

Критерій χ^2 ґрунтується на групуваннях. Статистична його характеристика — це сума стандартизованих квадратів відхилень емпіричних частот від теоретичних :

$$\chi^2 = \sum_{j=1}^{m} \frac{(f_j - \hat{f}_j)^2}{\hat{f}_j}.$$

Розрахунок χ^2 можна здійснити також на основі часток розподілу — емпіричних d_j і теоретичних $\hat{d}_j$:

$$\chi^2 = n \sum_{j=1}^{m} \frac{(d_j - \hat{d}_j)^2}{\hat{d}_j}.$$

Фактичне значення критерію χ^2 порівнюється з критичним для ймовірності 1 — α і відповідного числа ступенів вільності $k = m - q - 1$, де m — кількість інтервалів групування, q — кількість параметрів функції (для нормального розподілу $q = 2$). Якщо $\chi^2 < \chi^2_{1-\alpha}(k)$, відхилення емпіричного розподілу від нормального визнається неістотним. У разі, коли фактичне значення критерію перевищує критичне, гіпотеза про узгодженість розподілів відхиляється. Критичні значення критерію χ^2 наведено в додатку 2.

Висновки за критерієм χ^2 значною мірою залежать від кількості груп, частота яких має бути не менша 5. Окрім того, χ^2 не враховує послідовність знаків відхилень частот (часток), за наявності серій знаків надійність висновку зменшується. Цих вад поз-

бавлений критерій Колмогорова-Смирнова *d*. Статистичною х
рактеристикою останнього є максимальне за модулем відхилен
між кумулятивними частками емпіричного F_j і теоретичного $\hat{F}_j$
розподілів:

$$d = \max \mid F_j - \hat{F}_j \mid .$$

При використанні кумулятивних частот розподілів макси
мальний модуль відхилення між ними необхідно розділити на об
сяг сукупності *n*. Критичні значення *d* наведено в додатку 3. Дл
$n > 30$ більш точними вважаються межі критерію, визначені з
відношенням Лілліефорса.

У системі *Statistica* закономірність одномірного розподіл
можна аналізувати за допомогою процедур *Distribution* стартово
панелі *Descriptive statistics* (модуль *Basic Statistics and Tables*
Розподіл сукупності за варіаційною ознакою подається у вигляд
таблиці *Frequency Table*, де вказуються інтервали групувань, час
тоти і частки розподілу по інтервалах, а також кумулятивні час
тоти і частки. Опція *Normal expected frequencies* додає в таблиц
теоретичні частоти і частки (групові й кумулятивні). Перевірк
на нормальність розподілу здійснюється за опцією *K-S an*
Lilliefors test for normality.

У табл. 3.1 наведено розподіл 120 фірм, які брали участь
міжнародній виставці, за розміром витрат на рекламу (у % до за
гальної суми витрат) — VAR1. Основні характеристики розподілу

Descriptive Statistics (s_____ sta)						
	Mean	Minimum	Maximum	Std Dev	Skewness	Kurtosis
VAR1	2,295	0,7	3,9	0,6812	0,1361	–0,473

За даними групування максимальне відхилення кумулятивни
часток припадає на п'ятий інтервал і становить у %: |67,5 – 61,83|
= 5,67, тобто $d = 0{,}0567$. Такий же результат маємо при викори
танні кумулятивних частот: $d = |81{,}0 - 74{,}2| : 120 = 0{,}0567$. У таб
3.1 у верхньому лівому куті наведено дещо відмінне значен
критерію $d = 0{,}06713$. Розбіжності між ними зумовлені тим, що
процедурою *K-S and Lilliefors test for normality* розрахунок *d* зді
снюється за незгрупованими даними. Проте обидва значення
значно менші за критичне $\left(1{,}22 : \sqrt{120}\right) = 0{,}11$ при $\alpha = 0{,}10$, отж
гіпотеза про нормальний розподіл фірм за часткою витрат на р
кламу не відхиляється.

Таблиця 3.1

K – S d = ,06713, p > 20, Lilliefors p > 20								
Continue	Count	Cumul Count	Percent of Valid	Cumul % of Valid	Expected Count	Cumul Expected	Percent Expected	Cumul % Expected
,0 < x < = ,5	0	0	0	0	0,5	0,5	0,42	0,42
,5 < x < = 1,0	2	2	1,67	1,67	2,9	3,4	2,44	2,86
1,0 < x < = 1,5	15	17	12,50	14,17	11, 2	14,6	9,29	12,15
1,5 < x < = 2,0	26	43	21,67	35,83	25,3	39, 9	21,11	33,26
2,0 < x < = 2,5	38	81	31,67	67,50	34,3	74,2	28,57	61,83
2,5 < x < = 3,0	22	103	18,33	85,83	27,8	102,0	23,14	84,97
3,0 < x < = 3,5	13	116	10,83	96,67	13,4	115,4	11,19	96,16
3,5 < x < = 4,0	4	120	3,33	100	3,9	119,3	3,23	99,39

Узгодженість емпіричного розподілу фірм за рівнем витрат на рекламу з нормальним розподілом видно на рис. 3.1.

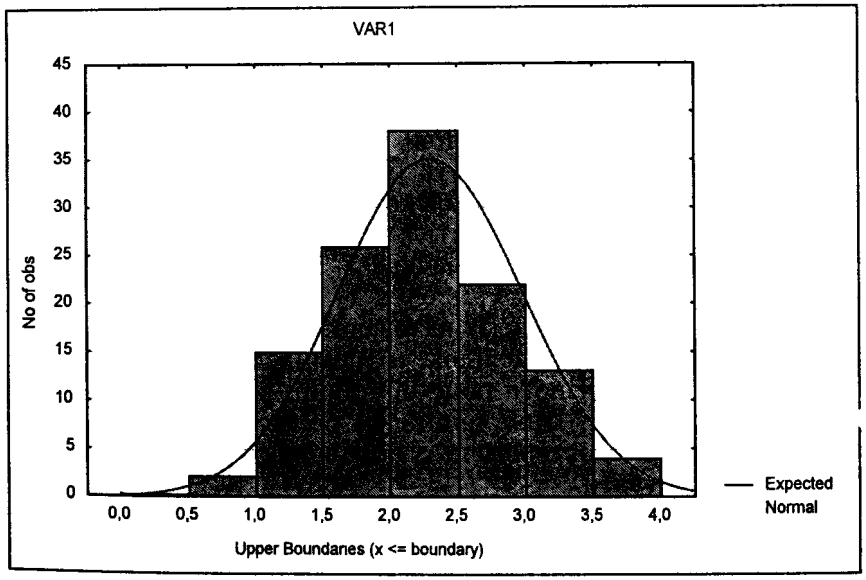


Рис. 3.1 Емпіричний і теоретичний розподіл фірм за рівнем витрат на рекламу

Перевіримо гіпотезу про узгодженість емпіричного і теоретичного розподілів, застосувавши критерій χ^2. Розрахунок його характеристики наведено в табл. 3.2.

Таблиця 3.2

Інтервал групування	f_j	$\hat{f}_j$	$\frac{(f_j-\hat{f}_j)^2}{\hat{f}_j}$
1	0	0,5	0,5
2	2	2,9	0,28
3	15	11,2	1,29
4	26	25,3	0,02
5	38	34,3	0,34
6	22	27,8	1,21
7	13	13,4	0,01
8	4	3,9	0,00
Разом	X	X	3,65

Фактичне значення χ^2 = 3,65 значно менше за критичне $\chi_{1-0,05}(5) = 9,24$, що з імовірністю 0,95 підтверджує висновок про нормальний розподіл сукупності фірм за витратами на рекламу.

Якщо необхідно перевірити гіпотезу про узгодженість даних з іншими розподілами (лог-нормальним, експоненційним тощо), використовується модуль *Nonparametrics / Distribution* — Непараметричні статистики / Розподіли.

У складі сукупності можуть бути окремі одиниці, в яких значення варіюючої ознаки далеко віддалені від центра розподілу й нетипові для сукупності в цілому, *аномальні*. Це може бути максимальне x_n чи мінімальне x_1 значення в упорядкованому ряду спостережень $x_1 \le x_2 \le x_3 \le \ldots \le x_n$. Походження аномальних спостережень (їх називають викидами) різне. Вони можуть виникнути як наслідок: а) помилок спостережень; б) випадкового збігу різного роду обставин; в) порушення однорідності сукупності. Щоб прийняти об'єктивне рішення про вилучення таких спостережень з подальшого аналізу, необхідно їх ідентифікувати.

Оцінити істотність відхилень від основної маси можна за допомогою критерію Граббса T_n, статистичною характеристикою якого є стандартизоване граничне відхилення аномального значення x_n (x_1) від середньої $\bar{x}$. Якщо йдеться про максимальне значення x_n, то

$$T_n = \frac{x_n - \bar{x}}{\sigma},$$

де $\bar{x}$ і σ визначаються для сукупності в цілому.

Критичні значення $T_{1-\alpha}(n)$, де n — обсяг сукупності, наведено в додатку 4. Якщо фактичне значення T_n менше за критичне, то відхилення з імовірністю $(1 - \alpha)$ визнається випадковим, неістотним, а якщо перевищує критичне, то відхилення визнається істотним, а отже, значення — аномальним, нетиповим для сукупності в цілому. В такому разі це значення вилучається і критерій застосовується до x_{n-1} і т. д., доки не буде визнано, що викидів немає, а отже, сукупність однорідна.

Приміром, перевіримо на аномальність максимальний рівень втрат цукру при переробці сировини по сукупності цукрових заводів (табл. 2.2). Саме за цією ознакою (VAR3) коефіцієнт асиметрії виявився найбільшим (1,275). Максимальне значення ознаки $x_n = 2,68$, середня $\bar{x} = 2,23$, $\sigma = 0,179$. Звідси

$$T_n = \frac{2,68 - 2,23}{0,179} = 2,51,$$

що менше за критичне $T_{0,95}(15) = 2,705$. Отже, максимальне значення ознаки неістотно відрізняється від основної маси, і сукупність визнається однорідною.

Однією з поширених форм неоднорідності сукупностей соціально-економічних явищ є внутрішня їх розшарованість. Це зумовлено нерівномірністю розвитку окремих одиниць сукупності (різний вік, належність до різних поколінь тощо) і своєрідністю умов, у яких вони функціонують (природних, технологічних тощо). Одні сукупності поділяються на чітко визначені, ізольовані класи (групи, типи), іншим властива латентна, прихована структура.

Поділ сукупності на однорідні класи (групи, кластери) називають *класифікацією*. Ідея класифікації ґрунтується на поняттях подібності і відмінності. Методологічний принцип класифікації містить два фундаментальних положення:

- в один клас об'єднуються подібні, схожі між собою одиниці сукупності;
- ступінь подібності, схожості одиниць, які належать до одного класу, вища, ніж ступінь подібності одиниць, віднесених до різних класів.

Оцінювання подібності здійснюється на основі однієї чи декількох ознак, які, на думку експертів, формують «образ класу». В традиційній схемі класифікації ці ознаки ієрархічно впорядко-

вуються за своєю вагомістю. Наприклад, класифікація шахт за гірничо-геологічними умовами виробництва: потужністю пласта, нахилом його залягання, глибиною розробки лав, загазованістю лав і т. д. Саме так будується більшість комбінаційних групувань. На кожному кроці поділу сукупності до уваги береться лише одна ознака, тобто відбувається послідовне формування, покрокове уточнення, детальніше описування класів. У невеликих за обсягом сукупностях можливості використання такої схеми класифікації обмежені.

Друга схема класифікації використовує множину класифікаційних ознак одночасно. Будь-яка одиниця сукупності, описана множиною ознак, геометрично інтерпретується як точка у багатовимірному просторі, а близькість двох точок розглядається як подібність їх, однорідність. Існують різні варіанти реалізації багатовимірної схеми класифікації. Їх можна об'єднати в два блоки:

- конструювання багатовимірних інтегральних оцінок (індексів, рейтингів), на основі яких проводиться класифікація за традиційною схемою;
- автоматична багатовимірна класифікація методами кластерного аналізу, коли поняття однорідності задається певними метриками.

Слід зазначити, що класифікація за будь-якою схемою є певною мірою суб'єктивною, оскільки результати її визначаються передусім множиною класифікаційних ознак та їхніми розмежувальними властивостями.

3. 2. КЛАСТЕРНІ ПРОЦЕДУРИ КЛАСИФІКАЦІЇ

Кластер — це група, клас однорідних одиниць сукупності. Основне завдання кластерного аналізу — формування таких груп у багатовимірному просторі. Однорідність сукупності задається правилом обчислення певної метрики, що характеризує ступінь подібності (схожості) j-ї та k-ї одиниць сукупності. Такою метрикою може бути *відстань* між ними c_{jk} або *коефіцієнт подібності* r_{jk}. Близькі, схожі за вибраними метриками одиниці вважаються належними до одного типу, однорідними. Вибір метрики є вузловим моментом кластерного аналізу, від якого залежить кінцевий варіант поділу сукупності на класи.

На ознаках метричної шкали формується *матриця відстаней* розміром $n \cdot n$ з нульовими діагональними елементами. Використовують різні метрики відстані, з-поміж яких найбільш відома *Евклідова відстань*:

$$c_{jk} = \left[\sum_{i=1}^{m}\left(z_{ij} - z_{ik}\right)^2\right]^{1/2},$$

де z_{ij} і z_{ik} — стандартизовані значення i-ї ознаки в j-ї та k-ї одиниць сукупності.

Приклад розрахунку *Евклідової відстані* на ознаковій множині $m = 5$ наведено в табл. 3.3. Згідно з даними $c_{jk} = \sqrt{0{,}49} = 0{,}7$.

Таблиця 3.3

Одиниця сукупності	Стандартизовані значення ознаки					Разом
	z_{1j}	z_{2j}	z_{3j}	z_{4j}	z_{5j}	
j	0,4	0,9	0,3	1,1	0,5	X
k	0,8	0,6	0,5	0,9	0,7	X
$(z_{ij} - z_{ik})^2$	0,16	0,09	0,04	0,16	0,04	0,49

Якщо ознаки x_i різновагомі, то розраховується зважена *Евклідова відстань* з вагами ω_i:

$$c_{jk} = \left[\sum_{i=1}^{m}\omega_i\left(z_{ij} - z_{ik}\right)^2\right]^{1/2}.$$

За своєю квадратичною формою *Евклідова відстань* вписується у традиційні статистичні конструкції, проте на практиці використовують й інші метрики, зокрема *Манхеттенську відстань*

$$c_{jk} = \sum_{i=1}^{m}\left|z_{ij} - z_{ik}\right|.$$

Інформаційною базою кластерного аналізу є матриця відстаней. Приклад такої матриці для сукупності $n = 5$ наведено в табл. 3.5. За способом кластеризації розрізняють ієрархічні та ітераційні процедури. З-поміж ієрархічних найбільш відома і вживана *агломеративна* (об'єднувальна) процедура, суть якої — послідовне об'єднання двох найближчих одиниць сукупності. В матриці відстаней це одиниці, що мають мінімальну відстань c_{jk}. На першому кроці об'єднання всі одиниці сукупності розглядаються як окремі кластери; після кожного кроку розмірність матриці зменшується на одиницю. Повна кластеризація n одиниць відбувається за $(n - 1)$ кроків.

Іноді кластерні процедури вводять обмеження зверху на максимальну відстань між об'єктами одного класу. Таке обмеження

називають *порогом*. Якщо при формуванні кластерів відстань між об'єктами перевищує поріг c_0, то ці об'єкти за певними правилами відносяться до різних кластерів. Порогове значення вибирається суб'єктивно або за певною схемою, може бути постійним або змінюватися, скажімо, монотонно зростаючи на кожному кроці формування кластерів.

Загальну схему агломеративної кластер-процедури на матриці відстаней можна представити як повторення трьох операцій:

1) пошук мінімальної відстані між j-им і k-им кластерами;

2) об'єднання j та k в один кластер і надання останньому спільного індексу q;

3) розрахунок відстаней від сформованого кластера q до інших одиниць сукупності c_{qs} за формулою

$$c_{qs} = a_1 c_{js} + a_2 c_{ks} + a_3 c_{jk} + a_4(c_{js} - c_{ks}).$$

Значення коефіцієнтів a_1, a_2, a_3, a_4 залежать від алгоритму формування кластерів. Для трьох алгоритмів їх наведено в табл. 3.4. Так, за алгоритмом одиничного зв'язку (близького сусіда) одиниця s приєднується до кластера q, якщо вона близька хоча б до одного представника цього кластера. В алгоритмі повного зв'язку (далекого сусіда) відстань між кластером q і s-ю одиницею визначається як відстань до найвіддаленішого представника кластера q. Алгоритм середнього зв'язку використовує середню відстань між кандидатом на включення в кластер q і представниками існуючого кластера.

Таблиця 3.4

Алгоритм	a_1	a_2	a_3	a_4
Одиничного зв'язку	0,5	0,5	0	–0,5
Повного зв'язку	0,5	0,5	0	0,5
Середнього зв'язку	0,5	0,5	0	0

Застосуємо алгоритм одиничного зв'язку до матриці відстаней (табл. 3.5). Оскільки мінімальною є відстань $c_{12} = 0{,}6$, то перша і друга одиниці сукупності об'єднуються в один кластер $q = 1 \cap 2$. Перераховані відстані від новоутвореного кластера до інших одиниць сукупності становлять:

$c_{q3} = 0{,}5 \times 2{,}8 + 0{,}5 \times 1{,}7 - 0{,}5 \times (2{,}8 - 1{,}7) = 1{,}7$;

$c_{q4} = 0{,}9$;

$c_{q5} = 2{,}0$.

Таблиця 3.5

j	1	2	3	4	5
1	0	0,6	2,8	3,3	2,0
2	0,6	0	1,7	0,9	2,5
3	2,8	1,7	0	0,7	4,2
4	3,3	0,9	0,7	0	1,8
5	2,0	2,5	4,2	1,8	0

Як бачимо, відстань будь-якої одиниці сукупності до кластера q дорівнює мінімальній з тих двох відстаней, за якими велися розрахунки. Нову матрицю наведено в табл. 3.6.

Таблиця 3.6

j	$q = 1 \cap 2$	3	4	5
$q = 1 \cap 2$	0	1,7	0,9	2,0
3	1,7	0	0,7	4,2
4	0,9	0,7	0	1,8
5	2,0	4,2	1,8	0

У новій матриці мінімальна відстань $c_{34} = 0{,}7$, отже, об'єднанню підлягають третя і четверта одиниці сукупності.

Результати ієрархічних процедур кластеризації оформляються у вигляді деревоподібних діаграм — *дендрограм*. На одній осі дендрограми зазначаються номери об'єктів, на другій — відстані, за якими відбувається об'єднання. Дендрограма відображує ієрархію структур: кожний кластер можна розглядати як елемент іншого, з більшим значенням відстані c_{jk} (рис. 3.2).

У системі *Statistica* ієрархічну процедуру класифікації — *Joining* (*Tree clustering*) — реалізовано в модулі *Cluster Analysis*. Діалогове вікно процедури пропонує вибрати установки аналізу:

- ознакову множину;
- тип первинних даних: *Raw Data* — дані типу «об'єкт—ознака» чи *Distance Matrix* — матриця відстаней;
- варіант класифікації: за стовпцями (*columns*) — класифікація ознак чи за рядками (*rows*) — класифікація об'єктів;
- алгоритм об'єднання — *Amalgamation (linkage) Rules*; за умовчування — алгоритм одиничного зв'язку — *Single linkage (nearest neighbor)*;
- метрику відстаней — *Distance measure: Euclidean distances, City-block (Manhattan) distance*, інші.

За командою на виконання вибраних установок система видає *Joining Results* з опціями виду дендрограми — горизонтальної чи вертикальної. На рис. 3.1 наведено вертикальну дендрограму класифікації дітей за рівнем інтелектуального розвитку (алгоритм одиничного зв'язку, Евклідова відстань).

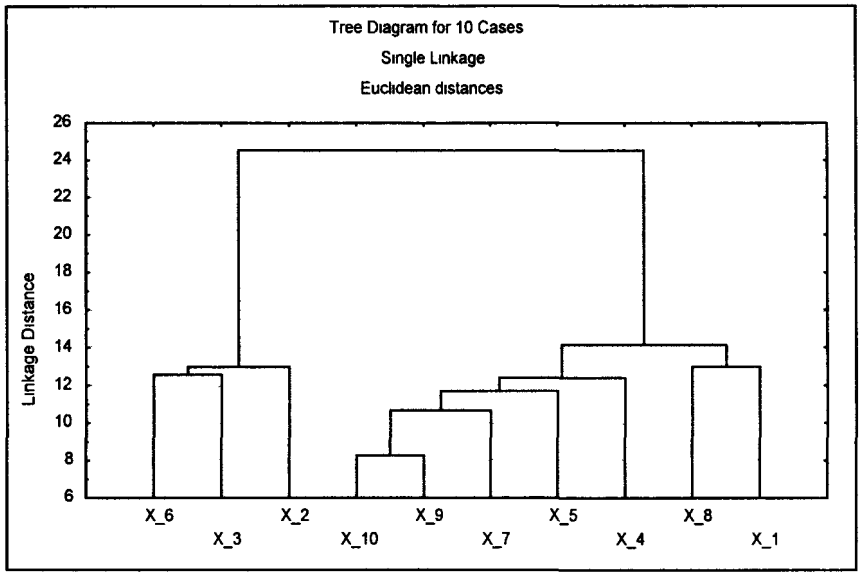


Рис 3 2 Дендрограма

Ознакова множина класифікації включає результати тестування (тести логічного мислення, змістової та асоціативної пам'яті, просторової візуалізації). На графіку чітко виділяються два кластери: до першого потрапили троє дітей, до другого — семеро.

Вісь дендрограми має розмірність метрики. За допомогою опції *Scale tree to dlink/dmax**100 розмірність осі можна нормувати процентним співвідношенням.

Якщо значення ознаки представлені двійковим кодом, для оцінювання ступеня близькості об'єктів використовують коефіцієнт подібності r_{jk}. Розрахунок його ґрунтується на співвідношеннях кількості ознак, значення яких збігаються чи не збігаються. Наприклад, оцінюється якість продукції за m параметрами. Кожна одиниця сукупності характеризується вектором значень цих параметрів якості. Для параметра, що відповідає стандарту, $x = 1$, для параметра, що не відповідає стандарту, $x = 0$ (табл. 3.7).

Таблиця 3.7

Одиниця сукупності	Параметр якості							
	а	б	в	г	ґ	д	е	є
j	0	1	1	0	1	0	0	1
k	0	0	1	1	1	1	0	1

Таблиця 3.8

Значення ознаки	1	0
1	a	b
0	c	d

Частоти однакових і різних пар значень ознак зручно подавати у вигляді чотириклітинкової таблиці (табл. 3.8). У нашому прикладі кількість пар однакових значень ознаки: одиничних a (1, 1) = 3; нульових d (0, 0) = 2. Кількість пар ознак, значення яких не збігаються: b (1, 0) = 1; c (0, 1) = 2.

За умови, що одиничні та нульові ознаки рівновагомі, використовують відношення

$$r_{jk} = \frac{a+d}{a+b+c+d}.$$

Якщо значущими вважаються лише одиничні ознаки, то частоту a відносять або до загальної кількості ознак (коефіцієнт Рао) або до загальної кількості одиничних значень (коефіцієнт Жаккара):

$$r_{jk} = \frac{a}{a+b+c+d}; \qquad r_{jk} = \frac{a}{a+b+c}.$$

Іноді важливо надати подвійну вагу одиничним ознакам. Скажімо, коли «1» позначає відхилення i-го параметра від нормативу (коефіцієнт Дейка):

$$r_{jk} = \frac{2a}{2a+b+c}.$$

У практиці використовують багато інших оцінок ступеня подібності. Значення r_{jk} коливаються в межах $0 \le r_{jk} \le 1$. Очевидно, що різні коефіцієнти, розраховані для тих самих об'єктів, за величиною будуть різними. Вибір коефіцієнта r_{jk} визначається відносною значущістю одиничних і нульових ознак, важливістю порозрядного збігу чи незбігу їхніх значень, а отже, певною мірою є суб'єктивним.

У матриці коефіцієнтів подібності діагональні елементи представлено одиницями. Агломеративна процедура послідовно об'єднує найближчі об'єкти, що мають максимальний коефіцієнт r_{jk}.

Ієрархічна кластер-процедура досить проста і прийнятна для інтерпретації. Проте для численної сукупності вона виявляється громіздкою. У таких випадках перевагу віддають ітераційним процедурам.

На відміну від ієрархічної процедури, яка потребує розрахунку і збереження матриці подібності, ітераційна процедура оперує безпосередньо первинними даними; формуються кластери одного рангу, ієрархічно не підпорядковані. Основні риси ітераційних кластер-процедур розглянемо на прикладі алгоритму *k-середніх*, який реалізує ідею утворення груп за принципом «найближчого центра».

На першому кроці ітераційного процесу здійснюється орієнтовний поділ сукупності на класи і визначаються центри тяжіння (багатовимірні середні) цих класів.

На другому кроці визначаються Евклідові відстані одиниць сукупності до центрів тяжіння виділених кластерів, і кожна з них відноситься до того кластера, центр тяжіння якого найближчий.

На третьому кроці розраховуються нові центри тяжіння кластерів.

Кроки 2 і 3 повторюються доти, доки склад кластерів не стабілізується. Ітерації за принципом *k-середніх* у явному вигляді не використовують критеріїв якості класифікації, проте неявно вони мінімізують внутрішньогрупові дисперсії, забезпечуючи тим самим однорідність сформованих кластерів.

3.3. КЛАСИФІКАЦІЯ НА ОСНОВІ ДИСКРИМІНАНТНОЇ ФУНКЦІЇ

З-поміж методів розпізнавання образів особливе місце посідає дискримінантний аналіз. На відміну від кластерного аналізу дискримінантний не утворює нових класів, а допомагає виявити різницю між існуючими класами і віднести новий (нерозпізнаний) об'єкт до одного з них за принципом максимальної схожості. Наприклад, банк, спираючись на певну систему характеристик фінансового стану клієнтів, які звертаються за позиками, класифікує їх на дві категорії: надійні та ненадійні. Дискримінантний аналіз використовується в медичній діагностиці, при визначенні ризику відмови приладів у технічних системах тощо. Основна проблема — звести помилку класифікації до мінімуму.

Дискримінантна функція — це лінійна комбінація певної множини ознак, які називаються класифікаційними і на основі яких ідентифікуються класи. Особливість дискримінантної функції полягає в тому, що класи представляються шкалою найменувань, а класифікаційні ознаки x_i, де $i = 1, 2, \ldots, m$, вимірюються метричною шкалою. Кількість останніх не може перевищувати $(n - 2)$, де n — обсяг сукупності. Функціонально зв'язані та висококорельовані ознаки до ознакового простору моделі не включаються.

Дискримінантна функція f_j визначається для кожного j-го класу ($j = 1, 2, \ldots, p$):

$$f_j = a_{0j} + a_{1j}\bar{x}_{1j} + a_{2j}\bar{x}_{2j} + \ldots + a_{mj}\bar{x}_{mj},$$

де a_{ij} — коефіцієнт функції (змістовної інтерпретації не має);

$\bar{x}_{ij}$ — середнє значення i-ї ознаки в j-му класі.

Коефіцієнти функції a_{ij} можна розрахувати за формулами [16, с. 113]:

$$a_{ij} = (n - p)\sum_{k=1}^{m} b_{ik}\bar{x}_{kj},$$

де b_{ik} — елемент матриці, оберненої до внутрішньогрупової матриці сум попарних добутків $W_{ik} = \sum_{j=1}^{p}\sum_{h=1}^{n_j}(x_{ijh} - \bar{x}_{ij})(x_{kjh} - \bar{x}_{kj})$;

константа $a_{0j} = -0{,}5\sum_{k=1}^{m} a_{kj}\bar{x}_{kj}$.

У геометричній інтерпретації f_j — це уявна точка m-вимірного Евклідового простору, координатами якої є середні значення класифікаційних ознак j-го класу. Значення f_j для p класів розглядаються як центри їх тяжіння і називаються *центроїдами*.

Процедура класифікації ґрунтується на геометричній близькості h-ї одиниці (з координатами значень ознак x_{ih}) до центроїдів виділених класів. Належність її до того чи іншого класу визначається на основі *відстані Махаланобіса*, яку можна записати так:

$$D^2 = (n - p)\sum_{i=1}^{m}\sum_{k=1}^{m} b_{ik}(x_{ijh} - \bar{x}_{ij})(x_{kjh} - \bar{x}_{kj}).$$

Дискримінантна функція максимізує різницю між класами і мінімізує дисперсію всередині класу. Критерієм оптимального поділу сукупності на класи є максимум відношення міжкласової варіації до внутрішньокласової.

Міжкласову варіацію характеризує квадрат різниці центроїдів $(f_j - f_s)^2$, а внутрішньокласову — середній квадрат відстаней між точками, що належать j-му класу x_{ijh}, і центроїдами цих класів f_j:

$$\sigma_f^2 = \frac{\sum_{j=1}^{p}\sum_{h=1}^{n_j} a_{ij}(x_{ijh} - \bar{x}_{ij})^2}{\sum_{j=1}^{p} n_j - p},$$

де n_j — кількість одиниць j-го класу.

Отже, критерій оптимального поділу на класи можна представити відношенням

$$D^2 = \frac{(f_j - f_k)^2}{\sigma_f^2},$$

яке називають *узагальненою міжкласовою відстанню Махаланобіса.*

Для оцінювання спроможності дискримінантної функції розпізнавати класи у багатовимірному ознаковому просторі використовують також λ-статистику Вілкса (*Wilks lambda*):

$$\lambda = \prod_{j=1}^{p} \frac{1}{1+\lambda_j},$$

де λj — властиві значення матриці коваріацій.

λ-статистика враховує як відмінності між класами, так і однорідність кожного класу. Оскільки λ розраховується як обернена величина, то чим більше різняться центроїди, тим менше її значення, і навпаки, якщо центроїди збігаються, то λ прямує до 1. Отже, близькі до 0 значення λ свідчать про високу розпізнавальну спроможність дискримінантної функції. Істотність різниці значень центроїдів перевіряється також за допомогою критерію χ^2 чи дисперсійного F-критерію, які функціонально зв'язані з λ-статистикою.

У системі *Statistica* процедури дискримінантного аналізу об'єднані в модулі *Discriminant Analysis* — Дискримінантний аналіз. Порядок використання модуля розглянемо на умовному прикладі професійної психодіагностики, методика якої передбачає дискримінацію претендентів на заміщення вакансій на дві групи: відповідають (група С) і не відповідають (група NC) вимогам професії. Діагностичні ознаки: VAR2 — оперативна пам'ять, VAR3 — концентрація уваги. Значення цих ознак у балах наведено в табл. 3.9.

Таблиця 3.9

	VAR1	VAR2	VAR3
1	C	72	75
2	C	57	70
3	C	59	62
4	C	67	72
5	C	75	59
6	C	62	73
7	NC	67	50
8	NC	56	59
9	NC	58	54
10	NC	47	60

За командами на стартовій панелі модуля проведемо селекцію ознак: незалежні (*independent variable list*) — VAR2 та VAR3; ідентифікатор груп (*grouping variable*) — VAR1; вкажемо метод аналізу — *Standart*. За результатами аналізу в інформаційній частині діалогового вікна вказується кількість класифікаційних ознак, значення λ-статистики та F-критерію:

```
Discriminant Function Analysis Results
Number of variables in the model: 2
Wilks'Lambda:,270128 approx.F(2,7)=9,45681 p < ,01024.
```

Згідно з даними дискримінантна функція спроможна визначити професійно придатних осіб з мінімальною ймовірністю помилки. Параметри дискримінантної функції за кожним з виділених класів визначимо за допомогою процедури *Classification functions* (значення їх наведено в табл. 3.10).

Таблиця 3.10

Classification Functions; grouping: VAR1 (new. sta)		
Continue...	C p = ,60	NC p = ,40
VAR2	1,9867	1,6938
VAR3	2,9689	2,4539
Constant	–167,0933	–117,5931

Установки аналізу *Distances between groups* і *Squared Mahalanobis distances* визначають міжкласову та внутрішньокласові відстані. Так, узагальнена міжкласова відстань Махаланобіса становить 11,258. Відстані окремих одиниць сукупності до центроїдів груп наведено в табл. 3.11. Частка правильно класифікованих одиниць сукупності становить 90 % (одна неправильно класифікована одиниця маркірована).

Таблиця 3.11

Squared Mahalanobis Distances from Group Centroids (new.sta)			
Incorrect classifications are marked with *			
Continue...	Observed Classif.	C p =,600	NC p =,400
1	C	3,180	22,786
2	C	1,227	6,898
*3	C	3,051	1,653
4	C	0,586	14,174
5	C	3,140	8,023
6	C	0,616	12,305
7	NC	11,076	2,004
8	NC	6,560	0,315
9	NC	10,313	0,090
10	NC	12,275	1,789

Нові, нерозпізнані об'єкти відносяться до того класу, для якого індивідуальні значення дискримінантної функції більші. Скажімо, в нашому прикладі новий претендент на заміщення вакансії набрав 65 балів по тесту «оперативна пам'ять» і 68 балів по тесту «концентрація уваги». Значення дискримінантної функції для групи *C* становить 163,957, для групи *NC* — 159,39. Оскільки перше значення функції більше, то претендент належить до групи *C*.

Розглянуту процедуру класифікації можна використати й тоді, коли кількість класів $m > 2$. Важливо, щоб кількість одиниць у кожному класі була не менша 2. Іноді метою дискримінантного аналізу є не віднесення об'єктів до того чи іншого класу, а визначення апостеріорних імовірностей належності до цих класів. Результати такого аналізу дає установка *Posterior Probabilities*.

Завдання для самоконтролю

1. Розподіл 400 домогосподарств за рівнем середньодушового доходу характеризується даними:

Номер групи	Частка емпіричного розподілу, %	Імовірність теоретичного розподілу, %	
		нормального	лог-нормального
1	5,5	6,7	5,3
2	9,4	8,0	11,6
3	17,2	12,7	16,3
4	15,3	16,6	17,6
5	15,0	17,8	15,0
6	13,0	15,5	12,7
7	10,2	11,2	9,0
8	8,5	6,6	6,5
9	3,8	3,1	4,0
10	2,0	1,8	2,0
Разом	100	100	100

За допомогою критеріїв χ^2 та Колмогорова-Смирнова *d* перевірте, з нормальним чи лог-нормальним розподілом узгоджується розподіл домогосподарств за середньодушовим доходом. Висновок зробіть з імовірністю 0,95.

