МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

ДЕРЖАВНИЙ УНІВЕРСИТЕТ «ЖИТОМИРСЬКА ПОЛІТЕХНІКА»

Бондарчук В.М., Головня Р.М., Громовий О.А., Давидова І.В.,

Ткачук А.Г.

Математичні методи аналізу даних

Навчальний посібник

Житомирська політехніка

2023

Рекомендовано до друку Вченою радою Державного університету «Житомирська політехніка»

(протокол № 9 від 30 червня 2023 р.)

Рецензенти:

Анатолій МАЦУЙ – доктор технічних наук, професор кафедри автоматизації виробничих процесів Центральноукраїнського національного технічного університету;

Олександр Сарана – кандидат фізико-математичних наук, доцент кафедри математичного аналізу Житомирського державного університету імені Івана Франка;

Юрій Подчашинський – доктор технічних наук, професор, завідувач кафедри метрології та інформаційно-вимірювальної техніки Державного університету «Житомирська політехніка».

Математичні методи аналізу даних: навч. посібник / Бондарчук В.М., Головня Р.М., Громовий О.А., Давидова І.В., Ткачук А.Г.. – Житомир: Державний університет «Житомирська політехніка», 2023. – 145 с.

Посібник містить теоретичний матеріал і велику кількість розв’язаних прикладів. Посібник призначений, у першу чергу, для забезпечення самостійної роботи студентів денної та заочної форм навчання. Він може використовуватись також як збірник задач під час проведення практичних занять зі студентами денної форми навчання та для виконання ними контрольних, індивідуальних домашніх або розрахунково-графічних робіт.

Вступ

Пропонований посібник укладено для студентів факультету комп'ютерно-інтегрованих технологій, мехатроніки та робототехніки денної та заочної форм навчання з метою забезпечення якісного засвоєння предмету «Багатофакторний аналіз. Дана методична розробка допоможе студентам опрацювати дисципліну, виробити уміння та навички розв'язання задач, що у свою чергу забезпечить успішне засвоєння навчального матеріалу.

Посібник містить велику кількість завдань, які можна використовувати як індивідуальні домашні завдання, завдання для типового розрахунку або завдання контрольних чи самостійних робіт. У посібнику містяться основні теоретичні відомості та приклади детального розв'язання типових задач з чітким порядком дій та правильним оформленням виконаного завдання.

Призначення навчального посібника – допомогти студентам глибше вивчити навчальну дисципліну.

Посібник може бути застосований і як доповнення до лекційного матеріалу, і для самостійної роботи студентів. У результаті вивчення матеріалу посібника студенти повинні знати основні поняття та формули, а також вміти застосовувати набуті знання при розв'язуванні прикладних задач.

Сучасна наука з різноманітністю її підходів та засобів спостереження, методів обробки інформації та моделювання екологічних, еколого-економічних систем є міждисциплінарним утворенням, що акумулює результати багатьох дисциплін, таких як математика й інформатика, статистика і теорія ймовірності та інші. Математичні методи аналізу даних широко застосовується для вирішення багатьох актуальних задач. Довгострокові прогнози, дослідження впливу на навколишнє середовище, моделі походження життя, вивчення людського організму, завдання генетики — ось далеко не повний перелік завдань, вирішення яких в даний час немислимо без застосування математичного моделювання.

Перші дослідження в області популяційного моделювання з'явилися в 20-і роки XX століття. Але бурхливий розвиток цей напрям отримав, починаючи з 1950-х років, що, безумовно, пов'язано з появою і швидким розвитком обчислювальної техніки. Серед великої кількості різноманітних моделей, розроблених на першому етапі, можна виділити такі класи моделей, як моделі з віковою структурою, просторово-розподілені моделі, дискретні відображення, статистичні моделі.

Необхідність застосування моделей під час опису природних об'єктів пов'язана з тим, що природні системи керуються одночасно багатьма факторами різної фізичної природи і не піддаються строгому кількісному опису. На відміну від закону, що має характер абсолютної істини, модель дає лише наближене уявлення про об'єкті, точніше, про ті його властивості, для вивчення яких здійснювалось моделювання. Створення математичної моделі здійснюється у наступній послідовності:

1) Отримання вихідних даних про об'єкт або явище шляхом вимірювання та визначення його властивостей.

2) Створення екологічної моделі об'єкта та формулювання екологічного завдання.

3) Вираження поставленої задачі в математичній формі.

Створення математичної моделі. При цьому може виникнути необхідність отримання додаткових даних або уточнення геологічних уявлень про об'єкт.

4) Математичні розрахунки відповідно до прийнятої моделі.

5) Перевірка відповідності отриманих результатів фактичним даним. Якщо екологічних моделей було кілька (це звичайний випадок), можна оцінити, яка з них краще відповідає дійсності.

Оскільки отримана модель враховує лише окремі властивості об'єкта, її можна послідовно ускладнювати та деталізувати. Чим складніша модель, тим достовірніше вона відображає об'єкт, що вивчається і дозволяє більш надійно прогнозувати його властивості. Однак у реальних умовах існує оптимальна міра складності математичних моделей, що визначається з урахуванням вимог до точності розв'язання поставленого завдання. Ступінь складності моделі може також обмежуватися можливостями аналітичних рішень та електронно-обчислювальної техніки.

Зрозуміло, що далеко не вся досліджувана сукупність доступна для спостереження. Науковцю найчастіше доводиться задовольнятися лише окремими параметрами, що характеризують частину досліджуваного об'єкта. Звідси зрозуміло, що необхідно розрізняти досліджувану та випробувану сукупність і завжди усвідомлювати, наскільки друга представницька стосовно першої. Проте джерело можливих помилок не обмежується розбіжністю досліджуваної та випробуваної сукупності. Остання також не може бути досліджена в повному обсязі. Еколог зазвичай обмежується певною кількістю зразків, проб, вимірів тощо. Множина всіх зроблених над випробуваною сукупністю спостережень утворює вибіркову сукупність, або просто вибірку. Очевидно, що вибіркова сукупність у багато разів менша випробуваної. Водночас саме за

результатами вибіркових спостережень робляться висновки не тільки щодо випробуваної, а й щодо всієї досліджуваної сукупності. Цю обставину завжди потрібно мати на увазі, роблячи будь-які висновки, інакше найточніші обчислення не врятують від помилки.

До вибіркових даних пред'являються наступні вимоги:

1) вибірка повинна складатися зі спостережень, отриманих у однакових умов;

2) спостереження мають бути незалежними один від одного.

Можливість поширення висновків, отриманих за вибірковими даними, на всю досліджувану сукупність забезпечується застосуванням методів математичної статистики.

Основні фактори, що враховуються при моделюванні екологічних систем, можна розділити на такі дві групи: а) фактори зовнішнього впливу: кліматичні зіни (температура, опади тощо); антропогенне втручання і таке ін.; б) внутрішні фактори: конкуренція; паразитизм; хижацтво; захворюваність та її поширення; трофічні ланцюги.

При цьому потрібно враховувати, що вплив таких факторів характеризується наявністю: ефекту запізнення; кумулятивного ефекту; граничних ефектів. Як правило, математичний опис впливу факторів зв'язаний з великою кількістю взаємозалежних змінних, зв'язаних між собою нелінійними співвідношеннями, що сильно ускладнює задачу і вимагає застосування ЕОМ. При побудові моделей екологічних процесів застосовують наступні основні принципи.

1) Принцип системності. Внаслідок пересиченості екосистем зв'язками екологічні об'єкти являють собою єдину систему. З цієї причини в екології виявилося необхідним злиття методів системного аналізу і математичного моделювання. Це призвело до створення інтегрального методу системного моделювання — вищого етапу в розвитку екологічного моделювання. Принцип системності полягає в усвідомленні цілісності об'єктів світу, їхньої стійкості, взаємодії із зовнішнім світом тощо; інший аспект цього принципу — динамічна багатогранність, єдність якості й кількості, теорії та практики.

2) Принцип єдності структурності та ієрархічності . Фундаментальна риса екосистем — наявність у них складних ієрархічних структур. Звідси випливає вимога єдності структурності й ієрархічності систем них екологічних моделей. Відповідно виникає проблема структурування моделі, тобто виділення істотних підсистем і елементів із сукупності всіх зв'язків і компонентів. Звичайно систему організують найбільш залежні одне від одного елементи (підсистеми). Інші впливають на поводження

системи слабко, а через їхню велику їх можна розглядати як інтегровані зовнішні чи внутрішні фактори впливу.

3) Принцип багатомодельного опису . Через динамізм і складність екологічних об'єктів, що виникають у результаті множинності мети антропогенного втручання, на сьогодні немає можливості побудови єдиної теорії соціоекосистеми в класичному розумінні, тобто як дедуктивної моделі, з якої можна вивести всі можливі наслідки. Тому наука йде по шляху створення множинних взаємодоповнюючих моделей.

4) Принцип єдності формалізованого і неформалізованого опису. Досвід перших глобальних моделей розвитку світової соціоекосистеми, побудованих за замовленням Римського клуба, показав: єдиного формалізованого (математичного) опису недостатньо для адекватного моделювання соціоекосистеми. Для цього необхідно враховувати неформальні фактори і доповнювати формалізований опис (з позицій історичного, психологічного та інших підходів) неформалізованим описом.

5) Принцип визнання фундаментальності екологічних процесів . Екологічні процеси неможливо звести до простої сукупності біологічних, фізичних, економічних процесів, оскільки всі вони тісно переплетені між собою. У цьому переплетенні виникають нові, екологічні закономірності. Звідси випливає самостійна значимість екологічних цінностей.

6) Принцип єдності теорії та практики. Благополуччя соціоекосистеми, частиною якої є Людина, має для неї найважливіше значення. Тому екологія є не тільки фундаментальною, але і прикладною наукою, що поєднує пізнання екологічних закономірностей із практичним їхнім застосуванням у повсякденній діяльності Людини. Значення моделювання в екології дуже велике. За допомогою моделювання одержують можливість оцінювання потенційних наслідків застосування різних стратегій оперативного керування, впливу на екосистему, користування природними ресурсами (біотичними й абіотичними), оптимізації екосистем. Моделювання дозволяє глибоко проникнути в сутність явищ, зрозуміти їхню справжню природу

Статистика - це наука, що вивчає закономірності, яким підпорядковані масові випадкові явища. Із цього визначення випливає, що використання математичної статистики для моделювання властивостей екологічних об'єктів та явищ можливе лише в тому випадку, коли екологічні спостереження задовольняють умову масовості (тобто їх можна багаторазово повторювати за одних і тих самих умов), можуть бути представлені у вигляді схеми випадкових подій та виражені випадковою

величиною. Проведення екологічних досліджень зазвичай полягає в вимірах значень досліджуваної властивості чи об'єкта в довільних точках простору. Тому ці виміри можна розглядати як серію випадкових подій, а отримані результати (числові значення) як випадкові величини, оскільки їх неможливо передбачити заздалегідь. Ці виміри можна повторювати багаторазово. Отже, явища, вивчені в процесі екологічних досліджень, можуть розглядатися як випадкові та масові та для них правомірно використання статистичних методів.

Теоретичною базою математичної статистики є теорія ймовірностей, окремі положення якої ми розглянемо нижче.

1. Основи теорії ймовірностей

2.1. Основні визначення та поняття

У статистичному моделюванні одним із головних є поняття про ймовірність випадкової події. Предметом теорії ймовірностей є математичний аналіз випадкових явищ, тобто явищ з невизначеними результатами. Наприклад, невизначені результати таких явищ, як підкидання монети, постріл з гармати у ціль, вибірковий контроль якості продукції, тривалість безвідмовної роботи приладу, визначення ціни акції під час торгів на фондовій біржі тощо.

Випадкові явища формалізуються в теорії ймовірностей у понятті випадкового експерименту. Це такий експеримент, який можна повторити скільки завгодно разів при заданих умовах, але його результати передбачити неможливо. Результати випадкового експерименту називають випадковими подіями. Усі події поділяються на достовірні, неможливі та випадкові. Достовірною називається подія, яка неодмінно відбудеться при кожному випробуванні.. Неможливою – подія, яка ніколи не реалізується при даному випробуванні. Події третього типу характеризуються тим, що вони можуть відбутися в цьому випробуванні, а можуть і не відбутися. Розрізняють випадкові події елементарні (ті, які не розкладаються на простіші) та складені. Для елементарних подій використовують, як правило, позначення ω, ω_1, ω_2 тощо, для складених – A, B, A_1, B_1 тощо. Сукупність усіх елементарних подій в даному випадковому експерименті називають простором елементарних подій. Його позначають літерою Ω. Якщо виділено простір елементарних подій, то довільна складена подія A – це сукупність тих елементарних подій, які сприяють її появі.

Розглянемо приклад випадкової події. По одному з рудних тіл мідного родовища відібрано за рівномірною сіткою 1000 проб, вміст міді в яких коливається від 0,1% до 5%. Кондиційним є вміст 2%. Наявність міді у будь-якій навмання взятій пробі буде подією достовірною, а ось вміст у ній міді понад 2% - подія випадкова. Якщо ми розділимо кількість проб з кондиційним вмістом на загальну кількість проб, то отримаємо величину коефіцієнта рудоносності для цього рудного тіла. Ця величина буде змінюватися від одного рудного тіла до іншого, причому заздалегідь не можна передбачити, яке значення вона набуде в кожному конкретному випадку, тобто це величина випадкова.

У різних практичних задачах доводиться розглядати числові функції, задані на просторі елементарних подій Ω . Такі функції називають випадковими величинами. Їх позначають X , Y , X_1 , Y_1 тощо. Наприклад, число появ герба при двох підкиданнях монети, число бракованих виробів серед відібраних для вибіркового контролю, тривалість безвідмовної роботи приладу тощо.

Розрізняють випадкові величини дискретні (ті, що набувають скінченного або зліченного числа значень) та неперервні (ті, що можуть набути довільного значення з деякого проміжку).

Якщо випадкова величина X – дискретна, то її вичерпною характеристикою є закон розподілу $p_1 = p\left(X = x_1\right)$, $p_2 = p\left(X = x_2\right)$, …, $p_n = p\left(X = x_n\right)$, де x_1, x_2, …, x_n – значення, яких набуває X .

Для неперервних випадкових величин доводиться використовувати складніше поняття функції розподілу $F\left(x\right) = P\left(X < x\right)$, $-\infty < x < +\infty$. Функція розподілу неперервної випадкової величини *X* може бути подана у вигляді

$F\left(x\right) = \int\limits_{-\infty}^{x} f\left(t\right)dt$, де функцію $f\left(t\right)$ називають щільністю розподілу. Очевидно, що $f\left(x\right) = \dfrac{dF\left(x\right)}{dx} = F'\left(x\right)$ у точках неперервності функції $f\left(x\right)$.

У розглянутому вище прикладі вміст міді в пробах коливається від 0,1 до 5,0%. Всередині цього інтервалу величина вмісту міді теоретично може приймати нескінченну кількість значень, тому є величиною неперервною.

Числовою характеристикою випадкової події є її ймовірність. Вона характеризує ступінь об'єктивної можливості появи випадкової події. Ймовірність випадкової події A позначають $P\ A$. Ймовірності повинні задовольняти умови:

1) $0 \le P\left(A\right) \le 1,$

2) $P\left(\Omega\right) = 1,$

3) ймовірність суми двох несумісних випадкових подій A та B дорівнює сумі їх ймовірностей: $P\left(A + B\right) = P\left(A\right) + P\left(B\right).$

Ці умови задовольняють, зокрема, ймовірності, що обчислюються за класичним визначенням, та геометричні ймовірності. Класичне означення ймовірності свідчить, що ймовірність появи події А дорівнює відношенню числа випадків, що сприяють появі події А, до загального числа випадків.

На практиці класичне визначення часто не застосовується, тому що загальна кількість випадків зазвичай невідома чи нескінченна. Крім того, далеко не завжди можна уявити результати досліду у вигляді рівноможливих та несумісних подій. Тим часом давно було помічено, що частота появи подій при багаторазовому повторенні досвіду має тенденцію стабілізації близько якоїсь постійної величини. Це свідчить про те, що дані події теж мають певним ступенем можливості появи у досліді, міру якої можна уявити у вигляді відносної частоти.

Відносна частота - це відношення числа дослідів, що сприяли події А, до загального числа виконаних випробувань. Швейцарським ученим Яковом Бернуллі доведено, що при великій кількості випробувань відносна частота прагне відтворити ймовірність і в границі збігається з нею. Отже, ймовірність $P\left(A\right)$ це відносна частота появи події у n проведених випробуваннях (статистичне означення ймовірності): $P\left(A\right) = \frac{m}{n}.$

Чим більше число n , тим ймовірність, визначена за цією формулою, ближче до її справжнього значення. Тому на практиці завжди треба прагнути того, щоб вибірка була достатньо представницькою.

2.2. Закон розподілу випадкової величини

Законом розподілу випадкової величини називається залежність між усіма можливими значеннями випадкової величини та відповідними їм ймовірностями. Закон розподіл може бути заданий у вигляді таблиці, графіка або функції розподілу. Табличний спосіб найпростіший, виглядає він в такий спосіб.

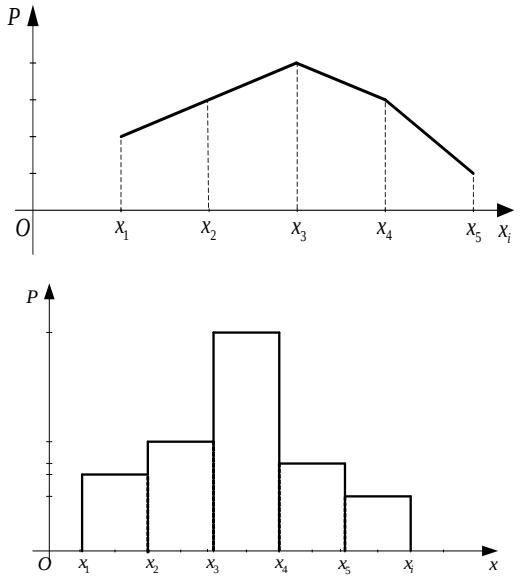
Таблиця 1

Задання закону розподілу

X	x_1	x_2	x_3	…	x_n
P	p_1	p_2	p_3	…	p_n

Зрозуміло, що табличне завдання закону розподілу можливо тільки для дискретної випадкової величини з скінченим числом значень.

На практиці неперервну випадкову величину зазвичай розбивають на ряд інтервалів та потім оперують з центрами інтервалів як з дискретною випадковою величиною. Графічне зображення такого ряду розподілу виглядає так.

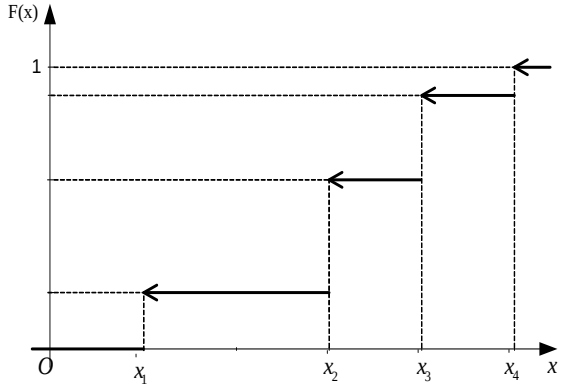


Мал. 1. Графічне зображення рядів розподілу

Якщо справжнє значення ймовірностей невідомо, по осі ординат відкладають відносну частоту появи кожного із значень.

Найбільш загальною формою завдання закону розподілу є функція розподілу Вона визначає ймовірність того, що випадкова величина x набуде значення, менше якогось фіксованого значення X . Ця ймовірність залежить від X і, отже, є функцією від X , тобто $F\left(x\right)=P\left(X<x\right)$.

Якщо ми побудуємо графічний вираз $F\left(x\right)=P\left(X<x\right)$ табличним даним, то отримаємо графік.



Мал. 2. Графік інтегральної функції розподілу дискретної випадкової величини

Неперервна випадкова величина має графік функції розподілу у вигляді плавної кривої.

Мал. 3. Графік інтегральної функції розподілу неперервної випадкової величини

Описана функція носить назву інтегральної функції розподілу. Відзначимо її основні властивості:

1) $F\left(x\right)$, як і будь-яка ймовірність, змінюється в межах від 0 до 1. $F\left(-\infty\right)=0$, $F\left(+\infty\right)=1$;

2) ймовірність попадання випадкової величини в інтервал від A до B дорівнює різниці ординат у точках A й B, тобто

$$P(A<x<B)=F(B)-F(A)$$

Неперервна випадкова величина може бути задана не тільки інтегральною, але і диференціальною функцією (або функцією щільності розподілу ймовірності). Вона є першою похідною від інтегральної функції:

$$f(x)=\frac{dF(x)}{dx}=F'(x)$$

Мал. 4. Графічне зображення функції густини ймовірності (диференціальна функція розподілу)

Виділимо на осі x елементарну ділянку dx. Ймовірність попадання випадкової величини на цю ділянку дорівнює $dF(x)=f(x)\cdot dx$

Тобто, це площа елементарного прямокутника з сторонами dx та $f(x)$ (рис. 4). Звідси випливає висновок у тому, що ймовірність влучення випадкової величини в інтервал від A до B чисельно дорівнює площі криволінійної трапеції, обмеженої графіком $f(x)$, віссю x та перпендикулярами в точках A та B. З курсу вищої математики ми знаємо, що ця площа дорівнює інтегралу функції $f(x)$ в межах від A до B: $S=\int_A^B f(x)dx$.

Отже, відзначимо основні властивості диференціальної функції розподілу.

1. Оскільки $f(x)$ неспадна функція, то її перша похідна завжди більша або дорівнює нулю. Це означає, що графік $f(x)$ повністю розташований вище осі x.

2. Інтегральна функція може бути виражена через диференціальну за формулою $F(x)=\int_{-\infty}^{x} f(t)dt$.

3. Імовірність того, що випадкова величина потрапить у інтервал від A до B дорівнює

$$P\left(A<x<B\right)=\int_{A}^{B}f\left(x\right)dx$$

4. Вся площа фігури, що знаходиться під кривою $f\left(x\right)$, характеризує повну ймовірність, тому дорівнює 1: $\int_{-\infty}^{+\infty}f\left(x\right)dx=1$

2.3.Основні характеристики положення та розсіювання випадкової величини

Закон розподілу повністю характеризує випадкову величину з ймовірнісного погляду. На практиці при розгляді випадкових величин часто обмежуються простішими характеристиками, ніж закон розподілу чи щільність розподілу. Зручніше користуватися деякими кількісними показниками, які у стислій формі дають досить повну інформацію про випадкову величину.

Найбільш суттєві особливості розподілу випадкової величини можуть бути виражені за допомогою числових характеристик положення і розсіювання. До найважливіших характеристикам положення відносяться математичне сподівання, мода та медіана.

Математичне сподівання характеризує положення випадкової величини на числовій осі, визначаючи собою деяке середнє значення, біля якого зосереджені всі можливі значення випадкової величини. Тому математичне очікування іноді називають просто середнім значенням випадкової величини. Математичне очікування дискретної випадкової величини можна визначити як середнє з її значень, зважених на ймовірність їх появи: $M\left(x\right)=\frac{\sum_{i=1}^{n}x_i p_i}{\sum_{i=1}^{n}p_i}$.

Так як $\sum_{i=1}^{n}p_i=1$ оскільки це є повна ймовірність. Отже, $M\left(x\right)=\sum_{i=1}^{n}x_i p_i$, тобто математичне очікування дискретної випадкової

величини є сума добутків її всіх можливих значень на відповідні їм ймовірності.

Можна довести, що зі збільшенням кількості випробувань середнє арифметичне $(\overline{x})$ дедалі більше наближається до $M(x)$, а при $n=\infty$ вони збігаються.

Модою (Mo) випадкової величини називається найбільш ймовірне її значення. Геометрично мода – це абсциса точки максимуму диференціальної кривої розподілу. Криві розподілу можуть бути одно- і багатомодальними.

Є також криві, які не мають максимуму, але мають мінімум. Вони називаються антимодальними (рис 5).

Мал. 5. Одномодальна (а), багатомодальна (б) та антимодальна (в) криві розподілу випадкової величини.

Медіана (Me) випадкової величини називається таке її значення, для якого ймовірність зустрічі більших та менших значень однакова:

$$F(Me)=P(X<Me)=P(X>Me)=0,5$$

З геометричної точки зору (Me) - це абсциса точки, якою площа, обмежена кривою розподілу, ділиться навпіл. Для визначення медіани дискретної випадкової величини можна розмістити всі її значення в порядку зростання (спадання). У разі парного числа значень, медіана дорівнює напівсумі двох середніх (по порядку) значень. Якщо крива розподілу симетрична щодо середнього значення, то $M(x)$, Mo та Me рівні між собою; в загальному випадку вони не збігаються.

Як характеристики розсіювання випадкової величини щодо середнього значення зазвичай використовують дисперсію, середнє квадратичне відхилення та коефіцієнт варіації.

