Валерій Єрьоменко, Андрій Алілуйко,
Катерина Березька, Олеся Мартинюк

ЕКОНОМЕТРИКА

Навчальний посібник

Тернопіль
Видавництво «Підручники і посібники»
2023

УДК 330.43
Е45

Рецензенти: **Притула Микола Миколайович** — доктор фізико-математичних наук, професор, завідувач кафедри дискретного аналізу та інтелектуальних систем Львівського національного університету імені Івана Франка

Мазко Олексій Григорович — доктор фізико-математичних наук, професор, провідний науковий співробітник Інституту математики НАН України

Федорейко Валерій Степанович — доктор технічних наук, академік Академії економічних наук України, професор Тернопільського національного педагогічного університету імені Володимира Гнатюка

Розглянуто і рекомендовано до видання вченою радою Західноукраїнського національного університету, протокол № 9 від 31.05.2023 р.

Єрьоменко В.

Е45 Економетрика. Навчальний посібник / В. Єрьоменко, А. Алілуйко, К. Березька, О. Мартинюк. — Тернопіль: Підручники і посібники, 2023. — 168 с.

У навчальному посібнику викладено основи економетрики. Особлива увага приділена класичній парній і множинній, а також узагальненій моделям лінійної регресії, класичному й узагальненому методам найменших квадратів, аналізу часових рядів. Наведено приклади розв'язування типових задач, а також дидактичні завдання для самостійної роботи.

Посібник призначено для вивчення навчальної дисципліни «Економетрика» студентами першого (бакалаврського) освітнього рівня економічних та управлінських спеціальностей вищих навчальних закладів.

УДК 303.43

Вступ

Останні десятиліття характеризуються, зокрема, стрімким розвитком наукової дисципліни економетрика. Разом із ростом числа наукових досліджень із застосуванням економетричних методів прийшло всесвітнє визнання економетрики. Свідченням цього є присудження за найбільш видатні розробки у цій галузі Нобелівських премій по економіці Р. Фрішу, і Я. Тінбергу (1969), Л. Клейку (1980), Т. Хаавельмо (1989), Дж. Хейману і Д. Макфалдену (2000).

Дослідження сучасної економічної науки зумовили нові вимоги до вищої професійної освіти економістів. У 90-х роках минулого століття економетрика стала однією із основних нормативних дисциплін у підготовці бакалаврів з економічних та управлінських спеціальностей вищих навчальних закладів.

Відсутність усталеної назви дисципліни економетрика (економетрія) характерна і для формулювання єдиного загальноприйнятого визначення її предмета. Ми будемо дотримуватися наступного означення.

Економетрика вивчає кількісні закономірності та взаємозв'язки економічних об'єктів і процесів за допомогою математичних моделей та методів економічної і математичної статистики.

Економетрика поєднує в собі економічну теорію, математичну економіку, економічну і математичну статистику. При цьому слід враховувати, що економічна теорія пропонує твердження чи гіпотези, які по суті є, як правило, якісними. Тобто, економетрика забезпечує кількісну сторону економічної теорії.

Математична економіка описує економічну теорію в математичній формі без мети вимірювання чи емпіричного підтвердження теорії. Натомість економетрист часто використовує математичні співвідношення, отримані математиком-економістом, однак перетворює їх у форму, найбільш придатну для емпіричного тестування. При цьому побудова економетричних моделей на підставі математичних рівнянь вимагає винахідливості, практичних навичок та обережності.

Суттєвою є відмінність економетрики від економічної статистики, завдання якої в основному полягає у зборі, обробці та зображення даних у вигляді таблиць та діаграм. Отримані дані використовуються економетристом. При цьому слід враховувати, що така інформація часто містить помилки вимірювання, а тому економетрист повинен розробляти спеціальні методи для аналізу подібних помилок.

Економетрика разом з тим використовує інформацію з багатьох розділів математики: теорії імовірностей, математичної статистики, диференціального числення, математичного програмування, лінійної, і зокрема матричної алгебри.

На практиці обсяг вибіркових сукупностей є дуже великим (іноді декілька тисяч). Зрозуміло, що виконувати «вручну» обчислення такого обсягу неможливо. Комп'ютерні економетричні (їх ще називають регресійними) пакети містять готові підпрограми для виконання основних етапів обчислень. У даному посібнику розглядається реалізація на ПК економетричних моделей за допомогою програмної системи EXCEL.

У посібнику вивчається одне (скалярне) рівняння регресії. Однак адекватне відображення реальних взаємозв'язків в економічних процесах досягається за рахунок розгляду систем регресійних рівнянь. Дослідження цих питань, а також ряду інших залишилося за межами посібника. Це зумовлено програмою для бакалаврату («Економетрика–1»). В університетах ФРН в магістратурі викладаються курси «Економетрика–1, 2, 3»).

При викладенні матеріалу в більшості випадків здійснюється доведення тверджень, оскільки метою посібника на думку авторів є створення базису, необхідного для поглибленого вивчення економетрики, виконання курсових і дипломних проектів, а також забезпечення студентських науково-дослідних робіт. При першому прочитанні можна пропустити доведення, зосередившись на усвідомленні ключових понять і основних тверджень.

§ 1. ПОНЯТТЯ ПРО ЕКОНОМЕТРИЧНІ МОДЕЛІ

1. ***Кореляційно-регресійний аналіз в економіці.***
2. ***Економетрична модель та її елементи.***
3. ***Інформаційна база економетричних досліджень.***
4. ***Нормальний розподіл і основні вибіркові розподіли (χ^2, Ст'юдента, Фішера-Снедокора), пов'язані з ним.***
5. ***Кореляція. Вибірковий коефіцієнт кореляції.***

1. У багатьох практично важливих задачах потрібно встановити та оцінити залежність деякого економічного показника від одного або кількох інших показників. Проте будь-які економічні показники, як правило, перебувають під впливом випадкових чинників, а тому з математичної точки зору вони інтерпретуються як випадкові величини.

З теорії ймовірностей відомо, що випадкові величини можуть бути або незалежними, або пов'язані функціонально, або ж між ними існує стохастична залежність. Строга функціональна залежність в економіці реалізується рідко. Переважно спостерігається так звана стохастична залежність, коли зміна можливих значень однієї випадкової величини (внаслідок проведення випробувань) призводить до зміни **умовного закону розподілу імовірностей** іншої (див. [4, §7]). Частковим випадком стохастичної залежності є кореляційна залежність, коли зміна можливих значень однієї величини приводить до зміни **умовного математичного сподівання** ([4, п.7.5]) іншої випадкової величини (якщо вони розглядаються на прогнозний період).

Нагадаємо про два типи кореляційної залежності двох випадкових величин. У першому випадку вони рівноправні і зв'язок між ними двосторонній. У другому випадку змінні нерівноправні, тобто є економічний сенс розглядати лише односторонній зв'язок, коли зміна тільки однієї із них приводить до зміни умовного математичного сподівання іншої. Наприклад, середня урожайність залежить від кількості внесених добрив, а не навпаки.

Покажемо, яким чином рівняння регресії, що вивчалися в математичній статистиці, приводять до економетричних моделей.

Нехай з певних економічних міркувань встановлено, що деякий економічний показник x впливає на інший показник y. Статистичні дані по кожному із них інтерпретуються як деякі реалізації (можливі значення)

випадкових величин X та Y. Тоді кореляційну залежність між ними або **залежність в середньому** можна зобразити у вигляді співвідношення

$$M(Y \mid x) = f(x), \qquad (1.1)$$

де $M(Y \mid x) = M(Y \mid X = x)$ — умовне математичне сподівання випадкової величини Y при умові, що X набирає можливе значення x. Функція $f(x)$ називається **функцією регресії** Y на X, а її графік — **лінією регресії**. При цьому X називається **незалежною змінною**, а Y — **залежною**. Розглядаючи залежність двох випадкових величин, говорять про **парну регресію**.

Залежність Y від кількох змінних, що описується рівнянням

$$M(Y \mid x_1, x_2, ..., x_m) = F(x_1, x_2, ..., x_m), \qquad (1.2)$$

називають **множинною регресією**.

Оскільки реальні (спостережені) значення залежної змінної не завжди співпадають із умовним математичним сподіванням, то ліві частини рівнянь (1.1) і (1.2) можна записати таким чином:

$$M(Y \mid x) = Y - U, \qquad M(Y \mid x_1, x_2, ..., x_m) = Y - U,$$

де U — випадкові величини. Із врахуванням цих співвідношень рівняння (1.1) і (1.2) набудуть такого виду:

$$Y = f(x) + U, \qquad (1.3)$$

$$Y = F(x_1, x_2, ..., x_m) + U, \qquad (1.4)$$

Означення. *Зв'язки між залежною та незалежною (незалежними) змінними, що описуються співвідношеннями (1.3) або (1.4) називаються* ***економетричними (регресійними) моделями (рівняннями)***. *При цьому модель (1.3) називається* ***однофакторною****, а (1.4) —* ***багатофакторною*** *(*m*-факторною).*

Виникає питання про причини обов'язкової присутності в регресійних моделях випадкової складової U, яку називають **відхиленням**, **залишком** або **збуренням**. Серед них виокремимо найістотніші.

1) **Уведення в модель не всіх пояснюючих змінних**. Будь-яка регресійна модель — це спрощення реальної ситуації. Остання завжди є складною композицією різних чинників, багато з яких у моделі не враховуються, що призводить до відхилення реальних значень залежної змінної від її модельних (розрахункових) значень. При цьому у деяких випадках заздалегідь невідомо, які чинники за умов, що склалися, насправді є визначальними, а якими можна знехтувати.

2) **Неправильний вибір функціональної форми моделі**. Визначальною особливістю моделей (1.3) та (1.4) є те, що **функції** $f(x)$ **та** $F(x_1, x_2, ..., x_m)$ **є невідомими**, а тому вони шукаються у певному класі функцій і немає ніяких гарантій, що обраний клас функцій є кращим за інші.
3) **Помилки вимірювання**. Якою б якісною не була модель, помилки вимірювання змінних впливатимуть на розбіжності між модельними та емпіричними даними.
4) **Обмеженість статистичних даних та їх випадковий характер**.
5) **Непередбачуваність людського фактора**.

2. Вище було показано, яким чином рівняння регресії приводять до економетричної моделі.

Означення. *Економетрична модель — це скалярне або векторне рівняння, яке описує кореляційно-регресійний зв'язок між економічними показниками, причому залежно від причинних зв'язків між ними один або кілька із цих показників розглядаються як залежні змінні, а інші — як незалежні.*

У загальному випадку економетрична модель має такий вигляд:

$$Y = f(x_1, x_2, ..., x_m, a_0, a_1, ..., a_k, U), \qquad (1.5)$$

де $Y = (Y_1, Y_2, ..., Y_s)'$, $f = (f_1, f_2, ..., f_s)'$, $(.)'$ — операція транспонування матриці, $Y_1, Y_2, ..., Y_s$ — залежні (пояснювані, регресанди, ендогенні) змінні, $x_1, x_2, ..., x_m$ — незалежні (пояснюючі, регресори, екзогенні) змінні, $a_0, a_1, ..., a_k$ — параметри (невідомі) моделі, U — l-вимірна випадкова величина, яка називається відхиленням (залишком флуктуації або збуренням).

Відмітимо, що у серйозних емпіричних економічних і соціальних дослідженнях вимірність s вектора Y часто буває дуже великою. Наприклад, економетричні моделі, які в теперішній час використовуються в Німеччині інститутами економічних досліджень і Федеральним банком для прогнозування кон'юнктури, містять, як правило, більше 100 рівнянь [1-2].

Означення. *Дослідження опису явища чи процесу, тобто вибір* ***конкретної*** *функції* f, *називається* ***специфікацією моделі***.

Означення. *Знаходження значень параметрів* $a_0, a_1, ..., a_k$ ***обраної*** *форми статистичного зв'язку змінних на підставі статистичних даних*

називається ***параметризацією*** *економетричної моделі або* ***оцінюванням параметрів***.

Залежно від методу, з допомогою якого здійснюється оцінювання невідомих параметрів моделі, додатково накладаються певні обмеження як на параметри, так і на випадкову складову. Такого роду обмеження також належать до специфікації моделі. Припускається, що параметри (числа) залишаються незмінними протягом усього періоду спостереження. Зміна незалежних змінних приводить модель у рух, зумовлює перехід системи до нового стану.

Зауваження. *В багатьох економетричних моделях є такі незалежні змінні, які можуть бути змінені керівними органами (державним регулюванням чи керівництвом фірмою). Ці керовані змінні можуть впливати на подальший розвиток процесу.*

Побудова якісної економетричної моделі, яка узгоджується з емпіричними даними і відповідає цілям дослідника, є досить складним процесом, для якого можна виділити наступні характерні етапи [12].

1. Початковий аналіз економічного процесу, що розглядається (його теоретичний опис з відображенням існуючих тенденцій і якісний економічний аналіз).
2. Визначення мети дослідження, досягнення якої вимагає залучення моделі; введення припущень та обмежень.
3. Вибір факторів, істотних з погляду мети дослідження (специфікація змінних).
4. Специфікація моделі.
5. Збір і попередній аналіз даних для змінних, що входять у модель.
6. Вибір методу оцінювання невідомих параметрів з урахуванням припущень про імовірнісні властивості випадкового відхилення.
7. Реалізація процедури оцінювання, яка забезпечує найкраще наближення модельних значень змінних до їхніх значень, що спостерігалися.
8. Перевірка виконання основних припущень (передумов).
9. Інтерпретація отриманих результатів, визначення їхньої адекватності поставленим цілям.
10. Аналіз неузгодженості і корегування моделі.
11. Прийняття рішень щодо наступного циклу дослідження з урахуванням альтернативних можливостей.

Основна мета дослідження економетричної моделі (1.5) — **теоретично обґрунтований і статистично надійний точковий чи інтервальний**

прогноз значень залежної змінної Y або її математичного сподівання (середньої).

Економетричні моделі можуть бути **статичними** та **динамічними**. У статичних моделях зв'язки розглядаються у фіксований момент часу і часові зміни в них ролі не відіграють. У динамічній моделі час є необхідним фактором змін.

Моделі розрізняються також за рівнем агрегування змінних (мікро, мезо- чи макроекономічні показники), за способом відображення змінних (у поточних чи постійних цінах, у абсолютних значеннях чи приростах показників), за кількістю змінних (одно- чи багатофакторні моделі), за часом спостереження (річні, квартальні чи місячні дані).

Класифікують моделі також за призначенням та метою використання (аналітичні, імітаційні, прогностичні).

3. Будь-яке економетричне дослідження поєднує теорію (якісний економічний аналіз, на основі якого будуються математичні моделі) і практику (статистичні дані). За допомогою моделей принципово описують і пояснюють процеси, що вивчаються, а статистичні дані використовуються для параметризації та обґрунтування моделей.

Економічні дані звичайно поділяються на три види: **просторові**, **часові** (динамічні) **ряди** та **перехресні** дані.

Просторовими є дані за деякими економічними показниками, які отримані для різних однотипних об'єктів (фірм, регіонів) або в один і той самий період часу, або ж часова приналежність несуттєва.

Часові ряди характеризують один і той самий об'єкт у різні рівновіддалені моменти часу. Послідовні значення часових рядів можуть бути пов'язані між собою певними залежностями: спостерігаються деякі закономірності у відхиленнях від загальної тенденції розвитку (тренду) чи виявляються часові зсуви показників (часові лаги). У зв'язку з цим методи обробки таких даних дещо відрізняються від методів, що використовуються для опрацювання просторових даних.

Перехресні ряди є поєднанням просторових та часових рядів.

Коректність висновків, які можна зробити в результаті економетричного моделювання, зумовлюється якістю вхідних даних, тобто їх повнотою та достовірністю. Тому, формуючи сукупність спостережень, слід забезпечити порівнянність даних у просторі і часі. Це означає, що дані вхідної сукупності повинні мати:

– однаковий ступінь агрегування;

– однорідну структуру одиниць сукупності;
– одні і ті ж самі методи розрахунку показників у часі чи просторі;
– однакову періодичність обліку окремих змінних;
– порівнянні ціни та однакові інші зовнішні економічні умови.

Особливої уваги заслуговує обсяг вибіркової сукупності. Результати багатьох досліджень підтверджують [16], що число спостережень повинно у 6-7 разів перевищувати число параметрів моделі. Це означає, що досліджувати парну економетричну модель з двома параметрами при наявності менше семи спостережень взагалі не має сенсу. Враховуючи, що економетричні моделі часто будуються за даними рядів динаміки, обмеженими по тривалості (10, 20, 30 років), то при виборі специфікації моделі перевагу слід віддати моделі з меншим числом параметрів.

4. Економетрика використовує, зокрема, основні поняття і методи теорії імовірностей та математичної статистики. Із усього обсягу теоретичного матеріалу цієї дисципліни виокремимо основні закони розподілу випадкових величин, якими доведеться постійно користуватися.

Неперервна випадкова величина X розподілена за **нормальним законом**, якщо її густина розподілу імовірностей має такий вигляд:

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}}. \qquad (1.6)$$

Імовірнісний зміст параметрів a і σ визначається рівностями: $a=M(X)$, $\sigma^2=D(X)$. Надалі будемо використовувати компактний запис (1.6): $X \sim N(a,\sigma)$.

Нормальний закон розподілу з параметрами $a=0$, $\sigma=1$, тобто $N(0,1)$, називається **стандартним** або **нормованим**. Використовуючи властивості математичного сподівання і дисперсії, можна переконатись у тому, що коли $X \sim N(a,\sigma)$, то $U=\frac{X-a}{\sigma} \sim N(0,1)$, тобто випадкова величина U розподілена за стандартним нормальним законом.

Нормальний закон розподілу є найбільш вивченим, тому його намагаються використовувати і при дослідженні випадкових величин, розподіл яких відмінний від нормального. Один із шляхів — розподіл досліджуваної випадкової величини замінюють приблизно нормальним. Наприклад, для **достатньо великого обсягу вибірки** середня вибіркова і вибіркова

частка для нефіксованої вибірки є нормально розподіленими із достатнім ступенем точності [5, с. 52, 58].

Інший підхід у припущенні нормального розподілу результатів спостережень полягає у побудові статистик, які мають відомі закони розподілу. **Статистикою** називається довільна функція результатів спостережень, яка не залежить від невідомих статистичних характеристик (наприклад, $\overline{x}_в$, $D_в$ тощо).

χ^2-розподіл (розподіл Пірсона). Нехай $Z_1, Z_2, ..., Z_k$ — незалежні стандартні нормальні випадкові величини ($Z_i \sim N(0,1), i = \overline{1,k}$). Тоді випадкова величина

$$\chi^2 = \sum_{i=1}^{k} Z_i^2$$

розподілена за законом χ^2 (Пірсона) із k ступенями вільності. χ^2-розподіл не містить невідомих параметрів і залежить тільки від k. При цьому $M(\chi^2) = k$, $D(\chi^2) = 2k$.

Виявляється, що деякі статистики вибірки мають χ^2-розподіл. Наприклад, якщо σ^2 — дисперсія нормального розподілу і S^2 — її незміщена оцінка із $k = n - l$ ступенями вільності (тут n — обсяг вибірки, l — число зв'язків у виразі для S^2), то статистика

$$\chi^2 = \frac{kS^2}{\sigma^2} \tag{1.7}$$

має χ^2-розподіл із k ступенями вільності [5, с.81].

Відмітимо, що поняття «число зв'язків» буде уточнене в §2 (п.4). Зокрема, якщо $X_i \sim N(a,\sigma)$ — спостережені значення ($i = \overline{1,n}$), то вибіркова (виправлена) дисперсія $S^2 = \frac{n}{n-1} D_в$ має $k = n - 1$ ступенів вільності. Відповідно статистика (1.7) має χ^2-розподіл із числом ступенів вільності $k = n - 1$.

Для $k > 30$ випадкова величина $V = \sqrt{2\chi^2} - \sqrt{2k-1}$ приблизно розподілена за стандартним нормальним законом, тому при $k > 30$ значення χ_p^2, які задовольняють рівнянню

$$P(\chi^2 \le \chi_p^2) = p \tag{1.8}$$

і називаються **квантилем рівня** p (або p**-квантилем**), можна обчислювати за наближеною формулою

$$\chi_p^2 = \frac{1}{2}\left(\sqrt{2k-1} + u_p\right)^2, \tag{1.9}$$

де u_p — корінь рівняння $\Phi(u_p) + 0{,}5 = p$, $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt$ — функція Лапласа.

Розподіл Ст'юдента (t**-розподіл)**. Нехай Z — випадкова величина, розподілена за стандартним нормальним законом, а χ^2 — незалежна від Z випадкова величина, яка має χ^2-розподіл із k ступенями вільності. Тоді розподіл випадкової величини

$$t = \frac{Z}{\sqrt{\frac{1}{k}\chi^2}} = \frac{Z}{\chi}\sqrt{k} \tag{1.10}$$

називається **розподілом Ст'юдента** або t**-розподілом** із k ступенями вільності.

При $k \to \infty$ t-розподіл наближається до стандартного нормального розподілу. При цьому практично вже при $k > 30$ t-розподіл можна вважати приблизно стандартним нормальним.

Побудуємо конкретні статистики Ст'юдента. Нехай деяка оцінка Θ^* параметра Θ лінійна за спостереженнями $X_1, X_2, ... X_n$, тобто

$$\Theta^* = c_0 + \sum_{i=1}^{n} c_i X_i .$$

Має місце таке твердження: якщо $X_i \sim N(a, \sigma)$ $\forall$ $i = \overline{1,n}$, тоді $\Theta^* \sim N\left(\Theta, \sigma_{\Theta^*}\right)$, причому

$$M\left(\Theta^*\right) = c_0 + a\sum_{i=1}^{n} c_i , \;\; \sigma_{\Theta^*}^2 = \sigma^2 \sum_{i=1}^{n} c_i^2 , \tag{1.11}$$

а згідно із (1.6)

$$Z = \frac{\Theta^* - \Theta}{\sigma_{\Theta^*}} \sim N(0,1) . \tag{1.12}$$

В якості χ^2-статистики візьмемо величину (1.7). Можна довести, що введені таким чином випадкові величини Z і χ є незалежними. А тому статистика

$$t = \frac{Z}{\chi}\sqrt{k} = \frac{\dfrac{\Theta^* - \Theta}{\sigma_{\Theta^*}}}{\sqrt{\dfrac{kS^2}{\sigma^2}}} \cdot \sqrt{k} = \frac{\Theta^* - \Theta}{\sigma_{\Theta^*}} \cdot \frac{\sigma}{S} \tag{1.13}$$

має розподіл Ст'юдента, причому згідно з другою рівністю (1.11) вона не залежить від генеральної дисперсії σ^2.

Зокрема, якщо $\Theta^* = \bar{x}$, $\Theta = a$, тоді $\sigma_{\Theta^*} = \sigma/\sqrt{n}$, і статистика (1.13) набуває такого вигляду:

$$t = \frac{\bar{x} - a}{S/\sqrt{n}}, k = n - 1. \tag{1.14}$$

Розподіл Фішера або F**-розподіл (розподіл Фішера-Снедокора)**. Нехай випадкові величини χ_1^2 і χ_2^2 мають χ^2-розподіл із k_1 і k_2 ступенями вільності відповідно. Розподіл величини

$$F(k_1, k_2) = \frac{\chi_1^2}{k_1} : \frac{\chi_2^2}{k_2} \tag{1.15}$$

називається F-розподілом або розподілом Фішера із k_1 і k_2 ступенями вільності. Із (1.15) отримується:

$$F(k_2, k_1) = \frac{1}{F(k_1, k_2)}.$$

Універсальність F-розподілу підкреслюється зв'язками з іншими розподілами. При $k_1 = 1$, $k_2 = k$ корінь квадратний величини $F(1, k)$ має розподіл Ст'юдента із k ступенями вільності. Якщо ж $k_1 = k$, $k_2 = \infty$, то має місце тотожність

$$F(k, \infty) = \frac{\chi^2(k)}{k}.$$

Розподіл Фішера відіграє фундаментальну роль у математичній статистиці і досліджувався у першу чергу як розподіл частки двох вибіркових дисперсій.

Нехай дві випадкові величини X та Y розподілені нормально за законом $N(a_1, \sigma_1)$ та $N(a_2, \sigma_2)$ відповідно, а S_1^2 та S_2^2 — незміщені оцінки із ступенями вільності k_1 і k_2 генеральних дисперсій σ_1^2 та σ_2^2. Тоді згід-

но (1.7) випадкові величини $\chi_1^2 = \frac{k_1 S_1^2}{\sigma_1^2}$ і $\chi_2^2 = \frac{k_2 S_2^2}{\sigma_2^2}$ мають χ^2-розподіл із k_1 і k_2 ступенями вільності відповідно. На підставі (1.15) статистика

$$F(k_1, k_2) = \frac{\chi_1^2}{k_1} : \frac{\chi_2^2}{k_2} = \frac{S_1^2}{\sigma_1^2} : \frac{S_2^2}{\sigma_2^2} \tag{1.16}$$

розподілена за законом Фішера із k_1 і k_2 ступенями вільності.

Зокрема, якщо розглядається одна і та ж генеральна сукупність, тобто $a_1 = a_2$, $\sigma_1^2 = \sigma_2^2$, і S_1^2 та S_2^2 — вибіркові дисперсії при обсягах вибірок n_1 і n_2 відповідно, тоді з (1.16) одержуємо:

$$F(n_1 - 1, n_2 - 1) = \frac{S_1^2}{S_2^2}. \tag{1.17}$$

5. При умові наявності стохастичного зв'язку між двома випадковими величинами зміна можливих значень однієї з них приводить до зміни умовного закону розподілу іншої. Виявлення стохастичного зв'язку і оцінювання його сили — важлива і складна задача математичної статистики.

Якщо X та Y **незалежні** випадкові величини, тоді $D(X+Y) = D(X) + D(Y)$. Тому якщо $D(X+Y) \neq D(X) + D(Y)$, то це свідчить про наявність залежності між X та Y. У випадку виконання цієї нерівності будемо вважати, що випадкові величини X та Y **корельовані**, при цьому **кореляція** тим сильніша, чим більша різниця по модулю між лівою і правою частинами нерівності.

З'ясуємо аналітичний вираз для величини кореляції між X та Y. Використавши означення дисперсії, а також властивості математичного сподівання, отримаємо:

$$\begin{aligned} D(X+Y) &= M\left[X+Y-M(X+Y)\right]^2 = M\left\{\left[X-M(X)\right]+\left[Y-M(Y)\right]\right\}^2 = \\ &= M\left\{\left[X-M(X)\right]^2 + 2\left[X-M(X)\right]\left[Y-M(Y)\right]+\left[Y-M(Y)\right]^2\right\} = \\ &= M\left[X-M(X)\right]^2 + 2M\left\{\left[X-M(X)\right]\left[Y-M(Y)\right]\right\} + M\left[Y-M(Y)\right]^2 = \\ &= D(X) + 2\operatorname{cov}(X,Y) + D(Y), \end{aligned}$$

де

$$\operatorname{cov}(X,Y) = M\left\{\left[X-M(X)\right]\left[Y-M(Y)\right]\right\}. \tag{1.18}$$

Таким чином, величина кореляції визначається величиною $\operatorname{cov}(X,Y)$, яку називають **коваріацією** (спільною варіацією) або **кореляційним моментом**. Відмітимо також ще таку формулу для обчислення коваріації:

$$\operatorname{cov}(X,Y)=M(X\cdot Y)-M(X)\cdot M(Y), \qquad (1.18^*)$$

яка отримується із (1.18) з допомогою простих перетворень.

Коваріація залежить від одиниць виміру величин X та Y, тому доцільно використовувати безрозмірну величину

$$\rho=\frac{\operatorname{cov}(X,Y)}{\sigma(X)\cdot\sigma(Y)}, \qquad (1.19)$$

яка називається **коефіцієнтом кореляції**.

Зауваження. *Якщо випадкові величини X та Y незалежні, то $\rho=0$. Проте обернене твердження неправильне: з рівності $\rho=0$ не випливає незалежність випадкових величин* (див. задачу 7.8 [4]).

Відмітимо такі властивості коефіцієнта кореляції [4, с.148]: 1) $|\rho|\le 1$; 2) якщо $\rho=\pm 1$, то між X та Y існує **лінійна функціональна** залежність ($Y=\alpha_0+\alpha_1 X$, де α_0 і α_1 — дійсні числа).

Якщо $\rho\neq 0$, то коефіцієнт кореляції своєю величиною характеризує не тільки наявність, але й силу **лінійного стохастичного** зв'язку між випадковими величинами: чим ближче $|\rho|$ до одиниці, тим сильніший лінійний зв'язок; чим ближче $|\rho|$ до нуля, тим слабший лінійний зв'язок.

Випадкові величини X та Y називаються **корельованими**, якщо $\rho\neq 0$, і **некорельованими**, якщо $\rho=0$.

Недоліком коефіцієнта кореляції є те, що при $0<|\rho|<1$ між X та Y може бути як **стохастичний**, так і **функціональний нелінійний** зв'язок.

Оцінкою **невідомого** коефіцієнта кореляції ρ є **вибірковий коефіцієнт кореляції**, який обчислюється за формулою:

$$r=\frac{\overline{xy}-\bar{x}\cdot\bar{y}}{\sigma_x\cdot\sigma_y}. \qquad (1.20)$$

Однак r є безпосередньою оцінкою ρ **лише у випадку двовимірного нормального закону розподілу** випадкових величин X та Y [4, с. 152]. Проте навіть і у цьому випадку при достатньо великому обсязі вибірки n оцінити похибку, що виникає, дуже важко. Але це і не обов'язково, тому що точне значення ρ у розрахунках практично не використовується, а потрібне лише як показник наявності кореляції між X та Y. Вибірковий коефіцієнт кореляції r застосовується в основному для перевірки статистичної гіпотези про наявність кореляції між величинами, що спостерігаються, не вдаючись у детальні оцінки сили цієї кореляції.

Випадковість вибірки може призвести до нерівності $r \neq 0$ навіть у випадку некорельованості випадкових величин. Тому для перевірки гіпотези про некорельованість величин потрібно перевірити, чи значуще r відрізняється від нуля. Відправною точкою такої перевірки є те, що при некорельованості величин ($\rho = 0$) статистика $t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ має t-розподіл Ст'юдента із $k = n-2$ ступенями вільності.

Вибірковий коефіцієнт кореляції r значущий на рівні α (гіпотеза $H_0 : \rho = 0$ відкидається), якщо

$$\left|t_{спост.}\right| = \frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}} > t(1-\alpha; n-2), \tag{1.21}$$

де $t(1-\alpha; n-2)$ — табличне значення критерія Ст'юдента для рівня значущості α при числі ступенів вільності $n-2$.

І тільки у випадку, коли $H_0 : \rho = 0$ відкидається, будуються довірчі інтервали для оцінювання невідомого коефіцієнта кореляції r.

Нарешті, відмітимо інші модифікації формули для знаходження вибіркового коефіцієнта кореляції r:

$$r = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{n\sigma_x\sigma_y}, \tag{1.20*}$$

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}. \tag{1.20**}$$

Для практичних розрахунків найбільш зручною є формула (1.20**), оскільки за нею r знаходиться безпосередньо із даних спостережень і на значення r не впливають заокруглення даних, зумовлені розрахунком середніх і відхилень від них.

§ 2. КЛАСИЧНА НОРМАЛЬНА ЛІНІЙНА МОДЕЛЬ ПАРНОЇ РЕГРЕСІЇ.

1. ***Основні передумови нормальної класичної лінійної моделі парної регресії.***
2. ***Знаходження оцінок параметрів.***
3. ***Властивості МНК-оцінок.***
4. ***Точкова оцінка дисперсії збурень. Число ступенів вільності.***
5. ***Довірчий інтервал дисперсії збурень.***
6. ***Перевірка значущості коефіцієнтів регресії.***
7. ***Довірча зона функції регресії.***
8. ***Довірчі інтервали коефіцієнтів регресії.***
9. ***Критерії якості лінійної моделі. Коефіцієнт детермінації.***
10. ***Прогнозування за класичною нормальною лінійною моделлю.***
11. ***Задача.***

1. Припустимо, що попереднім результатом специфікації моделі (1.3) є висновок про лінійну залежність між X та Y. Тоді модель (1.3) набере такого вигляду:

$$Y = \alpha_0 + \alpha_1 X + U, \qquad (2.1)$$

де α_0, α_1 — невідомі (теоретичні) детерміновані параметри, U — невідома випадкова величина (збурення).

Приклади:

1) Y — річний обсяг заощаджень родини, X — річний дохід родини;

2) рівняння Кейнса: Y — індивідуальне споживання, X — наявний прибуток, α_0 — величина автономного споживання, α_1 — гранична схильність до споживання ($0 < \alpha_1 < 1$);

3) Y — річний товарообіг однієї філії торгівельного підприємства, X — торгівельна площа цієї філії;

4) Y — валовий випуск продукції, X — вартість основних виробничих фондів підприємства.

При моделюванні різних процесів розрізняють **класичну** і **економетричну регресійні моделі**.

У класичній лінійній моделі **незалежна змінна вважається детермінованою величиною**. В цьому параграфі будемо розглядати класичну лінійну модель

$$Y = \alpha_0 + \alpha_1 x + U\,. \qquad (2.2)$$

Зауваження. *Передумова стосовно детермінованості незалежної змінної не виконується для багатьох прикладних регресійних моделей в економіці і соціології, в які часто включаються випадкові неконтрольовані величини (наприклад, ціни і кількість пропонованих товарів або товарів, що користуються попитом).*

Нехай x набирає значення $x_1, x_2, ..., x_n$, де n — обсяг вибірки. Ці статистичні дані можуть бути або просторовими, або часовими рядами, або ж перехресними рядами. Тоді із (2.2) отримується система n рівнянь

$$Y_i = \alpha_0 + \alpha_1 x_i + U_i,\ \ i = \overline{1,n}\,. \qquad (2.3)$$

Повна специфікація моделі (2.2) передбачає виконання певних умов стосовно випадкової складової правої частини (2.3).

Передумова 1. *Математичне сподівання збурень дорівнює нулю:*

$$M(U_i) = 0,\ \ i = \overline{1,n}\,. \qquad (2.4)$$

Ця передумова означає, що збурення **в середньому** не здійснюють на Y ніякого впливу. Справді, за властивостями математичного сподівання для $i = \overline{1,n}$:

$$M(Y_i) = M(\alpha_0 + \alpha_1 x_i + U_i) = \alpha_0 + \alpha_1 x_i + M(U_i) = \alpha_0 + \alpha_1 x_i\,.$$

Передумова 2. *Збурення мають однакову дисперсію:*

$$D(U_i) = \sigma_u^2,\ \ i = \overline{1,n}, \qquad (2.5)$$

де σ_u^2 — **невідоме число**, яке підлягає оцінюванню.

Якщо виконуються рівності (2.5), то говорять, що збурення **гомоскедастичні**, якщо ж ні — то збурення **гетероскедастичні**. Наведені терміни зумовлені тим, що функція $g(x) = D(Y \mid x)$ називається **функцією скедастичності**.

Передумова 3. *Збурення U_i і U_j при $i \neq j$ не корелюють між собою:*

$$\operatorname{cov}(U_i, U_j) = 0\ \ \forall\ i, j = \overline{1,n},\ \ i \neq j\,. \qquad (2.6)$$

Враховуючи рівності (2.4) і формулу (1.18*), із (2.6) отримаємо рівності:

$$\operatorname{cov}(U_i, U_j) = M(U_i \cdot U_j) - M(U_i)M(U_j) = M(U_i \cdot U_j) = 0\,,$$

тобто

$$M(U_i \cdot U_j) = 0,\ \ i \neq j,\ \ i, j = \overline{1,n}\,. \qquad (2.7)$$

Передумова 4. *Випадкові збурення U_i, $i = \overline{1,n}$, розподілені за нормальним законом.*

Згідно із передумовами 1, 2: $U_i \sim N(0,\sigma_u)$, $i=\overline{1,n}$. Наслідком передумов 1-4 є нормальність розподілу випадкових величин Y_i:

$$Y_i \sim N(\alpha_0+\alpha_1 x_i,\sigma_u),\ \ i=\overline{1,n}. \qquad (2.8)$$

Зауваження. *Передумови 1-4 стосовно моделі (2.3) разом визначають* ***класичну нормальну лінійну модель регресії****. Якщо ж передумова 4 не виконується, то має місце* ***класична лінійна модель.***

2. Перший крок на шляху дослідження моделі (2.2) полягає у параметризації цієї моделі, тобто, щоб за **конкретною** вибіркою $\{(x_i, y_i), i=\overline{1,n}\}$ обсягом n знайти такі значення оцінок невідомих параметрів α_0, α_1, для яких побудована пряма регресії була б найкращою в певному сенсі серед усіх інших прямих.

Нехай a_0 та a_1 — оцінки невідомих параметрів α_0 та α_1 відповідно. Тоді оцінкою моделі (2.3) за вибіркою є n рівнянь

$$y_i = a_0 + a_1 x_i + u_i,\ \ i=\overline{1,n}, \qquad (2.9)$$

або

$$y_i = \hat{y}_i + u_i,\ \ i=\overline{1,n}, \qquad (2.10)$$

де $\hat{y}_i$ — **групова середня**, або **умовна середня**, знайдена за рівнянням регресії

$$\hat{y} = a_0 + a_1 x, \qquad (2.11)$$

тобто $\hat{y}_i = a_0 + a_1 x_i$, u_i — **вибіркова оцінка збурення** U_i або **залишок регресії**, або ж **відхилення**.

Згідно із (2.8) $\hat{y}_i$ можна назвати також **розрахунковим** або **оціночним значенням** Y_i при $x=x_i$, оскільки $M(Y_i)=\alpha_0+\alpha_1 x_i$, а a_0 та a_1 — оцінки α_0 та α_1 відповідно.

Мірою якості оцінок a_0, a_1 можуть бути визначені композиції відхилень $u_i = y_i - \hat{y}_i$ [3, 6, 7-10]. Найпоширенішим і теоретично обґрунтованим є метод, при якому мінімізується $\sum_{i=1}^{n} u_i^2$, і який має назву **метод найменших квадратів** (МНК). Перевагами його є оптимальні властивості оцінок (незміщеність, ефективність, спроможність), а також зручність з обчислювальної точки зору.

Використаємо МНК для знаходження a_0, a_1. Квадратична функція змінних a_0, a_1

$$Q(a_0, a_1) = \sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} (a_0 + a_1 x_i - y_i)^2$$

є неперервною і обмеженою знизу ($Q \geq 0$), тому має мінімум. Необхідною умовою існування екстремуму неперервно диференційованої функції двох змінних є рівність нулю її частинних похідних:

$$\begin{cases} \dfrac{\partial Q}{\partial a_0} = 2\sum\limits_{i=1}^{n} (a_0 + a_1 x_i - y_i) = 0, \\ \dfrac{\partial Q}{\partial a_1} = 2\sum\limits_{i=1}^{n} (a_0 + a_1 x_i - y_i) x_i = 0. \end{cases}$$

Цю систему рівнянь можна записати в такому вигляді:

$$\begin{cases} n a_0 + \left(\sum\limits_{i=1}^{n} x_i \right) a_1 = \sum\limits_{i=1}^{n} y_i, \\ \left(\sum\limits_{i=1}^{n} x_i \right) a_0 + \left(\sum\limits_{i=1}^{n} x_i^2 \right) a_1 = \sum\limits_{i=1}^{n} x_i y_i. \end{cases}$$

Поділимо кожне із рівнянь на n:

$$\begin{cases} a_0 + \bar{x} a_1 = \bar{y}, \\ \bar{x} a_0 + \overline{x^2} a_1 = \overline{xy}, \end{cases} \tag{2.12}$$

де $\bar{x} = \sum x_i / n$, $\bar{y} = \sum y_i / n$, $\overline{xy} = \sum x_i y_i / n$, $\overline{x^2} = \sum x_i^2 / n$.

Використавши формулу Крамера, остаточно отримаємо єдиний розв'язок системи (2.12):

$$a_1 = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2}, \quad a_0 = \bar{y} - a_1 \bar{x}. \tag{2.13}$$

Отже, функція $Q(a_0, a_1)$ має єдину критичну точку. Виявляється [5, с.166], що в цій точці виконується і достатня умова існування мінімуму.

Оскільки оцінки (2.13) знайдені з допомогою МНК, то їх називають **МНК-оцінками**. Система (2.12) називається **системою нормальних рівнянь**.