2. Виробничі потужності 12 металургійних комбінатів характеризуються такими даними, млн. т/рік:

Продукція	Максимальний рівень	Мінімальний рівень	Середній рівень	Середнє квадратичне відхилення
Чавун	10,4	3,2	8,5	1,3
Сталь	17,6	4,4	9,8	2,2
Прокат	12,4	3,7	7,6	1,9

За допомогою критерію Граббса перевірте однорідність сукупності металургійних комбінатів за виробничими потужностями. Висновок зробіть з імовірністю 0,95.

3. За наведеними даними визначте Евклідові відстані між агрогосподарствами за рівнем забезпеченості технікою (в розрахунку на 100 га):

Господарство	Стандартизовані значення показників забезпеченості			
	колісними тракторами	зернозбиральними комбайнами	знаряддями поверхневого обробітку ґрунту	транспортними засобами
А	1,15	0,72	–0,16	0,26
Б	–0,36	0,58	–0,43	0,27
В	0,64	1,45	1,02	0,38

4. За даними матриці відстаней, використовуючи ієрархічну кластер-процедуру (алгоритм одиничного зв'язку), здійсніть класифікацію агрогосподарств за рівнем забезпеченості технікою, побудуйте дендрограму:

Господарство	1	2	3	4	5	6
1	0	0,60	2,17	1,42	1,96	3,32
2		0	0,85	2,58	1,75	2,24
3			0	2,03	3,13	1,40
4				0	2,46	1,08
5					0	0,65
6						0

5. За даними про відповідність окремих параметрів робочого місця нормативам («0» — відповідає, «1» — не відповідає) обчисліть попарні міри подібності, на основі їх складіть матрицю подібності:

Параметр робочого місця	Робоче місце				
	1	2	3	4	5
Безпека	1	0	0	1	0
Шум	1	0	1	1	0
Вібрація	1	1	0	1	0
Температура	0	1	1	0	1
Загазованість	0	1	0	1	0
Освітленість	1	0	0	0	0

Обґрунтуйте, яку міру подібності необхідно використати.

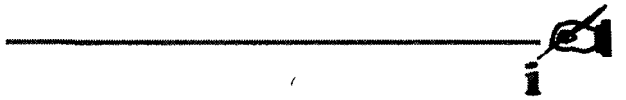
6. За даними про приплив води в шахту ($м^3$/год) і кількість метану на 1 т середньомісячного видобутку вугілля ($м^3$) шахти поділяються на дві групи. Перша об'єднує шахти з безпечними умовами праці, друга — з небезпечними. Дискримінантна функція класифікації шахт за рівнем безпеки праці має такий вигляд:

$$f = 0{,}35\,x_1 + 2{,}42\,x_2.$$

Визначте центроїди груп і процедуру класифікації шахт; класифікуйте нові шахти за наведеними нижче даними:

Показник	Середній рівень у групі		Нові шахти	
	1	2	А	Б
Приплив води в шахту	120	150	125	118
Кількість метану на 1 т видобутку вугілля	16	25	24	22

7. Як оцінити якість класифікації? Обґрунтуйте вибір критерію якості.



Розділ 4

МОДЕЛЮВАННЯ ТА ПРОГНОЗУВАННЯ ДИНАМІКИ

4.1. ОСНОВНІ ЗАСАДИ МОДЕЛЮВАННЯ ДИНАМІКИ

Ряди динаміки характеризують процеси розвитку соціально-економічних явищ. Цим процесам властиві дві взаємопов'язані риси: динамічність та інерційність. Динамічність проявляється зміною рівнів і варіації показників, що характеризують процес, інерційність — сталістю механізму формування процесу, напрямку та інтенсивності динаміки протягом певного часу. Поєднуючи ці риси, динамічний ряд у будь-який момент t містить залишки минулого, основи сучасного і зародки майбутнього.

Діалектична єдність мінливості й сталості, динамічності й інерційності формує закономірність розвитку. Під впливом безлічі факторів довгострокової і короткострокової дії в одних рядах рівні протягом тривалого часу зростають або зменшуються з різною інтенсивністю, в інших зростання і зменшення рівнів чергуються з певною періодичністю (наприклад, одинадцятирічні цикли градових опадів, зумовлені циклами сонячної активності). З року в рік більш-менш регулярно повторюються сезонні піднесення і спади (використання виробничих потужностей і робочої сили, попит на ринку споживчих товарів тощо). Окрім закономірних коливань рівнів, динамічним рядам притаманні також випадкові коливання, пов'язані з масовим процесом.

Ряди, в яких рівні коливаються навколо постійної середньої, називаються стаціонарними. Економічні ряди, як правило, нестаціонарні. Для більшості з них характерна систематична зміна рівнів з нерегулярними коливаннями, коли піки і западини чергуються з різною інтенсивністю. Скажімо, економічні цикли (промислові, будівельні, фондового ринку тощо) повторюються з різною тривалістю і різною амплітудою коливань. Рисунок 4.1 ілюструє характер динаміки виплат страхового відшкодування VAR2, коливання якого залежать від кількості постраждалих об'єктів. Поквартальні ($n = 18$) обсяги виплат коливаються від 7,9 до 19,2 млн. грн., на графіку вони представлені відхиленнями від мінімального рівня.

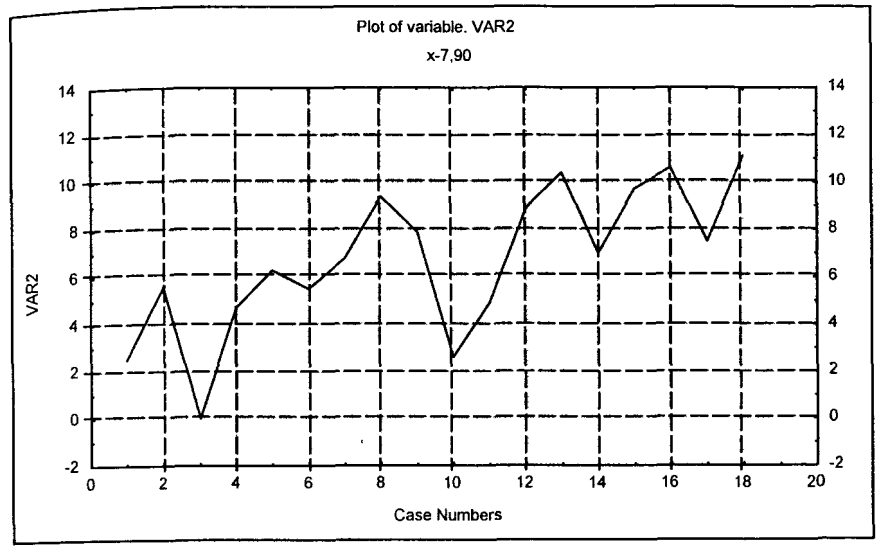


Рис. 4.1. Динаміка виплат страхового відшкодування

При моделюванні динамічних процесів причинний механізм формування властивих їм особливостей у явному вигляді не враховується. Будь-який процес розглядається як функція часу. Певна річ, час не є фактором конкретного соціально-економічного процесу, змінна часу t просто акумулює комплекс постійно діючих умов і причин, які визначають цей процес.

У моделях динаміки процес умовно поділяється на чотири складові:

- довгострокову, детерміновану часом еволюцію — тренд $f(t)$;
- періодичні коливання різних частот C_t;
- сезонні коливання S_t;
- випадкові коливання e_t.

Зв'язок між цими складовими представляється адитивно (сумою) або мультиплікативно (добутком):

$$y_t = f(t) + C_t + S_t + e_t,$$
$$y_t = f(t)\, C_t\, S_t\, e_t.$$

Така умовна конструкція дає змогу, залежно від мети дослідження, вивчати тренд, елімінуючи коливання, або вивчати коливання, елімінуючи тренд. При прогнозуванні здійснюється зведення прогнозів різних елементів в один кінцевий прогноз.

Характерною властивістю будь-якого динамічного ряду є залежність рівнів: значення y_t певною мірою залежить від попере-

дніх значень: y_{t-1}, y_{t-2} і т. д. Для оцінювання ступеня залежності рівнів ряду використовують *коефіцієнти автокореляції* r_p з часовим лагом $p = 1, 2, \ldots, m$.

Коефіцієнт r_p характеризує щільність зв'язку між первинним рядом динаміки і цим же рядом, зсуненим на p моментів. У табл. 4.1 наведено зсунені ряди динаміки з лагами $p = 1, 2, 3$. Як видно, із збільшенням лага p кількість пар корельованих рівнів зменшується. Так, при $p = 1$ довжина корельованих рядів менша за первинний ряд на один рівень, при $p = 2$ — на два рівні і т. д. Через це на практиці при визначенні автокореляційної функції дотримуються правила, за яким кількість лагів $m \le \frac{n}{2}$.

Таблиця 4.1

Змінна часу t	Рівень ряду y_t	$p = 1$	$p = 2$	$p = 3$
1	y_1	—	—	—
2	y_2	y_1	—	—
3	y_3	y_2	y_1	—
…	…	…	…	…
$n-2$	y_{n-2}	y_{n-3}	y_{n-4}	y_{n-5}
$n-1$	y_{n-1}	y_{n-2}	y_{n-3}	y_{n-4}
n	y_n	y_{n-1}	y_{n-2}	y_{n-3}

Значення коефіцієнта автокореляції r_p визначається величиною лага p і не виходить за межі ± 1:

$$r_p = \frac{c_p}{c_0},$$

де $c_p = \frac{1}{n}\sum_{t=1}^{n-p}(y_t - \bar{y})(y_{t+p} - \bar{y})$; $c_0 = \frac{1}{n}\sum_{t=1}^{n}(y_t - \bar{y})^2$.

Послідовність коефіцієнтів r_p називають *автокореляційною функцією* і зображують графічно у вигляді автокорелограми з абсцисою p та ординатою r_p. На рис. 4.2 наведено автокореляційну функцію динамічного ряду виплат страхового відшкодування. Функція має вигляд згасаючих коливань — від $r_1 = 0{,}354$ до $r_9 = 0{,}007$. Точками позначені дві паралельні прямі, що визначають 95%-ні довірчі межі істотності r_p. Якщо r_p не виходить за довірчі межі, ряд вважається стаціонарним.

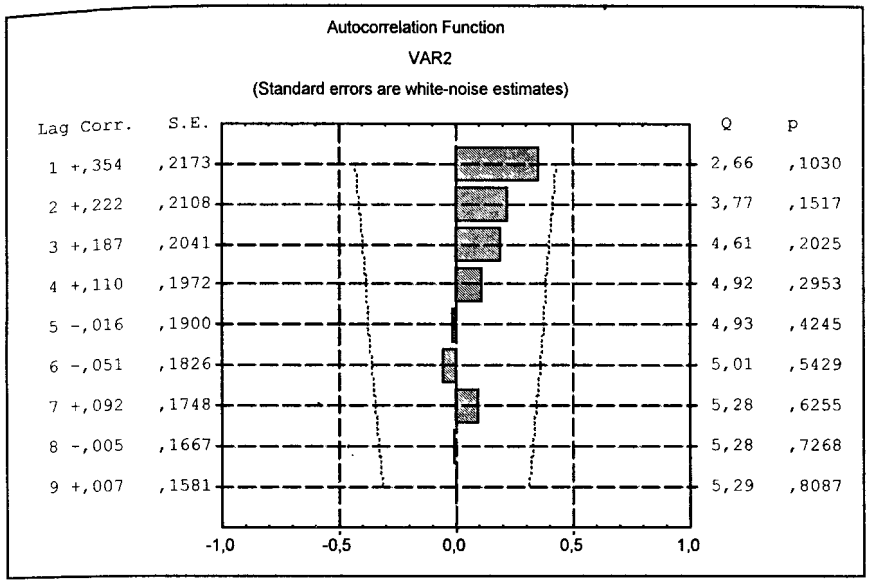


Рис. 4.2. Автокореляційна функція

За швидкістю згасання автокореляційної функції можна зробити висновок про характер динаміки. Найчастіше використовується значення r_1. Характеризуючи ступінь залежності двох послідовних членів ряду, коефіцієнт автокореляції є мірою неперервності цього ряду. Якщо $r_1 \to 1$, то ряду динаміки властива тенденція розвитку, якщо $r_1 \to 0$ — рівні ряду незалежні. Відносно високі значення коефіцієнта автокореляції при $p = k$, $2k$, $3k$ … свідчать про регулярні коливання.

На відміну від детермінованої складової випадкова складова не зв'язана із зміною часу. Аналіз цієї складової є основою перевірки гіпотези про адекватність моделі реальному процесу. За умови, що модель вибрано правильно, випадкова складова являє собою стаціонарний процес з математичним сподіванням $M(e) = 0$ і дисперсією

$$s_e^2 = \frac{\sum_{t=1}^{n}(y_t - f_t)^2}{n - m},$$

де m — число параметрів функції.

Для оцінювання стаціонарності випадкової складової використовують циклічний коефіцієнт автокореляції першого порядку r_1. Корелюються ряди залишкових величин:

$e_1, e_2, e_3, \ldots, e_n$ та $e_2, e_3, e_4, \ldots, e_1$.

Припускаючи, що $\sum e_e = \sum e_{e+1} = 0$, формула розрахунку спрощується:

$$r_1 = \frac{\sum_1^n e_t e_{t+1}}{\sum_1^n e_t^2}.$$

Існують таблиці критичних значень циклічного коефіцієнта автокореляції для додатних і від'ємних значень (додаток 7). Якщо фактичне значення r_1 менше за критичне, автокореляція вважається неістотною, а випадкова складова — стаціонарним процесом. У разі, коли фактичне значення r_1 перевищує критичне, можна зробити висновок про неадекватність детермінованої складової реальному процесу.

У системі *Statistica* моделювання та прогнозування динамічних процесів можна здійснити за процедурами модулів *Multiple Regression*, *Time Series / Forecasting*, *Nonlinear Estimation*. Моделювання трендів і трендова екстраполяція здійснюються за процедурами модулів *Multiple Regression* та *Nonlinear Estimation*; комплексний аналіз динамічних процесів, ідентифікація моделей, адаптивне прогнозування — за процедурами модуля *Time Series / Forecasting*.

Стартова панель модуля *Time Series / Forecasting* має дві частини: верхню — інформаційну і нижню — функціональну. У функціональній частині стартової панелі представлені методи обробки даних:

- *Arima & autocorrelation functions* — Модель авторегресії та проінтегрованої ковзної середньої;
- *Interrupted time series analysis* — Аналіз розірваного динамічного ряду (моделі інтервенції для *Arima*);
- *Exponential smoothing & forecasting* — Експоненційне згладжування та прогнозування;
- *Seasonal decomposition (1, 2)* — Сезонна декомпозиція 1 і 2 (помісячна й поквартальна);
- *Distributed lags analysis* — Аналіз розподілених лагів (регресійна модель для двох динамічних рядів);
- *Spectral (Fourier) analysis* — Спектральний (Фур'є) аналіз.

Верхня інформаційна частина стартової панелі містить команди, за допомогою яких відкривається файл даних (*Open data*) і вибираються ознаки для аналізу (*Variables*). Одномірний динамічний ряд характеризується однією ознакою y, ім'я та довге ім'я якої висвічуються у вікні стартової панелі, біля імені з'являється символ *L*. Це означає, що ознака «замкнена на ключ» і її неможливо вилучити.

За необхідності вибрану ознаку можна трансформувати, використовуючи послідовність команд: *OK (transformations, autocorrelations, crosscorrelations, plots)* → *OK (Transform highlighted variable)* → *Transformations of variables*. У вікні *Time series transformations* пропонується широкий спектр трансформацій динамічного ряду, зокрема:

- *Add a constant* — додати константу до значень ряду;
- *Power* — піднести до степеня;
- *Inverse power* — добути корінь;
- *Natural log* — логарифмування за основою натурального логарифма;
- *Exponent* — піднесення до степеня за експонентою;
- *Mean subtract* — відхилення від середнього рівня;
- *Standardize* — стандартизація значень ознаки.

Усі варіанти трансформації застосовують послідовно один за одним лише для висвіченої ознаки. В двох останніх варіантах можна вказати середні та стандартні відхилення або визначити їх автоматично, користуючись опцією *Estimate mean & std.dev.from data*.

Модуль трансформації передбачає усунення лінійного тренда *Trend subtract*, вводячи його параметри власноруч або використовуючи команду автоматичного розрахунку *Estimate a/b from data*. Аналогічно здійснюється усунення автокореляції з відповідним лагом за опцією *Autocorr*.

У групі опцій *Shift relative starting point of series* пропонуються варіанти зсунення ряду вперед чи назад на певний лаг. Опція *Differencing* дає можливість визначити різниці між поточним y_t і зсуненим на лаг p рівнями ряду ($y_t - y_{t+p}$).

Важливе значення має група опцій *Smoothing* — згладжування.

Усі трансформації висвічуються в інформаційній частині стартової панелі, максимальна їх кількість — дев'ять. Якщо та чи інша трансформація в подальшому аналізі не використовується, її можна вилучити за командою *Delete highlighted variable*. Команда *Save*, навпаки, зберігає визначені ознаки в окремому файлі.

4.2. ТИПИ ТРЕНДОВИХ МОДЕЛЕЙ

Важливою складовою динамічних процесів є тенденція середньої, тобто основний напрям розвитку. В аналізі динамічних рядів тенденцію представляють у вигляді плавної траєкторії та описують певною функцією, яку називають *трендом* $Y_t = f(t)$, де $t = 1, 2, \ldots, n$ — змінна часу. На основі такої функції здійснюється вирівнювання динамічного ряду і прогнозування подальшого розвитку процесу.

Процедура вирівнювання динамічних рядів включає два етапи: обґрунтування (вибір) типу функції, яка б адекватно описувала характер динаміки, та оцінювання параметрів функції. На практиці переважно використовують функції, параметри яких мають конкретну інтерпретацію залежно від характеру динаміки. Найбільш поширені поліноми (многочлени), різного роду експоненти та логістичні криві. Так, параметри полінома p-го ступеня $Y_t = a + bt + ct^2 + dt^3 \ldots$ характеризують:

a — рівень динамічного ряду при $t = 0$;

b — абсолютну швидкість зміни рівнів ряду (ординат);

$2c$ — прискорення (приростання абсолютної швидкості);

d — зміну приростання тощо.

Поліном 1-го ступеня, тобто лінійний тренд $Y_t = a + bt$, описує процеси, які рівномірно змінюються в часі і мають стабільні прирости ординат. Поліном 2-го ступеня (парабола) $Y_t = a + bt + ct^2$ здатний описати процес, характерною особливістю якого є рівноприскорене зростання або зменшення ординат. Форма параболи визначається параметром c: при $c > 0$ гілки параболи спрямовані вгору — парабола має мінімум, при $c < 0$ гілки параболи спрямовані вниз — парабола має максимум. При визначенні екстремуму (max, min) похідну параболи прирівнюють до нуля і розв'язують систему рівнянь відносно t. Наприклад, динаміка захворювань при епідемії грипу (чол.) описується параболою $Y_t = 264 + 45t - 1{,}5t^2$. Похідна параболи $45 - 2{,}25t = 0$, а $t = 20$. Максимум захворювань буде зафіксовано через 20 днів від початку відліку часу ($t = 0$) і становитиме $Y_{\max} = 264 + 45 \cdot 20 - 1{,}5 \cdot 20^2 = 564$ чол. У полінома 3-го ступеня $Y_t = a + bt + ct^2 + dt^3$ знак приростання ординати може змінюватися один чи два рази.

Якщо характерною властивістю процесу є стабільна відносна швидкість (темпи приросту), такий процес описується експонентою, яка може набувати різних еквівалентних форм. Основна (показникова) форма експоненти

$$Y_t = ab^t,$$

де b — середня відносна швидкість зміни ординати: при $b > 1$ ордината зростає з постійним темпом, при $b < 1$, навпаки, зменшується. Абсолютний приріст пропорційний досягнутому рівню.

Експоненту можна представити у формі

$$Y_t = ae^{\lambda t} \quad \text{або} \quad Y_t = e^{a+bt},$$

де $\lambda = \ln b$, $e = 2{,}718$ — основа натурального логарифма, $\ln e = 1$.

Експоненти приводяться до лінійного виду заміною y_t десятковими або натуральними логарифмами:

$$\lg Y = \lg a + t \lg b,$$
$$\ln Y = \ln a + \lambda t \ln e = \ln a + \lambda t,$$
$$\ln Y = \ln e^a + \ln e^{bt} = \ln a + \ln bt = \ln a + \lambda t.$$

Оцінювання параметрів трендових рівнянь найчастіше здійснюється *методом найменших квадратів* (МНК), основною умовою якого є мінімізація суми квадратів відхилень фактичних значень y_t від теоретичних Y_t, визначених за трендовим рівнянням

$$\sum_1^n (y_t - Y_t)^2 = \min.$$

Параметри поліноміального тренда визначаються безпосередньо розв'язуванням систем $p + 1$ нормальних рівнянь. Експонента, як показано вище, приводиться до лінійного виду логарифмуванням; розраховані параметри підлягають потенціюванню.

Побудова трендових моделей МНК є ефективною в модулі *Multiple Regression*, який детально описано в 5.3. Стартова панель модуля дає можливість відкрити необхідний файл даних і вибрати ознаки для аналізу, а за необхідності трансформувати значення цих ознак.

Як приклад, розглянемо динаміку обороту універсальної біржі VAR1 за 18 місяців (табл. 4.2). Інтенсивне нарощення біржового обороту можна описати експонентою. Для цього обсяги біржового обороту через опцію *Current Specs* (див. 2.1) замінимо десятковими логарифмами.

У модулі *Multiple Regression* здійснимо селекцію ознак: незалежна — порядковий номер місяця t, залежна — $\lg y_t$ (LOGV1). Після команди на виконання програми *OK* вибираємо опцію *Regression summary*. Параметри моделі наведено в табл. 4.3.

Таблиця 4.2

Порядковий номер місяця, t	Біржовий оборот, млн. грн., VAR1	LOG V1
1	78,4	1,894316
2	81,2	1,909556
3	85,6	1,932474
4	89,4	1,949878
5	93,8	1,972203
6	102,5	2,011147
7	106,2	2,026125
8	112,8	2,052309
9	120,4	2,080626
10	126,3	2,101403
11	130,7	2,116276
12	140,9	2,148911
13	146,7	2,166430
14	153,6	2,186391
15	163,5	2,213518
16	170,6	2,231979
17	179,5	2,254064
18	185,2	2,267641

Таблиця 4.3

Regression Summary for Dependent Variable: LOGV1						
Continue...	R = ,9989 RI = ,9978 Adjusted RI = ,9977					
	F (1,16) = 7384,6 p < ,000 Std. Error of estimate: ,00585					
N = 18	BETA	St. Err. of BETA	B	St. Err. of B	t(16)	p-level
Intercpt			1,8670	0,00287	648,5	8,63E-37
LOGV1	0,9989	0,0116	0,0229	0,00027	85,9	9,38E-23

За даними розрахунку експонента має вигляд $\lg Y_t = 1{,}867 + 0{,}0229t$, стандартна похибка $s_e = 0{,}00585$. Після потенціювання $Y_t = 70{,}6 \cdot 1{,}054^t$. Тобто, біржовий оборот щомісячно зростав у середньому на 5,4%, $s_e = 0{,}0386$.

Виявлену тенденцію можна продовжити за межі динамічного ряду. Така процедура називається *екстраполяцією* тренда. Принципова можливість екстраполяції ґрунтується на припущенні, що умови, які визначали тенденцію у минулому, не зазнають істотних змін у майбутньому. Формально операцію екстраполяції можна представити як визначення функції

$$Y_{t+v} = f(Y_t^*, v),$$

де Y_{t+v} — прогнозне значення на період упередження v;

Y_t^* — база екстраполяції, найчастіше це останній, визначений за трендом рівень ряду.

Якщо, скажімо, в червні ($t_n = 18$) теоретичний рівень біржового обороту становив (млн. грн): $Y_t = 70{,}6 \cdot 1{,}054^{18} = 181{,}9$, то в липні можна очікувати $Y_{18+1} = 181{,}9 \cdot 1{,}054 = 191{,}7$.

У модулі *Multiple Regression* екстраполяція здійснюється за опцією *Predict dependent var*. Для цього у вікні *Specify dependent for indep. vars.* треба вказати значення v.

Екстраполяція тренда дає точковий прогноз. Очевидно, що «влучення в точку» малоймовірне. Адже тренду властива невизначеність, передусім через похибки параметрів. Джерелом цих похибок є обмежена сукупність спостережень y_t, кожне з яких містить випадкову компоненту e_t. Зсунення періоду спостереження лише на один крок веде до зсунення оцінок параметрів. Випадкова компонента буде присутня і за межами динамічного ряду, а отже, її необхідно врахувати. Для цього визначають довірчий інтервал, який би з певною ймовірністю окреслив межі можливих значень Y_{t+v}. Точковий інтервал перетворюється в інтервальний. Ширина інтервалу залежить від варіації рівнів динамічного ряду навколо тренда та ймовірності висновку $(1 - \alpha)$:

$$Y_{t+v} \pm t_{1-\alpha} s_p,$$

де s_p — середня квадратична похибка прогнозу, значення якої залежить від дисперсії тренда s_Y^2 та дисперсії відхилень від тренда s_e^2.

Зокрема, для лінійного тренда

$$s_Y^2 = \frac{s_e^2}{n} + \frac{s_e^2}{\sum (t - \bar{t})^2} (t - \bar{t})^2 .$$

Якщо база прогнозування — останній рівень ряду, то $\sum(t-\bar{t})^2 = \frac{n(n^2-1)}{12}$, а $(t-\bar{t})^2$ замінюється на $(2v+n-1)^2$. Після нескладних алгебраїчних перетворень похибку прогнозу за лінійним трендом можна представити так:

$$s_p = s_e\sqrt{1+\frac{1}{n}+\frac{3(2v+n-1)^2}{n(n^2-1)}}$$

або, позначивши підкореневий вираз символом z, $s_p = s_e z$.

Тобто похибка прогнозу залежить від залишкової дисперсії s_e^2, довжини динамічного ряду (передісторії) n та періоду упередження v. Чим довший період передісторії, тим похибка менша, а збільшення періоду упередження, навпаки, веде до зростання похибки прогнозу.

При визначенні інтервального прогнозу для лінійного тренда (експоненти) можна скористатися таблицями Є. Четиркіна. У додатку 6 наведено значення $Z^* = t_{0,90}z$ для $n = 7 \ldots 25$ та $v = 1 \ldots 5$. Як бачимо, вони прямо пропорційні періоду упередження v та обернено пропорційні довжині динамічного ряду n.

Скористаємося значеннями Z^* для визначення довірчих меж прогнозу біржового обороту. При $n = 18$ і $v = 1$ значення $Z^* = 1{,}9455$. Якщо $s_e = 0{,}0384$, то похибка прогнозу біржового обороту становить $s_p = 0{,}0384 \cdot 1{,}9455 = 0{,}075$, а довірчі межі прогнозу на липень $191{,}7 \pm 0{,}075$. Отже, з імовірністю 0,90 можна стверджувати, що в липні оборот біржі буде щонайменше 191,6 млн. і не перевищить 191,8 млн. грн.

4.3. КОРОТКОСТРОКОВЕ ПРОГНОЗУВАННЯ НА ОСНОВІ КОВЗНИХ СЕРЕДНІХ

Досить поширеним і простим методом аналізу динаміки є згладжування ряду. Суть його полягає в заміні фактичних рівнів y_t середніми за певними інтервалами. Варіація середніх порівняно з варіацією рівнів первинного ряду значно менша, а тому характер динаміки проявляється чіткіше. Процедуру згладжування називають фільтруванням, а оператори, за допомогою яких вона здійснюється, — фільтрами. На практиці використовують переважно лінійні фільтри, з-поміж яких найпростіший — *ковзна середня* з інтервалом згладжування $m < n$. Інтервали поступово зміщуються на один елемент:

$y_1, y_2, \ldots, y_m$;
$y_2, y_3, \ldots, y_{m+1}$;
$y_3, y_4, \ldots, y_{m+2}$ і т. д.

Для кожного з них визначається середня $\bar{y}_t$, яка припадає на середину інтервалу. Якщо m — непарне число, тобто $m = 2p + 1$, а ваги членів ряду в межах інтервалу однакові $a_r = \frac{1}{(2p+1)}$, то

$$\bar{y}_t = \frac{1}{2p+1}\sum_{t-p}^{t+p} y_t,$$

де y_t — фактичне значення рівня в i-й момент;
i — порядковий номер рівня в інтервалі.

При парному m середина інтервалу знаходиться між двома часовими точками і тоді проводиться додаткова процедура *центрування* (усереднення кожної пари значень). Так, за допомогою ковзної середньої згладимо ряд динаміки виплат страхового відшкодування (рис. 4.1). У вікні *Time series transformations* виберемо опцію *N-pts mov.averg.* $N = 4$, а оскільки m — парне число, то слід передбачити процедуру центрування — *Prior*. На рис. 4.3 ковзну середню представлено ламаною лінією *VAR2tmsfrmd(L)*, амплітуда коливань якої значно менша порівняно з рядом первинних даних.

Ковзна середня з однаковими вагами a_r при згладжуванні динамічного ряду погашає не лише випадкові, а й властиві конкретному процесу періодичні коливання. Припускаючи наявність таких коливань, використовують зважену ковзну середню, тобто кожному рівню в межах інтервалу згладжування надають певну вагу. Способи формування вагової функції різні. В одних випадках ваги відповідають членам розкладання бінома $\left(\frac{1}{2}+\frac{1}{2}\right)^{2p}$, при $m = 3$, скажімо, $a_r = 1/4, 1/2, 1/4$. В інших випадках до даних інтервалу згладжування добирається певний поліном, наприклад, парабола $\bar{y}_t = a + b_t + c_t^2$, де $i = -p, \ldots, p$. Тоді вагова функція така:

для $m = 5$ $\quad a_r = \frac{1}{35}(-3, 12, 17, 12, -3)$;

для $m = 7$ $\quad a_r = \frac{1}{21}(-2, 3, 6, 7, 6, 3, -2)$ і т. д.

Як видно з формул, ваги симетричні відносно центра інтервалу згладжування, сума їх з урахуванням винесеного за дужки множника дорівнює $\sum a_r = 1$.

Основна перевага ковзної середньої — наочність і простота тлумачення тенденції. Проте не слід забувати, що ряд ковзних середніх коротший за первинний ряд на $2p$ рівнів, а отже, втрачається інформація про крайні члени ряду. І чим ширший інтервал згладжування, тим відчутніші втрати, особливо нової інформації. Окрім того, маючи спільну основу розрахунку, ковзні середні виявляються залежними, що при згладжуванні значних коливань навіть за відсутності циклів у первинному ряду може вказувати на циклічність процесу (ефект Слуцького).

У симетричних фільтрах стара і нова інформація рівновагомі, а при прогнозуванні важливішою є нова інформація. У такому разі використовують асиметричні фільтри. Найпростіший з них — ковзна середня, яка замінює не центральний, а останній член ряду (*адаптивна середня*):

$$\bar{y}_t = \bar{y}_{t-1} + \frac{y_t - y_{t-m}}{m}.$$

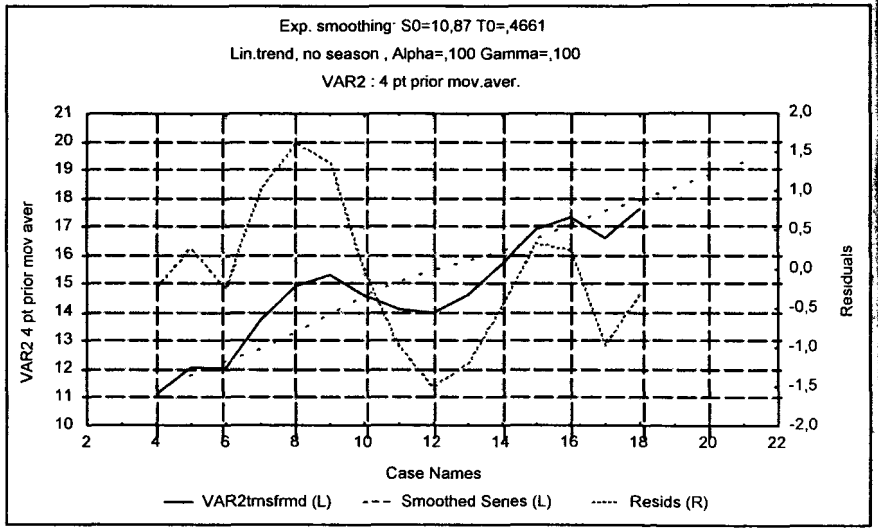


Рис. 4.3. Ковзна та експоненційна середні виплат страхового відшкодування

У наведеній формулі перший елемент характеризує інерцію розвитку, другий — адаптує середню до нових умов. Таким чином, середня $\bar{y}_t$ з кожним кроком ніби оновлюється. Ступінь оновлення визначається постійною вагою $\frac{1}{m}$. При використанні зважених асиметричних фільтрів вагова функція формується з урахуванням ступеня новизни інформації. Такою є середня з експоненційно розподіленими вагами:

$$Y_t = \sum_{r=0}^{t} a(1-a)^r y_{t-r},$$

де Y_t — *експоненційна середня*, тобто згладжене значення рівня динамічного ряду на момент t; $a(1-a)^r$ — вага $(t-r)$-го рівня; a — параметр згладжування, який визначає вагу t-го рівня, значення його коливаються в межах від 0 до 1.