Дисперсія $D(x)$ або σ^2 служить головною характеристикою розсіювання:

$$\sigma^2 = \frac{\sum_{i=1}^{n}\left(x_i - M\left(x\right)\right)^2}{n} = \sum_{i=1}^{n}\left(x_i - M\left(x\right)\right)^2 \cdot p_i$$

Можна використати й іншу формулу:

$$\sigma^2 = \sum_{i=1}^{n} x_i^2 \cdot p_i - \left(M\left(x\right)\right)^2$$

Оскільки дисперсія має розмірність квадрата випадкової величини, для оцінки розсіяння значень зазвичай використовують похідну від неї характеристику – середнє квадратичне відхилення: $\sigma = \sqrt{\sigma^2}$

Середнє квадратичне відхилення виражається у тих самих одиницях, як і випадкова величина та наочно показує розсіювання її значень. Однак для порівняння ступеня розсіювання двох величин, що мають різну розмірність, середнє квадратичне відхилення застосувати неможливо. В цьому випадку використовують безрозмірний показник – коефіцієнт варіації (ν):

$$\nu = \frac{\sigma}{M\left(x\right)} \cdot 100\%$$

Коефіцієнт варіації з успіхом використовується для порівняння ступеня мінливості різних екологічних об'єктів та явищ.

Криві розподілу випадкової величини можуть бути симетричними та асиметричними (рис. 6), стиснутими та розтягнутими (рис. 7). Ці їхні властивості відображаються у показниках асиметрії A і ексцесу E:

$$A = \frac{\sum_{i=1}^{n}\left(x_i - M\left(x\right)\right)^3}{n\sigma^3} = \frac{\sum_{i=1}^{n}\left(x_i - M\left(x\right)\right)^3 \cdot p_i}{\sigma^3}$$

$$E = \frac{\sum_{i=1}^{n}\left(x_i - M\left(x\right)\right)^4}{n\sigma^4} - 3 = \frac{\sum_{i=1}^{n}\left(x_i - M\left(x\right)\right)^4 \cdot p_i}{\sigma^4} - 3$$

Мал. 6. Симетричність та асиметричність кривих розподілу

Мал. 7. Стисліcть і розтягнутість кривих розподілу

Розрахунок характеристик положення та розсіювання випадкової величини можна здійснити не тільки за наведеними формулами, а й з допомогою моментів випадкової величини. В окремих випадках це виявляється зручнішим, особливо при застосуванні обчислювальної техніки.

2.4. Деякі теоретичні закони розподілу випадкової величини

Для наближеного опису розподілу властивостей екологічних об'єктів, що спостерігаються емпірично, у практиці застосовують найрізноманітніші теоретичні закони розподіл випадкової величини. При цьому часто обмежуються використанням чотирьох основних законів: нормального, логнормального, біномінального, Пуассона.

Закон нормального розподілу, так званий закон Гаусса, - один з найпоширеніших законів. Це фундаментальний закон у теорії ймовірностей і в її застосуванні. Нормальний розподіл найчастіше зустрічається у вивченні природних і соціально-економічних явищ. Інакше кажучи, більшість статистичних сукупностей у природі і суспільстві підпорядковується закону нормального розподілу. Відповідно можна сказати, що сукупності значної частини великих за обсягом вибірок підпорядковуються закону нормального розподілу. Ті із сукупностей, які відхиляються від нормального розподілу в результаті спеціальних перетворень, можуть бути наближені до нормального. У зв'язку з цим слід пам'ятати, що принципова особливість цього закону стосовно інших законів розподілу полягає в тому, що він є граничним законом, до якого наближаються інші закони розподілу в певних (типових) умовах.

Нормальним називають розподіл ймовірностей неперервної випадкової величини, для якого інтегральна функція розподілу має вигляд:

$$F\left(x\right)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{\left(x-M\left(x\right)\right)^{2}}{2\sigma^{2}}}dx$$

Відповідно щільність розподілу має вигляд:

$$f\left(x\right)=\frac{1}{\sigma\sqrt{2\pi}}\cdot e^{-\frac{\left(x-M\left(x\right)\right)^{2}}{2\sigma^{2}}}$$

де $M(x)$ - математичне сподівання або середня величина. Як видно, нормальний розподіл визначається двома параметрами: $M(x)$ і σ . Щоб задати нормальний розподіл, досить знати математичне сподівання, або середнє і середнє квадратичне відхилення. Ці дві величини визначають центр групування і форму кривої на графіку. Графік функції розподілу називається нормальною кривою (крива Гаусса) з параметрами $M(x)$ і σ (рис. 8).

При нормальному розподілі середня арифметична, мода і медіана будуть рівними між собою.

Форма нормальної кривої має вид одновершинної симетричної кривої, вітки якої асимптотично наближаються до осі абсцис. Найбільша ордината кривої відповідає $x = M(x)$. У цій точці на осі абсцис розміщується чисельне значення ознак, яке дорівнює середній арифметичній, моді і медіані. По обидві сторони від вершини кривої її вітки спадають, змінюючи в певних точках форму випуклості на увігнутість.

Рис.12. Крива нормального розподілу (крива Гауса)

Форму і положення нормальної кривої зумовлюють значення середньої і середнього квадратичного відхилення. Математично доведено, що зміна величини середньої (математичного сподівання) не змінює форми нормальної кривої, а призводить лише до її зміщення вздовж осі абсцис. Крива зрушується вправо, якщо $M(x)$ зростає, і вліво, якщо $M(x)$ спадає.

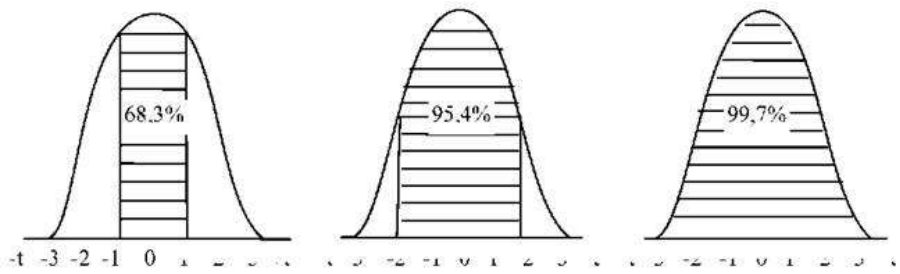


Рис.13. Нормальний розподіл з одно-, дво- та трьосигмовими границями

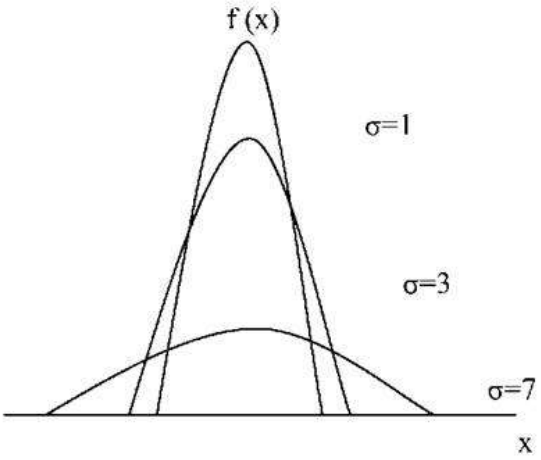


Рис.14. Криві нормального розподілу з різними значеннями параметра σ.

Про зміну форми графіка нормальної кривої при зміні середнього квадратичного відхилення можна судити по максимуму диференціальної функції нормального розподілу, який дорівнює 1.

Як видно, при зростанні величини σ максимальна ордината кривої буде зменшуватися. Отже, крива нормального розподілу буде стискуватися до осі абсцис і приймати більш плосковершинну форму. І, навпаки, при зменшенні параметра σ нормальна крива витягується в додатному напрямку осі ординат, а форма "дзвона" стає більш гостровершиною. Відзначимо, що незалежно від величини параметрів $M\left(x\right)$ і σ площа, обмежена віссю абсцис і кривою, завжди дорівнює одиниці (властивість щільності розподілу).

Припустимо, що $M\left(x\right)=a$. Якщо замість випадкової величини розглянути нову випадкову величину

$$t=\frac{x-a}{\sigma},$$

то нова величина t буде також розподілена нормально з середнім значенням, що дорівнює нулю та дисперсією, що дорівнює 1:

Щільність ймовірності величини t має вигляд:

$$f\left(t\right)=\frac{1}{\sqrt{2\pi}}\cdot e^{-\frac{t^2}{2}}.$$

При $M\left(x\right)=a$ і $\sigma=1$ нормальну криву називають нормованою кривою або нормальним розподілом у канонічному вигляді. Перетворений таким чином розподіл називається нормованим або стандартним нормальним розподілом.

Перехід від нормального до стандартного нормального розподілу полягає в перенесенні центру розподілу в початок координат з виразом випадкової величини у частках її стандарту. Необхідність такого перетворення полягає в тому, що обчислення ймовірностей за формулою (12) є собою дуже трудомістку операцію, а скласти таблиці для всіх можливих значень випадкової величини не є можливим. Такі таблиці складені для нормованої (безрозмірної) величини t для якої, як ми побачимо нижче, цілком достатньо мати таблицю значень $F\left(t\right)$ на інтервалі $\left[-3;3\right]$. Перехід від реальних значень випадкової величини до нормованих за формулою (14) не представляє жодних труднощів.

У довідниках наводяться таблиці для $f\left(t\right)$, $F\left(t\right)$ та $\Phi\left(t\right)$:

$$f\left(t\right)=\frac{1}{\sqrt{2\pi}}\cdot e^{-\frac{t^2}{2}}$$

$$F\left(t\right)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{t^2}{2}}dt$$

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

Інтеграл (16) визначає ймовірність попадання випадкової величини на інтервалі від $-\infty$ до t, а інтеграл (17) від $-t$ до $+t$.

Щоб визначити ймовірність попадання випадкової величини в інтервал від A до B, необхідно спочатку нормувати межі інтервалу:

$$t_1=\frac{A-M(x)}{\sigma},\ t_2=\frac{B-M(x)}{\sigma},$$

а потім знайти відповідні значення $F(t)$ у таблиці та обчислити шукану ймовірність: $P(A\le x\le B)=F(t_2)-F(t_1)$

Якщо A і B розташовані симетрично щодо $M(x)$, то завдання спрощується: знаходимо $t=t_1=t_2$ і визначаємо за таблицею $\Phi(t)$ шукану ймовірність: $P=\Phi(t)$.

Названі вище особливості прояву "нормальності" розподілу дозволяють виділити ряд загальних властивостей, які мають криві нормального розподілу:

1) будь-яка нормальна крива досягає максимуму в точці $x=M(x)$ (або $t=0$), спадає неперервно вправо і вліво від неї, поступово наближаючись до осі абсцис. При $t=\pm\infty$ функція $f(t)$ прагне до нуля. Власне, вже при $t>3$, $f(t)$ практично дорівнює 0: $\Phi(t=1)=0,6827$, $\Phi(t=2)=0,9545$, $\Phi(t=3)=0,9973$). Інакше кажучи, практично всі значення випадкової величини (99,73%) укладаються в інтервал $M(x)\pm3\sigma$. На цій властивості засновано широко використовуване в геохімії правило "трьох сигм", згідно з яким концентрації елементів, що перевищують фон більш ніж на три середні квадратичні відхилення, вважаються аномальними.

2) будь-яка нормальна крива симетрична відносно прямої, паралельної осі ординат і проходить через точку максимуму $x = M\left(x\right)$ (або $t = 0$, якщо розподіл нормований);

3) будь-яка нормальна крива має форму "дзвона", має випуклість, яка направлена вверх до точки максимуму. У точках $M\left(x\right) - 0$ і $M\left(x\right) + 0$ вона змінює випуклість, і , чим менше σ, тим гостріше "дзвін", а чим більше σ, тим більш похилою стає вершина "дзвону" (рис.14).

4)При $t = 0$ щільність ймовірності максимальна: $f\left(t\right) = 0,3989$

Розглянемо приклад використання таблиць нормального розподілу:

На одному із досліджуваних об'єктів встановлено, що середній вміст забруднюючих речовин становить 7,5 $г/м$ при $\sigma = 3,5$. Яка ймовірність того, що в навмання взятому зразку вміст забруднюючих речовин коливатиметься від 11 до 18 $г/м$.

Розв'язання: В нашому випадку $M\left(x\right) = 7,5\ г/м$, $\sigma = 3,5$.

Щоб визначити ймовірність попадання випадкової величини в інтервал від 11 до 18, необхідно спочатку нормувати межі інтервалу:

$$t_1 = \frac{11 - 7,5}{3,5} = 1,\ t_2 = \frac{18 - 7,5}{3,5} = 3,$$

$$P\left(A \le x \le B\right) = F\left(t_2\right) - F\left(t_1\right) = F\left(3\right) - Fd\left(1\right) = 0,9986 - 0$$

Це означає, що у кожних 15-16 пробах зі 100 взятих навмання з даного досліджуваного об'єкта вміст забруднюючих речовин складе від 11 до 18 $г/м$.

У тісному зв'язку з нормальним перебуває логарифмічно нормальний (логнормальний) закон розподілу, який дуже широко застосовується у екології. Встановлено, що цим законом задовільно описується розподіл ряду хімічних елементів у зразках, розподіл вмісту речовин в розчинах, розподіл діаметру часток при дробленні і т.д. При логнормальному розподілі нормальному закону підпорядковані не самі значення випадкової величини, які її логарифми. Тому спочатку знаходять

натуральні (або десяткові) логарифми всіх значень випадкової величини, а потім усі операції проводять з логарифмами, як з звичайними числами: обчислюють їх статистичні характеристики та за таблицями нормального розподілу визначають ймовірність. У разі, якщо у вихідній сукупності зустрічаються нульові значення, їх замінюють мінімальними або половиною чутливості аналізу, оскільки логарифмувати нульові значення не можна.

Крива щільності ймовірності логнормального розподілу, побудована не за логарифмами, а за вихідними значенням, є асиметричною та описується наступним виразом:

$$f\left(x\right)=\frac{1}{x\sigma_{\ln x}\sqrt{2\pi}}\cdot e^{-\frac{\left(\ln x-M_{\ln x}\right)^2}{2\sigma_{\ln x}^{\ 2}}},$$

де $M_{\ln x}$ і $\sigma_{\ln x}$ математичне сподівання та середнє квадратичне відхилення (стандарт) логарифмів значень.

Ця функція досягає максимуму в точці

$$Mo=e^{M_{\ln x}-\sigma_{\ln x}^2}$$

Медіана (або середнє геометричне) дорівнює

$$Me=e^{M_{\ln x}}$$

Математичне сподівання дорівнює

$$M\left(x\right)=e^{M_{\ln x}+\frac{\sigma_{\ln x}}{2}}$$

Дисперсія визначається співвідношенням

$$\sigma_x^2=e^{2M_{\ln x}}\left(e^{2\sigma_{\ln x}}-e^{\sigma_{\ln x}}\right).$$

Асиметрія та ексцес функції позитивні.

Таблиці для логнормального розподілу відсутні, тому теоретичну криву щільності ймовірності будують безпосередньо за формулою. Крива логнормального розподілу завжди додатня і має правобічну скошеність (асиметрично), тобто вона вказує на велику імовірність відхилення вгору.

Для моделювання звичайно використовується зв'язок з нормальним розподілом. Тому, достатньо згенерувати нормально розподілену випадкову величину, наприклад, використовуючи перетворення Бокса-Мюллера, і обчислити її експоненту.

Моделювання значень випадкової величини з логнормальним розподілом (з параметрами $M(x)$ і σ) проводиться за формулою $X = e^Y$, де Y має нормальний розподіл з тими ж параметрами.

Біноміальний закон розподілу використовується у тих випадках, коли в результаті одного випробування подія A може або з'явитися з ймовірністю p, або не з'явитися з ймовірністю $q = 1 - p$. Подібна схема випробування називається схемою Бернуллі. Я. Бернуллі винайшов закон біноміального розподілу, згідно з яким ймовірність того, що подія A відбудеться в n випробуваннях рівно x разів дорівнює:

$$P_n(x) = C_n^x \cdot p^x \cdot q^{n-x} = \frac{n!}{x!(n-x)!} \cdot p^x \cdot (1-p)^{n-x}$$

Тут n і p є параметрами біноміального розподілу.

Основні характеристики біноміального розподілу визначаються наступними формулами:

$$M(x) = np$$

$$\sigma_x^2 = np(1-p)$$

$$A = \frac{q-p}{\sqrt{npq}} = \frac{1-2p}{\sqrt{np(1-p)}}$$

$$A = \frac{1-6pq}{npq} = \frac{1-6p(1-p)}{np(1-p)}$$

За допомогою біноміального закону розподілу описується лише розподіл дискретних величин. Коефіцієнти C_n^x при $x = 1, 2, 3\ldots$ утворюють ряд коефіцієнтів розкладе бінома Ньютона, тому розподіл і називається біноміальним. Ці коефіцієнти можна знайти за спеціальними таблицями, або за трикутником Паскаля (якщо $x \le 12$).

У тих випадках, коли n дуже великі числа, обчислення ймовірності за формулою (19) є громіздкими. У цьому випадку рекомендується застосування наближеної формули Муавра-Лапласа:

$$p_n(x) \approx \frac{1}{\sqrt{npq}} f(t)$$

де $f(t)$ - функція щільності ймовірності стандартного нормального розподілу,

$$t = \frac{x - M(x)}{\sigma} = \frac{x - np}{\sqrt{npq}}$$

Значення $f(t)$ беруть із таблиці.

Розглянемо приклад.

На досліджуваному об'єкті було відібрано 600 проб, з них з кондиційним вмістом домішок 302 проби. Необхідно визначити ймовірність того, що з 10 навмання взятих проб кондиційних буде 0,1,2... 10 проб.

Розв'язування: Знайдемо ймовірність появи проби з кондиційним вмістом домішок. За класичним означеням ймовірності:

$$p = \frac{m}{n} = \frac{302}{600} = 0,5$$

Оскільки проб мало, обчислення ведемо за формулою Бернуллі: $P_n(x) = C_n^x \cdot p^x \cdot q^{n-x}$.

x_i	0	1	2	3	4	5	6	7	8	9	10
p_i	0,001	0,01	0,044	0,117	0,205	0,246	0,205	0,117	0,044	0,01	0,001

Наприклад:

$$P_{10}(x=0) = C_{10}^0 \cdot 0,5^0 \cdot 0,5^{10-0} = \frac{10!}{0! \cdot (10-0)!} \cdot 0,5^0 \cdot (1-0,5)^{10}$$

$$P_{10}(x=3) = C_{10}^3 \cdot 0,5^3 \cdot 0,5^{10-3} = \frac{10!}{3! \cdot (10-3)!} \cdot 0,5^3 \cdot (1-0,5)^7 =$$

$$P_{10}\left(x=5\right)=C_{10}^{5}\cdot 0,5^{5}\cdot 0,5^{10-5}=\frac{10!}{5!\cdot\left(10-5\right)!}\cdot 0,5^{5}\cdot\left(1-0,5\right)^{5}=$$

$$P_{10}\left(x=8\right)=C_{10}^{8}\cdot 0,5^{8}\cdot 0,5^{10-8}=\frac{10!}{8!\cdot\left(10-8\right)!}\cdot 0,5^{8}\cdot\left(1-0,5\right)^{2}=$$

Припустимо тепер, що у нас виникла потреба визначити ймовірність того, що зі 100 взятих проб кондиційними виявляться 55. Формула (19) у цьому випадку виявляється малопридатною, тому скористаємося локальною формулою Муавра-Лапласа:

$$p_n\left(x\right)\approx\frac{1}{\sqrt{npq}}f\left(t\right)$$

Знайдемо $f\left(t\right)$:

$$t=\frac{x-np}{\sqrt{npq}}=\frac{55-100\cdot 0,5}{\sqrt{100\cdot 0,5\cdot\left(1-0,5\right)}}=1$$

За таблицею значень диференціальної функції Лапласа: $f\left(1\right)=0,2420$

Отже, $p_{100}\left(55\right)\approx\frac{1}{\sqrt{100\cdot 0,5\cdot\left(1-0,5\right)}}\cdot 0,2420=0,048$.

При $n\to\infty$ біноміальний розподіл прагне до нормального, але якщо при цьому p або q прямує до нуля, то випадкова величина починає підкорятися розподілу Пуассона. Формула Муавра-Лапласа у цьому випадку стає малопридатною, а при $p\to 0$ втрачає сенс. Вираз, що визначає ймовірність появи малоймовірної події в серії з n випробувань, було знайдено Пуассоном:

$$p_n\left(x=m\right)\approx\frac{\lambda^{m}\cdot e^{-\lambda}}{m!}$$

де $\lambda = n \cdot p$ є єдиним параметром розподілу Пуассона. Можна легко переконатися, що

$$M(x) = \sigma_x^2 = \lambda = np$$

$$A = \frac{1}{\sqrt{\lambda}} = \frac{1}{\sqrt{np}}$$

$$E = \frac{1}{\lambda} = \frac{1}{np}$$

Функція розподілу такої випадкової величини є сумою:

$$P_n(x \le m) = \sum_{k=0}^{m} \frac{\lambda \cdot e^{-\lambda}}{k!},$$

де $k = 0,1,2,...,m$.

Якщо n недостатньо велике, а поодинока ймовірність p недостатньо мала $(> 0,1)$, то ймовірність, що обчислюється за формулою Пуассона містить помітну похибку. Для цих випадків А.Н. Колмогоровим запропоновано виправлену формулу:

$$P_m' = \frac{\lambda^m \cdot e^{-\lambda}}{m!} - \frac{b}{2} \cdot \frac{\lambda^{m-2} \cdot e^{-\lambda}}{2(m-2)!} \cdot \left(\frac{\lambda^2}{m(m-1)} - \frac{2\lambda}{m-1} + 1 \right)$$

Формула Колмогорова враховує і можливу зміну одиничної ймовірності, в такому випадку:

$$\lambda = p_1 + p_2 + p_3 + ... + p_n$$

$$b = p_1^2 + p_2^2 + p_3^2 + ... + p_n^2$$

Тобто запропоноване Пуассоном значення $\lambda = pn$ є окремим випадком, коли всі p рівні між собою.

Розглянемо приклад розподілу Пуассона.

У басейні одного з водотоків відібрано 150 проб, у деяких з них є вміст радіоактивних ізотопів (табл.4).

Розв'язання: Розрахуємо середнє вибіркове та дисперсію:

$$\bar{x} = \frac{0\cdot 32 + 1\cdot 51 + 2\cdot 36 + 3\cdot 19 + 4\cdot 8 + 5\cdot 2 + 6\cdot 1 + 7\cdot 0}{150} = 1{,}52$$

$$S^2 = \frac{0^2\cdot 32 + 1^2\cdot 51 + 2^2\cdot 36 + 3^2\cdot 19 + 4^2\cdot 8 + 5^2\cdot 2 + 6^2\cdot 1 + 7^2\cdot 0}{150} - 1{,}52^2 = 1{,}55$$

Як бачимо, $\bar{x} \approx S^2$, що є однією з ознак розподілу Пуассона. Підставивши $\lambda = 1{,}52$ у формулу (21),розрахуємо p_m. Потім розраховуємо теоретичну частоту, округляючи її до цілих чисел. Майже повний збіг теоретичної та фактичної частот свідчить про те, що розподіл кількості радіоактивних ізотопів в пробах даного водотоку дійсно підпорядковується закону Пуассона.

Кількість радіоактивних ізотопів	0	1	2	3	4	5	6	7
Кількість проб	32	51	36	19	8	2	1	0
p_m	0,22	0,33	0,25	0,12	0,04	0,01	0,01	0
Np_m	33,4	50,1	37,6	18,8	7,1	2,1	1,2	0,1
Теоретична частота	33	50	38	19	7	2	1	0

Крім розглянутих чотирьох законів розподілу на геології використовуються й інші, зокрема, розподіли, похідні від логнормального, розподіл Пойа, Лапласа, рівномірний та інші.

3. Статистика випадкових величин

3.1. Статистичні оцінки невідомих параметрів

Кожна сукупність може бути поділена на досліджувану, випробувану та вибіркову. З цього виходить що досліджувана та випробувана сукупності характеризуються деякими невідомими нам значеннями досліджуваних властивостей, найчастіше середніми величинами та дисперсіями, про які ми можемо говорити з урахуванням вибіркових даних. Вибірки часто бувають обмежені за обсягом, тому питання щодо їх використання для судження про невідомі параметри генеральної сукупності стоїть особливо гостро.

Отримані за вибірковими даними наближені характеристики будь-яких властивостей досліджуваної сукупності називаються їх оцінками. Наприклад, як оцінка невідомого середнього значення найчастіше

використовується середнє арифметичне за вибіркою, хоча можливі і інші варіанти оцінок цього параметра: середнє геометричне, середнє гармонійне та ін. У зв'язку з цим завжди виникає питання про вибір з набору можливих варіантів оцінок параметрів тих із них, які задовольняють деяким вимогам якості.

Статистичні оцінки можуть бути точковими і інтервальними. При точковій оцінці невідома характеристика оцінюється деяким числом, а за інтервальної оцінки вказується певний інтервал значень, в межах якого із заданою ймовірністю має бути справжнє значення оцінюваної величини.

Основна задача математичної статистики полягає в знаходженні розподілу досліджуваної випадкової величини X за даними вибірки. В багатьох випадках вигляд розподілу X можна вважати відомим і задача зводиться до визначення невідомих параметрів цього розподілу. Нехай розподіл досліджуваної випадкової величини X містить один невідомий параметр θ.