3. Для **нефіксованої** вибірки обсягом n згідно із (2.3) результуюча ознака Y набирає значення $Y_1, Y_2, ..., Y_n$, які є випадковими величинами. Тоді за формулами (2.13)

$$a_1 = \frac{\overline{xY} - \bar{x}\cdot\bar{Y}}{\overline{x^2} - \left(\bar{x}\right)^2}, \quad a_0 = \bar{Y} - a_1\bar{x}, \tag{2.13*}$$

тобто **ці оцінки є** також **випадковими величинами**.

Необхідно з'ясувати якість цих статистичних оцінок.

Властивість 1. *МНК-оцінки* a_0, a_1 *є лінійними комбінаціями спостережень* $Y_1, Y_2, \ldots Y_n$:

$$a_1 = \sum_{i=1}^{n} \lambda_i Y_i, \quad a_0 = \sum_{i=1}^{n} \mu_i Y_i, \tag{2.14}$$

де

$$\lambda_i = \frac{x_i - \bar{x}}{n\sigma_x^2}, \quad \mu_i = \frac{1}{n} - \lambda_i\bar{x}, \quad i = \overline{1,n}, \quad \sigma_x^2 = \sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \Big/ n. \tag{2.15}$$

□ Справді,

$$a_1 = \frac{\overline{xY} - \bar{x}\cdot\bar{Y}}{\overline{x^2} - \left(\bar{x}\right)^2} = \frac{1}{n\sigma_x^2}\left(\sum_{i=1}^{n} x_i Y_i - \bar{x}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n\sigma_x^2}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)Y_i = \sum_{i=1}^{n}\frac{x_i - \bar{x}}{n\sigma_x^2}Y_i,$$

$$a_0 = \bar{Y} - a_1\bar{x} = \frac{1}{n}\sum_{i=1}^{n} Y_i - \bar{x}\sum_{i=1}^{n}\frac{x_i - \bar{x}}{n\sigma_x^2}Y_i = \sum_{i=1}^{n}\left(\frac{1}{n} - \lambda_i\bar{x}\right)Y_i. \qquad \blacksquare$$

Зауваження. *Вагові коефіцієнти* λ_i *та* μ_i *залишаються незмінними при переході від вибірки до вибірки. Вони задовольняють такі співвідношення:*

$$\sum_{i=1}^{n}\lambda_i = 0, \ \sum_{i=1}^{n}\lambda_i x_i = 1, \ \sum_{i=1}^{n}\lambda_i^2 = \frac{1}{n\sigma_x^2}, \ \sum_{i=1}^{n}\mu_i = 1, \ \sum_{i=1}^{n}\mu_i x_i = 0, \ \sum_{i=1}^{n}\mu_i^2 = \frac{\overline{x^2}}{n\sigma_x^2}. \tag{2.16}$$

Пропонується самостійно перевірити правильність цих співвідношень.

Властивість 2. *МНК-оцінки* a_0, a_1 *є незміщеними оцінками відповідних параметрів* α_0, α_1.

□ Використавши (2.3), (2.4), (2.14), (2.16) і властивості математичного сподівання, отримаємо:

$$M(a_1) = M\left(\sum_{i=1}^{n}\lambda_i Y_i\right) = \sum_{i=1}^{n}\lambda_i M(Y_i) = \sum_{i=1}^{n}\lambda_i(\alpha_0 + \alpha_1 x_i) = \alpha_0\sum_{i=1}^{n}\lambda_i + \alpha_1\sum_{i=1}^{n}\lambda_i x_i = \alpha_1,$$

$$M(a_0) = M\left(\sum_{i=1}^{n}\mu_i Y_i\right) = \sum_{i=1}^{n}\mu_i M(Y_i) = \alpha_0\sum_{i=1}^{n}\mu_i + \alpha_1\sum_{i=1}^{n}\mu_i x_i = \alpha_0,$$

тобто

$$M(a_1)=\alpha_1,\quad M(a_0)=\alpha_0. \qquad \blacksquare \quad (2.17)$$

Властивість 3 (теорема Гаусса-Маркова). *Із усіх лінійних незміщених оцінок параметрів α_0, α_1 тільки МНК-оцінки a_0, a_1 є ефективними і, отже, найкращими лінійними незміщеними оцінками.*

$\square$ Обчислимо спочатку $D(a_1)$, $D(a_0)$, $\operatorname{cov}(a_0,a_1)$. Для цього встановимо некорельованість випадкових величин Y_i та Y_j для $i\neq j$, $i,j=\overline{1,n}$, використавши (2.3), (2.4) та (2.7) і детермінованість α_0, α_1 та x_i:

$$\operatorname{cov}(Y_i,Y_j)=M(Y_i\cdot Y_j)-M(Y_i)M(Y_j)=$$

$$=M[(\alpha_0+\alpha_1x_i+U_i)(\alpha_0+\alpha_1x_j+U_j)]-M(\alpha_0+\alpha_1x_i+U_i)M(\alpha_0+\alpha_1x_j+U_j)=$$

$$=M[(\alpha_0+\alpha_1x_i)(\alpha_0+\alpha_1x_j)]+M[(\alpha_0+\alpha_1x_i)U_j]+M[(\alpha_0+\alpha_1x_j)U_i]+$$

$$+M(U_i\cdot U_j)-(\alpha_0+\alpha_1x_i)(\alpha_0+\alpha_1x_j)=0.$$

А тому із (2.14)-(2.16) та (2.8) отримаємо:

$$D(a_1)=D\left(\sum_{i=1}^n\lambda_iY_i\right)=\sum_{i=1}^n\lambda_i^2D(Y_i)=\sum_{i=1}^n\lambda_i^2\sigma_u^2=\sigma_u^2\sum_{i=1}^n\lambda_i^2=\frac{\sigma_u^2}{n\sigma_x^2}; \qquad (2.18)$$

$$D(a_0)=D\left(\sum_{i=1}^n\mu_iY_i\right)=\sum_{i=1}^n\mu_i^2D(Y_i)=\sum_{i=1}^n\mu_i^2\sigma_u^2=\sigma_u^2\sum_{i=1}^n\mu_i^2=\frac{\sigma_u^2\overline{x^2}}{n\sigma_x^2}; \quad (2.19)$$

$$\operatorname{cov}(a_0,a_1)=M\{[a_0-M(a_0)][a_1-M(a_1)]\}=$$

$$=M\left[\left(\sum_{i=1}^n\mu_iY_i-\alpha_0\right)\left(\sum_{i=1}^n\lambda_iY_i-\alpha_1\right)\right]=$$

$$=M\left\{\left[\sum_{i=1}^n\mu_i(\alpha_0+\alpha_1x_i+U_i)-\alpha_0\right]\left[\sum_{i=1}^n\lambda_i(\alpha_0+\alpha_1x_i+U_i)-\alpha_1\right]\right\}=$$

$$=M\left[\left(\alpha_0\sum_{i=1}^n\mu_i+\alpha_1\sum_{i=1}^n\mu_ix_i+\sum_{i=1}^n\mu_iU_i-\alpha_0\right)\left(\alpha_0\sum_{i=1}^n\lambda_i+\alpha_1\sum_{i=1}^n\lambda_ix_i+\sum_{i=1}^n\lambda_iU_i-\alpha_1\right)\right]=$$

$$=M\left[\left(\sum_{i=1}^n\mu_iU_i\right)\left(\sum_{i=1}^n\lambda_iU_i\right)\right]=M\left(\sum_{i=1}^n\mu_i\lambda_iU_i^2+\sum_{\substack{i,j=1\\ i\neq j}}^n\mu_i\lambda_jU_iU_j\right)=$$

$$=\sum_{i=1}^n\mu_i\lambda_iM(U_i^2)+\sum_{\substack{i,j=1\\ i\neq j}}^n\mu_i\lambda_jM(U_iU_j)=\sigma_u^2\sum_{i=1}^n\left(\frac{1}{n}-\lambda_i\bar{x}\right)\lambda_i=$$

$$= \sigma_u^2 \left(\frac{1}{n} \sum_{i=1}^{n} \lambda_i - \bar{x} \sum_{i=1}^{n} \lambda_i^2 \right) = -\frac{\sigma_u^2 \bar{x}}{n\sigma_x^2} \neq 0.$$

Отже, оцінки a_0 і a_1 корелюють між собою, при цьому **коваріаційна матриця** цих оцінок має такий вигляд:

$$\sum_{a_0, a_1} = \begin{pmatrix} \sigma_{a_0}^2 & \sigma_{a_0, a_1} \\ \sigma_{a_1, a_0} & \sigma_{a_1}^2 \end{pmatrix} = \begin{pmatrix} \frac{\sigma_u^2 \overline{x^2}}{n\sigma_x^2} & -\frac{\sigma_u^2 \bar{x}}{n\sigma_x^2} \\ -\frac{\sigma_u^2 \bar{x}}{n\sigma_x^2} & \frac{\sigma_u^2}{n\sigma_x^2} \end{pmatrix}$$

або більш компактно:

$$\sum_{a_0, a_1} = \frac{\sigma_u^2}{n\sigma_x^2} \begin{pmatrix} \overline{x^2} & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}.$$

Зауваження. *Символи* $\sigma_{a_0}^2$ *та* σ_{a_0, a_1} *позначають* $D(a_0)$ *і* $\operatorname{cov}(a_0, a_1)$ *відповідно. Натомість символ* $\sum_{a_0, a_1}$ *(велика літера «сигма» грецького алфавіту) позначає* ***дисперсійно-коваріаційну матрицю*** *(коротше:* ***коваріаційну*** *матрицю) двовимірної випадкової величини* (a_0, a_1), ***втрачаючи значення*** *символу* ***суми***.

Розглянемо тепер довільні інші лінійні оцінки $\hat{a}_0$, $\hat{a}_1$ (відмінні від a_0, a_1) параметрів α_0, α_1 такі, що

$$\hat{a}_0 = \sum_{i=1}^{n} B_i Y_i, \quad M(\hat{a}_0) = \alpha_0, \quad \hat{a}_1 = \sum_{i=1}^{n} C_i Y_i, \quad M(\hat{a}_1) = \alpha_1.$$

Згідно із формулюванням властивості 3, потрібно довести, що $D(\hat{a}_0) > D(a_0)$, $D(\hat{a}_1) > D(a_1)$, тобто ефективність оцінок a_0, a_1.

Доведемо, наприклад, другу нерівність. Нехай

$$C_i = \lambda_i + d_i, \quad \sum_{i=1}^{n} d_i^2 > 0,$$

де d_i — зсув ваги λ_i, визначеної (2.15), а друга нерівність зумовлена тим, що $a_1 \neq \hat{a}_1$.

Із лінійності оцінки $\hat{a}_1$, співвідношень (2.3) та (2.4) отримуємо:

$$M(\hat{a}_1) = M\left(\sum_{i=1}^{n} C_i Y_i \right) = \sum_{i=1}^{n} C_i M(Y_i) = \sum_{i=1}^{n} C_i M(\alpha_0 + \alpha_1 x_i + U_i) =$$

$$= \alpha_0 \sum_{i=1}^{n} C_i + \alpha_1 \sum_{i=1}^{n} C_i x_i \,.$$

З другого боку, незміщеність $\hat{a}_1$ дає рівність $M(\hat{a}_1) = \alpha_1$, тому повинні виконуватись рівності

$$\sum_{i=1}^{n} C_i = 0, \quad \sum_{i=1}^{n} C_i x_1 = 1 \quad \text{або} \quad \sum_{i=1}^{n} (\lambda_i + d_i) = 0, \quad \sum_{i=1}^{n} (\lambda_i + d_i) x_i = 1,$$

звідки з урахуванням перших двох формул (2.16) отримуємо необхідні умови для зсувів d_i:

$$\sum_{i=1}^{n} d_i = 0, \quad \sum_{i=1}^{n} d_i \, x_i = 0. \tag{2.20}$$

За аналогією з обчисленням $D(a_1)$:

$$D(\hat{a}_1) = \sigma_u^2 \sum_{i=1}^{n} C_i^2 = \sigma_u^2 \sum_{i=1}^{n} (\lambda_i + d_i)^2 = \sigma_u^2 \left(\sum_{i=1}^{n} \lambda_i^2 + 2 \sum_{i=1}^{n} \lambda_i d_i + \sum_{i=1}^{n} d_i^2 \right).$$

Але згідно із першою формулою (2.15) та (2.20):

$$\sum_{i=1}^{n} \lambda_i d_i = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{n\sigma_x^2} d_i = \frac{1}{n\sigma_x^2} \left(\sum_{i=1}^{n} d_i x_i - \bar{x} \sum_{i=1}^{n} d_i \right) = 0,$$

а відповідно до (2.18):

$$D(a_1) = \sigma_u^2 \sum_{i=1}^{n} \lambda_i^2 = \frac{\sigma_u^2}{n\sigma_x^2}.$$

Тому остаточно:

$$D(\hat{a}_1) = D(a_1) + \sum_{i=1}^{n} d_i^2 > D(a_1),$$

що свідчить про ефективність оцінки a_1.

Аналогічно доводиться ефективність оцінки a_0. ▣

Властивість 4. *МНК-оцінки є спроможними оцінками.*

□ Нагадаємо [5, с.51], що оцінка Θ^* називається спроможною оцінкою параметра Θ, якщо для як завгодно малого $\varepsilon > 0$ має місце граничний перехід

$$\lim_{n \to \infty} P\left(\left| \Theta - \Theta^* \right| \le \varepsilon \right) = 1.$$

Згідно із нерівністю Чебишева: $P\left(\left| Z - M(Z) \right| \le \varepsilon \right) \ge 1 - \dfrac{D(Z)}{\varepsilon^2}$.

Для випадкової величини a_1 виконуються рівності $M(a_1)=\alpha_1$, $D(a_1)=\dfrac{\sigma_u^2}{n\sigma_x^2}$. Тому $D(a_1)\to 0$ при $n\to\infty$, тобто $P(|a_1-\alpha_1|\le\varepsilon)\to 1$ при $n\to\infty$ для як завгодно малого $\varepsilon>0$.

Аналогічно доводиться спроможність оцінки a_0. ▣

Висновки. *При встановленні властивостей статистичних МНК-оцінок параметрів регресії в рамках класичної лінійної регресії були використані чотири умови стосовно випадкового збурення U, які часто називаються* ***умовами Гаусса-Маркова****, а саме — передумови 1-3 і умова* $\operatorname{cov}(U_i,x_i)=0\ \ \forall\ \ i=\overline{1,n}$*, яка виконується, оскільки за припущенням x є детермінованою змінною.*

При невиконанні хоча б однієї з цих умов МНК-оцінки втрачають бажані властивості. Подолання складностей, які при цьому виникають, а також одержання більш надійних результатів — найважливіші задачі економетричного моделювання.

4. Вище були знайдені **теоретичні** числові характеристики МНК-оцінок. Для здійснення аналізу побудованої моделі (2.9), яка є оцінкою теоретичної моделі (2.3), потрібно знайти незміщені оцінки дисперсій (2.18), (2.19) для a_1 та a_0 відповідно. Кожна з цих дисперсій містить невідомий співмножник σ_u^2. Тому необхідно знайти незміщену точкову оцінку σ_u^2.

Вихідною інформацією для оцінювання σ_u^2 є значення $u_i=y_i-\hat{y}_i$, $i=\overline{1,n}$, тобто $\overline{u^2}-(\bar{u})^2$ — можлива оцінка σ_u^2. Але

$$\bar{u}=\sum_{i=1}^{n}u_i/n=0, \tag{2.21}$$

оскільки

$$\sum_{i=1}^{n}u_i=\sum_{i=1}^{n}(y_i-a_0-a_1x_i)=\frac{1}{2}\frac{\partial Q}{\partial a_0}=0.$$

Відмітимо, що **рівність** (2.21) **відповідає передумові 1** і **вона виконується**, взагалі кажучи, **тільки у випадку** $\alpha_0\neq 0$ (пропонується самостійно переконатися в цьому).

Отже, в оцінці σ_u^2 фігуруватиме тільки складова $\overline{u^2}$. Наступний крок — відкоригувати цю оцінку з тим, щоб вона стала незміщеною. Можна довести [5, с.207], що $M\left(\sum_{i=1}^{n} u_i^2\right) = \sigma_u^2(n-2)$.

Тому

$$S_u^2 = \frac{1}{n-2}\sum_{i=1}^{n} u_i^2 \quad \left(= \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right) \tag{2.22}$$

є **незміщеною оцінкою** невідомої σ_u^2.

Звернемо увагу на знаменник у (2.22). Виникає питання, як його **пояснити**, враховуючи, що незміщеною оцінкою (виправленою дисперсією) невідомої генеральної дисперсії кількісної ознаки X є

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2, \tag{2.23}$$

тобто знаменник у цьому випадку вже дорівнює $n-1$. Справа у тому, що для знаходження (2.23) вхідною інформацією є n чисел $x_1, x_2, ..., x_n$. Але серед чисел $x_1 - \bar{x}, x_2 - \bar{x}, ..., x_n - \bar{x}$ незалежними є тільки $n-1$, оскільки $\sum_{i=1}^{n}\left(x_i - \bar{x}\right) = 0$, тобто одне із чисел $x_1, x_2, ..., x_n$ можна виразити з допомогою останньої рівності через решту $n-1$ чисел.

У випадку (2.22) вхідною інформацією є n чисел $y_1, y_2, ..., y_n$, які задовольняють двом рівностям (2.14). Тепер уже два числа можна виразити через решту, використавши ці рівності, тобто лінійно незалежними є вже $n-2$ чисел.

Число l (для (2.22) $l=2$, а для (2.23) $l=1$) називається **числом зв'язків**, накладених на вибірку, при знаходженні дисперсії.

Число $k = n - l$ називається **числом ступенів вільності**.

5. Згідно із (1.7) та (2.22) випадкова величина

$$\chi^2 = \frac{(n-2)S_u^2}{\sigma_u^2} \tag{2.24}$$

розподілена за законом χ^2 із числом ступенів вільності $k = n-2$. Оскільки розподіл χ^2 несиметричний [4, с.126], то це приводить до несиметричності і довірчого інтервалу для випадкової величини, розподіленої за цим законом.

Для побудови довірчого інтервалу $\left(\chi_1^2, \chi_2^2\right)$, в який з імовірністю γ потрапить можливе значення χ^2, виберемо його межі таким чином, щоб

$$P\left(\chi^2 < \chi_1^2\right) = P\left(\chi^2 > \chi_2^2\right) = (1-\gamma)/2. \qquad (2.25)$$

Для цього значення χ_1^2 і χ_2^2 знайдемо за табл. 4 додатків, використовуючи рівняння

$$P\left(\chi^2(k) > \chi_1^2(p;k)\right) = p, \qquad p = (1+\gamma)/2, \qquad (2.26)$$

$$P\left(\chi^2(k) > \chi_2^2(p;k)\right) = p, \qquad p = (1-\gamma)/2. \qquad (2.27)$$

Із співвідношень (2.26), (2.27) з використанням протилежних подій і того, що для можливого значення χ_1^2 $P\left(\chi^2 = \chi_1^2\right) = 0$, випливає виконання рівностей (2.25).

Отже, за побудовою подвійна нерівність

$$\chi_1^2 < \frac{(n-2)S_u^2}{\sigma_u^2} < \chi_2^2 \qquad (2.28)$$

виконується з імовірністю γ.

З першої нерівності (2.28) отримаємо рівносильну нерівність $\sigma_u^2 < \frac{(n-2)S_u^2}{\chi_1^2}$, а з другої: $\sigma_u^2 > \frac{(n-2)S_u^2}{\chi_2^2}$, об'єднання яких дає шуканий довірчий інтервал для оцінки σ_u^2:

$$\frac{(n-2)S_u^2}{\chi_2^2} < \sigma_u^2 < \frac{(n-2)S_u^2}{\chi_1^2}. \qquad (2.29)$$

З урахуванням (2.22) цей інтервал можна записати у такому вигляді:

$$\frac{\sum u_i^2}{\chi_2^2} < \sigma_u^2 < \frac{\sum u_i^2}{\chi_1^2}. \qquad (2.29^*)$$

6. Вище були знайдені точкові МНК-оцінки невідомих параметрів регресії α_0, α_1, а також їх числові характеристики. На перший погляд, наступний крок мав би полягати у побудові інтервальних оцінок для цих параметрів. Проте в полі зору необхідно тримати **правомірність** самої **специфікації вихідної моделі** (2.2), тобто теоретично можливі такі варіанти:
1) $\alpha_0 \neq 0$, $\alpha_1 \neq 0$; 2) $\alpha_0 = 0$, $\alpha_1 \neq 0$; 3) $\alpha_0 \neq 0$, $\alpha_1 = 0$; 4) $\alpha_0 = \alpha_1 = 0$.
При цьому **безпосередньо** на підставі вибірки неможливо діагностувати кожний із цих варіантів, оскільки вибірка організовується випадковим чи-

ном. У той же час немає сенсу будувати довірчі інтервали для α_1 у випадках 3 та 4, а для α_0 — у випадках 2 та 4.

Відмітимо важливість діагностування рівності $\alpha_1 = 0$. У цьому випадку взагалі слід переглянути висунуту модель. Разом з тим, якщо $\alpha_0 = 0$ (при $\alpha_1 \neq 0$), то це, як буде встановлено нижче, необхідно враховувати при розгляді питання про якість моделі.

Отже, актуальним є питання про **з'ясування значущості** коефіцієнтів регресії α_0 і α_1 на підставі наявних статистичних даних.

Висунемо статистичну нульову гіпотезу $H_0 : \alpha_m = 0$, де $m = 0;1$. Альтернативна гіпотеза $H_1 : \alpha_m \neq 0$. Задамо також рівень значущості α.

Згідно із (2.17) і (2.14) МНК-оцінки a_m є незміщеними і лінійними відносно випадкових величин $Y_1, Y_2, ..., Y_n$, які у відповідності із (2.8) розподілені нормально ($Y_i \sim N(\alpha_0 + \alpha_1 x_i, \sigma_u)$, $i = \overline{1,n}$). Тому і a_m мають нормальний розподіл:

$$a_m \sim N\left(\alpha_m, \sigma_{a_m}\right), \quad m = 0;1,$$

де $\sigma^2_{a_m}$ визначаються формулами (2.18), (2.19). Незміщеною оцінкою σ^2_u є S^2_u (формула (2.22)). Тому з урахуванням (2.18), (2.19) незміщені оцінки $\sigma^2_{a_m}$ (вони невідомі, бо невідомим є σ^2_u!) визначаються формулами

$$S^2_{a_0} = \frac{S^2_u \overline{x^2}}{n\sigma^2_x}, \quad S^2_{a_1} = \frac{S^2_u}{n\sigma^2_x}. \tag{2.30}$$

Розглянемо випадкову величину (1.13), де $Z = Z_m = \dfrac{a_m - \alpha_m}{\sigma_{a_m}}$, $k = n - 2$, $\chi = \sqrt{\chi^2}$, χ^2 визначається за формулою (2.24), тобто

$$t_m = \frac{a_m - \alpha_m}{\sigma_{a_m}} \cdot \frac{\sigma_u}{S_u}, \; m = 0;1, \tag{2.31}$$

має t-розподіл із $k = n - 2$ ступенями вільності і вже **не залежить від** σ_u.

Враховуючи (2.18), (2.19) та (2.30), із (2.31) отримаємо такі t-розподіли:

$$t_0 = \frac{a_0 - \alpha_0}{S_{a_0}}, \quad t_1 = \frac{a_1 - \alpha_1}{S_{a_1}}. \tag{2.32}$$

Зміст основної гіпотези $H_0(\alpha_m = 0)$ та альтернативної $H_1(\alpha_m \neq 0)$ дозволяє сформулювати **двосторонній критерій значущості оцінок** a_0, a_1:

якщо виконується нерівність

$$\left|\frac{a_m}{S_{a_m}}\right| > t_{кр.}, \tag{2.33}$$

де $t_{кр.} = t_{двост.кр.}(\alpha, n-2)$, $\alpha = 1-\gamma$, *— критична точка розподілу Ст'юдента (табл. 3 додатків), тоді на рівні значущості* α *приймається гіпотеза* H_1, *тобто вважається, що* $\alpha_m \neq 0$.

7. Нехай нерівність (2.33) виконується для $m = 1$, що означає існування стохастичного зв'язку між змінними Y та x.

Побудуємо **довірчий інтервал для функції регресії**, тобто для умовного математичного сподівання $M(Y \mid x)$, який із заданою надійністю γ покриває невідоме значення $\alpha_0 + \alpha_1 x$.

Знайдемо дисперсію величини $\hat{y}$, яка є статистичною оцінкою $M(Y \mid x)$. Для цього емпіричне рівняння регресії (2.11) запишемо у такому вигляді

$$\hat{y} = \bar{y} + a_1(x - \bar{x}), \tag{2.34}$$

підставивши другу рівність (2.13) в (2.11). Перевагою рівняння (2.34) є некорельованість його доданків справа. Доведення рівності $\operatorname{cov}(\bar{y}, a_1) = 0$ (за припущенням $x - \bar{x}$ є детермінованою величиною) здійснюється з використанням (2.14), (2.4) і (2.16). Тому дисперсія лівої частини (2.34) дорівнює сумі дисперсій двох **некорельованих** доданків правої:

$$D(\hat{y}) = D(\bar{y}) + (x - \bar{x})^2 D(a_1), \tag{2.35}$$

де $(x - \bar{x})^2$ отримується внаслідок винесення детермінованого множника за знак дисперсії, піднесеного до квадрату.

Згідно із (2.18) $D(a_1) = \dfrac{\sigma_u^2}{n\sigma_x^2}$, а використання (2.8) дає (з урахуванням некорельованості $Y_1, Y_2, \ldots, Y_n$)

$$\sigma_{\bar{y}}^2 = D(\bar{y}) = D\left(\frac{\sum_{i=1}^{n} Y_i}{n}\right) = \frac{\sigma_u^2}{n}.$$

Отже, з (2.35)

$$D(\hat{y}) = \left[1 + \frac{(x-\bar{x})^2}{\sigma_x^2}\right]\frac{\sigma_u^2}{n},$$

а незміщена оцінка $D(\hat{y})$ знаходиться за формулою:

$$S_{\hat{y}}^2 = \left[1 + \frac{(x-\bar{x})^2}{\sigma_x^2}\right]\frac{S_u^2}{n}, \qquad (2.36)$$

де S_u^2 визначена формулою (2.22).

Розглянемо статистику (1.10)

$$t = \frac{Z}{\sqrt{\frac{1}{k}\chi^2}} = \frac{Z}{\chi}\sqrt{k},$$

поклавши $Z = \dfrac{\hat{y} - M(Y \mid x)}{\sigma_{\hat{y}}}$, $\chi^2 = \dfrac{kS_{\hat{y}}^2}{\sigma_{\hat{y}}^2}$, $k = n-2$.

Використання передумов 1-4 та (1.7) дозволяє показати, що $Z \sim N(0,1)$, а χ^2 має розподіл χ^2 із k ступенями вільності, причому Z та χ^2 незалежні випадкові величини. На підставі П.4 §1 можна показати, що випадкова величина

$$T = \frac{\hat{y} - M(Y \mid x)}{S_{\hat{y}}}$$

має t-розподіл Ст'юдента із $k = n-2$ ступенями вільності. Оскільки густина $g_k(t)$ розподілу Ст'юдента парна, то

$$P\left(\left|\frac{\hat{y} - M(Y \mid x)}{S_{\hat{y}}}\right| < t\right) = P(|T| < t) = 2\int_0^t g_k(t)dt = \gamma. \qquad (2.37)$$

У табл. 2 додатків наведені значення $t = t(\gamma; k)$ як кореня рівняння (2.37) в залежності від заданої довірчої імовірності γ і від числа ступенів вільності k.

Таким чином, довірчий інтервал для невідомого умовного математичного сподівання $M(Y \mid x)$ має такий вигляд:

$$\hat{y} - t(\gamma; n-2)S_{\hat{y}} < M(Y \mid x) < \hat{y} + t(\gamma; n-2)S_{\hat{y}}, \qquad (2.38)$$

де $S_{\hat{y}} = \sqrt{S_{\hat{y}}^2}$, $S_{\hat{y}}^2$ визначається формулою (2.36).

Якщо урахувати, що $M(Y \mid x) = \alpha_0 + \alpha_1 x$, а x змінюється, то подвійна нерівність (2.38) дає **довірчу зону** для прямої регресії $y = \alpha_0 + \alpha_1 x$. Тому потрібно з'ясувати, яка поведінка «ширини» цієї зони при зміні x. Із формул (2.36) і (2.38) видно, що «ширина» зони мінімальна при $x = \overline{x}$, а при віддаленні x від $\overline{x}$ вона збільшується (рис. 2.1).

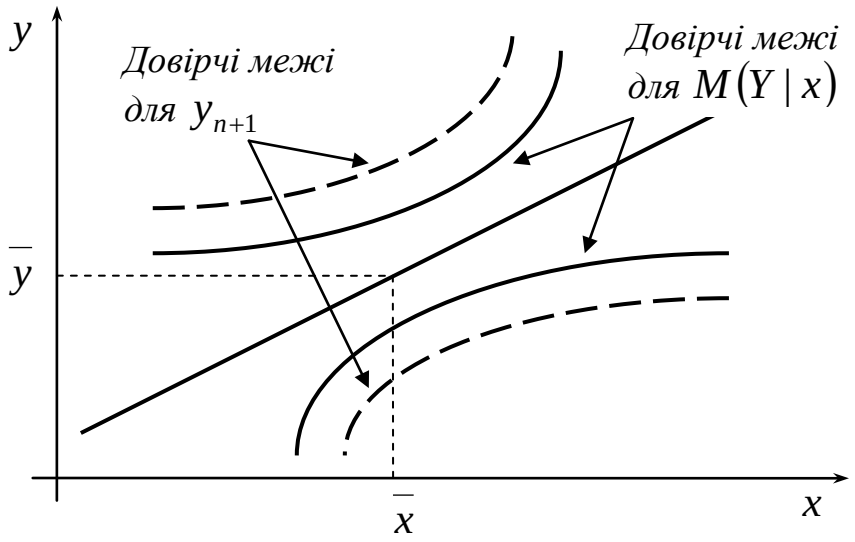


Рисунок 2.1.

Таким чином, **прогноз** значень залежної змінної Y (в середньому) може призвести до **значних похибок**.

Побудова довірчої зони для функції регресії (або $M(Y \mid x)$) у випадку **конкретної** вибірки (іншими словами, довірчої зони для значень регресії в базисних точках) передбачає побудову точок з координатами $\{x_i; \hat{y}_i - t(\gamma, n-2)S_{\hat{y}_i}\}$, $i = \overline{1,n}$, з наступним з'єднанням сусідніх (по індексу i) точок прямолінійними відрізками, а потім здійснення аналогічної процедури для послідовності точок $\{x_i; \hat{y}_i + t(\gamma, n-2)S_{\hat{y}_i}\}$.

Зауваження. Якщо у вибірці є пари чисел із однаковими значеннями змінної x, тоді $\hat{y}_i$ покладається рівним середньому арифметичному тих $\hat{y}$, які відповідають цим значенням, рівних x.

8. Нехай МНК-оцінки a_0 та a_1 є значущими, тобто виконуються нерівності (2.33). Побудуємо довірчі інтервали для невідомих параметрів α_0 та α_1, використавши випадкові величини (2.32), які розподілені за законом Ст'юдента із $k = n - 2$ ступенями вільності.

Позначимо $t_0(\gamma, k)$, $t_1(\gamma, k)$ корені рівнянь

$$P(|t_0| < t) = \gamma, \quad P(|t_1| < t) = \gamma$$

відповідно, які знаходяться за табл. 2 додатків. Тоді із надійністю γ виконуються нерівності

$$\left|\frac{a_0 - \alpha_0}{S_{a_0}}\right| < t_0(\gamma, k), \quad \left|\frac{a_1 - \alpha_1}{S_{a_1}}\right| < t_1(\gamma, k),$$

звідки отримуються довірчі інтервали

$$a_0 - t_0(\gamma, k)S_{a_0} < \alpha_0 < a_0 + t_0(\gamma, k)S_{a_0}, \tag{2.39}$$

$$a_1 - t_1(\gamma, k)S_{a_1} < \alpha_1 < a_1 + t_1(\gamma, k)S_{a_1}, \tag{2.40}$$

де S_{a_0}, S_{a_1} визначаються формулами (2.30).

9. Основна мета дослідження моделі (2.2) — це статистично надійний точковий чи інтервальний прогноз значень залежної змінної Y або її математичного сподівання. Для досягнення цієї мети необхідно з'ясувати **якість моделі** або **значущість моделі**, тобто встановити, чи відповідає модель експериментальним даним і чи достатньо включених в рівняння незалежних (пояснюючих) змінних (однієї або кількох) для описання залежної змінної. Перевірка значущості рівняння регресії проводиться на основі дисперсійного аналізу, який у даному випадку використовується як допоміжний засіб для вивчення якості регресійної моделі.

Розглянемо питання про **декомпозицію дисперсій**. Дисперсія спостережених значень залежної змінної має такий вигляд:

$$\sigma_y^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 / n.$$

Відхилення $y_i - \bar{y}$ запишемо таким чином:

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i). \tag{2.41}$$

У статистиці різницю $y_i - \bar{y}$ називають **загальним відхиленням**, $\hat{y}_i - \bar{y}$ **— відхиленням, яке можна пояснити моделлю**, $y_i - \hat{y}_i$ **— непоясненим відхиленням** (яке не можна пояснити моделлю).

Піднесемо обидві частини (2.41) до квадрату та підсумуємо за індексом i:

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

Якщо в моделі (2.2) $\alpha_0\neq 0$, тоді другий доданок в правій частині дорівнює нулю з урахуванням системи нормальних рівнянь (2.12):

$$\begin{aligned}\sum\left(\hat{y}_i-\bar{y}\right)\left(y_i-\hat{y}_i\right)&=\sum\hat{y}_i\left(y_i-\hat{y}_i\right)-\bar{y}\sum\left(y_i-\hat{y}_i\right)=\\&=\sum\left(a_0+a_1x_i\right)\left(y_i-a_0-a_1x_i\right)-\bar{y}\sum\left(y_i-a_0-a_1x_i\right)=\\&=a_0\sum\left(y_i-a_0-a_1x_i\right)+a_1\sum\left(y_i-a_0-a_1x_i\right)x_i=0.\end{aligned}$$

Тому остаточно отримаємо рівність

$$\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2=\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2+\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2 \qquad (2.42)$$

або

$$СКЗ=СКП+СКН,$$

де $СКЗ$ — загальна сума квадратів, $СКП$ — пояснена сума квадратів, $СКН$ — непояснена сума квадратів.

Поділивши (2.42) на n, отримаємо:

$$\sigma^2_{заг.}=\sigma^2_{регр.}+\sigma^2_{пом.}, \qquad (2.43)$$

де $\sigma^2_{заг.}$ — загальна дисперсія, $\sigma^2_{регр.}$ — дисперсія, що пояснює регресію, $\sigma^2_{пом.}$ — дисперсія помилок.

Для одержання **незміщених** оцінок дисперсій, які фігурують в рівності (2.43), потрібно знайти відповідні ступені вільності.

Для обчислення $СКЗ$ використовуються n чисел $\left\{\left(y_1-\bar{y}\right),\left(y_2-\bar{y}\right),...,\left(y_n-\bar{y}\right)\right\}$, серед яких лінійно незалежними є $n-1$, оскільки $\sum_{i=1}^{n}\left(y_i-\bar{y}\right)=0$. Тому ступінь вільності $СКЗ$ дорівнює $n-1$.

З урахуванням (2.13) рівняння регресії можна записати у такому вигляді:

$$\hat{y}_i-\bar{y}=a_1\left(x_i-\bar{x}\right),$$

звідки $СКП=\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2=a_1^2\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2$.

Отже, *СКП* утворюється з використанням однієї одиниці незалежної інформації — a_1, тому ступінь вільності її дорівнює 1 або $m-1$, де m — число параметрів моделі ($m=2$).

Нарешті, $СКН=\sum_{i=1}^{n}(y_i-\hat{y})^2$ має (див. п.4) $n-2$ ступенів вільності.

Означення. *Число, яке отримується діленням суми квадратів на відповідний ступінь вільності, називається* ***середнім квадратом***.

Середні квадрати

$$S_R^2=\overline{СКП}=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2/1,\ \ S_u^2=\overline{СКН}=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2/(n-2) \qquad (2.44)$$

є **незміщеними оцінками** дисперсій залежної змінної, обумовлених відповідно регресією (пояснюючою змінною) і дією неврахованих факторів та помилок.

Повернемося до рівності (2.43). Якщо обидві її частини розділити на $\sigma^2_{заг.}$, то отримаємо

$$1=\frac{\sigma^2_{регр.}}{\sigma^2_{заг.}}+\frac{\sigma^2_{пом.}}{\sigma^2_{заг.}}. \qquad (2.45)$$

Означення. *Перший доданок в правій частині (2.45) називається* ***коефіцієнтом детермінації*** *і позначається* R^2 (або d)*:*

$$R^2=\frac{\sigma^2_{регр.}}{\sigma^2_{заг.}}=\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{n\sigma_y^2}. \qquad (2.46)$$

Із рівності (2.45) отримаємо рівносильну формулу для означення коефіцієнта детермінації:

$$R^2=1-\frac{\sigma^2_{пом.}}{\sigma^2_{заг.}}=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n\sigma_y^2}=1-\frac{\sum_{i=1}^{n}u_i^2}{n\sigma_y^2}. \qquad (2.46^*)$$

R^2 є однією із **найбільш ефективних оцінок адекватності регресійної моделі**, **мірою її якості** або **характеристикою прогностичної сили моделі**.

Величина R^2 показує, яка частина (частка) варіації залежної змінної обумовлена варіацією пояснюючої змінної згідно з моделлю.

Оскільки доданки в рівності (2.45) невід'ємні, то з урахуванням (2.46) отримується подвійна нерівність $0\le R^2\le 1$.

Чим ближче значення R^2 до одиниці, тим краще модель апроксимує емпіричні дані, тим ближче спостереження знаходяться по відношенню до прямої регресії. Якщо $R^2 = 1$, то всі емпіричні точки (x_i, y_i) лежать на прямій регресії і між змінними Y та x існує лінійна функціональна залежність. Якщо ж $R^2 = 0$, то варіація залежної змінної повністю обумовлена дією неврахованих у моделі змінних і пряма регресії паралельна осі абсцис.

Зауваження. *Коефіцієнт детермінації R^2 є сенс розглядати тільки у випадку наявності вільного члена в моделі (2.2), тобто $\alpha_0 \neq 0$, оскільки тільки у цьому випадку, як це вище відзначалося, виконується рівність (2.42), а, отже, і співвідношення (2.46*).*

Зміст R^2 нагадує зміст вибіркового коефіцієнта кореляції r. Це обумовлено рівністю

$$R^2 = r^2, \tag{2.47}$$

яка отримується із використанням (2.46), (1.20), (2.13):

$$R^2 = \frac{\sigma^2_{регр.}}{\sigma^2_{заг.}} = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \overline{y}\right)^2}{n\sigma_y^2} = \frac{\sum_{i=1}^{n} a_1^2\left(x_i - \overline{x}\right)^2}{n\sigma_y^2} = \frac{a_1^2\sigma_x^2}{\sigma_y^2} =$$

$$= \left(\frac{\overline{xy} - \overline{x}\cdot\overline{y}}{\sigma_x^2}\right)^2 \frac{\sigma_x^2}{\sigma_y^2} == \left(\frac{\overline{xy} - \overline{x}\cdot\overline{y}}{\sigma_x\sigma_y}\right)^2 = r^2.$$

Числові характеристики R^2 та r є точковими статистичними оцінками відповідних невідомих чисел. У зв'язку із цим навіть у випадку $\alpha_1 = 0$ R^2 та $|r|$ можуть бути відмінними від нуля. А тому виникає необхідність перевірити значущість R^2 та r, отриманих для конкретної вибірки.

Значущість R^2 з'ясовується з допомогою статистики Фішера, а r – статистики Ст'юдента.

Якщо $\alpha_1 = 0$, тобто відсутня лінійна залежність між залежною і пояснюючою змінними, тоді випадкові величини

$$S_R^2 = \sum_{i=1}^{n}\left(\hat{y}_i - \overline{y}\right)^2 \Big/ 1 \quad \text{та} \quad S_u^2 = \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \Big/ (n-2)$$

мають χ^2-розподіли відповідно із 1 та $n-2$ ступенями вільності, а їх відношення (згідно із (1.15)) — розподіл Фішера з тими ж ступенями вільно-

сті. Тому рівняння регресії значуще на рівні значущості α, якщо виконується нерівність

$$F_{спост.} = \frac{СКП \cdot (n-2)}{СКН \cdot 1} = \frac{S_R^2}{S_u^2} > F_{кр.}(\alpha; k_1; k_2), \qquad (2.48)$$

де $F_{кр.}(\alpha; k_1; k_2)$ — табличне значення F-критерія Фішера-Снедокора, визначене на рівні значущості α при $k_1 = 1$ і $k_2 = n-2$ ступенях вільності.