Розклавши формулу за елементами суми, маємо

$$Y_t = ay_t + a(1-a)y_{t-1} + a(1-a)^2 y_{t-2} + \ldots + a(1-a)^t y_0,$$

або

$$Y_t = ay_t + (1-a)\left[ay_{t-1} + a(1-a)y_{t-2} + \ldots + a(1-a)^{t-1} y_0\right].$$

Друга складова останньої формули є не що інше, як експоненційна середня для $(t-1)$-го моменту. Отже, експоненційну середню можна представити як лінійну комбінацію фактичного рівня t-го моменту та експоненційної середньої $(t-1)$-го моменту:

$$Y_t = ay_t + (1-a)Y_{t-1}.$$

Чим віддаленіший від t-го моменту рівень ряду, тим менша його відносна вага і вклад у тенденцію. Так, при $a = 0{,}2$ ваги становлять: для t-го моменту — 0,2, для $(t-1)$-го моменту — $0{,}2(1-0{,}2) = 0{,}16$; для $(t-2)$-го моменту — $0{,}2(1-0{,}2)^2 = 0{,}128$ і т. д. Надаючи більшу вагу новій інформації, експоненційна середня адаптується до нових умов, що робить її досить ефективним і надійним методом *короткострокового прогнозування*.

Для розрахунку експоненційної середньої Y_t необхідно визначити початкові умови: початкову величину Y_0 і параметр a. Як початкову величину можна використати середній рівень за минулий (до динамічного ряду) період, або за відсутності таких даних, перший рівень ряду, тобто $Y_0 = y_1$. Щодо параметра a, то на практиці найчастіше використовують його значення в інтервалі від 0,1 до 0,3. Оскільки від параметра a залежить сума вагових коефіцієнтів $\sum a_r$ на певному часовому інтервалі m, то можна за наперед заданим значенням цих величин орієнтовно визначити параметр a:

$$a = 1 - \sqrt[m]{1 - \sum_{1}^{m} a_r}\,.$$

Наприклад, якщо часовий інтервал $m = 10$ місяців, а сума ва $\sum a_r = 0{,}90$, то $a = 1 - \sqrt[10]{1 - 0{,}9} \approx 0{,}2$. Тобто, при $a = 0{,}2$ десять чле нів динамічного ряду визначать 90% величини експоненційно середньої.

У моніторингу валютного ринку використовують 12-денні і 26-денні експоненційні середні курсових цін закриття з парамет рами згладжування відповідно 0,15 і 0,075. Вони розглядаютьс як швидка і повільна лінії тренда (лінії підтримки та опору). Зна чне відхилення між цими середніми свідчить про силу тренда, а перетинання дає сигнал про можливі його зміни. Якщо швидка середня перетинає повільну зверху, це сигналізує про народжен ня нового спадного тренда, якщо знизу — про народження зрос таючого тренда.

При прогнозуванні процесу вдаються до багаторазового згла джування. Якщо період упередження $v = 1$, то використовуют подвійне згладжування. Експоненційна середня другого порядку Y_t^* визначається за такою ж самою рекурентною формулою на основі згладженого ряду Y_t:

$$Y_t^* = aY_t + (1-a)Y_{t-1}^*.$$

Якщо припустити наявність лінійного тренда, прогнозний рі вень Y_{t+1} можна розрахувати за формулою

$$Y_{t+1} = \frac{(2-a)Y_t - Y_t^*}{1-a}.$$

Для певних значень параметра a ця формула набуває вигляду:

- для $a = 0{,}1$ $Y_{t+1} = 2{,}111Y_t - 1{,}111Y_t^*$;
- для $a = 0{,}2$ $Y_{t+1} = 2{,}250Y_t - 1{,}250Y_t^*$;
- для $a = 0{,}3$ $Y_{t+1} = 2{,}429Y_t - 1{,}429Y_t^*$.

Довірчі межі прогнозного рівня визначаються традиційно:

$$Y_{t+1} \pm t_{1-\alpha}\sigma_y\sqrt{1 + \frac{a}{2-a}},$$

де $\sigma_y = \sqrt{\frac{\sum(y_t - \bar{y})^2}{n-1}}$ — дисперсія рівнів первинного динамічного ряду; t — квантиль розподілу Стьюдента для ймовірності $(1 - \alpha)$.

Очевидно, що за умови значної варіації рівнів динамічного ряду довірчі межі будуть досить широкими.

Проведемо експоненційне згладжування ряду динаміки виплат страхового відшкодування. На стартовій панелі модуля *Time Series Analysis* ініціюємо кнопку процедури *Exponential Smoothing & Forecasting*. У вікні *Seasonal end non-seasonal exponential smoothing* вибираємо установки без сезонної компоненти — *None*, зазначаємо вид тренда — *Linear trend* і період упередження *Forecast* — 3. Результати експоненційного згладжування ряду ілюструються на рис. 4.3. Експоненційна середня позначена пунктирною лінією *Smoothed series (L)*. За межами динамічного ряду представлені прогнозні рівні для періоду упередження v:

$v = 1$	$v = 2$	$v = 3$
19,1	19,6	20,0

Базову модель експоненційного згладжування можна використати при моделюванні рядів, які мають сезонну компоненту.

4.4. ОЦІНЮВАННЯ СЕЗОННОЇ КОМПОНЕНТИ

Сезонні коливання формуються під впливом не лише природно-кліматичних, але й соціально-економічних факторів. Сила і напрям дії окремих факторів формує різну конфігурацію сезонної хвилі. За своїм характером сезонна компонента може бути адитивною або мультиплікативною. Для адитивної компоненти характерні сталі коливання навколо середнього рівня чи тренда, для мультиплікативної — зростання амплітуди коливань з часом.

Кожний рівень ряду y_t належить до певного сезонного циклу s, довжина якого становить 12 місяців, або 4 квартали. Відношення y_t до середнього рівня за цикл називається *індексом сезонності*:

$$I_s = \frac{y_t}{\bar{y}}.$$

За умови, що вплив несезонних факторів елімінований, середня з індексів j-го циклу становить 1, або 100 %.

У нестаціонарних рядах замість середньої використовують лінію тренда $Y_t = y(t)$, яка плавно проходить через ряд динаміки і, як і середня, елімінує його нерівномірності. Сукупність індексів сезонності в межах циклу характеризує сезонний ритм.

Прогнозування сезонних процесів ґрунтується на декомпозиції динамічного ряду. Припускають, що у майбутньому збережеться тенденція і такий же характер коливань. За таких умов прогноз на будь-який місяць (квартал), визначений методом екстраполяції тренда, коригується індексом сезонності: $Y^*_{t+v} = I_t \cdot Y_{t+v}$, де v — період упередження. Скажімо, поквартальна динаміка обсягів імпорту пального (тис. т) за два роки ($n = 8$, $t_1 = -3{,}5$, $t_n = 3{,}5$) описується трендом $Y_t = 923{,}7 + 33{,}8t$, за яким теоретичний обсяг імпорту у восьмому кварталі становить 1042,0 тис. т, а в 1-му кварталі наступного року ($v = 1$) передбачається $Y_{t+v} = 1042{,}0 + + 33{,}8 \cdot 1 = 1075{,}8$. Якщо середній індекс сезонності 1-го кварталу $I_1 = 1{,}34$, то скоригований на сезонність прогнозний рівень дорівнює $Y^*_{t+1} = 1{,}34 \cdot 1075{,}8 = 1441{,}6$ тис. т.

Динаміка більшості показників не виявляє чітко вираженої тенденції розвитку. Через постійний перерозподіл впливу факторів, які формують динаміку процесу, змінюється інтенсивність динаміки, частота та амплітуда коливань. До таких фактичних даних більш еластичною виявляється ковзна середня, інтервал згладжування якої дорівнює сезонному циклу (4 або 12). Коригування ковзної середньої на сезонність здійснюється так само, як коригування лінійного тренда.

На використанні експоненційної середньої *ґрунтується сезонно-декомпозиційна модель Холта-Вінтера*, в якій поєднуються моделі стаціонарності, лінійності та сезонності. Послідовність операцій така:

1. Визначаються індекси сезонності I_t.

2. Ряд динаміки фільтрується від сезонних коливань діленням y_t на коефіцієнт сезонності з лагом s; ряд $u_t = y_t : I_{t-s}$ називається декомпозиційним.

3. Перші різниці декомпозиційного ряду $b_t = (u_t - u_{t-1})$ розглядаються як характеристики лінійного тренда.

Кожна з компонент моделі згладжується за допомогою експоненційної середньої. При комбінації лінійної та сезонно-адитивної моделей тренда:

$$u_t = A\frac{y_t}{I_{t-s}} + (1-A)(u_{t-1} + b_{t-1});$$

$$b_t = B(u_t - u_{t-1}) + (1-B)b_{t-1};$$

$$I_t = C\frac{y_t}{u_t} + (1-C)I_{t-s}.$$

Значення параметрів згладжування A (*Alpha*), B (*Delta*) і C (*Gamma*) в системі Statistica за умовчування визначаються на рівні 0,1, в [9] рекомендуються: $A = 0{,}2$; $B = 0{,}2$; $C = 0{,}5$.

За умови ізольованої оцінки трьох факторів прогноз на період упередження v визначається як скоригована на сезонність сума прогнозного рівня u_t і лінійного тренда:

$$Y^*_{t+v} = (u_t + b_t v)I_{t-s+v}.$$

При комбінації лінійного та сезонно-мультиплікативного трендів кінцевий прогноз визначається за формулою

$$Y^*_{t+v} = u_t(1+b_t)^v I_{t-s+v},$$

де $b_t = B\frac{u_t - u_{t-1}}{u_{t-1}} + (1-B)b_{t-1}$.

За наявності періодичних коливань ряду помісячної динаміки використовують також моделі сезонної хвилі на основі *гармонійного аналізу*. Основними її характеристиками є: амплітуда, фаза, період і частота коливань.

Амплітуда A характеризує відстань від середнього рівня максимуму (мінімуму) сезонної хвилі, *період коливань T* — тривалість циклу, *частота f* — кількість циклів в одиницю часу, тобто $f = 1/T$. Якщо $T = 12$ місяців, то $f = 1/12$ циклу в місяць. Відстань між початком відліку часу з точкою $t = 0$ і найближчим піком називають *фазою* Θ. Сезонну хвилю з періодом T можна описати функцією:

$$Y = a + b\cos\omega t + d\sin\omega t,$$

де ω — кутова частота гармоніки; вимірюється радіанами в одиницю часу $\omega = 2\pi f = 2\pi/T$ і змінюється в інтервалі $0 \le \omega \le 2\pi$;

b, d — коефіцієнти гармоніки, функціонально зв'язані з амплітудою: $A = \sqrt{b^2 + d^2}$.

Коефіцієнти гармоніки визначаються методом найменших квадратів. Завдяки властивостям ортогональності функцій синуса і косинуса система нормальних рівнянь приводиться до тотожностей

$$\sum y = an$$

$$\sum y\cos\omega t = \tfrac{1}{2}nb$$

$$\sum y\sin\omega t = \tfrac{1}{2}nd$$

Звідси: для $n = 12$ $a = \frac{\sum y}{12}$; $b = \frac{\sum y\cos\omega t}{6}$; $d = \frac{\sum y\sin\omega t}{6}$.

Отже, a — це не що інше, як середньомісячний рівень ряду. Коефіцієнти b і d визначають амплітуду коливань навколо середнього рівня.

Очевидно, що чим більша амплітуда коливань, тим вагоміший вклад гармоніки в загальну дисперсію процесу. Оцінкою такого вкладу слугує дисперсійне відношення

$$R^2 = \frac{\delta^2}{\sigma^2},$$

де $\delta^2 = 0{,}5A^2$ — дисперсія гармоніки,

$\sigma^2 = \frac{\sum(y-\hat{y})^2}{n}$ — загальна дисперсія процесу.

У модель гармонійного аналізу можна включити декілька гармонік з різними періодами коливань. Скажімо, перша гармоніка з періодом 12, друга — з періодом 6, третя — з періодом 4 і т. д.

Гармонійна функція розкладає часовий ряд на правильні періодичні хвилі — синусоїди. Адекватність її реальному процесу залежить від того, наскільки сталими є частота й амплітуда коливань. Відносно сталий характер внутрішньорічної динаміки притаманний ринку сезонних товарів. У табл. 4.4 наведено помісячну динаміку середньої ціни свіжих огірків (грн.).

Таблиця 4.4

Місяць	Ціна, грн.	Місяць	Ціна, грн.	Місяць	Ціна, грн.
1	5,56	5	2,60	9	0,76
2	5,70	6	1,38	10	1,43
3	4,72	7	0,70	11	4,36
4	3,68	8	0,57	12	5,89

Для побудови гармонійної функції на стартовій панелі модуля ініціюємо кнопку *Spectral (Fourier) analysis* — Спектральний аналіз. У діалоговому вікні *Fourier (Spectral) analysis* виберемо опцію *Single series Fourier analysis* — Аналіз Фур'є одиничного ряду. Визначені за опцією *Summary* коефіцієнти гармонік наведено в табл. 4.5.

Таблиця 4.5

Spectral analysis: VAR3 (_____.sta)				
No. Of cases: 12				
Continue...	Frequncy	Period	Cosine Coeffs	Sine Coeffs
0	0		–3,7E-17	–0
1	0,083	12	2,493	0,055
2	0,167	6	0,098	–0,678
3	0,25	4	–0,328	–0,456
4	0,333	3	–0,455	–0,037
5	0,417	2,4	–0,289	0,060
6	0,5	2	–0,176	–0

Найвагомішою виявилася перша гармоніка, яка з амплітудою $A = \sqrt{2{,}493^2 + 0{,}055^2} = 2{,}494$ пояснює 76,5% варіації ряду. Друга гармоніка пояснює 5,8% варіації ряду. Внесок решти гармонік — 17,7%.

Визначену амплітуду коливань можна використати при прогнозуванні сезонного процесу.

4.5. МОДЕЛЬ *ARIMA*

Внутрішня структура динамічного ряду, залежність рівня y_t від попередніх його значень $y_{t-1}, y_{t-2}, \ldots, y_{t-p}$ описується *авторегресійною функцією*:

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \ldots + a_p y_{t-p} + e_t,$$

де p — порядок авторегресії; a_p — коефіцієнт авторегресії.

Процес авторегресії порядку p функціонально зв'язаний з автокореляційною функцією

$$r_p = a_1 r_{p-1} + a_2 r_{p-2} + \ldots + a_p,$$

де $p = 1, 2, \ldots, m$ — лаг автокореляції (зсування y_t на p значень назад); $r_0 = 1$.

Згідно з цим співвідношенням єдиний коефіцієнт авторегресії першого порядку $y_t = a_1 y_{t-1} + e_t$ дорівнює коефіцієнту автокореля-

ції першого порядку, тобто $a_1 = r_1$. Для авторегресії другого порядку $y_t = a_1 y_{t-1} + a_2 y_{t-2} + e_t$ маємо систему рівнянь

$$r_1 = a_1 + a_2 r_1$$
$$r_2 = a_1 r_1 + a_2.$$

Звідси

$$a_1 = -\frac{r_1(1-r_2)}{1-r_1^2},\ a_2 = -\frac{r_2 - r_1^2}{1-r_1^2}.$$

Отже, коефіцієнт авторегресії, як і коефіцієнт автокореляції, змінюється в межах від –1 до +1.

При моделюванні нестаціонарних за своєю природою економічних процесів авторегресійна функція об'єднується з іншими методами аналізу динаміки: ковзною (експоненційною) середньою, трендом, сезонною хвилею. Об'єднання різних моделей в єдине ціле суттєво розширює сферу практичного їх використання. Окрім того, об'єднані моделі формуються на основі одних і тих же статистичних характеристик — автокореляційних функцій, розробляється один алгоритм розрахунку параметрів моделі і визначення прогнозів. Моделі такого класу називають *об'єднаними (інтегрованими) моделями авторегресії — ковзної середньої* або скорочено *ARIMA*. Методику їх побудови грунтовно викладено в [2, розділи 4, 9].

У моделі *ARIMA* рівень динамічного ряду y_t визначається як зважена сума попередніх його значень і значень залишків e_t — поточних і попередніх. Вона об'єднує модель авторегресії порядку p і модель ковзної середньої залишків порядку q. Тренд включається в *ARIMA* за допомогою оператора кінцевих різниць ряду y_t. У модулі *Time Series/Forecasting* для цього передбачено процедуру трансформації *Differencing* ($x = x - x(\text{lag})$). Так, для фільтрації лінійного тренда використовують різниці першого порядку $d_1 = y_t - y_{t-1}$ (лаг = 1), для фільтрації параболічного тренда — різниці другого порядку і т. д. Різниця d має бути стаціонарною.

Вид моделі *ARIMA*, адекватність її реальному процесу та прогнозні властивості залежать від порядку авторегресії p і порядку ковзної середньої q. Через те ключовим моментом моделювання вважається процедура *ідентифікації* — обґрунтування виду моделі. В стандартній методиці *ARIMA* [2] ідентифікація зводиться до візуального аналізу автокорелограм і грунтується на принципі економії, за яким $(p + q) \le 2$.

Модель *ARIMA* порядку (p, d, q) досить гнучка і описує широкий спектр несезонних процесів. За наявності сезонних коливань

у моделі враховується їх періодичність з лагом s (для квартальних даних $s = 4$, для помісячних $s = 12$) і аналогічного змісту параметрами $(P, D, Q)_s$. Порядок мультиплікативної *ARIMA* становить $(p, d, q) \cdot (P, D, Q)_s$. Для ідентифікації моделі у діалоговому вікні *Single Series ARIMA* передбачено спеціальну групу опцій *Arima model parameters* — Параметри *ARIMA*:

p — Autoregressive — параметр авторегресії (регулярний);
P — Seasonal — сезонний параметр авторегресії;
q — Moving average — параметр ковзної середньої (регулярний);
Q — Seasonal — сезонний параметр ковзної середньої.

Необхідно вказати принаймні один із зазначених параметрів. Найпростіші види моделей *ARIMA*:

(1, 0, 0) — авторегресійна функція;
(0, 0, 1) — ковзна середня;
(1, 0, 1) — комбінована модель авторегресії і ковзної середньої;
(0, 1, 1) — експоненційна середня;
(1, 1, 1) — нестаціонарний процес з лінійним трендом;
(0, 1, 1) · (0, 1, 1) — мультиплікативна модель сезонного процесу.

Практична реалізація моделей можлива лише на рядах довжиною не менше 50 спостережень. Проілюструємо види моделей *ARIMA* на класичних прикладах [2, додатки: G, B'].

Перший стосується помісячної (n = 131) динаміки перевезень авіапасажирів (у тис.). До первинного ряду застосовано трансформацію *Natural Log*, а також *Difference* з лагом 1. Враховуючи сезонність авіаперевезень, зазначимо сезонний лаг 12. Порядок моделі $(0, 1, 1) \cdot (0, 1, 1)_{12}$. Параметри моделі оцінюються методом максимальної правдоподібності, необхідно лише в нижньому лівому куті діалогового вікна *Single series ARIMA* задати обчислювальну процедуру: *Approximate* — наближена чи *Exact* — точна. За командою *Begin parameter estimation* здійснюється ітераційна процедура визначення параметрів моделі і за умови їх прийнятності через команду *OK* відкривається вікно результатів оцінювання:

```
Single Series ARIMA Results
   Variable: SERIES_G: Monthly passenger totals (in 1000's)
Transformations: ln(x),D(1),D(12)
   Model: (0,1,1)(0,1,1) Seasonal lag: 12
No.of obs.: 131 Initial SS= .273 Final SS= .183(66,95%) MS = .0014
Parameters (p/Ps-Autoregressive, q/Qs-Moving aver.); p < .05
              q(1)      Qs(1)
Estimate:     .40182    .55694
Std. Err.:    .09069    .07395
```

Ініціювавши кнопку *Parameters estimates*, результати аналізу можна отримати у вигляді таблиці. Наприклад, 434-денна динаміка курсу акцій компанії IBM описується моделлю *ARIMA* порядку (0, 1, 1). Значення параметра q(1), його асимптотична стандартна похибка, *t*-критерій та *p-level* — фактичний рівень істотності, наведені в табл. 4.6, свідчать про адекватність моделі.

Таблиця 4.6

Input: VAR1(ibm1.sta)				
Continue...	Transformations: D(1) Model:(0,1,1) MS Residual=338.95			
Param.	Param.	Asympt. Std.Err.	Asympt. t (432)	p
q(1)	.60668	.03671	16.52	0.00

Для визначення прогнозів необхідно ініціювати кнопку *Forecast cases* — Прогнозні спостереження. В табл. 4.7 наведено прогнозні рівні курсу акцій IBM з 90%-ми довірчими межами на період упередження $v = 3$.

Таблиця 4.7

Forecasts: Model(0,1,1) Seasonal lag: 12 (ibm1.sta)			
Continue...	Input: VAR1 Start of origin: 1 End of origin:434		
Case No.	Forecast	Lover 90,0%	Upper 90,0%
435	351,5	321,1	381,8
436	351,5	318,9	384,1
437	351,5	316,7	386,2

Як видно з даних таблиці, точковий прогноз на період упередження не змінюється, проте довірчі межі його розширюються.

Для візуалізації результатів моделювання і прогнозування у діалоговому вікні *Single Series ARIMA Results* передбачено опції *Review and plot variables*.

Якщо характер динаміки стрімко змінюється під впливом зовнішніх факторів, то до такого ряду застосовують модель *Interrupted ARIMA* — Перервана *ARIMA*.

4.6. МОДЕЛЮВАННЯ ПОВНИХ ЦИКЛІВ

Свої особливості має моделювання динамічних процесів з ефектом насичення, коли темпи зростання (зниження) уповільнюються і рівень наближується до певної межі (питомі витрати ресурсів, споживання продуктів харчування на душу населення тощо). Для їх описування використовують клас кривих, що мають горизонтальну асимптоту $K \neq 0$. Найпростішою з-поміж них є модифікована експонента:

$$Y_t = K + ab^t,$$

де параметр a — різниця між ординатою Y_t при $t = 0$ та асимптотою K. Якщо $a < 0$, асимптота знаходиться вище кривої, якщо $a > 0$ — асимптота нижче кривої. Параметр b характеризує співвідношення послідовних приростів ординати. За умови рівномірного розподілу ординати по осі часу ці співвідношення є сталими: $b = \dfrac{Y_{t+1} - Y_t}{Y_t - Y_{t-1}} = const$.

Модифікована експонента описує процеси, на які діє певний обмежувальний фактор, і вплив цього фактора зростає зі зростанням Y_t. У разі, коли обмежувальний фактор впливає лише після певного моменту, до якого процес розвивався за експоненційним законом, то такий процес найкраще апроксимується *S*-подібною функцією з точкою перегину P, в якій прискорене зростання змінюється уповільненням. Наприклад, попит на новий товар попервах незначний; потім, після визнання споживачами, він стрімко зростає, але у міру насичення ринку темпи зростання уповільнюються, згасають. Попит стабілізується на певному рівні. Аналогічні фази розвитку мають процеси нововведень і винаходів, ефективність використання ресурсів тощо. З-поміж *S*-подібних кривих, що описують повний цикл розвитку, найпоширенішою є функція Перла-Ріда — *логістична крива*:

$$Y_t = \frac{K}{1 + be^{-at}}.$$

Якщо показник процесу — частка, що змінюється в межах від 0 до 1, то формула логістичної функції спрощується:

$$Y_t = \frac{1}{e^{a+bt} + 1} \text{ або } \frac{1}{Y_t} = 1 + e^{a+bt}.$$

У страховій і демографічній статистиці використовують інш*S*-подібну функцію — *криву Гомперца*:

$$Y_t = Ka^{b^t} \text{ або в логарифмах } \lg Y_t = \lg K + b^t \lg a .$$

Тобто крива Гомперца приводиться до модифікованої експоненти, у якої сталими є відношення приростів ординат у логарифмах.

Оцінювання параметрів функцій, які мають асимптоти, порівняно з поліномами та експонентами значно складніше. Тут можливі два варіанти.

За першим варіантом асимптота у вигляді нормативу, стандарту тощо визначається апріорі — K^*. Тоді модифіковану експоненту можна представити так:

$$\left(Y_t - K^*\right) = ab^t .$$

Замінивши $\left(Y_t - K^*\right)$ на z і прологарифмувавши рівняння, дістанемо лінійну функцію логарифмів $\lg z = \lg a + t\lg b$. Аналогічно приводиться до лінійного виду логістична функція $1/Y_t = K^* + ab^t$, яка при заміні $\left(1/Y_t - K^*\right)$ на z у логарифмах набуває такого ж вигляду: $\lg z = \lg a + t\lg b$. Параметри приведених до лінійного виду функцій, як і параметри поліномів, можна оцінити методом найменших квадратів, використовуючи процедури модуля *Multiple Regression* (див. 4.2). Прогноз та його довірчі межі визначаються традиційно, хоча довірчі межі прогнозу за кривими повного циклу мають умовний характер.

За другим варіантом асимптота невідома, отже, необхідно визначити усі три параметри: K, a, b. У літературі для кожної кривої запропоновано різні процедури, що реалізують МНК. Оскільки логістична крива і крива Гомперца приводяться до модифікованої експоненти, то доцільно розглянути універсальний для трьох функцій метод, описаний Бріантом [9]. За цим методом спершу визначається параметр b, а потім a та K. Формули розрахунку параметрів модифікованої експоненти такі:

$$b = \frac{(n-1)\sum_1^{n-1} y_t y_{t+1} - \sum_1^{n-1} y_t \sum_1^{n-1} y_{t+1}}{(n-1)\sum_1^{n-1} y_t^2 - \left(\sum_1^{n-1} y_t\right)^2};$$

$$a = \frac{n\sum_1^n b^t y_t - \sum_1^n b^t \sum_1^n y_t}{n\sum_1^n b^{2t} - \left(\sum_1^n b^t\right)^2};$$

$$K = \frac{\sum_1^n y_t - a\sum_1^n b^t}{n}.$$

У системі *Statistica* розрахунок параметрів *S*-подібних кривих можна здійснити в модулі *Nonlinear Estimation* — Нелінійне оцінювання, скориставшись процедурою *User-specified regression*, яка передбачає визначення виду функції користувачем самостійно. Приміром, застосуємо логістичну криву до даних ряду динаміки населення мегаполіса (табл. 4.8).

Таблиця 4.8

Рік	Млн чол	Рік	Млн чол	Рік	Млн чол
1950	3,48	1970	4,78	1990	6,14
1955	3,86	1975	5,13	1995	6,37
1960	4,17	1980	5,52	2000	7,04
1965	4,56	1985	5,90		

У діалоговому вікні *Estimated function & loss function* задамо вид функціонального виду кривої: $v2 = b1/(1 + b2 \cdot \exp(-b3 \cdot v1))$. Параметри її означають: $b1 = K$, $b2 = a$, $b3 = b$. Щодо функції втрат, то можна обмежитися залишковою девіатою, яка визначається системою за умовчування. Через кнопку *Variables* ідентифікуємо ознаку, динаміка якої моделюється (у даному прикладі — $v2$), і метод оцінювання параметрів моделі (*Quasi-Newton*). По закінченні ітераційної процедури оцінювання параметрів за командою *OK* відкривається вікно *Results*. Значення індексу кореляції $R = 0{,}997$ свідчить про високу апроксимуючу властивість моделі. Оцінки параметрів — *Parameter estimates* представлені в табл. 4.9.

Таблиця 4.9

Model: v2 = b1 / (1 + b2*exp(–b3*v1)) (_____.sta)			
Final loss: ,0723 R=,997 Variance explained: 99,428%			
	B1	B2	B3
Estimate	12,37	2,79	0,114

Згідно з даними приріст населення мегаполісу за п'ятиріччя становить в середньому 11,4%, наближаючись до межі — 12,37 млн. чол.

Отже, клас моделей динаміки досить широкий, і вони описують різні процеси розвитку. Вибір типу моделі у конкретному дослідженні ґрунтується передусім на теоретичному аналізі специфіки процесу, його внутрішньої структури, взаємозв'язків з іншими процесами. На основі такого аналізу в загальних рисах визначається характер динаміки (рівномірний, рівноприскорений, з насиченням тощо) та окреслюється коло функцій, здатних апроксимувати цей процес. Серйозною підмогою при виборі конкретної моделі слугують формальні методи. Скажімо, для поліномів — це аналіз послідовних різниць. Рівність різниць p-го порядку розглядається як симптом того, що процес описується поліномом p-го порядку. Якщо приблизно однакові різниці 1-го порядку $\Delta'_t = y_t - y_{t-1}$, використовують лінійний тренд, якщо однакові різниці 2-го порядку — $\Delta''_t = \Delta_t - \Delta_{t-1}$, — параболу і т. д. Певні складнощі можуть виникнути при виборі експоненти. Адже S-подібна крива до точки перегину описує експоненційний тренд, а сама точка перегину може бути за межами динамічного ряду. Отже, якщо межа насичення теоретично можлива і процес у майбутньому може згасати або існують певні обмеження для процесу (правові, матеріальних ресурсів, виробничих потужностей тощо), то перевага віддається S-подібній кривій.

Оскільки первинним рядам динаміки властива значна варіація рівнів y_t, то аналіз послідовних різниць більш коректно проводити на основі рядів ковзних середніх. У табл. 4.10 наведено основні характеристики такого аналізу (апріорні тести), за якими визначається конкретний тип моделі повного циклу.

Таблиця 4.10

Характеристика	Властивості характеристик	Тип трендової моделі
Δ'_t	Приблизно однакові	Поліном 1-го ступеня
Δ'_t	Лінійно змінюються	Поліном 2-го ступеня
Δ'_t / y_{t-1}	Приблизно однакові	Експонента
$\lg \Delta'_t$	Лінійно змінюються	Модифікована експонента
$\lg \Delta'_t / (y_{t-1})^2$	Лінійно змінюються	Логістична крива
$\lg \Delta'_t / y_{t-1}$	Лінійно змінюються	Крива Гомперца

При зворотному напрямку тенденції різниці розраховуються, починаючи з кінця. За наявності від'ємних різниць логарифмування неможливе, тому необхідно збільшити інтервал згладжування ковзних середніх.

Завдання для самоконтролю

1. Ситуація на ринку праці характеризується навантаженням незайнятого населення на одну вакансію. За даними семи місяців визначте прогнозний рівень цього показника на початок серпня та стандартне відхилення.

На початок місяця	Рівень навантаження, чол.	На початок місяця	Рівень навантаження, чол.
Січня	3,7	Травня	6,1
Лютого	4,2	Червня	6,9
Березня	4,8	Липня	7,5
Квітня	5,3	Серпня	?

Вибір функції тренда обґрунтуйте.

2. Динаміка перевезення вантажів залізницею (млн. т) описується трендовим рівнянням $Y = 14{,}9 + 0{,}9t$, де $t = 1, 2, \ldots, 7$, із стандартною похибкою 0,275. Визначте прогнозний обсяг перевезень вантажів на період упередження $v = 1$ та довірчі межі прогнозу з імовірністю 0,90.

3. Динаміка витрат компанії (млн. грн.) на модернізацію діючого устаткування за 1995—1999 рр. описується трендовим рівнянням $Y = 22{,}2 - 2{,}7t$. Оцініть автокореляцію залишкових величин з лагом 1. Зробіть висновки про адекватність лінійного тренда реальному процесу.

Роки	1995	1996	1997	1998	1999
Витрати	29	24	21	19	18

4. Динаміка експорту олії характеризується даними:

№ року	1	2	3	4	5	6	7	8	9
Тис. т	86	92	97	106	115	127	142	154	170

Опишіть тенденцію ряду експонентою, поясніть зміст параметрів. Визначте прогнозні рівні експорту олії на період упередження $v = 1, 2, 3$.

5. Динаміка біржових цін акцій компанії на торгах минулого тижня характеризується даними:

День	1	2	3	4	5
Ціна акції, грн.	23	19	20	22	18

Використовуючи метод експоненційного згладжування ($a = 0{,}2$), визначте прогнозний рівень біржової ціни акції на понеділок наступного тижня.

6. Сезонні коливання обсягів відпущеної теплоенергії в регіоні характеризується даними:

Місяць	тис. Гкал	Місяць	тис. Гкал.	Місяць	тис. Гкал.
1	69	5	14	9	11
2	62	6	8	10	29
3	60	7	8	11	58
4	31	8	9	12	72

Опишіть динаміку обсягів відпущеної теплоенергії моделлю гармонійного аналізу. Оцініть адекватність моделі.

7. Реалізація плодоовочевих консервів характеризується такими даними (тис. ум. банок):

Рік	Квартал			
	1	2	3	4
1	8,6	9,0	5,4	9,2
2	9,4	9,9	5,8	9,8
3	10,5	11,2	6,3	10,4
4	10,8	11,5	6,7	11,3
5	11,6	12,4	7,2	12,5

Визначте прогнозні рівні реалізації плодоовочевих консервів на кожний квартал наступного року, скоригувавши їх на сезонність.

8. Динаміку захворювання раком щитовидної залози у дітей (0—14 років), які проживають на забруднених радіонуклідами територіях, опишіть модифікованою експонентою:

Рік	Кількість хворих, чол.	Рік	Кількість хворих, чол.	Рік	Кількість хворих, чол.
1986	5	1990	20	1994	70
1987	7	1991	36	1995	75
1988	8	1992	50	1996	80
1989	10	1993	62	1997	78

Поясніть зміст параметрів моделі.

9. Динаміка витрат на кінцеве споживання по рахунку сектора загальнодержавного управління характеризується даними:

Рік	1	2	3	4	5	6	7	8
У % до ВВП	16,5	17,4	18,0	19,4	21.3	21,8	22,6	23,9

Опишіть динаміку витрат на кінцеве споживання логістичною кривою, поясніть зміст параметрів.

Розділ 5

ОСНОВИ МОДЕЛЮВАННЯ ВЗАЄМОЗВ'ЯЗКІВ

5.1. ТИПИ МОДЕЛЕЙ ВЗАЄМОЗВ'ЯЗКУ

Усі явища навколишнього світу взаємопов'язані й взаємозумовлені. У складному переплетенні всеохоплюючого взаємозв'язку будь-яке з них є наслідком дії певної множини причин і водночас причиною інших явищ.

Логічний зміст і практичну значущість статистичних моделей взаємозв'язку слід розглядати саме в площині співвідношення причинності і зв'язків, що вимірюються статистичними методами. Суть причинності полягає в породженні одного явища іншим. Причина — активна основа, що примушує інше явище змінюватися. Сама по собі причина не визначає наслідку. Останній залежить і від умов, у яких діє причина. Через нерозпізнаність причин і умов при моделюванні вони об'єднуються в одне поняття «фактор», а наслідок розглядається як результат дії факторів. Отже, в рамках моделі досліджується детермінованість результату факторами.