Точковою оцінкою параметра θ називається його наближене значення $\overline{\theta}$, одержане за вибіркою. Оскільки елементи вибірки $\xi_1, \xi_2, ..., \xi_n$ можна розглядати як незалежні випадкові величини, то число $\overline{\theta}$ є значенням деякої функції $\overline{\theta} = \overline{\theta}\left(\xi_1, \xi_2, ..., \xi_n\right)$, яку називають ***статистичною оцінкою*** параметра θ. Значення цієї функції змінюються від вибірки до вибірки, а тому її можна розглядати як випадкову величину. Функцію $\overline{\theta}\left(\xi_1, \xi_2, ..., \xi_n\right)$ потрібно вибрати такою, щоб випадкова величина давала найкраще наближення параметра θ, а це можливо за умов її незміщеності, ефективності та обґрунтованості.

Незміщеною називається статистична оцінка $\overline{\theta}$, математичне сподівання якої дорівнює оцінюваному параметру θ при довільному об’ємі вибірки: $M\left(\overline{\theta}\right) = \theta$

Ефективною називається статистична оцінка, яка при заданому об’ємі вибірки має найменшу можливу дисперсію.

Обґрунтованою називається статистична оцінка $\overline{\theta}$, яка при $n \to \infty$ прямує за ймовірністю до оцінюваного параметру θ, тобто для довільного $\varepsilon > 0$ $\lim\limits_{n\to\infty} P\left(\left|\overline{\theta} - \theta\right| < \varepsilon\right) = 1$.

Простішим методом статистичного оцінювання є ***метод підстановки*** або ***аналогії***, який полягає в тому, що в якості тієї чи іншої числової характеристики генеральної сукупності беруть відповідну характеристику розподілу вибірки – вибіркову характеристику.

За методом підстановки в якості оцінки $\tilde{M}\left(X\right)$ математичного сподівання досліджуваної випадкової величини X потрібно взяти вибіркову середню, в якості оцінки $\tilde{D}\left(X\right)$ дисперсії – вибіркову дисперсію і т.д.

Вибіркова середня є незміщеною і обґрунтованою оцінкою математичного сподівання, вибіркова дисперсія – зміщеною оцінкою дисперсії, тобто $M\left(D_в\right) \neq D\left(X\right)$, а незміщеною оцінкою дисперсії є ***виправлена*** вибіркова дисперсія: $S^2 = \dfrac{n}{n-1} D_в$

На практиці користуються виправленою вибірковою дисперсією, коли об'єм вибірки $n < 30$.

Точність оцінки та довірчі інтервали.

При малому об'ємі вибірки точкові оцінки можуть мати значне розходження із значеннями параметра, що вивчається. Більш точними є інтервальні оцінки.

Інтервальною називають оцінку, яка визначається двома числами – кінцями інтервалу.

Нехай за даними вибірки знайдена статистична оцінка $\overline{\theta}$ деякого невідомого параметра θ. Очевидно, що $\overline{\theta}$ тим точніше визначає параметр θ, чим менша за абсолютною величиною різниця $\overline{\theta} - \theta$. Число δ, для якого виконується нерівність $\left|\overline{\theta} - \theta\right| < \delta$ називають *точністю* оцінки, а ймовірність γ, з якою вона виконується – *довірчою ймовірністю* або *надійністю* оцінки:

$$\gamma = P\left(\left|\overline{\theta} - \theta\right| < \delta\right)$$

Надійність оцінки задають наперед і в якості значення γ беруть число близьке до 1, наприклад, 0,95, 0,99, 0,999 і т.д.

Нерівність $\left|\overline{\theta} - \theta\right| < \delta$ рівносильна нерівності

$\overline{\theta} - \delta < \theta < \overline{\theta} + \delta$, а тому формулу (9.2) можна записати у вигляді:

$$\gamma = P\left(\overline{\theta} - \delta < \theta < \overline{\theta} + \delta\right)$$

Інтервал $\left(\overline{\theta} - \delta; \overline{\theta} + \delta\right)$, який із заданою надійністю γ покриває невідомий параметр θ, називають *довірчим* інтервалом. У прикладних статистичних задачах довжина надійного інтервалу грає важливу роль: чим менша його довжина, тим точніша його оцінка. Якщо довжина цього інтервалу велика, то значимість такої оцінки незначна.

Надійний інтервал для оцінки математичного сподівання нормального розподілу при відомій дисперсії

Нехай кількісна ознака X генеральної сукупності розподілена за нормальним законом і відоме середнє квадратичне відхилення σ. Треба знайти довірчий інтервал, що покриває математичне сподівання $M\left(x\right) = a$ генеральної сукупності з заданою надійністю γ.

Надійний інтервал для математичного сподівання будь-якого розподілу легко знайти за допомогою центральної граничної теореми. Якщо $x_1, x_2, \ldots, x_k$ - вибірка з генеральної сукупності X, розподіленої за довільним законом з математичним сподіванням a і дисперсією σ^2, то $x_1, x_2, \ldots, x_k$ є взаємно незалежними й однаково розподіленими випадковими величинами з математичним сподіванням a і дисперсією σ^2. З центральної граничної теореми випливає, що при довірчій ймовірності $P\left(\left|\overline{x} - a\right| < \dfrac{\sigma t}{\sqrt{n}}\right) \approx 2\Phi\left(t\right)$, де $\Phi\left(t\right)$ - функція Лапласа, тобто при рівні надійності $\gamma = 2\Phi\left(t\right)$ надійним інтервалом

для математичного сподівання при відомому σ є інтервал

$P\left(\overline{x}_в - \delta; \overline{x}_в + \delta\right) = 2\Phi\left(t\right) = \gamma$, де δ - точність оцінки, $\delta = \frac{\sigma t}{\sqrt{n}}$.

Приклад 4. Побудувати надійний інтервал для оцінки з надійністю $\gamma = 0,99$ невідомого математичного сподівання a нормально розподіленої генеральної сукупності X, якщо $\sigma = 2$, $\overline{x} = 15,35$ і $n = 16$.

Розв'язання. Шуканий надійний інтервал має вигляд

$$\overline{x} - \frac{t_\gamma \sigma}{\sqrt{n}} < a < \overline{x} + \frac{t_\gamma \sigma}{\sqrt{n}}$$

де t_γ – значення аргументу функції Лапласа

$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_0^t e^{-\frac{x^2}{2}} dx$, при якому $\Phi(t_\gamma) = \frac{\gamma}{2}$.

Знаходимо t_γ зі співвідношення $\Phi\left(t_\gamma\right) = \frac{0,99}{2} = 0,495$: за таблицею значень функції Лапласа (додаток 2) маємо $t_\gamma = 2,58$.

Підставляючи $\sigma = 2$, $\overline{x} = 15,35$, $n = 16$, $t_\gamma = 2,58$ в формулу для знаходження надійного інтервалу, отримаємо надійний інтервал:

$$15,35 - \frac{2,58 \cdot 2}{\sqrt{16}} < a < 15,35 + \frac{2,58 \cdot 2}{\sqrt{16}}$$

$$14,06 < a < 16,64$$

Приклад 5. Визначити, з якою надійністю γ можна гарантувати точність оцінки, яка не перевищує $\delta = 10$, якщо $\sigma = 16$ та $n = 15$.

Розв'язання. З формули $\delta = \frac{\sigma t}{\sqrt{n}}$ знаходимо t

$$t=\frac{\delta\sqrt{n}}{\sigma}=\frac{10\sqrt{15}}{16}=2,42.\ 2\Phi\left(2,42\right)=\gamma.$$

Отже, шукана ймовірність буде дорівнювати $\gamma=0,9844$.

Надійний інтервал для оцінки математичного сподівання нормального розподілу при невідомій дисперсії

Нехай потрібно оцінити математичне сподівання генеральної сукупності, яка має нормальний розподіл, при невідомій дисперсії σ^2. Так як σ невідома, то не можна скористатися результатами, в якому σ припускалось відомим.

Обчислимо за вибіркою виправлену дисперсію $S^2=\frac{n}{n-1}\sigma_в^2$.

Визначимося з надійною ймовірністю γ та знайдемо таке число ε, щоб виконувалося співвідношення $P\left(\overline{x}_в-\varepsilon<a<\overline{x}_в+\varepsilon\right)=\gamma$.

Виявляється, що за даними вибірки можна побудувати випадкову величину $T=\frac{\left(\overline{x}-a\right)\sqrt{n}}{S}$, що має розподіл Стьюдента зі степенями вільностей $k=n-1$, яка не залежить від параметра σ.

Ймовірність виконання нерівності $T<\varepsilon$ визначається наступним чином:

$$P\left(\left|\frac{\left(\overline{x}-a\right)\sqrt{n}}{S}\right|<\varepsilon\right)=\gamma.$$

За таблицею розподілу Стьюдента (додаток 3) отримаємо значення $\varepsilon=t$, при якому виконується рівність

$$P\left(\left|\frac{\left(\overline{x}-a\right)\sqrt{n}}{S}\right|<\frac{tS}{\sqrt{n}}\right)=\gamma$$

При рівні надійності $\gamma=2\Phi\left(t\right)$ для a беруть надійний інтервал

$$P\left(\bar{x}-t\frac{S}{\sqrt{n}}<a<\bar{x}+t\frac{S}{\sqrt{n}}\right)=\gamma,$$

де $t=t\left(\gamma,n\right)$ – значення, яке береться з таблиці розподілу Стьюдента (додаток 3).

Приклад 6. Побудувати надійний інтервал для оцінки з надійністю $\gamma=0,99$ невідомого математичного сподівання a нормально розподіленої генеральної сукупності *X*, якщо $S=2$, $\bar{x}=15,35$ і $n=16$.

Розв'язання. Шуканий надійний інтервал має вигляд

$$\bar{x}-\frac{tS}{\sqrt{n}}<a<\bar{x}+\frac{tS}{\sqrt{n}}$$

де $t=t\left(\gamma,n\right)$ – значення, яке береться з таблиці розподілу Стьюдента.

Знаходимо $t=t\left(0,99;16\right)=2,95$.

Підставляючи $S=2$, $\bar{x}=15,35$, $n=16$, $t=2,95$ в формулу для знаходження надійного інтервалу, отримаємо надійний інтервал:

$$15,35-\frac{2,95\cdot 2}{\sqrt{16}}<a<15,35+\frac{2,95\cdot 2}{\sqrt{16}}$$

$$13,88<a<16,83$$

Надійний інтервал для оцінки середнього квадратичного відхилення σ

Нехай треба оцінити середнє квадратичне відхилення σ нормально розподіленої генеральної сукупності N за виправленим середнім квадратичним відхиленням S. Для цього проведемо k спостережень і за результатами $x_1,x_2,...,x_k$ обчислимо $\bar{x}$, оцінку S^2 невідомої дисперсії σ^2 і оцінку S для σ. Задаючи надійність γ інтервальної оцінки, знайдемо таке число ε, щоб виконувалося співвідношення $P\left(S-\varepsilon<\sigma<S+\varepsilon\right)=\gamma$.

Перетворимо подвійну нерівність $S-\varepsilon<\sigma<S+\varepsilon$ до виду

$S(1-q)<\sigma<S(1+q)$, де $q=\dfrac{\varepsilon}{S}$

Отже, шуканий довірчий інтервал для оцінки середнього квадратичного відхилення нормального розподілу за виправленим середньоквадратичним відхиленням S має такий вигляд:

$s(1-q)<\sigma<s(1+q)$, якщо $q<1$;

$0<\sigma<s(1+q)$, якщо $q\geq 1$;

де $q=q(\gamma,n)$ знаходиться за таблицею додатку 3 за заданими γ і n.

Надійний інтервал для оцінки середнього квадратичного відхилення σ можна також отримати за допомогою розподілу χ^2 або розподілу Пірсона.

Ознайомимося попередньо з розподілом χ^2. Якщо у формулу вибіркової дисперсії замість нормально розподіленої величини x ввести нову випадкову величину $\omega=x-M(x)$, то значення S^2 не зміниться, а випадкова величина ω також підкорятиметься нормальному закону з $M(\omega)=0$ та дисперсією δ^2.

Отже, $S^2=\dfrac{1}{n}\sum_{i=1}^{n}\left(x_i-M(x)\right)^2=\dfrac{1}{n}\sum_{i=1}^{n}\omega_i^2$. Звідси,

$nS^2=\sum_{i=1}^{n}\omega_i^2$ /

Розділимо обидві частини на δ^2, тоді $\dfrac{nS^2}{\delta^2}=\sum_{i=1}^{n}\left(\dfrac{\omega_i}{\delta}\right)^2$,

$\dfrac{\omega_i}{\delta}=\dfrac{x_i-M(x)}{\delta}=t$.

Оскільки випадкова величина ω підпорядковується нормальному закону з параметрами $(0,\delta)$, то t також має нормальний закон

розподілення з параметрами $(0,1)$. Значення $t_1, t_2, ..., t_n$ незалежні між собою, отже, незалежні і їх квадрати.

Позначимо $\chi^2 = \frac{nS^2}{\delta^2} = \sum_{i=1}^{n} t_i^2$.

Отже, випадкова величина, що є сумою квадратів незалежних випадкових величин, кожна з яких підпорядковується нормальному закону розподілу з параметрами $(0,1)$, називається випадковою величиною з χ^2 - розподілом і $k = n - 1$ ступенями свободи. Число ступенів свободи дорівнює числу незалежних змінних мінус число зв'язків, що накладаються на ці змінні.

Диференціальна функція розподілу χ^2 має вигляд

$$f(\chi^2) = L_n \cdot \chi^{n-2} \cdot e^{\frac{\chi^2}{2}}$$

де L_n - коефіцієнт, що залежить від n. Як бачимо, розподіл χ^2 не залежить від $M(x)$ і δ^2, а залежить лише від об'єму вибірки.

Математичне сподівання розподілу χ^2 дорівнює числу ступенів волі $M(x) = k$. Можна також довести, що дисперсія $\delta_\chi^2 = 2k$.

Для функції розподілу χ^2 складені таблиці, за якими можна вирахувати ймовірність того, що випадкова величина, що підпорядковується закону χ^2 з відомим числом n, не перевищить фіксоване значення $\chi^2_{k,\alpha}$. Побудова надійного інтервалу дисперсії при заданій довірчій ймовірності $p = 1 - \alpha$ (α -рівень значимості) здійснюється за допомогою виразу:

$$P\left(\frac{nS^2}{\chi_2^2} < \delta^2 < \frac{nS^2}{\chi_1^2}\right) = 1 - \alpha.$$

Для середнього квадратичного відхилення надійний інтервал має вигляд:

$$P\left(\sqrt{\frac{nS^2}{\chi_2^2}}<\delta<\sqrt{\frac{nS^2}{\chi_1^2}}\right)=1-\alpha.$$

Приклад 7. Потрібно побудувати надійний інтервал з ймовірністю $p=0,96$ для дисперсії випадкової величини x, розподіленої нормально, якщо $S^2=10$, $n=20$.

Розв'язання: За таблицею χ^2-розподілу нам необхідно вибрати два таких значення, щоб площа, що знаходиться під кривою $f\left(\chi^2\right)$ в інтервалі $\left(\chi_1^2;\chi_2^2\right)$, дорівнювала $1-\alpha$; χ_1^2 і χ_2^2 зазвичай вибирають так, щоб $P\left(\chi^2<\chi_2^2\right)=P\left(\chi^2>\chi_1^2\right)=\frac{\alpha}{2}$

Рис. 11. Вибір точок χ_1^2 і χ_2^2 для знаходження надійного інтервалу для δ^2

У прикладі $\alpha=1-p=0,04$, $\frac{\alpha}{2}=0,02$. Знаходимо за таблицями значення χ_1^2 і χ_2^2 при $p_1=0,98$, $p_2=0,02$ та $k=n-1=20-1=19$:

$\chi_1^2=8,6$

$\chi_2^2=33,7$

Надійний інтервал для δ^2 запишеться таким чином:

$$\frac{20\cdot10}{33,7}<\delta^2\le\frac{20\cdot10}{8,6}$$

або $5,94<\delta^2\le23,6$

Надійний інтервал для середнього квадратичного відхилення:

$$\sqrt{5,94} < \delta \le \sqrt{23,6}$$
$$2,44 < \delta \le 4,82.$$

Приклад 8. Побудувати надійний інтервал для оцінки з надійністю $\gamma = 0,95$ невідомого середнього квадратичного відхилення σ нормально розподіленої генеральної сукупності *X*, якщо $S = 0,7$ і $n = 20$.

Розв'язання. Шуканий надійний інтервал має вигляд:

$s(1-q) < \sigma < s(1+q)$, якщо $q < 1$;

$0 < \sigma < s(1+q)$, якщо $q \ge 1$;

де $q = q(\gamma, n)$ знаходиться за таблицею додатку 3 за заданими γ і n.

При $\gamma = 0,95$ і $n = 20$ за таблицею знаходимо $q = 0,37$.

Підставляючи $q = 0,37$, $S = 0,7$, $n = 20$ в відповідну формулу, отримаємо надійний інтервал:

$$0,7(1-0,37) < \sigma < 0,7(1+0,37)$$

$$0,441 < \sigma < 0,959.$$

Приклад 9. З генеральної сукупності отримали вибірку

№ п/п	1	2	3	4	5	6	7	8	9	10
x_i	21,0	19,5	28,5	30,0	25,5	18,8	30,8	33,0	22,5	21,0
№ п/п	11	12	13	14	15	16	17	18	19	20
x_i	22,5	21,0	26,3	21,0	23,3	19,5	24,0	28,5	39,0	21,0

Оцінити з надійністю $\gamma = 0,95$ істинне значення нормально розподіленої ознаки генеральної сукупності за вибіркової середньою та

для оцінки невідомого середнього квадратичного відхилення σ за допомогою надійних інтервалів.

Розв'язання. Істинне значення ознаки ототожнюється з математичним сподіванням a. Для оцінки a при невідомому σ скористаємося формулою $\overline{x} - \frac{t_\gamma S}{\sqrt{n}} < a < \overline{x} + \frac{t_\gamma S}{\sqrt{n}}$. Об'єм вибірки $n = 20$. Далі обчислимо

Середнє вибіркове:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{496,7}{20} = 24,83$$

Знайдемо вибіркову дисперсію:

$$\sigma^2 = \frac{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2}{n} = \frac{539,87}{20} = 26,99$$

Знайдемо виправлену вибіркову дисперсію:

$$S^2 = \frac{n}{n-1}\sigma^2 = \frac{20}{20-1} \cdot 26,99 = 28,41$$

Знайдемо виправлене середнє квадратичне відхилення

$$S = \sqrt{S^2} = \sqrt{28,41} = 5,33$$

Надійний інтервал з надійністю $\gamma = 0,95$ (рівні значущості $\alpha = 1 - \gamma = 0,05$).

Для середнього вибіркового:

$$\overline{x} - \frac{tS}{\sqrt{n}} < a < \overline{x} + \frac{tS}{\sqrt{n}}$$

де $t = t\left(\gamma, n\right)$ – значення, яке береться з таблиці розподілу Стьюдента.

Знаходимо $t = t\left(0,95; 20\right) = 2,093$.

$$24,83-\frac{2,093\cdot 5,33}{\sqrt{20}}<a<24,83-\frac{2,093\cdot 5,33}{\sqrt{20}}$$

$$22,34<a<27,32$$

Для середнього квадратичного відхилення:

За таблицею χ^2 -розподілу нам необхідно вибрати два таких значення, щоб площа, що знаходиться під кривою $f\left(\chi^2\right)$ в інтервалі $\left(\chi_1^2;\chi_2^2\right)$, дорівнювала $1-\alpha$; χ_1^2 і χ_2^2 зазвичай вибирають так, щоб

$$P\left(\chi^2<\chi_2^2\right)=P\left(\chi^2>\chi_1^2\right)=\frac{\alpha}{2}$$

У прикладі $\alpha=1-p=0,05$, $\frac{\alpha}{2}=0,025$. Знаходимо за таблицями значення χ_1^2 і χ_2^2 при $p_1=0,975$, $p_2=0,025$ та $k=n-1=20-1=19$:

$$\chi_1^2=8,91$$

$$\chi_2^2=32,9$$

Надійний інтервал для середнього квадратичного відхилення запишеться таким чином:

$$\sqrt{\frac{20\cdot 28,41}{32,9}}<\delta\le\sqrt{\frac{20\cdot 28,41}{8,91}}$$

$$4,15<\delta\le 7,99.$$

2. Прості лінійні регресійні моделі

Прості лінійні регресійні моделі встановлюють лінійну залежність між двома змінними. При цьому одна із змінних вважається залежною змінною (Y) та розглядається як функція від незалежної змінної (X).

Регресійна модель називається лінійною, якщо вона лінійна за своїми параметрами.

У загальному вигляді проста вибіркова регресійна модель запишеться так:

$$Y = a_0 + a_1 X + u$$

де Y – вектор спостережень за залежною змінною;

X – вектор спостережень за незалежною змінною;

a_0, a_1 – невідомі параметри регресійної моделі;

u – вектор випадкових величин (помилок).

Рівняння парної регресії характеризує зв'язок між двома змінними, який виявляється як певна закономірність лише в середньому для всієї сукупності спостережених даних, а для кожного окремого спостереження може не виконуватись.

Побудова парної лінійної регресійної моделі

1. Вибір форми залежності

Нехай відомо n спостережень двох економічних показників X та Y: $X = (x_1; x_2; ...; x_n)$, $Y = (y_1; y_2; ...; y_n)$, де X є факторною (незалежною) змінною, а Y- результативною (залежною) змінною.

Вибір формули зв'язку змінних називається специфікацією моделі регресії. У випадку парної регресії вибір формули звичайно здійснюється за графічним зображенням реальних статистичних даних точками $(x_i; y_i)$,

$(i = 1, 2, ..., n)$ в декартовій системі координат. Такий графік називається кореляційним полем (діаграмою розсіювання).

Якщо точки кореляційного поля гуртуються навколо деякої прямої лінії, то залежність між змінними можна описати лінійним рівнянням $Y = a_0 + a_1 X + u$.

2. Оцінювання параметрів

Для параметризації таких рівнянь найчастіше застосовують метод найменших квадратів (1МНК).

Ідея методу найменших квадратів полягає в тому, щоб підібрати такі значення параметрів моделі, при яких сума квадратів відхилень між заданими y_i і розрахунковими $\hat{y}_i$ значеннями результативної змінної буде найменшою, тобто

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

Нехай рівняння регресії записано у вигляді $Y = a_0 + a_1 X + u$

Для парної лінійної моделі $Y = a_0 + a_1 X$

Розрахункові значення обчислюються за формулою $\hat{y}_i = a_0 + a_1 x_i$, де x_i - задані значення незалежної (факторної) змінної.

Значення x_i, y_i відомі і незмінні, тобто є константами, а параметри a_0, a_1 - невідомі і вважаються змінними. Отже, функція S залежить від двох змінних: $S = S(a_0; a_1)$.

Функція S набуває вигляду $S = \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i \right)^2 \to \min$. Для визначення екстремуму функції S обидві її частинні похідні прирівнюють до нуля:

$$\begin{cases} \dfrac{\partial S}{\partial a_0} = -2 \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i \right) = 0 \\ \dfrac{\partial S}{\partial a_1} = -2 \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i \right) x_i = 0 \end{cases}$$

і в результаті отримують систему нормальних рівнянь:

$$\begin{cases} a_0 n + a_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases}$$

Оцінювання параметрів моделі $Y = a_0 + a_1 X$ за методом найменших квадратів (МНК) зводиться до розв'язання системи нормальних рівнянь.

$$\begin{cases} a_0 + a_1 \bar{x} = \bar{y} \\ a_0 \bar{x} + a_1 \overline{x^2} = \overline{xy} \end{cases}$$

Розв'язки системи рівнянь можна знайти у вигляді:

$$\begin{cases} a_1 = \dfrac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \left(\bar{x} \right)^2} = r_{xy} \cdot \dfrac{S_y}{S_x} \\ a_0 = \bar{y} - a_1 \bar{x} \end{cases}$$

де $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, - середні арифметичні значення;

$S_x^2 = \overline{x^2} - \left(\bar{x}\right)^2$, $S_y^2 = \overline{y^2} - \left(\bar{y}\right)^2$ - вибіркові дисперсії.

3. Оцінка тісноти та значимості зв’язку між змінними моделі

Після вибору виду рівняння регресії та знаходження його параметрів розпочинають наступний етап – кореляційний аналіз, тобто дають оцінку тісноти та значимості зв'язку змінних у регресійній моделі.

Тісноту зв'язку між залежною змінною Y та незалежною змінною X оцінюють за допомогою статистичних характеристик: коефіцієнт детермінації, коефіцієнт кореляції. За допомогою цих коефіцієнтів перевіряється відповідність побудованої регресійної моделі (теоретичної) фактичним даним.

Після встановлення тісноти зв'язку між змінними моделі характеризують значимість зв'язку, яка в кореляційному аналізі частіше всього здійснюється за допомогою F-критерію Фішера.