Проте виявляється [10, с.97], що **F-тест (2.48) рівносильний t-тесту Ст'юдента при перевірці значущості параметра** α_1 (у випадку парної лінійної моделі).

В ряді задач потрібно оцінити **значущість коефіцієнта кореляції** r. На рівні значущості α він вважається значущим (тобто відкидається гіпотеза $H_0 : \rho = 0$), якщо виконується нерівність (1.21). Однак неважко показати, що отримувані значення t-критерію при перевірці гіпотез $\alpha_1 = 0$ по (2.33) і $\rho = 0$ по (1.21) однакові.

Отже, **якщо на рівні $\alpha = 1-\gamma$ зроблено висновок про значущість α_1, то на тому ж рівні вважається значущим і генеральний (теоретичний) коефіцієнт кореляції ρ і навпаки**.

Наведемо інші прості показники якості лінійної регресії, які використовуються як додаткова інформація при виборі найкращої моделі з можливих.

Абсолютна середня відсоткова помилка MAPE (mean absolute percentage error):

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \cdot 100\%. \qquad (2.49)$$

Цей показник використовується при порівнянні точності прогнозів різнорідних об'єктів, бо характеризує відносну точність прогнозу. При цьому вважається, що значення MAPE, менше 10%, дає високу точність прогнозу, а, отже, і якість моделі; від 10% до 20% — добру точність; від 20% до 50% — задовільну точність; понад 50% — незадовільну точність.

Середня відсоткова помилка MPE (mean percentage error):

$$MPE = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{y}_i - y_i}{y_i} \cdot 100\%. \qquad (2.50)$$

Це показник незміщеності прогнозу. З точки зору практики для **якісних моделей** цей показник повинен бути «малим», тобто не перевищувати 5%.

Зауваження. *Показники (2.49) та (2.50) — невизначені, якщо серед* $y_1, y_2, ..., y_n$ *є нульове значення.*

10. Якщо встановлено, що побудована модель є адекватною, тоді можна знаходити **прогнозні значення залежної змінної**. При цьому можна отримати два типи прогнозів: точковий та інтервальний. Нехай задається значення x_{n+1} незалежної змінної. Тоді точковий прогноз для значення залежної змінної за моделлю (2.11) має такий вигляд:

$$\hat{y}_{n+1} = a_0 + a_1 x_{n+1}. \tag{2.51}$$

Разом з тим **дійсне** значення залежної змінної для прогнозного періоду згідно із (2.3) дорівнює:

$$Y_{n+1} = \alpha_0 + \alpha_1 x_{n+1} + U_{n+1}, \tag{2.52}$$

де стосовно випадкової величини U_{n+1} природно вимагати виконання передумов 1-4, тобто

$$\operatorname{cov}(U_i, U_{n+1}) = 0,\ i = \overline{1,n} \quad U_{n+1} \sim N(0, \sigma_u). \tag{2.53}$$

Отже, $\hat{y}_{n+1}$ є точковою оцінкою **невідомого числа** y_{n+1}, яке є реалізацією (можливим значенням) випадкової величини (2.52).

Згідно із (2.51), (2.52) помилка прогнозу:

$$u_{n+1} = y_{n+1} - \hat{y}_{n+1} = (\alpha_0 - a_0) + (\alpha_1 - a_1) x_{n+1} + U_{n+1}. \tag{2.54}$$

Потрібно знайти числові характеристики u_{n+1} та закон розподілу цієї величини. Незміщеність оцінок a_0, a_1 і (2.53) призводять до рівності $M(u_{n+1}) = 0$. Оскільки U_{n+1} не корелює із $U_1, U_2, ..., U_n$, то згідно із (2.3) U_{n+1} не корелює і з $Y_1, Y_2, ..., Y_n$, а, отже, і з a_0 та a_1. Тому з урахуванням детермінованості α_0, α_1, x_{n+1}, (2.14), (2.18), (2.19) отримаємо:

$$\begin{aligned} D(u_{n+1}) &= D(-a_0) + D(-x_{n+1}a_1) + 2\operatorname{cov}(-a_0, -x_{n+1}a_1) + D(U_{n+1}) = \\ &= D(a_0) + x_{n+1}^2 D(a_1) + 2x_{n+1}\operatorname{cov}(a_0, a_1) + \sigma_u^2 = \\ &= \sigma_u^2 \frac{\overline{x^2}}{n\sigma_x^2} + x_{n+1}^2 \frac{\sigma_u^2}{n\sigma_x^2} - 2x_{n+1}\frac{\bar{x}}{n\sigma_x^2}\sigma_u^2 + \sigma_u^2 = \\ &= \left[\overline{x^2} - \left(\bar{x}\right)^2 + \left(\left(\bar{x}\right)^2 - 2\bar{x}x_{n+1} + x_{n+1}^2\right)\frac{1}{n\sigma_x^2} + 1\right]\sigma_u^2 = \left[1 + \frac{1}{n} + \frac{\left(x_{n+1} - \bar{x}\right)^2}{n\sigma_x^2}\right]\sigma_u^2. \end{aligned}$$

Неважко переконатися також у лінійній залежності u_{n+1} від збурень $U_1, U_2, ..., U_{n+1}$. А тому остаточно отримуємо:

$$u_{n+1} \sim N\left(0, \sigma_{u_{n+1}}\right),$$

де

$$\sigma_{u_{n+1}} = \sigma_u \sqrt{1 + \frac{1}{n} + \frac{\left(x_{n+1} - \bar{x}\right)^2}{n\sigma_x^2}}.$$

Незміщена оцінка $D\left(u_{n+1}\right) = \sigma^2_{u_{n+1}}$ знаходиться за формулою

$$S^2_{u_{n+1}} = S^2_u \left[1 + \frac{1}{n} + \frac{\left(x_{n+1} - \bar{x}\right)^2}{n\sigma_x^2} \right], \qquad (2.55)$$

де S^2_u визначена формулою (2.22).

За аналогією із побудовою інтервальної зони функції регресії (п.7) остаточно можна отримати довірчий інтервал для прогнозного значення залежної змінної:

$$\hat{y}_{n+1} - t\left(\gamma; n-2\right) S_{u_{n+1}} < y_{n+1} < \hat{y}_{n+1} + t\left(\gamma; n-2\right) S_{u_{n+1}}. \qquad (2.56)$$

11. ***Задача 2.1****. Торговельне підприємство має велику кількість філій і його керівництво вивчає питання про залежність Y (річний товарообіг однієї філії, млн. грн.) від x (торгівельної площі, тис. м^2). Для десяти філій за певний рік зафіксовані такі значення показників Y і x:*

i	1	2	3	4	5	6	7	8	9	10
y_i	1,5	2,9	3,1	3,2	4,3	5,7	5,8	7	7,2	7,5
x_i	0,2	0,3	0,5	0,6	0,8	1	1,1	1,2	1,3	1,4

На обсяг товарообігу впливають такі чинники: середньоденна інтенсивність потоку покупців, об'єм основних фондів, їх структура, середньосписькова чисельність працівників, площа підсобних приміщень тощо. Припускається, що в досліджуваній групі філій значення цих чинників приблизно однакові, тому вплив відмінностей їх значень на зміну обсягу товарообігу є незначним.

Вважаючи, що виконуються передумови 1-4*, потрібно:*

1) знайти статистичні оцінки параметрів лінійного рівняння регресії;

2) точкову оцінку та довірчий інтервал дисперсії збурень із надійністю $\gamma = 0{,}9$;

3) для рівня значущості $\alpha = 0{,}05$ *перевірити значущість коефіцієнтів регресії* α_0 *та* α_1*;*

4) знайти довірчі інтервали коефіцієнтів регресії з надійністю $\gamma = 0{,}95$*;*

5) знайти вибіркові коефіцієнт детермінації, коефіцієнт кореляції, а також інші показники якості лінійної регресії (MAPE, MPE);

6) знайти та побудувати довірчу зону функції регресії з надійністю $\gamma = 0{,}95$*;*

7) знайти прогнозне значення річного товарообігу для нової філії, торгівельна площа якої складає 1,8 тис. м^2*, а також із надійністю* $\gamma = 0{,}95$ *побудувати довірчий інтервал для цього прогнозного значення.*

○ **1)** Статистичні оцінки a_0, a_1 параметрів α_0 та α_1 лінійного рівняння регресії задовольняють системі нормальних рівнянь (2.12):

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

Для знаходження коефіцієнтів цієї системи складемо розрахункову табл. 2.1, останній стовпець якої потрібний для обчислення σ_y.

Таблиця 2.1

i	x_i	y_i	x_i^2	$x_i\, y_i$	y_i^2
1	0,2	1,5	0,04	0,3	2,25
2	0,3	2,9	0,09	0,87	8,41
3	0,5	3,1	0,25	1,55	9,61
4	0,6	3,2	0,36	1,92	10,24
5	0,8	4,3	0,64	3,44	18,49
6	1	5,7	1	5,7	32,49
7	1,1	5,8	1,21	6,38	33,64
8	1,2	7	1,44	8,4	49
9	1,3	7,2	1,69	9,36	51,84
10	1,4	7,5	1,96	10,5	56,25
Σ	8,4	48,2	8,68	48,42	272,22

Використовуючи нижній рядок табл. 2.1, отримаємо (обсяг вибірки $n = 10$):

$$\overline{x} = \sum_{i=1}^{10} x_i / n = 8{,}4/10 = 0{,}84; \qquad \overline{y} = \sum_{i=1}^{10} y_i / n = 48{,}2/10 = 4{,}82;$$

$$\overline{x^2} = \sum_{i=1}^{10} x_i^2 / n = 8{,}68/10 = 0{,}868; \qquad \overline{xy} = \sum_{i=1}^{10} x_i y_i / n = 48{,}42/10 = 4{,}842;$$

$$\overline{y^2} = \sum_{i=1}^{10} y_i^2 \,/ n = 272{,}22/10 = 27{,}222;$$

$$\begin{cases} a_0 + 0{,}84a_1 = 4{,}82, \\ 0{,}84a_0 + 0{,}868a_1 = 4{,}842. \end{cases}$$

Єдиний розв'язок цієї системи рівнянь згідно із формулами (2.13):

$$a_1 = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \left(\overline{x}\right)^2} = \frac{4{,}842 - 0{,}84 \cdot 4{,}82}{0{,}868 - \left(0{,}84\right)^2} = \frac{0{,}7932}{0{,}1624} = 4{,}884,$$

$$a_0 = \overline{y} - a_1 \overline{x} = 4{,}82 - 0{,}84 \cdot 4{,}884 = 0{,}717.$$

Отже, емпіричне рівняння регресії має такий вигляд:

$$\hat{y} = 0{,}717 + 4{,}884x. \qquad (2.57)$$

2) Незміщену точкову оцінку S_u^2 невідомої дисперсії збурень σ_u^2 знайдемо за формулою (2.22):

$$S_u^2 = \frac{1}{n-2} \sum_{i=1}^{n} u_i^2 = \frac{1}{n-2} \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2,$$

попередньо обчисливши $\hat{y}_i = 0{,}717 + 4{,}884x_i$ та $u_i^2 = \left(y_i - \hat{y}_i\right)^2$, $i = \overline{1{,}10}$, (табл. 2.2).

Таблиця 2.2

i	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	Σ
y_i	1,5	2,9	3,1	3,2	4,3	5,7	5,8	7	7,2	7,5	–
$\hat{y}_i$	1,6938	2,1822	3,159	3,6474	4,6242	5,601	6,0894	6,5778	7,0662	7,5546	–
u_i	-0,1938	0,7178	-0,059	-0,4474	-0,3242	0,099	-0,2894	0,4222	0,1338	-0,0546	0,0044
u_i^2	0,0376	0,5152	0,0035	0,2002	0,1051	0,0098	0,0838	0,1783	0,0179	0,003	1,1544

Зауваження. *Згідно із (2.21):* $\overline{u} = 0$, *в той час як у нашому випадку* $\sum u_i / 10 = 0{,}00044$. *Цим значенням можна ігнорувати (вважати практично рівним нулю). Разом з тим з'ясуємо причину такого відхилення від нуля. Значення* a_1 *та* a_0 *з точністю до шести знаків після коми відповідно складають* 4,884236 *та* 0,717242, *тобто обидва ці значення (хай і несуттєво) більші тих, які взяті у моделі (2.57). Накопичення додатних похибок у різницях* $y_i - \hat{y}_i$ *і привело до того, що* $\sum\left(y_i - \hat{y}_i\right)$ *незначно перевищує нуль. Відмітимо також, що значення* $a_1 = 4{,}88$, $a_0 = 0{,}72$ *приводять до рівності* $\sum u_i = 0{,}008$.

Використавши підсумок останнього рядка, отримаємо:

$$S_u^2 = \frac{1}{10-2} \cdot 1{,}1544 = 0{,}1443.$$

Ліва і права межі довірчого інтервалу для σ_u^2 визначаються згідно (2.29) за формулами відповідно $\frac{(n-2)S_u^2}{\chi_2^2}$ і $\frac{(n-2)S_u^2}{\chi_1^2}$, де у відповідності із (2.26) та (2.27) χ_1^2 та χ_2^2 є коренями рівнянь

$$P\left(\chi^2(k) > \chi_1^2(p;k)\right) = p, \quad p = (1+\gamma)/2 = 0{,}95,$$
$$P\left(\chi^2(k) > \chi_2^2(p;k)\right) = p, \quad p = (1-\gamma)/2 = 0{,}05.$$

За табл. 4 додатків для $k = n-2 = 8$ знайдемо: $\chi_1^2(0{,}95;8) = 2{,}73$ і $\chi_2^2(0{,}05;8) = 15{,}51$. Тоді ліва межа довірчого інтервалу дорівнює $\frac{8 \cdot 0{,}1443}{15{,}51} = 0{,}0744$, а права — $\frac{8 \cdot 0{,}1443}{2{,}73} = 0{,}4229$. Тобто остаточно з надійністю 0,9

$$0{,}0744 < \sigma_u^2 < 0{,}4229.$$

3) Згідно з п.8, якщо виконується нерівність (2.33): $\left|\frac{a_m}{S_{a_m}}\right| > t_{кр.}$ $(m = 0, m = 1)$, тоді на рівні значущості α приймається гіпотеза $H_1 : a_m \neq 0$. Значення S_{a_0} та S_{a_1} знайдемо із виразів (2.30):

$$S_{a_0} = \sqrt{\frac{S_u^2 \overline{x^2}}{n\sigma_x^2}} = \frac{S_u}{\sigma_x}\sqrt{\frac{\overline{x^2}}{n}} = \frac{\sqrt{0{,}1443}}{\sqrt{\overline{x^2} - (\overline{x})^2}}\sqrt{\frac{\overline{x^2}}{10}} = \frac{\sqrt{0{,}1443}}{\sqrt{0{,}868 - (0{,}84)^2}}\sqrt{\frac{0{,}868}{10}} = 0{,}2777$$

;

$$S_{a_1} = \sqrt{\frac{S_u^2}{n\sigma_x^2}} = \frac{S_u}{\sigma_x\sqrt{n}} = \frac{\sqrt{0{,}1443}}{\sqrt{10[0{,}868 - (0{,}84)^2]}} = 0{,}2981.$$

Тоді спостережені значення критерію:

$$\left|\frac{a_0}{S_{a_0}}\right| = \frac{0{,}717}{0{,}2777} = 2{,}582, \qquad \left|\frac{a_1}{S_{a_1}}\right| = \frac{4{,}884}{0{,}2981} = 16{,}384.$$

Критична точка для двосторонньої критичної області $t_{кр.} = t_{двост.}(\alpha, k)$ при значеннях $\alpha = 0{,}05$, $k = n-2 = 8$ знаходиться за верхньою частиною табл. 3 додатків: $t_{кр.} = 2{,}306$.

Оскільки $2{,}582 > t_{кр.} = 2{,}306$ і $16{,}384 > t_{кр.} = 2{,}306$, то на рівні значущості $\alpha = 0{,}05$ робимо висновки, що $\alpha_0 \neq 0$ і $\alpha_1 \neq 0$.

4) Згідно з (2.39) та (2.40) довірчі інтервали з надійністю γ для невідомих параметрів регресії a_0 та a_1 мають такий вигляд:

$$a_m - t_m(\gamma, k)S_{a_m} < \alpha_m < a_m + t_m(\gamma, k)S_{a_m},$$

де $m = 0{,}1$, $t_m = t_m(\gamma, k)$ — корінь рівняння $P(|t_m| < t) = \gamma$, t_0 та t_1 — випадкові величини, розподілені за законом Ст'юдента.

У нашому випадку $\gamma = 0{,}95$, число ступенів вільності $k = n - 2 = 8$. За табл. 2 додатків знаходимо $t_0(0{,}95;8) = t_1(0{,}95;8) = 2{,}306$. Тоді з врахуванням знайдених значень $S_{a_0} = 0{,}2777$, $S_{a_1} = 0{,}1132$ отримаємо:

$$0{,}717 - 2{,}306 \cdot 0{,}2777 < \alpha_0 < 0{,}717 + 2{,}306 \cdot 0{,}2777,$$

$$4{,}884 - 2{,}306 \cdot 0{,}2981 < \alpha_1 < 4{,}884 + 2{,}306 \cdot 0{,}2981$$

або остаточно

$$0{,}0766 < \alpha_0 < 1{,}3574,$$

$$4{,}1966 < \alpha_1 < 5{,}5714.$$

5) Коефіцієнт детермінації R^2 знайдемо за формулою (2.46*):

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

Із табл. 2.2 (останнє число нижнього рядка) $\sum_{i=1}^{10} u_i^2 = \sum_{i=1}^{10}(y_i - \hat{y}_i)^2 = 1{,}1544$.

Для знаходження σ_y^2 використаємо табл. 2.1:

$$\sigma_y^2 = \overline{y^2} - (\overline{y})^2 = 272{,}22/10 - (48{,}2/10)^2 = 1{,}9896.$$

Отже,

$$R^2 = 1 - \frac{1{,}1544}{10 \cdot 1{,}9896} = 0{,}9711.$$

Таким чином, варіація залежної змінної Y на 97,11% пояснюється варіацією пояснюючої змінної.

Вибірковий коефіцієнт кореляції згідно із (2.47):

$$r = \sqrt{R^2} = \sqrt{0{,}9711} = 0{,}9854.$$

При цьому додатний знак цього числа обрано в зв'язку з тим, що $a_1 > 0$.

Обчислимо абсолютну середню відсоткову помилку MAPE за формулою (2.49):

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \cdot 100\%.$$

Для цього використаємо другий і четвертий рядки табл. 2.2:

$$\sum_{i=1}^{10}\left|\frac{\hat{y}_i - y_i}{y_i}\right| = \frac{0{,}1938}{1{,}5} + \frac{0{,}7178}{2{,}9} + \frac{0{,}059}{3{,}1} + \frac{0{,}4474}{3{,}2} + \frac{0{,}3242}{4{,}3} + \frac{0{,}099}{5{,}7} +$$

$$+ \frac{0{,}2894}{5{,}8} + \frac{0{,}4222}{7} + \frac{0{,}1338}{7{,}2} + \frac{0{,}0546}{7{,}5} = 0{,}1292 + 0{,}2475 + 0{,}019 + 0{,}1398 +$$

$$+ 0{,}0754 + 0{,}0174 + 0{,}0499 + 0{,}0603 + 0{,}0186 + 0{,}0073 = 0{,}7644.$$

Отже, $MAPE = \frac{1}{10} \cdot 0{,}7644 \cdot 100\% = 7{,}644\% < 10\%$, тобто відповідає високій точності прогнозу за моделлю.

Середню відсоткову помилку MPE знайдемо за формулою (2.50):

$$MPE = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{y}_i - y_i}{y_i} \cdot 100\%,$$

використавши розрахунки при обчисленні MAPE:

$$\sum_{i=1}^{10}\frac{\hat{y}_i - y_i}{y_i} = 0{,}1292 - 0{,}2475 + 0{,}019 + 0{,}1398 + +0{,}0754 - 0{,}0174 +$$

$$+ 0{,}0499 - 0{,}0603 - 0{,}0186 + 0{,}0073 = 0{,}0768.$$

Остаточно

$$MPE = \frac{1}{10} \cdot 0{,}0768 \cdot 100\% = 0{,}768\% < 5\%.$$

Висновок*: всі знайдені показники вказують на високу якість моделі.*

6) Побудова довірчої зони для функції регресії передбачає побудову точок з координатами $\{x_i; \hat{y}_i - t(\gamma, n-2)S_{\hat{y}_i}\}$, $i = \overline{1,n}$, з наступним з'єднанням сусідніх (по індексу i) точок прямолінійними відрізками, а потім здійснення аналогічної процедури для послідовності точок $\{x_i; \hat{y}_i + t(\gamma, n-2)S_{\hat{y}_i}\}$.

Величину $S_{\hat{y}_i}$ знайдемо із формули (2.36) $S_{\hat{y}_i} = S_u\sqrt{\left[1 + \frac{(x_i - \bar{x})^2}{\sigma_x^2}\right]\frac{1}{n}}$.

Використовуючи табл. 2.1 і знайдене значення $S_u = \sqrt{0{,}1443} = 0{,}3799$, отримаємо:

$$\sigma_x^2 = \overline{x^2} - \left(\overline{x}\right)^2 = 0{,}1624;$$

$$S_{\hat{y}_1} = 0{,}3799\sqrt{\left[1 + \frac{(0{,}2 - 0{,}84)^2}{0{,}1624}\right]\frac{1}{10}} = 0{,}2255;$$

$$S_{\hat{y}_2} = 0{,}3799\sqrt{\left[1 + \frac{(0{,}3 - 0{,}84)^2}{0{,}1624}\right]\frac{1}{10}} = 0{,}201;$$

$$S_{\hat{y}_3} = 0{,}1572; \quad S_{\hat{y}_4} = 0{,}1398; \quad S_{\hat{y}_5} = 0{,}1207; \quad S_{\hat{y}_6} = 0{,}1293;$$

$$S_{\hat{y}_7} = 0{,}143; \quad S_{\hat{y}_8} = 0{,}1611; \quad S_{\hat{y}_9} = 0{,}1823; \quad S_{\hat{y}_{10}} = 0{,}2057.$$

За табл. 2 додатків знайдемо $t(0{,}95;8) = 2{,}306$. Використовуючи табл. 2.2 і знайдені $S_{\hat{y}_i}$, отримаємо ординати точок нижньої межі довірчої зони:

$$\hat{y}_1 - tS_{\hat{y}_1} = 1{,}6938 - 2{,}306 \cdot 0{,}2255 = 1{,}1738;$$

$$\hat{y}_2 - tS_{\hat{y}_2} = 2{,}1822 - 2{,}306 \cdot 0{,}201 = 1{,}7187;$$

$$\hat{y}_3 - tS_{\hat{y}_3} = 3{,}159 - 2{,}306 \cdot 0{,}1572 = 2{,}7965;$$

$$\hat{y}_4 - tS_{\hat{y}_4} = 3{,}6474 - 2{,}306 \cdot 0{,}1398 = 3{,}325;$$

$$\hat{y}_5 - tS_{\hat{y}_5} = 4{,}6242 - 2{,}306 \cdot 0{,}1207 = 4{,}3459;$$

$$\hat{y}_6 - tS_{\hat{y}_6} = 5{,}601 - 2{,}306 \cdot 0{,}1293 = 5{,}3028;$$

$$\hat{y}_7 - tS_{\hat{y}_7} = 6{,}0894 - 2{,}306 \cdot 0{,}143 = 5{,}7596;$$

$$\hat{y}_8 - tS_{\hat{y}_8} = 6{,}5778 - 2{,}306 \cdot 0{,}1611 = 6{,}2063;$$

$$\hat{y}_9 - tS_{\hat{y}_9} = 7{,}0662 - 2{,}306 \cdot 0{,}1823 = 6{,}6458;$$

$$\hat{y}_{10} - tS_{\hat{y}_{10}} = 7{,}5546 - 2{,}306 \cdot 0{,}2057 = 7{,}0803.$$

Тоді ординати точок верхньої межі довірчої зони набирають таких значень:

$$\hat{y}_1 + tS_{\hat{y}_1} = 1{,}6938 + 2{,}306 \cdot 0{,}2255 = 2{,}2138;$$

$$\hat{y}_2 + tS_{\hat{y}_2} = 2{,}1822 + 2{,}306 \cdot 0{,}201 = 2{,}6457;$$

$$\hat{y}_3 + tS_{\hat{y}_3} = 3{,}159 + 2{,}306 \cdot 0{,}1572 = 3{,}5215;$$

$$\hat{y}_4 + tS_{\hat{y}_4} = 3{,}6474 + 2{,}306 \cdot 0{,}1398 = 3{,}9698;$$

$$\hat{y}_5 + tS_{\hat{y}_5} = 4{,}6242 + 2{,}306 \cdot 0{,}1207 = 4{,}9025;$$

$$\hat{y}_6 + tS_{\hat{y}_6} = 5{,}601 + 2{,}306 \cdot 0{,}1293 = 5{,}8992;$$

$\hat{y}_7 + tS_{\hat{y}_7} = 6{,}0894 + 2{,}306 \cdot 0{,}143 = 6{,}4192;$

$\hat{y}_8 + tS_{\hat{y}_8} = 6{,}5778 + 2{,}306 \cdot 0{,}1611 = 6{,}9493;$

$\hat{y}_9 + tS_{\hat{y}_9} = 7{,}0662 + 2{,}306 \cdot 0{,}1823 = 7{,}4866;$

$\hat{y}_{10} + tS_{\hat{y}_{10}} = 7{,}5546 + 2{,}306 \cdot 0{,}2057 = 8{,}0289.$

Довірча зона (з надійністю 0,95) для функції регресії зображена на рис. 2.2.

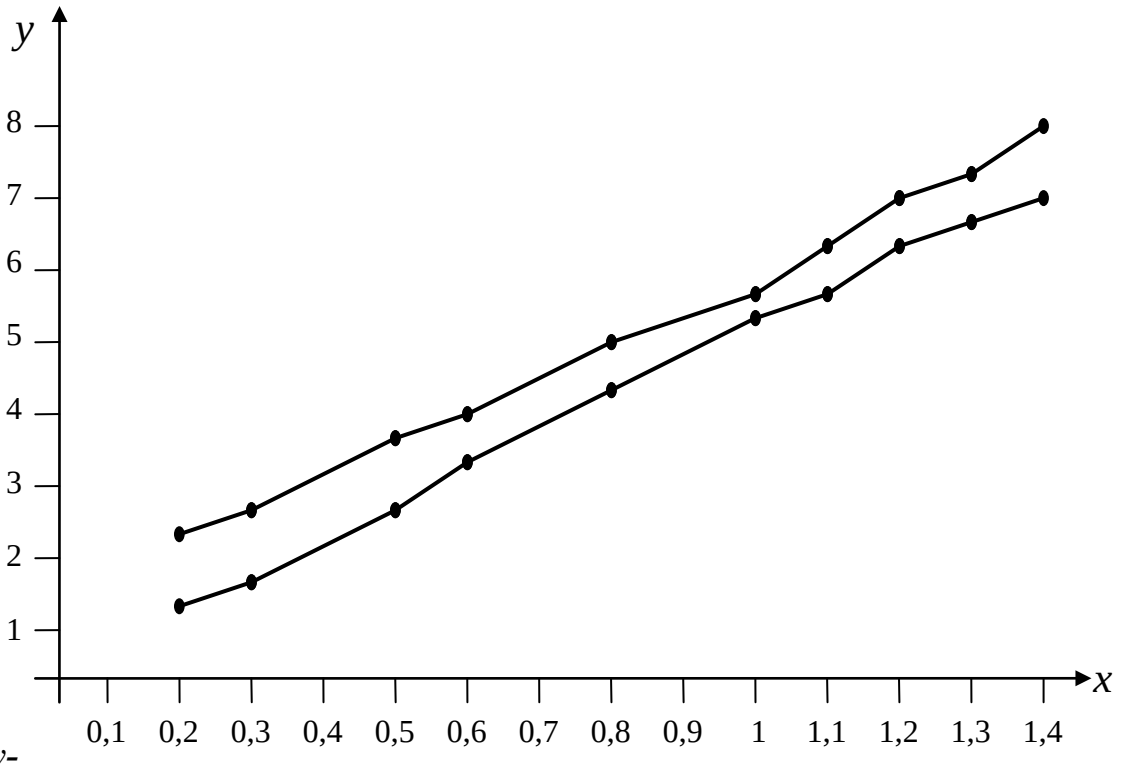


Рисунок 2.2.

7) Прогнозне значення річного товарообігу для нової філії із торгівельною площею 1,8 тис. м^2 знайдемо із рівняння (2.57):

$$\hat{y}_{n+1} = 0{,}717 + 4{,}884 \cdot 1{,}8 = 9{,}508.$$

Довірчий інтервал для прогнозного значення y_{n+1} із надійністю $\gamma = 0{,}95$ визначається (2.56). З допомогою виразу (2.55) знайдемо

$$S_{u_{n+1}} = S_u \sqrt{1 + \frac{1}{n} + \frac{\left(x_{n+1} - \bar{x}\right)^2}{n\sigma_x^2}} = 0{,}3799 \sqrt{1 + \frac{1}{10} + \frac{\left(1{,}8 - 0{,}84\right)^2}{10 \cdot 0{,}1624}} = 0{,}4906.$$

Тоді шуканий довірчий інтервал має вид

$$9{,}508 - 2{,}306 \cdot 0{,}4906 < y_{n+1} < 9{,}508 + 2{,}306 \cdot 0{,}4906$$

або остаточно

$$8{,}3767 < y_{n+1} < 10{,}6393.$$

ІНДИВІДУАЛЬНЕ ЗАВДАННЯ № 1

Торгівельне підприємство має велику кількість філій і керівництво цього підприємства вивчає питання про залежність Y (річний товарообіг однієї філії, млн. грн.) від x (торгівельної площі, тис. $м^2$). Для десяти філій за певний рік зафіксовані такі значення показників Y і x:

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14
y_i	2,8	5,2	6,8	7,1	7,3	8,3	4,3	5,8	7,7	3,2	1,5	3,7	5,4	2,6
x_i	0,3	0,9	1,2	1,3	1,2	0,8	0,8	0,9	1,3	0,5	0,3	0,6	1,1	0,2
i	15	16	17	18	19	20	21	22	23	24	25	26	27	28
y_i	5,8	8,1	7,5	8,4	4,2	5,6	7,6	3,5	1,4	3,9	6,4	7,3	7,6	8,3
x_i	0,8	1,6	1,5	0,9	0,9	0,8	1,4	0,7	0,4	0,8	1,1	1,4	1,5	0,9
i	29	30	31	32	33	34	35	36	37	38	39	40	41	42
y_i	4,4	5,6	7,5	3,2	1,5	3,5	2,7	4,3	6,9	7,1	7,3	8,4	4,2	5,8
x_i	0,8	0,7	1,4	0,6	0,3	0,7	0,2	0,6	1,4	1,3	1,2	0,9	0,7	0,7
i	43	44	45	46	47	48	49	50						
y_i	7,5	3,5	1,4	4,2	2,6	5,8	4,2	3,1						
x_i	1,5	0,8	0,3	0,9	0,4	0,7	0,8	0,6						

На обсяг товарообігу впливають такі чинники: середньоденна інтенсивність потоку покупців, об'єм основних фондів, їх структура, середньоспискова чисельність працівників, площа підсобних приміщень тощо. Припускається, що в досліджуваній групі філій значення цих чинників приблизно однакові, тому вплив відмінностей їх значень на зміну обсягу товарообігу є незначним.

Потрібно:

1. знайти статистичні оцінки параметрів лінійного рівняння регресії;
2. точкову оцінку та довірчий інтервал дисперсії збурень із надійністю $\gamma = 0{,}9$;
3. для рівня значущості $\alpha = 0{,}05$ перевірити значущість коефіцієнтів регресії α_0 та α_1;
4. знайти довірчі інтервали коефіцієнтів регресії з надійністю $\gamma = 0{,}95$;
5. знайти вибіркові коефіцієнт детермінації, коефіцієнт кореляції, а також інші показники якості лінійної регресії (MAPE, MPE);
6. знайти та побудувати довірчу зону функції регресії з надійністю $\gamma = 0{,}95$;

7. знайти прогнозне значення річного товарообігу для нової філії, торгівельна площа якої складає 1,8 *тис.м*^{2}, а також із надійністю $\gamma = 0{,}95$ побудувати довірчий інтервал для цього прогнозного значення.

§ 3. ПЕРЕВІРКА ВИКОНАННЯ ПЕРЕДУМОВ КЛАСИЧНОЇ НОРМАЛЬНОЇ ЛІНІЙНОЇ МОДЕЛІ ПАРНОЇ РЕГРЕСІЇ.

1. ***Відповідність вибірки нормальному розподілу.***
2. ***Гетероскедастичність та її наслідки.***
3. ***Діагностування гетероскедастичності та її вилучення.***
   - ***3.1. Тест рангової кореляції Спірмена.***
   - ***3.2. Тест Голдфелда-Квандта.***
   - ***3.3. Тест Уайта.***
   - ***3.4. Тест Глейзера.***
   - ***3.5. Усунення гетероскедастичності.***
4. ***Автокореляція залишків часового ряду.***
5. ***Авторегресія першого порядку. Статистика Дарбіна-Уотсона.***
6. ***Тести на наявність автокореляції.***

При розв'язуванні задачі §2 було зроблено висновок про високу якість побудованої лінійної моделі парної регресії. Проте суттєвою обставиною є те, що всі оцінки і висновки отримані при умові, що виконуються передумови 1-4. Тому для достовірності прийнятих рішень необхідно перевірити виконання цих передумов. При цьому слід мати на увазі, що перевірці при МНК-оцінюванні підлягають лише передумови 2, 3, 4 (див. зауваження до рівності (2.21)). У випадку констатації невиконання хоча б однієї передумови модель класичної парної регресії перетворюється на **економетричну модель** і актуальними стають шляхи її дослідження.

1. Розпочнемо із перевірки виконання передумови 4 про нормальність розподілу випадкового збурення U. Це зумовлено тим, що використання статистик Ст'юдента, Пірсона, Фішера-Снедокора, а також деяких тестів на відсутність гетероскедастичності передбачає нормальність розподілу U.

Якщо обсяг вибірки є дуже великим, тоді згідно з центральною граничною теоремою Ляпунова є підстави стверджувати виконання цієї передумови. У випадку малих вибірок ($n \le 30$) вже неправомірно робити такий висновок. Проблема також полягає у тому, що випадкова величина U є **неспостережуваною**. Для отримання інформації про можливі значення цієї величини будемо виходити з двох обставин:

1) лінійності моделі (2.9)

$$y_i = a_0 + a_1 x_i + u_i, \quad i = \overline{1,n};$$

2) отримання оцінок коефіцієнтів регресії a_0, a_1 з допомогою МНК.

Тоді

$$u_i = y_i - \hat{y}_i = y_i - a_0 - a_1 x_i, \quad i = \overline{1,n} \tag{3.1}$$

можна тлумачити як «спостережені» значення збурення U. Наявність лапок в слові «спостережені» зумовлене врахуванням двох попередніх обставин (лінійна модель може виявитись не найкращою, а оцінки a_0 та a_1 можуть отримуватися з допомогою інших методів).

Отже, будемо вважати, що n чисел $u_1, u_2, ..., u_n$, визначені (3.1), відомі. Вони є можливими (спостереженими) значеннями випадкової величини U. Ставиться задача про перевірку статистичної гіпотези H_0: U розподілена за нормальним законом ($U \sim N(0, \sigma_u)$). Якщо обсяг вибірки великий (див. [4]), то можна використати критерії узгодженості Пірсона або Колмогорова. Проте в багатьох випадках обсяги вибірки є малими, тому використаємо критерій Фішера.

В якості основних характеристик розподілу зручніше всього брати коефіцієнт асиметрії і ексцес:

$$A = \frac{\mu_3}{\sigma^3}, \qquad E = \frac{\mu_4}{\sigma^4} - 3,$$

де μ_3 та μ_4 — центральні моменти третього та четвертого порядків відповідно. Для випадку нормального розподілу $\mu_3 = 0$, $\mu_4 = 3\sigma^4$. Тому для цього розподілу виконуються рівності

$$A = 0, \qquad E = 0. \tag{3.2}$$

Відповідні вибіркові (емпіричні) коефіцієнти асиметрії і ексцесу визначаються формулами:

$$A^* = \frac{\mu_3^*}{\sigma_u^3}, \qquad E^* = \frac{\mu_4^*}{\sigma_u^4} - 3, \tag{3.3}$$

де центральний емпіричний момент m-го порядку μ_m^* визначається за формулою (для даного випадку позначень):

$$\mu_m^* = \frac{1}{n} \sum_{i=1}^{n} \left(u_i - \bar{u}\right)^m, \quad m = 3,4. \tag{3.4}$$

Емпіричні коефіцієнт асиметрії і ексцес, як і всі числові характеристики нефіксованої вибірки, є випадковими величинами і тому навіть для нормального розподілу можуть відрізнятися від нуля. Закони розподілу

A^* і E^* дуже складні і мало вивчені. Фішер запропонував модифікацію оцінок коефіцієнта асиметрії A^* та ексцесу E^*:

$$\hat{A}^* = \frac{\mu_3^*}{\sigma_u^3}\sqrt{\frac{n(n-1)}{(n-2)^2}}, \qquad \hat{E}^* = \frac{n^2-1}{(n-2)(n-3)}\left(\frac{\mu_4^*}{\sigma_u^4} - 3\frac{n-1}{n+1}\right). \tag{3.5}$$

При невеликих обсягах вибірок $\hat{A}^*$ та $\hat{E}^*$ помітно відрізняються від A^* та E^*. Виявляється, що у випадку нормального розподілу досліджуваної випадкової величини оцінки $\hat{A}^*$ та $\hat{E}^*$ мають з великим ступенем точності нормальні розподіли, причому $M(\hat{A}^*)=0$, $M(\hat{E}^*)=0$, а дисперсії визначаються виразами:

$$\sigma^2_{\hat{A}^*} = \frac{6n(n-1)}{(n+3)(n+1)(n-2)}, \quad \sigma^2_{\hat{E}^*} = \frac{24n(n-1)^2}{(n+5)(n+3)(n-2)(n-3)}. \tag{3.6}$$

Отже, задача полягає у відповіді на питання: чи значуще оцінки $\hat{A}^*$ і $\hat{E}^*$ відрізняються від своїх математичних сподівань, тобто від нуля?

На практиці можна користуватися таким наближеним критерієм узгодженості:

$$\left|\hat{A}^*\right| \le 2\sigma_{\hat{A}^*}, \qquad \left|\hat{E}^*\right| \le 2\sigma_{\hat{E}^*}. \tag{3.7}$$

Задача 3.1. *На основі статистичних даних задачі 2.1 здійснити перевірку виконання передумови 4 про нормальність розподілу випадкової величини* U.

○ Для знаходження $\hat{A}^*$ і $\hat{E}^*$ згідно з формулами (3.5) потрібно обчислити μ_3, μ_4, σ_u^3, σ_u^4. Згідно з останнім рядком табл. 2.2 $\sum_{i=1}^{10} u_i^2 = 1{,}1544$. Тому

$\sigma_u^2 = \overline{u^2} - (\overline{u})^2 = \frac{1}{10}\cdot 1{,}1544 - 0^2 = 0{,}11544$, $\sigma_u = 0{,}33976$, $\sigma_u^3 = 0{,}0392208$, $\sigma_u^4 = 0{,}0133263$. Значення μ_3^* та μ_4^* знайдемо за формулами (3.4). Для цього складемо розрахункову табл. 3.1, дані першого стовпця якої взяті з четвертого рядка табл. 2.2.