Методологічні проблеми побудови моделей взаємозв'язку можна об'єднати в дві групи:

- формування *ознакової множини* моделі, себто визначення кількості факторів та їх числових еквівалентів;
- *модельна специфікація* — вибір функціонального виду моделі, ідентифікація та оцінювання параметрів.

При формуванні ознакової множини моделі різноманітні прояви причинно-наслідкових зв'язків доцільно представляти візуально у вигляді спеціальних конструкцій — *графів зв'язку*, елементами яких є вершини та орієнтовані ребра (дуги). Вершини графа відповідають ознакам, а дуги показують відношення між ознаками. На рис. 5.1 ілюструється граф зв'язку чотирьох ознак. За дугами графа можна простежити систему відношень між ними: x впливає на y прямо, безпосередньо, z — прямо та опосередковано двома шляхами: $z \to x \to y$ та $z \to v \to y$. У такій логічній конструкції ознака y є результатом, а x, z і v — факторами, що визначають результат.

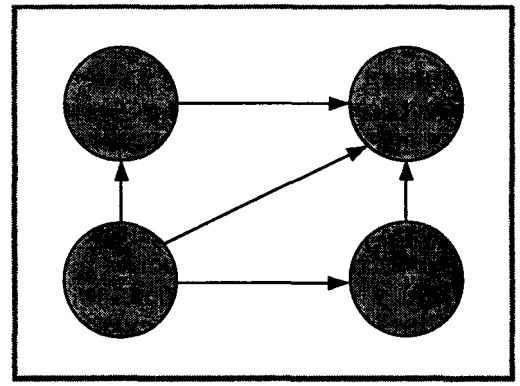


Рис. 5.1. Граф зв'язку

Граф відображує теоретично обґрунтовану систему відношень між ознаками. Кожна ланка цієї системи розглядається як окрема гіпотеза, що підлягає перевірці в подальшому аналізі на усіх етапах побудови моделі. Основна мета моделей взаємозв'язку — виявити і кількісно виміряти вплив факторів на результат. Очевидно, щоб визначити ефект впливу i-го фактора, необхідно елімінувати (усунути) вплив інших факторів, умовно зафіксувавши їх шляхом відповідних розрахунків на одному і тому ж рівні.

На етапі модельної специфікації враховується характер зв'язку та особливості наявної інформації. За своїм характером зв'язки поділяються на *стохастичні*, різновидом яких є *кореляційні* зв'язки, та жорстко детерміновані (*функціональні*). Перші відображують стохастичний характер причинно-наслідкових відношень, другі — адитивні чи мультиплікативні зв'язки між елементами розрахункових формул показників. Відповідно вибирається функціональна форма моделі: кореляційні зв'язки описуються переважно регресійними моделями, функціональні — балансовими або індексними.

У моделях, що описують функціональні зв'язки, ступінь вільності при формуванні ознакової множини обмежена, маневрувати можна лише кількістю факторів, укрупнюючи їх чи деталізуючи. Для регресійних моделей характерна багатоваріантність як ознакової множини, так і функціональної форми моделі.

Інформаційна база моделі залежить від того, як представлено об'єкт моделювання. Якщо він розглядається як сукупність елементів у просторі, то інформація подається просторовими рядами у вигляді матриці обсягом $(n \cdot m)$, де n — обсяг сукупності, m —

кількість включених у модель факторів. Класична регресія передбачає однорідність сукупності, тобто всі одиниці сукупності мають бути однотипними щодо комплексу умов існування, а властиві їм закономірності однаковими для усіх одиниць без винятку. Якщо сукупність внутрішньо диференційована, має у своєму складі певні групи (класи) одиниць зі специфічним характером зв'язку, в моделі слід врахувати неоднорідність за принципом структурної подібності. Методи відображення неоднорідності залежать від характеру та сталості міжгрупових розбіжностей.

Моделі, побудовані у просторовій площині, охоплюють одиничний, фіксований інтервал часу. Серія такого типу моделей за певний період дає можливість простежити динаміку взаємозв'язків, оцінити зміну потужності впливу окремих факторів, його перерозподіл.

Якщо об'єкт моделювання розглядається як первинний, неподільний елемент (галузь економіки, регіон, країна), то інформаційна база представляється багатовимірним динамічним рядом у вигляді матриці обсягом $(m \cdot T)$, де T — довжина динамічного ряду. В такому разі в моделі необхідно відобразити властиві процесу закономірності динаміки, як-от: тенденції, коливання, запізнення впливу тощо. За умови, що об'єкт моделювання нечисленний, а довжина динамічного ряду обмежена, просторові та динамічні ряди об'єднуються.

На практиці використовують переважно автономно побудовані моделі, тобто моделі одного показника-функції. Специфікація моделі залежить від її призначення, природи і структури взаємозв'язків, специфіки об'єкта моделювання, наявної інформації. Поєднання, комбінація усіх цих елементів визначає безліч типів моделей.

В автономних регресійних моделях (одного рівняння) відбувається складний процес елімінування впливів між включеними в модель факторами і виокремлення безпосереднього впливу кожного з них на результат. Фактичне використання такої моделі передбачає, що в разі необхідності рівні факторів можна змінювати незалежно один від одного. Проте в реальних умовах зміна одного фактора не може відбуватися за незмінності інших, вона спричиняє ланцюгову реакцію в усій системі взаємозв'язаних показників. Поряд з безпосереднім прямим впливом має місце опосередкований вплив, часом за різними напрямами, що потребує оцінювання сумарного впливу. Іноді одна й та сама змінна виступає водночас причиною і наслідком. Тоді виникає необхідність одночасного оцінювання прямого і зворотного впливів.

Складне переплетення взаємозв'язків соціально-економічних явищ потребує і складних інструментів аналізу. З-поміж таких інструментів є системи рівнянь, заміна множин висококорельованих ознак інтегральними факторами (головними компонентами) тощо. Методологічні засади модельної специфікації розглядаються за принципом «від простого до складного».

5.2. БАГАТОФАКТОРНІ ІНДЕКСНІ МОДЕЛІ

При вивченні функціональних зв'язків між показниками широко використовуються індексні моделі. Основою індексної моделі є мультиплікативний зв'язок між певною множиною показників; один з них розглядається як результат y, інші — як фактори x_i:

$$y = x_1 x_2 x_3 \dots x_n .$$

Послідовність факторів у моделі не може бути довільною, вона визначається економічним змістом показників і методикою їх розрахунку. Кожний наступний фактор-множник розраховується на одиницю попереднього, а отже, добуток будь-якої кількості факторів є економічно змістовною величиною. Наприклад, прибутковість активів компанії y є функцією прибутковості продажу продукції x_1 та оборотності мобільних активів z_1, тобто $y = x_1 z_1$. Оборотність мобільних активів z_1, в свою чергу, є функцією оборотності матеріальних запасів x_2 і частки матеріальних запасів у мобільних активах z_2. Отже, $y = x_1 x_2 z_2$.

Схематично послідовність розширення моделі можна представити так:

$$y = x_1 z_1 = x_1 x_2 z_2 = x_1 x_2 x_3 z_3 \text{ і т. д.}$$

Характерною рисою мультиплікативної моделі є взаємозв'язок факторів: чисельник розрахункової формули одного фактора є знаменником розрахункової формули наступного. Введення в ланцюгову схему нового фактора означає лише деталізацію функціонального зв'язку і не змінює його сутності. Ступінь деталізації залежить від мети дослідження.

При побудові індексної моделі функція $y = x_1 x_2 x_3 \dots x_m$ розглядається для двох періодів:

базисного $y_0 = x_{10} x_{20} x_{30} \dots x_{m0}$

і поточного $y_1 = x_{11} x_{21} x_{31} \dots x_{m1}$.

Абсолютну і відносну зміну показника-функції y можна розкласти за факторами-множниками x_i. Оцінювання ступеня та абсолютного розміру впливу кожного з них на динаміку функції здійснюється в рамках індексної моделі, в якій відтворюються взаємозв'язки між показниками:

$$Iy = Ix_1\, Ix_2\, Ix_3 \ldots Ix_m .$$

При розрахунку частинного індексу Ix_i необхідно елімінувати вплив інших включених у модель факторів. Задля цього всі фактори-множники, окрім x_i, фіксуються на постійному рівні. Найчастіше фактори, розміщені в ланцюгу зліва від x_i, фіксуються на рівні поточного періоду, а розміщені справа від x_i — на рівні базисного періоду. Скажімо, в моделі $y = x_1\, x_2\, x_3$ принцип послідовно-ланцюгового елімінування впливу фактора x_2 реалізується таким чином:

$$Ix_2 = (x_{11}\, x_{21}\, x_{30}) : (x_{11}\, x_{20}\, x_{30}).$$

За такою ж схемою визначається абсолютний вплив його на y:

$$A_2 = x_{11}\, (x_{21} - x_{20})\, x_{30}.$$

Абсолютний вплив факторів можна визначити з використанням відповідних частинних індексів. При послідовному множенні (за ланцюговою схемою) базисного рівня показника-функції на індекси факторів визначаються розрахункові рівні, тобто такі рівні, які мав би показник y під впливом i-го фактора і при незмінному рівні решти факторів. Якщо базисний його рівень позначити y_0, розрахунковий рівень для першого фактора — y', для другого — y'' і т. д., то порядок розрахунку абсолютного впливу i-го фактора A_i схематично можна представити так:

$$y_0 \rightarrow (y' = Ix_1 y_0) \rightarrow (y'' = Ix_2 y') \rightarrow (y''' = Ix_3 y'')$$
$$\underbrace{\qquad}_{A_1} \qquad \underbrace{\qquad}_{A_2} \qquad \underbrace{\qquad}_{A_3}$$

Методику побудови багатофакторної індексної моделі розглянемо на прикладі взаємозв'язку показника прибутковості капіталу з індикаторами фінансового стану та платоспроможності підприємства. Для окремої компанії (фірми, корпорації) прибутковість капіталу розраховується відношенням чистого прибутку до власного капіталу. Динаміку цього показника можна розкласти за такою множиною факторів:

a — чистий прибуток на одиницю валового обороту (реалізації продукції, послуг);

b — оборотність поточних активів;

c — поточна ліквідність;

d — частка поточних пасивів у залучених коштах (коефіцієнт заборгованості);

f — співвідношення залучених і власних коштів.

Взаємозв'язок між ними має вигляд:

$$\frac{\text{Чистий прибуток}}{\text{Власний капітал}} = \frac{\text{Чистий прибуток}}{\text{Валовий оборот}} \cdot \frac{\text{Валовий оборот}}{\text{Поточні активи}} \cdot \frac{\text{Поточні активи}}{\text{Поточні пасиви}} \cdot \frac{\text{Поточні пасиви}}{\text{Залучені кошти}} \cdot \frac{\text{Залучені кошти}}{\text{Власний капітал}} .$$

Наприклад, прибутковість капіталу умовної фірми становила: в базисному періоді — 115,1%, у поточному — 129,0%, тобто прибутковість зросла на 13,9 процентного пункту, індекс прибутковості — 1,121. Індекси включених у модель факторів-множників і розрахунок внеску кожного з них в абсолютний приріст прибутковості капіталу наведено в табл. 5.1.

Таблиця 5.1

Фактор	Індекс фактора	Розрахунковий рівень прибутковості	Абсолютний внесок фактора в приріст прибутковості
a	1,057	121,7	+6,6
b	0,986	120,0	–1,7
c	1,012	121,4	+1,4
d	1,025	124,4	+3,0
f	1,037	129,0	+4,6
Разом	X	X	+13,9

Абсолютний приріст прибутковості в розмірі 13,9 процентного пункту розкладено за факторами. Всі фактори, окрім оборотності поточних активів, мали позитивний вплив на динаміку прибутковості. З-поміж них найвагоміший вплив фактора a — чистого прибутку на одиницю валового обороту, на другому місці фактор f — співвідношення власних і залучених коштів, на третьому — фактор d — коефіцієнт заборгованості.

Систему взаємозв'язаних показників можна представити у матричному вигляді. На головній діагоналі матриці за певною стра-

тегією розміщуються m абсолютних величин q_i, на основі яких можна визначити $m(m-1)$ відносних величин $x_{ij} = \frac{q_i}{q_j}$, де $i \neq j$. Очевидно, що недіагональні елементи, симетрично розташовані щодо головної діагоналі, є оберненими одна до одної величинами, тобто $x_{ji} = \frac{1}{x_{ij}}$. Система взаємозв'язаних абсолютних і відносних величин утворює квадратну матрицю. Аналогічно складається матриця індексів.

У табл. 5.2 наведено індексно-матричну модель економічного розвитку умовної країни за певний період. На головній діагоналі розміщено індекси макропоказників (D — національний дохід, M — матеріальні витрати, F — виробничі фонди, T — чисельність зайнятих працівників). Вони ранжовані за економічною нормаллю, згідно з якою темпи зростання кінцевих результатів мають бути вищими за темпи зростання витрат і ресурсів, тобто

$$I_D > I_M > I_F > I_T .$$

Таблиця 5.2

Показник нормали	D	M	F	T
D	1,142			
M	$I_m = 1{,}005$	1,136		
F	$I_f = 0{,}935$	$I_n = 0{,}930$	1,222	
T	$I_q = 1{,}171$	$I_l = 1{,}165$	$I_r = 1{,}253$	0,975

За даними таблиці економічна нормаль порушена у двох ланках: $I_M < I_F$ та $I_D < I_F$. Значення індексів свідчать про фондоємкий трудозберігаючий тип відтворення. Піддіагональні елементи матриці — це результат бінарних відношень між індексами, на перетині яких знаходиться відповідний елемент. За змістом вони характеризують динаміку показників інтенсивності та ефективності економіки: I_q — продуктивності праці, I_f — фондовіддачі, I_m — матеріаловіддачі, I_r — фондоозброєності праці, I_n — співвідношення матеріальних витрат і вартості основних фондів. Аналізуючи співвідношення цих індексів, можна виявити диспропорції у використанні живої та уречевленої праці.

В індексно-матричній моделі ранжування показників і ступінь їх деталізації цілковито залежить від економічної стратегії та мети дослідження.

5.3. КЛАСИЧНА РЕГРЕСІЯ

Регресійна модель описує об'єктивно існуючі між явищами кореляційні зв'язки. За своїм характером кореляційні зв'язки надзвичайно складні та різноманітні. В одних випадках результат y зі зміною фактора x_i зростає чи зменшується рівномірно, в інших — нерівномірно. Іноді зростання може змінитися зменшенням і навпаки. Простежити всі ці взаємозв'язки і встановити точний функціональний вид практично неможливо. А тому при виборі типу функції йдеться лише про апроксимацію відносно простими функціями незрівнянно більш складних за своєю природою взаємозв'язків. На практиці перевагу віддають моделям, які є лінійними або приводяться до лінійного виду шляхом перетворення змінних, наприклад логарифмуванням. Такий підхід, безперечно, містить у собі певну умовність, оскільки передбачає однаковий характер зв'язку з усіма факторами. Проте використання надто складних функцій неминуче веде до збільшення кількості параметрів, а отже, зменшує точність вимірювання та ускладнює інтерпретацію результатів.

При обґрунтуванні типу функції слід враховувати й той факт, що межі варіації корельованих ознак у конкретних умовах простору і часу, в конкретній сукупності значно вужчі за їх можливі значення, і в цих межах варіації навіть лінійна функція може задовільно апроксимувати зв'язок.

У лінійному щодо параметрів рівнянні регресії індивідуальне значення результативного показника y_j (де j — порядковий номер одиниці сукупності) записується так:

$$y_j = b_0 + \sum_{i=1}^{m} b_i x_i + e_j ,$$

де b_0 — вільний член рівняння; економічного змісту, як правило, не має, лише окреслює область існування моделі;

b_i — *коефіцієнт регресії*; показує, як в середньому змінюється y зі зміною x_i на одиницю її шкали вимірювання за незмінності інших включених в модель факторів і за інших рівних умов;

$e_j = y_j - Y_j$ — *залишкова величина*.

У регресійній моделі основне навантаження покладається на коефіцієнт регресії b_i, він розглядається як своєрідна міра «очищеного» впливу x_i на y і називається *ефектом впливу*.

Процедура оцінювання параметрів регресійної моделі ґрунтується на *методі найменших квадратів* (МНК). Оскільки алгорит-

ми МНК описано в математико-статистичній літературі й реалізовано в комп'ютерних програмах, наведемо лише загальну схему розрахунку статистичних характеристик моделі, акцентуючи увагу на їх змістовній інтерпретації.

Первинна інформація представляється як матриця факторних ознак X розміром $(n \cdot m)$ і вектора результативної ознаки y розміром $(n \cdot 1)$. Задля зручності використання алгоритмів МНК матриця X розширюється за рахунок додатково введеної *фіктивної змінної* x_0, вектор якої представлений одиницями. Параметри моделі — вектор $B = |b_0, b_1, b_2, \ldots, b_m|$ визначаються розв'язуванням системи нормальних рівнянь, яка записується так:

$$X'XB = X'y, \text{ де } X'X \text{ — матриця розміром } n\,(m+1).$$

Послідовність розрахунків включає етапи:

- обчислення матриці $X'X$ і вектора $X'y$;
- обертання матриці $C = (X'X)^{-1}$;
- розрахунок параметрів $B = CX'y$;
- визначення *теоретичних значень результативної ознаки* $Y = \sum_0^m b_i x_i$ та залишків $e_j = y_j - Y_j$.

Значення коефіцієнтів регресії певною мірою залежать від складу введених у модель факторів. З розширенням ознакової множини моделі відбувається перерозподіл впливу попередньо введених факторів. Чим вагоміший вплив нововведеного фактора, тим помітніші зміни. Ілюстрацією перерозподілу впливу факторів може слугувати регресійна модель урожайності рису, ц/га [11]. У модель послідовно вводились агротехнічні фактори: x_1 — попередник, балів; x_2 — внесення добрив під основний обробіток, центнерів поживної речовини (ц п. р.) на 1 га посіву; x_3 — передпосівний обробіток, га м'якої оранки; x_4 — підживлення, ц п. р.; x_5 — норма висіву; x_6 — кількість прополювань. Відповідно отримано такі рівняння регресії:

1. $Y = 30{,}432 + 3{,}001x_1$;
2. $Y = 26{,}208 + 2{,}049x_1 + 5{,}995x_2$;
3. $Y = 21{,}563 + 1{,}970x_1 + 4{,}610x_2 + 2{,}906x_3$;
4. $Y = 22{,}332 + 1{,}321x_1 + 4{,}558x_2 + 1{,}465x_3 + 9{,}791x_4$;
5. $Y = 18{,}960 + 1{,}342x_1 + 4{,}483x_2 + 1{,}347x_3 + 9{,}545x_4 + 1{,}756x_5$;
6. $Y = 19{,}387 + 0{,}965x_1 + 3{,}400x_2 + 0{,}501x_3 + 7{,}500x_4 + 1{,}731x_5 + 3{,}433x_6$.

Як бачимо, введення кожного нового фактора спричиняє зменшення впливу попередньо введених факторів, таку ж тенденцію має й вільний член рівняння.

Оскільки факторні ознаки мають, як правило, різні одиниці вимірювання, то для порівняння ефектів їх впливу в рамках моделі використовують *стандартизовані коефіцієнти регресії* $\beta_i = b_i \frac{\sigma_{x_i}}{\sigma_y}$ (бета-коефіцієнти) або коефіцієнти еластичності $\gamma_i = b_i \frac{\bar{x}_i}{\bar{y}}$. *Бета-коефіцієнт* характеризує ефект впливу x_i на y в середньоквадратичних відхиленнях, *коефіцієнт еластичності* — в процентах. У табл. 5.2 наведено бета-коефіцієнти останнього (шостого) варіанта моделі врожайності рису. Згідно із значеннями β_i найвагоміший вплив на врожайність рису мають: прополювання ($\beta_6 = 0{,}360$), підживлення ($\beta_4 = 0{,}264$), внесення добрив під основний обробіток ($\beta_2 = 0{,}248$).

Для оцінювання *адекватності* регресійної моделі використовують:

- стандартне відхилення;
- множинні коефіцієнти детермінації та кореляції;
- частинні коефіцієнти детермінації та кореляції;
- коефіцієнти окремої детермінації;
- критерії перевірки істотності зв'язку.

Стандартне відхилення характеризує варіацію залишкових величин

$$s_e = \sqrt{\frac{\sum_1^n e^2}{n-(m+1)}},$$

де n — обсяг сукупності, m — кількість коефіцієнтів регресії.

Розрахунок характеристик щільності зв'язку ґрунтується на *декомпозиції (розкладанні) варіації y* за джерелами формування:

$$S_y^2 = S_Y^2 + S_e^2,$$

де $S_y^2 = \sum_1^n (y - \bar{y})^2$ — *загальна сума квадратів відхилень*, зумовлена впливом усіх можливих факторів;

$S_Y^2 = \sum_1^n (Y - \bar{y})^2$ — *факторна сума квадратів відхилень*, зумовлена впливом включених у модель факторних ознак x_i;

$S_e^2 = \sum_1^n (y - Y)^2$ — *залишкова сума квадратів відхилень*, розмір якої залежить від потужності впливу не включених у модель факторів.

Відношення факторної суми квадратів до загальної характеризує частку варіації y, пов'язану з варіацією включених у модель факторів, і називається *множинним коефіцієнтом детермінації*

$$R^2 = \frac{S_Y^2}{S_y^2} = \frac{S_y^2 - S_e^2}{S_y^2} = 1 - \frac{S_e^2}{S_y^2}.$$

За відсутності зв'язку $R^2 = 0$. Якщо зв'язок функціональний, то $R^2 = 1$. Очевидно, що R^2 пов'язаний із стандартним відхиленням s_e. При зменшенні s_e значення R^2 зростатиме і навпаки. Корінь квадратний із коефіцієнта детермінації називають *коефіцієнтом кореляції* $R = \sqrt{R^2}$. Для моделі врожайності рису $R = 0{,}8394$, $R^2 = 0{,}7029$, тобто 70,29% варіації врожайності рису лінійно пов'язані з агротехнічними факторами, включеними в модель.

Окрім названих множинних коефіцієнтів щільності зв'язку, в комп'ютерних програмах передбачено розрахунок R^2 з урахуванням числа ступенів вільності:

$$R_k^2 = 1 - \frac{s_e^2}{s_y^2} = 1 - \left(1 - R^2\right)\frac{n-1}{n-(m+1)},$$

де $s_y^2 = \frac{\Sigma(y-\bar{y})^2}{n-1}$ — оцінка дисперсії результативної ознаки y;

s_e^2 — оцінка залишкової дисперсії.

Скоригований коефіцієнт множинної детермінації R_k^2 відрізняється від R^2 співвідношенням числа ступенів вільності дисперсій: залишкової $(n - m + 1)$ і загальної $(n - 1)$. Для розглянутої моделі це співвідношення становить $(34 - 1) : (34 - 6 - 1) = 1{,}2222$, а $R_k^2 = 1 - (1 - 0{,}7029) \cdot 1{,}2222 = 0{,}6369$.

У моделях множинної регресії поряд з оцінкою сукупного впливу всіх включених у модель факторів вимірюється кореляція між функцією y та кожним окремим фактором x_i при елімінуванні впливу інших факторів. Для цього використовують *частинні коефіцієнти детермінації* R_i^2. Схему розрахунку R_i^2 розглянемо на прикладі фактора x_6 моделі врожайності рису. До введення його в модель п'ять факторів пояснювали 64,61% варіації врожайності ($R^2 = 0{,}6461$), не поясненими залишалися $(1 - 0{,}6461) \cdot 100 = 35{,}39\%$ варіації. Фактор x_6 додатково пояснив $0{,}7029 - 0{,}6461 = 0{,}0568$ варіації y, що відносно не поясненої іншими факторами варіації становить $0{,}0568 : 0{,}3539 = 0{,}1605$. Це і є частинним коефіцієнтом детермінації фактора x_6.

Отже, розрахунок R_i^2 ґрунтується на порівнянні двох регресійних моделей: повної, з урахуванням фактора x_i, і скороченої, у якій фактор x_i відсутній. Чисельник R_i^2 дорівнює різниці сукупних коефіцієнтів детермінації цих моделей, знаменник — одиниці мінус сукупний коефіцієнт детермінації скороченої моделі. Загальну схему його розрахунку можна представити як відношення сум квадратів: частинної S_i^2 і залишкової S_e^2:

$$R_i^2 = \frac{S_i^2}{S_i^2 + S_e^2},$$

де $S_i^2 = \frac{b_i^2}{c_{ii}}$;

c_{ii} — діагональний елемент оберненої матриці.

Корінь квадратний із частинного коефіцієнта детермінації називають *частинним коефіцієнтом кореляції*.

Іноді для характеристики ролі кожного фактора у відтворенні варіації y сукупний коефіцієнт детермінації розкладають на складові:

$$R^2 = \sum_1^m d_i^2,$$

де $d_i^2 = \beta_i r_{i0}$ — *коефіцієнт окремої детермінації*, який залежить від потужності впливу i-го фактора на y та щільності зв'язку між ними (r_{i0} — парний коефіцієнт кореляції).

Ефекти впливу факторів на врожайність рису та характеристики щільності зв'язку наведено в табл. 5.3.

Таблиця 5.3

Фактор	r_{i0}	b_i	β_i	d_i^2	R_i^2
x_1	0,597	0,965	0,192	0,1146	0,0727
x_2	0,614	3,400	0,248	0,1521	0,1160
x_3	0,489	0,501	0,045	0,0221	0,0039
x_4	0,638	7,500	0,264	0,1687	0,1168
x_5	0,411	1,730	0,029	0,0119	0,0020
x_6	0,716	3,443	0,362	0,2335	0,1605

У таблиці для кожного фактора наведено три характеристики щільності зв'язку: парний коефіцієнт r_{i0}, частинний R_i^2 і коефіці-

єнт окремої детермінації d_i^2. Найбільші значення мають парні коефіцієнти кореляції. Це пояснюється тим, що фактори взаємозалежні, і парний коефіцієнт кореляції акумулює вплив інших факторів. Частинні коефіцієнти характеризують відносну зміну залишкової дисперсії за рахунок відповідного фактора; для кожного з них база порівняння інша, а тому аналітичні можливості їх обмежені. Коефіцієнти окремої детермінації, сума яких дорівнює множинному коефіцієнту детермінації $R^2 = 0{,}7029$, упорядковуючи фактори за потужністю впливу, практично дублюють висновки, які можна зробити на основі бета-коефіцієнтів.

Перевірка істотності зв'язку статистично формулюється як перевірка нульових гіпотез: $H_0 : R^2 = 0$; $H_0 : b_i = 0$. Гіпотеза H_0 відхиляється чи визнається допустимою на основі статистичних критеріїв, зокрема дисперсійного F-критерію, статистична характеристика якого розраховується відношенням оцінок факторної і залишкової дисперсій:

$$F = \frac{S_Y^2}{S_e^2} : \frac{m-1}{n-(m-1)} \text{ або } F = \frac{R^2}{1-R^2} : \frac{m-1}{n-(m-1)}.$$

Критичні значення $F_{1-\alpha}(k_1, k_2)$, де α — рівень істотності, $k_1 = m-1$, $k_2 = n-(m-1)$ — числа ступенів вільності чисельника та знаменника, наведено в додатку 10. Оскільки F-критерій функціонально зв'язаний з коефіцієнтом детермінації R^2, то перевірку істотності зв'язку можна здійснити, використовуючи безпосередньо критичні значення $R^2_{1-\alpha}(k_1, k_2)$, наведені в додатку 11.

Паралельно з оцінюванням адекватності моделі проводиться *перевірка істотності впливу окремих факторів* x_i на y за допомогою t-критерію:

$$t = \frac{b_i}{\mu_{b_i}},$$

де $\mu_{b_i} = \sqrt{s_e^2 c_{ii}}$ — стандартна похибка коефіцієнта регресії;

s_e^2 — оцінка залишкової дисперсії;

c_{ii} — діагональний елемент оберненої матриці C.

Критичні значення $t_{1-\alpha}(k)$, де $k = n-1$ наведено в додатку 5. Ефект впливу i-го фактора визнається істотним, якщо $t_i > t_{1-\alpha}(k)$. Так, при $\alpha = 0{,}05$ і $k = 20$ коефіцієнт b_i в 2,15 раза перевищує стандартну похибку μ_{b_i}, що свідчить про його значущість (істотність).

Довірчі межі ефекту впливу визначаються за правилами вибіркового методу $b_i \pm t_{1-\frac{\alpha}{2}}\mu_{b_i}$, де $t_{1-\frac{\alpha}{2}}$ — значення двостороннього t-критерію.

Процедури регресійного аналізу об'єднано в модулі *Multiple Regression* — Множинна регресія. Як приклад розглянемо модель залежності виходу цукру з 1 т сировини в кг (y) від цукристості буряка (x_1), втрат сировини при транспортуванні та зберіганні (x_2) та втрат цукру при переробці сировини (x_3). Первинні дані наведено в табл. 2.1.

На стартовій панелі модуля відкриваємо файл даних і проводимо селекцію ознак на залежну (*Dependent var.*) та незалежні (*Independent Variable list*). За командою на виконання програми з'являється вікно результатів аналізу — *Multiple Regression Results*. У верхній, інформаційній частині цього вікна вказується назва залежної ознаки та обсяг сукупності; наводяться значення коефіцієнтів щільності зв'язку: множинної кореляції R, множинної детермінації R^2 та R_k^2 (у таблицях відповідно $R1$ та Adjusted $R1$), значення F-критерію, стандартної похибки s_e — *St.errou*, вільного члена рівняння регресії b_0 — *Intercept* та його похибки, значення β_i-коефіцієнтів.

У нижній, функціональній частині вікна пропонуються опції, за допомогою яких можна провести всебічний аналіз результатів регресійного аналізу. Так, опція *Regression Summary* видає таблицю, в якій, окрім зазначених характеристик, наведено для всіх включених у модель факторів β_i-коефіцієнти і коефіцієнти регресії b_i із стандартними похибками, значення t-критерію і фактичні рівні істотності *p-level*. У табл. 5.4 наведено характеристики регресійної моделі виходу цукру з 1 т сировини.

Таблиця 5.4

Regression Summary for Dependent Variable: VAR4 (new.sta)						
Continue...	R = ,919228 RI = ,844981 Adjusted RI = ,802703 F(3,11)=19,986 p<,00009 Std.Error of estimate: ,36406					
N = 15	BETA	St. Err. of BETA	B	St. Err. of B	t(11)	p-level
Intercpt			9,812	8,287	1,184	0,261
VAR1	0,332	0,146	0,953	0,420	2,267	0,044
VAR2	–0,507	0,157	–10,084	3,128	–3,223	0,008
VAR3	–0,377	0,130	–1,729	0,598	–2,888	0,014

Згідно з даними таблиці рівняння регресії має такий вигляд:

$$Y = 9{,}812 + 0{,}953x_1 - 10{,}084x_2 - 1{,}729x_3.$$

Із збільшенням цукристості буряка на 1%, за умови незмінності інших факторів, вихід цукру з 1 т сировини зростає в середньому на 0,953%; щодо порушень технології зберігання та переробки сировини, то вони мають негативний вплив, особливо порушення технології зберігання. Включені в модель фактори пояснюють 84,5% варіації виходу цукру з 1 т сировини; ефекти впливу усіх факторів істотні.

Опція *Analysis of variance* пропонує таблицю декомпозиції варіації показника-функції, де вказані суми квадратів *Sums of Squares*: факторна *Regress.*, залишкова *Residual* та загальна *Total*, число ступенів вільності *df*, оцінки дисперсій *Mean Squares*, значення *F*-критерію та *p-level* (табл. 5.5).

Таблиця 5.5

Analysis of Variance, DV VAR4 (new sta)					
Continue	Sums of Squares	df	Mean Squares	F	p-level
Regress	7,947	3	2,649	19,986	9,27E-05
Residual	1,458	11	0,132		
Total	9,405				

За опцією *Partial correlation* визначаються частинні коефіцієнти кореляції *Partial Cor.* для кожної змінної. У таблиці результатів (табл. 5.6), окрім коефіцієнтів частинної і напівчастинної (*Semipart Cor.*) кореляції, пропонується тест толерантності, за яким оцінюється ступінь зв'язку x_i з іншими включеними в модель факторами. Якщо x_i є лінійною комбінацією інших факторів, то *R-square* наближується до 1, а *Tolerance* $(1 - R^2)$ — до 0. Фактор з малою толерантністю не несе додаткової інформації, і включення його в модель не виправдане.

Таблиця 5.6

Variables currently in the Equation, DV VAR4 (new sta)							
Continue	Beta in	Partial Cor	Semipart Cor	Tolerance	R-square	t(11)	p-level
VAR1	0,332	0,564	0,269	0,656	0,344	2,267	0,045
VAR2	–0,507	–0,697	–0,383	0,570	0,430	–3,223	0,008
VAR3	–0,377	–0,657	–0,343	0,826	0,174	–2,889	0,015

5.4. ЗАБЕЗПЕЧЕННЯ АДЕКВАТНОСТІ РЕГРЕСІЙНОЇ МОДЕЛІ

Адекватність регресійної моделі означає здатність її правильно описати реальну структуру взаємозв'язків між ознаками x_i та y. Методологічною основою вирішення проблеми адекватності є теоретичний, змістовний аналіз матеріальної природи процесу (явища) та обґрунтування типу й структури моделі, яка описує механізм його формування. Практично з метою забезпечення адекватності моделі змістовний аналіз поєднується з формальними процедурами перевірки гіпотез щодо дотримання логіко-статистичних умов використання МНК.

Мірою адекватності моделі, як уже зазначалося в 5.3, слугують відхилення фактичних значень від теоретичних $e_j = y_j - Y_j$. На величину цих відхилень впливає весь комплекс умов, зокрема:

- обсяг та однорідність сукупності;
- незалежність спостережень;
- інформативність включених у модель факторів;
- стабільність не включених у модель факторів;
- тип моделі.