Коефіцієнт детермінації

Коефіцієнт детермінації показує, якою мірою варіація залежної змінної (результативного показника) Y визначається варіацією незалежної змінної (вхідного показника) X. Тобто дається відповідь на запитання, чи справді зміна значення Y лінійно залежить саме від зміни значення X, а не відбувається під впливом різних випадкових факторів. Він використовується як при лінійному, так і при нелінійному зв'язку між змінними та розраховується за формулою:

$$R^2 = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \vec{y}\right)^2} = 1 - \frac{S_u^2}{S_y^2}$$

де y_i – теоретичні значення залежної змінної на підставі побудованої регресійної моделі; $\bar{y}$ – загальна середня фактичних даних залежної змінної; $\hat{y}_i$ – фактичні індивідуальні значення залежної змінної.

Коефіцієнт детермінації приймає значення від 0 (відсутній лінійний зв'язок між показниками) до 1 (відсутній кореляційний зв'язок між показниками).

Коефіцієнт кореляції (індекс кореляції)

Найпростішим критерієм, який дає кількісну оцінку зв’язку між двома показниками, є коефіцієнт кореляції (або індекс кореляції). Він розраховується за такою формулою:

$$R = \pm\sqrt{\frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \vec{y}\right)^2}}$$

Чим ближче коефіцієнт кореляції до одиниці, тим тісніше зв'язок між незалежною та залежною змінними.

Іноді для спрощення розрахунків тісноту кореляційного зв'язку характеризують коефіцієнтом кореляції, який розраховується за формулою:

$R = \sqrt{R^2}$

Значення R лежить у діапазоні від –1 до +1. При $R = 0$ змінні не можуть мати лінійного кореляційного зв'язку. Ступінь тісноти їх лінійної залежності зростає при наближенні R до ±1. Кореляційний зв'язок між показниками відсутній при $R = \pm 1$ Коли $R > 0$, то зв'язок між показниками прямий, якщо $R < 0$ – обернений.

В залежності від значення коефіцієнта кореляції зв’язок між змінними класифікується так:

Значення коефіцієнта кореляції	0	0,1–0,3	0,3–0,5	0,5–0,7	0,7–0,9	0,9–0,99	1
Висновок про силу кореляційного зв’язку	відсутній	слабкий	помірний	середній	високий	досить високий	близький до функціонального

Після оцінки точності моделі перевіряють статистичні гіпотези:

про значущість коефіцієнта детермінації (тобто перевіряють адекватність моделі);

про значущість кореляційного зв’язку;

про значущість окремих параметрів моделі.

Перевірка значущості коефіцієнта детермінації

Для перевірки статистичної значущості коефіцієнта детермінації R^2 висувається нульова гіпотеза $H_0: R^2 = 0$. Це означає, що досліджуване рівняння не пояснює зміну залежної змінної (Y) під впливом відповідних незалежних змінних. У такому разі всі коефіцієнти при

незалежних змінних мають дорівнювати нулю. При цьому нульову гіпотезу можна подати у вигляді:

$$H_0: a_0 = a_1 = ... = a_n = 0$$

Альтернативною до неї є H_α: значення хоча б одного параметра моделі відмінне від нуля $(a_i \neq 0)$, тобто хоча б один із факторів впливає на зміну залежної змінної.

Для перевірки цих гіпотез застосовують F-критерій Фішера з $m-1$ та $n-m$ ступенями свободи. За отриманими в моделі значеннями коефіцієнта детермінації R^2 обчислюють експериментальне значення F-статистики:

$$F_{\text{ðîç}} = \frac{R^2}{1-R^2} \cdot \frac{n-m}{m-1},$$

де n – кількість спостережень, m – кількість параметрів рівняння регресії, які оцінюються.

Отримане значення $F_{\text{ðîç}}$ порівнюють з табличним значенням розподілу Фішера при заданому рівні значущості α (як правило, $\alpha = 0,05$ або $\alpha = 0,01$). За таблицями F-розподілу Фішера знаходимо $F_{\text{êð}}$ – квантиль порядку $p = 1-\alpha$ розподілу Фішера з $m-1$ та $n-m$ степенями вільності. Якщо $F_{\text{ðîç}} > F_{\text{êð}}$ нульова гіпотеза відхиляється, тобто існує такий коефіцієнт у регресійному рівнянні, який суттєво відрізняється від нуля, а відповідний фактор виливає на досліджувану змінну. Відхилення нуль-

гіпотези свідчить про адекватність побудованої моделі. У протилежному випадку модель вважається неадекватною.

Перевірка значущості коефіцієнта кореляції

Коефіцієнт кореляції, як вибіркова характеристика, перевіряється на значущість за допомогою t-критерію Ст'юдента. Фактичне значення t-статистики обчислюється за формулою

$$t_{\text{îáñ çð}} = \frac{R\sqrt{n-m}}{\sqrt{1-R^2}}$$

і порівнюється з табличним значенням t-розподілу з $n-m$ ступенями свободи та при заданому рівні значущості $\frac{\alpha}{2}$ (такий рівень зумовлений тим, що критична область складається з двох проміжків).

Якщо абсолютна величина експериментального значення t-статистики перевищує табличне, тобто $\left|t_{\text{îáñ çð}}\right| > t_{\text{êð}}$, можна зробити висновок, що коефіцієнт кореляції достовірний (значущий), а зв'язок між залежною змінною та всіма незалежними факторами суттєвий.

Оцінка статистичної значущості параметрів моделі

Окрім загальних показників адекватності моделі існують також оцінки, що дають змогу встановити якість окремих частин рівняння, зокрема одного чи кількох коефіцієнтів регресії.

Як і в попередніх випадках, рішення відносно якості коефіцієнтів приймають на основі відповідних статистичних критеріїв.

Статистичну значущість кожного параметра моделі можна перевірити за допомогою t-критерію. При цьому нульова гіпотеза має вигляд

$$H_0 : a_0 = a_1 = ... = a_n = 0 .$$

альтернативна

$$H_\alpha : a_i \neq 0 .$$

Експериментальне значення t-статистики для кожного параметра моделі обчислюється за формулою

$$t_i = \frac{a_i}{S_{a_i}} \quad (i = 0,1),$$

де $S_{a_0} = S_u \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}}$, $S_{a_1} = S_u \sqrt{\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$

Експериментальне значення t_j-критерію порівнюється з табличним значенням t-розподілу з $n-m$ ступенями свободи та при заданому рівні значущості $\alpha/2$. Якщо значення t-статистики потрапляє до критичної області (за абсолютним значенням перевищує $t_{êð}$), то приймається альтернативна гіпотеза про значущість відповідного параметра. Інакше робиться висновок про статистичну незначущість параметра a_i j, а це означає, що відповідна незалежна змінна не впливає суттєво на зміну залежної змінної.

Розглянемо приклад.

Маємо вибірку даних за 12 років, які характеризують продуктивність праці ($Y_{факт}$) групи однорідних підприємств в залежності від простоїв основного обладнання (X). Побудувати парну лінійну регресійну модель виду $Y = a_0 + a_1 X$.

Побудуйте кореляційне поле і за розташуванням точок на графіку визначте форму залежності між X та Y.

Оцініть за 1МНК параметри рівняння лінійної регресії.

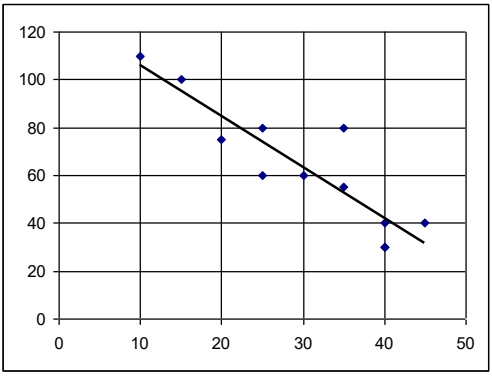
Оцініть вибірковий коефіцієнт кореляції r_{XY} .

Проінтерпретуйте результати.

Після оцінювання параметрів парних моделей провести їх економетричне дослідження/

Рік	1	2	3	4	5	6	7	8	9	10	11	12
$Y_{факт}$	10	20	15	25	30	35	40	35	25	40	45	40
X	110	75	100	80	60	55	40	80	60	30	40	30

Для визначення виду залежності побудуємо кореляційне поле. Реальні спостереження Y зобразимо точками у системі координат (X,Y). Візуально можна припустити, що між даними є лінійна залежність, тобто їх можна апроксимувати прямою лінією: $Y = a_0 + a_1 X$.



Для наочності розрахунків по МНК побудуємо таблицю.

i	x_i	y_i	x_i^2	$x_i y_i$	y_i^2
1	10	110	100	1100	12100
2	20	75	400	1500	5625
3	15	100	225	1500	10000
4	25	80	625	2000	6400
5	30	60	900	1800	3600
6	35	55	1225	1925	3025
7	40	40	1600	1600	1600
8	35	80	1225	2800	6400
9	25	60	625	1500	3600
10	40	30	1600	1200	900
11	45	40	2025	1800	1600

12	40	30	1600	1200	900
Сума	360	760	12150	19925	55750
Середнє	30	63,3	1012,5	1660,42	4645,83

Згідно розрахункових формул 1МНК, маємо

$$\begin{cases} a_1 = \dfrac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \left(\bar{x}\right)^2} = \dfrac{1660,42 - 30 \cdot 63,3}{1012,5 - 30^2} = -2,12 \\ a_0 = \bar{y} - a_1 \cdot \bar{x} = 63,3 - \left(-2,12\right) \cdot 30 = 126,9 \end{cases}$$

Отже, рівняння парної лінійної регресії має вигляд: $Y = 126,9 - 2,12 \cdot X$. Зобразимо дану пряму регресії на кореляційному полі. За цим рівнянням розрахуємо модельні значення результативного показника $\hat{y}_i$, а також залишки $u_i = y_i - \hat{y}_i$.

Для аналізу щільності лінійної залежності обчислимо коефіцієнт кореляції:

$$r_{xy} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{\overline{x^2} - \left(\bar{x}\right)^2} \sqrt{\overline{y^2} - \left(\bar{y}\right)^2}} = \frac{1660,42 - 30 \cdot 63,3}{\sqrt{1012,5 - 30^2} \sqrt{4645,83 - 63,3^2}} = -0,89$$

Значення коефіцієнта кореляції близьке до одиниці, що свідчить про тісну (обернену) лінійну залежність між змінними X та Y. Це також підтверджується розташуванням точок на кореляційному полі.

Коефіцієнт a_1 показує, на яку величину зміниться продуктивність праці, якщо кількість простоїв обладнання збільшується на одну одиницю. В нашому випадку значення $a_1 = -2{,}12$.

Після оцінювання параметрів парних моделей проведемо їх економетричне дослідження. Для цього виконаємо наступні дії:

Обчислимо модельні значення результативного показника $\hat{y}_i = a_0 + a_1 x_i$ та залишки моделі $u_i = y_i - \hat{y}_i$.

Обчислимо стандартну похибку рівняння $S_u = \sqrt{\frac{\sum u_i^2}{n}}$, незміщену дисперсію залишків $\sigma_u^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(u_i - \bar{u}\right)^2 \approx \frac{\sum_{i=1}^{n} u_i^2}{n-2}$ і відповідне середньоквадратичне відхилення залишків $\sigma_u = \sqrt{\frac{\sum u_i^2}{n-2}}$.

Обчислимо коефіцієнт детермінації $R^2 = 1 - \frac{S_u^2}{S_y^2}$, де $S_y^2 = \sigma_y^2$.

Перевіримо статистичні гіпотези:

про значущість коефіцієнта детермінації (тобто перевіримо адекватність моделі);

про значущість кореляційного зв’язку;

про значущість окремих параметрів моделі.

Побудуємо розрахункову таблицю:

i	x_i	y_i	$\hat{y}_i$	$u_i = y_i - \hat{y}_i$	u_i^2	$\left(y_i - \overline{y}\right)^2$	$\left(x_i - \overline{x}\right)^2$
1	10	110	105,7	4,3	18,49	2177,78	400
2	20	75	84,5	-9,5	90,25	136,11	100
3	15	100	95,1	4,9	24,01	1344,44	225
4	25	80	73,9	6,1	37,21	277,78	25
5	30	60	63,3	-3,3	10,89	11,11	0
6	35	55	52,7	2,3	5,29	69,44	25
7	40	40	42,1	-2,1	4,41	544,44	100
8	35	80	52,7	27,3	745,29	277,78	25
9	25	60	73,9	-13,9	193,21	11,11	25
10	40	30	42,1	-12,1	146,41	1111,11	100
11	45	40	31,5	8,5	72,25	544,44	225
12	40	30	42,1	-12,1	146,41	1111,11	100
Сума	360	760			1494,1	7616,67	1350
Середнє	30	63,3					

Стандартна похибка рівняння

$$S_u = \sqrt{\frac{\sum_{i=1}^{n} u_i^2}{n}} = \sqrt{\frac{1494,1}{12}} = 11,16$$

Незміщена дисперсія залишків

$$\sigma_u^2 = \frac{\sum_{i=1}^{n} u_i^2}{n-2} = \frac{1494,1}{12-2} = 149,41$$

Середньоквадратичне відхилення залишків

$$\sigma_u = \sqrt{\frac{\sum_{i=1}^{n} u_i^2}{n-2}} = \sqrt{\frac{1494,1}{12-2}} = 12,22 .$$

Коефіцієнт детермінації:

$$S_y^2 = \sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \frac{1}{12}\cdot 7616,67 = 634,72$$

$$R^2 = 1 - \frac{S_u^2}{S_y^2} = 1 - \frac{124,51}{634,72} = 0,80 .$$

Значення коефіцієнта детермінації близьке до одиниці, що свідчить про тісну лінійну залежність між змінними X та Y.

Перевіримо статистичні гіпотези:

1) про значущість коефіцієнта детермінації:

Експериментальне значення F-критерію обчислимо за формулою:

$$F_{\text{ðîçð}} = \frac{R^2}{1-R^2}\left(n-2\right) = \frac{0,80\left(12-2\right)}{1-0,80} = 40,0$$

і порівняємо з табличним значенням $F_{\text{êð}} = F\left(\alpha; k_1; k_2\right)$, де $\alpha = 0,05$ - рівень значущості, $k_1 = m-1 = 2-1 = 1$, $k_2 = n-m = 12-2 = 10$ - степені свободи. Звідси, $F_{\text{êð}} = F\left(0,05;1;10\right) = 4,96$

Маємо $F_{\text{îá ç÷}} > F_{\text{êð}}$, тобто модель пояснює зміну результату впливом заданого фактора. Отже, коефіцієнт детермінації значущий.

2) Оцінимо значущість коефіцієнта кореляції.

Обчислюємо $R = \sqrt{R^2}$ - коефіцієнт кореляції (характеризує тісноту лінійного зв'язку незалежної змінної X із залежною змінною Y.

$$R = \sqrt{R^2} = \sqrt{0,80} = 0,89$$

Знайдемо t-статистику за формулою $t_{\text{îá ç÷}} = \dfrac{R\sqrt{n-2}}{\sqrt{1-R^2}}$

$$t_{\text{îá ç÷}} = \frac{0,89 \cdot \sqrt{12-2}}{\sqrt{1-0,80}} = 6,29$$

Порівняємо t-статистику з

$$t_{\text{êð}}\left(\frac{\alpha}{2}, n-m\right) = t_{\text{êð}}\left(\frac{0,05}{2}; 12-2\right) = t_{\text{êð}}\left(0,025; 10\right) = 2,228$$

Отже, $t_{розр} > t_{кр}$, тобто коефіцієнт кореляції між залежною Y і незалежною X змінними значущий.

3) Оцінимо значущість окремих параметрів моделі

Обчислюємо t-статистику за формулою $t = \dfrac{a_i}{S_{a_i}} \quad (i = 0,1)$

Для того, щоб здійснити таке оцінювання, потрібно визначити стандартні похибки параметрів. Для парної регресії ці значення відповідно дорівнюють.

$$S_{a_0} = S_u \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}} = 11{,}16\sqrt{\frac{12150}{12 \cdot 1350}} = 9{,}66,$$

$$S_{a_1} = S_u \sqrt{\frac{1}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}} = 11{,}16\sqrt{\frac{1}{1350}} = 0{,}30$$

Звідси, маємо

$$t_0 = \frac{a_0}{S_{a_0}} = \frac{126{,}6}{9{,}66} = 13{,}106$$

$$t_1 = \frac{a_1}{S_{a_1}} = \frac{-2{,}12}{0{,}30} = -7{,}067.$$

За таблицями t-статистики знаходимо критичне значення $t_{кр}\left(\frac{\alpha}{2}, n-m\right)$ і порівнюємо її з обчисленою t-статистикою:

$$t_{кр}\left(\frac{\alpha}{2}, n-m\right) = t_{кр}\left(\frac{0{,}05}{2}; 12-2\right) = t_{кр}\left(0{,}025; 10\right) = 2{,}228$$

Звідси, $|t_0| > t_{êð}$, $|t_1| > t_{êð}$.

Отже, приймаємо, що коефіцієнти a_0 та a_1 значущі.

3. Методи побудови нелінійних економетричнх моделей

Залежність багатьох економічних показників не лінійна. Метод найменших квадратів не призначений для оцінки параметрів нелінійних регресій, крім поліноміальної.

Найбільш загальною формою нелінійних регресійних моделей є моделі, нелінійні як за параметрами, так і за вхідними змінними. Відразу ж відзначимо, що загальних методів аналізу таких моделей не існує. Більш того, не існує навіть скільки-небудь систематичної класифікації моделей такого типу.

Однак цілий клас нелінійних залежностей можна привести до залежностей лінійним перетворенням незалежної і / або залежною змінною.

Таке перетворення називається лінеаризацією.

Розглянемо тут деякі приклади, які при всій їх різноманітності зводяться до лінеаризації нелінійної регресійної моделі.

Завдання побудови нелінійної моделі регресії полягає в наступному

Задана нелінійна специфікація моделі $y = f(x, a, b, u)$,

де y - залежна, пояснювальна змінна; x – незалежна змінна; a, b - параметри моделі, для яких повинні бути отримані оцінки; u - адитивний або мультиплікативний випадковий фактор.

Потрібно

1. Перетворити вихідні дані $x \to x^*$, $y \to y^*$ так, щоб специфікація модифікованої регресії була лінійною:

$$y^* = a^* + b^* x^*$$

2. Методом найменших квадратів отримати оцінки параметрів a^*, b^*.

3. За оцінками a^*, b^* обчислити шукані оцінки параметрів a, b вихідної регресії.

Види рівнянь регресії:

Функції		Аналітичний вираз
1.	Лінійна	$Y = a_0 + a_1 X$
2.	Параболічна	$Y = a_0 + a_1 X^2$
3.	Гіперболічна	$Y = a_0 + a_1 \frac{1}{X}$
4.	Степенева	$Y = a_0 x^{a_1}$
5.	Показникова	$Y = ab^x$
6.	Модифікована експоненціальна	$Y = ae^{bx}$
7.	Показниково-степенева	$Y = ax^b c^x$
8.	Екологічна	$Y = ae^{-b^2(x-c)^2}$
9.	Логістична	$Y = \frac{a}{1+be^{cx}}$
10.	Гомперця	$Y = e^{ab^x + c}$

11.	Ірраціональна	$Y = \sqrt{a + bx + cx^2}$
12.	Обернена до квадратичної	$Y = \dfrac{1}{a + bx + cx^2}$
13.	Дрібно-раціональна	$Y = \dfrac{x}{a + bx + cx^2}$
14.	Функція Джонсона	$\log Y = -\dfrac{a}{b + x} + c$
15.	Функції Тронквіста	$Y = \dfrac{ax}{b + x}$ $Y = \dfrac{a(x - b)}{x + c}$ $Y = \dfrac{ax(x - b)}{x + c}$

Нелінійні зв'язки, як правило, певними перетвореннями (заміною змінних чи логарифмуванням) зводять до лінійного вигляду або апроксимують (наближують) лінійними функціями.

Використання (1МНК) для оцінки теоретичних параметрів моделі a_i $(i = 0,1,2)$ найпростіших рівнянь парної регресії приводить до таких систем нормальних рівнянь:

а) лінійна залежність $Y = a_0 + a_1 X$

$$\begin{cases} a_0 n + a_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases}$$

б) параболічна залежність $Y = a_0 + a_1 X^2$

$$\begin{cases} a_0 n + a_1 \sum_{i=1}^{n} x_{1i} = \sum_{i=1}^{n} y_i \\ a_0 \sum_{i=1}^{n} x_{1i} + a_1 \sum_{i=1}^{n} x_{1i}^2 = \sum_{i=1}^{n} x_{1i} y_i \end{cases}$$

де $x_1 = x^2$

в) гіперболічна залежність $Y = a_0 + a_1 \frac{1}{X}$

$$\begin{cases} a_0 n + a_1 \sum_{i=1}^{n} x_{1i} = \sum_{i=1}^{n} y_i \\ a_0 \sum_{i=1}^{n} x_{1i} + a_1 \sum_{i=1}^{n} x_{1i}^2 = \sum_{i=1}^{n} x_{1i} y_i \end{cases}$$

де $x_1 = \frac{1}{x}$.

г) степенева залежність $Y = a_0 x^{a_1}$

$$\begin{cases} b_0 n + a_1 \sum_{i=1}^{n} x_{1i} = \sum_{i=1}^{n} y_{1i} \\ b_0 \sum_{i=1}^{n} x_{1i} + a_1 \sum_{i=1}^{n} x_{1i}^2 = \sum_{i=1}^{n} x_{1i} y_{1i} \end{cases}$$

де $b_0 = \ln a_0$, $x_1 = \ln x$, $y_1 = \ln y$

4. Методи побудови множинних економетричнх моделей.

Множинна регресія являє собою узагальнення простої регресійної моделі для випадку, коли змінна Y залежить не від одного, а від кількох факторів (від n факторів).

Специфікація моделі множинної регресії:

$$Y = a_0 + a_1 X_1 + a_2 X_2 + ... + a_m X_m$$

У якості незалежних змінних можуть застосовуватись різні техніко-економічні показники роботи підприємства.

У рівнянні Y – залежна змінна, якою може бути будь-який з результуючих показників діяльності підприємства.

До вигляду без особливих зусиль можна звести більшість рівнянь, що практично застосовуються в якості виробничих функцій.

У розгорнутому вигляді модель виглядатимете так:

$$\begin{cases} \sum Y = na_0 + a_1 \sum X_1 + a_2 \sum X_2 + ... + a_m \sum X_m \\ \sum X_1 Y = a_0 \sum X_1 + a_1 \sum X_1 X_1 + a_2 \sum X_1 X_2 + ... + a_m \sum X_1 X_m \\ \sum X_2 Y = a_0 \sum X_2 + a_1 \sum X_1 X_2 + a_2 \sum X_2 X_2 + ... + a_m \sum X_2 X_m \\ ..... \\ \sum X_m Y = a_0 \sum X_m + a_1 \sum X_m X_1 + a_2 \sum X_m X_2 + ... + a_m \sum X_m X_m \end{cases}$$

Наведена система дозволяє порівняно легко скласти нормальні рівняння для розрахунку параметрів будь-яких виробничих функцій, що зводяться до вигляду.

Матрична форма економетричної моделі.

У загальному матричному вигляді економетрична модель для фактичних даних записується так:

$$Y = AX + u,$$

де A – матриця параметрів моделі розміром $m \times n$ (m – кількість незалежних змінних, n – число спостережень);

Y – матриця значень залежної змінної;

X – матриця незалежних змінних;

u – матриця випадкової складової.

Випадкові складові u називають ще помилками або залишками. Вони є наслідками помилок спостережень, містять у собі вплив усіх випадкових факторів, а також факторів, які не входять у модель.

Теоретичні (розрахункові) значення залежних змінних Y для моделі будуть представлені у вигляді: $Y = AX$

Сукупність виразів для фактичних і теоретичних значень залежних змінних визначає економетричну модель загального виду:

$$\left(X^T X\right) A = X^T Y$$

Це система нормальних рівнянь.

Розв'язок системи нормальних рівнянь в матричному записі буде мати вигляд:

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

де A – вектор параметрів лінійної моделі, X^T – матриця транспонована до матриці X.

Оцінка тісноти та значимості зв'язку між змінними у множинній регресії.

Тіснота зв'язку загального впливу всіх незалежних змінних X на залежну змінну Y визначається коефіцієнтами детермінації і множинної кореляції, парними коефіцієнтами кореляції, а також частинними коефіцієнтами кореляції.

Вигляд коефіцієнта детермінації у множинній регресії ідентичний коефіцієнту детермінації простої регресії.

Таким чином, коефіцієнт детермінації R^2 дорівнює відношенню суми квадратів відхилень розрахункових значень Y від його середнього до суми квадратів відхилень фактичних значень Y від його середнього значення:

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

Оскільки введення нових незалежних змінних $X_i \ \ (i = 1, 2, ..., m)$ у множинну регресію, а значить і ступенів вільності моделі, приводить до зменшення коефіцієнта детермінації, то його розрахунок повинен бути

відкоригований з урахуванням ступенів вільності, дисперсії залишок та загальної дисперсії.