Таблиця 3.1

i	u_i	u_i^3	u_i^4
1	-0,1938	-0,0072788	0,0014106
2	0,7178	0,369837	0,265469
3	-0,059	-0,0002054	0,0000121
4	-0,4474	-0,0895546	0,0400667
5	-0,3242	-0,0340752	0,0110472

6	0,099	0,0009703	0,0000961
7	-0,2894	-0,0242379	0,0070145
8	0,4222	0,0752583	0,0317741
9	0,1338	0,0023953	0,0003205
10	-0,0546	-0,0001628	0,0000089
Σ	0,0044	0,2929462	0,3572197

Врахувавши, що $\bar{u} = 0$, отримаємо:

$$\mu_3^* = \frac{1}{n}\sum_{i=1}^{10} u_I^3 = \frac{1}{10}\cdot 0{,}2929463 = 0{,}0292946,$$

$$\mu_4^* = \frac{1}{n}\sum_{i=1}^{10} u_I^4 = \frac{1}{10}\cdot 0{,}357219 = 0{,}0357220.$$

За формулами (3.5), (3.6) знаходимо:

$$\hat{A}^* = \frac{0{,}0292946}{0{,}0392208}\sqrt{\frac{10\cdot 9}{8^2}} = 0{,}8857321,$$

$$\hat{E}^* = \frac{10^2-1}{8\cdot 7}\left(\frac{0{,}0357219}{0{,}0133263} - 3\cdot\frac{9}{11}\right) = 0{,}3997819,$$

$$\sigma^2_{\hat{A}^*} = \frac{6\cdot 10\cdot 9}{13\cdot 11\cdot 8} = 0{,}4720280, \qquad \sigma_{\hat{A}^*} = 0{,}6870429,$$

$$\sigma^2_{\hat{E}^*} = \frac{24\cdot 10\cdot 9^2}{15\cdot 13\cdot 8\cdot 7} = 1{,}7802198, \qquad \sigma_{\hat{E}^*} = 1{,}3342488.$$

Обидві нерівності (3.7) виконуються:

$$0{,}8857321 < 2\cdot 0{,}6870429, \qquad 0{,}3997819 < 2\cdot 1{,}3342488,$$

а тому згідно з цим критерієм згоди гіпотеза H_0 про нормальність закону розподілу U приймається. ◎

2. Друга передумова (гомоскедастичність або «однаковий розкид») передбачає виконання n рівностей

$$D(U_i) = \sigma_u^2 = const, \quad i = \overline{1,n}. \tag{3.8}$$

Якщо хоча б одна з цих рівностей не виконується, тобто

$$D(U_i) = f(x_i), \quad i = \overline{1,n}, \tag{3.9}$$

тоді має місце гетероскедастичність («неоднаковий розкид»).

Появу проблеми гетероскедастичності часто можна передбачити заздалегідь, ґрунтуючись на значенні характеру даних. Припущення про гомоскедастичність виправдане в тих випадках, коли досліджувані об’єкти є

достатньо **однорідними**. Якщо ж досліджуються неоднорідні об'єкти, то, як правило, виникає проблема гетероскедастичності.

Приклади:

1) Якщо вивчається залежність прибутку фірми від розміру основних фондів, то природно очікувати, що для великих фірм коливання прибутку буде вищим, ніж для малих.

2) Якщо досліджується залежність витрат на харчування в родині від загального доходу, то розкид у даних буде більшим для родин із більш високим доходом.

Графічна форма розкиду спостережень залежить від форми зв'язку між $\sigma_{u_i}^2$ та x_i. Приклади зображені на рис. 3.1.

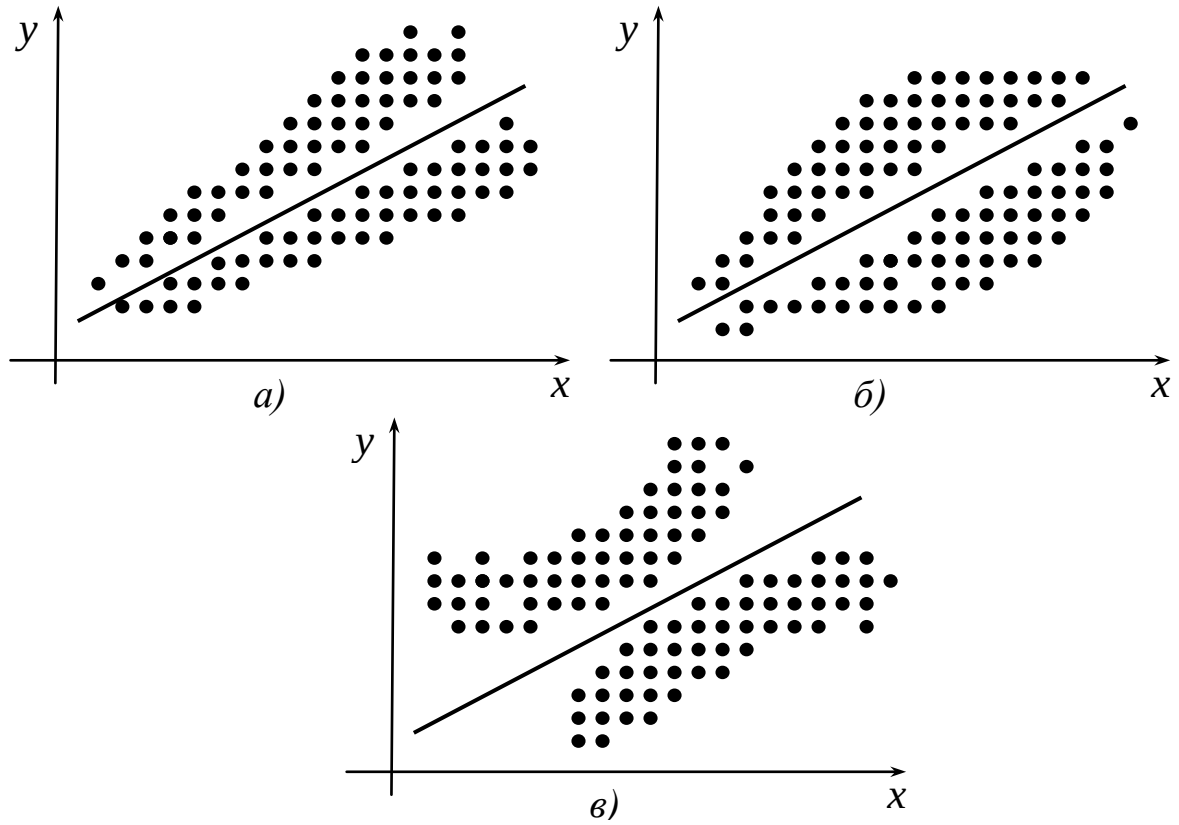


Рисунок 3.1.

У прикладних дослідженнях, як правило, використовується зручне припущення, що $f(x_i) = k^2 x_i^2$, де k — стала, яку потрібно оцінити.

Розглянемо наслідки порушення передумови про гомоскедастичність. Можна довести, що оцінки коефіцієнтів регресії залишаються лінійними і незміщеними, але вже не володіють властивістю ефективності, тобто їх дисперсії вже не будуть мінімальними в класі лінійних оцінок. У зв'язку з цим розширюються довірчі інтервали. Як наслідок, тести Ст'юдента і Фі-

шера-Снедокора дають неточні результати. Крім того, формулу для оцінки σ_u, строго кажучи, застосовувати вже не можна.

Отже, гетероскедастичність є серйозною проблемою. Досліднику потрібно знати: є вона чи немає. У випадку тестування гетероскедастичності вихідну модель необхідно модифікувати.

3. Виявляється, **єдині з правил діагностування гетероскедастичності немає**, а є різні тести з своїми недоліками та перевагами. Розглянемо найпростіші з них за змістом та розрахунками. В кожному тесті в якості нульової гіпотези розглядається H_0 — гіпотеза про відсутність гетероскедастичності.

3.1. Тест рангової кореляції Спірмена передбачає найбільш загальні припущення про залежність дисперсій помилок регресії від значень незалежної змінної:

$$\sigma_{u_i}^2 = f_i(x_i), \quad i = \overline{1,n}. \tag{3.10}$$

При цьому ніяких додаткових припущень відносно виду функцій f_i не робиться. Відмітимо також, що **відсутнє обмеження стосовно закону розподілу помилок**.

Ідея тесту полягає в тому, що абсолютні величини залишків регресії $|u_i|$ розглядаються як оцінки σ_{u_i}, тому при наявності гетероскедастичності $|u_i|$ і значення x_i будуть корелювати. Проте кореляція в цьому випадку передбачається **ранговою**.

Рангова кореляція досліджується тоді, коли необхідно встановити силу зв'язку між **ординальними** (**порядковими**) **змінними**. Прикладами ординальних змінних є житлові умови, тестові бали, екзаменаційні оцінки. Джерелом ординальних змінних можуть бути і **кількісні** змінні, для яких здійснюється процес ранжування. Наприклад, кожну з двох множин чисел $\{|u_1|, |u_2|, ..., |u_n|\}$, $\{x_1, x_2, ..., x_n\}$ можна ранжувати в порядку зростання. В результаті i-тий об'єкт характеризується двома рангами s_i та l_i по змінних u та x. Тоді **коефіцієнт рангової кореляції Спірмена** знаходиться за формулою

$$\rho = 1 - \frac{6\sum_{i=1}^{n}(s_i - l_i)^2}{n^3 - n}. \tag{3.11}$$

Якщо ранги всіх об'єктів рівні між собою, тобто $s_i = l_i \ \ \forall \ \ i = \overline{1,n}$, то $\rho = 1$. Цей випадок називається **повним прямим зв'язком**. При **повному оберненому зв'язку**, коли ранги об'єктів по обох змінних розташовані в оберненому порядку, можна довести, що $\rho = -1$. У решті випадків $|\rho| < 1$.

При перевірці значущості ρ виходять із того, що у випадку правильності нульової гіпотези (про відсутність кореляційного зв'язку між змінними) при $n > 10$ статистика

$$t = \frac{\rho\sqrt{n-2}}{\sqrt{1-\rho^2}} \tag{3.12}$$

має t-розподіл Ст'юдента із $n-2$ ступенями вільності. Тому ρ значущий на рівні α, якщо

$$|t_{спост.}| > t_{двост.\ \ кр.}(\alpha; n-2), \tag{3.13}$$

де $t_{спост.}$ визначається лівою частиною (3.12).

Задача 3.2. *На основі статистичних даних задачі 2.1 на рівні* $\alpha = 0{,}05$ *перевірити виконання передумови 2 з допомогою тесту рангової кореляції Спірмена.*

○ Здійснимо ранжування змінних $|u_i|$ та x_i. Для цього складемо табл. 3.2, використавши дані табл. 2.1 (другий стовпець) та табл. 2.2 (передостанній рядок). При заповненні останнього стовпця вибирається найменше число в третьому стовпці і поруч з ним записується 1. Найменшому з тих чисел, що залишилися, відповідає 2, і т.д.

Таблиця 3.2

x_i	l_i *(ранг* x_i*)*	$\|u_i\|$	s_i *(ранг* $\|u_i\|$*)*
0,2	1	0,1938	5
0,3	2	0,7178	10
0,5	3	0,059	2
0,6	4	0,4474	9
0,8	5	0,3242	7
1	6	0,099	3
1,1	7	0,2894	6
1,2	8	0,4222	8
1,3	9	0,1338	4
1,4	10	0,0546	1

За формулою (3.11) знайдемо коефіцієнт рангової кореляції Спірмена:

$$\rho = 1 - \frac{6}{10^3 - 10}\Big[(5-1)^2 + (10-2)^2 + (2-3)^2 + (9-4)^2 + (7-5)^2 +$$
$$+ (3-6)^2 + (6-7)^2 + (8-8)^2 + (4-9)^2 + (1-10)^2\Big] =$$
$$= 1 - 0{,}00606(16 + 64 + 1 + 25 + 4 + 9 + 1 + 25 + 81) = -0{,}3696.$$

Згідно із (3.12)

$$|t_{спост.}| = \frac{|-0{,}3696|\sqrt{10-2}}{\sqrt{1-(-0{,}3696)^2}} = 1{,}1251.$$

За табл. 3 додатків $t_{двост.\ кр.}(0{,}05;8) = 2{,}306$.

Оскільки на рівні $\alpha = 0{,}05$ нерівність (3.13) не виконується, то нульова гіпотеза про відсутність кореляційного зв'язку є правильною. Отже, на цьому ж рівні приймається і гіпотеза H_0 про відсутність гетероскедастичності (виконання передумови 2). ◎

3.2. Тест Голдфелда-Квандта використовується у тому випадку, коли помилки регресії U_i можна вважати **нормально розподіленими** випадковими величинами. При цьому спостережень має бути хоча б удвічі більше, ніж число оцінюваних параметрів. Як правило, тест застосовується до великих вибірок.

Впорядкуємо n спостережень в порядку зростання значень x_i і виберемо m перших і m останніх спостережень (число m визначимо пізніше). Тоді гіпотеза про гомоскедастичність буде рівносильна тому, що значення $u_1, u_2, ..., u_m$ та $u_{n-m+1}, u_{n-m+2}, ..., u_n$ є вибірковими спостереженнями **нормально розподілених** випадкових величин, які мають **однакові дисперсії**.

Зауваження. *Для знаходження $u_i = y_i - \hat{y}_i$ для двох груп ($i = \overline{1,m}$ та $i = \overline{n-m+1,n}$) необхідно попередньо знайти два емпіричні рівняння для кожної з груп.*

Гіпотеза про рівність дисперсій двох нормально розподілених сукупностей, як відомо [4], перевіряється з допомогою критерія Фішера-Снедокора. Нульова гіпотеза про рівність дисперсій двох сукупностей по m спостережень (тобто гіпотеза про відсутність гетероскедастичності) відкидається на рівні α, якщо

$$F_{спост.}=\frac{\max\left\{\sum_{i=1}^{m}u_i^2,\ \sum_{i=n-m+1}^{n}u_i^2\right\}}{\min\left\{\sum_{i=1}^{m}u_i^2,\ \sum_{i=n-m+1}^{n}u_i^2\right\}}>F_{кр.}\left(\alpha;m-1;m-1\right). \qquad (3.14)$$

Відмітимо, що чисельник і знаменник в (3.14) слід було розділити на відповідне число ступенів вільності, проте в даному випадку ці числа однакові і рівні $m-1$.

Виявляється, що коли вибрати m порядку $n/3$, тоді **потужність тесту**, тобто імовірність відкинути гіпотезу про наявність гомоскедастичності, коли насправді гетероскедастичність є, буде **максимальною**.

Задача 3.3. *На рівні значущості* $\alpha=0{,}05$ *для задачі 2.1 перевірити виконання передумови 2 (ігноруючи малість вибірки) з допомогою тесту Голдфелда-Квандта.*

○ За аналогією із розв'язуванням задачі 2.1. (п.1) знайдемо статистичні оцінки параметрів $a_1^{(1)}$, $a_0^{(1)}$ та $a_1^{(2)}$, $a_0^{(2)}$, виходячи з двох груп даних ($m=4$):

i	1	2	3	4
y_i	1,5	2,9	3,1	3,2
x_i	0,2	0,3	0,5	0,6

i	7	8	9	10
y_i	5,8	7	7,2	7,5
x_i	1,1	1,2	1,3	1,4

$$\bar{x}^{(1)}=\left(0{,}2+0{,}3+0{,}5+0{,}6\right)/4=0{,}4;$$
$$\bar{x}^{(2)}=\left(1{,}1+1{,}2+1{,}3+1{,}4\right)/4=1{,}25;$$
$$\overline{x^2}^{(1)}=\left(0{,}2^2+0{,}3^2+0{,}5^2+0{,}6^2\right)/4=0{,}185;$$
$$\overline{x^2}^{(2)}=\left(1{,}1^2+1{,}2^2+1{,}3^2+1{,}4^2\right)/4=1{,}575;$$
$$\bar{y}^{(1)}=\left(1{,}5+2{,}9+3{,}1+3{,}2\right)/4=2{,}675;$$
$$\bar{y}^{(2)}=\left(5{,}8+7+7{,}2+7{,}5\right)/4=6{,}875;$$
$$\overline{xy}^{(1)}=\left(1{,}5\cdot0{,}2+2{,}9\cdot0{,}3+3{,}1\cdot0{,}5+3{,}2\cdot0{,}6\right)/4=1{,}16;$$
$$\overline{xy}^{(2)}=\left(5{,}8\cdot1{,}1+7\cdot1{,}2+7{,}2\cdot1{,}3+7{,}5\cdot1{,}4\right)/4=8{,}66;$$
$$a_1^{(1)}=\frac{\overline{xy}^{(1)}-\bar{x}^{(1)}\cdot\bar{y}^{(1)}}{\overline{x^2}^{(1)}-\left(\bar{x}^{(1)}\right)^2}=\frac{1{,}16-0{,}4\cdot2{,}675}{0{,}185-\left(0{,}4\right)^2}=3{,}6;$$
$$a_0^{(1)}=\bar{y}^{(1)}-a_1^{(1)}\bar{x}^{(1)}=2{,}675-3{,}6\cdot0{,}4=1{,}235;$$

$$a_1^{(2)} = \frac{\overline{xy}^{(2)} - \bar{x}^{(2)} \cdot \bar{y}^{(2)}}{\overline{x^2}^{(2)} - \left(\bar{x}^{(2)}\right)^2} = \frac{8{,}66 - 1{,}25 \cdot 6{,}875}{1{,}575 - (1{,}25)^2} = 5{,}3;$$

$$a_0^{(2)} = \bar{y}^{(2)} - a_1^{(2)} \bar{x}^{(2)} = 6{,}875 - 5{,}3 \cdot 1{,}25 = 0{,}25.$$

Отже, емпіричне рівняння для першої групи має такий вигляд:

$$\hat{y}^{(1)} = 1{,}235 + 3{,}6x,$$

а другої

$$\hat{y}^{(2)} = 0{,}25 + 5{,}3x.$$

Тоді

$$\sum_{i=1}^{4} u_i^2 = \sum_{i=1}^{4} \left(y_i - \hat{y}_i^{(1)}\right)^2 = (1{,}5 - 1{,}235 - 3{,}6 \cdot 0{,}2)^2 + (2{,}9 - 1{,}235 - 3{,}6 \cdot 0{,}3)^2 +$$

$$+ (3{,}1 - 1{,}235 - 3{,}6 \cdot 0{,}5)^2 + (3{,}2 - 1{,}235 - 3{,}6 \cdot 0{,}6)^2 =$$

$$= 0{,}207 + 0{,}3422 + 0{,}0042 + 0{,}038 = 0{,}5914;$$

$$\sum_{i=7}^{10} u_i^2 = \sum_{i=7}^{10} \left(y_i - \hat{y}_i^{(2)}\right)^2 = (5{,}8 - 0{,}25 - 5{,}3 \cdot 1{,}1)^2 + (7 - 0{,}25 - 5{,}3 \cdot 1{,}2)^2 +$$

$$+ (7{,}2 - 0{,}25 - 5{,}3 \cdot 1{,}3)^2 + (7{,}5 - 0{,}25 - 5{,}3 \cdot 1{,}4)^2 =$$

$$= 0{,}0784 + 0{,}1521 + 0{,}0036 + 0{,}0289 = 0{,}263;$$

$$\max\left\{\sum_{i=1}^{4} u_i^2, \sum_{i=7}^{10} u_i^2\right\} = 0{,}5914; \quad \min\left\{\sum_{i=1}^{4} u_i^2, \sum_{i=7}^{10} u_i^2\right\} = 0{,}263;$$

$$F_{спост.} = \frac{\sum_{i=1}^{4} u_i^2}{\sum_{i=7}^{10} u_i^2} = \frac{0{,}5914}{0{,}263} = 2{,}2487.$$

Врахувавши, що $m = 4$, $\alpha = 0{,}05$, за табл. 5 додатків знайдемо $F_{кр.}(0{,}05;3;3) = 9{,}28$.

Оскільки $F_{спост.} = 2{,}2487 < F_{кр.}(0{,}05;3;3) = 9{,}28$, тобто нерівність (3.14) не виконується, то робимо висновок, що гіпотеза H_0 про відсутність гетероскедастичності на рівні $\alpha = 0{,}05$ приймається.

Отже, за тестом Голдфелда-Квандта на рівні $\alpha = 0{,}05$ для статистичних даних задачі 2.1. передумова 2 виконується. ◎

3.3. Тест рангової кореляції Спірмена і тест Голдфелда-Квандта дозволяють лише **виявити наявність гетероскедастичності**, але вони не

дають можливості з'ясувати кількісний характер залежності дисперсій помилок регресії від значень незалежної змінної, і, отже, не дають методів усунення гетероскедастичності.

Для досягнення цієї мети необхідні деякі додаткові припущення стосовно характеру гетероскедастичності. Справді, без цих припущень, очевидно, неможливо було б оцінити n дисперсій помилок регресії $\sigma_{u_i}^2$ ($i=\overline{1,n}$) з допомогою n спостережень.

Найбільш простий і часто використовуваний тест на гетероскедастичність — **тест Уайта**. При його використанні припускається, що дисперсії помилок регресії є **значеннями однієї і тієї ж функції** від спостережених значень незалежної змінної, тобто рівняння (3.10) набирають такого виду:

$$\sigma_{u_i}^2 = f(x_i), \quad i=\overline{1,n}. \tag{3.15}$$

Найчастіше функція f обирається квадратичною:

$$f(x_i)=\alpha_0+\alpha_1 x_i+\alpha_2 x_i^2, \tag{3.16}$$

що відповідає тому, що σ_{u_i} залежить від x_i приблизно лінійно. У випадку гомоскедастичності $f=const$, тобто вибіркові коефіцієнти регресії a_1, a_2, які є оцінками невідомих чисел α_1, α_2 відповідно, незначуще відрізняються від нуля.

Ідея тесту Уайта полягає в оцінці функції f з (3.15) за допомогою відповідного рівняння регресії для квадратів залишків:

$$u_i^2 = f(x_i)+\varepsilon_i, \quad i=\overline{1,n}, \tag{3.17}$$

де ε_i випадкова величина (за аналогією з U_i із рівняння (2.3)).

Гіпотеза про відсутність гетероскедастичності (умова $f=const$) приймається у випадку незначущості регресії (3.17) в цілому (тобто одночасної незначущості теоретичних коефіцієнтів регресії α_1 та α_2).

Знаходження стандартних помилок коефіцієнтів регресії a_1 і a_2 можна здійснити шляхом розгляду моделі з двома незалежними змінними $z_1=x$, $z_2=x^2$ (див. §4).

3.4. Тест Глейзера аналогічний тесту Уайта, тільки в якості залежної змінної для вивчення гетероскедастичності вибирається не квадрат залишків, а їх абсолютна величина, тобто розглядається регресія

$$|u_i| = f(x_i)+\varepsilon_i, \quad i=\overline{1,n}, \tag{3.18}$$

де зазвичай

$$f(x)=\beta_0+\beta_1 x^{\delta}, \tag{3.19}$$

Регресія (3.18) вивчається при різних значеннях δ, а потім обирається те конкретне значення, при якому коефіцієнт β_1 виявляється найбільш значущим, тобто має найбільше значення t-статистики. При цьому в якості значень δ беруться числа: 1, 2, 3, 1/2, 1/3 тощо. Якщо ж β_1 незначущий для всіх розглянутих значень δ (випадок $f \equiv const$), тоді робиться висновок про відсутність гетероскедастичності.

Задача 3.4. *На рівні значущості* $\alpha=0{,}05$ *для задачі 2.1 перевірити виконання передумови 2 за тестом Глейзера.*

○ Покладемо у функції (3.19) $\delta=1$. Нехай b_0 і b_1 — оцінки невідомих параметрів β_0 і β_1 відповідно. Тоді оцінкою моделі (3.18) за статистичними даними $\{|u_i|, x_i\}_{i=\overline{1,n}}$ ($|u_i|$ знайдені в задачі 2.1, а, отже, відомі) є n рівнянь

$$|u_i|=b_0+b_1x_i+\hat{\varepsilon}_i, \quad i=\overline{1,n},$$

де $b_0+b_1x_i=|\hat{u}_i|$ – емпіричне рівняння, $\hat{\varepsilon}_i$ — вибіркова оцінка збурення ε_i.

За формулами (2.12), (2.13) знайдемо МНК-оцінки b_0 та b_1, використавши дані табл. 2.1 і 2.2:

$$b_1=\frac{\overline{x|u|}-\bar{x}\overline{|u|}}{\overline{x^2}-\left(\bar{x}\right)^2}=\frac{0{,}198576-0{,}84\cdot 0{,}27412}{0{,}868-(0{,}84)^2}=-0{,}1951,$$

$$b_0=\overline{|u|}-b_1\bar{x}=0{,}27412-(-0{,}1951)\cdot 0{,}84=0{,}438.$$

Незміщену точкову оцінку S_ε^2 невідомої дисперсії збурень σ_ε^2 знайдемо за аналогом формули (2.22):

$$S_\varepsilon^2=\frac{1}{n-2}\sum_{i=1}^{n}\left(\overline{|\hat{u}_i|}-|\hat{u}_i|\right)^2,$$

попередньо обчисливши

$|\hat{u}_1|=|0{,}438-0{,}1951\cdot 0{,}2|=0{,}39899,$

$|\hat{u}_2|=|0{,}438-0{,}1951\cdot 0{,}3|=0{,}37948,$

… … … … … … … … … … … … … …,

$|\hat{u}_{10}|=|0{,}438-0{,}1951\cdot 1{,}3|=0{,}16486.$

Отже, $S_\varepsilon^2 = 0{,}0426$, $S_\varepsilon = 0{,}2064$.

Для визначення значущості β_1 знайдемо S_{b_1} за формулою (2.30):

$$S_{b_1} = \sqrt{\frac{S_\varepsilon^2}{n\sigma_x^2}} = \frac{S_\varepsilon}{\sqrt{n}\sigma_x} = \frac{0{,}2064}{\sqrt{10}\sqrt{0{,}868 - (0{,}84)^2}} = 0{,}162.$$

Тоді $$\left|\frac{b_1}{S_{b_1}}\right| = \frac{0{,}1951}{0{,}162} = 1{,}2043.$$

Але оскільки $\left|\frac{b_1}{S_{b_1}}\right| = 1{,}2043 < t_{двостор.}(0{,}05;8) = 2{,}306$, то на рівні $\alpha = 0{,}05$ робимо висновок, що $\beta_1 = 0$. Таким чином, у випадку $\delta = 1$

$$f(x) = \beta_0 \equiv const.$$

При розгляді випадків $\delta = 2,\ 3,\ 1/2,\ 1/3$ доцільно досліджувати рівняння

$$|u_i| = \beta_0 + \beta_1 z_i + \varepsilon_i, \quad i = \overline{1,n},$$

де $z_i = x_i^\delta$, за вище розглянутою схемою. Відповідні чисельні розрахунки для цих випадків наведені в §5.

Виявляється, що і в цих випадках коефіцієнт регресії β_1 незначущий (перевірте самостійно), тобто для статистичних даних задачі 2.1 передумова 2 виконується. ◎

Зауваження. *Припустимо, що значення t-статистики $\left|b_1/S_{b_1}\right|$ при $\delta = 1$ більше $t_{двостор.}(\alpha; n-2)$, що на рівні α означає наявність гетероскедастичності. Тоді необхідно з'ясувати, який вид функції (3.19), тобто, якому значенню δ слід віддати перевагу. Для цього потрібно при різних значеннях δ обчислити значення t-статистики або значення коефіцієнта детермінації. Знайдене максимальне значення і відповідатиме шуканому значенню δ.*

3.5. Повернемося до моделі (2.3)

$$Y_i = \alpha_0 + \alpha_1 x_i + U_i, \quad i = \overline{1,n}, \tag{3.20}$$

для якої встановлено існування гетероскедастичності, а за припущенням збурення U_i не корелюють між собою ($\operatorname{cov}(U_i, U_j) = 0\ \forall\ i, j = \overline{1,n},\ i \neq j$). Надалі буде з'ясовано, яким чином перевірити це припущення. Поділимо

ліву і праву частину рівнянь (3.20) на σ_i, де $\sigma_1, \sigma_2, ..., \sigma_n$ є фіксованими додатними числами, не рівними між собою. Тоді для моделі

$$\frac{Y_i}{\sigma_i} = \frac{\alpha_0}{\sigma_i} + \alpha_1 \frac{x_i}{\sigma_i} + \frac{U_i}{\sigma_i}, \quad i = \overline{1,n}, \tag{3.21}$$

вже виконується умова гомоскедастичності, оскільки при $\sigma_i = \sigma_{u_i}$

$$D\left(\frac{U_i}{\sigma_{u_i}}\right) = \frac{1}{\sigma_{u_i}^2} D(U_i) = \frac{1}{\sigma_{u_i}^2} \left\{M\left(U_i^2\right) - \left[M(U_i)\right]^2\right\} = \frac{1}{\sigma_{u_i}^2} \sigma_{u_i}^2 = 1.$$

Нехай a_0 і a_1 — оцінки невідомих параметрів α_0 і α_1 відповідно. Тоді оцінкою моделі (3.21) за вибіркою $\left\{(x_i, y_i), i = \overline{1,n}\right\}$ є n рівнянь

$$\frac{y_i}{\sigma_{u_i}} = \frac{a_0}{\sigma_{u_i}} + a_1 \frac{x_i}{\sigma_{u_i}} + \frac{u_i}{\sigma_{u_i}}, \quad i = \overline{1,n}.$$

Оцінки a_0, a_1 знайдемо із умови мінімізації функції

$$\tilde{Q}(a_0, a_1) = \sum_{i=1}^{n} \left(\frac{a_0 + a_1 x_i - y_i}{\sigma_{u_i}} \right)^2.$$

За аналогією із п.2 §2 отримаємо систему нормальних рівнянь

$$\begin{cases} \left(\sum_{i=1}^{n} \frac{1}{\sigma_{u_i}^2} \right) a_0 + \left(\sum_{i=1}^{n} \frac{x_i}{\sigma_{u_i}^2} \right) a_1 = \sum_{i=1}^{n} \frac{y_i}{\sigma_{u_i}^2}, \\ \left(\sum_{i=1}^{n} \frac{x_i}{\sigma_{u_i}^2} \right) a_0 + \left(\sum_{i=1}^{n} \frac{x_i^2}{\sigma_{u_i}^2} \right) a_1 = \sum_{i=1}^{n} \frac{x_i y_i}{\sigma_{u_i}^2}, \end{cases} \tag{3.22}$$

яка однозначно визначає невідомі оцінки a_0, a_1.

Описаний метод знаходження оцінок параметрів регресії називається **методом зважених найменших квадратів** (МЗНК). Можна безпосередньо перевірити, що МЗНК покращує точність моделі: оцінки коефіцієнтів для моделі (3.21) (її називають **зваженою регресією**) більш ефективні в порівнянні з оцінками звичайної регресії (3.20).

Для практичної реалізації МЗНК потрібно в системі рівнянь (3.22) замість σ_{u_i} підставити знайдені за допомогою теста Глейзера значення $f(x_i)$, які є статистичними оцінками невідомих σ_{u_i}, або $\sigma_{u_i} = \sqrt{f(x_i)}$ за тестом Уайта.

Зауваження. *На практиці процедура усунення гетероскедастичності може викликати технічні труднощі. Справа в тому, що внаслідок вико-*

ристання оцінок σ_{u_i} модель (3.21) не обов'язково виявиться гомоскедастичною з огляду на такі причини. По-перше, далеко не завжди виявляється правильним саме припущення (3.16) або (3.18). По-друге, функція f у формулі (3.15), взагалі кажучи, не обов'язково степенева (і, тим більше, не обов'язково квадратична), і у цьому випадку її підбір може виявитися далеко не таким простим.

Другим недоліком тестів Уайта і Глейзера є те, що факт не виявлення ними гетероскедастичності, взагалі кажучи, не означає її відсутності. Справді, приймаючи гіпотезу H_0, ми приймаємо лише той факт, що відсутня певного виду залежність дисперсій помилок регресії від незалежної змінної.

4. Передумова 3 передбачає відсутність кореляції між збуреннями U_i та U_j при $i \neq j$, тобто $\operatorname{cov}(U_i, U_j) = 0 \quad \forall \quad i, j = \overline{1,n}, \; i \neq j$. Моделі, для яких не виконується ця передумова, називаються **моделями із наявністю автокореляції**.

Відмітимо, що у випадку просторової вибірки відсутність автокореляції **постулюється**. Разом з тим при використанні комп'ютерних регресійних пакетів (незалежно від виду вибірки) наводиться значення статистики Дарбіна-Уотсона стосовно наявності автокореляції між **сусідніми** членами вибірки.

Для часових рядів модель (2.3) запишеться у такому вигляді:

$$Y_t = \alpha_0 + \alpha_1 x_t + U_t, \quad t = \overline{1,n}. \tag{3.23}$$

Якщо збурення U_t в різні моменти часу t корелюють між собою, тоді між значеннями збурення існує залежність, яку у загальному випадку можна записати у вигляді

$$U_t = \rho_1 U_{t-1} + \rho_2 U_{t-2} + \ldots + \rho_s U_{t-s} + \varepsilon_t, \quad t = \overline{s+1,n}, \tag{3.24}$$

де ρ_i — i-тий коефіцієнт кореляції, ε_t — випадкова величина, s — величина запізнення, $s = \overline{1,n-1}$.

При виконанні співвідношення (3.24) із $\rho_i \neq 0$, $i = \overline{1,s}$, говорять, що послідовність $\left\{U_t, t = \overline{1,n}\right\}$ утворить **авторегресійний процес s-го порядку**. Назва «авторегресійний» зумовлена тим, що U_t визначається значеннями цієї ж самої величини в попередніх моментах часу (запізненнями).

МНК при наявності корельованості збурень регресії дає незміщені оцінки коефіцієнтів регресії, але вони вже є неефективними. Більше того, оцінки їх дисперсій зміщені (як правило, у бік заниження), тобто результати тестування гіпотез виявляються недостовірними.

Таким чином, актуальними є такі питання:

1) як діагностувати наявність автокореляції?
2) яким чином модель із автокореляцією можна було б привести до класичної регресійної моделі?

Відповідь на друге питання пов'язана з необхідністю розгляду **узагальненого методу найменших квадратів** (УМНК), який буде викладено в §6.

5. Якщо в моделі (3.23) присутня автокореляція, тоді, як правило, найбільший вплив на наступне спостереження виявляє результат попереднього спостереження. Наприклад, коли розглядається ряд значень курсу деякої акції, то, очевидно, саме результат останніх торгів слугує відправною точкою для формування курсу на наступних торгах.

Ситуація, коли на значення спостереження Y_t виявляє основний вплив не результат Y_{t-1}, а більш ранні значення, є достатньо рідкісною.

Таким чином, відсутність кореляції між сусідніми членами служить вагомою підставою вважати, що кореляція відсутня в цілому, і звичайний МНК дає адекватні і ефективні результати.

Тест Дарбіна-Уотсона визначає наявність автокореляції між сусідніми членами (часового) ряду. Він ґрунтується на простій ідеї: якщо кореляція збурень (залишків) регресії U_t існує, то вона присутня у залишках регресії u_t, які отримуються в результаті використання звичайного МНК. У тесті Дарбіна-Уотсона для оцінки кореляції використовується статистика виду

$$d = (DW) = \frac{\sum_{t=2}^{n} \left(u_t - u_{t-1}\right)^2}{\sum_{t=1}^{n} u_t^2}. \qquad (3.25)$$

Можна довести, що статистика Дарбіна-Уотсона пов'язана з вибірковим коефіцієнтом кореляції між сусідніми спостереженнями таким чином:

$$d \approx 2(1-r).$$

Природно, що у випадку відсутності автокореляції вибірковий коефіцієнт кореляції r є близьким до нуля, а значення статистики d буде близьким до двох.

Якщо спостережене значення $d \approx 0$, то $r \approx 1$, тобто спостерігається додатна автокореляція, а якщо $d \approx 4$, то $r \approx -1$, тобто наявна від'ємна кореляція. У зв'язку з цим бажано мати відповідні порогові (критичні) значення, присутні у статистичних критеріях, і які або дозволяють прийняти гіпотезу, або примушують її відкинути. Проте, на жаль, такі критичні або порогові значення однозначно вказати неможливо.

Тест Дарбіна-Уотсона має один суттєвий недолік — розподіл статистики d залежить не тільки від числа спостережень n, але і від значень незалежної змінної x_t. Це означає, що тест Дарбіна-Уотсона, взагалі кажучи, **не можна віднести до статистичних критеріїв**, оскільки не можна вказати критичну область, яка дозволяла б відкинути гіпотезу про відсутність кореляції, якби виявилося, що в цю область потрапило б спостережене значення статистики d.

Проте існують два порогові значення $d_в$ і $d_н$, які залежать **тільки** від n і рівня значущості, і такі, що виконуються наступні твердження.

Якщо спостережене значення d:

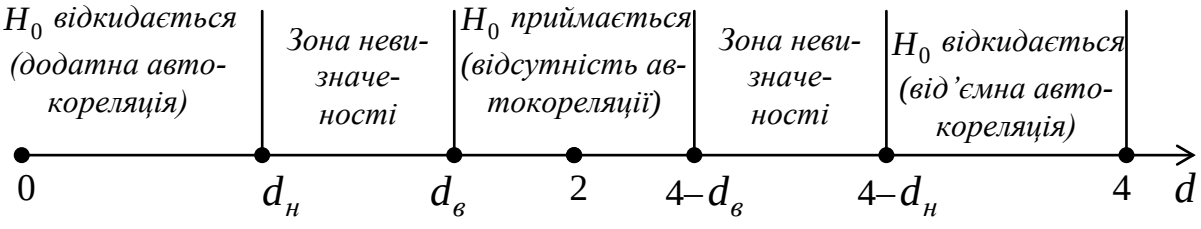
а) $d_в < d < 4 - d_в$, то гіпотеза H_0 про відсутність автокореляції приймається;

б) $d_н < d < d_в$ або $4 - d_в < d < 4 - d_н$, то питання про відхилення або прийняття гіпотези залишається відкритим (значення d потрапляє у область невизначеності критерія);

в) $0 < d < d_н$, то приймається альтернативна гіпотеза H_1 про додатну автокореляцію;

г) $4 - d_н < d < 4$, то приймається альтернативна гіпотеза H_1 про від'ємну автокореляцію.

Зобразимо наведені результати графічно:



Для d-статистики знайдені верхня $d_в$ і нижня $d_н$ межі на рівнях значущості $\alpha = 0{,}01;\ 0{,}025$ і $0{,}05$.

У табл. 6 додатків наведені значення статистик $d_в$ і $d_н$ критерія Дарбіна-Уотсона для рівня значущості $\alpha = 0{,}05$.

Недоліком критерія Дарбіна-Уотсона є наявність зон невизначеності цього критерія, а також те, що критичні значення d-статистики визначені для обсягів вибірки $n \geq 15$. Тим не менше, тест Дарбіна-Уотсона є найбільш використовуваним.

При використанні комп'ютерних регресійних пакетів значення статистики d наводиться автоматично при оцінюванні моделі методом найменших квадратів.

Задача 3.5. *На рівні значущості* $\alpha = 0{,}05$ *для задачі 2.1 перевірити виконання передумови 3 за тестом Дарбіна-Уотсона.*

○ Підсумок останнього рядка табл. 2.2 визначає знаменник дробу (3.25): $\sum_{t=1}^{10} u_t^2 = 1{,}1544$. Для обчислення чисельника цього ж дробу використаємо дані передостаннього рядка табл. 2.2:

$$\sum_{t=2}^{10}(u_t - u_{t-1})^2 = [0{,}7178 - (-0{,}1938)]^2 + (-0{,}059 - 0{,}7178)^2 +$$

$$+ [-0{,}4474 - (-0{,}059)]^2 + [-0{,}3242 - (-0{,}4474)]^2 + [0{,}099 - (-0{,}3242)]^2 +$$

$$+ (-0{,}2894 - 0{,}099)^2 + [0{,}4222 - (-0{,}2894)]^2 + (0{,}1338 - 0{,}4222)^2 +$$

$$+ (-0{,}0546 - 0{,}1338)^2 = 0{,}831 + 0{,}6034 + 0{,}1509 + 0{,}0152 + 0{,}1791 +$$

$$+ 0{,}1509 + 0{,}5064 + 0{,}0832 + 0{,}0355 = 2{,}5556.$$

Спостережене значення статистики (3.25):

$$d = \frac{2{,}5556}{1{,}1544} = 2{,}2138.$$

За табл. 6 додатків при $n = 15$ критичні значення $d_н = 1{,}08$, $d_в = 1{,}36$, тобто $d = 2{,}2138$ знаходиться в межах від $d_в$ до $4 - d_в$ $(1{,}36 < d < 2{,}64)$. Як було відмічено вище, при $n < 15$ критичних значень d-статистики в таблиці немає, проте згідно з тенденцією їх змін із зменшенням n, можна припустити, що знайдене значення залишиться в інтервалі $(d_в; 4 - d_в)$, тобто для статистичних даних задачі 2.1 на рівні значущості $\alpha = 0{,}05$ гіпотеза про відсутність автокореляції збурень не відхиляється (приймається). ◎

6. Статистика Дарбіна-Уотсона є найбільш важливим індикатором наявності автокореляції. Проте недоліками цієї статистики, окрім наявності зон невизначеності, є також обмеженість результату, яка полягає у тому, що кореляція виявляється тільки між **сусідніми** членами. Ця обставина приводить до необхідності використання інших тестів на наявність автокореляції. У всіх цих тестах в якості основної гіпотези H_0 розглядається гіпотеза про відсутність автокореляції.