Репрезентативність оцінок регресійного аналізу прямо пропорційна обсягу та однорідності сукупності. Саме недостатній обсяг сукупності та її неоднорідність вважаються найвагомішими чинниками неадекватності моделей. Тому при формуванні ознакової множини моделі слід враховувати співвідношення між обсягом вибірки і кількістю включених у модель факторів (воно має бути приблизно 8 : 1).

Оцінювання однорідності сукупності здійснюється на етапі розвідувального аналізу даних (див. 3.1). Так, наявність аномальних значень, які не узгоджуються з розподілом основної маси даних, може бути наслідком помилок спостереження або результатом незвичайної комбінації причин і умов, у яких функціонує одиниця сукупності. Ідентифікація таких спостережень дає можливість усунути помилки, а якщо це неможливо, то вилучити аномальний об'єкт з подальшого аналізу. Якщо сукупність розшарована на групи (кластери), то в моделі можна врахувати таку неоднорідність (див. 6.2).

Інформативність включених у модель факторних ознак залежить як від соціально-економічного змісту, так і від шкали вимірювання ознаки. Якщо ознака за змістом не інформативна, то ні-

який спосіб моделювання не забезпечить належних результатів. Так само результати аналізу будуть суттєво різнитися залежно від того, якою шкалою представлено одну й ту саму ознаку (метричною, ранговою чи номінальною).

Ті властивості, що безпосередньо не вимірюються або не мають єдиного вимірника, включаються в модель у вигляді інтегральних оцінок. Наприклад, погодні умови характеризуються середньодобовою температурою повітря, кількістю опадів, тривалістю сонячного світла, хмарністю і т. ін. Усі ці характеристики агрегуються в індексі погодних умов.

Важливою умовою регресійного аналізу є відсутність мультиколінеарності, яка веде до зсунення оцінок параметрів моделі та унеможливлює коректну інтерпретацію результатів. Два фактори вважаються колінеарними, якщо коефіцієнт кореляції між ними перевищує сукупний коефіцієнт кореляції, тобто $r_{ik} > R$. Найпростіший спосіб усунення мультиколінеарності — виключити одну із корельованих ознак із моделі або замінити її іншою. Часом колінеарні фактори агрегуються в одну узагальнюючу оцінку (див. 9.1).

Стабільність не включених у модель факторів означає, що вплив їх на варіацію y незначний і врівноважується, він однаковий в усіх частинах сукупності (умова *гомоскедастичності*). Математичною основою дотримання цих передумов МНК слугує ймовірнісний розподіл залишків e_j. Передбачається, що:

- для кожного спостереження залишок e_j — випадкова величина, яка має нормальний розподіл. Умова нормальності необхідна для визначення довірчих меж коефіцієнтів регресії і для перевірки гіпотез щодо їх істотності;
- математичне сподівання залишків $M(e) = 0$;
- дисперсія залишків однакова в усіх частинах сукупності: $s_e^2 = \text{const}$. Ця умова пов'язана з однорідністю сукупності;
- залишки незалежні, тобто відсутня серійна кореляція чи автокореляція даних.

У модулі *Multiple Regression* процедури аналізу залишків *Residual analysis* передбачають як візуальні методи, так і статистичні. На рис. 5.2 наведено один з типів графіків залишків (*Normal probability plot of residuals*) для моделі виходу цукру з 1 т сировини. Пряма відповідає нормальному закону розподілу. По тому, як коливаються залишки навколо прямої, можна зробити висновок щодо дотримання умови нормальності їх розподілу, гомоскедастичності і незалежності. Тобто вплив не врахованих в моделі факторів незначний, спільна їх дія є однаковою в усіх частинах сукупності, варіація виходу цукру з 1 т сировини не залежить від рівня не врахованих факторів. Все це свідчить про адекватність моделі реальному процесу.

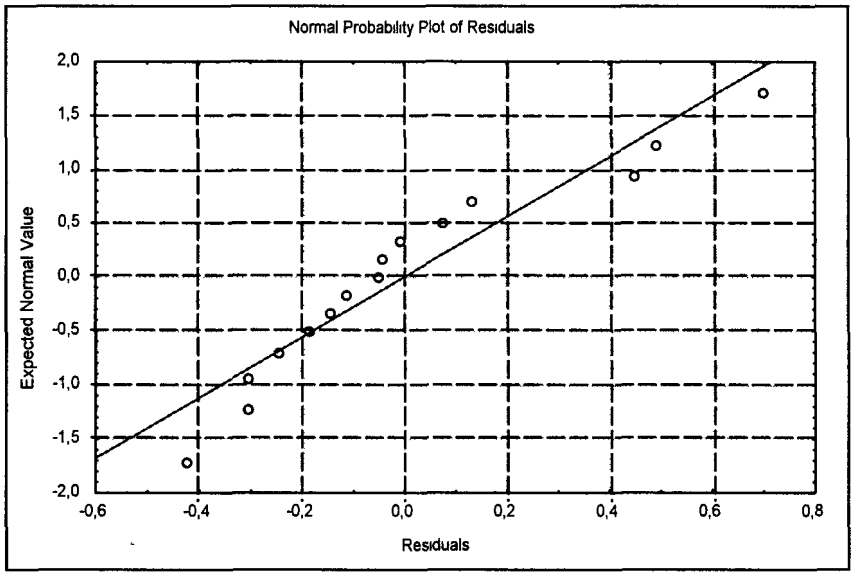


Рис. 5.2. Графік залишків регресійної моделі

У великих за обсягом сукупностях візуальні методи аналізу залишків необхідно поєднувати із статистичними оцінками, які представлені групою опцій *Statistics*.

На основі адекватної регресійної моделі можна здійснити прогноз показника-функції y, задаючи певні значення факторів x_i. Такий прогноз уможливлює опція *Predict dependent var.* — Прогноз залежної ознаки. Значення x_i задаються у вікні *Specify values for indep.vars*. Так, визначивши, що $x_1 = 15{,}6$, $x_2 = 0{,}9$ і $x_3 = 2{,}0$, маємо прогнозне значення виходу цукру з 1 т сировини $Y = 12{,}14$ кг.

Використовуючи параметри моделі, можна також оцінити потенційно можливі рівні показника-функції для кожної одиниці сукупності, визначити резерви збільшення (зменшення) показника y за рахунок факторів, які піддаються регулюванню (суб'єктивних факторів). У нашому прикладі — це збільшення виходу цукру з 1 т сировини за рахунок зменшення втрат при зберіганні цукрового буряка і в процесі його переробки. Така оцінка, при-

родно, орієнтована на кращі досягнення в галузі. Ефект регулювання i-го фактора на j-му об'єкті визначається за формулою

$$\Delta x_i = b_i (x_{ij} - x_{i0}),$$

де x_{i0} — база порівняння,

b_i — коефіцієнт регресії i-го фактора.

Застосовуючи цю методику, визначимо резерв збільшення виходу цукру з 1 т сировини для j-го заводу (табл. 5.7).

Таблиця 5.7

Фактор	Рівень втрат, %		Відхилення	Коефіцієнт регресії	Ефект регулювання фактора
	фактичний	мінімальний			
x_2	1,06	0,90	0,16	–10,084	–1,613
x_3	2,68	2,0	0,68	–1,729	–1,175
Разом	X	X	X	X	–2,788

Якщо мінімальні втрати цукрового буряка при переробці — 2,0%, а на j-му заводі — 2,68%, то ефект доведення втрат до мінімального рівня становить $(2,68 - 2,0)(-1,729) = -1,175$. Зменшення втрат при зберіганні цукрового буряка дає ефект $(1,06 - 0,90)(-10,084) = -1,613$. Отже, сумарний ефект за рахунок обох факторів –2,788, а потенційно можливий вихід цукру з 1 т сировини за незмінності цукристості буряка, яка є зовнішнім, об'єктивним фактором, становить 11,91 кг. Відношення фактичного рівня до потенційно можливого характеризує *ступінь використання об'єктивних можливостей*. У розглянутому прикладі це відношення становить $9,13 : 11,91 = 0,777$, тобто ефективність використання сировини на заводі нижча за потенційно можливу на 23,3%. При визначенні резервів збільшення (зменшення) показника-функції за рахунок регулювання суб'єктивних факторів базою порівняння може бути середня величина, норматив, стандарт тощо.

Завдання для самоконтролю

1. За наведеними даними (в млн. грн.) побудуйте 4-факторну індексну модель ефективності комерційної діяльності фірми, вимірником якої є балансова рентабельність виробничого капіталу. Оцініть абсолютний вплив на динаміку цього показника кожного фактора.

Показник	Базисний період	Поточний період
Балансовий прибуток	4,0	3,8
Виручка від реалізації продукції	22,6	24,0
Витрати на виробництво продукції	19,0	20,5
Виробничий капітал	43,7	45,9
У т. ч. оборотний капітал	5,8	5,3

2. Динаміка матеріальних витрат на виробництво продукції залежить від матеріаломісткості продукції, оборотності та розміру оборотного капіталу. За поточний квартал матеріальні витрати зросли з 200 до 221 млн. грн. Визначте абсолютний вплив металомісткості та оборотності капіталу на динаміку матеріальних витрат.

Показник	Індекс
Матеріальні витрати	1,105
Оборотний капітал	1,12
Матеріаломісткість продукції	1,05
Оборотність капіталу	0,94

3. За минулий рік темпи приросту макропоказників становили: ВВП — 0,7%, матеріальних витрат — 1,5%, енерговитрат — 2,3%, кількості робочих місць — 0,2%. Проведіть діагностику збалансованості економічного розвитку за умови енергозберігаючої економічної стратегії.

4. За наведеними даними побудуйте індексно-матричну модель розвитку промисловості регіону і зробіть висновки щодо збалансованості динаміки показників інтенсивності та ефективності промислового виробництва.

Показник	Індекс
Товарна продукція промисловості	0,90
Основні виробничі фонди	1,02
Матеріальні витрати	0,95
Споживання електроенергії	1,07
Витрати праці, людино-годин	0,98

5. Залежність питомих витрат газу в чорній металургії від обсягу виробництва прокату чорних металів VAR1 та споживання вугілля VAR2 описується параметрами:

Regression Summary for Dependent Variable VAR3 (new sta)				
R= ,8960 RI= ,8028 Adjusted RI= ,737 F(2,10)=20,355 p<,000 Std Error of estimate· ,0206				
N =12	BETA	B	St Err of B	t(11)
Intercpt		4,192	2,551	1,64
VAR1	–0,847	–48,48	7,916	–6,124
VAR2	0,534	2,814	0,651	4,322

Поясніть зміст параметрів, зробіть висновок щодо адекватності моделі.

6. Функція питомих витрат енергоресурсів на залізницях має вигляд:

$$Y = -0{,}173 + 0{,}0073x_1 - 0{,}04x_2,$$

де x_1 — співвідношення сумарного споживання енергії і вантажообороту залізниці, x_2 — частка електрифікованої довжини залізниць. Стандартні похибки коефіцієнта регресії становлять відповідно 0,07 і 0,02.

Перевірте істотність ефектів впливу факторів для $n = 10$. Висновок зробіть з імовірністю 0,95.

7. Залежність споживання яловичини (кг на душу населення за рік) від ціни P та середньодушового доходу D (грн.) характеризується коефіцієнтами регресії:

Фактор	b_i	$\bar{x}_i$
P	–3,2	6
D	0,16	120

Середній рівень споживання яловичини — 24 кг. Порівняйте ефекти впливу факторів.

8. У регресійну модель собівартості продукції за даними 25 підприємств взуттєвої промисловості покроково вводилися такі фактори: x_1 — продуктивність праці в умовних одиницях трудомісткості; x_2 — витрати хромової шкірсировини на одну пару знеособленого взуття; x_3 — фондоозброєність праці. Сукупні коефіцієнти кореляції становили відповідно 0,820; 0,875; 0,902.

Визначте частинні коефіцієнти кореляції, перевірте їх істотність. Висновок зробіть з імовірністю 0,95.

9. Регресійна модель фондовіддачі на машинобудівних підприємствах має вигляд: $Y = -21{,}33 + 0{,}063x_1 + 1{,}78x_2 + 5{,}59x_3 + 0{,}133x_4$.

Факторні ознаки: x_1 — ступінь використання потужностей підприємства, %; x_2 — капітальні витрати на 1000 грн. потужності, тис. грн.; x_3 — частка устаткування в загальній вартості основних виробничих фондів, %; x_4 — ступінь автоматизації виробничих процесів, %.

Визначте резерви підвищення фондовіддачі в групі підприємств, які використовують застарілу технологію, якщо значення факторів довести до рівня модернізованих підприємств.

Факторна ознака	Середньогрупові значення факторних ознак	
	Група підприємств, які потребують модернізації	Група модернізованих підприємств
x_1	63,6	93,4
x_2	4,0	4,3
x_3	0,72	0,70
x_4	48,3	54,2
y	0,60	3,32

10. Регресійна модель продуктивності праці (тис. грн. на працівника) має вигляд: $Y = -3{,}42 + 0{,}32x_1 - 6{,}36x_2 + 0{,}18x_3$, де x_1 — ступінь завантаженості устаткування; x_2 — оборотність матеріальних запасів; x_3 — енергоозброєність праці. Для конкретних підприємств визначте ступінь використання об'єктивних можливостей виробництва:

Підприємство	Значення факторів			y
	x_1	x_2	x_3	
1	76	0,4	20,4	23,1
2	68	0,5	18,7	17,6

11. Оцінки толерантності двох факторних ознак становлять: 0,65 і 0,12. Яку з них слід включити в ознакову множину регресійної моделі і чому?

Розділ 6

РОЗШИРЕНА РЕГРЕСІЯ

6.1. РЕГРЕСІЯ НА ЗМІШАНИХ ФАКТОРНИХ МНОЖИНАХ

У моделях класичної регресії факторні ознаки x_i належать до метричної шкали вимірювання — виражаються числом, і значення їх варіюють у певних межах. У соціально-економічних дослідженнях часто стикаються з ситуацією, коли окремі властивості явищ — нечислові, текстові (форма власності, професія тощо). Це ознаки номінальної шкали — шкали найменувань, градацій. Використання таких ознак у регресійному аналізі передбачає їх оцифровування, тобто приписування кожній градації певного числа. Можливі різні варіанти оцифровування, проте на практиці найчастіше застосовують двійкову систему, коли приписане k-й градації число u_{ik} має лише два значення (0; 1).

Оцифровування передбачає дотримання двох умов:

- повноту шкали градацій;
- неперетинальність градацій.

Повнота шкали градацій дає: $\Sigma\, u_k = f_k$, де f_k — частота k-ї градації. Для кожної з них середнє значення дорівнює частці $\overline{u} = \frac{f_k}{n} = d_k$ Оскільки величина u_{ik} є характеристикою розподілу сукупності, то в подальшому будемо її називати *структурною змінною*. В математичній літературі таку змінну називають фіктивною, дихотомічною, бінарною.

Умова неперетинальності виключає одночасну належність одиниці сукупності до двох градацій: $\Sigma\, u_k u_s = 0$, де k, s — градації $(k \neq s)$.

Структурна змінна розглядається як умовний код, що вказує на належність (1) чи неналежність (0) j-ї одиниці сукупності до k-ї градації. Для ознаки, що має p градацій $x_1, x_2, \ldots, x_p$, ставиться у відповідність $(p-1)$ величин $u_1, u_2, \ldots, u_{p-1}$. У регресійному аналізі до матриці ознакової множини X додається матриця структурних змінних $U = [u_1, u_2, \ldots, u_{p-1}]$, а модель включає додаткові члени $a_1u_1 + a_2u_2 + \ldots + a_{p-1}u_{p-1}$. Параметри a_k оцінюються одночасно з коефіцієнтами регресії b_i при метричних ознаках. Так, наприклад, за даними агропідприємств моделюється залежність ефективності використання землі y від якості ґрунтів x_1 і виробничої спеціалізації господарств x_2. Перший фактор вимірюється балами, другий — належить до номінальної шкали і має три градації: а) овочево-молочну, б) буряківництво і в) зернову. В ознакову множину моделі другий фактор x_2 вводиться двома структурними змінними:

$$u_{21} = \begin{cases} 1 — \text{для овочево-молочних,} \\ 0 — \text{для інших;} \end{cases}$$

$$u_{22} = \begin{cases} 1 — \text{для буряківництва,} \\ 0 — \text{для інших.} \end{cases}$$

Відповідно формуються два вектори значень цих величин (табл. 6.1). При такому варіанті оцифровування третя спеціалізація (зернова) дістає числові еквіваленти (0; 0) і стає базою порівняння для перших двох. Регресійна модель ефективності використання землі з урахуванням спеціалізації господарств має вигляд:

$$Y = a_0 + a_{21}u_{21} + a_{22}u_{22} + b_1x_1.$$

Параметр b_1 характеризує чистий ефект впливу якості ґрунтів на ефективність використання землі за умови однакової спеціалізації;

a_{21} показує різницю в ефективності використання землі в господарствах овочево-молочної спеціалізації порівняно з господарствами зернового спрямування за умови однакової якості ґрунтів;

a_{22} має таку ж інтерпретацію для господарств, які спеціалізуються на буряківництві;

a_0 — вільний член рівняння.

Отже, теоретичний рівень ефективності використання землі для відповідної спеціалізації визначається так:

$Y = a_0 + b_1x_1$ — для зернової;

$Y = (a_0 + a_{21}) + b_1x_1$ — для овочево-молочної;

$Y = (a_0 + a_{22}) + b_1x_1$ — для буряківництва.

Таблиця 6.1

Номер агрогосподарства	Спеціалізація	Числовий еквівалент	
		u_1	u_2
1	а	1	0
2	в	0	0
3	б	0	1
4	а	1	0
…	…	…	…
n	б	0	1

Загальний вигляд регресійної моделі із структурними змінними:

$$Y = a_0 + \sum_{i=1}^{q} b_i x_i + \sum_{i=q+1}^{m} \sum_{k=1}^{p-1} a_{rk} u_{rk}.$$

Ознакова множина такої моделі складається з двох блоків: перший — блок факторних ознак метричної шкали обсягом $(q \cdot n)$, другий — блок структурних змінних для ознак номінальної шкали обсягом $[(m-q) \cdot n]$.

Коефіцієнти регресії вимірюють:

b_i — чистий, елімінований від взаємозв'язків всередині моделі, ефект впливу фактора x_i;

a_{rk} — вплив k-ї градації r-го фактора $(r \neq i)$ на функцію y; алгебраїчно — це різниця середніх значень функції y між k-ю градацією і градацією, взятою за базу порівняння.

При моделюванні використовуються процедури модуля *Multiple Regression* (див. 5.3). Специфікація текстових ознак передбачає їх оцифровування. В системі *Statistica* ця процедура здійснюється для кожної ознаки окремо за командами: *Current Specs* (кнопка VARS) → *Text Values*. У діалоговому вікні *Text Values Manager* — Менеджер текстових значень — вказуються числові еквіваленти (*Text Value — Numeric*).

Як приклад розглянемо модель, що описує залежність вартості будівництва атомних електростанцій з реактором водяного охолодження від номінальної потужності електростанцій, використання нагрівальної башти та силової установки виробництва фірми B-W [3]. Два останніх фактори представлені текстовими ознаками і підлягають оцифровуванню. В табл. 6.2 наведено дані по 23 електростанціях: VAR1 — вартість електростанції, млн. дол. США, VAR2 — потужність електростанції, МВт, VAR3 — приписані значення 1 і 0 залежно від того, використовує чи не використовує електростанція нагрівальну башту, VAR4 — аналогічно приписані значення стосовно використання силової установки виробництва фірми B-W.

Модель вартості будівництва електростанцій має вигляд:

$$Y = a_0 + b_1 x_1 + a_1 u_1 + a_2 u_2.$$

Значення параметрів наведено в табл. 6.3. Коефіцієнт детермінації становить 0,506, тобто включені в модель фактори пояснюють 50,6% варіації вартості атомних електростанцій. Значення F-критерію і *p-level* свідчать про адекватність моделі, а t-критерію — про істотний вплив кожного фактора.

Таблиця 6.2

Номер електростанції	VAR1	VAR2	VAR3	VAR4
1	460	687	0	0
2	453	1065	0	1
3	443	1065	0	1
4	642	1065	1	1
5	272	822	0	0
6	317	457	0	0
7	457	822	0	0
8	350	560	0	0
9	402	790	0	0
10	412	530	1	0
11	394	850	0	1
12	423	778	0	0
13	712	845	0	0
14	881	1090	0	0
15	491	1050	0	0
16	568	913	1	1
17	621	786	1	0
18	473	538	1	0
19	207	745	0	0
20	284	886	0	1
21	217	745	0	0
22	345	514	1	0
23	280	886	0	1

Таблиця 6.3

Regression Summary for Dependent Variable: VAR1

R= ,7114 RI= ,5061 Adjusted RI= ,4281
F(3,19)=6,49 p<,0033 Std.Error of estimate: 123,00

$N = 23$	BETA	St. Err. of BETA	B	St. Err. of B	t(19)	p-level
Intercpt			– 129,295	137,448	– 0,9406	0,3587
VAR2	0,841	0,2054	0,714	0,174	4,0947	0,0006
VAR3	0,495	0,1708	179,342	61,883	2,8980	0,0092
VAR4	–0,493	0,1989	– 170,317	68,753	– 2,4772	0,0228

Аналізуючи параметри моделі, слід зазначити, що найвагоміший вплив на вартість будівництва має потужність електростанцій, значення β-коефіцієнта для цього фактора становить 0,841. Ефекти впливу використання нагрівальної башти і силових установок приблизно однакові, але напрямок дії різний. На електростанціях, які використовують нагрівальні башти, вартість будівництва в середньому на 179,342 млн. дол. вища, тоді як використання силових установок фірми B-W, навпаки, зменшує капітальні витрати в середньому на 170,317 млн. дол.

Розглянута методика використання структурних змінних передбачає, що усі одиниці сукупності мають градації існуючої шкали. Якщо ця умова не виконується, то можна ввести додаткову групу для невизначених градацій.

Не завжди виконується й умова неперетинальності груп — та сама одиниця сукупності може одночасно належати до різних градацій. Скажімо, робітник має декілька професій, і щоб забезпечити умову неперетинальності, його відносять до градації, яка відповідає основній професії. Аналогічна проблема виникає при обробці даних соціологічних обстежень, програмою яких передбачені питання-набори. Наприклад, респондент може вказати декілька джерел інформації про валютний ринок: телебачення, преса, особисті спостереження. Кожна градація набору розглядається як альтернативна ознака і може самостійно включатися в модель.

6.2. АДАПТАЦІЯ РЕГРЕСІЙНОЇ МОДЕЛІ ДО НЕОДНОРІДНОЇ СУКУПНОСТІ

За допомогою структурних змінних можна адаптувати регресійну модель до неоднорідної сукупності. Якщо неоднорідність проявляється розшаруванням сукупності на p ізольованих класів (груп), то кожен клас розглядається як градація номінальної ознаки і тим одиницям, що належать до j-го класу, приписується структурна змінна $u_j = 1$, а тим, що не належать, — $u_j = 0$. Параметри при структурних змінних класів інтерпретуються так само, як і при градаціях текстових ознак.

Специфіка моделювання процесів у неоднорідній сукупності зумовлена своєрідністю внутрішньокласової варіації і характером взаємозв'язків. Ефекти впливу одного й того ж фактора

x_i на y по класах можуть істотно різнитися. Наприклад, у вугільній промисловості виділяються класи шахт за гірничо-геологічними умовами: потужністю та нахилом залягання пластів, їх газоносністю, глибиною розробки лав тощо. Кожному типу цих природних умов відповідають певна технологічна схема і певний рівень механізації виробничих процесів. Вплив механізації на трудомісткість, скажімо, очисних робіт залежить від класу шахти. Залежність сили впливу одного фактора від рівня іншого називається взаємодією. В неоднорідних сукупностях йдеться про *взаємодію* факторів і специфічних умов окремих класів. Для цього використовують змінні взаємодії $x_i u_j$, значення яких дорівнює добутку значень відповідних ознак.

Отже, при моделюванні взаємозв'язків у неоднорідних сукупностях ознакова множина моделі включає, окрім факторних ознак x_i, два типи інструментальних змінних: структурні змінні u_j, які відображують особливості класів, і змінні взаємодії $x_i u_j$, що характеризують особливості взаємозв'язків в окремих класах. За рахунок цих змінних модель регресійного аналізу розширюється:

$$Y = a_0 + \sum_{i=1}^{m} b_i x_i + \sum_{j=1}^{p-1} a_j u_j + \sum_i \sum_j c_{ij} x_i u_j \ .$$

Зміст параметрів моделі: b_i — чистий ефект впливу i-го фактора в середньому по сукупності;

a_j — відхилення середнього значення показника-функції в j-му класі від середнього його рівня в класі, взятому за базу порівняння;

c_{ij} — відхилення ефекту впливу i-го фактора в j-му класі від середнього по сукупності.

Істотність параметрів a_j та c_{ij} свідчить про неоднорідність сукупності.

Модель такого типу Е. Маленво назвав *коваріаційною*. В ній поєднуються регресія на факторних ознаках метричної шкали і модель дисперсійного аналізу міжкласових відмінностей. Вона уможливлює одночасну оцінку декількох рівнянь, і така оцінка може бути точнішою, ніж оцінки покласових регресій.

Як приклад розглянемо модель продуктивності праці робітників очисних вибоїв за даними 21 шахти, з-поміж яких за гірничо-геологічними умовами 12 належать до першого класу (пологі пласти), дев'ять — до другого класу (крутопадаючі пласти). Ак-

центуючи увагу на особливостях коваріаційної моделі, обмежимося одним фактором — швидкістю просування лави. Первинні дані наведено в табл. 6.4.

Таблиця 6.4

Перший клас ($u_1 = 1$)			Другий клас ($u_1 = 0$)		
Номер шахти	Швидкість просування лави, м/міс.	Продуктивність праці, т/місяць	Номер шахти	Швидкість просування лави, м/міс.	Продуктивність праці, т/міс.
1	46	139	1	36	129
2	64	183	2	28	104
3	48	165	3	43	132
4	62	175	4	25	106
5	41	147	5	30	128
6	76	192	6	32	98
7	57	149	7	21	83
8	65	158	8	17	76
9	87	190	9	29	117
10	80	175			
11	48	153			
12	82	198			

Модель має вигляд: $Y = a_0 + b_1 x_1 + a_1 u_1 + c_{11} x_1 u_1$. Параметри її наведено в табл. 6.5. Коефіцієнт детермінації показує, що 92,1% варіації продуктивності праці робітників очисних вибоїв (VAR1) пояснюється класом шахт (VAR3) і швидкістю просування лави (VAR2). Адекватність моделі підтверджується значеннями *F*-критерію та *p-level*, істотність впливу факторів — характеристиками *t*-критерію. На шахтах з пологими пластами (клас 1) середня місячна продуктивність праці в очисних вибоях на 57,5 т вища, ніж на шахтах, що мають крутопадаючі пласти (клас 2). Зі збільшенням швидкості просування лави на 1 м продуктивність праці зростає в середньому на 2,23 т, на шахтах першого класу ефект впливу цього фактора на 1,15 т менший за середній.

Теоретичний рівень продуктивності праці визначається так:
для шахт першого класу $Y = (43{,}525 + 57{,}5) + (2{,}227 - 1{,}153)\, x_1$,
для шахт другого класу $Y = 43{,}525 + 2{,}227\, x_1$.

Таблиця 6.5

Regression Summary for Dependent Variable: VAR1						
R= ,96 RI= ,921 Adjusted RI= ,907 F(3,17)=66,4 p<,000 Std.Error of estimate: 11,05						
$N = 21$	BETA	St. Err. of BETA	B	St. Err. of B	t(17)	p-level
Intercpt			43,525	15,078	2,89	0,010
VAR2	1,31	0,296	2,227	0,504	4,42	0,000
VAR3	0,80	0,286	57,500	20,484	2,81	0,012
VAR4	–1,08	0,512	–1,153	0,548	–2,11	0,050

Аналогічно здійснюється модельна специфікація за наявності трьох і більше класів. Наприклад, ознакова множина моделі включає дві факторні ознаки (x_1, x_2) та дві структурні змінні (u_1, u_2):

Структурні змінні		Специфікація моделі
u_1	u_2	
0	0	$Y = a_0 + b_1 x_1 + b_2 x_2$
1	0	$Y = (a_0 + a_1) + (b_1 + c_{11})\, x_1 + b_2 x_2$
0	1	$Y = (a_0 + a_2) + b_1 x_1 + (b_2 + c_{22})\, x_2$
1	1	$Y = (a_0 + a_1 + a_2) + (b_1 + c_{11})\, x_1 + (b_2 + c_{22})\, x_2$

За допомогою структурних змінних можна врахувати в моделі нетиповість певної групи одиниць, які класифікуються як аномальні. Належним до такої групи елементам сукупності приписується $u_j = 1$.

Якісна однорідність є однією з умов моделювання динаміки. Вона виявляється неперервністю ряду, сталістю тенденції розвитку. Проте в рядах соціально-економічних показників ця умова часом порушується; спостерігаються розриви однорідності рядів через зміни в причинному комплексі формування тенденцій. Скажімо, зміна форми власності, фінансова криза, несприятливі погодні умови тощо. Тоді ряд динаміки в точці розриву t_p поділяється на інтервали за допомогою структурної змінної u_t. У першому інтервалі (до точки розриву) змінній u_t, яка вводиться в трендову модель лінійно, приписується значення «0». У другому інтервалі (після змін) $u_t = 1, 2, \ldots, m$, де m — довжина інтервалу.

Значення змінної часу t, навпаки, у першому інтервалі зростають, у другому фіксуються на рівні t_p.

Наприклад, у ряду динаміки y_t, де $t = 1, \ldots, 5$, зміни відбулися при $t = 3$. Файл первинних даних можна сформувати так:

y_t	y_1	y_2	y_3	y_4	y_5
t	1	2	3	3	3
u_t	0	0	1	2	3

Трендова модель має вигляд:

$$Y = a + bt + cu_t.$$

Параметри при змінній часу b і структурній змінній c характеризують абсолютну швидкість динаміки відповідно до і після змін.

Якщо в межах динамічного ряду зафіксовано два і більше зрушень, то відповідно збільшується кількість введених у модель структурних змінних.

6.3. РЕГРЕСІЯ НА ГРУПУВАННЯХ

Модель із структурними змінними і змінними взаємодії можна застосувати до комбінаційних групувань. Традиційно для аналізу взаємозв'язків за даними комбінаційних групувань використовується модель дисперсійного аналізу *Anova/Manova*. Основне завдання дисперсійного аналізу — виявити джерела варіації; дисперсійні комплекси орієнтовані переважно на обробку даних запланованих експериментів з однаковими частотами груп і підгруп. У соціально-економічних дослідженнях будь-яке комбінаційне групування є результатом статистичного спостереження, тобто «незапланованого» експерименту, а отже, забезпечити однакові частоти груп і підгруп практично неможливо. У такому разі перевага віддається моделям множинної лінійної регресії. Особливості застосування регресії до задач дисперсійного аналізу розглянемо на прикладі двофакторної класифікації (фактори A і B).

У регресійній моделі, як і в дисперсійному аналізі, значення i-ї ознаки у h-ї одиниці сукупності, яка належить до j-ї групи, представляється сумою загальної середньої μ, ефекту кожного фактора $(a_i + b_j)$ та ефектів їх взаємодії $(ab)_{ij}$:

$$Y_{ijh} = \mu + a_i + b_j + (ab)_{ij} + e_{ijh},$$

де e_{ijh} — залишок;

i — рівень фактора A;

j — рівень фактора B.

Щоб забезпечити однозначність МНК-оцінок параметрів моделі, формулюються додаткові обмеження:

$$\sum_i a_i = \sum_j b_j = 0;$$

$$\sum_i (ab)_{ij} = \sum_j (ab)_{ij} = 0.$$

На основі введених обмежень можна представити одні ефекти моделі як лінійну комбінацію інших і записати модель з мінімальною кількістю ефектів.

Наприклад, за фактором A виділено три групи, за фактором B — дві. Тоді в моделі, окрім ефектів факторів $(a_1, a_2, a_3, b_1, b_2)$, необхідно врахувати шість ефектів взаємодії: $(ab)_{11}$, $(ab)_{12}$, $(ab)_{21}$, $(ab)_{22}$, $(ab)_{31}$, $(ab)_{32}$. Сформулюємо додаткові обмеження для зазначених ефектів моделі:

- $a_1 + a_2 + a_3 = 0$; $b_1 + b_2 = 0$;
- $(ab)_{11} + (ab)_{21} + (ab)_{31} = (ab)_{12} + (ab)_{22} + (ab)_{32} = (ab)_{11} + (ab)_{12} = (ab)_{21} + (ab)_{22} = (ab)_{31} + (ab)_{32} = 0$.

Звідси маємо:

$a_3 = -a_1 - a_2$; $b_2 = -b_1$; $(ab)_{12} = -(ab)_{11}$; $(ab)_{22} = -(ab)_{21}$;

$(ab)_{31} = -(ab)_{11} - (ab)_{21}$; $(ab)_{32} = -(ab)_{31} = (ab)_{11} + (ab)_{21}$.

Визначальними виявляються параметри: μ, a_1, a_2, b_1, $(ab)_{11}$ та $(ab)_{21}$. Це мінімальна їх кількість, і модель з цими параметрами записується так:

$$Y = \mu + a_1 x_1 + a_2 x_2 + b_1 x_3 + (ab)_{11} x_4 + (ab)_{21} x_5.$$

Порядок формування файла первинних даних розглянемо на прикладі моделі тривалості перерви в роботі безробітних ($n = 12$). Фактор A — вікова група безробітних: A_1 — до 30 років; A_2 — від 30 до 50; A_3 — 50 років і старші. Фактор B — стать безробітного: B_1 — чоловіки; B_2 — жінки. В табл. 6.6 належність безробітного до відповідної підгрупи за цими факторами вказується подвійним індексом ij, де $i = 1, 2, 3$; $j = 1, 2$; y — тривалість перерви в роботі (міс.).