Скоригований коефіцієнт детермінації розраховується за формулою:

$$R^2_{\text{ñêîð}} = 1 - \frac{\sigma_u^2}{\sigma_y^2}$$

де σ_u^2 — дисперсія залишків моделі, $\sigma_u^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n - m - 1}$,

σ_y^2 – загальна дисперсія моделі, $\sigma_y^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n - 1}$

Множинний коефіцієнт кореляції R розраховується за формулою $R = \sqrt{R^2}$

Для нього характерна така сама зміна числового значення, як і для коефіцієнта детермінації.

Парні коефіцієнти кореляції

Інформацію про парну залежність може дати симетрична матриця коефіцієнтів парної регресії між змінними:

$$r_{X_i X_j} = X^{*T} X^*$$

де X^* – матриця нормалізованих змінних; X^{*T} – матриця, транспонована до X*.

Матриці нормалізованих змінних:

$$Y^* = \begin{pmatrix} y_1^* \\ y_2^* \\ \dots \\ y_n^* \end{pmatrix}, \qquad X^* = \begin{pmatrix} x_{11}^* & x_{12}^* & \dots & x_{1m}^* \\ x_{21}^* & x_{22}^* & \dots & x_{2m}^* \\ . & . & \dots & . \\ x_{n1}^* & x_{n2}^* & \dots & x_{nm}^* \end{pmatrix}.$$

Елементи нормалізованих векторів розраховують за формулами:

$$y_i^* = \frac{y_i - \bar{y}}{\sqrt{n\sigma_y^2}}; \qquad x_{ik}^* = \frac{x_{ik} - \overline{x_k}}{\sqrt{n\sigma_{x_k}^2}}, \qquad x_{ij}^* = \frac{x_{ij} - \overline{x_j}}{\sqrt{n\sigma_{x_j}^2}}$$

де n – кількість спостережень;

y_i^*, x_{ik}^*, x_{ij}^* – нормалізовані (стандартизовані) змінні.

$\bar{y}$, $\overline{x_k}$, $\overline{x_j}$ – середні значення залежної та незалежних змінних;

σ_y, σ_{x_k}, σ_{x_j} – середньоквадратичні відхилення змінних.

Дисперсії кожної змінної мають такі значення:

$$\sigma_y^2 = \frac{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}{n}, \qquad \sigma_{x_k}^2 = \frac{\sum_{i=1}^{n}\left(x_{ik} - \overline{x_k}\right)^2}{n}, \qquad \sigma_{x_j}^2 = \frac{\sum_{i=1}^{n}\left(x_{ij} - \overline{x_j}\right)^2}{n}$$

Кореляційна матриця (матриця парних коефіцієнтів кореляції) має наступний вигляд:

$$r_{X_i X_j} = \begin{pmatrix} r_{YY} & r_{YX_1} & r_{YX_2} & \dots & r_{YX_m} \\ r_{X_1 Y} & r_{X_1 X_1} & r_{X_1 X_2} & \dots & r_{X_1 X_m} \\ r_{X_2 Y} & r_{X_2 X_1} & r_{X_2 X_2} & \dots & r_{X_2 X_m} \\ . & . & . & \dots & . \\ r_{X_m Y} & r_{X_m X_1} & r_{X_m X_2} & \dots & r_{X_m X_m} \end{pmatrix}.$$

Парні коефіцієнти кореляції дають оцінку тісноти зв'язку між парами змінних:

незалежною X_k та залежною Y – $r_{x_k y}$;

незалежною X_j та залежною у – $r_{x_j y}$;

незалежними змінними X_k та X_j – $r_{x_k x_j}$.

Кореляційна матриця коефіцієнтів парної регресії симетрична і має розмірність $m \times m$:

Діагональні елементи характеризують тісноту зв'язку однойменних змінних, тому вони дорівнюють одиниці.

Користуючись цими коефіцієнтами, можна зробити висновок про наявність кореляційного зв'язку між змінними.

Частинні коефіцієнти кореляції

Частинні коефіцієнти кореляції, як і парні, характеризують тісноту зв'язку між двома змінними, але за умови, що решта змінних сталі.

Розрахунок частинних коефіцієнтів кореляції базується на обернений матриці до матриці $r_{X_i X_j}$ (матриця С):

$$r_{kj} = \frac{-c_{kj}}{\sqrt{c_{kk} \cdot c_{jj}}},$$

де c_{kj} – елемент матриці С, що міститься в k-му рядку і j-му стовпці; c_{kk}, c_{jj} – діагональні елементи матриці С.

Частинні коефіцієнти кореляції характеризують рівень тісноти зв'язку між двома змінними за умови, що решта змінних на цей зв'язок не впливає.,

Значимість зв'язку

Значимість зв'язку між залежною Y та незалежними змінними X у випадку множинної регресії можна перевірити за допомогою F-критерію Фішера з урахуванням ступенів вільності:

Перевірка значимості коефіцієнта детермінації, коефіцієнта кореляції та оцінок параметрів моделі множинної регресії.

Перевірка значимості коефіцієнта детермінації

Коефіцієнт детермінації R^2 перевіряється на значущість за допомогою F-критерію Фішера.

Висувається нульова гіпотеза H_0: $R^2 = 0$. Це означає, що досліджуване рівняння не пояснює зміну залежної змінної (Y) під впливом відповідних незалежних факторів.

У такому разі всі коефіцієнти при незалежних змінних мають дорівнювати нулю: H_0: $a_0 = a_1 = ... = a_n = 0$.

Альтернативною до неї є $H_\alpha: (a_i \neq 0)$, значення хоча б одного параметра моделі відмінне від нуля (тобто хоча б один із факторів впливає на змінювання залежної змінної).

За отриманими в моделі значеннями коефіцієнта детермінації R^2 обчислюють експериментальне значення F-статистики:

$$F_{\text{îá ç÷}} = \frac{R^2}{1-R^2} \cdot \frac{n-m-1}{m}$$

яке порівнюють з табличним значенням розподілу Фішера при заданому рівні значущості α (як правило, α = 0,05 або α = 0,01). Якщо $F_{\text{îá ç÷}} > F_{\text{êð}}$, то нульова гіпотеза відхиляється.

Відхилення нуль-гіпотези свідчить про значимість коефіцієнта детермінації.

Перевірка значимості коефіцієнта кореляції

У кореляційному аналізі для характеристики відхилень коефіцієнта кореляції, як вибіркової величини, від свого "істотного" значення вимагається перевірка його значимості за t-критерієм Ст'юдента:

$$t = \frac{R\sqrt{n-m\text{-}1}}{\sqrt{1-R^2}}$$

де R^2 – коефіцієнт детермінації моделі; R – коефіцієнт кореляції;

(n–m-1) – число ступенів вільності.

Розраховане за формулою фактичне значення t-критерію порівнюється з табличним значенням $t_{êð}$. Останнє обирається за статистичними таблицями на підставі прийнятого рівня значимості α та розрахованого числа ступіней вільності (n–m-1). Якщо $|t| > t_{êð}$, то можна зробити висновок про значимість коефіцієнта кореляції між змінними.

Перевірка значимості оцінок параметрів моделі множинної регресії

У кореляційному аналізі може перевірятись також значимість оцінок параметрів моделі А із знаходженням їх довірчих інтервалів.

Припустивши, що залишки u розподілені за нормальним законом, приймається, що параметри моделі А також задовольняють нормальному розподілу. Тоді перевірку гіпотези про значимість оцінок параметрів моделі проводять згідно з t-критерієм Ст'юдента:

$$t_{a_j} = \frac{|a_j|}{\sqrt{\sigma_u^2 c_{jj}}}$$

де a_j – індивідуальні параметри матриці A; σ_u^2 – дисперсія залишків;

c_{jj} – діагональні елементи матриці $\left(X^T X\right)^{-1}$;

$S_{a_j} = \sqrt{\sigma_u^2 c_{jj}}$ – стандартна помилка оцінки параметра моделі.

Обчислене значення t-критерію порівнюється з табличним значенням $t_{êð}$ при вибраному рівні значимості α і (n–m-1) ступенях

вільності. Якщо $|t_{a_i}| > t_{êð}$, то оцінка значимості відповідного параметру моделі є достовірною.

На підставі t-критерію і стандартної помилки встановлюються довірчі інтервали для параметра a_j:

$$a_j - t_\alpha \sqrt{\sigma_u^2 c_{jj}} \le a_j \le a_j + t_\alpha \sqrt{\sigma_u^2 c_{jj}}$$

Приклад дослідження багатофакторної моделі

За статистичними даними

Y	X1	X2	X3
10,4	12,7	30,9	60,9
11,5	11,8	31,3	65,1
12,7	10,3	32,8	70,4
13,8	9,1	33,9	74,3
14,3	9,0	34,5	78,7
14,9	8,5	36,0	84,9
15,1	7,4	40,1	85,5
16,3	7,0	45,8	86,6
17,9	6,3	46,7	87,0
18,4	5,1	45,7	87,0
20,1	4,2	48,8	89,1
24,3	4,0	49,3	89,4

25,5	2,1	50,1	89,7
26,9	1,5	52,4	89,7
26,3	1,5	50,0	97,2

Побудувати множинну лінійну регресійну модель залежності Y (продуктивності праці, тис. грн./чол.) від X1 – втрат робочого часу, тис. год./рік, X2 – коефіцієнту використання потужностей, %, X3 – рівня механізації і автоматизації виробництва, %.

Побудувати рівняння регресії.

Провести оцінку точності та імовірності моделі: розрахувати коефіцієнт кореляції; розрахувати середньоквадратичну та відносну похибки; розрахувати критерій Фішера; розрахувати коефіцієнт еластичності.

Зробити загальний економічний аналіз моделі

Порядок виконання завдання

1. Коефіцієнти рівняння регресії $\hat{y}_i = a_0 + a_1 x_{1i} + a_2 x_{2i} + a_3 x_{3i}$ знаходимо за формулою $A = \left(X^T X \right)^{-1} \left(X^T Y \right)$. В даній задачі матриця X та Y мають вигляд:

1	12,7	30,9	60,9
1	11,8	31,3	65,1
1	10,3	32,8	70,4
1	9,1	33,9	74,3
1	9,0	34,5	78,7

10,4
11,5
12,7
13,8
14,3

$X =$

1	8,5	36,0	84,9
1	7,4	40,1	85,5
1	7,0	45,8	86,6
1	6,3	46,7	87,0
1	5,1	45,7	87,0
1	4,2	48,8	89,1
1	4,0	49,3	89,4
1	2,1	50,1	89,7
1	1,5	52,4	89,7
1	1,5	50,0	97,2

$Y =$

14,9
15,1
16,3
17,9
18,4
20,1
24,3
25,5
26,9
26,3

Знайдемо транспортну матрицю X^T

1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
12,7	11,8	10,3	9,1	9	8,5	7,4	7	6,3	5,1	4,2	4	2,1	1,5	1,5
30,9	31,3	32,8	33,9	34,5	36	40,1	45,8	46,7	45,7	48,8	49,3	50,1	52,4	50
60,9	65,1	70,4	74,3	78,7	84,9	85,5	86,6	87	87	89,1	89,4	89,7	89,7	97,2

Знайдемо добуток матриць $X^T X$

15,00	100,50	628,30	1235,50
100,50	854,69	3830,19	7804,05
628,30	3830,19	27188,53	52762,26
1235,50	7804,05	52762,26	103244,17

Знайдемо обернену матрицю – матрицю похибок $\left(X^T X\right)^{-1}$

71,55	-2,30	-0,63	-0,36
-2,30	0,08	0,02	0,01
-0,63	0,02	0,01	0,00
-0,36	0,01	0,00	0,00

Знайдемо добуток матриць $X^T Y$

268,40
1526,94
11805,28
22772,55

Знайдемо невідомі параметри регресії $A = \left(X^T X\right)^{-1}\left(X^T Y\right)$

46,68
-2,00
-0,02
-0,18

Отже, наша регресійна модель має вигляд:

$$Y = 46,68 - 2X_1 - 0,02X_2 - 0,18X_3$$

Далі знаходяться відповідні розрахункові значення $\hat{Y} = X \cdot A$.

Y	$\hat{Y}$	$Y - \hat{Y}$	$Y - \overline{Y}$	$\hat{Y} - \overline{Y}$
10,4	9,92	0,48	-7,49	-7,97
11,5	10,97	0,53	-6,39	-6,93

12,7	12,99	-0,29	-5,19	-4,90
13,8	14,67	-0,87	-4,09	-3,22
14,3	14,08	0,22	-3,59	-3,82
14,9	13,94	0,96	-2,99	-3,95
15,1	15,96	-0,86	-2,79	-1,93
16,3	16,47	-0,17	-1,59	-1,42
17,9	17,79	0,11	0,01	-0,11
18,4	20,20	-1,80	0,51	2,31
20,1	21,57	-1,47	2,21	3,68
24,3	21,91	2,39	6,41	4,01
25,5	25,64	-0,14	7,61	7,74
26,9	26,80	0,10	9,01	8,90
26,3	25,50	0,80	8,41	7,60
Сума квадратів		14,91	428,79	413,88

Коефіцієнт детермінації моделі знайдемо за формулою

$$R^2 = \frac{\sum_{i=1}^{n}\left(\hat{Y} - \overline{Y}\right)^2}{\sum_{i=1}^{n}\left(Y - \overline{Y}\right)^2} = \frac{413,88}{428,79} = 0,97$$

або

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(Y - \hat{Y}\right)^2}{\sum_{i=1}^{n}\left(Y - \overline{Y}\right)^2} = 1 - \frac{14,91}{428,79} = 0,97$$

Так, як коефіцієнт детермінації близький до 1, то між змінними Y, X_1, X_2, X_3 існує тісний кореляційний зв'язок.

Варіація змінної Y на 97 % визначається варіацією змінних X_1, X_2, X_3 і лише на 3% визначається впливом випадкових факторів.

Множинний коефіцієнт кореляції $R = \sqrt{R^2} = \sqrt{0,97} = 0,985$,

Так як коефіцієнт кореляції близький до 1, то між змінною Y та змінними X_1, X_2, X_3 існує тісний кореляційний зв'язок.

Парні коефіцієнти кореляції розраховуються за формулою матриці коефіцієнтів парної регресії між змінними:

Визначимо елементи нормалізованих векторів:

$$y_i^* = \frac{y_i - \overline{y}}{\sqrt{n\sigma_y^2}}, \qquad x_{1i}^* = \frac{x_{1i} - \overline{x_1}}{\sqrt{n\sigma_{x_1}^2}}, \qquad x_{2i}^* = \frac{x_{2i} - \overline{x_2}}{\sqrt{n\sigma_{x_2}^2}}, \qquad x_{3i}^* = \frac{x_{3i} - \overline{x_3}}{\sqrt{n\sigma_{x_3}^2}}$$

Дисперсії змінних мають такі значення:

$$\sigma_y^2 = \frac{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)}{n} = \frac{428,79}{15} = 28,59,$$

$$\sigma_{x_1}^2 = \frac{\sum_{i=1}^{n}\left(x_{1i} - \overline{x_1}\right)}{n} = \frac{181,3}{15} = 12,09,$$

$$\sigma^2_{x_2} = \frac{\sum_{i=1}^{n}\left(x_{2i} - \overline{x_2}\right)}{n} = \frac{871,1}{15} = 58,08$$

$$\sigma^2_{x_3} = \frac{\sum_{i=1}^{n}\left(x_{3i} - \overline{x_3}\right)}{n} = \frac{1480,2}{15} = 98,68$$

Тоді знаменники для нормалізації кожної змінної будуть такими:

$$\sqrt{n\sigma^2_y} = \sqrt{15 \cdot 28,59} = 20,7$$

$$\sqrt{n\sigma^2_{x_1}} = \sqrt{15 \cdot 12,09} = 13,5$$

$$\sqrt{n\sigma^2_{x_2}} = \sqrt{15 \cdot 58,08} = 29,5$$

$$\sqrt{n\sigma^2_{x_3}} = \sqrt{15 \cdot 98,68} = 38,5$$

Для зручності обчислень побудуємо розрахункову таблицю:

Y	X_1	X_2	X_3	$Y-\overline{Y}$	$X_1-\overline{X_1}$	$X_2-\overline{X_2}$	$X_3-\overline{X_3}$	Y^*	X_1^*	X_2^*	X_3^*
10,4	12,7	30,9	60,9	-7,5	6,0	-11,0	-21,5	-0,36	0,45	-0,37	-0,56
11,5	11,8	31,3	65,1	-6,4	5,1	-10,6	-17,3	-0,31	0,38	-0,36	-0,45
12,7	10,3	32,8	70,4	-5,2	3,6	-9,1	-12,0	-0,25	0,27	-0,31	-0,31
13,8	9,1	33,9	74,3	-4,1	2,4	-8,0	-8,1	-0,20	0,18	-0,27	-0,21
14,3	9,0	34,5	78,7	-3,6	2,3	-7,4	-3,7	-0,17	0,17	-0,25	-0,10
14,9	8,5	36,0	84,9	-3,0	1,8	-5,9	2,5	-0,14	0,13	-0,20	0,07
15,1	7,4	40,1	85,5	-2,8	0,7	-1,8	3,1	-0,13	0,05	-0,06	0,08
16,3	7,0	45,8	86,6	-1,6	0,3	3,9	4,2	-0,08	0,02	0,13	0,11
17,9	6,3	46,7	87,0	0,0	-0,4	4,8	4,6	0,00	-0,03	0,16	0,12
18,4	5,1	45,7	87,0	0,5	-1,6	3,8	4,6	0,02	-0,12	0,13	0,12
20,1	4,2	48,8	89,1	2,2	-2,5	6,9	6,7	0,11	-0,19	0,23	0,18
24,3	4,0	49,3	89,4	6,4	-2,7	7,4	7,0	0,31	-0,20	0,25	0,18
25,5	2,1	50,1	89,7	7,6	-4,6	8,2	7,3	0,37	-0,34	0,28	0,19

26,9	1,5	52,4	89,7	9,0	-5,2	10,5	7,3	0,43	-0,39	0,36	0,19
26,3	1,5	50,0	97,2	8,4	-5,2	8,1	14,8	0,41	-0,39	0,27	0,39
17,9	6,7	41,9	82,4	428,8	181,3	871,1	1480,2				

Матриця нормалізованих змінних:

$X^* =$

-0,36	0,45	-0,37	-0,56
-0,31	0,38	-0,36	-0,45
-0,25	0,27	-0,31	-0,31
-0,20	0,18	-0,27	-0,21
-0,17	0,17	-0,25	-0,10
-0,14	0,13	-0,20	0,07
-0,13	0,05	-0,06	0,08
-0,08	0,02	0,13	0,11
0,00	-0,03	0,16	0,12
0,02	-0,12	0,13	0,12
0,11	-0,19	0,23	0,18
0,31	-0,20	0,25	0,18
0,37	-0,34	0,28	0,19
0,43	-0,39	0,36	0,19
0,41	-0,39	0,27	0,39

Матриця, транспонована до X^*:

-0,36	-0,31	-0,25	-0,20	-0,17	-0,14	-0,13	-0,08	0,00	0,02	0,11	0,31	0,37	0,43	0,41
0,45	0,38	0,27	0,18	0,17	0,13	0,05	0,02	-0,03	-0,12	-0,19	-0,20	-0,34	-0,39	-0,39
-0,37	-0,36	-0,31	-0,27	-0,25	-0,20	-0,06	0,13	0,16	0,13	0,23	0,25	0,28	0,36	0,27
-0,56	-0,45	-0,31	-0,21	-0,10	0,07	0,08	0,11	0,12	0,12	0,18	0,18	0,19	0,19	0,39

Записуємо шукану кореляційну матрицю $r_{X_i X_j}$:

1	-0,97	0,92	0,84
-0,97	1	-0,95	-0,91
0,92	-0,95	1	0,89
0,84	-0,91	0,89	1

Кожний елемент цієї матриці характеризує тісноту зв’язку однієї змінної з іншою.

Оскільки діагональні елементи характеризують тісноту зв’язку кожної змінної з цією самою зміною, то вони дорівнюють одиниці. Решта елементів матриці $r_{X_i X_j}$ визначають зв'язок між парами змінних.: Користуючись цими коефіцієнтами, можна зробити висновок:

$r_{x_1 y}$	-0,97	між змінними Y та X_1 існує сильний обернений кореляційний зв’язок
$r_{x_2 y}$	0,92	між змінними Y та X_2 існує сильний прямий кореляційний зв’язок
$r_{x_3 y}$	0,84	між змінними Y та X_3 існує сильний прямий кореляційний зв’язок
$r_{x_1 x_2}$	-0,95	між змінними X_1 та X_2 існує сильний обернений кореляційний зв’язок
$r_{x_1 x_3}$	-0,91	між змінними X_1 та X_3 існує сильний обернений кореляційний зв’язок
$r_{x_2 x_3}$	0,89	між змінними X_2 та X_3 існує сильний обернений кореляційний зв’язок

Розрахунок частинних коефіцієнтів кореляції базується на обернений матриці до матриці $r_{X_i X_j}$ (матриця С):

28,76	37,36	0,67	9,55
37,36	63,09	10,74	16,93
0,67	10,74	11,53	-1,01
9,55	16,93	-1,01	9,41

Визначимо частинні коефіцієнти кореляції: $r_{kj} = \frac{-c_{kj}}{\sqrt{c_{kk} \cdot c_{jj}}}$

$r_{x_1 y}$ -0,88

$r_{x_2 y}$ -0,04

$r_{x_3 y}$ -0,58

$r_{x_1 x_2}$ -0,40

$r_{x_1 x_3}$ 0,69

$r_{x_2 x_3}$ 0,10

Частинні коефіцієнти кореляції, як і парні, характеризують тісноту зв'язку між двома змінними, але за умови, що решта змінних сталі.

Перевіримо значимість зв'язку між змінними моделі.

$$F = \frac{\dfrac{\sum_{i=1}^{n}\left(\hat{y}_i - \overline{Y}\right)^2}{m}}{\dfrac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n-m}} = \frac{\dfrac{413,88}{3}}{\dfrac{14,91}{15-3}} = 111,03$$

Порівняємо отримане значення з табличним значенням $F_{\hat{e}\partial} = F\left(\alpha; k_1; k_2\right)$, де $\alpha = 0,05$ - рівень значущості, $k_1 = m = 3$, $k_2 = n - m = 15 - 3 = 12$

$$F_{\hat{e}\partial} = F\left(0,05; 3; 12\right) = 3,49$$

Так як, $F > F_{\hat{e}\partial}$, то модель приймаємо , тобто припускаємо присутність лінійного зв'язку для рівня надійності α=0,95 .

Перевіримо на значущість коефіцієнт детермінації, коефіцієнт кореляції та оцінки параметрів моделі множинної регресії.

Перевірка значимості коефіцієнта детермінації

Висунемо нульову гіпотезу: $H_0: R^2 = 0$, $\left(a_0 = a_1 = ... = a_m = 0\right)$.

Альтернативна гіпотеза: $H_\alpha: R^2 \neq 0$, $\left(a_i \neq 0\right)$.

За отриманими в моделі значеннями коефіцієнта детермінації R^2 обчислюємо експериментальне значення F-статистики:

$$F_{\text{ðî çð}} = \frac{R^2}{1-R^2} \cdot \frac{n-m}{m} = \frac{0,97}{1-0,97} \cdot \frac{15-3}{3} = 129,33$$

Порівняємо отримане значення з табличним значенням $F_{\text{êð}} = F\left(0,05;3;12\right) = 3,49$

$F > F_{\text{êð}}$. Нульова гіпотеза відхиляється, тобто коефіцієнт детермінації значущий.

Перевірка значимості коефіцієнта кореляції

Коефіцієнт кореляції, як вибіркова характеристика, перевіряється на значущість за допомогою t-критерію Ст'юдента.

$$t = \frac{R\sqrt{n-m}}{\sqrt{1-R^2}} = \frac{0,985\sqrt{15-3}}{\sqrt{1-0,97}} = 19,7$$

Задамо рівень значущості α=0,05 та визначимо табличне значення t-критерію Ст'юдента:

$$t_{\text{êð}}\left(\frac{\alpha}{2}, n-m\right) = t_{\text{êð}}\left(\frac{0,05}{2};15-3\right) = t_{\text{êð}}\left(0,025;12\right) = 2,179$$

Отже, $|t| > t_{\text{êð}}$, тобто коефіцієнт кореляції значущий а зв'язок між залежною змінною та всіма незалежними факторами суттєвий.

Перевірка значимості оцінок параметрів моделі множинної регресії

Для оцінки значимості кожного параметра моделі перевіряємо їх за допомогою t-критерію Ст'юдента:

$$t_{a_j} = \frac{|a_j|}{\sqrt{\sigma_u^2 c_{jj}}}$$

де a_j – індивідуальні параметри матриці A;

σ_u^2 – дисперсія залишків;

c_{jj} – діагональні елементи матриці $\left(X^T X\right)^{-1}$;

$S_{a_j} = \sqrt{\sigma_u^2 c_{jj}}$ – стандартна помилка оцінки параметра моделі.

$$\sigma_u^2 = \frac{\sum_{i=1}^{n} u_i^2}{n-m} = \frac{14,91}{12} = 1,24$$

$c_{00} = 71,55$

$c_{11} = 0,08$

$c_{22} = 0,01$

$c_{33} = 0,004$

$S_{a_0} = \sqrt{\sigma_u^2 c_{00}} = \sqrt{1,24 \cdot 71,55} = 9,42$

$S_{a_1} = \sqrt{\sigma_u^2 c_{11}} = \sqrt{1,24 \cdot 0,08} = 0,31$

$S_{a_2} = \sqrt{\sigma_u^2 c_{22}} = \sqrt{1,24 \cdot 0,01} = 0,11$

$S_{a_3} = \sqrt{\sigma_u^2 c_{33}} = \sqrt{1,24 \cdot 0,004} = 0,07$

Висунемо нульову гіпотезу: $H_0 : a_i = 0$.