Тест серій (Бреуша-Годфрі) ґрунтується на такій ідеї: якщо існує кореляція між сусідніми спостереженнями, то природно очікувати, що у рівнянні

$$U_t = \rho U_{t-1} + \varepsilon_t, \quad t = \overline{2,n} \tag{3.26}$$

коефіцієнт ρ виявиться значуще відмінним від нуля. Відмітимо, що рівняння (3.26) є частковим видом рівняння (3.24) при $s = 1$.

Практичне використання тесту серій полягає в оцінюванні методом найменших квадратів моделі (3.26).

Припустимо, що випадкова величина ε_t задовольняє передумовам 1-4. Порівняння моделей (3.26) і (2.3) дозволяє зробити висновок, що $\alpha_0 = 0$, $\alpha_1 = \rho$. Крім цього, вибіркою обсягом $n-1$ є дані

$$\left\{ y_i = u_i; x_i = u_{i-1}, t = \overline{2,n} \right\}. \tag{3.27}$$

Тоді згідно з МНК із другого рівняння (2.12), де $a_0 = 0$, отримаємо оцінку параметра ρ:

$$\hat{\rho} = \frac{\overline{xy}}{\overline{x^2}} = \frac{\sum\limits_{i=2}^{n} x_i y_i}{\sum\limits_{i=2}^{n} x_i^2} \tag{3.28}$$

або з урахуванням (3.27)

$$\hat{\rho} = \frac{\sum\limits_{i=2}^{n} u_{i-1} u_i}{\sum\limits_{i=2}^{n} u_{i-1}^2}. \tag{3.28*}$$

За аналогією із формулою (2.22) незміщеною оцінкою дисперсії збурень ε_t є

$$S_\varepsilon^2 = \frac{1}{n-2} \sum_{i=2}^{n} (u_i - \hat{u}_i)^2, \tag{3.29}$$

де $\hat{u}_i = \hat{\rho} u_{i-1}$, а величина знаменника зумовлена тим, що обсяг вибірки $n-1$ і число оцінюваних параметрів дорівнює 1.

Із (3.28) та (3.27) отримаємо з урахуванням передумов 2, 3 та детермінованості x_i і ρ

$$\sigma_{\hat{\rho}}^2 = D(\hat{\rho}) = \frac{\sum_{i=2}^{n} x_i^2}{\left(\sum_{i=2}^{n} x_i^2\right)^2} \sigma_y^2 = \frac{\sigma_\varepsilon^2}{\sum_{i=2}^{n} u_{i-1}^2}.$$

Незміщена оцінка $\sigma_{\hat{\rho}}^2$ з урахуванням (3.29) має такий вигляд:

$$S_{\hat{\rho}}^2 = \frac{\sum_{i=2}^{n} (u_i - \hat{u}_i)^2}{(n-2)\sum_{i=2}^{n} u_{i-1}^2}, \qquad (3.30)$$

Зміст основної гіпотези $H_0(\rho = 0)$ та альтернативної $H_1(\rho \neq 0)$ дозволяє сформулювати двосторонній критерій значущості оцінки $\hat{\rho}$:

якщо виконується нерівність

$$\left| \frac{\hat{\rho}}{S_{\hat{\rho}}} \right| > t_{кр.}, \qquad (3.31)$$

де $t_{кр.} = t_{двост.кр.}(\alpha; n-2)$ — критична точка розподілу Ст'юдента, тоді на рівні значущості α приймається гіпотеза H_1, тобто вважається, що $\rho \neq 0$.

Перевагою тесту Бреуша-Годфрі в порівнянні з тестом Дарбіна-Уотсона є те, що він перевіряється з допомогою **статистичного критерія**. Друга перевага полягає у можливості досліджувати авторегресійний процес (3.24) s-го порядку. Наприклад: $U_t = \rho_1 U_{t-1} + \rho_2 U_2 + \varepsilon_t$ для $s = 2$.

Задача 3.6. *На рівні значущості* $\alpha = 0{,}05$ *для задачі 2.1 перевірити наявність авторегресії першого порядку з допомогою теста Бреуша-Годфрі.*

○ Знайдемо оцінку $\hat{\rho}$ за формулою (3.28*), використавши дані двох останніх рядків табл. 2.2:

$$\hat{\rho} = -0{,}1474041/1{,}1168 = -0{,}132.$$

Для обчислення S_ε^2 за формулою (3.29) складемо допоміжну табл. 3.3, використавши дані передостаннього рядка табл. 2.2 і співвідношення $\hat{u}_i = -0{,}132u_{i-1}$, де $i = \overline{2{,}10}$.

Таблиця 3.3

i	u_i	$\hat{u}_i$	$u_i - \hat{u}_i$	$(u_i - \hat{u}_i)^2$
1	-0,1938	–	–	–
2	0,7178	0,02558	0,69222	0,47917
3	-0,059	-0,09475	0,03575	0,00128
4	-0,4474	0,00779	-0,45519	0,2072
5	-0,3242	0,05906	-0,38326	0,14689
6	0,099	0,04279	0,05621	0,00316
7	-0,2894	-0,01307	-0,2512	0,0631
8	0,4222	0,03820	0,384	0,14746
9	0,1338	-0,05573	0,18953	0,03592
10	-0,0546	-0,01766	-0,03694	0,00137
Σ	–	–	–	1,08555

Тоді

$$S_\varepsilon^2 = \frac{1}{10-2}\sum_{i=2}^{10}(u_i - \hat{u}_i)^2 = \frac{1}{8}\cdot 1{,}08555 = 0{,}1357.$$

За формулою (3.30) знайдемо незміщену оцінку дисперсії вибіркового коефіцієнта регресії $\hat{\rho}$:

$$S_{\hat{\rho}}^2 = \frac{\sum_{i=2}^{10}(u_i - \hat{u}_i)^2}{(10-2)\sum_{i=2}^{10}u_{i-1}^2} = \frac{1{,}08555}{8\cdot 1{,}1168} = 0{,}1215,$$

звідки $S_{\hat{\rho}} = 0{,}34857$.

Тоді

$$\left|\frac{\hat{\rho}}{S_{\hat{\rho}}}\right| = \left|\frac{-0{,}132}{0{,}34857}\right| = 0{,}3787.$$

Але оскільки

$$\left|\frac{\hat{\rho}}{S_{\hat{\rho}}}\right| = 0{,}3787 < t_{двост.кр.}(0{,}05;8) = 2{,}306,$$

то на рівні $\alpha = 0{,}05$ робимо висновок, що $\rho = 0$, тобто для даних задачі 2.1 відсутня авторегресія першого порядку. ◎

Зауваження. *В більшості сучасних комп'ютерних пакетів застосування тесту серій здійснюється спеціальною командою, а тому не обов'язково оцінювати регресію типу (3.26) «вручну».*

Відмітимо також ефективність Q**-тесту Л'юінга-Бокса**. Розгляд цього тесту можна здійснити після наведення більш детальної інформації про часові ряди.

ІНДИВІДУАЛЬНЕ ЗАВДАННЯ № 2

За умовами індивідуального завдання №1 для рівня значущості $\alpha = 0{,}05$ перевірити виконання передумов 1-4, використавши всі наведені вище тести. При наявності гетероскедастичності — усунути її за допомогою методу зважених найменших квадратів.

§ 4. МНОЖИННИЙ РЕГРЕСІЙНИЙ АНАЛІЗ

1. Класична нормальна лінійна модель множинної регресії.
2. Оцінка параметрів класичної регресійної моделі методом найменших квадратів.
3. Коваріаційна матриця вектора оцінок параметрів регресії та її вибіркова оцінка.
4. Оцінка дисперсії збурень.
5. Знаходження довірчих інтервалів для коефіцієнтів регресії і для базисних та прогнозних даних.
6. Оцінка значущості множинної регресії. Коефіцієнти детермінації R^2 і $\hat{R}^2$.
7. Мультиколінеарність у множинних регресійних моделях.
7.1. Парна і часткова кореляція.
7.1.1. Випадок двох регресорів ($m = 2$).
7.1.2. Загальний випадок.
7.2. Алгоритм Фаррара-Глобера дослідження мультиколінеарності.
8. Усунення або зменшення мультиколінеарності.

Припустимо, що R^2, обчислений для парної регресії, суттєво менший 1. Одна із причин – відсутність у моделі ще однієї або кількох інших пояснюючих змінних, які впливають на результуючу змінну. Узагальнимо постановку задачі.

1. Економічні явища та процеси визначаються великим числом діючих чинників. Це зумовлює необхідність дослідження залежності однієї результуючої змінної y від пояснюючих (незалежних) змінних $x_1, x_2, \ldots, x_m$. Наприклад, валовий регіональний продукт (y) залежить від: величини основних засобів, величини оборотних фондів, величини інвестицій в основний капітал, кількості людей, зайнятих на підприємствах регіону, технологій, які використовуються на підприємствах, ефективності управлінських рішень тощо. Така задача розв'язується з допомогою **множинного регресійного аналізу**.

Об'єктом вивчення цієї теми є така модель множинної лінійної регресії:

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_m x_m + u, \qquad (4.1)$$

де y та u — випадкові величини (u — збурення або залишок), $\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_m$ — невідомі детерміновані параметри, u відображає вплив на y інших факторів, помилки вимірів, помилки вибору моделі.

Нехай $x_{i1}, x_{i2}, \ldots, x_{im}$ $(i = \overline{1,n})$ — спостережені значення пояснюючих змінних. Тоді модель (4.1) набере такого виду:

$$Y_i = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \ldots + \alpha_m x_{im} + U_i, \qquad i = \overline{1,n} \qquad (4.2)$$

Систему n рівнянь (4.2) запишемо у векторно-матричному вигляді:

$$Y_M = X\alpha + U, \qquad (4.2^*)$$

де

$$Y_M = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1m} \\ 1 & x_{21} & x_{22} & \ldots & x_{2m} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nm} \end{pmatrix}, \quad \alpha = \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_m \end{pmatrix}, \quad U = \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_n \end{pmatrix}.$$

Оцінкою цієї моделі за вибіркою $\left\{ y_i, x_{i1}, x_{i2}, \ldots, x_{im}; i = \overline{1,n} \right\}$ обсягом n є векторно-матричне рівняння

$$Y = Xa + E, \qquad (4.3)$$

де $Y = (y_1, y_2, \ldots, y_n)'$, $a = (a_0, a_1, \ldots, a_m)'$, $E = (e_1, e_2, \ldots, e_n)'$, штрих означає операцію транспонування матриці.

Покладемо, що стосовно моделі (4.2*) виконуються такі припущення.

Передумова 1. U — випадковий вектор, змінні $x_1, x_2, \ldots, x_m$ — детерміновані величини, а тому X — детермінована матриця.

Передумова 2. $M(U) = \mathbf{0}_n = (0,0,\ldots,0)'$.

Передумова 3. $\sum_U = M(UU') = \sigma^2 I_n$, де I_n — одинична матриця порядку n, σ — додатна стала, яка підлягає оцінюванню.

Передумова 4. U — нормально розподілений випадковий вектор, тобто $U \sim N_n(\mathbf{0}, \sigma I_n)$.

Передумова 5. Ранг матриці X дорівнює $m+1 < n$.

Означення. *Модель (4.2*), яка задовольняє передумовам 1-5, називається* ***класичною нормальною лінійною моделлю множинної регресії****; якщо ж не виконується тільки передумова 4, то модель називається* ***класичною лінійною моделлю множинної регресії****.*

Відмітимо, що згідно з передумовами 2 і 3 для компонент вектора U виконуються рівності

$$D(U_i) = M(U_i^2) - [M(U_i)]^2 = \sigma^2 - 0 = \sigma^2,$$
$$\operatorname{cov}(U_i, U_j) = 0 \ \ \forall \ i \neq j \ \ i, j = \overline{1,n}.$$

Нагадаємо також, що $\sum_U$ позначає, як і у §2, коваріаційну матрицю n-вимірної випадкової величини U, а

$$U \cdot U' = \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_n \end{pmatrix} (U_1, U_2, \ldots, U_n) = \begin{pmatrix} U_1^2 & U_1U_2 & \ldots & U_1U_n \\ U_1U_2 & U_2^2 & \ldots & U_2U_n \\ \ldots & \ldots & \ldots & \ldots \\ U_nU_1 & U_nU_2 & \ldots & U_n^2 \end{pmatrix}.$$

Нарешті, передумова 5 означає, що обсяг вибірки повинен бути більшим від $m+1$, а всі стовпці матриці X є лінійно незалежними.

Таким чином, при $m = 1$ всі передумови лінійної моделі парної регресії виконуються.

2. Детерміновану складову моделі (4.3) позначимо $\hat{Y} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)'$, тобто

$$\hat{Y} = Xa, \tag{4.4}$$

Тоді критерієм вибору вектора оцінок a згідно з методом найменших квадратів є мінімізація суми квадратів залишків:

$$Q(a) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 = E'E = (Y - Xa)'(Y - Xa) \to \min. \tag{4.5}$$

Оскільки для будь-яких матриць A та B узгодженої вимірності $(A \cdot B)' = B' \cdot A'$, то після розкриття дужок в (4.5) отримаємо

$$Q(a) = Y'Y - a'X'Y - Y'Xa + a'X'Xa.$$

Добуток $Y'Xa$ є скалярною величиною, тому він не змінюється від транспонування:

$$Y'Xa = (Y'Xa)' = a'X'Y.$$

З урахуванням цього умова (4.5) набуде такого вигляду:

$$Q(a) = Y'Y - 2a'X'Y + a'X'Xa \to \min. \tag{4.6}$$

Згідно з необхідною умовою екстремуму функції $Q(a)$ $m+1$ змінних $a_0, a_1, \ldots, a_m$ потрібно прирівняти до нуля всі частинні похідні першого порядку від $Q(a)$ по цих змінних або у матричній формі — вектор частинних похідних:

$$\frac{\partial Q}{\partial a}=\left(\frac{\partial Q}{\partial a_0},\frac{\partial Q}{\partial a_1},\ldots,\frac{\partial Q}{\partial a_m}\right).$$

Нехай b та c — k-вимірні вектор-стовпці, A — симетрична матриця порядку k. Можна довести такі рівності:

$$\frac{\partial}{\partial b}(b'c)=c', \qquad \frac{\partial}{\partial b}(b'Ab)=2b'A.$$

Тому, покладаючи $b=a,\ c=X'Y,\ A=X'X,\ k=m+1,$ із (4.6) отримаємо

$$\frac{\partial Q}{\partial a}=-2(X'Y)'+2a'X'X=-2Y'X+2a'X'X.$$

Врахувавши необхідну умову екстремуму

$$-2Y'X+2a'X'X=(0,0,\ldots,0),$$

або

$$-2X'Y+2X'Xa=(0,0,\ldots 0)',$$

прийдемо до системи **нормальних рівнянь у матричній формі для визначення вектора** a:

$$X'Xa=X'Y. \tag{4.7}$$

Згідно з передумовою 5 $(m+1)\times(m+1)$ — матриця $X'X$ є невиродженою, тому розв'язком рівняння (4.7) є вектор

$$a=(X'X)^{-1}X'Y, \tag{4.8}$$

де $(X'X)^{-1}$ — обернена матриця до матриці $X'X$.

Основні властивості отриманих оцінок (4.8) визначаються наступним твердженням.

Теорема Гаусса-Маркова. Нехай стосовно моделі (4.2*) виконуються передумови 1-3, 5. Тоді оцінки (4.8) вектора параметрів α мають найменшу дисперсію в класі лінійних незміщених оцінок.

Доведення цієї теореми можна знайти, зокрема, в [16].

Якщо вектор a знайдено, тоді **вибіркове рівняння множинної регресії** можна зобразити у такому вигляді:

$$\hat{y}=X_0'a=a_0+a_1x_1+a_2x_2+\ldots+a_mx_m, \tag{4.9}$$

де $\hat{y}$ — групова (умовна) середня змінної y при заданому векторі значень пояснюючих змінних,

$$X_0'=(1,x_{01},x_{02},\ldots,x_{0m})=(1,x_1,x_2,\ldots,x_m).$$

Задача 4.1. *Підприємство, що складається із багатьох філій, досліджує залежність свого річного товарообігу у (млн. грн.) від торгової*

площі своїх філій x_1 (тис. кв. м.) і середньоденної інтенсивності потоку покупців (тис. покупців за день). Дані 10 філій наведені у табл. 4.1. Оцінивши параметри лінійної регресійної моделі $y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + u$, знайти вибіркове рівняння множинної регресії.

Таблиця 4.1

Номер філії	Значення		
	y	x_1	x_2
1	6,9	1,2	10,8
2	7,1	1,3	9,9
3	7	1,1	13,7
4	8,4	1,5	13,9
5	4,3	0,8	8,5
6	5,8	0,9	12,4
7	7,7	1,3	12,3
8	3,2	0,5	11
9	1,5	0,2	8,3
10	3,1	0,6	9,3

○ Введемо позначення:

$$Y = \begin{pmatrix} 6,9 \\ 7,1 \\ 7 \\ 8,4 \\ 4,3 \\ 5,8 \\ 7,7 \\ 3,2 \\ 1,5 \\ 3,1 \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 1,2 & 10,8 \\ 1 & 1,3 & 9,9 \\ 1 & 1,1 & 13,7 \\ 1 & 1,5 & 13,9 \\ 1 & 0,8 & 8,5 \\ 1 & 0,9 & 12,4 \\ 1 & 1,3 & 12,3 \\ 1 & 0,5 & 11 \\ 1 & 0,2 & 8,3 \\ 1 & 0,6 & 9,3 \end{pmatrix},$$

$$X' = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1,2 & 1,3 & 1,1 & 1,5 & 0,8 & 0,9 & 1,3 & 0,5 & 0,2 & 0,6 \\ 10,8 & 9,9 & 13,7 & 13,9 & 8,5 & 12,4 & 12,3 & 11 & 8,3 & 9,3 \end{pmatrix}.$$

Знаходимо добутки $X' \cdot X$ та $X'Y$. В результаті отримаємо:

$$X'X = \begin{pmatrix} 10 & 9,4 & 110,1 \\ 9,4 & 10,38 & 108,44 \\ 110,1 & 108,44 & 1249,23 \end{pmatrix},\ X'Y = \begin{pmatrix} 55 \\ 60,24 \\ 637,13 \end{pmatrix}.$$

Обернена матриця до матриці $X'X$ матиме вигляд:

$$(X'X)^{-1} = \begin{pmatrix} 3,6924 & 0,6007 & -0,3776 \\ 0,6007 & 1,1320 & -0,1512 \\ -0,3776 & -0,1512 & 0,0472 \end{pmatrix}.$$

Значення елементів вектора a знайдемо, перемноживши матрицю $(X'X)^{-1}$ на вектор $X'Y$:

$$a = (X'X)^{-1} X'Y = \begin{pmatrix} -1,3121 \\ 4,8961 \\ 0,1962 \end{pmatrix}.$$

Отже, ми отримали вибіркове рівняння множинної регресії:

$$\hat{y} = -1,3121 + 4,8961x_1 + 0,1962x_2.$$

Нагадаємо, що добутком $m \times n$-матриці $A = (a_{ij})$, $i = \overline{1,m}$, $j = \overline{1,n}$, на $n \times k$-матрицю $B = (b_{ij})$, $i = \overline{1,n}$, $j = \overline{1,k}$, є $m \times k$-матриця $C = (c_{ij})$, $i = \overline{1,m}$, $j = \overline{1,k}$, для елементів якої виконуються рівності

$$c_{ij} = \sum_{s=1}^{n} a_{is} b_{sj},\ i = \overline{1,m},\ j = \overline{1,n}.$$

При цьому використовується позначення $C = AB$.

У випадку здійснення розрахунків «вручну» рекомендується використання табл. 4.2 та 4.3.

Елементи першого рядка матриці $X'X$ отримуються шляхом множення чисел першого рядка матриці X' на відповідні елементи стовпців матриці X з наступним додаванням добутків, тобто 10; 9,4; 110,1 (див. останній рядок табл. 4.2).

Елементи другого рядка матриці $X'X$ — це сума добутків чисел другого рядка на відповідні елементи стовпців матриці X: 9,4; 10,38; 108,44. Нарешті, елементи третього рядка матриці $X'X$ (110,1; 108,44; 1249,23) також визначаються із останнього рядка табл. 4.2.

Таблиця 4.2

i	x_{i1}	x_{i2}	x_{i1}^2	$x_{i1}x_{i2}$	x_{i2}^2
1	1,2	10,8	1,44	12,96	116,64
2	1,3	9,9	1,69	12,87	98,01
3	1,1	13,7	1,21	15,07	187,69
4	1,5	13,9	2,25	20,85	193,21
5	0,8	8,5	0,64	6,8	72,25
6	0,9	12,4	0,81	11,16	153,76
7	1,3	12,3	1,69	15,99	151,29
8	0,5	11	0,25	5,5	121
9	0,2	8,3	0,04	1,66	68,89
10	0,6	9,3	0,36	5,58	86,49
$\sum$	9,4	110,1	10,38	108,44	1249,23

Таблиця 4.3

i	y	$y_i x_{i1}$	$y_i x_{i2}$
1	6,9	8,28	74,52
2	7,1	9,23	70,29
3	7	7,7	95,9
4	8,4	12,6	116,76
5	4,3	3,44	36,55
6	5,8	5,22	71,92
7	7,7	10,01	94,71
8	3,2	1,6	35,2
9	1,5	0,3	12,45
10	3,1	1,86	28,83
$\sum$	55	60,24	637,13

Елементи вектор-стовпця $X'Y$ виписуються із останнього рядка табл. 4.2:

$$X'Y = \begin{pmatrix} 55 \\ 60{,}24 \\ 637{,}13 \end{pmatrix}.$$

Нагадаємо також послідовність дій при знаходженні оберненої матриці A:

1) обчислюється визначник $|A|$ (при цьому A^{-1} не існує, якщо $|A| = 0$);

2) для транспонованої матриці A' знаходиться приєднана матриця $\tilde{A} = (A_{ij})$, де A_{ij} — алгебраїчні доповнення елементів a_{ij} матриці A', тобто $A_{ij} = (-1)^{i+j} |.|$, $|.|$ — визначник, який отримується із $|A'|$ шляхом викреслення i-го рядка та j-го стовпця;

3) знаходиться $A^{-1} = \frac{1}{|A|}\tilde{A}$. ◎

3. Дисперсії оцінок параметрів визначають точність рівняння множинної регресії. Для їх вимірювання розглядають так звану **коваріаційну матрицю вектора оцінок параметрів** $\sum_a$, яка є матричним аналогом дисперсії однієї випадкової величини:

$$\sum_a = \begin{pmatrix} \sigma_{00} & \sigma_{01} & \dots & \sigma_{0m} \\ \sigma_{10} & \sigma_{11} & \dots & \sigma_{1m} \\ \dots & \dots & \dots & \dots \\ \sigma_{m0} & \sigma_{m1} & \dots & \sigma_{mm} \end{pmatrix},$$

де елементи σ_{ij} — коваріації оцінок параметрів a_i та a_j: $\sigma_{ij} = \operatorname{cov}(a_i, a_j),\ i \neq j,\ i, j = \overline{0, m},\ \sigma_{ii} = D(a_i),\ i = \overline{0, m}.$ Зауважимо, що Σ_a є симетричною матрицею, оскільки $\operatorname{cov}(a_i, a_j) = \operatorname{cov}(a_j, a_i)$.

Знайдемо матрицю $\sum_a$ на підставі конкретної вибірки. Із урахуванням (4.2*) для нефіксованої вибірки отримаємо зображення вектора оцінок (4.8):

$$a = (X'X)^{-1} X'Y_M = (X'X)^{-1} X'(X\alpha + U) = (X'X)^{-1} X'X\alpha + \\ + (X'X)^{-1} X'U = I_{m+1}\alpha + (X'X)^{-1} X'U$$

або

$$a = \alpha + (X'X)^{-1} X'U. \qquad (4.10)$$

Із цієї рівності випливають такі висновки:

1) оцінки параметрів (4.8), знайдені за **нефіксованою** вибіркою, будуть містити випадкові помилки;

2) згідно із передумовами 1 та 2

$$M(a) = \alpha,$$

тобто вектор a є незміщеною оцінкою вектора параметрів α.

У скороченому вигляді коваріаційна матриця вектора оцінок параметрів має такий вигляд:

$$\sum_a = M\left[(a - \alpha)(a - \alpha)'\right], \qquad (4.11)$$

оскільки

$$\sigma_{ij} = \operatorname{cov}(a_i, a_j) = M[(a_i - M(a_i))(a_j - M(a_j))] = M[(a_i - \alpha_i)(a_j - \alpha_j)],$$

$$(a - \alpha)(a - \alpha)' = \begin{pmatrix} a_0 - \alpha_0 \\ a_1 - \alpha_1 \\ \vdots \\ a_m - \alpha_m \end{pmatrix} (a_0 - \alpha_0, a_1 - \alpha_1, \dots, a_m - \alpha_m) =$$

$$= \begin{pmatrix} (a_0-\alpha_0)(a_0-\alpha_0) & (a_0-\alpha_0)(a_1-\alpha_1) & \dots & (a_0-\alpha_0)(a_m-\alpha_m) \\ (a_1-\alpha_1)(a_0-\alpha_0) & (a_1-\alpha_1)(a_1-\alpha_1) & \dots & (a_1-\alpha_1)(a_m-\alpha_m) \\ \dots & \dots & \dots & \dots \\ (a_m-\alpha_m)(a_0-\alpha_0) & (a_m-\alpha_m)(a_1-\alpha_1) & \dots & (a_m-\alpha_m)(a_m-\alpha_m) \end{pmatrix}.$$

Враховуючи (4.10), детермінованість матриці X, передумову 3 і симетричність матриці $X'X$, отримаємо із (4.11)

$$\sum_a = M\{[(X'X)^{-1}X'U][(X'X)^{-1}X'U]'\} = M\{(X'X)^{-1}X'UU'X[(X'X)^{-1}]'\} =$$
$$= (X'X)^{-1}X'M(UU')X(X'X)^{-1} = \sigma^2(X'X)^{-1}X'I_nX(X'X)^{-1} = \sigma^2(X'X)^{-1}.$$

Тобто, остаточно

$$\sum_a = \sigma^2(X'X)^{-1}. \qquad (4.12)$$

Зауважимо, що в ланцюжку рівностей використано тотожність $(ABC)' = C'B'A'$ для матриць з узгодженими вимірностями.

Висновок: *з допомогою матриці* $(X'X)^{-1}$ *визначається не тільки сам вектор* a *оцінок параметрів (4.8), але і дисперсії, і коваріації його компонент.*

Із (4.12) зокрема отримуємо:

$D(a_i) = \sigma^2[(X'X)^{-1}]_{ii}$, $i = \overline{0,m}$, *– відповідні діагональні елементи матриці* Σ_a.

4. Параметр σ^2, який фігурує у передумовах 3, 4, а також у рівності (4.12) — **невідомий**. Як і у випадку парної регресії його необхідно **оцінити.**

Розглянемо вектор залишків E, який визначається із (4.3):

$$E = Y - Xa.$$

Для **нефіксованої** вибірки $Y = Y_M$, а тому з урахуванням (4.2*) і (4.8)

$$E = X\alpha + U - X[(X'X)^{-1}X'(X\alpha + U)] =$$
$$= X\alpha + U - X(X'X)^{-1}X'X\alpha - X(X'X)^{-1}X'U =$$
$$= X\alpha + U - X\alpha - X(X'X)^{-1}X'U = U - X(X'X)^{-1}X'U,$$

тобто

$$E = U - X(X'X)^{-1}X'U,$$

звідки

$$E' = U' - U'X(X'X)^{-1}X'.$$

Тоді

$$M(E'E) = M\{[U' - U'X(X'X)^{-1}X'][U - X(X'X)^{-1}X'U]\} =$$
$$= M(U'U) - M[U'X(X'X)^{-1}X'U] - M[U'X(X'X)^{-1}X'U] +$$
$$+ M[U'X(X'X)^{-1}X'X(X'X)^{-1}X'U].$$

Оскільки два останні доданки знищуються, то

$$M(E'E) = M(U'U) - M[U'X(X'X)^{-1}X'U]. \qquad (4.13)$$

Перший доданок правої частини (4.13) дорівнює $n\sigma^2$, оскільки згідно із передумовою 3

$$M(U'U) = M\left(\sum_{i=1}^{n} U_i^2\right) = \sum_{i=1}^{n} M\left(U_i^2\right) = \sum_{i=1}^{n} \sigma^2 = n\sigma^2. \qquad (4.14)$$

Матриця $B = X(X'X)^{-1}X'$ симетрична, тому $U'BU$ – квадратична форма $\sum\limits_{i,j=1}^{n} b_{ij}U_iU_j$, де b_{ij} – детерміновані величини. Тоді

$$M(U'BU) = M\left(\sum_{i,j=1}^{n} b_{ij}U_iU_j\right) = \sum_{i,j=1}^{n} b_{ij}M\left(U_iU_j\right) =$$
$$= \sum_{i=1}^{n} b_{ii}M\left(U_i^2\right) + \sum_{\substack{i,j=1 \\ (i\neq j)}}^{n} b_{ij}M\left(U_iU_j\right).$$

Згідно із передумовою 3 останній доданок дорівнює нулю, а $M\left(U_i^2\right) = \sigma^2$ ($i = \overline{1,n}$). Позначимо $Sp\,B$ або слід квадратної матриці — сума діагональних елементів матриці B. Для будь-яких матриць K та M, для яких визначені добутки KM та MK, виконується рівність $Sp(KM) = Sp(MK)$, використавши яку, отримаємо

$$Sp\,B = Sp(X(X'X)^{-1}X') = Sp((X'X)^{-1}X'X) = Sp\,I_{m+1} = m+1.$$

Таким чином, $M(U'BU) = (m+1)\sigma^2$, і остаточно із (4.13) та (4.14) отримаємо

$$M(E'E) = (n-m-1)\sigma^2, \qquad (4.15)$$

звідки визначається незміщена оцінка S_e^2 параметра σ^2 або **вибіркова залишкова дисперсія** S_e^2:

$$S_e^2 = \frac{E'E}{n-m-1} = \frac{\sum_{i=1}^{n} e_i^2}{n-m-1} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-m-1} \qquad (4.16)$$

Число $n-m-1$ називається **числом ступенів вільності**.

Можна довести, що оцінки a і S_e^2 параметрів α і σ^2 не корелюють між собою як випадкові величини, тобто $\operatorname{cov}(a, S_e^2) = 0$. Тоді при виконанні передумови 4 a і S_e^2 є незалежними випадковими величинами.

Задача 4.2. *Для задачі 4.1 знайти точкову оцінку дисперсії збурень та точкову оцінку для коваріаційної матриці вектора оцінок параметрів.*

○ Незміщену точкову оцінку невідомого параметра σ^2 знайдемо за формулою (4.16), попередньо обчисливши

$\hat{y}_i = -1{,}3121 + 4{,}8961x_{i1} + 0{,}1962x_{i2}$ та $e_i^2 = (y_i - \hat{y}_i)^2$, $i = \overline{1,10}$, (табл. 4.4).

Таблиця 4.4

i	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	$\sum$
y_i	6,9	7,1	7	8,4	4,3	5,8	7,7	3,2	1,5	3,1	–
$\hat{y}_i$	6,6822	6,9952	6,7616	8,7592	4,2725	5,5273	7,4661	3,2942	1,2956	3,4502	–
e_i	0,2178	0,1048	0,2384	-0,3592	0,0275	0,2727	0,2339	-0,0942	0,2044	-0,3502	0,4959
e_i^2	0,0474	0,011	0,0568	0,129	0,0008	0,0744	0,0547	0,0089	0,0418	0,1226	0,5474

Зауваження*. Раніше вже говорилося (див. задачу 2.1) про накопичення додатних похибок у різницях $y_i - \hat{y}_i$, що приводить величину $\left|\sum (y_i - \hat{y}_i)\right|$ до незначного перевищення нуля. Значенням $\bar{e} = \sum e_i / 10 = 0{,}04959$ можна ігнорувати (вважати практично рівним нулю).*

Використавши підсумок останнього рядка, отримаємо:

$$S_e^2 = \frac{0{,}5474}{10-2-1} = 0{,}0782.$$

Оскільки в (4.12) параметр σ^2 невідомий, то використаємо його незміщену точкову оцінку S_e^2 і отримаємо оцінку для коваріаційної матриці Σ_a:

$$S_e^2 (X'X)^{-1} = 0{,}0782 \cdot \begin{pmatrix} 3{,}6924 & 0{,}6007 & -0{,}3776 \\ 0{,}6007 & 1{,}1320 & -0{,}1512 \\ -0{,}3776 & -0{,}1512 & 0{,}0472 \end{pmatrix} =$$

$$= \begin{pmatrix} 0{,}2887 & 0{,}045 & -0{,}0295 \\ 0{,}045 & 0{,}0885 & -0{,}0118 \\ -0{,}0295 & -0{,}0118 & 0{,}0037 \end{pmatrix}. \quad \text{◎}$$

5. З'ясуємо значущість коефіцієнтів регресії a_i $(i = \overline{0,m})$ і побудуємо довірчі інтервали для параметрів моделі α_i при умові їх значущості.

Згідно (4.12) і (4.16) незміщена оцінка $S^2_{a_i}$ дисперсії $\sigma^2_{a_i}$ коефіцієнта регресії a_i визначається за формулою

$$S^2_{a_i} = S^2_e[(X'X)^{-1}]_{ii},\ i = \overline{0,m},$$

де S^2_e – незміщена оцінка параметра σ^2, $[(X'X)^{-1}]_{ii}$ – i-й діагональний елемент матриці $(X'X)^{-1}$.

Середнє квадратичне відхилення або **стандартна помилка** коефіцієнта регресії a_i знаходиться за формулою

$$S_{a_i} = S_e\sqrt{[(X'X)^{-1}]_{ii}},\ i = \overline{0,m}. \qquad (4.17)$$

Оскільки випадкова величина $(a_i - \alpha_i)/S_{a_i}$ для нефіксованої вибірки розподілена за законом Ст'юдента із $k = n - m - 1$ ступенями вільності, то a_i значуще відрізняється від нуля (гіпотеза $H_0 : a_i = 0$ відхиляється) на рівні значущості α, якщо

$$|t_{спост.}| = \frac{|a_i|}{S_{a_i}} > t_{кр.}, \qquad (4.18)$$

де $t_{кр.} = t_{двост.кр.}(1-\alpha; n-m-1)$ – табличне значення t – критерія Ст'юдента, визначене на рівні значущості α при числі ступенів вільності $k = n - m - 1$.

Якщо виконується нерівність (4.18), тоді є сенс будувати довірчий інтервал для параметра α_i, який визначається подвійною нерівністю

$$a_i - t_{кр.}S_{a_i} < \alpha_i < a_i + t_{кр.}S_{a_i}. \qquad (4.19)$$

Базисними даними назвемо вектори $X'_i = (1, x_{i1}, x_{i2}, \ldots, x_{im})$, $i = \overline{1,n}$, які визначаються відповідними рядками матриці X. Тоді на цих базисних даних можна обчислити емпіричні значення регресії (у відповідності з (4.4)):

$$\hat{Y} = \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \dots \\ \hat{y}_n \end{pmatrix} = \begin{pmatrix} X_1' \\ X_2' \\ \dots \\ X_n' \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \dots \\ a_m \end{pmatrix} = Xa$$

та теоретичні значення регресії

$$Y^T = \begin{pmatrix} y_1^T \\ y_2^T \\ \dots \\ y_n^T \end{pmatrix} = \begin{pmatrix} X_1' \\ X_2' \\ \dots \\ X_n' \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \dots \\ \alpha_m \end{pmatrix} = X\alpha .$$

Для нефіксованої вибірки згідно з (4.10) компоненти вектора $\hat{Y}$ є випадковими величинами. Знайдемо коваріаційну матрицю вектора $\hat{Y}$, використавши (4.10) та передумови 2 і 3:

$$M(\hat{Y}) = M(Xa) = M\left\{X\left[\alpha + (X'X)^{-1}X'U\right]\right\} =$$

$$= M\left[X\alpha + X(X'X)^{-1}X'U\right] = X\alpha = Y^T,$$

$$\Omega_{\hat{Y}} = M\left[(\hat{Y} - Y^T)(\hat{Y} - Y^T)'\right] =$$

$$= M\left\{\left[X\alpha + X(X'X)^{-1}X'U - X\alpha\right]\left[X\alpha + X(X'X)^{-1}X'U - X\alpha\right]'\right\} =$$

$$= M\left[X(X'X)^{-1}X'UU'X(X'X)^{-1}X'\right] = \sigma^2 X(X'X)^{-1}X'.$$

Отже, коваріаційна матриця оціненої регресії $\hat{Y}$ на базисних даних має такий вигляд:

$$\sum_{\hat{Y}} = \sigma^2 X(X'X)^{-1}X',$$

а **оцінена** коваріаційна матриця вектора $\hat{Y}$ –

$$\hat{\sum}_{\hat{Y}} = S_e^2 X(X'X)^{-1}X'.$$

Діагональні елементи цієї матриці, рівні $S_{\hat{y}_1}^2, S_{\hat{y}_2}^2, \dots, S_{\hat{y}_n}^2$, є оцінками дисперсій $\sigma_{\hat{y}_1}^2, \sigma_{\hat{y}_2}^2, \dots, \sigma_{\hat{y}_n}^2$ на базисних даних, тобто

$$S_{\hat{y}_i}^2 = S_e^2\left[X(X'X)^{-1}X'\right]_{ii}, \; i = \overline{1,n}. \qquad (4.20)$$

Величини $S_{\hat{y}_1}, S_{\hat{y}_2}, \dots, S_{\hat{y}_n}$ називаються **стандартними** помилками значень $\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n$ регресії на базисних даних, тобто

$$S_{\hat{y}_i} = \sqrt{S_e^2 X_i'(X'X)^{-1}X_i}, \; i = \overline{1,n}. \qquad (4.21)$$

Довірчі інтервали з надійністю γ для значень залежної змінної на базисних даних мають такий вид:

$$\left(\hat{y}_i - t(\gamma; n-m-1)S_{\hat{y}_i};\ \hat{y}_i + t(\gamma; n-m-1)S_{\hat{y}_i}\right), i = \overline{1,n}. \qquad (4.22)$$

Якщо отримане рівняння множинної регресії адекватне реальній дійсності (узгоджується із даними вибірки), то з його допомогою можна здійснювати прогноз. Нехай незалежні змінні на прогнозний період набирають значення $x_{1,n+1}, x_{2,n+1}, \ldots, x_{m,n+1}$ відповідно. Тоді **точковий прогноз** залежної змінної має такий вид

$$\hat{y}_{n+1} = a_0 + a_1 x_{1,n+1} + a_2 x_{2,n+1} + \ldots + a_m x_{m,n+1}$$

або

$$\hat{y}_{n+1} = X_{n+1} a, \text{ де } X_{n+1} = \left(1, x_{1,n+1}, x_{2,n+1}, \ldots, x_{m,n+1}\right).$$

Прогнозне значення $\hat{y}_{n+1}$ є точковою оцінкою невідомого значення y_{n+1}.

За аналогією із випадком парної регресії можна довести, що довірчий інтервал, який з надійністю γ покриває невідоме значення y_{n+1}, має такий вигляд:

$$\left(\hat{y}_{n+1} - t(\gamma; n-m-1)S_{\hat{y}_{n+1}};\ \hat{y}_{n+1} + t(\gamma; n-m-1)S_{\hat{y}_{n+1}}\right), \qquad (4.23)$$

де

$$S_{\hat{y}_{n+1}} = \sqrt{S_e^2\left[1 + X_{n+1}(X'X)^{-1}X'_{n+1}\right]} \qquad (4.24)$$

і називається середньоквадратичною (стандартною) помилкою прогнозу.

Довірчий інтервал для параметра σ^2 в множинній регресії будується аналогічно парній регресії за формулою (2.29) із відповідною зміною числа ступенів вільності критерію χ^2.