Ознакова множина моделі X являє собою матрицю коефіцієнтів при відповідних ефектах впливу факторів та ефектах їх взаємодії. Так, x_1 відповідає ефекту a_1, тому $x_1 = 1$ для першої вікової групи, $x_1 = 0$ для другої вікової групи, а оскільки $a_1 + a_2 + a_3 = 0$,

то для третьої вікової групи $x_1 = -1$. Аналогічно визначається вектор x_2, який відповідає ефекту a_2. Вектор x_3 відноситься до ефекта b_1. Прийнявши $x_3 = 1$ для чоловіків, маємо $x_3 = -1$ для жінок. Вектори x_4 та x_5 відносяться до ефектів взаємодії: $x_4 = x_1 x_3$; $x_5 = x_2 x_3$.

Таблиця 6.6

ij	*y*	x_1	x_2	x_3	x_4	x_5
11	4	1	0	1	1	0
11	3	1	0	1	1	0
21	5	0	1	1	0	1
31	7	–1	–1	1	–1	–1
31	6	–1	–1	1	–1	–1
31	5	–1	–1	1	–1	–1
12	2	1	0	–1	–1	0
12	4	1	0	–1	–1	0
12	3	1	0	–1	–1	0
22	7	0	1	–1	0	–1
22	6	0	1	–1	0	–1
32	11	–1	–1	–1	1	1

До сформованого таким чином файла первинних даних застосуємо процедури модуля *Multiple Regression.* Параметри моделі наведено в табл. 6.7.

Таблиця 6.7

Regression Summary for Dependent Variable: *y*						
R= ,96 RI= ,922 Adjusted RI= ,857 F(5,6)=14,22 p<,0028 Std.Error of estimate: ,913						
N = 12	BETA	St. Err. of BETA	B	St. Err. of B	t(6)	p-level
Intercpt, μ			5,833	0,291	20,02	1,01E-06
x_1	– 0,962	0,141	– 2,583	0,378	– 6,84	0,000
x_2	– 0,027	0,143	– 0,083	0,435	– 0,19	0,854
x_3	– 0,432	0,126	– 1	0,291	– 3,43	0,014
x_4	0,448	0,135	1,25	0,378	3,31	0,016
x_5	0,078	0,136	0,25	0,435	0,57	0,586

Істотними виявилися ефекти впливу першої групи фактора *A* (вік до 30 років) і першої градації фактора *B* (чоловіки), а також ефект взаємодії цих факторів $(ab)_{11}$. Якщо середня тривалість перерви в роботі по сукупності в цілому становить 5,8 міс., то у віковій групі до 30 років цей показник на 2,6 міс. менший; на 1 міс. менша за середню тривалість перерви в роботі у чоловіків.

При збільшенні кількості факторів оцінювання ефектів впливу кожного з них й усіх можливих взаємодій за розглянутою методикою значно ускладнюється. У такому разі ефективною виявляється модель з адитивними ефектами. *Адитивність* означає незалежність впливу одного фактора від рівня іншого. Забезпечити її можна шляхом стандартизації комбінаційного групування, себто заміною частот емпіричного розподілу частотами певного стандартного розподілу. Зважені по частотах стандартного розподілу середні *j*-ї групи за *i*-им фактором Y_{ij} називаються *стандартизованими*, а відхилення цих середніх від загальної середньої $\bar{y}$ — стандартизованими (центрованими) ефектами $a_{ij} = Y_{ij} - \bar{y}$. Незсунені та з мінімальною дисперсією оцінки ефектів a_{ij} дає стандартизація групувань методом найменших квадратів.

6.4. МОДЕЛЬ СТАНДАРТИЗОВАНИХ ГРУПУВАНЬ

У моделі МНК стандартизована середня Y_{ij} подається як регресія на структурних змінних u_{ij}:

$$Y_{ij} = \bar{y} + \sum_{i=1}^{m} \sum_{j=1}^{p_i} a_{ij} u_{ij},$$

де m — кількість факторів;

p_i — кількість груп за *i*-им фактором.

Параметри моделі a_{ij} характеризують чистий, елімінований від взаємодії ефект впливу *j*-го рівня *i*-го фактора. Оцінювання параметрів моделі здійснюється розв'язуванням системи нормальних рівнянь, елементами якої є групові $\sum u_{ij}$ і підгрупові $\sum u_{ij} u_{rj}$ частоти. Щоб уникнути лінійної залежності рівнянь, ефекти однієї з груп за кожним фактором прирівнюються до нуля. Трансформована таким чином модель набуває вигляду:

$$Y_{ij} = b_0 + \sum_{i=1}^{m} \sum_{j=1}^{p_i - 1} b_{ij} u_{ij},$$

де b_{ij} — різниця між груповими середніми j-ї групи і групи з нульовими ефектами (бази порівняння за i-им фактором).

Як і в загальній моделі МНК, систему нормальних рівнянь такої моделі можна записати у матричному вигляді:

$$U'y = U'UB;$$

$$B = U'y(U'U)^{-1},$$

де $U'U$ — матриця структурних змінних.

Щоб визначити стандартизовані ефекти за i-им фактором a_{ij}, необхідно зафіксувати на середньому рівні значення інших факторів. Це досягається шляхом приписування усім структурним змінним, окрім змінної за i-им фактором, їх середніх значень — часток груп d_{ij}. Структурній змінній j-ї групи за i-им фактором приписується одиниця, іншим групам за цим фактором — нулі. У матричному вигляді розрахунок стандартизованих середніх записується у вигляді скалярного добутку вектора B на вектор коефіцієнтів D:

$$Y_{ij} = D'B.$$

Середня похибка стандартизованого ефекту оцінюється за формулою

$$\mu_Y = s_\varepsilon \sqrt{D'(U'U)^{-1}D},$$

де s_e^2 — залишкова дисперсія.

Отже, схема стандартизації комбінаційних групувань МНК об'єднує два блоки. У першому визначаються параметри трансформованої моделі b_{ij}, у другому — вектор коефіцієнтів переходу від параметрів b_{ij} до стандартизованих ефектів a_{ij}.

Реалізацію цієї схеми розглянемо на прикладі професійної мобільності зареєстрованих безробітних. Як оцінку професійної мобільності використаємо частку безробітних, які виявили бажання пройти перенавчання для роботи за іншою професією.

У табл. 6.8 подано комбінаційний розподіл безробітних за віком (фактор А) та освітою (фактор Б). Вікову структуру представлено двома групами: $А_1$ — до 30 років, $А_2$ — 30 років і старше; освіту — трьома рівнями: $Б_1$ — ПТУ, $Б_2$ — середня спеціальна, $Б_3$ — вища.

Таблиця 6.8

Вік, років	Рівень освіти			Разом	Направлено на перенавчання	Рівень професійної мобільності
	ПТУ	середня спеціальна	вища			
До 30	150	40	10	200	60	0,30
30 і старше	120	130	50	300	60	0,20
Разом	270	170	60	500	120	0,24
Направлено на перенавчання	53	43	24	120	X	X
Рівень професійної мобільності	0,20	0,25	0,40	0,24	X	0,24

Як видно з даних таблиці, молодь (до 30 років) є професійно мобільнішою. Водночас простежується залежність професійної мобільності від рівня освіти: чим вищий рівень освіти, тим вища готовність здобути нову професію. Очевидно, що ці два фактори взаємопов'язані. Стандартизація групування МНК передбачає передусім ідентифікацію структурних змінних. Вилучивши за кожним фактором останню групу, дістанемо одну структурну змінну за фактором А і дві — за фактором Б. Рівняння регресії має вигляд:

$$Y = b_0 + b_{11}u_{11} + b_{21}u_{21} + b_{22}u_{22}.$$

Враховуючи, що $u_{ij}^2 = u_{ij}$, а $u_{ij}u_{ik} = 0$, де $j \neq k$, елементами матриці системи нормальних рівнянь $U'U$ та вектора $U'y$ будуть такі частоти:

$$U'U = \begin{vmatrix} n & \sum u_{11} & \sum u_{21} & \sum u_{22} \\ \sum u_{11} & \sum u_{11} & \sum u_{11}u_{21} & \sum u_{11}u_{22} \\ \sum u_{21} & \sum u_{11}u_{21} & \sum u_{21} & 0 \\ \sum u_{22} & \sum u_{11}u_{22} & 0 & \sum u_{22} \end{vmatrix} = \begin{vmatrix} 500 & 200 & 270 & 170 \\ 200 & 200 & 150 & 40 \\ 270 & 150 & 270 & 0 \\ 170 & 40 & 0 & 170 \end{vmatrix}$$

$$(U'y)' = (\sum y, \sum yu_{11}, \sum yu_{21}, \sum yu_{22}) = (120; 60; 53; 43).$$

Розв'язавши систему рівнянь, дістаємо параметри $B' = (0{,}3470;\ 0{,}1557;\ -0{,}2642;\ -0{,}1577)$, які оцінюють ефекти відповідних груп за i-им фактором щодо вилученої групи. На основі цих параметрів визначаються стандартизовані середні Y_{ij} та ефекти a_{ij}:

$$Y_{ij} = b_0 + b_{ij}u_{ij} + \sum_{1}^{m-1} \sum_{1}^{p_r} b_{rj}d_{rj};$$

де d_{rj} – частка j-ї групи за r-им фактором ($i \neq r$).

$$a_{ij} = Y_{ij} - \bar{y}.$$

Сформуємо вектори коефіцієнтів переходу від регресійної моделі до стандартизованих середніх Y_{ij} за даними табл. 6.8 ($m = 2$; $p_1 = 2$; $p_2 = 3$):

$$D_{11} = (1, 1, 0, d_{21}, d_{22}, d_{23})$$
$$D_{12} = (1, 0, 1, d_{21}, d_{22}, d_{23})$$
$$D_{21} = (1, d_{11}, d_{12}, 1, 0, 0)$$
$$D_{22} = (1, d_{11}, d_{12}, 0, 1, 0)$$
$$D_{23} = (1, d_{11}, d_{12}, 0, 0, 1).$$

Комбінаційний розподіл безробітних (табл. 6.8) характеризується частками: $d_{11} = 200 : 500 = 0{,}4$; $d_{12} = 0{,}6$; $d_{21} = 270 : 500 = 0{,}54$; $d_{22} = 170 : 500 = 0{,}34$; $d_{23} = 0{,}12$. Звідси стандартизована середня для першої групи за фактором A становить:

$$Y_{11} = 0{,}3470 + 0{,}1557 + 0 + (-0{,}2642)\cdot 0{,}54 + (-0{,}1577)\cdot 0{,}34 + 0 = 0{,}306.$$

Відповідно центрований ефект цієї групи $a_{11} = 0{,}306 - 0{,}240 = +0{,}066$, тобто професійна мобільність молоді, незалежно від освіти, в середньому на 6,6% вища за середній рівень. Аналогічно визначені стандартизовані середні та центровані ефекти для інших груп наведено в табл. 6.9.

Таблиця 6.9

Номер групи, ij	11	12	21	22	23
Y_{ij}	0,306	0,152	0,145	0,251	0,409
a_{ij}	0,066	–0,088	–0,095	0,011	0,169

Згідно з даними таблиці професійна мобільність осіб старшого віку на 8,8% нижча за середній рівень. Щодо освіти, то незалежно від віку найбільший ефект професійної мобільності дає вища освіта (+16,9%). Для випускників ПТУ, навпаки, характерний низький рівень професійної мобільності, ефект цієї групи становить –9,5%.

Отже, за допомогою стандартизації ефекти впливу факторів, представлених ознаками різного типу (метричними, номінальними), приводяться до порівнянного виду. І це значно розширює аналітичні можливості регресійних моделей.

Завдання для самоконтролю

1. Вплив стану навколишнього середовища на здоров'я населення описується регресійною моделлю, ознакова множина якої включає: x_1 — середні викиди забруднюючих речовин в атмосферу на одну людину за рік, кг; x_2 — місце проживання: великі промислові центри (I група) та території з невисоким рівнем техногенного навантаження (II група). Коефіцієнти регресії моделі онкозахворювань на 100 000 чол. становлять: $b_1 = 0{,}414$; $a_1 = 112{,}380$. Поясніть зміст коефіцієнтів регресії.

2. Регресійна модель описує залежність продуктивності праці робітників очисних вибоїв (тонн на одного робітника за зміну) від потужності вугільного пласта (x_1) та типу вугільного комбайна (широкозахватні, вузькозахватні).
Параметри моделі становлять:

R^2	a_0	b_1	a_1	c_{11}
0,76	3,872	1,875	1,290	2,754

Поясніть зміст параметрів моделі.

3. Виконайте специфікацію моделей окупності витрат y залежно від продуктивності праці x_1 та оборотності обігових коштів x_2:

1) для птахофабрик, різних за спеціалізацією виробництва (м'ясні, яєчні);

2) для теплоелектростанцій, які використовують різні види палива (вугілля, природний газ, мазут);

3) для агрогосподарств, які мають переробні цехи (плодоконсервні, виноробні) та різняться за ступенем розвитку зовнішньоекономічних зв'язків (експортують продукцію, не експортують).

4. Через погіршання фінансового стану компанії чистий дохід на акції стрімко зменшився. Опишіть динаміку дохідності акцій лінійним трендом з урахуванням зрушень. Поясніть зміст параметрів моделі.

Рік	1	2	3	4	5	6	7
Дохід на акцію, грн.	1,12	1,23	1,33	1,46	1,52	1,56	1,62

5. За наведеними даними сформуйте матрицю ознакової множини моделі врожайності кукурудзи; здійсніть специфікацію моделі; визначте ефекти впливу та ефекти взаємодії факторів: A — режиму іригації, B — сорту, їх взаємодії.

Режим іригації	Урожайність сорту, ц/га		
	B_1	B_2	B_3
A_1	43	46; 52; 58	47; 53
A_2	56; 49; 50	66; 71	62

6. За допомогою моделі стандартизованих групувань проведено аналіз залежності розміру кредитно-інвестиційного портфеля комерційних банків від розміру капіталу x та еволюційних факторів розвитку банківської системи (змінна часу t). Центровані ефекти впливу цих факторів становлять:

Групи за ознакою x	Фактичний розмір КІП, млн. грн	Центрований ефект
1	6,2	–11,3
2	8,2	–9,1
3	10,4	–7,0
4	11,3	–6,1
5	20,9	3,8
6	33,8	16,7
7	9,8	–7,2
У середньому	17,3	X
Роки 1	15,0	–1,0
2	17,3	–0,2
3	19,5	1,3
У середньому	17,3	X

Визначте стандартизовані середні для кожної групи і, розглядаючи їх як теоретично можливі рівні, оцініть ступінь використання теоретичних можливостей банків щодо збільшення КІП.

7. Характеристикою потенційної професійної мобільності є частка робітників, які прагнуть змінити професію. На конкретному підприємстві цей показник становить 0,240. Механізм формування потенційної мобільності робітників підприємства описано регресійною моделлю, параметри якої наведено в таблиці:

Групи робітників за ознаками	Частка робітників	Коефіцієнт регресії
1. Вік, років:		
до 25	0,345	0,090
25—35	0,380	–0,067
35 і старші	0,275	—
2. Ступінь задоволеності професією:		
не подобається	0,233	0,117
ставлюся байдуже	0,093	0,054
подобається	0,674	—
Рівень кваліфікації (тарифний розряд):		
3	0,328	–0,126
4	0,252	0,073
5, 6	0,420	—

Вільний член рівняння $a_0 = 0{,}512$.

Визначте стандартизовані середні та центровані ефекти впливу факторів на професійну мобільність робітників. Зробіть висновки.

Розділ 7

БАГАТОФАКТОРНЕ ПРОГНОЗУВАННЯ

7.1. ОСОБЛИВОСТІ МОДЕЛЮВАННЯ ВЗАЄМОЗВ'ЯЗАНИХ ДИНАМІЧНИХ РЯДІВ

Якщо інформаційна база регресійної моделі представлена рядами динаміки, то виникають певні методологічні труднощі, спричинені залежністю рівнів, їх автокореляцією. Наявність останньої порушує одну з передумов регресійного аналізу — незалежність спостережень — і призводить до викривлення його результатів.

У практиці регресійного аналізу застосовують різні способи усунення автокореляції. Найпростішим є спосіб різницевих перетворень, коли замість первинних рівнів взаємозв'язаних рядів динаміки y_t, x_t використовують абсолютні прирости (різниці). Так, різниці першого порядку $\Delta y = y_t - y_{t-1}$ та $\Delta x = x_t - x_{t-1}$ усувають лінійний тренд, однофакторна регресія набуває такого вигляду:

$$\Delta y = a + b\Delta x_1,$$

де b інтерпретується як звичайний коефіцієнт регресії; a — вільний член рівняння.

Якщо тенденція нелінійна, доцільно застосувати спосіб відхилень від тенденції, коли первинні рівні y_t, x_t замінюються відхиленнями від тренда $d_y = y_t - f_{(t)}$; $d_x = x_t - f_{(t)}$.

Усуненню автокореляції сприяє також уведення фактора часу t у рівняння регресії $Y = f(x_1, t)$. Навантаження на змінну t залежить від комплексу включених у модель факторів. Зміст параметрів такої моделі розглянемо на прикладі взаємозв'язку динаміки імпорту нафти y_t і цін за барель нафти x_t на світовому ринку. За даними табл. 7.1, обсяги імпорту нафти в країну систематично зменшувалися, що зумовлено як зміною цін, так і внутрішніми факторами. Зв'язок між цими показниками можна подати лінійною функцією

$$Y = a + bx + ct,$$

де b — середній приріст результативної ознаки y на одиницю приросту факторної ознаки x;

c — середній щорічний приріст y під впливом зміни неідентифікованих факторів, які рівномірно змінюються в часі.

Таблиця 7.1

Порядковий номер року	Імпорт нафти, y_t, млн. барелів	Ціна за 1 барель, x_t, дол.	Y_t	$e_t = y_t - Y_t$
1	1749	13,48	1808	–59
2	1702	14,76	1743	–41
3	1769	18,92	1653	116
4	1600	22,97	1562	38
5	1431	30,29	1442	–11
6	1325	34,66	1349	–24
7	1302	30,77	1332	–30
8	1341	29,36	1292	49
9	1232	28,07	1251	–19
10	1180	26,40	1213	–33
11	1162	27,79	1147	15
Разом	15793	×	15793	0

Модель імпорту нафти описується рівнянням:

$$Y = 1984{,}340 - 2{,}497x_1 - 52{,}986t$$
$$(27{,}97) \quad (-2{,}50) \quad (-6{,}99).$$

Наведені в дужках значення t-критерію перевищують критичне $t_{0,95}(8) = 2{,}31$, що дає підстави з імовірністю 0,95 вважати вплив кожного фактора на обсяги імпорту істотним. Згідно із значеннями коефіцієнтів регресії підвищення ціни одного бареля нафти на 1 долар зменшує імпорт нафти в країну в середньому на 2,5 млн. барелів. За рахунок інших факторів, передусім політики енергозбереження, імпорт нафти щорічно зменшується в середньому на 53 млн. барелів.

Значення коефіцієнта детермінації $R^2 = 0{,}951$ та дисперсійного критерію $F(2,8) = 77{,}48$ свідчать про адекватність моделі.

Отже, за наявності лінійної тенденції в рядах у модель вводиться змінна часу

$$Y = a_0 + \sum_{1}^{m} b_i x_i + ct,$$

де b_i — чистий ефект впливу i-го фактора на y;

c — ефект неідентифікованих факторів, які формують тенденцію ряду.

У динамічній моделі можна відобразити не лише тенденцію, а й більш складні компоненти ряду, скажімо, періодичні чи сезонні коливання, перервність процесу тощо.

Особливістю регресійного аналізу динамічних рядів є оцінка автокореляції залишкових величин $e_t = y_t - Y_t$. Якщо автокореляція істотна, значить включені в модель фактори не повністю розшифровують механізм формування процесу, модель визнається неадекватною. Перевірку істотності автокореляції можна здійснити на основі циклічного коефіцієнта першого порядку r_1 (див. 4.1).

У програмних засобах для перевірки істотності автокореляції частіше використовують критерій Дарбіна-Ватсона, характеристика якого D функціонально зв'язана з r_1

$$D = \frac{\sum_{2}^{n}(e_t - e_{t+1})}{\sum_{1}^{n} e_t^2} = 2\,(1 - r_1).$$

За відсутності автокореляції між суміжними членами ряду значення D становить приблизно 2, при високій додатній автокореляції D наближається до 0, при високій від'ємній автокореляції — до 4. Визначені критичні межі його значень: нижня D_L і верхня D_U, на основі яких приймається або відхиляється гіпотеза про відсутність автокореляції: $H_0 : r_1 = 0$.

При перевірці гіпотези можливі три висновки:

- $D > D_U$ — автокореляція відсутня;
- $D < D_L$ — гіпотеза про відсутність автокореляції відхиляється;
- $D_L \le D \le D_U$ — висновок залишається невизначеним.

Критичні межі D залежать від кількості членів ряду n і кількості параметрів моделі m. У додатку 8 наведено критичні значення D для додатної автокореляції при $\alpha = 0{,}05$. Перевірка від'ємної автокореляції проводиться на основі значень $(4 - D)$.

У модулі *Multiple Regression* для перевірки істотності автокореляції залишкових величин у вікні *Residual Analysis* передбачена опція *Durbin-Watson stat*. За даними табл. 7.1 $D = 1{,}831$, що потрапляє в інтервал допустимих значень гіпотези H_0: $r_1 = 0$, а отже, істотність автокореляції не доведено. Аналогічний висновок дає перевірка гіпотези за допомогою циклічного коефіцієнта автокореляції, значення якого $r_1 = 0{,}085$ значно менше за критичне $r_{0,95}(11) = 0{,}353$. Відсутність автокореляції залишків підтверджує адекватність моделі.

Характерною рисою механізму формування варіації та динаміки соціально-економічних показників є запізнення впливу факторів, коли причина і наслідок розірвані в часі (наприклад, інвестиції в іригацію і введення в дію зрошувальних земель). Часові лаги зумовлені тривалістю виробничого циклу, інерційністю процесів, наявністю зворотного зв'язку тощо. Для оцінювання ефектів запізнення впливу i-го фактора в модель вводиться лагова змінна $x_{i,p}$. Фактори, що мають два і більше лагів (*розподілений у часі лаг*), вводяться в модель блоками лагових змінних. Загальний вигляд моделі з розподіленими лагами:

$$Y = a_{00} + \sum_{i=1}^{m}\sum_{p=0}^{k} b_{i,p} x_{i,p},$$

де $p = 0, 1, \ldots, k$ — лаги;

m — кількість включених у модель факторів.

Теоретично модель з розподіленими лагами можна узагальнити на будь-яку кількість факторів, проте практична реалізація такої моделі натикається на непереборні труднощі, зумовлені обмеженістю динамічних рядів і складністю внутрішньої їх структури. Як правило, в модель включаються такі лагові змінні, для яких лаги обґрунтовано теоретично і перевірено емпірично. Інструментом визначення лагів слугує взаємокореляційна функція, яка являє собою множину коефіцієнтів кореляції між рядами x_t та y, зсуненими відносно один до одного на лаг p. Зі збільшенням лага взаємокореляційна функція згасає. У табл. 7.2 наведено коефіцієнти кореляції між попитом на легкові автомобілі y та двома факторами: середньодушовим доходом x_1 та цінами x_2.

Таблиця 7.2

Лаг	x_1	x_2
0	0,823	0,612
1	0,646	0,441
2	0,416	0,187
3	0,098	0,098

Для фактора x_1 істотними виявилися лаги $p = 0,1,2$; для фактора x_2 — лаги $p = 0,1$. Модель набуває вигляду:

$$Y = a_0 + b_{10}x_1 + b_{11}x_{1,t-1} + b_{12}x_{1,t-2} + b_{20}x_2 + b_{21}x_{2,t-1} + ct,$$

де параметри $b_{1,p}$ і $b_{2,p}$ характеризують ефекти впливу факторів з відповідними лагами, параметр c — вплив неідентифікованих факторів (мода, смаки тощо).

7.2. ДИНАМІЧНА МОДЕЛЬ ДЛЯ СУКУПНОСТІ ОБ'ЄКТІВ

Через обмеженість динамічних рядів соціально-економічних явищ неможливо врахувати в моделі усі особливості розвитку процесу. Аби розширити інформаційну базу моделі, практикують об'єднання просторових і динамічних рядів. Скажімо, описується залежність $Y = f(x_1, x_2, x_3)$ за даними по 10 об'єктах за п'ять років. Можливі різні варіанти використання такої змішаної статично-динамічної інформації. Розглянемо два з них.

1. *Динамізація просторових моделей.* Для кожного t-го року визначається статична модель $Y_t = f(x_1, x_2, x_3)$. У нашому прикладі їх буде п'ять. Коефіцієнти регресії статичних моделей утворюють динамічні ряди. Якщо ефект впливу i-го фактора змінюється в часі, то така зміна виявиться трендом ряду b_i. Методом екстраполяції тренда можна визначити очікуваний ефект впливу на період упередження v. Водночас визначається прогнозний рівень самого фактора $x_{i,t+v}$. Поєднання цих прогнозів дає прогноз функції y:

$$Y_{t+v} = a_{0,\,t+v} + b_{1,\,t+v}x_{1,\,t+v} + b_{2,\,t+v}x_{2,\,t+v} + b_{3,\,t+v}x_{3,\,t+v}.$$

За відсутності тренда коефіцієнта регресії в прогнозній моделі використовують середнє його значення. В табл. 7.3 наведено фрагменти динамічних рядів параметрів регресійної моделі продуктивності праці в цементній промисловості (тонн на одного робітника). Фактори: x_1 — енергоозброєність праці, кВт-г; x_2 — продуктивність цементних печей т/г; x_3 — коефіцієнт використання календарного часу роботи цементних печей.

Таблиця 7.3

Рік	b_1	b_2	b_3
1	11,8	11,3	18,5
2	11,5	11,9	19,1
3	11,3	12,2	17,7
4	10,6	13,4	18,2
5	9,9	13,7	18,6

Як видно з даних таблиці, в цементній промисловості відбувається перерозподіл ефектів впливу факторів на продуктивність

праці: зменшується вплив енергоозброєності праці (x_1), збільшується вплив продуктивності устаткування (x_2) і практично незмінним залишається вплив використання календарного часу устаткування (x_3).

Прогнозування ефектів впливу факторів та їх рівнів можна здійснити у будь-який спосіб, обґрунтувавши функціональний вид прогнозної моделі (див. 4.2—4.3). Звісно, щоб характер динаміки чітко виявився, довжина динамічного ряду має бути достатньою. Умова достатності інформації стосується і просторового ряду.

2. *Модель об'єкто-періодів.* У невеликих за обсягом сукупностях просторові та динамічні ряди об'єднуються в один інформаційний масив, одиницею якого є об'єкто-період. Для 10 об'єктів і п'яти років маємо 10·5=50 об'єкто-періодів. Такий підхід до об'єднання просторово-динамічних рядів значно розширює інформаційну базу моделі, водночас наділяє її особливими властивостями. Головна особливість статично-динамічної інформації — залежність спостережень. Залежними виявляються не лише рівні динамічних рядів, але й ряди в цілому (і просторові, і часові), оскільки належність рівнів до того чи іншого ряду фіксована. Так, залежність між рядами динаміки — це результат просторової варіації, яка через інерційність процесів зберігається певний час. Залежність просторових рядів відбиває синхронність динаміки показників по окремих об'єктах, зумовлену спільними умовами розвитку. Ігнорування цих особливостей інформаційної бази моделювання призводить до помилкових висновків.

Особливості просторової варіації враховуються в моделі за допомогою структурних змінних окремих об'єктів u_j. Властивий усім об'єктам тренд функції y фільтрується за допомогою змінної часу t. Проте через нерівномірність розвитку окремих об'єктів сукупності поряд зі спільним трендом можуть виявитися істотними індивідуальні тренди. Для їх фільтрації можна використати змінні *динамічної взаємодії*: для факторів — $x_i\,t$, для об'єктів — $u_j t$. З урахуванням усіх цих особливостей регресійну модель для сукупності об'єкто-періодів можна записати так:

$$Y = a_0 + \sum_{i=1}^{m} b_i x_i + \sum_{i=1}^{m} c_i x_i t + \sum_{j=1}^{n-1} a_j u_j + \sum_{j=1}^{n-1} d_j u_j t + ft\,.$$

Параметри моделі вимірюють:

b_i — чистий, елімінований від взаємозв'язків у межах моделі, ефект впливу фактора x_i;

c_i — зміну ефектів впливу b_i у часі;

a_j — різницю між значеннями функції на j-му об'єкті та в цілому по сукупності;

d_j — зміну цих відмінностей у часі;

f — спільний для всіх об'єктів сукупності тренд — вплив неідентифікованих в моделі факторів;

a_0 — вільний член рівняння. Для кожного j-го об'єкта вільний член рівняння дорівнює сумі (a_0+a_j); на відміну від a_0 сума має економічний зміст — вимірює вплив факторів, які визначають специфіку цього об'єкта.

Отже, модель об'єкто-періодів включає дві групи параметрів. Одна з них представляє оцінки ефектів впливу факторів і зміну їх у часі, друга — особливості сукупності, специфіку розвитку окремих об'єктів. Уникнути перевантаження моделі і зберегти максимум інформації для оцінки параметрів можна, скориставшись алгоритмом покрокового регресійного аналізу.

Як приклад розглянемо параметри моделі продуктивності праці в агрогосподарствах, які спеціалізуються на вирощуванні винограду та фруктів і мають власні переробні цехи. Інформаційний масив сформовано за даними 18 господарств за п'ять років. До ознакової множини моделі включено фактори: x_1 — економічна оцінка сільськогосподарських угідь, бали; x_2 — частка садів і виноградників у загальній площі сільськогосподарських угідь; x_3 — частка праці механізаторів у загальній кількості відпрацьованих людино-днів. Для оцінювання тенденцій ефектів впливу кожного з цих факторів введено змінні динамічної взаємодії $x_i t$. Два нетипових (аномальних) господарства представлено в моделі структурними змінними u_j, а індивідуальні їх тренди — змінними динамічної взаємодії $u_j t$.

Параметри моделі визначалися за процедурами модуля *Multiple Regression*. Істотними виявилися ефекти впливу всіх факторів (b_1, b_2, b_3), параметр при змінній динамічної взаємодії другого фактора (c_{2t}), параметри при структурних змінних обох господарств (a_1, a_2), параметр при змінній динамічної взаємодії другого господарства (d_{2t}). Значення параметрів наведено в табл. 7.4.

Таблиця 7.4

Параметр моделі	b_1	b_2	b_3	c_2t	a_1	a_2	d_2t
Значення параметра	39,86	15.63	20,46	1,17	– 42,65	56,78	–3,52

Коефіцієнти регресії b_i інтерпретуються традиційно як чисті ефекти впливу включених у модель факторів. При цьому, як показує параметр c_2t , ефект впливу спеціалізації (частки садів і виноградників у загальній площі сільськогосподарських угідь) на продуктивність праці щорічно збільшується в середньому на 1,17 тис. грн. Істотність параметрів a_1 і a_2 підтверджує нетиповість господарств, представлених у моделі структурними змінними. За рахунок специфічних умов функціонування цих господарств рівень продуктивності праці на першому з них нижчий за середній на 42,65 тис. грн, на другому, навпаки, на 56,78 тис. грн. вищий. Останній параметр має тенденцію до зменшення щорічно в середньому на 3,52 тис. грн.

Отже, модель об'єкто-періодів більш універсальна і повніше використовує інформацію про взаємозв'язки порівняно зі схемою динамізації просторових моделей. Прогнозні можливості моделі реалізуються процедурою *Predict dependent var.* модуля *Multiple Regression.*

7.3. НЕЛІНІЙНА РЕГРЕСІЯ

При моделюванні взаємозв'язків на динамічних рядах широко використовуються відносні величини, передусім індекси. Це пояснюється більшою їх сталістю в часі порівняно з абсолютними величинами. Окрім того, з'являється можливість виключити мультиколінеарність та автокореляцію залишків. Описуються такі взаємозв'язки степеневою функцією

$$Y = Ax_1^{b_1} x_2^{b_2} x_3^{b_3} \ldots x_m^{b_m},$$

де b_i — коефіцієнт еластичності, який показує, на скільки % у середньому зміниться y зі зміною x_i на 1 % за умови незмінності інших факторів. Тобто коефіцієнт еластичності — це відносний ефект впливу i-го фактора на y.

Степенева функція лінійна в логарифмах, а тому параметри її визначаються МНК. Класичним прикладом такого типу нелінійної функції є виробнича функція Кобба-Дугласа, яка описує співвідношення між факторами і результатом виробництва на будь-якому рівні економічної діяльності (фірма, галузь, регіон, економіка в цілому):

$$Q = A\,K^{\alpha} L^{\beta},$$

де Q — результат виробництва;

K — основний капітал;

L — трудові затрати (кількість зайнятих).

Параметри α і β — коефіцієнти еластичності: α характеризує відносний приріст результату на одиницю приросту капіталу при $L = const$, а β — відносний приріст результату на одиницю приросту трудових затрат при $K = const$. Капітал і трудові затрати розглядаються як фактори екстенсивного розвитку (залучення нових ресурсів). При трудомісткому виробництві $\alpha > \beta$, при фондомісткому — $\beta > \alpha$. У виробничій функції закладено умову, за якою $(\alpha + \beta) = 1$, тобто результат зростає у такій же пропорції, як і фактори. Параметр A приводить масштаб (розмірність) факторів до масштабу результату. При використанні індексів $A = 1$, а тренд результату, зумовлений дією інших, неекстенсивних факторів, враховується в моделі змінною часу $e^{\lambda t}$. Модель набуває вигляду:

$$Q = A K^{\alpha} L^{\beta} e^{\lambda t},$$

де λ характеризує темп приросту функції за рахунок неекстенсивних факторів, зокрема неуречевлених факторів НТП (вдосконалення технології, зростання кваліфікації робітників тощо).

Цей варіант моделі називають виробничою функцією Тінбергена. Застосувавши до неї логарифмічне диференціювання, дістанемо модель, яка описує взаємозв'язок темпів приросту:

$$q = \alpha k + \beta l + \lambda,$$

де q, k, l — темпи приросту відповідно результату, капіталу й трудових затрат.

На основі такої моделі можна визначити внесок екстенсивних та інтенсивних факторів у розвиток процесів відтворення:

$$d_{екс} = (\alpha k + \beta l) / q;$$
$$d_{інт} = \lambda / q.$$

У табл. 7.5 наведено параметри макроекономічних функцій трьох індустріально розвинутих країн за період з 1950 по 1977 рр.: Q — валовий національний продукт; K — основні фонди, L — трудові ресурси [14, с. 41].