Альтернативна гіпотеза: $H_\alpha : a_i \neq 0$.

Розрахуємо значення t-критерію для кожного параметра і порівнюємо з табличним значенням t-критерію Ст'юдента:

$$t_{êð}\left(\frac{\alpha}{2}, n-m\right) = t_{êð}\left(\frac{0,05}{2}; 15-3\right) = t_{êð}\left(0,025; 12\right) = 2,179$$

Перевірка гіпотези H_0	Розраховане значення t-критерію	Висновки
$a_0 = 0$	$t_{a_0} = \frac{46,68}{9,42} = 4,96$	$\lvert t\rvert > t_{êð}$, змінна X_0 (вільний член) є значущою
$a_1 = 0$	$t_{a_1} = \frac{-2,0}{0,31} = -6,45$	$\lvert t\rvert > t_{êð}$, змінна X_1 є значущою
$a_2 = 0$	$t_{a_2} = \frac{-0,02}{0,11} = -0,18$	$\lvert t\rvert < t_{êð}$ змінна X_2 не значуща
$a_3 = 0$	$t_{a_3} = \frac{-0,18}{0,07} = -2,57$	$\lvert t\rvert > t_{êð}$, змінна X_3 є значущою

Знайдемо інтервали надійності для кожного окремого параметра за формулою:

Довірчі інтервали для параметрів a_i $(i = 0, 1, 2, 3)$ з надійністю γ мають вигляд

$$a_i^* - \mathrm{t_{p,n\text{-}m}} \cdot \mathrm{S}_{a_i} < a_i < a_i^* + \mathrm{t_{p,n\text{-}m}} \cdot \mathrm{S}_{a_i},$$

де $t_{p;n-m}$ – квантиль порядку $p = \frac{1+\gamma}{2}$ для розподілу Стьюдента з $n-m$ степенями вільності. В даній задачі $p = \frac{1+0,95}{2} = 0,975$, $n-m = 15-3 = 12$. За таблицями знаходимо $t_{0,975;12} = 2,179$.

Оскільки оцінки параметрів моделі a_i^*, $t_{p;n-m}$ і стандартні похибки параметрів моделі S_{a_i} обчислені, достатньо просто скористатися формулою для знаходження інтервалів:

$S_{a_0} = \sqrt{\sigma_u^2 c_{00}} = 9,42$, $S_{a_1} = \sqrt{\sigma_u^2 c_{11}} = 0,31$, $S_{a_2} = \sqrt{\sigma_u^2 c_{22}} = 0,11$, $S_{a_3} = \sqrt{\sigma_u^2 c_{33}} = 0,07$.

Знаходимо довірчі інтервали:

для параметра a_0: $(46,68 - 2,179 \cdot 9,42; 46,68 + 2,179 \cdot 9,42)$, тобто $(26,15; 67,21)$,

для параметра a_1: $(-2,0 - 2,179 \cdot 0,31; -2,0 + 2,179 \cdot 0,31)$, тобто $(-2,68; -1,32)$,

для параметра a_2: $(-0,02 - 2,179 \cdot 0,11; -0,02 + 2,179 \cdot 0,11)$, тобто $(-0,26; 0,22)$,

для параметра a_3: $(-0,18 - 2,179 \cdot 0,07; -0,18 + 2,179 \cdot 0,07)$, тобто $(-0,33; -0,03)$.

5. Мультиколінеарність

Однією з умов використання методу найменших квадратів (МНК) для знаходження параметрів економетричної моделі є те, що пояснювальні змінні у матриці X мають бути незалежними між собою, тобто $\left|X^T X\right| \neq 0$ (четверта умова застосування МНК). Проте на практиці можуть мати місце випадки, коли пояснювальні змінні пов'язані між собою, що стає перешкодою до використання МНК.

Явище існування тісної лінійної залежності, або сильної кореляції, між двома або більше пояснювальними змінними називається мультиколінеарністю. Термін "мультиколінеарність" вперше було впроваджено Р.Фрішем (1934 р,). Вона негативно впливає на кількісні характеристики економетричної моделі або взагалі робить неможливою її побудову.

Основні наслідки мультиколінеарності такі:

- падає точність оцінювання параметрів моделі;
- оцінки деяких параметрів моделі можуть показати порушення гіпотези про значимість зв'язку через наявність мільтиколінеарності пояснювальних змінних;

оцінки параметрів моделі стають дуже чутливими до розмірів сукупності спостережень і навіть збільшення цієї сукупності іноді може призвести до значних змін в оцінках параметрів.

Ознаки мультиколінеарності

1. Якщо серед парних коефіцієнтів кореляції незалежних змінних є такі, рівень яких наближається або дорівнює множинному коефіцієнту кореляції, то це свідчить про можливість існування мультиколінеарності. Інформацію про парну залежність може дати симетрична матриця коефіцієнтів парної кореляції, або кореляції нульового порядку:

$$r = \begin{pmatrix} r_{x_1x_1} & r_{x_1x_2} & \dots & r_{x_1x_m} \\ r_{x_2x_1} & r_{x_2x_2} & \dots & r_{x_2x_m} \\ . & . & \dots & . \\ r_{x_mx_1} & r_{x_mx_2} & \dots & r_{x_mx_m} \end{pmatrix}$$

Але якщо в моделі фігурує більше двох незалежних змінних, вивчення питання про мультиколінеарність не може обмежуватись інформацією, що дає ця матриця. Явище мультиколінеарності ні в якому разі не зводиться тільки до існування парної кореляції між незалежними змінними.

Більш загальна перевірка передбачає визначення визначника (детермінанта) матриці *r*, який називається детермінантом кореляції і позначається $|r|$. Числові значення детермінанта кореляції знаходяться на множині: $|r| \in [0,1]$.

2. Якщо $|r| = 1$, то існує повна мультиколінеарність, якщо $|r| = 0$ — мультиколінеарність відсутня, чим ближче $|r|$ до нуля, тим певніше можна стверджувати, що між незалежними змінними існує мультиколінеарність. Незважаючи на те, що на числове значення $|r|$ впливає дисперсія

незалежних змінних, цей показник можна вважати точковою мірою тісноти мультиколінеарності.

3. Якщо в економетричній моделі одержано мале значення параметра $\hat{a}_i$ при високому рівні коефіцієнта детермінації і при цьому F-критерій суттєво відрізняється від нуля, то це також свідчить про наявність мультиколінеарності.

4. Якщо коефіцієнт детермінації R^2, що розрахований для регресійних залежностей між однією незалежною змінною та іншими, має значення, яке близьке до одиниці, то можна говорити про наявність мультиколінеарності.

5. Якщо при побудові економетричної моделі на основі покрокової регресії включення нової незалежної змінної суттєво змінює оцінку параметрів моделі при незначному підвищенні (або зниженні) коефіцієнтів кореляції чи детермінації, то ця змінна, очевидно, знаходиться в лінійній залежності від інших, які введені в модель раніше.

Всі ці методи виявлення мультиколінеарності мають один загальний недолік: жоден із них не проводить чіткої межі між тим, що треба вважати «суттєвою» мультиколінеарністю, яку треба враховувати, і тим, коли мультиколінеарністю можна знехтувати.

Алгоритм Феррара—Глобера

Найбільш повне дослідження мультиколінеарності можна здійснити на основі алгоритму Феррара—Глобера. Цей алгоритм включає три види

статистичних критеріїв, на основі яких перевіряється мультиколінеарність всього масиву незалежних змінних (χ^2 , хі-квадрат); кожної незалежної змінної зі всіма незалежними змінними (*F*-критерій) і мультиколінеарність кожної пари незалежних змінних (*t*-критерій).

Всі ці критерії при порівнянні з їх критичними значеннями дають можливість зробити конкретні висновки відносно наявності чи відсутності мультиколінеарності незалежних змінних.

Опишемо алгоритм Феррара—Глобера.

Крок 1. Стандартизація (нормалізація) змінних.

Позначимо вектори незалежних змінних економетричної моделі через X_1, X_2,…, X_m. Елементи стандартизованих векторів розрахуємо за формулою:

$$x_{ik}^{*} = \frac{x_{ik} - \overline{x_k}}{\sqrt{n\sigma_{x_k}^2}},$$

де n – кількість спостережень;

x_{ik}^{*} – нормалізовані (стандартизовані) змінні.

$\overline{x_k}$ – середні незалежних змінних;

σ_{x_k} – середньоквадратичні відхилення змінних.

Крок 2. Знаходження кореляційної матриці (матриці моментів стандартизованої системи нормальних рівнянь): $R = X^{*T} X^{*}$

де X^{*} – матриця нормалізованих змінних; X^{*T} – матриця, транспонована до X*.

Крок 3. Визначення критерію χ^2 (хі-квадрат):

$$\chi^2 = -\left[n-1-\frac{1}{6}(2m+5) \right] \ln|R|,$$

де $|R|$ — визначник кореляційної матриці R.

Значення цього критерію порівнюється з табличним при $\frac{1}{2} m\,(m-1)$ ступенях свободи і рівні значущості α**.** Якщо $\chi^2_{\text{ô àêò}} < \chi^2_{\text{ô àáë}}$, то в масиві незалежних змінних не існує мультиколінеарності.

Крок 4. Визначення оберненої матриці C: $C = R^{-1} = (X^{*T} X^{*})^{-1}$.

Крок 5. Розрахунок *F*- критеріїв: $F_k = (c_{kk} - 1)\frac{n-m}{m-1}$**,**

де c_{kk} — діагональні елементи матриці C. Фактичні значення критеріїв F_k порівнюються з табличними при $n-m$ і $m-1$ ступенях свободи і рівні значущості α. Якщо $F_k > F_{\text{êð}}$, то відповідна *k*-та незалежна змінна мультиколінеарна з іншими.

Крок 6. Знаходження часткових коефіцієнтів кореляції:

$$r_{kj} = \frac{-c_{kj}}{\sqrt{c_{kk} \cdot c_{jj}}},$$

де c_{kj} — елемент матриці C, що заходиться в k-му рядку і j-му стовпці, c_{kk}, c_{jj} — діагональні елементи матриці C.

Крок 7. Розрахунок t критеріїв:

$$t_{kj} = \frac{r_{kj}\sqrt{n-m}}{\sqrt{1-r_{kj}^2}}.$$

Фактичні значення критеріїв t_{kj} порівнюються з табличними при $n-m$ ступенях свободи і рівні значущості α. Якщо $t_{kj} > t_{\text{êð}}$, між незалежними змінними X_k і X_j існує мультиколінеарність.

Дослідження наявності мультиколінеарності на основі алгоритму Феррара—Глобера

Розглянемо застосування алгоритму Феррара—Глобера для розв'язування конкретної задачі.

На середньомісячну заробітну плату впливає ряд факторів. Виділимо серед них продуктивність праці, фондомісткість та коефіцієнт плинності робочої сили. Щоб побудувати економетричну модель заробітної плати від наведених чинників на основі методу найменших квадратів, треба переконатись, що продуктивність праці, фондомісткість та коефіцієнт плинності робочої сили як незалежні змінні — не мультиколінеарні.

Вихідні дані наведені в таблиці.

Номер цеху	Продуктивність праці, млн.грн./ люд.	Фондомісткість, грн./грн.	Коефіцієнт плинності робочої сили, %
1	32	0,59	10,5
2	29	0,43	15,5
3	30	0,70	13,5
4	31	0,61	9,5
5	25	0,51	2,5
6	34	0,51	1,5
7	29	0,65	17,5
8	24	0,43	14,5
9	20	0,51	14,5
10	35	0,92	7,5

Розв’язання

Крок 1. Нормалізація змінних.

Позначимо вектори незалежних змінних — продуктивності праці, фондомісткості, коефіцієнтів плинності робочої сили — через відповідно X_1, X_2, X_3. Елементи стандартизованих векторів розрахуємо за формулою:

$$x_{ik}^{*} = \frac{x_{ik} - \overline{x_k}}{\sqrt{n\sigma_{x_k}^2}}$$

де n — кількість спостережень, $n = 10$; m — число незалежних змінних, $m = 3$; $\overline{X_k}$ — середня арифметична вектора X_k; $\sigma_{x_k}^2$ — дисперсія змінної X_k.

$$\overline{X_1} = \frac{\sum_{i=1}^{10} X_{1i}}{n} = \frac{287}{10} = 28,7\,, \qquad \overline{X_2} = \frac{\sum_{i=1}^{10} X_{2i}}{n} = \frac{5,86}{10} = 0,586\,,$$

$$\overline{X_3} = \frac{\sum_{i=1}^{10} X_{3i}}{n} = \frac{139}{10} = 13,9$$

Дисперсії кожної незалежної змінної мають такі значення:

$$\sigma_{X_1}^2 = \frac{\sum_{i=1}^{10} \left(X_{1i} - \overline{X_1}\right)^2}{n} = \frac{176,1}{10} = 17,61;$$

$$\sigma_{X_2}^2 = \frac{\sum_{i=1}^{10} \left(X_{2i} - \overline{X_2}\right)^2}{n} = \frac{0,19524}{10} = 0,0195;$$

$$\sigma_{X_3}^2 = \frac{\sum_{i=1}^{10} \left(X_{3i} - \overline{X_3}\right)^2}{n} = \frac{182,4}{10} = 18,24\,.$$

Тоді знаменник для стандартизації кожної незалежної змінної буде дорівнювати:

для X_1: $\sqrt{n\sigma^2_{X_1}} = \sqrt{10 \cdot 17,61} = \sqrt{176,1} = 13,27$;

для X_2: $\sqrt{n\sigma^2_{X_2}} = \sqrt{10 \cdot 0,0195} = \sqrt{0,195} = 0,44$;

для X_3: $\sqrt{n\sigma^2_{X_3}} = \sqrt{10 \cdot 18,24} = \sqrt{182,4} = 13,51$.

В наступній таблиці наведені всі розрахунки по стандартизації незалежних змінних X_1, X_2, X_3 згідно з наведеним співвідношенням.

$X_{1i} - \overline{X_1}$	$X_{2i} - \overline{X_2}$	$X_{3i} - \overline{X_3}$	$\left(X_{1i} - \overline{X_1}\right)^2$	$\left(X_{2i} - \overline{X_2}\right)^2$	$\left(X_{3i} - \overline{X_3}\right)^2$	X_{1i}^*	X_{2i}^*	X_{3i}^*
3.3	0,004	–3,4	10,89	0,000015	11,56	0,2487	0,0091	–0,2518
0.3	–0,156	1,6	0,09	0,024336	2,56	0,0226	–0,2531	0,1185
1.3	0,114	–0,4	1,69	0,012996	0,16	0,0979	0,2580	–0,0296
2.3	0,024	–4,4	5,29	0,000576	19,36	0,1733	0,0543	-0,3258
3.7	–0,076	9,6	13,09	0,005776	92,16	–0,2788	–0,1720	0,7108
5.3	–0,076	–1,4	23,09	0,005776	1,96	0,3994	–0,1720	–0,1037
0.3	0,064	3,5	10,09	0,004095	12,95	0,0226	0,1448	0,2666
–4.7	–0,156	0,6	22,09	0,024336	0,36	–0,3542	–0,3531	0,0444
–8.7	–0,076	0,6	75,69	0,005776	0,36	–0,6556	–0,1720	0,0444
4.3	0,334	–6,4	14,49	0,111556	40,95	0,3240	0,7559	–0,4739
$\sum$			176,1	0,19524	182,4			

Матриця стандартизованих змінних матиме вигляд:

$$X^* = \begin{pmatrix} 0,2487 & 0,0090 & -0,2518 \\ 0,0226 & -0,3531 & 0,1185 \\ 0,0979 & 0,2580 & -0,0296 \\ 0,1733 & 0,0543 & -0,3258 \\ -0,2788 & -0,1720 & 0,7108 \\ 0,3994 & -0,1720 & -0,1037 \\ 0,0226 & 0,1448 & 0,2666 \\ -0,3542 & -0,3531 & 0,0444 \\ -0,6556 & -0,1720 & 0,0444 \\ 0,3240 & 0,7559 & -0,4739 \end{pmatrix}.$$

Крок 2. Знаходження кореляційної матриці R:

$$R = X^{*T} X^*$$

Ця матриця симетрична і має розмір 3 х 3. Для даної задачі:

$$R = \begin{pmatrix} 1 & 0,494 & -0,551 \\ 0,494 & 1 & -0,517 \\ -0,551 & -0,517 & 1 \end{pmatrix}.$$

Кожен елемент цієї матриці характеризує тісноту зв'язку однієї незалежної змінної з іншою. Оскільки діагональні елементи характеризують тісноту зв'язку кожної незалежної змінної з цією самою змінною, то вони дорівнюють одиниці.

Інші елементи матриці R трактуються так:

$r_{X_1X_2} = 0,494$;

$r_{X_1X_3} = -0,551$;

$r_{X_2X_3} = -0,517$,

тобто вони є парними коефіцієнтами кореляції незалежних змінних. На основі цих коефіцієнтів можна зробити висновок, що між змінними X_1, X_2, X_3 існує зв'язок. Але чи можна стверджувати, що цей зв'язок є явищем мультиколінеарності і він негативно впливатиме на оцінку економетричної моделі?

Щоб відповісти на це запитання, треба продовжити розв'язання на основі алгоритму Феррара—Глобера і в результаті знайти статистичні критерії оцінки мультиколінеарності.

Крок 3. Знайдемо визначник кореляційної матриці R і критерій χ^2 (хі-квадрат):

$|R| = 0,466$;

$$\chi^2 = -\left[n - 1 - \frac{1}{6}(2m+5)\right]\ln|R| = -\left[10 - 1 - \frac{1}{6}(2\cdot 3+5)\right]\ln 0,466 = 2,37.$$

При ступені свободи $\frac{1}{2}m(m-1) = 3$ і рівні значущості $\alpha = 0,05$ маємо $\chi^2_{\text{òàáë}} = 7,8$. Так як $\chi^2_{\text{ôàêò}} < \chi^2_{\text{òàáë}}$, можна зробити висновок, що в масиві змінних не існує мультиколінеарність.

Крок 4. Знайдемо матрицю, обернену до матриці R:

$$C = R^{-1} = (X^{*T}\ X^{*})^{-1};$$

$$\tilde{N} = \begin{pmatrix} 1,572 & -0,448 & 0,634 \\ -0,448 & 1,492 & 0,524 \\ 0,634 & 0,524 & 1,620 \end{pmatrix}.$$

Крок 5. Використовуючи діагональні елементи матриці C, розрахуємо F-критерії:

$$F_1 = (c_{11} - 1)\frac{n-m}{m-1} = (1,572 - 1)\frac{10-3}{3-1} = 2,0$$

$$F_2 = (c_{22} - 1)\frac{n-m}{m-1} = (1,492 - 1)\frac{10-3}{3-1} = 1,72;$$

$$F_3 = (c_{33} - 1)\frac{n-m}{m-1} = (1,620 - 1)\frac{10-3}{3-1} = 2,17.$$

При рівні значущості $\alpha = 0,05$ і ступенях свободи $\gamma_1 = n - m = 10 - 3 = 7$ і $\gamma_2 = m - 1 = 3 - 1 = 2$ критичне (табличне) значення критерію $F_{êð} = F(0,05;2;7) = 4,74$.

Так як, $F_1 < F_{êð}$, $F_2 < F_{êð}$, $F_3 < F_{êð}$, то жодна із незалежних змінних не мультиколінеарна з двома іншими.

Щоб визначити наявність попарної мультиколінеарності, продовжимо розрахунок і перейдемо до кроку 6.

Крок 6. Розрахуємо часткові коефіцієнти кореляції, використавши елементи матриці C:

$$r_{12} = \frac{-c_{12}}{\sqrt{c_{11} \cdot c_{22}}} = \frac{0,448}{\sqrt{1,572 \cdot 1,492}} = 0,293;$$

$$r_{13} = \frac{-c_{13}}{\sqrt{c_{11} \cdot c_{33}}} = \frac{-0,634}{\sqrt{1,572 \cdot 1,620}} = -0,397;$$

$$r_{23} = \frac{-c_{23}}{\sqrt{c_{22} \cdot c_{33}}} = \frac{-0,524}{\sqrt{1,492 \cdot 1,620}} = -0,337.$$

Часткові коефіцієнти кореляції характеризують тісноту зв'язку між двома змінними за умови, що третя не впливає на цей зв'язок.

Порівнявши часткові коефіцієнти кореляції з парними, які наведені вище, можна помітити, що часткові коефіцієнти значно менше парних. Це ще раз підтверджує, що на основі парних коефіцієнтів кореляції не можна зробити висновки про наявність чи відсутність мультиколінеарності.

Крок 7. Визначимо *t*- критерії на основі часткових коефіцієнтів кореляції:

$$t_{12} = \frac{r_{12}\sqrt{n-m}}{\sqrt{1-r_{12}^2}} = \frac{0,293\sqrt{10-3}}{\sqrt{1-0,293^2}} = 0,811;$$

$$t_{13} = \frac{r_{13}\sqrt{n-m}}{\sqrt{1-r_{13}^2}} = \frac{0,397\sqrt{10-3}}{\sqrt{1-0,397^2}} = 1,144;$$

$$t_{23} = \frac{r_{23}\sqrt{n-m}}{\sqrt{1-r_{23}^2}} = \frac{0,337\sqrt{10-3}}{\sqrt{1-0,337^2}} = 0,947.$$

Табличне значення *t*- критерію при $n - m = 7$ ступенях свободи і рівні значущості $\alpha = 0,05$ дорівнює 2,365. Всі числові значення *t*- критеріїв, знайдених для кожної пари змінних, менше за їх табличне значення. Звідси

робимо висновок, що всі пари незалежних змінних не є мультиколінеарними.

Таким чином, незважаючи на те, що між незалежними змінними, що досліджуються, існує лінійна залежність, але вона не є явищем мультиколінеарності і не буде негативно впливати на кількісні параметри економетричної моделі.

Якщо *F*- критерій більше табличного значення, а це значить, що *k*-та змінна залежить від всіх інших в масиві, то необхідно вирішувати питання про її виключення з переліку змінних.

Якщо t_{kj} - критерій більше табличного, то ця пара змінних (*k* і *j*) тісно взаємопов'язані. Звідси, аналізуючи рівень обох видів критеріїв *F* і *t*, можна зробити обгрунтований висновок про те, яку із змінних необхідно виключити із дослідження чи замінити іншою. Але заміна масиву незалежних змінних завжди повинна узгоджуватись із економічною доцільністю, що випливає з мети дослідження.

6. Гетероскедастичність.

Передумови, які висуваються при оцінці параметрів моделі за методом 1МНК на практиці часто можуть порушуватись. Однією з таких передумов є незмінність дисперсії залишків для всіх спостережень вихідної сукупності. Це явище називається гомоскедастичністю. В практичних дослідженнях воно часто порушується. Наприклад, в економетричній моделі, що характеризує залежність витрат на споживання від доходу, дисперсія залишків може змінюватись для спостережень, які відносяться до різних груп населення за розміром доходів.

Якщо дисперсія залишків в економетричному моделюванні змінюється для кожного спостереження або для груп спостережень, то це явище називається гетероскедастичністю.

Наявність гетероскедастичності спричиняє порушення властивостей оцінок параметрів моделі при розрахунку їх за методом 1МНК. Тому завжди виникає необхідність вивчати це явище, і, якщо воно існує, для оцінки параметрів моделі використовувати узагальнений метод найменших квадратів (метод Ейткена).

Для визначення гетероскедастичності застосовуються наступні критерії: 1) критерій μ; 2) тест Уайта; 3) тест рангової кореляції Спірмена; 4) параметричний тест Гольдфельда—Квандта; 5) непараметричний тест Гольдфельда—Квандта; 6) тест Глейсера і т. д.

Критерій μ

Цей критерій застосовується в тих випадках, коли вихідна сукупність спостережень досить велика. Розглянемо цей алгоритм.

Крок 1. Вихідні дані залежної змінної Y розбиваються на k груп згідно із зміною рівня величини Y.

Крок 2. По кожній групі даних розраховується сума квадратів відхилень:

$$S_r = \sum_{i=1}^{n_r} (y_{ir} - \overline{y_r})^2 .$$

Крок 3. Розраховується сума квадратів відхилень у цілому по всій сукупності спостережень:

$$\sum_{r=1}^{k} S_r = \sum_{i=1}^{n_r} \sum_{r=1}^{k} \left(y_{ir} - \overline{y_r} \right)^2 .$$

Крок 4. Обчислюється параметр λ: $\lambda = \dfrac{\prod_{r=1}^{k} \left(\dfrac{S_r}{n_r} \right)^{\frac{n_r}{2}}}{\left(\dfrac{\sum_{r=1}^{k} S_r}{n} \right)^{\frac{n}{2}}}$,

де n — загальна сукупність спостережень;

n_r — кількість спостережень r-ї групи.