Задача 4.3. *На рівні значущості $\alpha = 0,05$ для задачі 3.1 перевірити значущість коефіцієнтів регресії α_0, α_1, α_2 та побудувати для них довірчі інтервали з надійністю $\gamma = 0,95$.*

○ Знайдемо середні квадратичні помилки коефіцієнтів регресії a_0, a_1, a_2 за формулою (4.17):

$$S_{a_0} = S_e\sqrt{[(X'X)^{-1}]_{00}} = \sqrt{0,0782}\sqrt{3,6924} = 0,5374;$$

$$S_{a_1} = S_e\sqrt{[(X'X)^{-1}]_{11}} = \sqrt{0,0782}\sqrt{1,132} = 0,2975;$$

$$S_{a_2} = S_e\sqrt{[(X'X)^{-1}]_{22}} = \sqrt{0,0782}\sqrt{0,0472} = 0,0608.$$

Тоді спостережені значення критерію:

$$\frac{|a_0|}{S_{a_0}} = \frac{1{,}3121}{0{,}5374} = 2{,}4416,\ \frac{|a_1|}{S_{a_1}} = \frac{4{,}8961}{0{,}2975} = 16{,}4575,\ \frac{|a_2|}{S_{a_2}} = \frac{0{,}1962}{0{,}0608} = 3{,}227.$$

Критична точка для двосторонньої критичної області $t_{кр.} = t_{двост.кр.}(\alpha; n-m-1)$ при значеннях $\alpha = 0{,}05,\ k = n-m-1 = 7$ знаходиться за верхньою частиною табл. 3 додатків: $t_{кр.} = 2{,}365.$

Оскільки, $2{,}4416 > t_{кр.} = 2{,}365,$ $16{,}4575 > t_{кр.} = 2{,}365$ і $3{,}227 > t_{кр.} = 2{,}365$, то на рівні значущості $\alpha = 0{,}05$ робимо висновок, що $a_0 \neq 0,\ a_1 \neq 0$ і $a_2 \neq 0.$

Згідно з (4.19) для параметрів регресії $a_0,\ a_1$ та a_2 матимемо довірчі інтервали:

$$-1{,}3121 - 2{,}365 \cdot 0{,}5374 < \alpha_0 < -1{,}3121 + 2{,}365 \cdot 0{,}5374,$$
$$4{,}8961 - 2{,}365 \cdot 0{,}2975 < \alpha_1 < 4{,}8961 + 2{,}365 \cdot 0{,}2975,$$
$$0{,}1962 - 2{,}365 \cdot 0{,}0608 < \alpha_2 < 0{,}1962 + 2{,}365 \cdot 0{,}0608$$

або остаточно

$$-2{,}5831 < \alpha_0 < -0{,}0411,\quad 4{,}1925 < \alpha_1 < 5{,}6,\quad 0{,}0524 < \alpha_2 < 0{,}34.$$ ◎

6. В п. 2.9 розглянуто **коефіцієнт детермінації** R^2 для оцінювання адекватності регресійної моделі, міри якості рівняння регресії, а також характеристики його прогностичної сили.

Коефіцієнт детермінації або **множинний коефіцієнт детермінації** визначається за формулами (2.46) або (2.46*):

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{n\sigma_y^2};\quad R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n\sigma_y^2} = 1 - \frac{\sum_{i=1}^{n} e_i^2}{n\sigma_y^2}, \qquad (4.25)$$

де $\hat{y}_i = a_0 + \sum_{j=1}^{m} a_j x_{ji},\ e_i = y_i - \hat{y}_i.$

Як і у випадку парної регресії R^2 характеризує частку варіації залежної змінної, обумовленої регресією. При цьому чим ближче R^2 до 1, тим краще регресія відображає залежність між пояснюючими (незалежними) та залежними змінними.

Слід врахувати **застереження**: якщо в моделі (4.1) $\alpha_0 = 0$, тоді використовувати R^2 неможливо: в загальному випадку R^2 може виходити навіть за межі проміжку [0; 1].

Критерій значущості на рівні α коефіцієнта детермінації або рівняння регресії має такий вид:

$$F = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 (n-m-1)}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 m} > F(\alpha; m; n-m-1), \qquad (4.26)$$

де $F(\alpha; m; n-m-1)$ – табличне значення F-критерія Фішера-Снедокора (табл. 5 додатків).

Якщо відомий коефіцієнт детермінації R^2, то критерій (4.26) можна записати у такому вигляді:

$$F = \frac{R^2 (n-m-1)}{(1-R^2)m} > F(\alpha; m; n-m-1). \qquad (4.26^*)$$

Задача 4.4. *За даними задачі 4.1 знайти* R^2 *і на рівні значущості* $\alpha = 0,05$ *перевірити його значущість.*

○ Обчислимо $n\sigma_y^2 = \sum_{i=1}^{n}(y_i - \overline{y})^2 = 48,54$, де $n = 10$. Використавши другу формулу (4.25) і підсумок останнього рядка табл. 4.4, отримаємо:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{n\sigma_y^2} = 1 - \frac{0,5474}{48,54} = 0,9887.$$

За табл. 5 додатків знаходимо критичну точку $F(0,05; 2; 7) = 19,35$, врахувавши, що $n = 10$, $m = 2$. Використаємо співвідношення (4.26*):

$$F = \frac{0,9887 \cdot 7}{(1-0,9887) \cdot 2} = 298,315 > 19,35.$$

Отже, на рівні $\alpha = 0,05$ можна зробити висновок про значущість коефіцієнта детермінації і рівняння регресії. ◎

Підсумком задач 4.1-4.4 є таке емпіричне рівняння регресії:

$$\hat{y} = \underset{(0,5374)}{-1,3121} + \underset{(0,2975)}{4,8961} x_1 + \underset{(0,0608)}{0,1962} x_2;\ R^2 = 0,9887, \qquad (4.27)$$

де в дужках вказані середні квадратичні (стандартні) помилки відповідних коефіцієнтів регресії.

Припустимо, що за даними задачі 4.1 досліджується залежність y тільки від однієї пояснюючої змінної x_1. Використовуючи взірець розв'язування задачі 2.1, можна отримати (перевірте!) емпіричне рівняння

$$\hat{y} = \underset{(0,4595)}{0,522} + \underset{(0,451)}{5,5311} x_1;\ R^2 = 0,948. \qquad (4.28)$$

Рівняння (4.27) показує, що при збільшенні тільки торговельної площі x_1 (при незмінному x_2) на 1 тис.м^2 річний товарообіг збільшиться в середньому на 4,8961 млн. грн., а при збільшенні тільки середньоденної інтенсивності потоку покупців (при незмінній x_1) на 1 тис. людей за день – в середньому на 0,1962 млн. грн.

Приєднання в регресійну модель (4.27) нової пояснюючої змінної x_2 змінило коефіцієнт регресії a_1 з 5,5311 для парної регресії до 4,8961 – для множинної регресії. В цьому нічого дивного немає, оскільки у випадку рівняння (4.27) коефіцієнт регресії оцінює приріст залежної змінної y при зміні на одиницю пояснюючої змінної x_1 в «чистому» виді, тобто незалежно від x_2. В той же час у випадку парної регресії (4.28) коефіцієнт a_1 враховує вплив на y не тільки змінної x_1, але й опосередковано кореляційно пов'язаної з нею змінної x_2.

Практично важливою є задача порівняння впливу на залежну змінну пояснюючих змінних з різними одиницями виміру. В таких випадках використовують **стандартизовані коефіцієнти регресії** a'_j і **коефіцієнти еластичності** E_j $(j=\overline{1,m})$:

$$a'_j = a_j \frac{\sigma_{x_j}}{\sigma_y}, \tag{4.29}$$

$$E_j = a_j \frac{\overline{x}_j}{\overline{y}}. \tag{4.30}$$

Стандартизований коефіцієнт регресії a'_j показує, на скільки величин σ_y зміниться в середньому залежна змінна y при збільшенні тільки змінної x_j на σ_{x_j}, а коефіцієнт еластичності E_j – на скільки відсотків (від середньої) зміниться в середньому y при збільшенні тільки x_j на 1%.

Задача 4.5. *За даними задачі 4.1 порівняти роздільний вплив на річний товарообіг двох факторів – торгової площі і середньоденної інтенсивності потоку покупців.*

○ Для порівняння впливу кожної із пояснюючих змінних за формулою (4.29) обчислимо стандартизовані коефіцієнти регресії:

$$a'_1 = 4,8961 \cdot \frac{0,3929}{2,2045} = 0,8726, \quad a'_2 = 0,1962 \cdot \frac{1,9243}{2,2045} = 0,1713,$$

а за формулою (4.30) – коефіцієнти еластичності:

$$E_1 = 4,8961 \cdot \frac{0,94}{5,5} = 0,8368, \;\; E_2 = 0,1962 \cdot \frac{11,01}{5,5} = 0,3928.$$

Тут ми скористалися даними останніх рядків таблиць 4.2 і 4.3 для розрахунків необхідних характеристик змінних:

$$\overline{x}_1 = 0,94, \; \overline{x}_2 = 11,01, \; \overline{y} = 5,5, \; \sigma_{x_1} = 0,3929, \; \sigma_{x_2} = 1,9243, \; \sigma_y = 2,2045.$$

Отже, збільшення торгової площі і середньодобової інтенсивності потоку покупців тільки на одно σ_{x_1} або на одно σ_{x_2} збільшує в середньому річний товарообіг відповідно на $0,8726\sigma_y$ або на $0,1713\sigma_y$, а збільшення цих змінних на 1% (від своїх середніх значень) призводить в середньому до росту річного товарообігу відповідно на 0,8368% і 0,3928%. Таким чином, по обох показниках на річний товарообіг більший вплив виявляє фактор «торгова площа» в порівнянні з фактором «середньоденна інтенсивність потоку покупців». ◎

7. Передумова 5 характерна для лінійної моделі **множинної** регресії і перевірка її виконання – значно складніший процес, ніж це може здатися на перший погляд.

Під **мультиколінеарністю** будемо розуміти високу корельованість незалежних змінних. Мультиколінеарність може проявлятися у **функціональній** (явній) і **стохастичній** (прихованій) формах.

При функціональній формі мультиколінеарності хоча б один із парних зв'язків між незалежними змінними є лінійним функціональним зв'язком. У цьому випадку передумова 5 не виконується. Це приводить до неможливості розв'язування системи нормальних рівнянь (матриця $X'X$ – вироджена), а отже, отримання оцінок параметрів регресійної моделі.

Однак в економічних дослідженнях мультиколінеарність частіше проявляється у **стохастичній формі**, коли між хоча б двома пояснюючими змінними існує тісний кореляційний зв'язок. Матриця $X'X$ у цьому випадку є не виродженою, але її визначник — дуже мале число. Разом з тим вектор оцінок a і його коваріаційна матриця $\sum_a$ у відповідності із формулами (4.8) і (4.12) пропорційні матриці $(X'X)^{-1}$, а тому їх елементи обернено пропорційні величині визначника $|X'X|$. В результаті отримують значні середні квадратичні відхилення (стандартні помилки) коефіцієнтів регресії a_0, a_1,..., a_m і оцінка їх значущості за критерієм Ст'юдента не має

сенсу, хоча у цілому регресійна модель може виявитися значущою за критерієм Фішера-Снедокора.

Більше того, оцінки стають дуже чутливими до незначних змін результатів спостережень і обсягу вибірки. Рівняння регресії у цьому випадку, як правило, не має реального сенсу, оскільки деякі із його коефіцієнтів можуть мати неправильні з точки зору економічної теорії знаки і невиправдано великі значення.

Для формулювання критерію наявності мультиколінеарності необхідно попередньо розглянути допоміжні означення.

7.1. Парна і часткова кореляція

7.1.1. Випадок двох регресорів ($m = 2$).

У випадку парної регресії у рамках лінійної моделі оцінкою міри сили зв'язку між змінними є вибірковий коефіцієнт кореляції

$$r_{yx} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{n \sigma_x \sigma_y}. \qquad (4.25)$$

Розглянемо лінійну модель із двома пояснюючими змінними

$$Y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + U \qquad (4.26)$$

на підставі вибірки $\left\{(y_i, x_{i1}, x_{i2}), i = \overline{1,n}\right\}$ обсягом n. Тоді за аналогією з (4.25) можна обчислити вибіркові коефіцієнти **звичайної** (парної) кореляції між Y і x_1, Y і x_2, x_1 і x_2:

$$r_{y1} = r(Y, x_1) = \frac{\overline{x_1 y} - \bar{x}_1 \cdot \bar{y}}{\sigma_1 \sigma_y}, \quad r_{y2} = r(Y, x_2) = \frac{\overline{x_2 y} - \bar{x}_2 \cdot \bar{y}}{\sigma_2 \sigma_y},$$
$$r_{12} = r(x_1, x_2) = \frac{\overline{x_1 x_2} - \bar{x}_1 \cdot \bar{x}_2}{\sigma_1 \sigma_2}, \qquad (4.27)$$

де $\sigma_1 = \sigma_{x_1}$, $\sigma_2 = \sigma_{x_2}$.

Розглянемо дві можливості: 1) кореляція, що спостерігається між, наприклад, залежною змінною Y і незалежною x_1, обумовлена **чистою залежністю** між ними; 2) друга незалежна змінна x_2 здійснює на кожну з них вплив, що і слугує однією із причин кореляції між першими двома змінними (Y і x_1). Виявляється, що в загальному випадку реалізується друга можливість.

Під **частковою кореляцією** між Y і x_1 будемо розуміти кореляцію між ними при умові, що вплив x_2 на кожну з цих величин усунуто. Аналогічно визначається поняття часткової кореляції між Y і x_2.

Припускаючи, що

$$|r_{y1}| \neq 1, \ |r_{y2}| \neq 1, \ |r_{12}| \neq 1, \tag{4.28}$$

виведемо аналітичний вираз для вибіркового часткового коефіцієнта кореляції. З допомогою МНК знаходимо рівняння прямих регресії Y на x_2

$$y_i - \overline{y} = r_{y2} \frac{\sigma_y}{\sigma_2}(x_{i2} - \overline{x}_2) + v_i, \ i = \overline{1,n},$$

та x_1 на x_2

$$x_{i1} - \overline{x}_1 = r_{12} \frac{\sigma_1}{\sigma_2}(x_{i2} - \overline{x}_2) + w_i, i = \overline{1,n},$$

де v_i – залишки, які не пояснюють регресію Y на x_2, w_i – залишки, які не пояснюють регресію x_1 на x_2, $\sigma_1 = \sigma_{x_1}$, $\sigma_2 = \sigma_{x_2}$.

Усуваємо вплив x_2, визначаючи залишки

$$v_i = y_i - \overline{y} - r_{y2} \frac{\sigma_y}{\sigma_2}(x_{i2} - \overline{x}_2), \ w_i = x_{i1} - \overline{x}_1 - r_{12} \frac{\sigma_1}{\sigma_2}(x_{i2} - \overline{x}_2), \ i = \overline{1,n}. \tag{4.29}$$

Вибірковим коефіцієнтом **часткової кореляції** між Y та x_1 при усуненні впливу x_2 називається вибірковий коефіцієнт звичайної (парної) кореляції між залишками v та w, спостережені значення яких визначаються (4.29):

$$r_{y1\bullet 2} = r(Y, x_1 \mid x_2) = r(v, w) = \frac{\overline{vw}}{\sqrt{\overline{v^2}}\sqrt{\overline{w^2}}}. \tag{4.30}$$

Зауваження. В правій частині (4.30) використані рівності $\overline{v} = \overline{w} = 0$ (порівняйте з (4.25), враховуючи розрахункову формулу для обчислення вибіркової дисперсії).

Виведемо більш зручну формулу для $r_{y1\bullet 2}$. Використавши (4.29) та (4.27), отримаємо

$$\sum_{i=1}^{n} v_i w_i = \sum_{i=1}^{n} (y_i - \overline{y})(x_{i1} - \overline{x}_1) - r_{12} \frac{\sigma_1}{\sigma_2} \sum_{i=1}^{n} (y_i - \overline{y})(x_{i2} - \overline{x}_2) -$$

$$- r_{y2} \frac{\sigma_y}{\sigma_2} \sum_{i=1}^{n} (x_{i1} - \overline{x}_1)(x_{i2} - \overline{x}_2) + r_{y2} r_{12} \frac{\sigma_y \sigma_1}{\sigma_2^2} \sum_{i=1}^{n} (x_{i2} - \overline{x}_2)^2.$$

Вирази (4.27) можна записати в еквівалентній формі виду другої рівності (4.25), звідки

$$\sum_{i=1}^{n}(y_i-\overline{y})(x_{i1}-\overline{x}_1)=nr_{y1}\sigma_y\sigma_1,\ \sum_{i=1}^{n}(y_i-\overline{y})(x_{i2}-\overline{x}_2)=nr_{y2}\sigma_y\sigma_2,$$

$$\sum_{i=1}^{n}(x_{i1}-\overline{x}_1)(x_{i2}-\overline{x}_2)=nr_{12}\sigma_1\sigma_2.$$

Врахувавши ці рівності і те, що $\sum_{i=1}^{n}(x_{i2}-\overline{x}_2)^2=n\sigma_2^2$, отримаємо:

$$\sum_{i=1}^{n}v_iw_i=n\sigma_y\sigma_1(r_{y1}-r_{y2}r_{12}).$$

Аналогічно отримуємо рівності

$$\sum_{i=1}^{n}v_i^2=n\sigma_y^2(1-r_{y2}^2),\ \sum_{i=1}^{n}w_i^2=n\sigma_1^2(1-r_{12}^2).$$

Тому остаточно (4.30) набере такого виду

$$r_{y1\bullet 2}=\frac{r_{y1}-r_{y2}r_{12}}{\sqrt{1-r_{y2}^2}\sqrt{1-r_{12}^2}}. \qquad (4.30^*)$$

Аналогічно можна отримати формули

$$r_{y2\bullet 1}=\frac{r_{y2}-r_{y1}r_{21}}{\sqrt{1-r_{y1}^2}\sqrt{1-r_{21}^2}},\ \ r_{12\bullet y}=\frac{r_{12}-r_{y1}r_{y2}}{\sqrt{1-r_{y1}^2}\sqrt{1-r_{y2}^2}}, \qquad (4.31)$$

де $r_{21}=r_{12}$.

З (4.30) та (4.29) випливає, що вибіркові часткові коефіцієнти кореляції набирають значень з проміжку [–1; 1], як і звичайні коефіцієнти кореляції.

Зауваження. Формули (4.30*), (4.31) мають сенс тільки при виконанні нерівностей (4.28).

7.1.2. Загальний випадок.

Для регресійної моделі (4.26) з двома незалежними змінними x_1 та x_2 введемо в розгляд симетричну кореляційну матрицю

$$B=\begin{pmatrix}b_{11}&b_{12}&b_{13}\\b_{21}&b_{22}&b_{23}\\b_{31}&b_{32}&b_{33}\end{pmatrix}=\begin{pmatrix}1&\rho_{Y1}&\rho_{Y2}\\\rho_{Y1}&1&\rho_{12}\\\rho_{Y2}&\rho_{12}&1\end{pmatrix},$$

складену з **теоретичних** парних коефіцієнтів кореляції. Знайдені оцінки теоретичних коефіцієнтів кореляції, тобто вибіркові коефіцієнти парної кореляції, утворюють **оцінену** симетричну кореляційну матрицю

$$K = \begin{pmatrix} k_{11} & k_{12} & k_{13} \\ k_{21} & k_{22} & k_{23} \\ k_{31} & k_{32} & k_{33} \end{pmatrix} = \begin{pmatrix} 1 & r_{y1} & r_{y2} \\ r_{y1} & 1 & r_{12} \\ r_{y2} & r_{12} & 1 \end{pmatrix}.$$

Аналіз формул (4.30*), (4.31) дозволяє записати формули для вибіркових часткових коефіцієнтів кореляції у такому вигляді

$$\begin{aligned} r_{y1\bullet 2} &= k_{12\bullet *} = -\frac{K_{12}}{\sqrt{K_{11}K_{22}}}, \\ r_{y2\bullet 1} &= k_{13\bullet *} = -\frac{K_{13}}{\sqrt{K_{11}K_{33}}}, \\ r_{12\bullet Y} &= k_{23\bullet *} = -\frac{K_{23}}{\sqrt{K_{22}K_{33}}}, \end{aligned} \qquad (4.32)$$

де K_{ij} – алгебраїчне доповнення елемента k_{ij} матриці K, $k_{ij\bullet *}$ – вибірковий частковий коефіцієнт кореляції між i-ою та j-ою ознаками при фіксованій ознаці, що залишилася, при цьому знаходження індексів, рівних 1, 2 і 3, відповідають ознакам Y, x_1 і x_2 відповідно.

Формули (4.32) узагальнено можна записати таким чином:

$$k_{ij\bullet *} = -\frac{K_{ij}}{\sqrt{K_{ii}K_{jj}}}, \qquad (4.33)$$

де $k_{ij\bullet *}$ – вибірковий частковий коефіцієнт кореляції між i-ою та j-ою ознаками, коли виключений вплив інших ознак.

Отримані формули (4.33) дозволяють отримати інформацію про вибіркові часткові коефіцієнти кореляції при виключенні впливу решти ознак у загальному випадку лінійної регресії

$$Y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_m x_m + U.$$

При цьому індекси i та j змінюються від 1 до $m+1$ ($i \neq j$).

7.2. Найповніше дослідити мультиколінеарність в моделі $Y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_m x_m + U$ можна за допомогою **алгоритма Фаррара-Глобера**. Цей алгоритм має три види статистичних критеріїв, згідно з якими перевіряються:

1) мультиколінеарність усього масиву незалежних змінних (χ^2);

2) кожної незалежної змінної з рештою змінних (F-критерій);

3) кожної пари незалежних змінних (*t*-критерій Ст'юдента).

Крок 1. Обчислюється кореляційна матриця

$$r = \begin{pmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1m} \\ r_{12} & 1 & r_{23} & \cdots & r_{2m} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ r_{1m} & r_{2m} & r_{3m} & \cdots & 1 \end{pmatrix},$$

де

$$r_{ij} = r_{x_i x_j} = \frac{\overline{x_i x_j} - \overline{x}_i \cdot \overline{x}_j}{\sigma_{x_i} \sigma_{x_j}},\ i \neq j;\ i, j = \overline{1, m}.$$

Крок 2. Знаходиться спостережене значення χ^2:

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

де $|r|$ – визначник матриці r.

Якщо

$$\chi^2_{спост.} > \chi^2_{кр.}\left(\alpha; \frac{1}{2} m(m-1) \right),$$

то на рівні значущості α робиться висновок: в масиві незалежних змінних існує мультиколінеарність.

Крок 3. Знаходиться матриця $C = r^{-1}$.

Крок 4. Обчислюються спостережені значення *F*-критеріїв:

$$F_{k.спост.} = (c_{kk} - 1)\frac{n-m}{m-1},$$

де c_{kk} – k-й діагональний елемент матриці *C*.

Якщо

$$F_{k.спост.} > F_{кр.}(\alpha; n-m; m-1),$$

то на рівні значущості α k-а незалежна змінна x_k мультиколінеарна з іншими незалежними змінними.

При цьому коефіцієнт детермінації для кожної змінної має такий вид:

$$R^2_{x_k} = 1 - \frac{1}{c_{kk}},\ k = \overline{1, m}.$$

Крок 5. Знаходяться часткові коефіцієнти кореляції:

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

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

Крок 6. Обчислюються спостережені значення t-критеріїв:

$$t_{kj.спост.} = \frac{\left|r_{kj\bullet *}\right|\sqrt{n-m-1}}{\sqrt{1-r_{kj\bullet *}^2}}.$$

Якщо

$$t_{kj.спост.} > t_{кр.}(\alpha; n-m-1),$$

то між x_k та x_j на рівні значущості α існує мультиколінеарність.

8. Найпростіший метод, однак не завжди можливий, полягає в тому, що із двох пояснюючих змінних, які мають високий коефіцієнт кореляції (більший 0,8), одну змінну виключають із розгляду. Визначення питання, яку змінну залишити, а яку вилучити із моделі, вирішують в першу чергу на підставі економічних міркувань. Якщо з економічної точки зору неможливо надати перевагу жодній, то залишають ту із двох змінних, яка має більший коефіцієнт кореляції із залежною змінною.

Другий метод усунення або зменшення мультиколінеарності полягає у переході від незміщених МНК-оцінок до зміщених оцінок, які володіють, однак, меншим розкидом відносно оцінюваного параметра, тобто меншою дисперсією. Справді, МНК-оцінки згідно з теоремою Гаусса-Маркова володіють мінімальними дисперсіями в класі всіх лінійних **незміщених** оцінок, однак при наявності мультиколінеарності ці дисперсії можуть виявитися занадто великими за рахунок того, що матриця $(X'X)^{-1}$ містить великий співмножник $\left[\det(X'X)^{-1}\right]$. Реалізацію другого підходу дає використання «рідж-регресії» або «гребеневої регресії», для якої замість нормальної оцінки $a = (X'X)^{-1}X'Y$ розглядають зміщену оцінку $\hat{a} = (X'X + pI_{m+1})^{-1}X'Y$, де p – деяке число, яке називається «гребенем» або «хребтом». При цьому додавання p до діагональних елементів матриці $X'X$ перетворюють оцінки параметрів моделі у зміщені, але при цьому збільшується визначник матриці системи нормальних рівнянь: замість $\det(X'X)$ використовується $\det(X'X + pI_{m+1})$.

Отже, стає можливим виключення мультиколінеарності у випадку, коли $\det(X'X)$ близький до нуля.

Третім із можливих методів усунення або зменшення мультиколінеарності є використання **покрокових процедур відбору найбільш інформативних змінних**. На першому кроці розглядається лише одна пояснююча змінна, яка має із залежною змінною Y найбільший коефіцієнт детермінації. На другому кроці в регресійну модель долучається нова пояснююча змінна, яка разом із початково обраною утворює пару пояснюючих змінних, яка має з Y найбільший скорегований коефіцієнт детермінації $\hat{R}^2$. На третьому кроці в модель вводиться ще одна пояснююча змінна, яка разом із двома попередньо відібраними утворює трійку пояснюючих змінних, що має з Y найбільший скорегований коефіцієнт детермінації $\hat{R}^2$.

Процедура введення нових змінних проводиться до тих пір, поки буде збільшуватися відповідний скорегований коефіцієнт детермінації $\hat{R}^2$.

Четвертий метод – це **метод головних компонентів** [10], який призначений для оцінювання моделей великого розміру, до яких входять мультиколінеарні змінні.

Задача 4.6. *За даними статистичних щорічників України [14] досліджується залежність змінної y – кількість збудованих квартир (у сільській і міській місцевості) на 1000 населення від ряду змінних: x_1 – доходи населення, млн. грн.; x_2 – кредити банків та інші позики на капітальні інвестиції, млн. грн.; x_3 – кошти населення на будівництво житла, млн. грн.; x_4 – середньомісячна номінальна заробітна плата, грн. Вихідні дані за 2011 – 2020 роки приведені в таблиці 4.5.*

Таблиця 4.5

Рік	y	x_1	x_2	x_3	x_4
2011	3,2	1266753	36651,9	17589,2	2633
2012	3,7	1457864	39724,7	22575,5	3026
2013	4,3	1548733	34734,7	24072,3	3265
2014	4,7	1516768	21739,3	22064,2	3480
2015	5,4	1772016	20740,1	31985,4	4195
2016	5	2051331	27106	29932,6	5183
2017	5,1	2652082	29588,9	32802,5	7104
2018	4,7	3248730	44825,4	34645,7	8865
2019	6	3744060	67232,6	32422	10497
2020	3,2	3972428	27894,5	20590,9	11591

Дослідити наявність мультиколінеарності між пояснюючими змінними за алгоритмом Фаррара-Глобера. У випадку виявлення мультиколінеарності здійснити заходи по її усуненню методом виключення змінної з розгляду.

○ *Крок 1.* Обчислюємо кореляційну матрицю. Для обчислення кореляційної матриці використовуємо в Excel вбудовану функцію CORREL (коефіцієнт парної кореляції), яка знаходиться в категорії СТАТИСТИКА або заходимо в Data – Аналіз даних – Кореляція.

$$r = \begin{pmatrix} 1 & 0,454 & 0,406 & 0,999 \\ 0,454 & 1 & 0,265 & 0,427 \\ 0,406 & 0,265 & 1 & 0,388 \\ 0,999 & 0,427 & 0,388 & 1 \end{pmatrix}.$$

Крок 2. Знаходимо спостережене значення χ^2. Спочатку обчислюємо визначник кореляційної матриці. Визначник зручно обчислити з допомогою вбудованої математичної функції MDETERM(блок кореляційної матриці).

$$|r| = 0,00039,\ \ln 0,00039 = -7,853.$$

$$\chi^2_{спост.} = -\left(10 - 1 - \frac{1}{6}\cdot 13\right)(-7,853) = 53,659.$$

Знайдене значення $\chi^2_{спост.}$ порівнюємо з $\chi^2_{кр.}\left(0,05; \frac{1}{2}\cdot 4\cdot(4-1)\right) = \chi^2_{кр.}\left(0,05;6\right) = 12,592$. Так як $\chi^2_{спост.} > \chi^2_{кр.}\left(0,05;6\right)$, то на рівні значущості 0,05 робимо висновок: в масиві незалежних змінних існує мультиколінеарність.

Крок 3. Знаходимо матрицю $C = r^{-1}$. Обернену матрицю зручно обчислити за допомогою MINVERSE (блок кореляційної матриці):

$$C = \begin{pmatrix} 1761,636 & -55,016 & -32,097 & -1724,150 \\ -55,016 & 2,959 & 0,857 & 53,372 \\ -32,097 & 0,857 & 1,779 & 31,014 \\ -1724,150 & 53,372 & 31,014 & 1688,818 \end{pmatrix}.$$

Крок 4. Обчислюємо спостережені значення *F*-критеріїв:

$$F_{1.спост.} = (1761,636 - 1)\frac{10-4}{4-1} = 3521,272.$$

$$F_{2.спост.} = (2,959-1)\frac{10-4}{4-1} = 3,918.$$

$$F_{3.спост.} = (1,779-1)\frac{10-4}{4-1} = 1,557.$$

$$F_{4.спост.} = (1688,818-1)\frac{10-4}{4-1} = 3375,637.$$

Знайдені значення порівнюємо з критичною величиною F-критерія при $n - m = 10 - 4 = 6$ та $m - 1 = 4 - 1 = 3$ ступенями вільності і рівні значущості $\alpha = 0,05$: $F_{кр.}(0,05;6;3) = 4,76$. Так як $F_{1.спост.} > 4,76$ та $F_{4.спост.} > 4,76$ то на рівні значущості $\alpha = 0,05$ незалежні змінні x_1 та x_4 мультиколінеарні з іншими незалежними змінними.

Обчислимо коефіцієнт детермінації для кожної змінної:

$$R^2_{x_1} = 1-\frac{1}{1761,636} = 0,999, \quad R^2_{x_2} = 1-\frac{1}{2,959} = 0,662, \quad R^2_{x_3} = 1-\frac{1}{1,779} = 0,438,$$

$$R^2_{x_4} = 1-\frac{1}{1688,818} = 0,999.$$

Значення коефіцієнтів детермінації теж підтверджують мультиколінеарність x_1 та x_2 з іншими незалежними змінними.

Крок 5. Знаходимо часткові коефіцієнти кореляції:

$$r_{12\cdot34} = -\frac{-55,016}{\sqrt{1761,636}\sqrt{2,959}} = 0,762, \quad r_{13\cdot24} = -\frac{-32,097}{\sqrt{1761,636}\sqrt{1,779}} = 0,573,$$

$$r_{14\cdot23} = -\frac{-1724,150}{\sqrt{1761,636}\sqrt{1688,818}} = 0,999, \quad r_{23\cdot14} = -\frac{0,857}{\sqrt{2,959}\sqrt{1,779}} = -0,374,$$

$$r_{24\cdot13} = -\frac{53,372}{\sqrt{2,959}\sqrt{1688,818}} = -0,755, \quad r_{34\cdot12} = -\frac{31,014}{\sqrt{1,779}\sqrt{1688,818}} = -0,566.$$

Враховуючи розраховані значення часткових коефіцієнтів кореляції, можна стверджувати, що найсильніший зв'язок є між змінними x_1 та x_4.

Крок 6. Обчислюємо спостережені значення t-критеріїв:

$$t_{12.спост.} = \frac{0,762\sqrt{10-4-1}}{\sqrt{1-0,762^2}} = 2,631, \quad t_{13.спост.} = \frac{0,573\sqrt{5}}{\sqrt{1-0,573^2}} = 1,565,$$

$$t_{14.спост.} = \frac{0,999\sqrt{5}}{\sqrt{1-0,999^2}} = 78,838, \quad t_{23.спост.} = \frac{0,374\sqrt{5}}{\sqrt{1-0,374^2}} = 0,900,$$

$$t_{24.спост.} = \frac{0,755\sqrt{5}}{\sqrt{1-0,755^2}} = 2,575, \quad t_{34.спост.} = \frac{0,566\sqrt{5}}{\sqrt{1-0,566^2}} = 1,535.$$

Обчислені значення порівнюємо з критичною величиною t-критерія при $n - m - 1 = 10 - 4 - 1 = 5$ ступенів вільності і рівні значущості $\alpha = 0{,}05$: $t_{кр.}(0{,}05;5) = 2{,}571$. Так як $t_{12.спост.} > 2{,}571$, $t_{14.спост.} > 2{,}571$ та $t_{24.спост.} > 2{,}571$ то на рівні значущості $\alpha = 0{,}05$ між змінними x_1 та x_2, x_1 та x_4 і x_2 та x_4 існує мультиколінеарність.

Необхідно виключити одну змінну з розгляду. Виключимо змінну x_4, оскільки $t_{14.спост.} = 78{,}838$ є найбільшим серед усіх спостережених значень t-критерію і

$$r_{yx_4} = 0{,}197 < r_{yx_1} = 0{,}212.$$

Примітка. Якщо виходити з економічних міркувань, то теж варто надати перевагу змінній x_1 (доходи населення, млн. грн.), а не x_4 (середньомісячна номінальна заробітна плата, грн.), оскільки x_4 є складовою x_1.

На наступному етапі знову робимо перевірку існування мультиколінеарності в масиві незалежних змінних x_1, x_2, x_3.

Крок 1. Обчислюємо кореляційну матрицю:

$$r = \begin{pmatrix} 1 & 0{,}454 & 0{,}406 \\ 0{,}454 & 1 & 0{,}265 \\ 0{,}406 & 0{,}265 & 1 \end{pmatrix}.$$

Крок 2. Знаходимо спостережене значення χ^2.

$$|r| = 0{,}657\,,\ \ln 0{,}657 = -0{,}421.$$

$$\chi^2_{спост.} = -\left(10 - 1 - \frac{1}{6}\cdot 11\right)(-0{,}421) = 3{,}015.$$

Знайдене значення $\chi^2_{спост.}$ порівнюємо з $\chi^2_{кр.}\left(0{,}05;\frac{1}{2}\cdot 3\cdot(3-1)\right) = \chi^2_{кр.}(0{,}05;3) = 7{,}8$. Так як $\chi^2_{спост.} < \chi^2_{кр.}(0{,}05;3)$, то на рівні значущості 0,05 робимо висновок: в масиві незалежних змінних мультиколінеарність відсутня.

Знайдемо вибіркове рівняння множинної регресії для змінних y – кількість збудованих квартир (у сільській і міській місцевості) на 1000 населення, x_1 – доходи населення, млн. грн., x_2 – кредити банків та інші позики на капітальні інвестиції, млн. грн., x_3 – кошти населення на будівництво житла, млн. грн. Приклад знаходження вибіркового рівняння множинної регресії приведено в прикладі 4.1 або можна скористатися функцією LINEST (алгоритм знаходження оцінок параметрів за цією функцією приведено в §8). Можна також скористатися пакетом аналізу Дані – Аналіз даних – Регресія.

Результат побудови лінійної регресійної моделі за функцією LINEST показано на рисунку:

0,000135	8,80785E-06	-1,89381E-07	1,04449044
3,54E-05	1,62759E-05	2,32213E-07	0,93116812
0,730648	0,591779409	#Н/Д	#Н/Д
5,42522	6	#Н/Д	#Н/Д
5,699783	2,101217215	#Н/Д	#Н/Д

Вибіркове рівняння множинної регресії має наступний вигляд:

$\hat{y} = 1,0444904 + 0,0000002x_1 + 0,0000088x_2 + 0,000135x_3, \quad R^2 = 0,7306.$ ◎

ІНДИВІДУАЛЬНЕ ЗАВДАННЯ № 3

Для торговельного підприємства, яке має велику кількість філій, результуюча змінна y (річний товарообіг однієї філії, млн. грн.) лінійно залежить від x_1 (торгівельної площі, тис. м^2) та від x_2 (середньоденної інтенсивності потоку покупців, тисяч людей за день). Для дванадцяти філій за певний рік зафіксовані такі значення показників y, x_1 та x_2:

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14
y_i	2,8	5,2	6,8	7,1	7,3	8,3	4,3	5,8	7,7	3,2	1,5	3,7	5,4	2,6
x_{i1}	0,3	0,9	1,2	1,3	1,2	0,8	0,8	0,9	1,3	0,5	0,3	0,6	1,1	0,2
x_{i2}	1,4	2,7	3,4	3,6	3,7	4,2	2,1	2,9	3,8	1,7	0,8	1,9	2,7	1,3
i	15	16	17	18	19	20	21	22	23	24	25	26	27	28
y_i	5,8	8,1	7,5	8,4	4,2	5,6	7,6	3,5	1,4	3,9	6,4	7,3	7,6	8,3
x_{i1}	0,8	1,6	1,5	0,9	0,9	0,8	1,4	0,7	0,4	0,8	1.1	1,4	1,5	0,9
x_{i2}	2,8	4,1	3,6	4,2	2,1	2,8	3,8	1,7	0,7	1,8	3,2	3,6	3,9	4,1
i	29	30	31	32	33	34	35	36	37	38	39	40	41	42
y_i	4,4	5,6	7,5	3,2	1,5	3,5	2,7	4,3	6,9	7,1	7,3	8,4	4,2	5,8
x_{i1}	0,8	0,7	1,4	0,6	0,3	0,7	0,2	0,6	1,4	1,3	1,2	0,9	0,7	0,7
x_{i2}	22	2,9	3,2	1,7	0,8	1,7	1,3	2,2	3,4	3,1	3,2	4,2	4,3	2,9
i	43	44	45	46	47	48	49	50						
y_i	7,5	3,5	1,4	4,2	2,6	5,8	4,2	3,1						
x_{i1}	1,5	0,8	0,3	0,9	0,4	0,7	0,8	0,6						
x_{i2}	3,7	1,8	0,7	2,0	1,4	2,9	2,0	1,6						

Потрібно:

1. Знайти статистичні оцінки параметрів теоретичної множинної лінійної регресії;

2. Обчислити вибірковий коефіцієнт детермінації;

3. Для рівняння значущості $\alpha = 0,05$ перевірити правильність статистичних гіпотез $a_1 = 0,\ a_2 = 0$;

4. З надійністю $\gamma = 0,95$ побудувати довірчі інтервали для параметрів теоретичного рівняння множинної регресії;

5. Знайти прогнозоване значення річного товарообігу для нової філії, введення в дію якої планується у відносно заселеному районі із середньоденною інтенсивністю потоку покупців 15 000 людей в день і торгівельною площею 1200 м^2, а також із надійністю $\gamma = 0,95$ побудувати довірчий інтервал для прогнозованого значення.

6. Обчислити вибіркові коефіцієнти звичайної і часткової кореляції між Y і x_1, Y і x_2, x_1 і x_2.

7. Здійснити аналіз отриманих результатів на підставі наступних тверджень:

а) якщо при усуненні впливу інших факторів кореляція двох величин зростає, то це означає, що ці фактори приховували істинну взаємозалежність досліджуваних двох величин;

б) якщо ж кореляція між двома величинами зменшується або стає ближчою до нуля при інших фіксованих величинах, то можна стверджувати, що взаємозалежність цих двох величин значною мірою має місце завдяки іншим факторам.

ІНДИВІДУАЛЬНЕ ЗАВДАННЯ № 4

Економічний показник середньомісячна заробітна плата (y) залежить від продуктивності праці (x_1), фондомісткості (x_2) і коефіцієнта плинності робочої сили (x_3). На основі статистичних даних за 10 років необхідно оцінити наявність мультиколінеарності між пояснюючими змінними. У випадку її присутності, виявити пари факторів між якими існує мультиколінеарність, один із факторів виключити з розгляду таких пар. Дослідження провести за методом Фаррара-Глобера.