Таблиця 7.5

Країна	α	β	λ
США	0,447	0,553	1,34
Великобританія	0,506	0,494	0,53
Японія	0,397	0,603	4,66

За значеннями параметрів функції можна зробити висновки про особливості економічного розвитку кожної з цих країн у по-

воєнні роки. Так, досягнення НТП найінтенсивніше впроваджувалися в економіці Японії: параметр λ вищий порівняно із США в 3,5 раза, порівняно з Великобританією — у 9 разів. Водночас японська економіка характеризується найнижчою капіталоємністю ($\alpha = 0{,}397$) і відносно високим рівнем ефективності використання трудових ресурсів ($\beta = 0{,}603$). Для американської економіки характерна збалансованість співвідношення еластичностей по капіталу і труду. Середньорічні темпи приросту в США у повоєнні роки становили (у %): валового національного продукту — 3,38, основних фондів — 2,79, трудових ресурсів — 1,46. Звідси внесок факторів у формування динаміки ВНП :

$$\text{екстенсивних } d_{екс} = \frac{0{,}447 \cdot 2{,}79 + 0{,}553 \cdot 1{,}46}{3{,}38} = 0{,}625;$$

$$\text{інтенсивних } d_{інт} = \frac{1{,}34}{3{,}38} = 0{,}375.$$

На практиці використовують різні модифікації виробничої функції. Наприклад, розділивши обидві її частини на L, отримаємо функцію продуктивності праці:

$$W = AF^{\alpha} e^{\lambda t},$$

де W — продуктивність праці;
F — фондоозброєність праці.

У темпах приросту ця функція записується так:

$$w = \alpha f + \lambda = \alpha (k - l) + \lambda.$$

Внесок екстенсивних та інтенсивних факторів у динаміку продуктивності праці визначається аналогічно:

$$d_{екс} = \alpha (k - l) / w;$$
$$d_{інт} = \lambda / w.$$

До темпів приросту застосовують класичну регресію. Як приклад розглянемо зв'язок між темпами приросту фондоозброєності (VAR1) і продуктивності праці (VAR2) в одній із галузей промисловості (табл. 7.6).

Таблиця 7.6

Показник	Ланцюгові темпи приросту, %						
	1993	1994	1995	1996	1997	1998	1999
VAR1	1,4	2,5	1,7	0,9	1,0	2,6	2,8
VAR2	2,6	2,7	2,8	1,5	1,8	2,4	3,2

Результати розрахунків за стандартною процедурою модуля *Multiple Regression* наведено в табл. 7.7. Параметри моделі становлять: $\alpha = 0{,}581$; $\lambda = 1{,}358$, тобто приріст фондоозброєності праці на 1 % спричиняє зростання продуктивності праці на 0,581%. За рахунок інших факторів щорічний приріст продуктивності праці становить у середньому 1,358 %. Середньорічні темпи приросту фондоозброєності праці — 1,8 %, продуктивності праці — 2,4 %. Звідси внесок екстенсивних факторів у динаміку продуктивності праці становить $(1{,}8 \cdot 0{,}581) : 2{,}4 = 0{,}435$, інтенсивних — $1{,}358 : 2{,}4 = 0{,}565$.

Таблиця 7.7

Regression Summary for Dependent Variable: VAR2 (---.sta)					
R= ,7762 RI= ,6024 Adjusted RI= ,5229 F(1,5)=7,577 p<,04019 Std.Error of estimate: ,408					
N = 7	BETA	B	St. Err. of B	t(5)	p-level
Intercpt		1,358	,418	3,245	,0223
VAR1	,776	,581	,211	2,753	,0402

Степеневою функцією описується також взаємозв'язок між попитом C, середньодушовим доходом населення D і цінами на товар P. Тренд попиту, зумовлений звичками, модою тощо, вводиться в модель змінною часу $e^{\lambda t}$:

$$C = A\, D^{\alpha}\, P^{\beta}\, e^{\lambda t},$$

де α і β — коефіцієнти еластичності попиту залежно від доходу та цін.

При побудові нелінійних моделей у вигляді степеневих функцій значення всіх показників, окрім ознаки часу t, замінюються власними логарифмами. В системі *Statistica* така заміна здійснюється на етапі специфікації ознак (див. 2.1).

Завдання для самоконтролю

1. Попит на сталь за помісячними даними описується моделлю:

$$Y = 1{,}5 - 1{,}27\, x_1 + 6{,}22\, x_1 t + 4{,}65\, x_2 - 0{,}03\, t,$$

де y — млн. т проданої сталі;

x_1 — ціна сталі, центів за фунт;

x_2 — індекс промислового виробництва.

Поясніть зміст параметрів моделі. Який з параметрів відображає кон'юнктуру ринку, а який — спекулятивний елемент попиту?

2. Модель особистого споживання за 1989—1999 рр. має вигляд:

$$Y = 240{,}1 + 0{,}72\, x_t + 0{,}12\, x_{t-1},$$

де y — середньодушове споживання, грн.;

x — середньодушовий дохід, грн.

Коефіцієнт детермінації становить 0,97; критерій Дарбіна-Ватсона — 2,24. Дайте інтерпретацію параметрів моделі, оцініть її адекватність.

Визначте прогнозний рівень середньодушового споживання в 2000 р. за умови, що в 1999 р. реальний середньодушовий дохід становив 2250 грн., а в 2000 р. дефльоване значення цього показника зросте на 10 %.

3. Модель урожайності зернових культур у регіоні за 1990—1999 рр. має вигляд:

$$Y = 2{,}10 + 0{,}272\, x_1 + 0{,}374\, x_2 - 3{,}15\, u + 0{,}210\, t,$$

де x_1 — родючість землі, балів;

x_2 — кількість внесених добрив на 1 га, центнерів поживної речовини (ц п. р.);

u — метеорологічні умови року (для сприятливих років $u = 0$, для несприятливих — $u = 1$).

Коефіцієнт детермінації становить 0,72, коефіцієнт автокореляції залишкових величин $r_1 = 0{,}36$. Дайте інтерпретацію параметрів моделі, перевірте істотність автокореляції залишкових величин.

Визначте теоретичний рівень урожайності для несприятливого 1999 р., якщо відомо значення факторів: $x_1 = 65{,}1$ бала; $x_2 = 35$ ц п. р.

Визначте прогнозні рівні врожайності на 2000 р. :

— оптимістичний, для сприятливих метеорологічних умов;

— песимістичний, для несприятливих метеорологічних умов.

У 2000 р. передбачається збільшення кількості внесених добрив порівняно з 1999 р. на 5 %. Родючість землі ймовірно не зміниться.

4. Динаміка військових витрат країни за даними 1989—1999 рр. описується моделлю:

$$Y = -23{,}463 + 0{,}87\, y_{t-1} + 0{,}0215\, x_{t-1},$$

де x — валовий національний продукт;

y — військові витрати.

Коефіцієнт детермінації становить 0,85, критерій Дарбіна-Ватсона — 1,72. Дайте інтерпретацію параметрів моделі, оцініть її адекватність.

Визначте прогнозні рівні військових витрат на 2000—2002 рр., якщо у 1999 р. валовий національний продукт становив 2310 млрд. дол. США, військові витрати — 133,2 млрд. дол. США. Очікуваний рівень валового національного продукту на 2000—2002 рр.:

Рік	2000	2001	2002
x	2346	2395	2412

5. Моделі товарного імпорту (без нафти та нафтопродуктів) описуються параметрами:

Фактори	Коефіцієнти еластичності	
	1981—1990 рр.	1991—2000 рр.
Обсяг валового національного продукту	2,057	2,280
Співвідношення імпортних і внутрішніх цін	–0,228	–0,521
Індекс завантаженості виробничих потужностей	0,004	–0,297

Дайте економічну інтерпретацію параметрів моделі; порівняйте їх у часі. Зробіть висновки щодо перерозподілу впливу факторів.

6. Залежність інфляції (динаміки споживчих цін) від темпів нарощування грошової маси (агрегат М0) описується функцією: $Y = 0{,}8x^{0{,}265}$.

Визначте очікуваний рівень інфляції, якщо грошову масу збільшити: а) на 10 %; б) на 20 %.

7. За наведеними даними визначте темпи приросту ВВП кожної країни, оцініть вплив на динаміку ВВП екстенсивних та інтенсивних факторів.

Країна	Основні фонди		Трудові ресурси		λ
	Темп приросту	Еластичність	Темп приросту	Еластичність	
A	4,45	0,47	1,2	0,53	1,43
B	6,32	0,35	1,7	0,65	4,60

8. За даними про темпи приросту та еластичність фондоозброєності праці в різних галузях промисловості визначте темпи приросту продуктивності праці, а також вплив на динаміку продуктивності праці екстенсивних та інтенсивних факторів.

Галузь промисловості	Еластичність фондоозброєності	Темп приросту фондоозброєності	λ
A	0,365	2,2	0,53
B	0,468	3,4	1,12
C	0,268	1,9	1,75

9. За трирічними поквартальними даними восьми підприємств цементної промисловості побудовано модель продуктивності праці (виробництво цементу на одного робітника, т). Ознакову множину моделі формують: x_1 — коефіцієнт забезпеченості основними напівфабрикатами (клінкером) власного виробництва; x_2 — коефіцієнт використання календарного фонду робочого часу цементних печей; x_3 — погодинна продуктивність цементних млинів, т. Істотними виявилися параметри при структурних змінних: u_1 — цементно-шиферного комбінату; u_2 — цементно-гірничого комбінату. Коефіцієнт детермінації — 0,925. Параметри моделі становлять:

Ознака	Коефіцієнт регресії		
	факторної ознаки	структурної змінної	змінної динамічної взаємодії
x_1	–0,52	—	—
x_2	0,85	—	0,04
x_3	1,04	—	—
u_1	—	67,12	– 4,32
u_2	—	–81,43	3,06

Поясніть зміст параметрів моделі. Зробіть специфікацію моделі для цементно-шиферного та цементно-гірничого комбінатів. Як Ви оцінюєте прогнозні властивості моделі?

Розділ 8

МОДЕЛЮВАННЯ ПРИЧИННИХ КОМПЛЕКСІВ

8.1. СТРУКТУРА ВЗАЄМОЗВ'ЯЗКІВ І СТРУКТУРНА ФОРМА МОДЕЛІ

Складність і багатогранність взаємозв'язків, наявність зворотного впливу зумовлюють необхідність використання моделей у вигляді системи взаємозалежних рівнянь. Розрізняють два типи таких систем. В одних системах рівняння описують послідовний ланцюг причинно-наслідкових зв'язків, що уможливлює послідовне їх розв'язування. Іншим системам властиві зворотні зв'язки, коли та сама змінна одночасно виступає і як причина, і як наслідок. У такому разі рівняння необхідно розв'язувати одночасно.

Логіко-методологічна схема побудови будь-якої системи рівнянь включає: формування логічного каркасу моделі та специфікацію рівнянь. Логічний каркас моделі можна представити геометрично у вигляді *графа зв'язку* або таблично у вигляді *матриці суміжності*. На рис. 8.1 зображено граф зв'язку з чотирма вершинами. Дуги графа відображують напрям зв'язку; послідовність дуг одного напряму називають *шляхом*, а кількість послідовно з'єднаних дуг — *довжиною шляху*. На основі графа можна простежити причинні ланцюги (прямі та опосередковані), виявити контури зворотного зв'язку тощо.

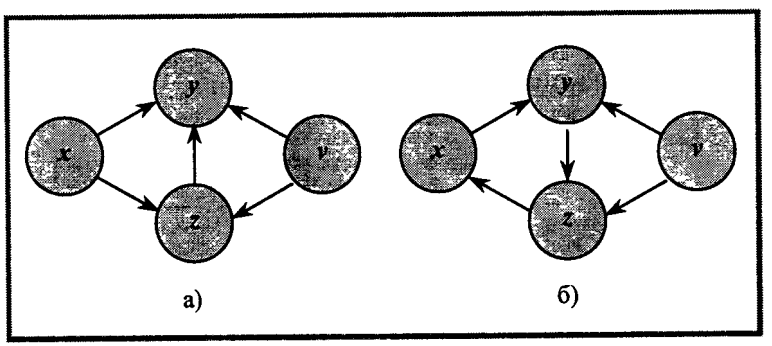


Рис. 8.1. Графи зв'язку:
а) послідовного прямого; б) одночасно реалізованого прямого і зворотного

При складному переплетенні взаємозв'язків шляхи впливу будь-якої довжини можна формалізувати за допомогою квадратної матриці порядку m з одиничними і нульовими елементами залежно від наявності (відсутності) дуг між вершинами x_i та x_k. Стовпці такої матриці асоціюються з причинами, а рядки — з наслідками. Якщо x_i впливає на x_k, на перетині i-го стовпця і k-го рядка ставиться одиниця. Нуль символізує відсутність впливу. Макет матриці суміжності порядку $m = 5$ наведено в табл. 8.1.

Таблиця 8.1

x_i	x_1	x_2	x_3	x_4	x_5
f_1	0	0	0	0	0
f_2	1	0	0	0	0
f_3	0	1	0	0	0
f_4	0	1	1	0	0
f_5	1	0	0	1	0

Як бачимо, x_1 впливає на x_2 та x_5; у свою чергу x_2 впливає на x_3 та x_4, а x_3 — на x_4 і т. д. Контури зворотного зв'язку відсутні. Зберігаючи лише ті дуги, які не можна продублювати іншими шляхами, отримаємо *мінімальний граф* зв'язку — гамільтонів шлях: $x_1 \to x_2 \to x_3 \to x_4 \to x_5$. Він найкомпактніше описує структуру взаємозв'язків і містить усю необхідну інформацію для побудови моделі у вигляді системи рівнянь.

Важливим етапом модельної специфікації є поділ змінних, що формують структуру моделі, на ендогенні та екзогенні. *Ендогенні*, тобто взаємозалежні, змінні зумовлені внутрішньою структурою процесу і є предметом аналізу. Кількість їх дорівнює кількості рівнянь і тотожностей моделі. Включені в модель незалежні змінні називаються *екзогенними.* Саме вони спричиняють зміни в системі взаємозв'язків, не зазнаючи на собі зворотного впливу. Класифікація змінних на ендогенні та екзогенні досить умовна і залежить від природи й суті явища, яке вивчається, та від мети дослідження. В динамічних моделях з'являються *лагові* змінні. Взаємозв'язок усіх типів змінних представимо такою системою рівнянь:

$$Y_{1,t} = f_1(y_{3,t}, y_{2,t-1}, x_{1,t})$$
$$Y_{2,t} = f_2(y_{1,t}, y_{3,t-1}, x_{2,t})$$
$$Y_{3,t} = f_3(y_{2,t}, y_{1,t-1}, x_{3,t}).$$

У системі стільки рівнянь, скільки ендогенних змінних ($y_{1,t}$, $y_{2,t}$, $y_{3,t}$). Усі вони взаємозалежні, і кожна з них окремо зазнає впливу як незалежних (екзогенних) змінних ($x_{1,t}$, $x_{2,t}$, $x_{3,t}$), так і ендогенних із запізненням ($y_{2,t-1}$, $y_{3,t-1}$, $y_{1,t-1}$). Лаговим змінним властиві такі ж риси, як і екзогенним, тому вони об'єднуються в один клас *визначених наперед* змінних z_j.

Окреме i-те рівняння системи можна записати так:

$$y_i = Y_i a_i + Z_j b_j + e_i,$$

де Y_i — вектор ендогенних змінних i-го рівняння ($i = 1, 2, \ldots, k_i$);

a_i — коефіцієнти при ендогенних змінних, що входять в i-те рівняння;

Z_j — вектор екзогенних і лагових змінних i-го рівняння ($j = 1, 2, \ldots, m_i$);

b_j — коефіцієнти при змінних z_j в i-му рівнянні.

Модель відображує структуру взаємозв'язків між змінними і тому називається *структурною*. Оскільки ті ж самі ендогенні змінні входять до різних рівнянь моделі, то це призводить до залежності залишків від ендогенних змінних, що ускладнює оцінювання параметрів моделі класичним МНК. Щоб виключити кореляцію залишків, структурну модель трансформують, приводять до скороченої, *приведеної* форми. У приведеній формі всі ендогенні змінні виражені виключно через визначені наперед (екзогенні та лагові) змінні:

$$y_i = Z_j r_j + v_i,$$

де r_j — коефіцієнти приведеної форми при змінних z_j, оцінюються класичним МНК;

v_i — залишок.

Проблема оцінювання параметрів структурної моделі і можливості її перетворення пов'язані з поняттям *ідентифікації моделі*. Модель називають ідентифікованою, якщо рівняння структурної форми однозначно описують зв'язок. Умова ідентифікації перевіряється для кожного i-рівняння за критерієм:

$$(k_i - 1) \le (m - m_i) .$$

В *ідентифікованому* рівнянні різниця між загальною кількістю екзогенних і лагових змінних усієї системи m і кількістю таких змінних в i-му рівнянні m_i на одиницю більша за кількість ендогенних змінних цього рівняння k_i. Кожне рівняння ідентифікованої системи відображує певну систему взаємозв'язків, не дублює і не може бути замінено ніякою комбінацією інших рівнянь.

Коли $(m - m_i) > (k_i - 1)$, оцінки параметрів моделі не можуть бути визначені однозначно, система вважається *надідентифікованою*. Якщо для i-го рівняння $(m - m_i) \le (k_i - 1)$, то така система є *неідентифікованою*, і визначити параметри статистичними методами неможливо. Перевіримо ідентифікованість наведеної вище системи рівнянь, у якій загальна кількість екзогенних і лагових змінних $m = 6$. Перше рівняння містить дві ендогенні і дві визначені наперед змінні z_j, тобто, $k_i - 1 = 2 - 1 = 1$, а $m - m_i = 6 - 2 = 4$, що свідчить про надідентифікованість рівняння. Аналогічно можна довести надідентифікованість другого і третього рівнянь.

Оцінювання параметрів надідентифікованих рівнянь здійснюється *двокроковим* МНК. На першому кроці (1) визначаються параметри приведеної системи рівнянь і теоретичні значення ендогенних змінних; на другому (2) — в системі структурних рівнянь значення ендогенних змінних замінюються значеннями, розрахованими в рамках приведеної системи. Для лінійної моделі:

1. $\hat{y}_i = \sum_{1}^{m_i} r_j z_j$;

2. $Y_i = \sum_{1}^{k_i} a_i \hat{y}_i + \sum_{1}^{m_i} b_j x_j$.

Побудовані у такий спосіб рівняння розглядаються як звичайні рівняння регресії, параметри їх визначаються МНК і використовуються при прогнозуванні. Якщо специфікація моделі не відповідає вимогам математичного апарату та емпіричним даним, а система рівнянь виявляється неідентифікованою, то можливі три шляхи видозмінення моделі:

- виключити з моделі деякі ендогенні змінні;
- ввести в модель додаткові екзогенні або лагові змінні;
- замінити певну множину взаємозалежних змінних багатовимірними оцінками, скажімо, головними компонентами.

У системі *Statistica* методи структурного моделювання реалізовано в модулі *Sepath*. Досить потужний математичний апарат модуля ефективний при побудові економетричних моделей. Ми розглянемо системи рівнянь, представлених однозначними причинними ланцюгами. Такі системи називаються рекурентними (рекурсивними).

8.2. РЕКУРЕНТНА МОДЕЛЬ

Рекурентна модель являє собою систему рівнянь, у якій залежна змінна *i*-го рівня виступає як незалежна в рівнянні $i + 1$ рівня. Скажімо, ознакова множина моделі містить $m - 1$ ендогенних змінних Y_i та екзогенну змінну Z_i. Тоді система рівнянь має вигляд

$$Y_1 = b_{10} + a_1 Z_1$$
$$Y_2 = b_{20} + b_{21} Y_1 + a_2 Z_2$$
$$Y_3 = b_{30} + b_{31} Y_1 + b_{32} Y_2 + a_3 Z_3$$
$$Y_4 = b_{40} + b_{41} Y_1 + b_{42} Y_2 + b_{43} Y_3 + a_4 Z_4$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$Y_m = b_{m0} + b_{m1} Y_1 + b_{m2} Y_2 + b_{m3} Y_3 + b_{m4} Y_4 + \ldots + b_{m-1} Y_{m-1} + a_m Z_m.$$

Розрахунок коефіцієнтів регресії рекурентної моделі здійснюється послідовно для кожного рівняння класичним МНК. Матриця коефіцієнтів регресії трикутна, тобто

$$\begin{vmatrix} -1 & 0 & 0 & 0 & \ldots & 0 \\ b_{21} & -1 & 0 & 0 & \ldots & 0 \\ b_{31} & b_{32} & -1 & 0 & \ldots & 0 \\ b_{41} & b_{42} & b_{43} & -1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ b_{m1} & b_{m2} & b_{m3} & b_{m4} & \ldots & -1 \end{vmatrix}.$$

Основана на принципах послідовного розкладання процесу за причинами його формування, рекурентна модель уможливлює оцінювання для кожної з них *повного ефекту* впливу. Останній дорівнює сумі *прямого* та *опосередкованого* впливів.

Методологічні засади вимірювання повного ефекту впливу розглянемо на прикладі двофакторної моделі собівартості 1 т цементу — x_3. Фактори собівартості: x_1 — погодинна продуктивність цементних млинів, т; x_2 — питомі витрати електроенергії на помел 1 т цементу, кВт · г. Граф взаємозв'язку цих ознак наведено на рис. 8.2. Дуги графа вказують потоки зв'язків, спрямованих від x_i до x_k, коефіцієнти регресії над дугами розглядаються як сила цих потоків. Сила послідовно з'єднаних на шляху потоків визначається як добуток відповідних коефіцієнтів регресії. За наявності двох і більше паралельних шляхів сила їх потоків підсумовується. Згідно з графом зв'язку собівартості цементу підвищення погодинної продуктивності цементних млинів на 1 т веде до зниження собівартості цементу в середньому на 2,418 грн., водночас зростає енергомісткість виробництва цементу на 0,483 кВт · г, що спричиняє зростання собівартості на 0,725 грн. (0,484 · 1,499). Тоді сумарний ефект впливу погодинної продуктивності цементних млинів на собівартість цементу становить: –2,418 + 0,725 = –1,693 грн.

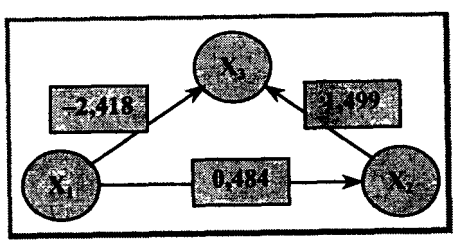


Рис 8.2 Граф зв'язку собівартості цементу

Отже, алгоритм розрахунку повного ефекту впливу фактора x_i включає дві операції:

- множення коефіцієнтів послідовно з'єднаних дуг;
- підсумовування коефіцієнтів паралельних дуг.

Такий розрахунок можна здійснити і на основі матриці коефіцієнтів регресії. Проілюструємо його на прикладі рекурентної моделі продуктивності корів (надій молока в центнерах на одну корову). Ознакову множину моделі представляють: x_1 — кількість корів на фермі; x_2 — структура раціону корів (частка концентрованих і соковитих кормів); x_3 — вміст протеїну в розрахунку на одну кормову одиницю; x_4 — витрати кормів на одну корову за рік, ц корм. од.; x_5 — яловість корів (телят на 100 корів і нетелей). У табл. 8.2 наведено коефіцієнти регресії: вище головної діагоналі — частинні, нижче — повні.

Таблиця 8.2

x_i	x_1	x_2	x_3	x_4	x_5	y
x_1	1	—	—	-0,136	—	—
x_2	—	1	0,255	—	0,450	0,014
x_3	—	0,255	1	0,110	0,052	0,105
x_4	–0,136	0,028	0,110	1	0,068	0,782
x_5	–0,009	0,465	0,059	0,068	1	0,280
y	–0,109	0,193	0,207	0,801	0,280	1

Повні ефекти впливу включених у модель факторів x_i на y (останній рядок таблиці) розраховані послідовним множенням справа наліво повних коефіцієнтів регресії на частинні за принципом рядок на рядок. Наприклад, повний ефект впливу x_4 стано-

вить: $1 \cdot 0{,}782 + 0{,}280 \cdot 0{,}068 = 0{,}801$. Аналогічний розрахунок для x_3 дає $1 \cdot 0{,}105 + 0{,}280 \cdot 0{,}052 + 0{,}801 \cdot 0{,}110 = 0{,}207$ і т. д. Як бачимо, повні коефіцієнти регресії для всіх факторів, окрім x_5, який не має опосередкованого впливу, відрізняються від частинних коефіцієнтів регресії (останній стовпець). Для факторів x_2 — x_4, які характеризують кормовий раціон, повні коефіцієнти за рахунок опосередкованих впливів перевищують частинні; щодо фактора x_1, то прямий його вплив виявився неістотним, а опосередкований — істотним.

За таким самим принципом можна розрахувати бета-коефіцієнти та коефіцієнти еластичності. Процедура перевірки істотності повних коефіцієнтів регресії і визначення довірчих меж здійснюється так само, як і в однорівневій моделі. Похибка вибірки визначається за формулою

$$\mu_b = \sqrt{\frac{S_k^2}{S_i^2(n-m_k)}},$$

де S_k^2, S_i^2 — залишкові суми квадратів відхилень по факторах, які в мінімальному графі зв'язків передують факторам x_i та x_k;

n — обсяг сукупності;

m — кількість змінних у рівнянні x_k.

Враховуючи прямі та опосередковані зв'язки, рекурентна модель значно розширює аналітичні можливості регресійного аналізу, розкриває механізм формування варіації моделюваного показника. Спроможність рекурентної моделі комплексно реагувати на будь-які зміни в системі взаємозв'язків робить її ефективним засобом при імітації варіантів економічних рішень.

Завдання для самоконтролю

1. На підсумкову успішність студентів впливає три групи факторів:

а) соціально-демографічні характеристики: вік, стать, сімейний стан, соціальний статус та освіта батьків;

б) стартові умови: тип і місце закінчення середнього навчального закладу, рівень довузівської підготовки, форма підготовки до вступних іспитів (самостійно, з репетитором, підготовчі курси), мотиви вибору професії;

в) трудова активність студента: поточна успішність, пропуски, заняття спортом, підробітки.

За допомогою графа зв'язку опишіть структуру взаємовпливу факторів на підсумкову успішність студентів.

2. Взаємозв'язок факторів, що впливають на результати роботи риболовецької флотилії за рік, подано матрицею суміжності: x_1 — вилов риби на одне судно; x_2 — кількість рейсів одного судна; x_3 — кількість підйомів трала на одне судно; x_4 — час чистого тралення; x_5 — тривалість знаходження судна на лові.

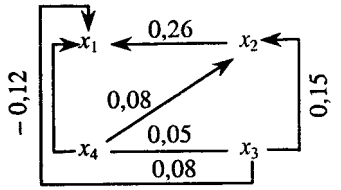
x_i	x_1	x_2	x_3	x_4	x_5
x_1	0	0	1	1	0
x_2	0	0	0	0	0
x_3	0	1	0	0	1
x_4	0	0	1	0	1
x_5	0	1	0	0	0

Визначте прямі та опосередковані зв'язки, побудуйте мінімальний граф зв'язку та на його основі — систему рівнянь.

3. На основі системи одночасних рівнянь класифікуйте показники на ендогенні та екзогенні, перевірте ідентифікованість рівнянь. У модель включено показники: x_1 — валовий внутрішній продукт; x_2 — основний капітал; x_3 — кількість зайнятих; x_4 — споживання населення; x_5 — індекс цін; x_6 — інвестиції в економіку; x_7 — грошова маса; x_8 — кредитна ставка; x_9 — курс національної валюти.

$$x_1 = f(x_2, x_3);$$
$$x_4 = f(x_1, x_5);$$
$$x_6 = f(x_1, x_{2-1}, x_8);$$
$$x_8 = f(x_1, x_7);$$
$$x_9 = f(x_1, x_8).$$

4. На рисунку наведено фрагмент графа зв'язку рекурентної моделі рентабельності тваринництва: x_1 — % рентабельності; x_2 — вихід валової продукції на одну умовну голову худоби, грн.; x_3 — витрати кормів на умовну голову худоби, корм. од.; x_4 — затрати праці на умовну голову худоби, людино-днів. Над дугами наведено частинні коефіцієнти регресії.



Визначте повні ефекти впливу на рентабельність тваринництва витрат кормів і затрат праці.

5. У таблиці наведено бета-коефіцієнти рекурентної моделі, яка описує залежність питомих витрат нафтопродуктів на транспорті (x_5). У модель включені фактори: x_1 — частка автомобільного транспорту в сумарному вантажообороті; x_2 — співвідношення вантажообороту автомобільного транспорту та ВВП; x_3 — споживання нафтопродуктів на одиницю ВВП; x_4 — частка світлих нафтопродуктів у загальному обсязі продукції нафтопереробки.

x_i	x_1	x_2	x_3	x_4	x_5
x_1	1	0,56	—	0,43	0,16
x_2		1	0,27	—	0,24
x_3			1	0,18	–0,35
x_4				1	0,62
x_5					1

На основі частинних бета-коефіцієнтів рекурентної моделі обчисліть повні, поясніть їх зміст.

Розділ 9

МОДЕЛЬ ГОЛОВНИХ КОМПОНЕНТ

9.1. КОНЦЕПЦІЯ МЕТОДУ ГОЛОВНИХ КОМПОНЕНТ

При моделюванні складних причинних комплексів часом стикаються з проблемою надлишковості інформації, коли екзогенні змінні x_i, включені в ознаковий простір моделі, висококорельовані (мультиколінеарні). Щоб забезпечити адекватність моделі реальному процесу, вдаються до заміни такого типу ознакової множини меншою кількістю некорельованих величин, які б зберігали всю інформацію щодо причинно-наслідкового механізму формування явища (процесу) і не впливали на точність результатів аналізу. Інструментом такої заміни є *метод головних компонент*.

Основне призначення методу головних компонент — виявити приховані (латентні) першопричини, які пояснюють кореляції між ознаками і змістовно інтерпретуються. Використання методу ґрунтується на припущенні, що ознаки x_i є лише індикаторами певних існуючих властивостей явища, які безпосередньо не вимірюються. Так, хвороба людини виявляється певними симптомами, рівень життя населення — умовами праці, побуту та дозвілля. Якщо таких першопричин декілька, в ознаковому просторі X виокремлюються групи висококорельованих ознак. Скажімо, сім ознак x_i поділяються на дві групи:

Група	Ознаки	Компонента
1	x_1 x_2 x_3 x_4	G_1
2	x_5 x_6 x_7	G_2

Першопричина кореляції ознак j-ї групи називається *компонентою* G_j. Ознаки, що належать до різних груп, некорельовані, а отже, і компоненти G_j незалежні (ортогональні). Суть методу головних компонент полягає у переході від численної множини x_i до мінімальної кількості максимально інформативних компонент G_j.

$$\underset{i=1,2,\ ,m}{x_i} \Rightarrow \underset{j=1,2,\ ,p}{G_j}\ .$$

Основні задачі методу головних компонент:

- виокремити та ідентифікувати компоненти G_j;
- визначити рівні G_j для окремих одиниць статистичної сукупності.

Ідентифікація компонент, тобто надання їм певного змісту, залежить від ознакової множини X. Як правило, її формують на основі теоретично обґрунтованої гіпотези щодо природи латентних властивостей явища. Якщо така гіпотеза відсутня, то використовують максимальну кількість ознак, покладаючись на можливості методу виявити такі властивості. Але в такому разі інтерпретація компонент ускладнюється.

Оскільки компоненти є гіпотетичними величинами, то виміряти їх можна лише опосередковано за допомогою спеціально сконструйованих моделей. У моделі головних компонент зв'язок між первинними ознаками і компонентами описується як лінійна комбінація

$$z_i = \sum_1^m a_{ij} G_j ,$$

де z_i — стандартизовані значення i-ї ознаки з одиничними дисперсіями; *сумарна дисперсія* дорівнює кількості ознак m;

a_{ij} — *факторне навантаження* j-ї компоненти на i-у ознаку.

Навантаження a_{ij} характеризує щільність зв'язку між i-ю ознакою та j-ю компонентою і як будь-яка міра щільності зв'язку змінюється в межах від 0 до ± 1.

У моделі головних компонент відсутні залишки, тобто апріорі передбачається, що всі m компонент повністю пояснюють сумарну дисперсію ознакової множини. За умови ортогональності компонент квадрат факторного навантаження a_{ij}^2 характеризує внесок j-ї компоненти у варіацію i-ї ознаки. Повний внесок j-ї компоненти у сумарну дисперсію m ознак становить $\lambda_j = \sum_1^m a_{ij}^2$.

У процесі компонентного аналізу сумарна варіація m первинних ознак x_i перерозподіляється між компонентами G_j з дисперсіями λ_j. Тобто сумарну дисперсію ознакової множини X можна представити як суму дисперсій компонент $\sum_1^m \lambda_j$ або через факторні навантаження.

$$m = \sum_1^m \lambda_j = \sum_1^m \sum_1^m a_{ij}^2 .$$

Схема декомпозиції сумарної дисперсії ознакової множини X наведено у вигляді матриці (табл. 9.1).

Таблиця 9.1

Z_i \ G_j	G_1	G_2	...	G_m	Дисперсія z_i
z_1	a_{11}^2	a_{12}^2	...	a_{1m}^2	1
z_2	a_{21}^2	a_{22}^2	...	a_{2m}^2	1
z_3	a_{31}^2	a_{32}^2	...	a_{3m}^2	1
...	...	...	...	...	...
z_m	a_{m1}^2	a_{m2}^2	...	a_{mm}^2	1
Дисперсія G_j	λ_1	λ_2	...	λ_m	m

Аналіз матриці по рядках показує, які компоненти і з якою вагою формують варіацію i-ї ознаки. Кожній ознаці властива своя *факторна структура*. Чим менше компонент навантажує ознаку, тим простішою вважається її факторна структура.