Крок 5. Розраховується критерій μ:

$$\mu = -2\ln\lambda\,,$$

який наближено буде відповідати розподілу χ^2 при ступенях свободи $k-1$, коли дисперсія всіх спостережень однорідна. Тобто, якщо значення μ менше табличного значення $\chi^2_{табл}$ при вибраному рівні значимості і ступені свободи $k-1$, то явище гетероскедастичності відсутнє.

Коли сукупність спостережень невелика, то розглянутий метод застосовувати неможливо.

Тест Уайта (White test)

Проводиться тест таким чином:

Крок 1. Припустимо, що вихідна модель має вигляд: $y_i = a_0 + a_1 x_{1i} + a_2 x_{2i} + a_3 x_{3i} + u_i$. У результаті оцінки даної моделі ми отримуємо регресійні залишки $u_i = y_i - \hat{y}_i$;

Крок 2. Оцінюється допоміжна регресія вигляду:

$$e_i^2 = b_0 + b_1 x_{1i} + b_2 x_{2i} + b_3 x_{3i} + b_4 x_{1i}^2 + b_5 x_{2i}^2 + \\ + b_6 x_{3i}^2 + b_7 x_{1i} x_{2i} + b_8 x_{1i} x_{3i} + b_9 x_{2i} x_{3i} + v_i\,,$$

де v_i – нормально розподілена помилка, незалежна від u_i. Допоміжна регресія дозволяє визначити, чи існує якась систематична можливість між змінами e_i^2 і будь-якою релевантною змінною моделі (щоб побачити, що релевантними є саме змінні, включені в допоміжну регресію, слід представити похибку у вигляді $u_i = y_i - a_0 - a_1 x_{1i} - a_2 x_{2i} - a_3 x_{3i}$ і піднести даний вираз у квадрат).

Крок 3. Досліджується статистика: $nR^2 \approx \chi^2 m$, де n – кількість спостережень; m – кількість регресорів у допоміжній регресії (за винятком постійного члена), тобто кількість параметрів біля змінних X_i. На основі статистики перевіряється нуль-гіпотеза:

$$H_0 : b_2 = b_3 = ... = b_m = 0$$

$$H_\alpha : b_i \neq 0 .$$

Якщо фактичні значення статистики перевищують критичні величини розподілу χ^2, то нульова гіпотеза про гомоскедастичність залишків відхиляється, тобто робиться висновок про наявність гетероскедастичності.

Тест рангової кореляції Спірмена.

При виконанні тесту рангової кореляції Спірмена передбачається, що дисперсія випадкового члена буде або збільшуватися, або зменшуватися по мірі збільшення x, і тому в регресії, що оцінюється за допомогою 1МНК, абсолютні величини залишків і значення x будуть корельованими. Дані по x і залишки упорядковується, і коефіцієнт рангової кореляції визначається як:

$$r_{x,e} = 1 - \frac{6\sum D_i^2}{n\left(n^2 - 1\right)},$$

де D – різниця між рангом x та рангом e. Якщо припустити, що коефіцієнт кореляції для генеральної сукупності дорівнює нулеві, то коефіцієнт рангової кореляції має нормальний розподіл з математичним сподіванням

0 і дисперсією $\frac{1}{n-1}$ у великих вибірках. Отже, відповідна тестова статистика дорівнює $r_{x,e}\sqrt{n-1}$ і при використанні двостороннього критерію нульова гіпотеза про відсутність гетероскедастичності буде відхилена при рівні значущості в 5 %, якщо вона перевищить 1,96, і при рівні значущості в 1 %, якщо вона перевищить 2,58. Якщо в моделі регресії є більше однієї пояснювальної змінної, то перевірка гіпотези може виконуватися з використанням будь-якої з них.

Або можна використати t-статистику, тоді: $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$, де n – кількість спостережень та $(n-2)$ – кількість ступенів вільності. При даних ступенях вільності знаходимо $t_{êð}$. Якщо $t > t_{êð}$, то це підтверджує гіпотезу про гетероскедастичність. Якщо $t < t_{êð}$, то в регресійній моделі правильним буде припущення про гомоскедастичність.

Алгоритм тесту рангової кореляції Спірмена:

Крок 1. Будується регресійне рівняння і оцінюються параметри регресії.

Крок 2. Спостереженням x_i за зростанням привласнюються ранги.

Крок 3. Похибкам u_i (за модулем) (відхилення від лінії регресії $u_i = y_i - \hat{y}_i$) також привласнюються ранги в порядку зростання.

Крок 4. Знаходимо D_i та D_i^2 – різниці рангів та їх квадрати.

Крок 5. Оцінюється значення рангового коефіцієнту кореляції: $r_{x,e} = 1 - \frac{6\sum D_i^2}{n\left(n^2 - 1\right)}$

Крок 6. Розраховується t-статистика: $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$.

Крок 7. Робиться висновок щодо наявності або відсутності гетероскедастичності.

Застосування тесту рангової кореляції Спірмена для визначення гетероскедастичності

Завдання 1. Нехай треба побудувати економетричну модель, яка характеризує залежність попиту на деяку продукцію від ціни. Для побудови цієї моделі використовується вихідна сукупність даних, яка включає 15 спостережень. Ці дані та розрахунки на їх основі наведені в таблиці.

Виходячи із сутності взаємозв'язку величини заощаджень та доходу населення, можна припустити, що дисперсія залишків не є постійною для кожного спостереження, тобто тут може існувати явище гетероскедастичності. Тому, щоб правильно вибрати метод для оцінки параметрів моделі, необхідно перевірити, чи властива гетероскедастичність для наведених вихідних даних.

Розрахункова таблиця для проведення тесту рангової кореляції Спірмена

Ранг за змінною X	Ціна X	Попит Y	Залишки	Ранг за залишком	Різниця рангів D_i	D_i^2
8	15,91	117,09	-0,34387	1	7	49
5	15,54	119,86	-0,39014	2	3	9
15	16,76	110,02	-0,84306	3	12	144
2	15,21	123,81	1,019821	4	-2	4
3	15,28	121,17	-1,11646	5	-2	4
9	15,92	116,17	-1,12322	6	3	9
10	15,95	118,34	1,257187	7	3	9
14	16,69	110,11	-1,31194	8	6	36
1	15,09	125,18	1,426776	9	-8	64
6	15,62	118,07	-1,5813	10	-4	16
11	16,31	116,21	1,847847	11	0	0
12	16,33	111,46	-2,67328	12	0	0
13	16,60	115,10	3,003645	13	0	0
4	15,49	116,91	-3,7319	14	-10	100
7	15,70	123,59	4,559903	15	-8	64
					Сума	508

Значення рангової кореляції Спірмена буде дорівнювати:

$$r_{x,e} = 1 - \frac{6 \cdot \sum_{i=1}^{n} D_i^2}{n \cdot \left(n^2 - 1\right)} = 1 - \frac{6 \cdot 508}{15 \cdot (225 - 1)} = 0,0928$$

А значення статистики буде $t = r_{x,e}\sqrt{n-1} = 0,0928 \cdot \sqrt{15-1} = 0,3472$

Виберемо рівень значимості 5 %, отримаємо критичну точку $t_{êð} = 2,16$.

Оскільки умова $t < t_{êð}$ виконується, то гіпотеза про наявність гетероскедастичності буде прийнята.

Параметричний тест Гольдфельда—Квандта

Гольдфельд і Квандт розглянули випадок, коли дисперсія залишків зростає пропорційно квадрату однієї із незалежних змінних моделі:

$$Y = XA + u.$$

Вони запропонували для виявлення наявності гетероскедастичності параметричний тест, в якому треба виконати наступні кроки.

Крок 1. Упорядкувати спостереження згідно з величиною елементів вектора x_i.

Крок 2. Відкинути c спостережень, які будуть знаходитись у центрі вектора. На основі експериментальних розрахунків автори вирахували оптимальні співвідношення між параметрами c і n, де n — кількість елементів вектора x_i.

$$\frac{c}{n} = \frac{4}{15} \text{ або } c = \frac{4n}{15}.$$

Крок 3. Побудувати дві економетричні моделі на основі 1МНК по двох створених сукупностях спостережень $\frac{n_1 - c}{2}$ за умови, що $\frac{n_2 - c}{2}$ перевищує кількість змінних m.

Крок 4. Знайти суму квадратів залишків за першою (1) і другою (2) моделях S_1 і S_2.

$S_1 = \hat{u}_{1i}^T \hat{u}_{1i}$, де $\hat{u}_{1i}$ — залишки по моделі (1);

$S_2 = \hat{u}_{2i}^T \hat{u}_{2i}$, де $\hat{u}_{2i}$ — залишки по моделі (2).

Крок 5. Розрахувати критерій R:

$$R = \frac{S_2}{S_1}, \text{ де } S_2 > S_1,$$

який при виконанні гіпотези про гомоскедастичність буде відповідати F-розподілу з $\frac{n_1 - c - 2m}{2}$, $\frac{n_2 - c - 2m}{2}$ ступенями свободи, де m - кількість параметрів моделі, що досліджується. Це означає, що розраховане значення R порівнюється з табличним значенням F-критерію при ступенях свободи $\frac{n_1 - c - 2m}{2}$ і $\frac{n_2 - c - 2m}{2}$ і вибраному рівні довіри. Якщо $R \le F_{\text{òàáë}}$, то гетероскедастичність відсутня.

Застосування параметричного тесту Гольдфельда—Квандта для визначення гетероскедастичності

Завдання 1. Нехай треба побудувати економетричну модель, яка характеризує залежність заощаджень від доходів населення. Для побудови цієї моделі використовується вихідна сукупність даних, яка включає 18 спостережень. Ці дані та розрахунки на основі їх наведені в таблиці. Виходячи із сутності взаємозв'язку величини заощаджень та доходу населення, можна припустити, що дисперсія залишків не є постійною для кожного спостереження, тобто тут може існувати явище гетероскедастичності. Тому, щоб правильно вибрати метод для оцінки параметрів моделі, необхідно перевірити, чи властива гетероскедастичність для наведених вихідних даних.

Рік	Заощадження Y	Дохід X
1	1,36	13,8
2	1,20	14,4
3	1,08	15,0
4	1,20	15,6
5	1,10	16,0
6	1,12	16,9
7	1,41	17,7
8	1,50	18,5
9	1,43	19,3
10	1,59	20,5

11	1,90	21,7
12	1,95	22,7
13	1,82	23,6
14	2,04	24,7
15	2,53	26,1
16	2,94	27,8
17	2,75	28,9
18	2,99	30,2

В таблиці дані впорядковані за величиною доходу, починаючи від меншого до більшого значення.

Розв'язання

1**.** Ідентифікація змінних:

$$Y = f\left(X, u\right),$$

де Y — залежна змінна (заощадження);

X — незалежна змінна (дохід);

u — стохастична складова.

2**.** Специфікація моделі:

$$Y = a_0 + a_1 X + u$$

$$\hat{Y} = \hat{a}_0 + \hat{a}_1 X$$

$$u = Y - \hat{Y}$$

3. Визначимо наявність гетероскедастичності. Для цього застосуємо алгоритм Гольдфельда-Квандта. Дану сукупність спостережень впорядкуємо по X від меншого до більшого значення. Відшукуємо c спостережень, які знаходяться всередині сукупності:

$$c = \frac{4n}{15} = \frac{4 \cdot 18}{15} = 4{,}8 \approx 4 .$$

Тоді отримаємо дві сукупності спостережень об'ємом $n_1 = n_2 = \frac{n-c}{2} = \frac{18-4}{2} = 7$.

Побудуємо розрахункову таблицю на основі отриманих спостережень

Рік	Заощадження Y	Дохід X		X^2	XY	$\hat{Y}$	$Y - \hat{Y}$	$\left(Y - \hat{Y}\right)^2$
1	1,36	13,8		190,44	18,768	1,201	0,159	0,025281
2	1,20	14,4		207,36	17,280	1,204	-0,004	0,000016
3	1,08	15,0		225,00	16,200	1,207	-0,127	0,016129
4	1,20	15,6	n1	243,36	18,720	1,21	-0,01	0,0001
5	1,10	16,0		256,00	17,600	1,212	-0,112	0,012544
6	1,12	16,9		285,61	18,928	1,2165	-0,0965	0,009312
7	1,41	17,7		313,29	24,957	1,2205	0,1895	0,03591
8	1,50	18,5						

9	1,43	19,3						
10	1,59	20,5						
11	1,90	21,7						
12	1,95	22,7		515,29	44,265	1,8401	0,1099	0,0121
13	1,82	23,6		556,96	42,925	1,9885	–0,1685	0,0284
14	2,04	24,7		610,09	50,388	2,1699	–0,1299	0,0169
15	2,53	26,1	n2	681,21	66,033	2,4008	0,1292	0,0167
16	2,94	27,8		772,84	81,732	2,6811	0,2589	0,0670
17	2,75	28,9		835,21	79,475	2,8625	–0,1125	0,0127
18	2,99	30,2		912,04	90,298	3,0769	–0,0869	0,0076

3.1. Розрахуємо економетричну модель для сукупності $n_1 = 7$.

Оцінимо кількісно параметри моделі на основі 1МНК.

$$\begin{cases} a_1 = \dfrac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \left(\bar{x}\right)^2} \\ a_0 = \bar{y} - a_1 \cdot \bar{x} \end{cases}$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n_1} = \frac{109,4}{7} = 15,63\,, \qquad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n_1} = \frac{8,47}{7} = 1,21$$

$$\overline{x^2} = \frac{\sum_{i=1}^{n} x_i^2}{n_1} = \frac{1721,06}{7} = 245,87 \qquad \overline{xy} = \frac{\sum_{i=1}^{n} x_i y_i}{n_1} = \frac{132,45}{7} = 18,92$$

$$\begin{cases} a_1 = \dfrac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \left(\bar{x}\right)^2} = \dfrac{18,92 - 15,63 \cdot 1,21}{245,87 - 15,63^2} = 0,005 \\ a_0 = \bar{y} - a_1 \cdot \bar{x} = 1,21 - 0,005 \cdot 15,63 = 1,132 \end{cases}$$

Звідси, $Y_1 = 1,132 + 0,005X$ - перша економетрична модель.

На основі моделі можна зробити висновок: і якщо дохід виросте на 1 одиницю, то заощадження збільшаться на 0,005 одиниці.

3.2. Розрахуємо економетричну модель для сукупності $n_2 = 7$

Оцінимо кількісно параметри моделі на основі 1МНК.

$$\begin{cases} a_1 = \dfrac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \left(\bar{x}\right)^2} \\ a_0 = \bar{y} - a_1 \cdot \bar{x} \end{cases}$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n_2} = \frac{184}{7} = 26,29\,, \qquad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n_2} = \frac{17,02}{7} = 2,43$$

$$\overline{x^2} = \frac{\sum_{i=1}^{n} x_i^2}{n_2} = \frac{4883,64}{7} = 697,66 \qquad \overline{xy} = \frac{\sum_{i=1}^{n} x_i y_i}{n_2} = \frac{455,14}{7} = 65,02$$

$$\begin{cases} a_1 = \dfrac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \left(\bar{x}\right)^2} = \dfrac{65,02 - 26,29 \cdot 2,43}{697,66 - 26,29^2} = 0,175 \\ a_0 = \bar{y} - a_1 \cdot \bar{x} = 2,43 - 0,175 \cdot 26,29 = -2,171 \end{cases}$$

Звідси, $Y_2 = -2,171 + 0,175X$ — друга економетрична модель.

На основі моделі можна зробити висновок: і якщо дохід виросте на 1 одиницю, то заощадження збільшаться на 0,175 одиниці для даної сукупності спостережень.

3.3. Для кожної моделі знайдемо суму квадратів залишків:

$$S_1 = u_1^T u_1 = \sum_{i=1}^{n_1} \left(y_{1i} - \hat{y}_{1i} \right)^2 ;$$

$$S_2 = u_2^T u_2 = \sum_{i=1}^{n_2} \left(y_{2i} - \hat{y}_{2i} \right)^2 ;$$

$$S_1 = 0{,}0992925 ;$$

$$S_2 = 0{,}166152$$

3.4. Знаходимо критерій R:

Так як $S_2 > S_1$, то $R = \dfrac{S_2}{S_1}$

$$R = \frac{S_2}{S_1} = \frac{0{,}166152}{0{,}0992925} \approx 1{,}67 .$$

Порівняємо цей критерій з табличним значенням F-критерію при ступенях свободи $k_1 = k_2 = \dfrac{n - c - 2m}{2} = \dfrac{18 - 4 - 2 \cdot 2}{2} = 5$ і рівні довіри $\alpha = 0{,}05$. Звідси, $F_{\text{òàáë}} = F\left(0{,}05; 5; 5\right) = 5{,}05$.

Гетероскедастичність відсутня, тому що $R < F_{\text{òàáë}}$.

Оцінка параметрів моделі на основі узагальненого методу найменших квадратів (методу Ейткена)

Завдання 2. Необхідно оцінити параметри економетричної моделі, яка характеризує залежність витрат на харчування від загальних затрат на основі даних, що наведені в таблиці.

Розв'язання

Виходячи з особливостей вихідної інформації, можна припустити, що порушується гіпотеза про незмінність дисперсії.

1**.** Ідентифікуємо змінні моделі:

$$Y = f\left(X, u\right),$$

де Y — залежна змінна (заощадження);

X — незалежна змінна (дохід);

u — стохастична складова.

2**.** Специфікація моделі:

$$Y = a_0 + a_1 X + u$$

$$\hat{Y} = \hat{a}_0 + \hat{a}_1 X$$

$$u = Y - \hat{Y}$$

3. Визначимо наявність гетероскедастичності. Для цього застосуємо алгоритм Гольдфельда-Квандта. Дану сукупність спостережень

впорядкуємо по X від меншого до більшого значення. Відшукуємо c спостережень, які знаходяться всередині сукупності:

$$c = \frac{4n}{15} = \frac{4 \cdot 18}{15} = 4{,}8 \approx 4 .$$

Тоді отримаємо дві сукупності спостережень об'ємом $n_1 = n_2 = \frac{n-c}{2} = \frac{18-4}{2} = 7$.

Побудуємо розрахункову таблицю:

№	Витрати на харчування	Загальні затрати	$\hat{Y}$	$u_i = Y - \hat{Y}$	u_i^2	$\lambda_i = \frac{1}{x_i}$
1	4,3	24	4,38	-0,082	0,0067	0,0417
2	4,5	25	4,52	-0,019	0,0004	0,0400
3	4,7	27	4,79	-0,094	0,0089	0,0370
4	4,8	27	4,79	0,006	0,0000	0,0370
5	5,2	29	5,07	0,130	0,0170	0,0345
6	5,6	31	5,34	0,255	0,0650	0,0323
7	5,7	35	5,90	-0,195	0,0382	0,0286
8	6,1	38				0,0263
9	6,3	41				0,0244
10	5,9	46				0,0217
11	6,7	57				0,0175
12	7,6	62	7,52	0,084	0,007	0,0161

13	8,4	69	7,68	0,722	0,521	0,0145
14	7,1	76	7,84	-0,740	0,547	0,0132
15	7,9	84	8,02	-0,125	0,016	0,0119
16	8,4	105	8,51	-0,111	0,012	0,0095
17	8,3	112	8,67	-0,373	0,139	0,0089
18	9,4	120	8,86	0,542	0,294	0,0083

На основі отриманих двох сукупностей спостережень (від першого до сьомого включно і від одинадцятого до вісімнадцятого значення) побудуємо дві економетричні моделі за методом 1МНК.

1-ша модель: $\hat{Y}_1 = 1,079 + 0,138X$;

2-га модель: $\hat{Y}_2 = 6,082 + 0,023X$.

Для кожної моделі знайдемо суму квадратів залишків:

$$S_1 = \sum_{i=1}^{n_1} \left(y_{1i} - \hat{y}_{1i} \right)^2 = 0,1362 ;$$

$$S_2 = \sum_{i=1}^{n_2} \left(y_{2i} - \hat{y}_{2i} \right)^2 = 1,5366 ;$$

Знаходимо критерій R :

Так як $S_2 > S_1$, то $R = \frac{S_2}{S_1} = \frac{1,5366}{0,1362} \approx 11,28$.

Порівняємо цей критерій з табличним значенням F-критерію при ступенях свободи $k_1 = k_2 = \frac{n-c-2m}{2} = \frac{18-4-2\cdot 2}{2} = 5$ і рівні довіри $\alpha = 0,05$. Звідси, $F_{\text{òàáë}} = F(0,05;5;5) = 5,05$.

Гетероскедастичність наявна в моделі, тому що $R > F_{\text{òàáë}}$.

4. При наявності гетероскедастичності оцінку параметрів моделі виконаємо методом Ейткена:

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

4.1. Запишемо матриці змінних, які входять в оператор Ейткена:

$$X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & ... & 1 \\ x_1 & x_2 & x_3 & x_4 & x_5 & ... & x_{18} \end{pmatrix};$$

$$X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & ... & 1 \\ 24 & 25 & 27 & 27 & 29 & ... & 120 \end{pmatrix}.$$

Визначимо матрицю S^{-1}, користуючись гіпотезою: $\lambda_i = \frac{1}{x_i}$, тобто

$$S^{-1} = \begin{pmatrix} \lambda_1 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \lambda_3 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \lambda_4 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & \lambda_5 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & \lambda_{18} \end{pmatrix}$$

$$S^{-1}=\begin{pmatrix} 0,0417 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0,04 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0,037 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0,037 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & 0,0345 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0,0083 \end{pmatrix}$$

4.2. Визначимо добутки матриць:

$$X^T S^{-1}=\begin{pmatrix} 0,0417 & 0,04 & 0,037 & 0,037 & 0,0345 & ... & 0,083 \\ 1 & 1 & 1 & 1 & 1 & ... & 1 \end{pmatrix};$$

$$X^T S^{-1} X=\begin{pmatrix} 0,4235 & 18 \\ 18 & 1008 \end{pmatrix};$$

4.3. Знайдемо обернену матрицю:

$$\left(X^T S^{-1} X\right)^{-1}=\begin{pmatrix} 9,7959 & -0,1749 \\ -0,1749 & 0,0041 \end{pmatrix};$$

і вектор:

$$X^T S^{-1} Y=\begin{pmatrix} 2,4581 \\ 116,9 \end{pmatrix}.$$

4.4. Обчислимо вектор оцінок параметрів моделі:

$$\hat{A}=\left(X^T S^{-1} X\right)^{-1}\left(X^T S^{-1} Y\right)=\begin{pmatrix} 9,7959 & -0,1749 \\ -0,1749 & 0,0041 \end{pmatrix}\begin{pmatrix} 2,4581 \\ 116,9 \end{pmatrix}=\begin{pmatrix} 3,6299 \\ 0,0511 \end{pmatrix}.$$

Звідси, $\hat{a}_0 = 3,6299$, $\hat{a}_1 = 0,0511$.

Економетрична модель витрат на харчування запишеться так:

$$\hat{Y} = 3,6299 + 0,0511X \ .$$

5. Економічний аналіз характеристик економетричної моделі.

5.1. Коефіцієнт детермінації $R^2 = 1 - \frac{\left(Y - XA^*\right)^T \left(Y - XA^*\right)}{\left(Y - \overline{Y}\right)^T \left(Y - \overline{Y}\right)} = 0,867$. Це означає, що на 86,7 % варіація витрат на харчування залежить від варіації загальних затрат.

5.2. Коефіцієнт кореляції: $R = \sqrt{R^2} = \sqrt{0,867} = 0,93$ свідчить про досить тісний зв'язок витрат на харчування і загальних затрат.

5.4. Параметр моделі $\hat{a}_1 = 0,0511$ свідчить про те, що збільшення загальних затрат на 1 одиницю сприятиме граничному зростанню витрат на харчування на 0,0511 одиниць.

7. Побудова моделі з автокорельованими залишками

В економетричних дослідженнях часто зустрічаються такі випадки, коли дисперсія залишків є постійною, але спостерігається їх коваріація. Це явище має назву автокореляції залишків.

Автокореляція залишків виникає частіше за все тоді, коли економетрична модель будується на основі часових рядів. Якщо існує кореляція між послідовними значеннями деякої незалежної змінної, то буде спостерігатись і кореляція послідовних значень залишків. Тобто в цьому випадку також порушується гіпотеза, згідно з якою $M\left(u^T u\right) = \sigma_u^2 E$, але при гетероскедастичності змінюється дисперсія залишків при відсутності їх коваріації, а при автокореляції — існує коваріація залишків при незмінній дисперсії.