Вихідні дані наводяться в табл. 4.6 та табл. 4.7

Таблиця 4.6

Номер цеху	Продуктивність праці, людино-дні	Фондомісткість, млн. грн.	Коефіцієнт плинності робочої сили, %
1	$32 + a_4$	$0.89 + a_9$	$19.5 + a_1$
2	$29 + a_5$	$0.43 + a_9$	$15.6 + a_2$
3	$30 + a_6$	$0.70 + a_9$	$13.5 + a_3$
4	$31 + a_7$	$0.61 + a_9$	$9.5 + a_1$
5	$25 + a_8$	$0.51 + a_9$	$23.5 + a_2$
6	$34 + a_4$	$0.71 + a_9$	$12.5 + a_3$

7	$29 + a_5$	$0.65 + a_9$	$17.5 + a_1$
8	$24 + a_6$	$0.43 + a_9$	$14.5 + a_2$
9	$20 + a_7$	$0.33 + a_9$	$14.5 + a_3$
10	$33 + a_8$	$0.92 + a_9$	$75 + a_1$

Таблиця 4.7

Варіант останнього цифра	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}
0	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	0.01	2.1
1	2.1	2.0	1.4	2.3	2.2	2.6	2.7	2.9	0.02	3.2
2	0.5	0.6	0.7	0.8	0.9	0.12	0.15	0.25	0.03	0.4
3	2.1	1.2	1.3	3.4	2.4	2.5	2.6	2.7	0.04	3.14
4	1.5	2.1	1.2	5.1	1.2	2.0	1.8	3.2	0.01	3.1
5	1.6	2.2	1.3	5.2	1.3	2.1	1.9	3.4	0.02	3.2
6	1.7	2.3	1.4	5.3	1.4	2.2	1.7	3.5	0.03	3.3
7	1.8	2.4	1.6	5.4	1.5	2.3	1.4	4.1	0.04	3.4
8	1.9	2.6	1.7	5.2	1.6	2.4	1.5	4.2	0.04	3.6
9	2.0	2.7	1.8	5.0	1.7	2.6	1.2	4.3	0.02	3.7

§ 5. НЕЛІНІЙНІ ЕКОНОМЕТРИЧНІ МОДЕЛІ

У реальних економічних умовах залежність між змінними може адекватно представлятися, як правило, у нелінійній формі [11-13]. Ця залежність описується формулою

$$Y = f(x) + U,$$

де $f(x)$ — нелінійна функція аргумента x, U — випадковий чинник.

Відповідна економетрична модель має вид:

$$\hat{y} = f(x).$$

Вид економетричної моделі вибирається на основі графічного зображення у системі координат (x, y) статистичної інформації (побудови діаграми розсіювання).

Розглянемо найважливіші нелінійні економетричні моделі.

Гіперболічна (зворотна) залежність має вид

$$\hat{y} = a_0 + \frac{a_1}{x}. \qquad (5.1)$$

Вона зводиться до лінійної заміною $z = \frac{1}{x}$. Одержимо

$$\hat{y} = a_0 + a_1 z.$$

Перевірка моделі на адекватність та побудова прогнозу здійснюється, як і для лінійної моделі, з урахуванням розглянутої заміни змінної x.

Задача 5.1. *Використати гіперболічну модель для дослідження залежності собівартості Y (гр.од./шт.) від кількості виготовленої продукції x (шт.). Наведена статистична інформація для показників Y і x:*

i	1	2	3	4	5	6	7	8	9	10
y_i	40	37	34	21	29	27	25	24	23	22
x_i	1	2	3	4	5	6	7	8	9	10

Потрібно знайти статистичні оцінки параметрів лінійного рівняння регресії.

○ Статистичні оцінки a_0, a_1 параметрів α_0 та α_1 гіперболічного рівняння регресії, із врахуванням заміни $z = 1/x$, задовольняють системі нормальних рівнянь (2.12):

$$\begin{cases} a_0 + \bar{z}a_1 = \bar{y}, \\ \bar{z}a_0 + \overline{z^2}a_1 = \overline{zy}. \end{cases}$$

Для знаходження коефіцієнтів цієї системи складемо розрахункову табл. 5.1.

Таблиця 5.1

i	x_i	y_i	z_i	z_i^2	$z_i\, y_i$
1	1	40	1	1	40
2	2	37	0,5	0,25	18,5
3	3	34	0,33	0,11	11,33
4	4	21	0,25	0,06	5,25
5	5	29	0,2	0,04	5,8
6	6	27	0,17	0,03	4,5
7	7	25	0,14	0,02	3,57
8	8	24	0,13	0,02	3
9	9	23	0,11	0,01	2,56
10	10	22	0,1	0,01	2,2
Σ		282	2,93	1,55	96,71

Використовуючи нижній рядок табл. 5.1, отримаємо (обсяг вибірки $n=10$):

$$\bar{z}=\sum_{i=1}^{10} z_i/n = 2{,}93/10 = 0{,}293; \qquad \bar{y}=\sum_{i=1}^{10} y_i/n = 282/10 = 28{,}2;$$

$$\overline{z^2}=\sum_{i=1}^{10} z_i^2/n = 1{,}55/10 = 0{,}155; \qquad \overline{zy}=\sum_{i=1}^{10} z_i y_i/n = 96{,}71/10 = 9{,}671;$$

$$\begin{cases} a_0 + 0{,}293 a_1 = 28{,}2, \\ 0{,}293 a_0 + 0{,}155 a_1 = 9{,}671. \end{cases}$$

Розв’язок цієї системи рівнянь згідно із формулами (2.13):

$$a_1 = \frac{\overline{zy}-\bar{z}\cdot\bar{y}}{\overline{z^2}-(\bar{z})^2} = \frac{9{,}671-0{,}293\cdot 28{,}2}{0{,}155-(0{,}293)^2} = \frac{1{,}4084}{0{,}0691} = 20{,}382,$$

$$a_0 = \bar{y} - a_1\bar{z} = 28{,}2 - 20{,}382\cdot 0{,}293 = 22{,}228.$$

Отже, емпіричне рівняння регресії має такий вигляд:

$$\hat{y} = 22{,}228 + \frac{20{,}382}{x}.$$

Для знаходження та оцінки значущості коефіцієнтів регресії α_0 та α_1, точкової оцінки дисперсії збурень, вибіркового коефіцієнта детермінації, коефіцієнта кореляції та побудови для них довірчих інтервалів можна використати розглянуті в §2 методи дослідження лінійної моделі парної регресії. При цьому необхідно здійснити потрібну заміну переходу від нелінійної моделі до лінійної. ◎

Степенева (мультиплікативна) залежність має наступний вид

$$\hat{y} = a_0 \cdot x^{a_1}, \ \ a_0 > 0, \ x > 0. \qquad (5.2)$$

Її графік зображено на рисунку 5.1. Степенева залежність використовується для моделювання ситуацій, в яких ріст витрат x деякого ресурсу обумовлює необмежене збільшення випуску Y.

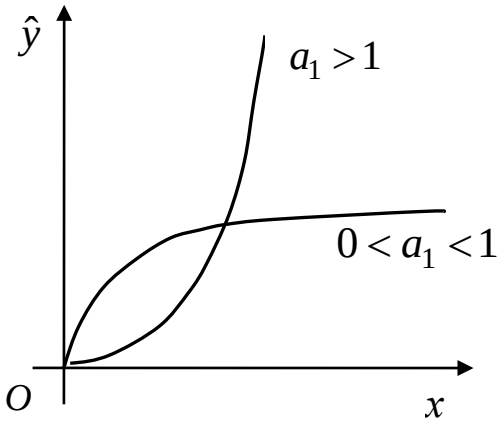


Рисунок 5.1.

Вона зводиться до лінійної моделі логарифмуванням з довільною основою, наприклад, e. Тоді отримаємо співвідношення

$$\ln \hat{y} = \ln a_0 + a_1 \ln x.$$

Застосуємо такі заміни:

$$\ln \hat{y} = \hat{y}^*, \ \ln a_0 = a_0^*, \ \ln x = x^*.$$

Отримаємо рівняння

$$\hat{y}^* = a_0^* + a_1 x^*.$$

Найвідомішою степеневою функцією є виробнича функція Кобба – Дугласа [15]. Виробнича функція Кобба – Дугласа для *m* факторів має вигляд:

$$\hat{y} = a_0 x_1^{a_1} x_2^{a_2} \ldots x_m^{a_m}.$$

Після логарифмування цього виразу і заміни величин:

$$\ln \hat{y} = \hat{y}^*, \ \ln a_0 = a_0^*, \ \ln x_1 = x_1^*, \ \ln x_2 = x_2^*, \ \ldots, \ \ln x_m = x_m^*,$$

отримаємо лінійну регресію

$$\hat{y}^* = a_0^* + a_1 x_1^* + a_2 x_2^* + \ldots + a_m x_m^*.$$

Задача 5.2. *За даними статистичних щорічників України [14] побудувати виробничу функцію Кобба-Дугласа, яка описує залежність ВВП України в переробній промисловості (Y, млн. грн.), від факторів: капітальні інвестиції в переробній промисловості (K, млн. грн.) та зайнятість населення в переробній промисловості (L, тис. люд.):*

$$\hat{Y} = a_0 K^{a_1} L^{a_2}.$$

Вихідні дані за 2010 – 2020 роки приведені в таблиці.

Рік	Y	K	L
2010	810843	30579	2402,3
2011	970116	42742	2326,8
2012	974924	43032	2321,6
2013	904052	46240	2275,5
2014	975675	42474	2022,2
2015	1206041	46219	1839,3
2016	1458786	62223	1791,7
2017	1805097	73884	1774,9
2018	2060485	100870	1786,3
2019	2142939	105878	1833,3
2020	2057221	84408	1737,2

○ Для оцінки параметрів вибіркового рівняння регресії прологарифмуємо обидві частини рівняння $\hat{Y} = a_0 K^{a_1} L^{a_2}$ при основі e:

$$\ln \hat{Y} = \ln a_0 + a_1 \ln K + a_2 \ln L .$$

Виконаємо заміну змінних: $\hat{Y}^* = \ln \hat{Y}$, $a_0^* = \ln a_0$, $K^* = \ln K$, $L^* = \ln L$. Отримуємо лінійну форму зв'язку: $\hat{Y}^* = a_0^* + a_1 K^* + a_2 L^*$.

Перетворюємо вхідні дані:

№	Y^*	K^*	L^*
1	13,6058	10,3281	7,7842
2	13,7852	10,6629	7,7522
3	13,7901	10,6697	7,7500
4	13,7146	10,7416	7,7300
5	13,7909	10,6566	7,6119
6	14,0029	10,7411	7,5171
7	14,1931	11,0385	7,4909
8	14,4061	11,2103	7,4815
9	14,5385	11,5216	7,4879
10	14,5777	11,5700	7,5139
11	14,5369	11,3434	7,4600

Статистичні оцінки a_0^*, a_1, a_2 рівняння регресії, із врахуванням попередніх замін, розраховують з системи нормальних рівнянь:

$$\begin{cases} a_0^* + \overline{K^*}a_1 + \overline{L^*}a_2 = \overline{Y^*}, \\ \overline{K^*}a_0^* + \overline{K^{*2}}a_1 + \overline{K^*L^*}a_2 = \overline{K^*Y^*}, \\ \overline{L^*}a_0^* + \overline{L^*K^*}a_1 + \overline{L^{*2}}a_2 = \overline{L^*Y^*}. \end{cases}$$

Для знаходження коефіцієнтів цієї системи складемо розрахункову таблицю 5.2.

Таблиця 5.2

i	Y^*	K^*	L^*	K^{*2}	L^{*2}	K^*L^*	K^*Y^*	L^*Y^*
1	13,6058	10,3281	7,7842	106,67	60,59	80,40	140,52	105,91
2	13,7852	10,6629	7,7522	113,70	60,10	82,66	146,99	106,87
3	13,7901	10,6697	7,7500	113,84	60,06	82,69	147,14	106,87
4	13,7146	10,7416	7,7300	115,38	59,75	83,03	147,32	106,01
5	13,7909	10,6566	7,6119	113,56	57,94	81,12	146,96	104,98
6	14,0029	10,7411	7,5171	115,37	56,51	80,74	150,41	105,26
7	14,1931	11,0385	7,4909	121,85	56,11	82,69	156,67	106,32
8	14,4061	11,2103	7,4815	125,67	55,97	83,87	161,50	107,78
9	14,5385	11,5216	7,4879	132,75	56,07	86,27	167,51	108,86
10	14,5777	11,5700	7,5139	133,87	56,46	86,94	168,66	109,53
11	14,5369	11,3434	7,4600	128,67	55,65	84,62	164,90	108,45
Σ	154,9417	120,4839	83,5797	1321,33	635,22	915,03	1698,57	1176,84

$\overline{Y^*} = 154,9417/11 = 14,09$, $\overline{K^*} = 120,4839/11 = 10,95$, $\overline{L^*} = 83,5797/11 = 7,60$,

$\overline{K^{*2}} = 1321,33/11 = 120,12$, $\overline{L^{*2}} = 635,22/11 = 57,75$,

$\overline{K^*L^*} = 915,03/11 = 83,18$, $\overline{K^*Y^*} = 1698,57/11 = 154,42$,

$\overline{L^*Y^*} = 1176,84/11 = 106,99$.

Отримуємо систему лінійних рівнянь:

$$\begin{cases} a_0^* + 10,95a_1 + 7,60a_2 = 14,09, \\ 10,95a_0^* + 120,12a_1 + 83,18a_2 = 154,42, \\ 7,60a_0^* + 83,18a_1 + 57,75a_2 = 106,99. \end{cases}$$

Розв'язавши її отримаємо: $a_0^* = 12,8463$, $a_1 = 0,6818$, $a_1 = -0,8197$.

Вибіркове рівняння множинної регресії має наступний вигляд:

$$\hat{Y}^* = 12,8463 + 0,6818K^* - 0,8197L^*.$$

Перейдемо до початкових змінних ($a_0 = e^{a_0^*} = e^{12,8463} = 379396,5$) і отримаємо виробничу функцію:

$$\hat{Y} = 379396,5 \cdot K^{0,6818} \cdot L^{-0,8197}.$$

Примітка. Приклад обчислення виробничої функції засобами Excel приведено в §8 в підрозділі «Нелінійна регресія». ◎

Експоненціальна (показникова) модель записується так:

$$\hat{y} = a_0 \cdot a_1^x, \ \ a_0 > 0, \ a_1 > 0, \ a_1 \neq 1. \qquad (5.3)$$

Для одержання лінійної залежності застосуємо логарифмування. Тоді

$$\ln \hat{y} = \ln a_0 + x \ln a_1.$$

Здійснивши заміну змінних $\ln \hat{y} = \hat{y}^*$, $\ln a_0 = a_0^*$, $\ln a_1 = a_1^*$, отримаємо

$$\hat{y}^* = a_0^* + a_1^* x.$$

Криві з границею росту і точкою перегину часто використовуються для статистичного аналізу попиту на деякі нові товари. Такою кривою є, наприклад, **крива Джонсона**:

$$\hat{y} = e^{a_0 - \frac{a_1}{x}}, \ \ a_0 > 0, \ a_1 > 0.$$

Її графік зображено на рисунку 5.2.

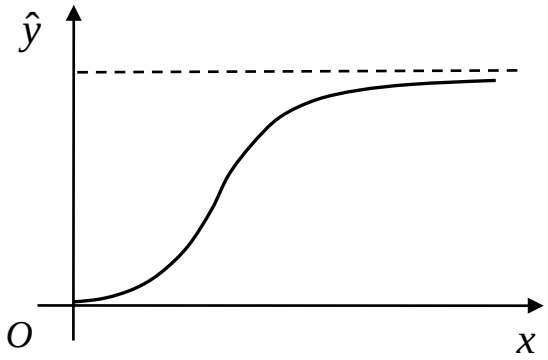


Рисунок 5.2.

Знайдемо логарифми обох частин кривої Джонсона:

$$\ln \hat{y} = a_0 - \frac{a_1}{x}.$$

Замінивши $\ln \hat{y} = \hat{y}^*$, $\frac{1}{x} = z$, одержимо лінійну залежність

$$\hat{y}^* = a_0 - a_1 x.$$

Для моделювання **немонотонних (коливних)** процесів набули широкого використання многочлени (поліноми)

$$\hat{y} = a_0 + a_1 x + a_2 x^2 + \ldots + a_m x^m.$$

Якщо всі статистичні значення x_i $(i = 1,2,\ldots,n)$ різні, то, як відомо з теорії інтерполяції, через n точок можна єдиним способом привести многочлен степені $n-1$.

Для одержання лінійної моделі використаємо заміну $x^m = x_m^*$. Одержимо

$$\hat{y} = a_0 + a_1 x_1^* + a_2 x_2^* + \ldots + a_m x_m^*.$$

Ця множинна лінійна залежність з числом змінних m, $m < n-1$.

При дослідженні залежності обсягу податкових надходжень Y від величини податкової ставки x застосовують **криву Лаффера**

$$\hat{y} = a_0 \cdot e^{a_1 (x - a_2)^2}. \qquad (5.4)$$

Тут a_0, a_1, a_2 — невідомі коефіцієнти, які визначаються на основі статистичної інформації. Логарифмуємо обидві частини цієї залежності. Маємо

$$\ln \hat{y} = \ln a_0 + a_1 x^2 - 2a_1 a_2 x + a_1 a_2^2.$$

Використаємо заміни змінних $\ln \hat{y} = \hat{y}^*$, $\ln a_0 + a_1 a_2^2 = a_0^*$, $-2a_1 a_2 = a_3$. Матимемо многочлен степені 2

$$\hat{y}^* = a_0^* + a_3 x + a_1 x^2.$$

Коефіцієнти a_0^*, a_3, a_1 знаходимо як розв'язок такої системи лінійних рівнянь

$$\begin{cases} a_0^* + a_3 \overline{x} + a_1 \overline{x^2} = \overline{y^*}, \\ \overline{x} a_0^* + \overline{x^2} a_3 + \overline{x^3} a_1 = \overline{x y^*}, \\ \overline{x^2} a_0^* + \overline{x^3} a_3 + \overline{x^4} a_1 = \overline{x^2 y^*}, \end{cases} \qquad (5.5)$$

де $y_i^* = \ln y_i$, n — число статистичних значень кожної із змінних x, y.

Графік кривої Лаффера зображено на рисунку 5.3.

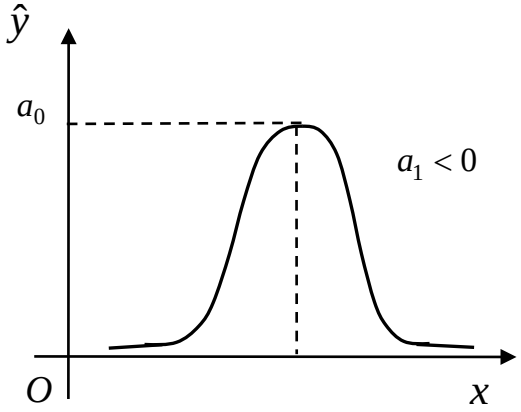


Рисунок 5.3.

Для опису процесів в демографії, маркетингу застосовують **криву Гомперця**

$$\hat{y} = e^{a_0 \cdot a_1^x + a_2},\ 0 < a_1 < 1.$$

Логарифмуванням ця крива зводиться до модифікованої експоненціальної моделі

$$\hat{y}^* = a_0 \cdot a_1^x + a_2,$$

де $\hat{y}^* = \ln \hat{y}$.

Графік цієї залежності наведено на рисунку 5.4.

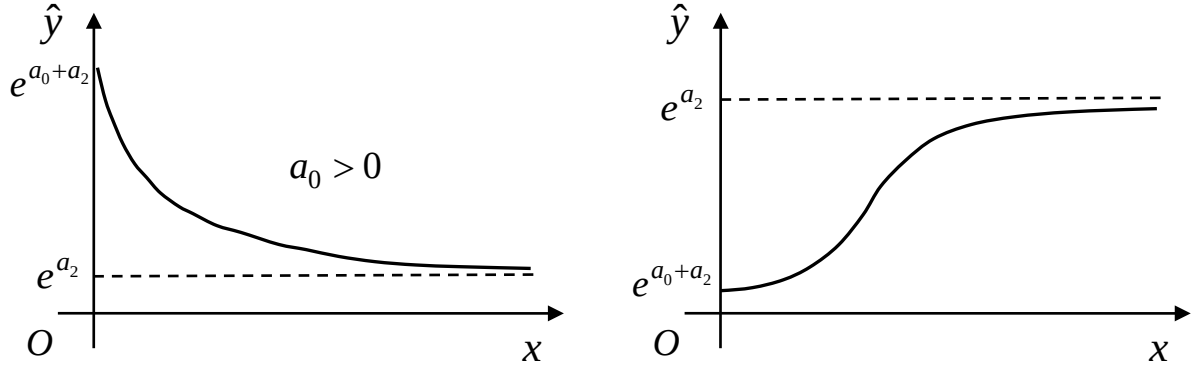


Рисунок 5.4.

Зворотною до модифікованої експоненти є **логістична крива**

$$\hat{y} = \frac{1}{a_0 \cdot a_1^x + a_2},\ 0 < a_1 < 1,\ a_0 > 0,\ a_2 > 0.$$

Її графік зображено на рисунку 5.5.

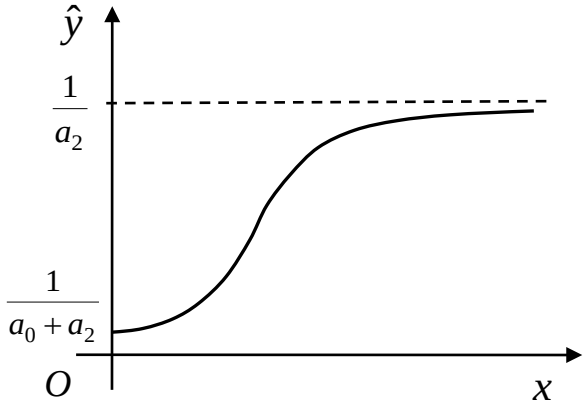


Рисунок 5.5.

Задача 5.3. *За даними таблиці побудувати експоненційну модель (5.3), яка описує залежність величини доходу (y, ум. од.) від рівня заощаджень (x, ум. од.).*

Місяць	Дохід, умовних одиниць	Заощадження, умовних одиниць
1	11,8	2,36

2	11,4	2,2
3	12	2,08
4	12,6	2,2
5	13	2,1
6	13,9	2,12
7	14,7	2,41
8	15,5	2,5
9	16,3	2,43
10	17,5	2,59
11	18,7	2,9
12	19,7	2,95

○ Для одержання лінійної залежності, як було показано вище, застосуємо логарифмування $\ln \hat{y} = \ln a_0 + x \ln a_1$ і заміну змінних $\ln \hat{y} = \hat{y}^*$, $\ln a_0 = a_0^*$, $\ln a_1 = a_1^*$. Лінійну залежність шукаємо у вигляді

$$\hat{y}^* = a_0^* + a_1^* x .$$

Статистичні оцінки a_0^*, a_1^* рівняння регресії, із врахуванням замін, задовольняють системі рівнянь:

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

Для знаходження коефіцієнтів цієї системи складемо розрахункову таблицю 5.3.

Таблиця 5.3

i	x_i	y_i	$y_i^* = \ln y$	x_i^2	$x_i y_i^*$
1	2,36	11,8	2,468	5,570	5,825
2	2,2	11,4	2,434	4,840	5,354
3	2,08	12	2,485	4,326	5,169
4	2,2	12,6	2,534	4,840	5,574
5	2,1	13	2,565	4,410	5,386
6	2,12	13,9	2,632	4,494	5,580
7	2,41	14,7	2,688	5,808	6,478
8	2,5	15,5	2,741	6,250	6,852
9	2,43	16,3	2,791	5,905	6,783
10	2,59	17,5	2,862	6,708	7,413
11	2,9	18,7	2,929	8,410	8,493
12	2,95	19,7	2,981	8,703	8,793
Σ	28,84	177,1	32,108	70,264	77,698

$\overline{x} = 28,84/12 = 2,403$, $\overline{y^*} = 32,108/12 = 2,676$, $\overline{x^{*2}} = 70,264/12 = 5,855$, $\overline{xy^*} = 77,698/12 = 6,475$.

Отримуємо систему лінійних рівнянь:

$$\begin{cases} a_0^* + 2,403a_1^* = 2,676, \\ 2,403a_0^* + 5,855a_1^* = 6,475. \end{cases}$$

Розв'язавши її отримаємо: $a_0^* = 1,334$, $a_1^* = 0,558$.

Вибіркове рівняння лінійної регресії має наступний вигляд:

$$\hat{y}^* = 1,334 + 0,558x.$$

Перейдемо до початкових змінних ($a_0 = e^{a_0^*} = e^{1,334} = 3,797$, $a_1 = e^{a_1^*} = e^{0,558} = 1,748$) і отримаємо експоненційну модель:

$$\hat{y} = 3,797 \cdot 1,748^x.$$

На рисунку 5.6 показано графік експоненційної моделі і діаграма розсіювання.

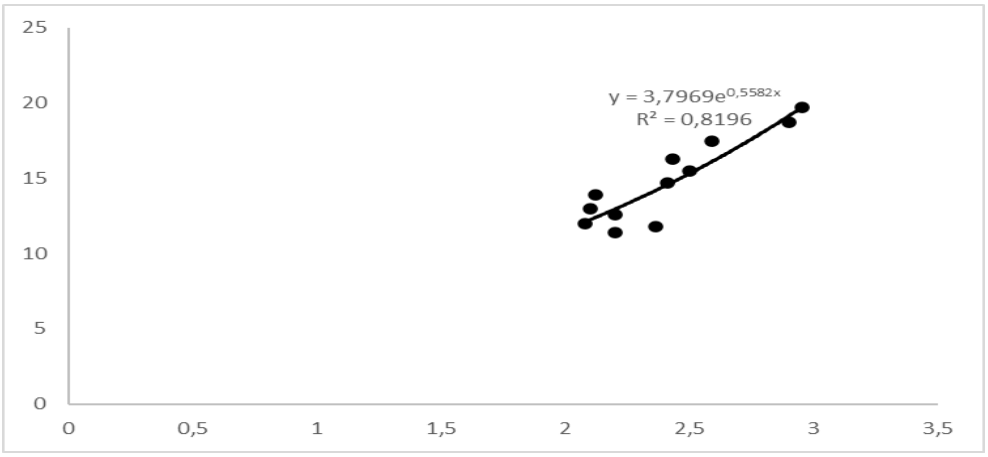


Рисунок 5.6

§6. СТАТИСТИЧНЕ ОЦІНЮВАННЯ І ТЕСТИ В УЗАГАЛЬНЕНИХ РЕГРЕСІЙНИХ МОДЕЛЯХ

1. Узагальнена лінійна регресійна модель.
2. Знаходження оцінок узагальненої моделі.
3. Прогноз на підставі узагальненої моделі.
4. Спеціальні форми коваріаційної матриці вектора збурень в узагальненій регресійній моделі.
4.1. Гетероскедастичність збурень.
4.2. Автокореляція збурень першого порядку.

1. Класична модель регресії, яка розглядалася в §4, була в принципі лише теоретичною основою для статистичної обробки даних, отриманих у строго контрольованих лабораторних умовах.

Якщо передумови класичної моделі не виконуються, тоді МНК-оцінки втрачають бажані статистичні властивості.

Стратегія дослідників у такому випадку полягає у наступному.

1) Класичну модель регресії узагальнюють так, щоб вона була максимально адаптована до умов емпіричних економічних і соціальних досліджень.

2) Модифікують методи оцінок і тестів, які в узагальненій моделі регресії по можливості максимально забезпечують бажані статистичні властивості. Такі узагальнені методи називаються **економетричними**.

Реалізуючи перший напрямок стратегії, розглянемо узагальнену лінійну модель множинної регресії

$$Y = X\alpha + U, \tag{6.1}$$

в якій змінні і параметри визначені аналогічно §4 і виконуються наступні передумови:

Передумова 1. *U – випадковий вектор, X – детермінована матриця;*

Передумова 2. $M(U) = 0$;

Передумова 3. $\sum_U M(UU') = \sigma^2\Omega$, (6.2)

де σ^2 – невідомий параметр, Ω – відома симетрична додатно визначена матриця порядку n.

Передумова 4. $rank(X) = m+1 < n$,

де m – число пояснюючих змінних, n – число спостережень.

Порівняння узагальненої моделі з класичною вказує на те, що вони відрізняються тільки видом коваріаційної матриці вектора U: замість $\sum_U = \sigma^2 I_n$ для класичної моделі покладається $\sum_U = \sigma^2 \Omega$ для узагальненої.

З допомогою таким чином узагальненої моделі можна вивчити, зокрема, проблему гетероскедастичності, проблему автокореляції тощо.

Зауваження. *Якщо елемент матриці* Ω $\sigma_{ij} > 0$ $(i \neq j)$, *то випадкові величини* U_i *та* U_j *варіюють в одному напрямку: додатна (від'ємна) зміна реалізації* u_i *пов'язана з такою ж тенденцією зміни реалізації* u_j. *Це означає, що змінні* U_i *та* U_j *додатно корельовані. Якщо* $\sigma_{ij} < 0$, *то змінні* U_i *та* U_j *від'ємно корельовані: додатна зміна реалізації* u_i *пов'язана із від'ємним проявом реалізації* u_j *і навпаки. Якщо ж* $\sigma_{ij} = 0$, *то збурення* U_i *та* U_j *некорельовані; вони будуть також незалежними, якщо розподілені за нормальним законом.*

2. Якщо формально використати МНК до узагальненої моделі, то отримаємо оцінку (див. (4.8))

$$a = (X'X)^{-1} X'Y \tag{6.3}$$

вектора a, яка володіє властивостями незміщеності і спроможності, однак не є ефективною.

Відповідь на питання про вид ефективної оцінки вектора α дає наступне твердження.

Теорема Айткена. В класі лінійних незміщених оцінок вектора α для узагальненої регресійної моделі компоненти оцінки

$$a^* = (X'\Omega^{-1}X)^{-1} X'\Omega^{-1}Y \tag{6.4}$$

мають найменші дисперсії.

□ З теорії матриць відомо, що для довільної невиродженої симетричної матриці існує невироджена матриця P така, що

$$\Omega = PP'. \tag{6.5}$$

Врахувавши властивості обернених квадратних матриць, тобто $(AB)^{-1} = B^{-1}A^{-1}$ і $(P')^{-1} = (P^{-1})'$, отримаємо

$$\Omega^{-1} = (P^{-1})'P^{-1}. \tag{6.6}$$

З (6.5) випливає рівність

$$P^{-1}\Omega(P^{-1})' = I_n. \tag{6.7}$$

Помноживши обидві частини узагальненої моделі (6.1) зліва на P^{-1}, отримаємо

$$Y_* = X_*\alpha + U_*, \tag{6.8}$$

де

$$Y_* = P^{-1}Y,\ X_* = P^{-1}X,\ U_* = P^{-1}U. \tag{6.9}$$

Модель (6.8) задовольняє всім вимогам класичної лінійної моделі множинної регресії, оскільки

$M(U_*) = M(P^{-1}U) = P^{-1}M(U) = O_n,\ \sum_{U_*} = M(U_*U_*') = M\left\{(P^{-1}U)(P^{-1}U)'\right\} =$

$= M\left\{P^{-1}UU'(P^{-1})'\right\} = P^{-1}M(UU')(P^{-1})' = \sigma^2 P^{-1}\Omega(P^{-1})' = \sigma^2 I_n$ (враховано рівність (6.7)), $rankX_* = rank(P^{-1}X) = rankX = m+1 < n,$ оскільки матриця P^{-1} невироджена. А тому на підставі теореми Гаусса-Маркова найбільш ефективною в класі лінійних незміщених оцінок є оцінка виду (6.3):

$$a^* = (X_*'X_*)^{-1}X_*'Y_*. \tag{6.10}$$

Повертаючись до вихідних спостережень X та Y і враховуючи (6.6), отримаємо шукану оцінку (6.4):

$$a^* = \left[(P^{-1}X)'(P^{-1}X)\right]^{-1}(P^{-1}X)'P^{-1}Y = \left[X'(P^{-1})'P^{-1}X\right]^{-1}X'(P^{-1})'P^{-1}Y =$$

$$= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y.$$

Нарешті, з (6.10) і (6.8) випливає рівність $M(a^*) = \alpha$, тобто незміщеність оцінки a^*.

Оцінка (6.10) мінімізує суму квадратів

$$Q = E_*'E_* = (Y_* - X_*a)'(Y_* - X_*a),$$

тобто є результатом використання МНК (звичайного). Якщо перейти до вихідних спостережень, то з урахуванням (6.6)

$$Q = \left[P^{-1}(Y - Xa)\right]'\left[P^{-1}(Y - Xa)\right] = (Y - Xa)'(P^{-1})'P^{-1}(Y - Xa) =$$

$$= (Y - Xa)'\Omega^{-1}(Y - Xa) = E'\Omega^{-1}E,$$

тобто a^* можна назвати оцінкою узагальненого метода найменших квадратів (УМНК), яка мінімізує узагальнений критерій $E'\Omega^{-1}E$.

Слід відмітити, що для узагальненої регресійної моделі, на відміну від класичної, коефіцієнт детермінації, обчислений за формулою

$$R^2 = 1 - \frac{(Y - Xa^*)'(Y - Xa^*)}{(Y - \bar{Y})'(Y - \bar{Y})},$$

де a^* визначене (6.4), не володіє задовільною мірою якості моделі. Справа в тому, що розклад загальної суми квадратів СКЗ на складові СКП і СКН здійснювався у припущенні наявності вільного члена в узагальненій моделі. Однак якщо у вихідній моделі (6.1) міститься вільний член, то не можна гарантувати його присутність у перетвореній моделі (6.8).

Використавши (6.8), (6.9) і (6.6), знайдемо коваріаційну матрицю вектора a^*:

$$\begin{aligned}\sum_{a^*} &= \sigma^2 (X_*' X_*)^{-1} = \sigma^2 \left[(P^{-1}X)' P^{-1} X \right]^{-1} = \\ &= \sigma^2 \left[X' (P^{-1})' P^{-1} X \right]^{-1} = \sigma^2 (X' \Omega^{-1} X)^{-1}.\end{aligned} \tag{6.11}$$

Незміщена оцінка $\hat{\sigma}^2$ для σ^2 з урахуванням (6.6) має такий вигляд:

$$\begin{aligned}\hat{\sigma}^2 &= \frac{E_*' E_*}{n-m-1} = \frac{1}{n-m-1} (Y_* - X_* a^*)' (Y_* - X_* a^*) = \\ &= \frac{1}{n-m-1} (Y - Xa^*)' \Omega^{-1} (Y - Xa^*) = \frac{1}{n-m-1} E' \Omega^{-1} E.\end{aligned} \tag{6.12}$$

Якщо матриця Ω відома, тоді на підставі оцінки a^* та її коваріаційної матриці стандартним чином можна сконструювати звичайні критерії значущості і довірчі інтервали для α_i $(i = \overline{0,m})$.

Узагальнена лінійна модель множинної регресії іноді специфікується у вигляді (6.1), де передумова 3 записується таким чином:

$$\sum_U = M(UU') = V, \tag{6.13}$$

де V – відома симетрична додатно визначена матриця, тобто відмінність полягає у заміні в (6.2) $\sigma^2 \Omega$ на V. Здійснивши відповідну підстановку в (6.4) і (6.11), отримаємо вирази для оцінки, знайдені узагальненим методом найменших квадратів:

$$a^* = (X' V^{-1} X)^{-1} X' V^{-1} Y. \tag{6.14}$$

і для її коваріаційної матриці:

$$\sum_{a^*} = (X' V^{-1} X)^{-1}. \tag{6.15}$$

3. У випадку узагальненої моделі проблема прогнозування вимагає спеціального дослідження. Розглянемо модель (6.1), відносно якої виконуються передумови 1-5 із заміною (6.2) на (6.13). Задача полягає у передбаченні прогнозного значення залежної змінної y_0 для заданого вектор-рядка x_0.

Ми можемо записати

$$y_0 = x_0\alpha + U_0, \tag{6.16}$$

де U_0 – справжнє, але невідоме значення збурення у прогнозний момент. Нехай

$$M(U_0) = 0, \tag{6.17}$$

$$M(U_0^2) = \sigma_0^2, \tag{6.18}$$

$$M(U_0U) = \begin{pmatrix} M(U_1U_0) \\ M(U_2U_0) \\ \cdots \\ M(U_nU_0) \end{pmatrix} = W, \tag{6.19}$$

де W – n-вимірний вектор коваріацій прогнозного збурення з вектором збурень U. Розглянемо лінійний прогноз

$$p = c'Y, \tag{6.20}$$

де c – вектор-стовпець, що складається із n констант. Для того щоб вектор p став найкращим лінійним незміщеним прогнозом, необхідно обрати вектор c, що мінімізує дисперсію прогнозу

$$\sigma_p^2 = M\left[(p - y_0)^2\right], \tag{6.21}$$

що досягається при $M(p - y_0) = 0$. Із (6.20), (6.1) і (6.16) отримаємо

$$p - y_0 = c'Y - x_0\alpha - U_0 = (c'X - x_0)\alpha + c'U - U_0.$$

Із умов незміщеності прогнозу випливає, що вектор c повинен задовольняти рівність

$$c'X - x_0 = 0. \tag{6.22}$$

Тоді для помилки прогнозування отримаємо

$$p - y_0 = c'U - U_0, \tag{6.23}$$

але оскільки ліва частина – скаляр, то дисперсія прогнозу з урахуванням (6.23), (6.13) і (6.19) дорівнює

$$\begin{aligned} \sigma_p^2 = M\left[(p - y_0)^2\right] = M\left[(p - y_0)(p - y_0)'\right] = M\left[(c'U - U_0)(c'U - U_0)'\right] = \\ = M\left[c'UU'c + U_0^2 - 2c'UU_0\right] = c'Vc + U_0^2 - 2c'W. \end{aligned} \tag{6.24}$$

Для мінімізації (6.24) при умові (6.22) утворимо функцію Лагранжа

$$\varphi(c, \lambda) = c'Vc - 2c'W - 2(c'X - x_0)\lambda,$$

де λ – $(m+1)$-вимірний вектор-стовпець, утворений множниками Лагранжа. Продиференціювавши функцію φ по c та λ і прирівнявши похідні нульовому вектору, отримаємо матричне рівняння

$$\begin{pmatrix} V & X \\ X' & 0 \end{pmatrix}\begin{pmatrix} c \\ -\lambda \end{pmatrix}=\begin{pmatrix} W \\ x_0' \end{pmatrix},$$

звідки

$$\begin{pmatrix} \hat{c} \\ -\hat{\lambda} \end{pmatrix}=\begin{pmatrix} V & X \\ X' & 0 \end{pmatrix}^{-1}\begin{pmatrix} W \\ x_0' \end{pmatrix}.$$

Використавши правило знаходження оберненої матриці до блочної матриці, знайдемо шуканий вектор c:

$$c = V^{-1}\left[I_n - X(X'V^{-1}X)^{-1}X'V^{-1} \right]W + V^{-1}X(X'V^{-1}X)^{-1}x_0'.$$

А тому згідно з (6.20) і (6.14) найкращим лінійним незміщеним прогнозом буде

$$\hat{p} = \hat{c}Y = x_0(X'V^{-1}X)^{-1}X'V^{-1}Y + W'V^{-1}Y - W'V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}Y =$$

$$= x_0 a^* + W'V^{-1}(Y - Xa^*)$$

або остаточно

$$\hat{p} = x_0 a^* + W'V^{-1}E, \qquad (6.25)$$

де $E = Y - Xa^*$ – вектор залишків, що відповідає узагальненому методу найменших квадратів.

4. Для застосування узагальненого метода найменших квадратів необхідне знання коваріаційної матриці (6.2) вектора збурень, що зустрічається досить рідко у практиці економетричного моделювання. Якщо вважати всі $n(n+1)/2+1$ елементів матриці Ω_U невідомими, враховуючи симетричність Ω_U і σ^2 (в доповненні до $m+1$ параметрів вектора α), то здійснити оцінку всіх невідомих на підставі n спостережень неможливо. Тому для практичної реалізації узагальненого метода найменших квадратів необхідно вводити додаткові умови на структуру матриці Ω.

Далі розглянемо найбільш важливі і часто досліджувані види структур матриці Ω або *V*, визначеної (6.13).

4.1. Нехай у вихідній моделі (4.2) на підставі тестів встановлена наявність гетероскедастичності збурень, а самі збурення не корельовані. Тоді коваріаційна матриця вектора збурень має такий вид

$$\sum_U = M(UU') = V = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_n^2 \end{pmatrix}. \qquad (6.26)$$

В якості діагональних елементів беруться значення $\hat{\sigma}_i$, знайдені за тестами Уайта або Глейзера. Тоді

$$V^{-1} = \begin{pmatrix} \sigma_1^{-2} & 0 & \dots & 0 \\ 0 & \sigma_2^{-2} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_n^{-2} \end{pmatrix}$$

і за формулою (6.14) знаходиться вектор оцінки a^*. Використання цієї формули, тобто узагальнений метод найменших квадратів для моделі із гетероскедастичністю з матрицею коваріацій збурень (6.26), називається ще **зваженим методом найменших квадратів**.