Аналіз матриці по стовпцях показує, які ознаки є індикаторами j-ї компоненти. Компоненти упорядковуються за значеннями дисперсій:

$$\lambda_1 > \lambda_2 > \lambda_3 > ... > \lambda_m .$$

Незважаючи на те, що замість m ознак визначається така ж кількість компонент, внесок більшості з них у сумарну варіацію виявляється незначним. Левова частка сумарної варіації припадає на декілька перших компонент. Як показує досвід, кількість таких вагомих компонент становить 10—15% від кількості первинних ознак. Саме вони називаються *головними компонентами* і підлягають змістовній інтерпретації.

Таким чином, модель головних компонент трансформує m-вимірний ознаковий простір у p-вимірний простір компонент ($p < m$). Сумарна дисперсія головних компонент менша за сумарну дисперсію ознакового простору. Відношення $\frac{\sum_1^p \lambda_j}{m}$ характеризує *повноту факторизації*.

Математичною основою методу головних компонент слугує кореляційна матриця R з одиницями на головній діагоналі. Недіа-

гональні елементи матриці представлені коефіцієнтами кореляції r_{ik}, які оцінюють не причинно-наслідкові, а супутні зв'язки між ознаками x_i та x_k, зумовлені наявністю спільної першопричини їх варіації.

У термінах матричної алгебри дисперсії компонент λ_j — це властиві числа кореляційної матриці R. Кожному з них відповідає властивий вектор V, який задовольняє рівняння $(R - \lambda E)V = 0$, де E — одинична матриця. Тобто виокремлення головних компонент є класичною задачею визначення властивих чисел λ та властивих векторів V кореляційної матриці R. Головними вважаються компоненти, для яких:

- за критерієм Кайзера $\lambda_j > 1$;
- повнота факторизації не менша, скажімо, 70%.

Наприклад, для кореляційної матриці

$$R = \begin{vmatrix} 1 & 0{,}8 & 0{,}2 \\ 0{,}8 & 1 & 0{,}6 \\ 0{,}2 & 0{,}6 & 1 \end{vmatrix}$$

властиві значення дорівнюють: $\lambda_1 = 2{,}1$; $\lambda_2 = 0{,}81$; $\lambda_3 = 0{,}09$. За критерієм Кайзера головною слід вважати першу компоненту ($\lambda_1 > 1$). Внесок цієї компоненти в сумарну варіацію трьох ознак становить 2,1 : 3 = 0,7, або 70%.

Розв'язування системи рівнянь

$$\begin{aligned} (1 - 2{,}1)V_1 + 0{,}8V_2 + 0{,}2V_3 &= 0 \\ 0{,}8V_1 + (1 - 2{,}1)V_2 + 0{,}6V_3 &= 0 \\ 0{,}2V_1 + 0{,}6V_2 + (1 - 2{,}1)V_3 &= 0 \end{aligned}$$

дає властивий вектор $V = (1{,}213;\ 1{,}428;\ 1{,}0)$.

Щоб задовольнити умову $\lambda_j = \sum_1^m a_{ij}^2$, властивий вектор нормується

$$a_{ij} = V_{ij} \sqrt{\frac{\lambda_j}{\sum_1^m V_{ij}^2}}.$$

Отже, факторні навантаження j-ї компоненти є не що інше, як нормований властивий вектор матриці R. У розглянутому прикладі: $\sum_1^3 V_{i1}^2 = 4{,}81064$, множник $\sqrt{\frac{2{,}1}{4{,}81064}} = 0{,}668267$. Звідси факторні навантаження:

$$a_{11} = 0{,}878;\ a_{21} = 0{,}975;\ a_{31} = 0{,}683.$$

Сума квадратів факторних навантажень дорівнює значенню $\lambda_1 = 2{,}1$.

Процедури методу головних компонент — *Principal components* — представлено в модулі *Factor Analysis* — Факторний аналіз. Інформаційною базою компонентного аналізу можуть бути як первинні ряди (*Raw data*), так і кореляційна матриця (*Correlation matrix*). Тип інформаційної бази вказується на стартовій панелі модуля (*Input file*).

Визначимо факторні навантаження за даними кореляційної матриці (табл. 9.2). Цей файл створено в модулі *Data Management* — Управління даними. Окрім коефіцієнтів кореляції він містить значення середніх і стандартних відхилень кожної ознаки, обсяг сукупності, за даними якої обчислено кореляційну матрицю, та кількість матриць (у нашому прикладі — одна).

Таблиця 9.2

DATA. FMOD1 STA 5v * 9c					
Correlation matrix					
Variable	VAR1	VAR2	VAR3	VAR4	VAR5
VAR1	1	0,839	0,927	0,871	0,753
VAR2	0,839	1	0,967	0,778	0,828
VAR3	0,927	0,967	1	0,845	0,852
VAR4	0,871	0,778	0,845	1	0,837
VAR5	0,753	0,828	0,852	0,837	1
means	112,2	59,4	76,8	107	66,8
st.dev	28,6	17,1	21	28,9	19,3
N.	12				
matrix	1				

Щільність взаємозв'язків між ознаками дає підстави зробити висновок про наявність однієї першопричини формування їх варіації. Цей висновок підтверджують розраховані факторні навантаження, наведені в табл. 9.3. Мінімальне значення — $a_{15} = 0{,}909$.

Таблиця 9.3

Factor Loadings (Unrotated) (fmod1.sta)	
Extraction: Principal components (Marked loadings are > ,700000)	
Variable	Factor 1
VAR1	0,9367
VAR2	0,9418
VAR3	0,9798
VAR4	0,9223
VAR5	0,9090
Expl.Var	4,4016
Prp.Totl	0,8803

Властиве значення кореляційної матриці *Expl.Var* становить 4,4, а ступінь факторизації *Prp.Totl* = 4.4 : 5 = 0,88.

9.2. ІДЕНТИФІКАЦІЯ ТА ВИМІРЮВАННЯ ГОЛОВНИХ КОМПОНЕНТ

У реальних багатовимірних сукупностях часто виокремлюється не одна, а декілька головних компонент, навантаження яких на окремі ознаки перетинаються. Складна факторна структура значно ускладнює ідентифікацію компонент. Пошук *простої факторної структури*, коли a_{ij} наближається до 1 або 0, здійснюється за допомогою різних процедур ортогонального чи косокутного *обертання*, в процесі якого значення одних факторних навантажень зростають, інших — зменшуються. Найчастіше використовують процедуру варімакс (*Varimax*), яка максимізує варіацію квадратів факторних навантажень для кожної компоненти, збільшуючи великі і зменшуючи малі значення a_{ij}.

В алгебраїчних термінах обертання означає перетворення матриці факторних навантажень A в матрицю простої факторної структури B. Необхідно знайти таку матрицю трансформації T, яка б забезпечила рівність $B = AT$. Матриця трансформації T залежить від кількості головних компонент і кута обертання Θ,

який не повинен перевищувати 45°. Для двох компонент при обертанні за годинниковою стрілкою

$$T = \begin{vmatrix} \cos\Theta & -\sin\Theta \\ \sin\Theta & \cos\Theta \end{vmatrix}.$$

Очевидно, що проста факторна структура недосяжна, але наближення до неї все ж спрощує ідентифікацію компонент. Наприклад, трансформуємо матрицю A з кутом обертання $\Theta = 30°$ ($\sin 30° = 0{,}500$; $\cos 30° = 0{,}866$):

$$B = AT = \begin{vmatrix} 0{,}60 & 0{,}40 \\ 0{,}40 & 0{,}50 \\ -0{,}30 & 0{,}60 \\ -0{,}20 & 0{,}80 \\ -0{,}10 & 0{,}70 \end{vmatrix} \cdot \begin{vmatrix} 0{,}866 & -0{,}500 \\ 0{,}500 & 0{,}866 \end{vmatrix} = \begin{vmatrix} 0{,}72 & 0{,}05 \\ 0{,}60 & 0{,}23 \\ 0{,}04 & 0{,}67 \\ 0{,}24 & 0{,}79 \\ 0{,}26 & 0{,}91 \end{vmatrix}.$$

На основі факторних навантажень матриці B можна зробити висновок, що перша компонента навантажує ознаки x_1 та x_2, друга — решту ознак. Зміст кожної компоненти визначається змістом ознак, які її представляють.

Отже, побудова моделі головних компонент здійснюється в три етапи:

- розрахунок кореляційної матриці R;
- виокремлення головних компонент і розрахунок факторних навантажень;
- ідентифікація головних компонент.

Розглянемо аналітичні можливості моделі за даними файла *faktor sta.* (піддиректорія *Examples*), в якому наведено результати соціологічного опитування 100 респондентів щодо ступеня задоволеності їх життям. Ініціювавши кнопку *Variables*, сформуємо ознакову множину моделі, надавши кожній ознаці соціально-економічний зміст:

1 — самооцінка професійного статусу респондента;
2 — оцінка умов праці;
3 — оцінка рейтингу компанії;
4 — оцінка можливостей самореалізації поза роботою;
5 — ефективність відпочинку;
6 — оцінка матеріального добробуту сім'ї;
7 — задоволеність соціальним статусом сім'ї;
8 — оцінка навколишнього середовища.

Після команди *OK* з'являється вікно *Define Method of Factor Extraction* — Визначити метод виокремлення факторів. У функ-

ціональній його частині з-поміж запропонованих методів вибираємо *Principal components* — Головні компоненти. Праворуч розміщено поля для установки параметрів моделі: *Maximum no. of factors* — максимальне число факторів і *Minimum eigenvalue* — мінімальне властиве число. За умовчування ці параметри становлять відповідно 2 і 1.

За командою на виконання програми з'являється вікно *Factor Analysis Results* — Результати факторного аналізу, в інформаційній частині якого вказується кількість ознак, метод аналізу, десятковий логарифм детермінанта кореляційної матриці, число виокремлених факторів і властиві значення матриці λ_j. Для детальнішого аналізу результатів скористаємося опціями функціональної частини вікна. Скажімо, для візуальної оцінки виокремлення головних компонент можна скористатися графічним критерієм «кам'янистий обвал» — *Scree plot* (рис. 9.1). Значення властивих чисел кореляційної матриці представлено на осі ординат. Як бачимо, ці значення стрімко зменшуються і лише два перших більші за одиницю.

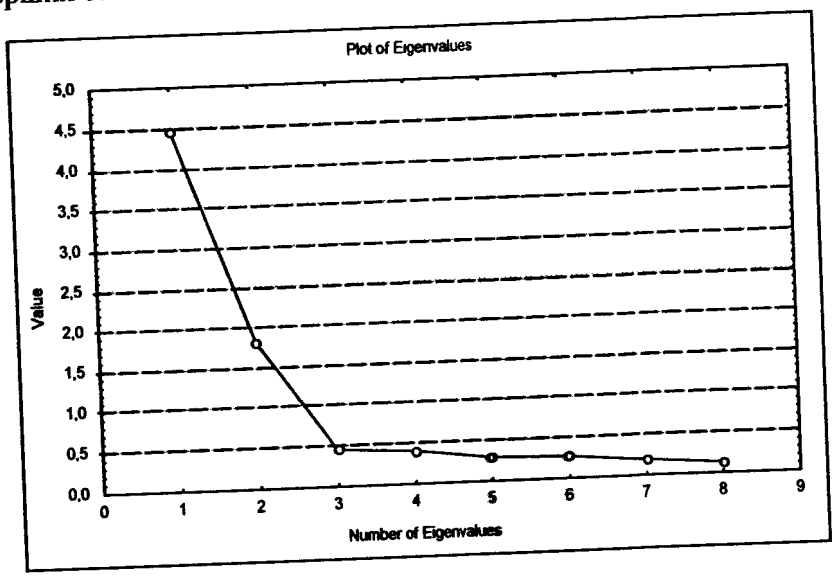


Рис. 9.1. Властиві числа кореляційної матриці

За установкою *Eigenvalues* система видає таблицю значень властивих чисел, які є дисперсіями головних компонент, а також внесок кожної з них у сумарну варіацію ознакової множини —

% *total Variance* (табл. 9.4). Внесок першої компоненти в сумарну дисперсію ознакової множини становить 55,6%, другої — 22,5%. Разом (*Cumul.%*) дві компоненти пояснюють 78,1% сумарної варіації, що свідчить про високий ступінь факторизації.

Таблиця 9.4

Eigenvalues (factor.sta)				
Continue...	Extraction: Principal components			
Value	Eigenval	% total Variance	Cumul Eigenval	Cumul. %
1	4,56	55,6	4,56	55,6
2	1,80	22,5	6,36	78,1

З-поміж процедур обертання факторів — *Factor rotation* вибираємо *Varimax normalized* — Варімакс нормалізований. За опцією *Factor loadings* маємо таблицю факторних навантажень, значення яких наближаються до 1 або до 0 (табл. 9.5). Ознаки, які навантажує кожна компонента, виділено.

Перша компонента зв'язана з ознаками 1—5, її можна ідентифікувати як ступінь задоволеності роботою і дозвіллям; друга компонента навантажує ознаки 6—8, які характеризують матеріальний добробут і соціальний статус сімей респондентів. Наведені в останніх рядках таблиці властиві значення і внесок окремих компонент у сумарну дисперсію визначені за трансформованими факторними навантаженнями, а тому відрізняються від первинних, проте сумарний їх внесок процедура обертання не змінює: *Prp.Totl.* = 0,418 + 0,363 = 0,781.

Таблиця 9.5

Factor Loadings (Varimax normalized) (factor sta)		
Continue ..	Extraction: Principal components (Marked loadings are > ,700000)	
Variable	Factor 1	Factor 2
WORK_1	**0,8425**	0,0196
WORK_2	**0,9023**	0,0958
WORK_3	**0,8700**	0,1185
HOBBY_1	**0,7109**	0,6075
HOBBY_2	**0,7182**	0,5165
HOME_1	0,0834	**0,8438**
HOME_2	0,1213	**0,8971**
HOME_3	0,1415	**0,8538**
Expl.Var	3,3438	2,9056
Prp.Totl	0,4180	0,3632

Поглиблений факторний аналіз складних соціально-економічних явищ передбачає вимірювання головних компонент для окремих одиниць сукупності. Процедура, за якою *h*-й одиниці сукупності надається певна оцінка латентної вели-

чини G, називають *факторним шкалюванням*. Значення компонент можна визначити, спираючись на зв'язок їх з первинними ознаками $Z = AG$, звідки

$$G = A^{-1}Z,$$

де A^{-1} — обернена матриця факторних навантажень m компонент.

Враховуючи, що в процесі факторного аналізу виокремлюється p головних компонент ($p < m$), вимірюванню підлягають саме ці компоненти:

$$G = \lambda^{-1}A'Z,$$

де λ^{-1} — дисперсії головних компонент.

Алгебраїчно ця процедура зводиться до підсумовування значень ознак x_i (у стандартизованому масштабі) з вагами, пропорційними факторним навантаженням (до обертання):

$$G_h = \sum_{i=1}^{m}\left(\frac{a_{ij}}{\lambda_j}z_{hi}\right).$$

Ділення факторних навантажень на λ_j забезпечує нульове математичне сподівання та одиничну дисперсію оцінок G. Знаки (+, –) свідчать про те, що рівень компоненти у h-ї одиниці сукупності вищий або нижчий за середній. Обчислені за даними файла *faktor sta.* значення обох головних компонент для семи респондентів наведено у табл. 9.6. Згідно з даними в одних респондентів обидві компоненти додатні, у других — обидві компоненти від'ємні, у третіх — знаки оцінок компонент протилежні.

Таблиця 9.6

Factor Scores (factor.sta)		
Rotation: Unrotated Extraction: Principal components		
Variable	Factor 1	Factor 2
1	–1,442	–1,243
2	1,007	0,773
3	0,200	1,612
4	–0,717	0,063
5	–0,007	–0,162
6	0,401	1,354
7	–2,556	0,859

Оцінки головних компонент застосовують при ранжуванні та типології одиниць сукупності, при вивченні закономірностей динаміки, при вимірюванні взаємозв'язків. У системах одночасних рівнянь, коли коефіцієнти регресії визначаються двокроковим МНК, головні компоненти використовуються на першому кроці як визначені наперед змінні приведеної форми моделі. Такий підхід значно спрощує розрахунки, не впливаючи на точність результатів аналізу.

Завдання для самоконтролю

1. За результатами факторного аналізу властиві значення п'яти компонент становлять: $\lambda_j = 2{,}52;\ 1{,}12;\ 0{,}85;\ 0{,}42;\ 0{,}09$.

Виокремте головні компоненти. Яку частку сумарної дисперсії вони пояснюють?

2. Для шести ознак маємо факторні навантаження двох компонент:

	x_1	x_2	x_3	x_4	x_5	x_6
a_{i1}	0,62	0,78	0,85	0,90	0,66	0,93
a_{i2}	0,55	0,48	0,52	0,43	0,68	0,45

Визначте внесок кожної компоненти в сумарну дисперсію ознак.

3. За результатами компонентного аналізу на семи показниках технічного стану підприємств виділено одну компоненту з дисперсією $\lambda = 5{,}2$ і властивим вектором $V = (0{,}90;\ 0{,}80;\ 0{,}74;\ 1{,}00;\ 0{,}82;\ 0{,}87;\ 0{,}96)$.

Визначте факторні навантаження кожного показника та оцініть адекватність моделі головних компонент.

4. Компонентний аналіз розвитку соціальної інфраструктури міста здійснено за даними динамічних рядів показників: x_1 — забезпеченість населення міста житлом; x_2 — частка комплексно упорядженого житла; x_3 — забезпеченість телефонними апаратами на 1000 мешканців; x_4 — надання побутових послуг на 1000 мешканців; x_5 — обсяг роздрібного товарообороту на одного мешканця. Виокремлено дві головні компоненти, факторні навантаження яких становлять:

Факторне навантаження	Показник				
	x_1	x_2	x_3	x_4	x_5
a_{i1}	0,72	0,68	0,59	0,44	0,37
a_{i2}	0,51	0,56	0,43	0,76	0,80

а) Поясніть зміст факторних навантажень.

б) Визначте внесок кожної компоненти в сумарну дисперсію.

в) З метою чіткішої інтерпретації виокремлених компонент проведіть факторне обертання проти годинникової стрілки на 30°.

г) Дайте економічну інтерпретацію компонент.

д) Оцініть адекватність моделі.

5. Компонента, яку можна ідентифікувати як рівень економічного розвитку країн, навантажує п'ять різнобічних показників. За наведеними даними оцініть: а) повноту факторизації моделі; б) рівень економічного розвитку двох країн. Зробіть висновки.

Показник	Факторне навантаження	Нормовані значення показників	
		Країна А	Країна В
Тривалість життя	0,68	0,346	0,115
Рівень освіти	0,72	0,428	0,010
Зайнятість населення	0,77	0,166	0,180
Споживання електроенергії на душу населення	0,87	0,211	0,315
Вартість життя	0,92	–0,036	–0,136

6. Кореляційна матриця характеризує взаємозв'язки показників експортного потенціалу країн: x_1 — ступінь відкритості економіки країни (співвідношення зовнішньоторговельного обороту і валового національного продукту); x_2 — експорт на душу населення; x_3 — частка спеціального експорту товарів вітчизняного виробництва та підданих переробці товарів іноземного виробництва; x_4 — збалансованість експорту та імпорту; x_5 — умови торгівлі (співвідношення індексів експортних та імпортних цін).

x_i	x_1	x_2	x_3	x_4	x_5
x_1	1				
x_2	0,85	1			
x_3	0,92	0,95	1		
x_4	0,74	0,67	0,80	1	
x_5	0,83	0,78	0,76	0,93	1

Сформуйте файл даних кореляційної матриці в модулі *Data Management* і за процедурами модуля *Factor Analysis*: а) визначте факторні навантаження на першу компоненту; б) оцініть рівень експортного потенціалу двох країн, для яких стандартизовані значення показників наведено в таблиці:

Країна	x_1	x_2	x_3	x_4	x_5
А	0,66	0,35	0,60	0,29	0,42
Б	1,25	0,82	0,43	-0,15	0,37

Зробіть висновки.

ЛІТЕРАТУРА

1. *Афифи А., Эйзен С.* Статистический анализ: Подход с использованием ЭВМ / Пер. с англ. — М.: Мир, 1982. — 488 с.
2. *Бокс Дж., Дженкинс Г.* Анализ временных рядов: Прогноз и управление. Вып. 1 / Пер. с англ. — М.: Мир, 1974. — 405 с.
3. *Боровиков В. П., Боровиков И. П.* Statistica® — Статистический анализ и обработка данных в среде Windows®. — М.: Информ.-издат. дом «Филинъ», 1998. — 608 с.
4. *Головач А. В., Ерина А. М., Трофимов В. П.* Критерии математической статистики в экономических исследованиях. — М.: Статистика, 1973. — 136 с.
5. *Дрейпер Н., Смит Г.* Прикладной регрессионный анализ: В 2 кн. / Пер. с англ. — М.: Финансы и статистика, 1986. — 366 с.
6. *Елисеева И. И., Рукавишников В. О.* Логика прикладного статистического анализа. — М.: Финансы и статистика, 1982. — 192 с.
7. *Ерина А. М.* Математико-статистические методы изучения экономической эффективности производства. — М.: Финансы и статистика, 1983. — 191 с.
8. *Кильдишев Г. С., Френкель А. А.* Анализ временных рядов и прогнозирование. — М.: Статистика, 1973. — 100 с.
9. *Льюис К. В.* Методы прогнозирования экономических показателей / Пер. с англ. Е. З. Демиденко. — М.: Финансы и статистика, 1986. — 133 с.
10. *Плюта В.* Сравнительный многомерный анализ в эконометрическом моделировании / Пер. с польск. — М.: Финансы и статистика, 1989. — 175 с.
11. *Политова И. Д.* Дисперсионный и корреляционный анализ в экономике: Учеб. пособие. — М.: Экономика, 1972. — 224 с.
12. Статистическое моделирование и прогнозирование: Учеб. пособие / Под ред. А. Г. Гранберга. — М.: Финансы и статистика, 1990. —383 с.
13. *Столяров Г. С., Ємшанов Д. Г., Ковтун Н. В.* ARM статистика: Навч. посібник. — К.: КНЕУ, 1999. — 268 с.
14. Теория и практика статистического моделирования экономики / Под ред. Е. М. Четыркина и А. Класа. — М.: Финансы и статистика, 1986. — 272 с.
15. *Трофимов В. П.* Логическая структура статистических моделей. — М.: Финансы и статистика, 1985. — 191 с.
16. Факторный, дискриминантный и кластерный анализ / Пер. с англ.; Дж.-О. Ким, Ч. У. Мьюллер, У. Р. Клекка и др.; Под ред. И. С. Енюкова. — М.: Финансы и статистика, 1989. — 215 с.
17. *Френкель А. А.* Прогнозирование производительности труда: методы и модели. — М.: Экономика, 1989. — 214 с.
18. *Четыркин Е. М.* Статистические методы прогнозирования. — М.: Статистика, 1977. — 199 с.

Додаток 1

Функція нормального розподілу $F_x = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{z^2}{2}} dz$.

z	0	1	2	3	4	5	6	7	8	9
0,0	0,500	504	508	512	516	520	524	528	532	536
0,1	540	544	548	552	556	560	564	567	571	575
0,2	580	583	587	591	595	599	603	606	610	614
0,3	618	622	626	629	633	637	641	644	648	652
0,4	655	659	663	666	670	674	677	681	684	688
0,5	691	695	698	702	705	709	712	716	719	722
0,6	726	729	732	736	739	742	745	749	752	755
0,7	758	761	764	767	770	773	776	779	782	785
0,8	788	791	794	797	800	802	805	808	811	813
0,9	816	819	821	824	826	829	831	834	836	839
1,0	841	844	846	849	851	853	855	858	860	862
1,1	864	867	869	871	873	875	877	879	881	883
1,2	885	887	889	891	893	894	896	898	900	901
1,3	903	905	907	908	910	911	913	915	916	918
1,4	919	921	922	924	925	926	928	929	931	932
1,5	933	934	936	937	938	939	941	942	943	944
1,6	945	946	947	948	950	951	952	953	954	954
1,7	955	956	957	958	959	960	961	962	962	963
1,8	964	965	966	966	967	968	969	969	970	971
1,9	971	972	973	973	974	974	975	976	976	977
2,0	977	978	978	979	979	980	980	981	981	982
2,1	982	983	983	983	984	984	985	985	985	986
2,2	986	986	987	987	987	988	988	988	989	989
2,3	989	990	990	990	990	991	991	991	991	992
2,4	992	992	992	992	993	993	993	993	993	994
2,5	994	994	994	994	994	995	995	995	995	995
2,6	995	995	996	996	996	996	996	996	996	996
2,8	997	998	998	998	998	998	998	998	998	998
2,9	998	998	998	998	998	998	998	999	999	999

Додаток 2

Критичні значення $\chi_{1-\alpha}(k)$

k	1	2	3	4	5	6	7	8	9	10	11
α = 0,10	2,71	4,61	6,25	7,78	9,24	10,64	12,02	13,36	14,68	15,99	17,28
α = 0,05	3,84	5,99	7,81	9,49	11,07	12,59	14,07	15,51	16,92	18,31	19,67

Додаток 3

Критичні значення критерію Колмогорова-Смирнова

Обсяг вибірки n	α = 0,10	α = 0,05	Обсяг вибірки n	α = 0,10	α = 0,05
5	0,510	0,563	15	0,304	0,338
6	470	521	16	295	328
7	438	486	17	286	318
8	411	457	18	278	309
9	388	432	19	272	301
10	368	409	20	264	294
11	352	391	25	240	264
12	338	375	30	220	242
13	325	361	$n > 30$	$\frac{1,22}{\sqrt{n}}$	$\frac{1,36}{\sqrt{n}}$
14	314	349			

Додаток 4

Критичні значення критерію Граббса T_n

№ спостереження	α = 0,05	α = 0,10	№ спостереження	α = 0,05	α = 0,10
3	1,41	1,41	15	2,49	2,33
4	1,69	1,65	16	2,52	2,35
5	1,87	1,79	17	2,55	2,38
6	2,00	1,89	18	2,58	2,40
7	2,09	1,97	19	2,60	2,43
8	2,17	2,04	20	2,62	2,45
9	2,24	2,10	21	2,64	2,47
10	2,29	2,15	22	2,66	2,49
11	2,34	2,19	23	2,68	2,50
12	2,39	2,23	24	2,70	2,52
13	2,43	2,26	25	2,72	2,54
14	2,46	2,30	26	2,73	2,55

Додаток 5

Квантилі t-розподілу Стьюдента $t_{1-0,05}(k)$: $|t|$ — двосторонній критерій; t — односторонній критерій

k	$\|t\|$	t	k	$\|t\|$	t
5	2,57	3,04	18	2,10	2,17
6	2,45	2,78	20	2,09	2,15
7	2,37	2,62	25	2,06	2,11
8	2,31	2,51	30	2,05	2,08
9	2,26	2,43	40	2,02	2,05
10	2,23	2,37	50	2,01	2,03
11	2,20	2,33	60	2,00	2,02
12	2,18	2,29	100	1,98	1,99
14	2,15	2,24	∞	1,96	1,96
16	2,12	2,20			

Додаток 6

Значення Z^* для оцінювання довірчих меж прогнозу (лінійний тренд)

n	v			n	v		
	1	2	3		1	2	3
5	1,366	1,524	1,702	10	1,211	1,270	1,335
7	1,309	1,427	1,558	11	1,191	1,239	1,293
8	1,267	1,358	1,459	12	1,174	1,215	1,260
9	1,236	1,308	1,389				

Додаток 7

Критичні значення циклічного коефіцієнта автокореляції ($\alpha = 0,05$)

n	Додатні значення	Від'ємні значення	n	Додатні значення	Від'ємні значення
5	0,253	–0,753	20	0,299	–0,399
6	0,345	–0,708	25	0,276	–0,356
7	0,370	–0,674	30	0,257	–0,356
8	0,371	–0,625	35	0,242	–0,300
9	0,366	–0,593	40	0,229	–0,279
10	0,360	–0,564	50	0,208	–0,248
11	0,353	–0,539	60	0,191	–0,225
12	0,348	–0,516	70	0,178	–0,207
13	0,341	–0,497	80	0,170	–0,195
14	0,335	–0,479	90	0,161	–0,184
15	0,328	–0,462	100	0,154	–0,174

Додаток 8

Критичні значення критерію Дарбіна-Ватсона

n	Кількість факторів											
	$m = 1$		$m = 2$		$m = 3$		$m = 4$		$m = 5$		$m = 6$	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
6	0,61	1,40										
7	0,70	1,36	0,47	1,90								
8	0,76	1,33	0,56	1,78	0,37	2,29						
9	0,82	1,32	0,63	1,70	0,46	2,13	0,30	2,59				
10	0,88	1,32	0,70	1,64	0,53	2,02	0,38	2,41	0,24	2,82		
11	0,93	1,32	0,76	1,60	0,60	1,93	0,44	2,28	0,32	2,65	0,12	2,89
12	0,97	1,33	0,81	1,58	0,66	1,86	0,51	2,18	0,38	2,51	0,16	2,67
13	1,01	1,34	0,86	1,56	0,72	1,82	0,57	2,09	0,45	2,39	0,21	2,49
14	1,05	1,35	0,91	1,55	0,77	1,78	0,63	2,03	0,51	2,30	0,26	2,35
15	1,08	1,36	0,95	1,54	0,81	1,75	0,69	1,98	0,56	2,22	0,30	2,24
16	1,11	1,37	0,98	1,54	0,86	1,73	0,73	1,94	0,62	2,16	0,35	2,15
18	1,16	1,39	1,05	1,54	0,93	1,70	0,82	1,87	0,71	2,06	0,44	2,02
20	1,20	1,41	1,10	1,54	1,00	1,68	0,89	1,83	0,79	1,99	0,52	1,92
22	1,24	1,43	1,15	1,54	1,05	1,66	0,96	1,80	0 86	1,94	1,57	1,85
25	1,29	1,45	1,21	1,55	1,12	1,65	1,04	1,77	0,95	1,89	1,68	1,78
28	1,33	1,48	1,26	1,56	1,18	1,65	1,10	1,75	1,03	1,85	0,76	1,73
32	1,37	1,50	1,31	1,57	1,24	1,65	1,18	1,73	1,11	1,82	0,86	1,69
35	1,40	1,52	1,34	1,58	1,28	1,65	1,22	1,73	1,16	1,80	0,91	1,67
40	1,44	1,54	1,39	1,60	1,34	1,66	1,29	1,72	1,23	1,79	1,00	1,64
45	1,48	1,57	1,43	1,62	1,38	1,67	1,34	1,72	1,29	1,78	1,07	1,64
50	1,50	1,59	1,46	1,63	1,42	1,67	1,38	1,72	1,34	1,77	1,12	1,64
60	1,55	1,62	1,51	1,65	1,48	1,69	1,44	1,73	1,41	1,77	1,21	1,64
70	1,58	1,64	1,55	1,67	1,53	1,70	1,49	1,74	1,46	1,77	1,28	1,65
80	1,61	1,66	1,59	1,69	1,56	1,72	1,53	1,74	1,51	1,77	1,34	1,65
90	1,64	1,68	1,61	1,70	1,59	1,73	1,57	1,75	1,54	1,78	1,38	1,66
100	1,65	1,69	1,63	1,72	1,61	1,74	1,59	1,76	1,57	1,78	1,42	1,67
150	1,72	1,75	1,71	1,76	1,69	1,77	1,68	1,79	1,67	1,80	1,54	1,71
200	1,76	1,78	1,75	1,79	1,74	1,80	1,73	1,81	1,72	1,82	1,61	1,74

Значення синусів і косинусів для 1—4 гармонік

Місяць	Радіанна міра	Гармоніка							
		перша		друга		третя		четверта	
		sin	cos	sin	cos	sin	cos	sin	cos
1	0	0	1	0	1	0	1	0	1
2	π/6	0,5	0,866	0,866	0,5	1	0	0,866	–0,5
3	π/3	0,866	0,5	0,866	–0,5	0	–1	–0,866	–0,5
4	π/2	1	0	0	–1	–1	0	0	1
5	2π/3	0,866	–0,5	–0,866	–0,5	0	1	0,866	–0,5
6	5π/6	0,5	–0,866	–0,866	0,5	1	0	–0,866	–0,5
7	π	0	–1	0	1	0	–1	0	1
8	7π/6	–0,5	–0,866	0,866	0,5	–1	0	0,866	–0,5
9	4π/3	–0,866	–0,5	0,866	–0,5	0	1	–0,866	–0,5
10	3π/2	–1	0	0	–1	1	0	0	1
11	5π/3	–0,866	0,5	–0,866	–0,5	0	–1	0,866	–0,5
12	11π/6	–0,5	0,866	–0,866	0,5	–1	0	–0,866	–0,05

Квантилі *F*-розподілу (α = 0,05)

k_2	k_1									
	1	2	3	4	5	6	7	8	9	10
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64
8	5,32	4,46	4,07	3,84	3,69	3,59	3,50	3,44	3,39	3,35
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30
24	4,26	3,40	3,01	2,78	2,62	2.51	2,42	2,36	2,30	2,25
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99
120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,96	1,91
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83

Додаток 11

Критичні значення коефіцієнта детермінації R^2 і кореляційного відношення η^2 для рівня істотності $\alpha = 0{,}05$

k_2 / k_1	1	2	3	4	5
5	0,569	699	764	806	835
6	500	632	704	751	785
7	444	575	651	702	739
8	399	527	604	657	697
9	362	488	563	618	659
10	332	451	527	582	624
12	283	394	466	521	564
14	247	348	417	471	514
16	219	312	378	429	477
18	197	283	345	394	435
20	179	259	318	364	404
24	151	221	273	316	353
28	130	193	240	279	314
32	115	171	214	250	282
36	102	153	192	226	256
40	093	139	176	207	234
50	075	113	143	170	194
60	063	095	121	144	165
80	047	072	093	110	127
100	038	058	075	090	103
120	032	049	063	075	087
200	019	030	038	046	053

ЗМІСТ

Навчальне видання

ЄРІНА Антоніна Михайлівна

СТАТИСТИЧНЕ МОДЕЛЮВАННЯ ТА ПРОГНОЗУВАННЯ

Навчальнийпосібник

Редактор *І.Стремовська*
Художник обкладинки *Т. Зябліцева*
Технічний редактор *Т. Піхота*
Коректор *А. Невзгляд*
Верстка *Т. Мальчевськоі*