При автокореляції залишків, як і при гетероскедастичності дисперсія залишків запишеться:

$$M\left(u^T u\right) = \sigma_u^2 S,$$

але матриця S матиме тут зовсім інший вигляд. Запишемо цю матрицю:

$$S = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \rho^2 & \dots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \rho & \dots & \rho^{n-3} \\ \rho^3 & \rho^2 & \rho & 1 & \dots & \rho^{n-4} \\ . & . & . & . & \dots & . \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \rho^{n-4} & \dots & 1 \end{pmatrix}$$

В даній матриці параметр ρ характеризує коваріацію кожного наступного значення залишків із попереднім. Так, якщо для залишків записати авторегресійну модель першого порядку:

$$u_t = \rho u_{t-1} + \varepsilon_t ,$$

то ρ характеризує силу зв'язку величини залишків у період *t* від величини залишків у період $t-1$.

Якщо проігнорувати матрицю S при визначенні дисперсії залишків, і для оцінки параметрів моделі застосувати метод 1МНК, то можливі такі наслідки:

1) оцінки параметрів моделі можуть бути незміщеними, але неефективними, тобто вибіркові дисперсії вектора оцінок $\hat{A}$ можуть бути невиправдано великими;

2) статистичні критерії *t* і *F*- статистики, які отримані для класичної лінійної моделі, практично не можуть бути використані для дисперсійного аналізу, бо їх розрахунок не враховує наявності коваріації залишків;

3) неефективність оцінок параметрів економетричної моделі, як правило, призводить до неефективних прогнозів, тобто прогнозні значення матимуть велику вибіркову дисперсію.

Перевірка наявності автокореляції

Для перевірки наявності автокореляції залишків можна застосувати чотири методи: 1) критерій Дарбіна—Уотсона, 2) критерій фон Неймана,

3) нециклічний коефіцієнт автокореляції, 4) циклічний коефіцієнт автокореляції.

***Критерій Дарбіна—Уотсона*:** $DW = \dfrac{\sum_{t=2}^{n}(u_t - u_{t-1})^2}{\sum_{t=1}^{n} u_t^{\,2}}$

Критерій Дарбіна—Уотсона може приймати значення на множині $DW \in [0;4]$. Якщо залишки u_t є випадковими величинами, тобто не автокорельовані, то значення DW знаходиться поблизу 2. При додатній автокореляції $DW < 2$, при від'ємній $DW > 2$.

Значення критерія DW табульовані на інтервалі $(DW_l; DW_u)$, де DW_l — нижня межа, DW_u — верхня межа. Фактичні значення критерію порівнюються з табличними (критичними) для числа спостережень *n* і числа незалежних змінних m при вибраному рівні довіри α.

Для перевірки наявності автокореляції в моделі використовують таблицю критичних точок розподілу Дарбіна-Уотсона.

Розглянемо зони автокореляційного зв'язку за критерієм Дарбіна-Уотсона, подані на рисунку.

Якщо $DW \in (0; DW_l)$, залишки мають додатню автокореляцію. Якщо $DW \in (DW_u; 4 - DW_u)$, приймається гіпотеза про відсутність автокореляції. Якщо $DW \in (4 - DW_l; 4)$, приймається гіпотеза про від'ємну автокореляцію Якщо $DW \in (DW_l; DW_u) \cup (4 - DW_u; 4 - DW_l)$ конкретних висновків зробити не можна.

***Критерій фон Неймана*:** $Q = \dfrac{\sum_{t=2}^{n} (u_t - u_{t-1})^2}{\sum_{t=1}^{n} u_t^2} \cdot \dfrac{n}{n-1}$.

Звідси $Q = DW \cdot \dfrac{n}{n-1}$, при $n \to \infty$, $Q = DW$. Фактичне значення критерію фон Неймана порівнюється з табличним при вибраному рівні довіри α і заданому числі спостережень n. Якщо $Q < Q_{\text{òàáë}}$, то існує додатня автокореляція.

Оцінка параметрів моделі з автокорельованими залишками

Оцінку параметрів моделі з автокорельованими залишками можна виконувати на основі чотирьох методів: 1) Ейткена; 2) перетворення вихідної інформації; 3) Кочрена—Оркатта; 4) Дарбіна.

Перші два методи доцільно застосовувати тоді, коли залишки описуються авторегресійною моделлю першого ступеня:

$$u_t = \rho u_{t-1} + \varepsilon_t .$$

Ітеративні методи Кочрена—Оркатта і Дарбіна можна застосовувати для оцінки параметрів економетричної моделі і тоді, коли залишки описуються авторегресійною моделлю більш високого ступеня:

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \varepsilon_t;$$

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \rho_3 u_{t-3} + \varepsilon_t.$$

Метод Ейткена. Оператор оцінювання цим методом запишеться так:

$$\hat{A} = \left(X^T S^{-1} X\right)^{-1} \left(X^T S^{-1} Y\right) \text{ або}$$

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

де S^{-1} — матриця, обернена до матриці S;

V^{-1} — матриця, обернена до матриці $V = \sigma_u^2 S$.

Оскільки в матриці S коваріація залишків ρ^s при $s > 2$ наближається до нуля, то матриця, обернена до матриці S, набуде наступного вигляду:

$$S^{-1} = \frac{1}{1-\rho^2}\begin{pmatrix} 1 & -\rho & 0 & 0 & ... & 0 \\ -\rho & 1+\rho^2 & -\rho & 0 & ... & 0 \\ 0 & -\rho & 1+\rho^2 & -\rho & ... & 0 \\ 0 & 0 & -\rho & 1+\rho^2 & ... & 0 \\ . & . & . & . & ... & . \\ 0 & 0 & 0 & 0 & ... & 1+\rho^2 \end{pmatrix}.$$

На практиці для розрахунку ρ використовується співвідношення:

$$\rho \approx r = \frac{\sum_{t=2}^{n} u_t u_{t-1}}{\sum_{t=1}^{n} u_t^{\,2}} \text{ або } \rho \approx r = \frac{\sum_{t=2}^{n} u_t u_{t-1}}{\sum_{t=1}^{n} u_t^{\,2}} \cdot \frac{n}{n-1}$$

Побудова та аналіз економетричної моделі з автокорельованими залишками

Завдання 6.1. На основі двох взаємопов'язаних часових рядів про роздрібний товарообіг та доходи населення побудувати економетричну модель, що характеризує залежність роздрібного товарообігу від доходу. Вихідні дані наведені в таблиці.

Рік	Роздрібний товарообіг	Дохід
1-й	24,0	27,1
2-й	25,0	28,2
3-й	25,7	29,3
4-й	27,0	31,3

5-й	28,8	34,0
6-й	30,8	36,0
7-й	33,8	38,7
8-й	38,1	43,2
9-й	43,4	50,0
10-й	45,5	52,1

Розв'язання

1**.** Ідентифікуємо змінні моделі:

Y_t — роздрібний товарообіг у період t, залежна змінна;

X_t — дохід у період t, пояснююча змінна;

Звідси, $Y_t = f\left(X_t, u_t\right)$,

2. Специфікуємо економетричну модель у лінійній формі:

$$Y_t = a_0 + a_1 X_t + u_t;$$

$$\hat{Y}_t = \hat{a}_0 + \hat{a}_1 X_t;$$

$$u_t = Y_t - \hat{Y}_t.$$

3. Визначимо параметри моделі $\hat{a}_0$, $\hat{a}_1$ на основі методу 1МНК, припустивши, що залишки u_t некорельовані:

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

де X^T — матриця, транспонована до матриці X.

$$X = \begin{pmatrix} 1 & 27,1 \\ 1 & 28,2 \\ 1 & 29,3 \\ 1 & 31,3 \\ 1 & 34,0 \\ 1 & 36,0 \\ 1 & 38,7 \\ 1 & 43,7 \\ 1 & 50,0 \\ 1 & 52,1 \end{pmatrix}; \qquad \left(X^T X\right) = \begin{pmatrix} 10 & 370,4 \\ 370,4 & 14441,62 \end{pmatrix};$$

$$\left(X^T Y\right) = \begin{pmatrix} 322,1 \\ 12555,09 \end{pmatrix}; \qquad \left(X^T X\right)^{-1} = \begin{pmatrix} 2,0002 & -0,0513 \\ -0,0513 & 0,0014 \end{pmatrix};$$

$$\hat{A} = \left(X^T X\right)^{-1}\left(X^T Y\right) = \begin{pmatrix} 2,0002 & -0,0513 \\ -0,0513 & 0,0014 \end{pmatrix}\begin{pmatrix} 322,1 \\ 12555,09 \end{pmatrix} = \begin{pmatrix} 0,172 \\ 0,865 \end{pmatrix};$$

$\hat{a}_0 = 0,172$; $\hat{a}_1 = 0,865$.

Економетрична модель має вигляд: $\hat{Y}_t = 0,172 + 0,865 X_t$.

4. Знайдемо розрахункові значення роздрібного товарообігу на основі моделі $\hat{Y}_t = 0,172 + 0,865 X_t$ і визначимо залишки u_t.

Рік	Y_t	$\hat{Y}_t$	u_t	u_t^2	$u_t - u_{t-1}$	$(u_t - u_{t-1})^2$	$u_t u_{t-1}$

1-й	24,0	23,6123	0,3877	0,1503	—	—	—
2-й	25,0	24,5637	0,4363	0,1903	0,0485	0,0024	0,1691
3-й	25,7	25,5152	0,4848	0,0342	–0,2515	0,0632	0,0806
4-й	27,0	27,2451	–0,2451	0,0601	–0,4299	0,1848	–0,0453
5-й	28,8	29,5805	–0,7805	0,6092	–0,5354	0,2866	0,1913
6-й	30,8	31,3104	–0,5104	0,2605	0,2701	0,0729	0,3984
7-й	33,8	33,6458	0,1542	0,0238	0,6646	0,4417	–0,0787
8-й	38,1	37,9706	0,1294	0,0167	–0,0248	0,0006	0,0199
9-й	43,4	43,4199	–0,0199	0,0004	–0,1492	0,0222	–0,0026
10-й	45,5	45,2363	0,2637	0,0695	0,2836	0,0804	–0,0052
$\sum$	322,1			1,4151		1,1550	0,7276

5. Знайдемо оцінку критерію Дарбіна—Уотсона:

$$DW = \frac{\sum_{t=2}^{n} \left(u_t - u_{t-1}\right)^2}{\sum_{t=1}^{n} u_t^{\,2}} = \frac{1{,}155}{1{,}4151} = 0{,}816 .$$

Порівняємо значення критерію DW з табличним при рівні значимості $\alpha = 0{,}05$ і кількості спостережень $n = 10$. Критичні значення критерію DW у цьому випадку:

$DW_l = 0{,}879$ — нижня межа;

$DW_u = 1,320$ — верхня межа.

Оскільки $DW < DW_l$, то при $\alpha = 0,05$ можна стверджувати, що залишки u_t мають додатню автокореляцію.

Наявність чи відсутність автокореляції залишків можна також визначити на основі критерію фон Неймана.

Критерій фон Неймана: $Q = DW \cdot \frac{n}{n-1} = 0,816 \frac{10}{10-1} = 0,906$. Це значення порівнюється з табличним $Q_{\text{òàáë}} = 1,18$ при рівні значимості $\alpha = 0,05$ і кількості спостережень $n = 10$. Оскільки $Q < Q_{\text{òàáë}}$, то існує додатня автокореляція залишків.

6. Використаємо метод Ейткена для оцінки параметрів економетричної моделі з автокорельованими залишками. Оператор оцінювання запишеться так:

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

Матриця S — матриця коваріацій залишків, яка має вигляд:

$$S = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \dots & \rho^9 \\ \rho & 1 & \rho & \rho^2 & \dots & \rho^8 \\ \rho^2 & \rho & 1 & \rho & \dots & \rho^7 \\ \rho^3 & \rho^2 & \rho & 1 & \dots & \rho^6 \\ . & . & . & . & \dots & . \\ \rho^9 & \rho^8 & \rho^7 & \rho^6 & \dots & 1 \end{pmatrix}.$$

Щоб сформувати матрицю S, необхідно визначити величину ρ, яка характеризує взаємозв'язок між послідовними членами ряду залишків.

Нехай залишки описуються автокореляційною моделлю першого ступеня:

$$u_t = \rho u_{t-1} + \varepsilon_t;$$

$$\rho \approx r = \frac{\sum_{t=2}^{n} u_t u_{t-1}}{\sum_{t=1}^{n} u_t^2} \cdot \frac{n}{n-1} + \frac{m+1}{n} = \frac{0,7276}{1,4156} \cdot \frac{10}{9} + \frac{2}{10} = 0,77 .$$

Таким чином, матриця S має вигляд:

$$S = \begin{pmatrix} 1 & 0,77 & 0,59 & 0,46 & 0,35 & 0,27 & 0,21 & 0,16 & 0,12 & 0,10 \\ 0,77 & 1 & 0,77 & 0,59 & 0,46 & 0,35 & 0,27 & 0,21 & 0,16 & 0,12 \\ 0,59 & 0,77 & 1 & 0,77 & 0,59 & 0,46 & 0,35 & 0,27 & 0,21 & 0,16 \\ 0,46 & 0,59 & 0,77 & 1 & 0,77 & 0,59 & 0,46 & 0,35 & 0,27 & 0,21 \\ 0,35 & 0,46 & 0,59 & 0,77 & 1 & 0,77 & 0,59 & 0,46 & 0,35 & 0,27 \\ 0,27 & 0,35 & 0,46 & 0,59 & 0,77 & 1 & 0,77 & 0,59 & 0,46 & 0,35 \\ 0,21 & 0,27 & 0,35 & 0,46 & 0,59 & 0,77 & 1 & 0,77 & 0,59 & 0,46 \\ 0,16 & 0,21 & 0,27 & 0,35 & 0,46 & 0,59 & 0,77 & 1 & 0,77 & 0,59 \\ 0,12 & 0,16 & 0,21 & 0,27 & 0,35 & 0,46 & 0,59 & 0,77 & 1 & 0,77 \\ 0,10 & 0,12 & 0,16 & 0,21 & 0,27 & 0,35 & 0,46 & 0,59 & 0,77 & 1 \end{pmatrix}$$

7. Згідно з оператором Ейткена розрахуємо:

$$X_t^T S^{-1} = \begin{pmatrix} 0,56 & 0,13 & 0,13 & 0,13 & 0,13 & 0,13 & 0,13 & 0,13 & 0,13 & 0,56 \\ 13,21 & 3,64 & 2,07 & 2,71 & 5,72 & 3,32 & 1,57 & 1,21 & 15,40 & 33,41 \end{pmatrix}$$

$$X^T S^{-1} X = \begin{pmatrix} 2{,}163 & 82{,}289 \\ 82{,}289 & 3544{,}819 \end{pmatrix};$$

$$X^T S^{-1} Y = \begin{pmatrix} 71{,}888 \\ 3089{,}962 \end{pmatrix};$$

$$\left(X^T S^{-1} X\right)^{-1} = \begin{pmatrix} 3{,}9536 & -0{,}0918 \\ -0{,}0918 & 0{,}0024 \end{pmatrix};$$

$$\hat{A} = \left(X^T S^{-1} X\right)^{-1} \left(X^T S^{-1} Y\right) = \begin{pmatrix} 3{,}9536 & -0{,}0918 \\ -0{,}0918 & 0{,}0024 \end{pmatrix} \begin{pmatrix} 71{,}888 \\ 3089{,}962 \end{pmatrix} = \begin{pmatrix} 0{,}625 \\ 0{,}857 \end{pmatrix}$$

Звідси, $\hat{a}_0 = 0{,}625$; $\hat{a}_1 = 0{,}857$.

Таким чином, економетрична модель має вигляд:

$$\hat{Y}_t = 0{,}625 + 0{,}857 X_t$$

8. Знайдемо розрахункові значення $\hat{Y}_t$ на основі побудованої економетричної моделі та визначимо залишки.

Рік	Y_t	$\hat{Y}_t$	v_t	v_t^2	$v_t - v_{t-1}$	$\left(v_t - v_{t-1}\right)^2$	$v_t\, v_{t-1}$
1-й	24,0	23,7835	0,2165	0,0468	—	—	—
2-й	25,0	24,7310	0,2690	0,0724	0,0526	0,0028	0,0582
3-й	25,7	25,6784	0,0216	0,0005	–0,2474	0,0612	0,0058
4-й	27,0	27,4011	–0,4011	0,1608	–0,4226	0,1786	–0,0086
5-й	28,8	29,7266	–0,9266	0,8586	–0,5255	0,2762	0,3716
6-й	30,8	31,4492	–0,6492	0,4215	0,2774	0,0769	0,6016

7-й	33,8	33,7746	0,0252	0,0006	0,6745	0,4549	–0,0164
8-й	38,1	38,0813	0,0187	0,0004	–0,0066	0,00004	0,0005
9-й	43,4	43,5076	–0,1076	0,01157	–0,1262	0,0159	–0,0020
10-й	45,5	45,3163	0,1837	0,0337	0,2912	0,0848	0,9908
∑				1,6069		1,1514	0,9908

9. Розрахуємо критерій Дарбіна—Уотсона:

$$DW = \frac{\sum_{t=2}^{n}\left(u_t - u_{t-1}\right)^2}{\sum_{t=1}^{n} u_t^{\,2}} = \frac{1{,}1514}{1{,}6069} = 0{,}716$$

Порівняємо отримане значення критерію DW з табличним при рівні значимості $\alpha = 0{,}05$ і кількості спостережень $n = 10$.

Оскільки $DW < DW_l$, то при $\alpha = 0{,}05$ можна стверджувати, що залишки v_t мають додатню автокореляцію. Приходимо до висновку, що ми не звільнились від автокореляції залишків. Це означає, що вихідна гіпотеза, коли залишки описуються авторегресійною схемою першого порядку, не дотримується. Якщо залишки описуються авторегресійною схемою більш високого порядку, то доцільно виконати оцінку параметрів моделі методом Кочрена—Оркатта або Дарбіна.

8. Економетричні моделі з якісними пояснювальними змінними.

При побудові економетричних моделей зустрічаються випадки, коли поряд з факторами, які набувають кількісних значень, мають місце якісні фактори (ознаки). Прикладами якісних факторів можуть бути: стать, сімейний стан, освіта, якість продукції, зміни в економічній політиці, соціологічні опитування, релігія, страйки, війни тощо. Такі фактори в регресійних моделях характеризуються якісними змінними, або атрибутивними, які в ролі пояснювальних змінних впливають на залежну змінну. Потрібно вміти вводити якісні змінні у регресійні моделі, оцінювати їх параметри та аналізувати отримані результати.

Часто якісні змінні є бінарними: вони отримують "значення 1" при наявності певної якості і "значення 0" при їх відсутності. Такі змінні називають Dummy-змінними.

Особливістю якісних змінних є те, що вони класифікують інформацію за моделлю на декілька підгруп (категорій), що базуються на атрибутивних ознаках, і окремо працюють з кожною підгрупою.

Взаємозв'язки між атрибутивними ознаками при парної регресії можуть бути проаналізовані на підставі таблиць взаємної спряженості, де можуть наводитися частоти розподілу f_{ij} - якісної ознаки за підгрупами. За даними таблиць дається оцінка тісноти зв'язку між змінними за показниками: відхиленню χ^2 Пірсона; коефіцієнту взаємної спряженості Чупрова; коефіцієнту контингенції (асоціації).

Пропорційні теоретичні частоти показників розраховуються за формулою:

$$F_{ij} = \frac{f_{io} f_{oj}}{n}$$

де f_{io} - підсумкові частоти за ознакою x; f_{oj} - підсумкові частоти за ознакою y; n - обсяг сукупності спостережень.

Абсолютну величину відхилень частот f_{ij} від F_{ij} характеризує квадратична спряженість χ^2 Пірсона.

$$x^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{\left(f_{ij} - F_{ij} \right)}{F_{ij}}$$

Фактичне значення χ^2 порівнюється з табличним. Останнє вибирається із статистичних таблиць χ^2 в залежності від рівня значимості α та числа ступенів вільності $k = \left(m_x - 1\right)\left(m_y - 1\right)$, де m_x - число груп за ознакою x, m_y - число груп за ознакою y.

Коефіцієнт взаємної спряженості найчастіше обчислюється за формулами Чупрова

$$C = \sqrt{\frac{x^2}{n\sqrt{\left(m_x - 1\right)\left(m_y - 1\right)}}}$$

або Крамера

$$C = \sqrt{\frac{x^2}{n\sqrt{(m_{\min} - 1)}}}$$

де $m_{\min}$ - мінімальне число груп.

При $0,3 < C < 1$ існує тісний зв'язок між ознаками. В разі наявності 4-клітиної таблиці взаємної співзалежності розрахований коефіцієнт C називається коефіцієнтом контингенції або асоціації. Цей коефіцієнт зв'язаний з χ^2 функціонально:

$$\chi^2 = nC^2$$

Dummy-змінні не обов'язково приймають значення $(0;1)$. Пара $(0;1)$ може трансформуватись у будь-яку іншу пару лінійним перетворенням $y = a + bz$ $(b \neq 0)$, де a та b - константи, а $z = 0$ або $z = 1$.

Dummy-змінні можуть використовуватись у регресійних моделях у чистому вигляді або поряд з кількісними змінними. Моделі тільки з якісними змінними називаються AOV-моделями. Якщо в економетричних моделях є випадок змішаних факторів - якісних і кількісних, то такі моделі називають ACOV-моделями.

Регресійні моделі з кількісними та якісними змінними

Найпростіша лінійна регресійна модель тільки з якісними змінними має вид такої парної регресії (AOV-модель):

$$y_i = \alpha_0 + \alpha_1 d_i + \varepsilon_i$$

le y_i - залежна змінна; d_i - Dummy-змінна, яка приймає значення 0 або 1; α_0 і α_1 - параметри, які характеризують математичне сподівання залежної змінної в залежності від якісних ознак у групах таких ознак: $M\left[y_i/(d_i=1)\right]=\alpha_0+\alpha_1$ або $M\left[y_i/(d_i=0)\right]=\alpha_0+\varepsilon_1$ - залишки (випадкові величини).

Базуючись на реальних даних, проводяться розрахунки за моделлю.

Більш поширеними моделями є такі, які містять у собі сукупність кількісних і якісних пояснювальних змінних. Найпростіша ACOV-модель описується таким рівнянням парної регресії:

$$y_i=\alpha_0+\alpha_1 d_{1i}+\beta_{1i}x_{1i}+\varepsilon_i$$

де d_{1i} - dummy-змінна (0,1); x_{1i} - кількісна пояснювальна змінна; α_0 і α_1 характеристики математичного сподівання в залежності від якісних ознак.

Кількісний параметр β_{1i} розраховується за допомогою МНК.

Введення до регресії dummy-змінних має свої особливості, які полягають у наступному: одна якісна змінна відокремлює дві атрибутивні ознаки; під час інтерпретації результатів моделей dummy-змінними важливо знати, які групи позначались 1, а які 0; група, позначена 0 (нулем), розглядається як базова категорія; коефіцієнт при dummy-змінній d_{1i} називається диференційним коефіцієнтом перетину і показує, наскільки значення першої групи відрізняється від значення базової категорії (групи).

У випадку порівняння двох або більше регресійних моделей з якісними змінними їх відмінність може бути в перетинах, нахилах або в обох випадках. У загальній методології знаходження відмінностей таких моделей може бути використаний Chow-тест (Чау-тест) або підхід з використанням dummy-змінної. Chow-тест заснований на використанні F-критерію Фішера у порівнянні, а підхід з використанням dummy-змінної - на аналізі параметрів моделей, які характеризують якісні ознаки за групами спостережень.

Економічні процеси дуже часто підпорядковані сезонним коливанням: різдвяний розпродаж товарів, попит на морозиво і напої влітку і т.д. В економічному аналізі інколи виникає проблема вилучення сезонних коливань з метою виявлення тенденції. Одним з методів вилучення сезонних коливань є використання dummy-змінних в регресійних моделях.

Список використаної літератури

1. Грубер Й. Економетрія: Вступ до множинної регресії та економетрії: У 2 т. - К: Нічлава, 1998-1999.
2. Джонстон Дж. Эконометрические методы. — М.: Статистика, 1980. — 444 с.
3. Доугерти К. Введение в эконометрику: Пер. с англ. — М.: ИНФРА-М, 1997. - 402 с.
4. Дрейпер П., Смит Г. Прикладной регрессионный анализ. — М.: Финансы и статистика, 1986. — Т. 1 — 365 с; Т. 2 — 379 с.
5. Єлейко В. Основи економетрії. — Львів: "Марка Лтд", 1995. — 191с.
6. Калберг К. Бизнес-анализ с помощью Exel . Киев-Москва: Диалектика, 1997. – 448 с
7. Корольов О. А. Економетрія: Навчальний посібник — К: Європейський ун-т,2002. - 660 с.
8. Лук'яненко І. Г., Краснікова Л. І. Економетрика: Підручник. — К.: Т-во "Знання", КОО, 1998. - 494 с
9. Магнус Я. Р., Катышев П. К., Пересецкий А. А. Эконометрика: Навч. курс. - М.: Дело, 1997. - 248 с.
10. Наконечний С. І., Терещенко Т .О., Романюк Т. П. Економетрія: Навч. посіб. - К: КНЕУ, 1997. - 352 с.
11. Фишер Ф. Проблема идентификации в эконометрии. — М.: Статистика, 1978. - 224 с.
12. Хеш Д. Причинный анализ в статистических исследованиях. — М.: Финансы и статистика, 1981. — 224 с.