4.2. Для класичної лінійної моделі згідно з передумовою 3 $\text{cov}(U_i, U_j) = 0, i \neq j$, що означає попарну незалежність збурень, якщо вони нормально розподілені. Для перехресних даних така гіпотеза рівносильна припущенню про відсутність впливу з боку збурень, діючих на якийнебудь один із спостережуваних об'єктів, на збурення, яким піддаються решта спостережувані об'єкти. При використанні часових рядів послідовні збурення, діючі в різні моменти часу, повинні бути незалежними.

Тим не менше доводиться зустрічатися із ситуаціями, в яких припущення про незалежність послідовних збурень виявляється не дуже правдоподібним. Наприклад, може бути вибрана помилкова специфікація форми залежності між змінними. Припустимо, що ми зупинилися на лінійній залежності між змінними *Y* і *x*, в той час як справжня залежність виявилася, наприклад, квадратичною. Навіть у тому випадку, коли збурення в істинному співвідношенні не будуть автокорельовані, ті квазізбурення, які відповідають лінійному зв'язку, будуть містити член, залежний від x^2. Якщо існує кореляція між послідовними значеннями деякої пояснюючої змінної, то буде спостерігатися і кореляція послідовних значень збурення. Цей приклад є частковим випадком проблеми впливу на модель неврахованих пояснюючих змінних. В загальному випадку ми включаємо в модель лише деякі із суттєвих змінних, а вплив виключених із розгляду величин повинен знайти відображення у зміні збурюючої дії.

Дві основні альтернативи дій при автокореляції.

1. Попробувати змінити специфікацію моделі таким чином, щоб вилучити автокореляцію збурень (наприклад, ввести одну або більше додаткових незалежних змінних).

2. Обрати такий метод оцінки параметрів, який при наявності автокореляції збурень міг би по можливості максимально забезпечити потрібні властивості отриманих оцінок, наприклад, метод Айткена.

Розглянемо реалізацію другої альтернативи на прикладі авторегресійного процесу першого порядку (п. 5 §3).

Нехай для всіх t

$$Y_t = \alpha_0 + \alpha_1 x_t + U_t, \tag{6.27}$$

де за припущенням збурення U_t задовольняють схемі авторегресії першого порядку

$$U_t = \rho U_{t-1} + \varepsilon_t, \tag{6.28}$$

для якої $|\rho| < 1$, а відносно випадкової величини ε_t виконуються умови

$$M(\varepsilon_t) = 0,\ M(\varepsilon_t \varepsilon_{t+s}) = \begin{cases} \sigma_\varepsilon^2,\ s = 0, \\ 0,\ s \neq 0. \end{cases} \tag{6.29}$$

Тоді

$$U_t = \rho U_{t-1} + \varepsilon_t = \rho(\rho U_{t-2} + \varepsilon_{t-1}) + \varepsilon_t = \ldots = \varepsilon_t + \rho\varepsilon_{t-1} + \rho^2 \varepsilon_{t-2} + \ldots,$$

так, що

$$U_t = \sum_{j=0}^{\infty} \rho^j \varepsilon_{t-j}, \tag{6.30}$$

Таким чином, згідно з (6.29) $M(U_t) = 0$ і $M(U_t^2) = (1 + \rho^2 + \rho^4 + \ldots)\sigma_\varepsilon^2$.

Отже, для всіх t (з урахуванням суми нескінченної геометричної прогресії)

$$\sigma_u^2 = \frac{\sigma_\varepsilon^2}{1 - \rho^2}, \tag{6.31}$$

а коваріація послідовних значень збурень:

$$M(U_t U_{t-1}) = M\left[(\varepsilon_t + \rho\varepsilon_{t-1} + \rho^2\varepsilon_{t-2} + \ldots)(\varepsilon_{t-1} + \rho\varepsilon_{t-2} + \rho^2\varepsilon_{t-3} + \ldots)\right] =$$

$$= M\left\{\left[\varepsilon_t + \rho(\varepsilon_{t-1} + \rho\varepsilon_{t-2} + \ldots)\right](\varepsilon_{t-1} + \rho\varepsilon_{t-2} + \ldots)\right\} = \rho M\left[(\varepsilon_{t-1} + \rho\varepsilon_{t-2} + \ldots)^2\right] =$$

$$= \rho\sigma_\varepsilon^2(1 + \rho^2 + \rho^4 + \ldots) = \rho\sigma_u^2.$$

Аналогічно $M(U_t U_{t-s}) = \rho^2 \sigma_u^2$,

а в загальному випадку

$$M(U_t U_{t-s}) = \rho^s \sigma_u^2. \tag{6.32}$$

Цей вираз дозволяє зробити висновок про еквівалентність схеми (6.28) умові

$$M(UU') = V = \sigma_u^2 \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{n-3} \\ \dots & \dots & \dots & \dots & \dots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \dots & 1 \end{pmatrix}, \qquad (6.33)$$

де σ_u^2 визначається виразом (6.31).

Для практичної реалізації необхідно знайти оцінки $\hat{\rho}$ і $\hat{\sigma}_\varepsilon^2$ параметрів ρ і σ_ε^2 відповідно на підставі моделі (6.28). Підставивши ці оцінки в (6.33), отримаємо матрицю $\hat{V}$. Використання формули Айткена (6.14) із заміною V на $\hat{V}$ дозволяє знайти оцінку a^* параметрів узагальненої моделі. Ця оцінка називається Айткен-подібною оцінкою, оскільки насправді значення ρ невідоме, а тільки оцінене.

§7. Часові ряди

1. Часові ряди і їх числові характеристики
2. Тести стаціонарності часового ряду
3. Виділення трендової складової часового ряду
4. Виділення сезонної складової часового ряду

1. При дослідженні поведінки економічної системи у часі незалежною змінною є часовий параметр (година, день, місяць, рік), яки будемо позначати t. Тоді залежна змінна буде поєднувати два фактори: а) при фіксованому часі t є фіксованою величиною; б) є функцією аргумента t. Таку величину називають **випадковою функцією** або **випадковим процесом** і її будемо позначати $Y(t)$.

Часовою вибіркою випадкового процесу називається сукупність спостережень $\{y_1, y_2, \ldots, y_n\}$ випадкової величини $Y(t)$ в дискретні моменти часу t, $(t = 1, 2, \ldots, n)$.

Часова залежність даних забезпечує порядок слідування спостережних значень y_t по часовій вибірці. Це означає, що перестановка y_t по часовій вибірці може суттєво вплинути на характеристики досліджуваної залежності $Y(t)$.

Часовим рядом називається сукупність випадкових величин $\{Y_1, Y_2, \ldots, Y_n\}$, яка побудована з випадкової величини $Y(t)$ в моменти t, $(t = 1, 2, \ldots, n)$.

Серед часових рядів виділяють **одновимірні**, які отримують в результаті спостереження одної фіксованої характеристики досліджуваного об'єкта, і, **багатовимірні** часові ряди, які є результатом спостереження декількох характеристик одного об'єкта, який вивчається протягом послідовності моментів часу. За часом спостереження часові ряди поділяються на **дискретні** і **неперервні**. Дискретні ряди, в свою чергу, поділяються на ряди з **рівновіддаленими** і **довільними моментами спостереження**.

Часові ряди бувають детермінованими і випадковими: перші отримані як значення деякої невипадкової функції, а другі – як реалізація випадкової величини.

Надалі будемо розглядати одновимірні, дискретні з рівновіддаленими моментами спостережень випадкові часові ряди.

Значення елементів часового ряду формуються під впливом ряду факторів, серед яких виділяють:

– **довготермінові**, які формують в тривалій перспективі загальну тенденцію досліджуваної ознаки; ця тенденція описується за допомогою деякої функції – **тренду**;

– **сезонні**, які формують періодично повторювані у визначений час року коливання досліджуваної ознаки. Дію сезонних факторів описують з допомогою невипадкової періодичної функції, в аналітичному записі якої присутні гармоніки (тригонометричні функції);

– **циклічні**, які формують зміни досліджуваного об'єкту в результаті дії циклів економічної, демографічної природи;

– **випадкові**, які не піддаються обліку, як результат дії випадкових зовнішніх факторів.

Для опису часового ряду використовують адитивну та мультиплікативну моделі:

$$Y(t) = f(t) + U, \tag{7.1}$$

$$Y(t) = f(t) \cdot U, \tag{7.2}$$

де детермінована складова $f(t)$ може включати одну або декілька із наступних компонент: трендову $\tau(t)$, сезонну $s(t)$ і циклічну $c(t)$; U – випадкова складова.

До **основних задач аналізу часових рядів** відносяться:

– визначити, які з невипадкових функцій $\tau(t)$, $s(t)$ і $c(t)$ присутні в $f(t)$;

– побудувати «найкращі» оцінки для тих невипадкових функцій, які присутні в $f(t)$;

– побудова моделі, яка б адекватно описувала поведінку випадкової складової U, і статистично оцінити параметри цієї моделі.

Часовий ряд називається **стаціонарним** (у вузькому розумінні), якщо сумісний розподіл імовірностей n спостережень $Y_1, Y_2, \ldots, Y_n$ такий же, як і n спостережень $Y_{1+l}, Y_{2+l}, \ldots, Y_{n+l}$ при будь-яких n, t і l. Іншими словами для стаціонарного часового ряду його математичне сподівання і дисперсія випадкової величини не залежать від часу t. Тому математичне сподівання $M(Y(t))$ і дисперсія $D(Y(t))$ можна оцінити по спостереженням y_t $(t = \overline{1,n})$ відповідно за формулами:

$$\bar{y} = \frac{\sum_{t=1}^{n} y_t}{n}; \qquad s_t^2 = \frac{\sum_{t=1}^{n} \left(y_t - \bar{y}\right)^2}{n} \tag{7.3}$$

Ступінь тісноти зв'язку між послідовностями $Y_1, Y_2, \ldots, Y_n$ і $Y_{1+l}, Y_{2+l}, \ldots, Y_{n+l}$ (зсунутих один відносно одного на l моментів часу, або, як кажуть з лагом l) можна визначити за допомогою **коефіцієнта автокореляції**

$$\rho(l) = \frac{\operatorname{cov}\left(Y(t), Y(t+l)\right)}{\sigma(t)\sigma(t+l)} = \frac{M\{[Y(t)-a][Y(t+l)-a]\}}{\sigma^2},$$

де $M(Y(t)) = M(Y(t+l)) = a$, $\sigma = \sigma(t) = \sigma(t+l)$.

Оскільки коефіцієнт $\rho(l)$ визначає кореляцію між членами одного і того ж ряду, то його називають коефіцієнтом автокореляції, а залежність $\rho(l)$ – **автокореляційною функцією**. Для стаціонарного часового ряду Y_t $(t=\overline{1,n})$ автокореляційна функція $\rho(l)$ залежить тільки від лагу l і $\rho(-l)=\rho(l)$. Тому при дослідженні $\rho(l)$ можна обмежитися розглядом тільки додатних значень l. Якщо $l=0$, то $\rho(0)=1$.

Оцінкою для $\rho(l)$ є **вибірковий коефіцієнт кореляції**, який обчислюється за формулою коефіцієнта кореляції (1.20**), в якій $x_i=y_t$, $y_i=y_{t+l}$, а n замінено на $n+l$:

$$r(l)=\frac{(n-l)\sum\limits_{t=1}^{n-l}y_t y_{t+l}-\left(\sum\limits_{t=1}^{n-l}y_t\right)\left(\sum\limits_{t=1}^{n-l}y_{t+l}\right)}{\sqrt{(n-l)\sum\limits_{t=1}^{n-l}y_t^2-\left(\sum\limits_{t=1}^{n-l}y_t\right)^2}\sqrt{(n-l)\sum\limits_{t=1}^{n-l}y_{t+l}^2-\left(\sum\limits_{t=1}^{n-l}y_{t+l}\right)^2}}. \qquad (7.4)$$

Функція $r(l)$ називається **вибірковою автокореляційною функцією**, а її графік – **корелограмою**.

Зауважимо, що із збільшенням l число $n-l$ пар спостережень y_t, y_{t+l} зменшується, тому лаг l повинен бути порівняно великим (рекомендують $l \le n/4$).

Для стаціонарного часового ряду із збільшенням лага l зв'язок членів часового ряду Y_t і Y_{t+l} слабшає і автокореляційна функція $\rho(l)$ повинна спадати (по абсолютній величині). В той же час для її оціночної функції $r(l)$ при невеликому числі пар спостережень $n-l$ властивість монотонного спадання при збільшенні l може порушуватися.

2. Для з'ясування стаціонарності часового ряду достатньо перевірити сталість математичного сподівання та дисперсії на всьому інтервалі визначення часового ряду.

Спочатку наведемо деякі критерії перевірки статистичної гіпотези $H_0: M(Y(t))=const$ та альтернативної гіпотези $H_1: M(Y(t))\neq const$.

І. Критерій Стьюдента. Часовий ряд Y_t $(t=\overline{1,n})$ розбивається на дві частини (не обов'язково одинакові) по кількості спостережень y_t. Нехай перша частина містить n_1 спостережень, а друга частина – містить n_2 спостережень.

Для кожної частини часового ряду обчислимо (використовуючи формули (7.3)) вибіркові середні $\overline{y}_1$, $\overline{y}_2$ і вибіркові дисперсії s_1^2, s_2^2.

Нульова гіпотеза про сталість математичного сподівання відхиляється на рівні значущості α, якщо виконується нерівність

$$t_{спост.} = \frac{\left|\overline{y}_1 - \overline{y}_2\right|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} > t_{двост.кр.}(1-\alpha; n_1 + n_2 - 2).$$

II. Критерій серій. Впорядкуємо члени ряду по зростанню $y_1, y_2, \ldots, y_n$. Визначимо медіану ряду

$$y_{med} = \begin{cases} y_{\frac{n+1}{2}}, & \textit{якщо } n \textit{ непарне,} \\ \frac{1}{2}\left(y_{\frac{n}{2}} + y_{\frac{n}{2}+1} \right), & \textit{якщо } n \textit{ парне.} \end{cases}$$

Утворимо послідовності плюсів і мінусів, тобто замість кожного члена y_t ставимо плюс, якщо $y_t > y_{med}$, і ставимо мінус, якщо $y_t < y_{med}$. Утворена послідовність плюсів і мінусів характеризується загальним числом серій ν і тривалістю самої серії μ. Під серією будемо розуміти послідовність підряд розміщених плюсів і підряд розміщених мінусів. Підраховуємо загальне число серій ν і протяжність найдовшої серії μ.

Нульова гіпотеза про сталість математичного сподівання відхиляється на рівні значущості $\alpha \in (0{,}05; 0{,}0975)$, якщо не виконується хоча б одна з наступних нерівностей:

$$\nu > \frac{1}{2}(n + 2 - 1{,}96\sqrt{n-1}), \quad \mu < 1{,}43\ln(n+1).$$

III. Критерій «висхідних» і «низхідних» серій. За аналогією до попереднього критерію досліджуються послідовності плюсів і мінусів. Правило побудови послідовності наступне: на t-му місці часового ряду ставиться плюс, якщо $y_{t+1} - y_t > 0$, і мінус, якщо $y_{t+1} - y_t < 0$ (якщо підряд йде кілька однакових спостережень, то до уваги береться тільки одне з них). Очевидно, що послідовність підряд розміщених плюсів відповідає зростанню результатів спостереження (висхідна серія), а послідовність мінусів – їх спаданню (низхідна серія).

Гіпотеза H_0 про сталість математичного сподівання відхиляється на рівні значущості $\alpha \in (0{,}05; 0{,}0975)$, якщо не виконується хоча б одна з наступних нерівностей:

$$\nu > \frac{1}{3}(2n-1) - 1{,}96\sqrt{\frac{16n-29}{90}}, \quad \mu < \mu_0,$$

де ν – загальне число серій, μ – протяжність найдовшої серії, μ_0 – величина, яка залежить від n наступним чином:

n	$n \le 26$	$26 < n \le 153$	$153 < n \le 1170$
μ_0	$\mu_0 = 5$	$\mu_0 = 6$	$\mu_0 = 7$

Задача 7.1. *Дані про кількість продукції (тис.од.), які продані підприємством протягом 16 останніх кварталів наведені у табл.7.1. Здійснити перевірку гіпотезу про сталість середнього обсягу реалізації продукції, використовуючи критерій «висхідних» і «низхідних» серій.*

Таблиця 7.1

Квартал, t	Об'єм продажів, y_t	Квартал, t	Об'єм продажів, y_t	Квартал, t	Об'єм продажів, y_t	Квартал, t	Об'єм продажів, y_t
1	6	5	7,2	9	8	13	9
2	4,4	6	4,8	10	5,6	14	6,6
3	5	7	6	11	6,4	15	7
4	9	8	10	12	11	16	10,8

○ Загальне число спостережень $n=16$. Побудуємо послідовність із плюсів та мінусів.

№	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
+/–	–	+	+	–	–	+	+	–	–	+	+	–	–	+	+	–

Аналіз отриманої послідовності плюсів і мінусів дає $\nu=9$ і $\mu=2$.

Оскільки $9>\frac{1}{3}(2\cdot16-1)-1{,}96\sqrt{\frac{16\cdot16-29}{90}}=7{,}22$ і $2<\mu_0=5$, то нульова гіпотеза про сталість математичного сподівання приймається. ◎

IV. Далі сформулюємо двосторонній критерій перевірки статистичної гіпотези H_0: $D(Y(t))=const$ та альтернативної гіпотези H_1: $D(Y(t))\neq const$.

Гіпотеза H_0 про сталість дисперсії відхиляється на рівні значущості α, якщо не виконується нерівність

$$F_{кр.}\left(\frac{\alpha}{2};n_1-1;n_2-1\right)\le F_{спост.}\le F_{кр.}\left(1-\frac{\alpha}{2};n_1-1;n_2-1\right),$$

де $F_{спост.}=\frac{s_1^2}{s_2^2}$, n_1, n_2 – кількість спостережень розбиття часового ряду в критерії I.

3. Як говорилося раніше, однією з найважливіших задач дослідження економічного часового ряду є виявлення детермінованої складової $f(t)$ моделі (7.1) або (7.2), тобто побудови оціночного рівняння регресії $\hat{f}(t)$

для функції $f(t)$ (або оцінок $\hat{f}_t$ для значень $f(t)$) по заданій вибірці $\{(t, y_t), t = \overline{1,n}\}$. Для розв'язування даної задачі можливі кілька методів.

Методи першого типу (**аналітичні**) ґрунтуються на припущені, що відомо загальний вигляд детермінованої складової $f(t)$ в моделі (7.1) або (7.2). Тоді задача виділення детермінованої складової (або задача згладжування часового ряду) зводиться до знаходження оцінок для невідомих параметрів функції $f(t)$.

Методи другого типу (**алгоритмічні**) не пов'язані обмеженням щодо аналітичного вигляду шуканої функції $f(t)$. Такі методи дають лише алгоритм обчислення оцінки $\hat{f}_t$ для значення $f(t)$ в будь-якій наперед заданій точці t без явного представлення функції $f(t)$.

Аналітичні методи виділення трендової складової часового ряду. Нехай детермінована складова $f(t)$ визначається лише трендовою компонентою $\tau(t)$, тобто інші складові s_t і c_t часового ряду відсутні. Тоді модель (7.1) перепишемо у вигляді

$$Y = \tau(t) + U. \tag{7.5}$$

З рівняння (7.5) при значеннях $t = \overline{1,n}$ отримується система n рівнянь

$$Y_t = \tau(t) + U_t, \quad t = \overline{1,n}. \tag{7.6}$$

Оцінкою моделі (6.6) по вибірці $\{(t, y_t), t = \overline{1,n}\}$ є система n рівнянь

$$y_t = \hat{\tau}_t + u_t, \quad t = \overline{1,n}, \tag{7.7}$$

де $\hat{\tau}_t$ – групова (середня) змінної Y, знайдена за рівнянням $\hat{\tau}_t = \hat{\tau}(t)$, u_t – вибіркова оцінка збурення U_t.

Для побудови тренду необхідно вибрати вид функції $\tau(t)$. Найчастіше використовують наступні функції:

– лінійна: $\tau(t) = \alpha_0 + \alpha_1 t$;

– степенева: $\tau(t) = \alpha_0 \cdot t^{\alpha_1}$;

– гіперболічна: $\tau(t) = \alpha_0 + \dfrac{\alpha_1}{t}$;

– експоненціальна: $\tau(t) = e^{\alpha_0 + \alpha_1 t}$;

– поліноміальна другого або більш високого порядку:

$\tau(t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \ldots + \alpha_p t^p$.

Вибір виду функції $\tau(t)$ часто ґрунтується на аналізі діаграми розсіювання, яка побудована по точках $\{(t, y_t), t = \overline{1,n}\}$.

Обчислення оцінок невідомих параметрів тренду виконується за допомогою МНК, якщо припустити виконання передумов стосовно випад-

кової складової U_t (див. §2). В якості залежної змінної виступає сукупність спостережень $y_1, y_2, \ldots, y_n$, а незалежної змінної є час $t = 1, 2, \ldots, n$.

Згідно МНК для лінійного рівняння регресії $\hat{\tau}_t = a_0 + a_1 t$ оцінки a_0, a_1 знаходяться із системи нормальних рівнянь

$$\begin{cases} a_0 + \bar{t} a_1 = \bar{y}, \\ \bar{t} a_0 + \overline{t^2} a_1 = \overline{ty} \end{cases} \tag{7.8}$$

за формулами

$$a_1 = \frac{\overline{ty} - \bar{t} \cdot \bar{y}}{\overline{t^2} - (\bar{t})^2}, \quad a_0 = \bar{y} - a_1 \bar{t}, \tag{7.9}$$

де $\bar{y} = \frac{1}{n}\sum_{t=1}^{n} y_t$, $\bar{t} = \frac{1}{n}\sum_{t=1}^{n} t$, $\overline{t^2} = \frac{1}{n}\sum_{t=1}^{n} t^2$, $\overline{ty} = \frac{1}{n}\sum_{t=1}^{n} t y_t$.

Врахувавши відомі математичні формули $\sum_{t=1}^{n} t = \frac{n(n+1)}{2}$, $\sum_{t=1}^{n} t^2 = \frac{n(n+1)(2n+1)}{6}$, отримаємо наступні співвідношення

$$\bar{t} = \frac{n+1}{2}, \quad \overline{t^2} = \frac{(n+1)(2n+1)}{6}. \tag{7.10}$$

Для знаходження оцінок невідомих параметрів нелінійних трендів необхідно провести лінеаризацію моделей (див. §5).

Якщо ж функція $\tau(t)$ має вигляд полінома p-го порядку

$$\tau(t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \ldots + \alpha_p t^p$$

і при цьому довжина часового ряду n суттєво перевищує степінь цього полінома (зазвичай вимагається виконання $n \geq 4p$), то після проведення замін змінних $x_{ti} = t^i$, $i = \overline{1, p}$ можна знайти оцінки $a_1, a_2, \ldots, a_p$ параметрів $\alpha_1, \alpha_2, \ldots, \alpha_p$ за МНК в матричній формі для лінійної моделі множинної регресії (див. п.2, §4)

$$a = (X'X)^{-1} X'Y,$$

де

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 1 & 1^2 & \ldots & 1^p \\ 1 & 2 & 2^2 & \ldots & 2^p \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & n & n^2 & \ldots & n^p \end{pmatrix}, \quad a = \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{pmatrix}.$$

При застосуванні поліноміальної функції важливо правильно визначити степінь полінома. Для цього можна використати **метод послідовних різниць** [17, 18], який полягає в обчисленні різниць:

– першого порядку $\Delta y_t = y_t - y_{t-1},\ t = \overline{1, n-1};$

– другого порядку $\Delta y_t^2 = \Delta y_t - \Delta y_{t-1},\ t = \overline{1, n-2};$

– k-го порядку $\Delta y_t^k = \Delta y_t^{k-1} - \Delta y_{t-1}^{k-1},\ t = \overline{1, n-k},$

а також величин

$$d^{(k)} = \frac{\dfrac{1}{n-k}\sum_{t=1}^{n-k}\left(\Delta y_t^k\right)^2}{C_{2k}^k},$$

де C_{2k}^k – число комбінацій.

Величина $d^{(k)}$ спадає із збільшенням k, а потім, починаючи з деякого значення k_0 стабілізуватися. Тоді степінь полінома визначається з формулою $p = k_0 - 1$.

Далі розглянемо **алгоритмічні методи виділення трендової складової часового ряду (методи ковзної середньої МКС).** В основі цих методів лежить наступне: якщо власне розсіювання значень члена часового ряду $Y(t)$ навколо свого середньої (згладженого) значення $a = M(Y(t))$ характеризується дисперсією $\sigma^2 = D(Y(t))$, то розкид середнього з n членів часового ряду $(Y_1 + Y_2 + \ldots + Y_n)/n$ навколо того ж значення a буде характеризуватися значно меншою величиною дисперсії, а саме значенням σ^2/n. А зменшення величини розкиду якраз і означає згладжування відповідної траєкторії.

Оцінка $\hat{\tau}_t$ для значення функції $\tau(t)$ в момент часу t будується як середньозважене значень $y_{t-m}, y_{t-m+1}, \ldots, y_t, y_{t+1}, \ldots, y_{t+m}$ за формулою

$$\hat{\tau}_t = \sum_{i=-m}^{m} c_t y_{t+i},\ t = m+1, m+2, \ldots, n-m, \qquad (7.11)$$

де m – деяке число (як правило $m < n/3$), яке залежить від специфікації вихідних даних, c_t – вагові коефіцієнти, які задовольняють умову

$$\sum_{i=-m}^{m} c_t = 1.$$

Довжина інтервалу сумування в (7.11) дорівнює $(2m+1)$ точки і цей інтервал «ковзає» по осі часу.

Визначення коефіцієнтів c_t полягає у наступному. Відповідно до теореми Вейєрштраса будь-яку гладку функцію $\tau(t)$ при певних припущеннях можна подати поліномом степеня p в околі точки t. Тому беремо перші $2m+1$ члени часового ряду $y_1, y_2, \ldots, y_{2m+1}$, будуємо з допомогою МНК поліном $\hat{\tau}_1(t)$ степеня p, який наближає поведінку цієї початкової частини часового ряду. Використовуємо цей поліном для знаходження

оцінки $\hat{\tau}_t$ згладжуваного значення $\tau(t)$ в середній (тобто $(m+1)$-й) точці цього відрізка часу, тобто вважаємо $\hat{\tau}_{m+1} = \hat{\tau}_1(m+1)$. Потім «ковзаємо» по осі часу на один такт і так само підбираємо поліном $\hat{\tau}_2(t)$ того ж степеня для відрізка часового ряду $y_2, y_3, \ldots, y_{2m+2}$, визначаємо оцінку $\hat{\tau}_{m+2} = \hat{\tau}_2(m+2)$, і т.д. В результаті будуть знайдені оцінки $\hat{\tau}_t$ при всіх t, крім $t = 1, 2, \ldots, m$ і $t = n, n-2, \ldots, n-m+1$.

Знайдемо коефіцієнти c_t для випадку лінійної функції тренду $\tau(t) = \alpha_0 + \alpha_1 t$. Не порушуючи загальності, введемо співвідношення $t' = t - (m+1)$, що дозволяє розглядати модель на новому часовому проміжку $t': -m, -m+1, \ldots, -1, 0, 1, \ldots, m-1, m$ ($t' = 0$ буде середньою точкою).

МНК-оцінки a_0 і a_1 знайдемо із системи нормальних рівнянь

$$\begin{cases} (2m+1)a_0 + \left(\sum\limits_{t'=-m}^{m} t'\right)a_1 = \sum\limits_{t'=-m}^{m} y_{t'}, \\ \left(\sum\limits_{t'=-m}^{m} t'\right)a_0 + \left(\sum\limits_{t'=-m}^{m} (t')^2\right)a_1 = \sum\limits_{t'=-m}^{m} t' y_{t'}. \end{cases}$$

Оскільки $\sum\limits_{t'=-m}^{m} t' = 0$ і згладжуване значення обчислюється в точці $t' = 0$, то

$$\hat{\tau}_{m+1} = a_0 + a_1 t'\,|_{t'=0} = a_0 = \frac{1}{2m+1}\sum_{t'=-m}^{m} y_{t'} = \frac{1}{2m+1}\sum_{t=1}^{2m+1} y_t\,.$$

Аналогічний результат можна отримати і для інших часових інтервалів. Тому у випадку лінійного наближення маємо оцінки

$$\hat{\tau}_t = \frac{1}{2m+1}\sum_{i=-m}^{m} y_{t+i}\,,\quad t = m+1, m+2, \ldots, n-m\,. \tag{7.12}$$

Якщо довжина часових інтервалів є парним числом $(2m)$, тоді ковзна середня знаходиться за формулою

$$\begin{aligned} \hat{\tau}_t = \frac{1}{2m}\Bigg(&\frac{1}{2}y_{t-m} + y_{t-m+1} + \ldots + y_{t-1} + \\ &y_t + y_{t+1} + \ldots + y_{t+m-1} + \frac{1}{2}y_{t+m}\Bigg), \qquad t = m+1, m+2, \ldots, n-m. \end{aligned} \tag{7.13}$$

Наведемо в таблиці значення вагових коефіцієнтів при різних m і p.

m	p	$c_{-m} = c_m$	$c_{-m+1} = c_{m+1}$	…	c_0
m_0	0 або 1	$\frac{1}{2m_0+1}$	$\frac{1}{2m_0+1}$	…	$\frac{1}{2m_0+1}$
5	2 або 3	$-\frac{3}{35}$	$\frac{12}{35}$		$\frac{17}{35}$
7	2 або 3	$-\frac{2}{21}$	$\frac{3}{21}$	$\frac{6}{21}$	$\frac{7}{21}$
9	2 або 3	$-\frac{21}{231}$	$\frac{14}{231}$	$\frac{39}{231}$ $\frac{54}{231}$	$\frac{59}{231}$
7	4 або 5	$\frac{5}{231}$	$-\frac{30}{231}$	$\frac{75}{231}$	$\frac{131}{231}$
9	4 або 5	$\frac{15}{429}$	$-\frac{55}{429}$	$\frac{30}{429}$ $\frac{135}{429}$	$\frac{179}{429}$

Розглянутий метод МКС передбачає застосування МНК, в якому усі статистичні дані мають однакову вагу, що не є правомірним в задачах прогнозування, де час має важливе значення. Врахувати часовий фактор при побудові оцінок моделі дозволяє **метод експоненціально зваженого ковзного середнього** (МЕЗКС).

Відповідно до цього метода оцінка згладженого значення $\hat{\tau}_t$ в точці t визначається із співвідношення

$$\hat{\tau}_t = \frac{1-\lambda}{1-\lambda^t} \sum_{i=0}^{t-1} \lambda^i y_{t-i}, \qquad (7.14)$$

де λ – коефіцієнт експоненціального згладжування ($0 < \lambda < 1$).

З (7.14) видно, що кожне спостереження y_{t-i} входить в оцінку $\hat{\tau}_t$ з вагою $\frac{1-\lambda}{1-\lambda^t}\lambda^i$, тобто по мірі віддалення «в минуле» від точки t вага спостереження y_{t-i} зменшується.

Можна довести справедливість рекурентного співвідношення

$$\hat{\tau}_t = \lambda\hat{\tau}_{t-i} + (1-\lambda)y_t, \quad t = \overline{1,n}. \qquad (7.15)$$

За початкове значення $\hat{\tau}_0$ можна вибрати середнє арифметичне всієї часової вибірки або лише її частини.

Співвідношення (7.15) перепишемо у вигляді

$$\hat{\tau}_t = y_t + \lambda(\hat{\tau}_{t-i} - y_t), \quad t = \overline{1,n}. \qquad (7.16)$$

З (7.16) видно, що значення $\hat{\tau}_t$ можна розглядати як прогнозне значення в момент t, яке складається з двох доданків: спостереженого значення y_t у даний момент часу і помилки прогнозування $\hat{\tau}_{t-i} - y_t$.

З виразу (7.14) видно, що зменшення λ веде до зростання згладжування. Рекомендується λ визначати за формулою

$$\lambda = \frac{n-1}{n+1}.$$

4. Для з'ясування наявності сезонних коливань можна використати коефіцієнт автокореляції, значення якого обчислюються по формулі (7.4). Тоді висновок про наявність або відсутність сезонних коливань робиться на основі корелограми. Якщо значення коефіцієнта автокореляції $r(l)$ міняються періодично, то цей період і буде періодом сезонних коливань.

Найпростіший спосіб моделювання часових рядів, які містять сезонні коливання, є побудова адитивної або мультиплікативної моделей часового ряду.

Вибір виду моделей ґрунтується на аналізі структури часового ряду. Якщо амплітуда сезонних коливань наближено стала, то будують адитивну модель

$$Y(t) = \tau(t) + s(t) + U, \qquad (7.17)$$

Якщо ж амплітуда коливань зростає або спадає, то будують мультиплікативну модель

$$Y(t) = \tau(t) \cdot s(t) \cdot U. \qquad (7.18)$$

Будемо вважати що в (7.17) і (7.18) циклічна складова $c(t)$ відсутня.

Процес побудови моделі часового ряду в цьому випадку включає наступні етапи:

1. Вирівнювання ряду методом ковзної середньої. Він передбачає сумування спостережень часового ряду послідовно по кварталах або місяцях із зміщенням на один момент часу. (для місячної динаміки це – сума 12 спостережень, для квартальної динаміки – сума 4 спостережень). Поділимо суми на кількість спостережень і знайдемо ковзні середні, які вже не містять сезонної компоненти. Ковзні середні також можна знаходити за формулами (7.9), (7.10). Щоб привести ці значення у відповідність з фактичними моментами часу, знаходимо середні значення з кожних двох сусідніх ковзних середніх (центровані ковзні середні). Оцінки сезонної компоненти знаходимо як різницю фактичного спостереження y_t та центрованого ковзного середнього для адитивної моделі та як частку від ділення фактичного спостереження y_t на центровані ковзні середні для мультиплікативної моделі. Далі за кожен місяць або квартал обчислюємо середню оцінку сезонної компоненти $\overline{S_i}$ ($i = \overline{1,12}$ або $i = \overline{1,4}$).

Сезонні впливи за період повинні взаємопогашатися. Для адитивної моделі це виражається в тому, що сума всіх сезонних компонент за всі періоди повинна дорівнювати нулю, а для мультиплікативної – числу періо-

дів в циклі. Корегуючі коефіцієнти для адитивної та мультиплікативної моделей знаходиться відповідно за формулами:

$$k = \frac{1}{l}\sum_{i=1}^{n} \overline{S}_i\,, \qquad k = l \Big/ \sum_{i=1}^{n} \overline{S}_i\,,$$

де l – довжина періоду.

Скореговані значення сезонної компоненти для адитивної та мультиплікативної моделей розраховуються відповідно за формулами:

$$S_i = \overline{S}_i - k\,, \qquad S_i = \overline{S}_i \cdot k\,,\ \ i = \overline{1,l}\,.$$

2. Виключення сезонної компоненти із початкового часового ряду для моделей (6.17) і (6.18) відповідно за формулами:

$$\tau_t = y_t - S_i\,, \qquad \tau_t = y_t / S_i\,,\ t = \overline{1,n}\,,\ i = \overline{1,l}\,.$$

3. Виділення лінійного тренду $\hat{\tau}(t) = a_0 + a_1 t$ за перетвореним рядом даних $y - S$ або y / S. Підставляючи в $\hat{\tau}(t)$ послідовно $t = 1, 2, \ldots, n$, отримаємо значення оцінок $\hat{\tau}_t$.

4. Обчислення оціночних значень часового ряду для моделей (7.17) і (7.18) відповідно за формулами:

$$\hat{y}_t = \hat{\tau}_t + S_i\,, \qquad \hat{y}_t = \hat{\tau}_t \cdot S_i\,, \qquad t = \overline{1,n}\,,\ i = \overline{1,l}\,.$$

Маючи значення сезонних компонент для різних періодів, можна здійснити прогноз в наступних періодах.

Задача. 7.2. *Побудувати модель часового ряду за даними задачі 7.1, попередньо виділивши сезонну компоненту.*

○ Подамо ряд графічно на рис. 7.1.

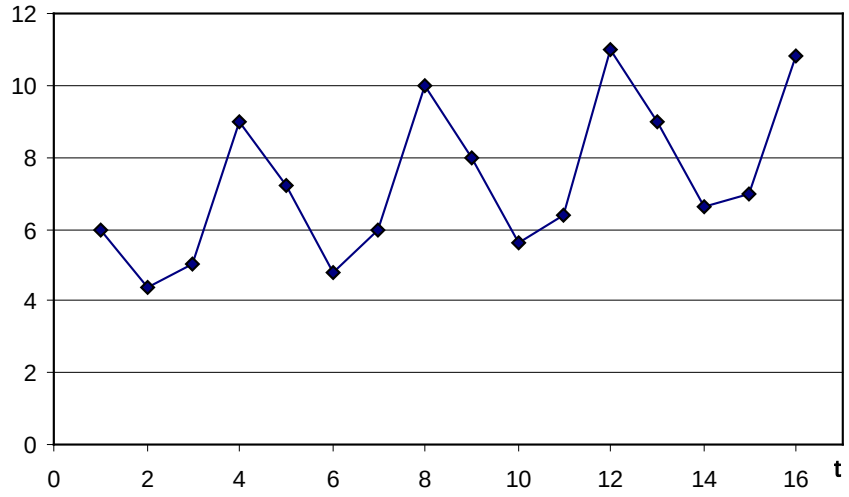


Рисунок 7.1.

Візуально видно, що амплітуда постійна, тому виберемо адитивну структуру часового ряду (7.17).

На першому етапі знайдемо ковзні середні:

$\overline{y}_{1-4} = (6 + 4{,}4 + 5 + 9)/4 = 6{,}1$; $\overline{y}_{2-5} = (4{,}4 + 5 + 9 + 7{,}2)/4 = 6{,}4$;

$\overline{y}_{36} = 6{,}5$, $\overline{y}_{47} = 6{,}75$; $\overline{y}_{58} = 7$; $\overline{y}_{69} = 7{,}2$; $\overline{y}_{7-10} = 7{,}4$; $\overline{y}_{8-11} = 7{,}5$;

$\bar{y}_{9–12} = 7{,}75$; $\bar{y}_{10–13} = 8$; $\bar{y}_{11–14} = 8{,}25$; $\bar{y}_{12–15} = 8{,}4$; $\bar{y}_{13–16} = 8{,}35$.

Далі знаходимо центровані ковзні середні:

$\bar{c}_3 = (6{,}1 + 6{,}4)/2 = 6{,}25$; $\bar{c}_4 = (6{,}4 + 6{,}5)/2 = 6{,}45$;

$\bar{c}_5 = 6{,}625$; $\bar{c}_6 = 6{,}875$; $\bar{c}_7 = 7{,}1$; $\bar{c}_8 = 7{,}3$; $\bar{c}_9 = 7{,}45$;

$\bar{c}_{10} = 7{,}625$; $\bar{c}_{11} = 7{,}875$; $\bar{c}_{12} = 8{,}125$; $\bar{c}_{13} = 8{,}325$; $\bar{c}_{14} = 8{,}375$.

Обчислимо оцінки сезонних компонент:

$s_3 = y_3 - \bar{c}_3 = 5 - 6{,}25 = -1{,}25$; $s_4 = y_4 - \bar{c}_4 = 9 - 6{,}45 = 2{,}55$;

$s_5 = 0{,}575$; $s_6 = -2{,}075$; $s_7 = -1{,}1$; $s_8 = 2{,}7$; $s_9 = 0{,}55$;

$s_{10} = -2{,}025$; $s_{11} = -1{,}475$; $s_{12} = 2{,}875$; $s_{13} = 0{,}675$; $s_{14} = -1{,}775$.

Результати обчислень подано в табл. 7.2.

Таблиця 7.2

Квартал, t	Об'єм продажів, y_t	Ковзні середні	Центровані ковзні середні	Оцінки сезонних компонент
1	6			
2	4,4	6,1		
3	5	6,4	6,25	– 1,25
4	9	6,5	6,45	2,55
5	7,2	6,75	6,625	0,575
6	4,8	7	6,875	– 2,075
7	6	7,2	7,1	– 1,1
8	10	7,4	7,3	2,7
9	8	7,5	7,45	0,55
10	5,6	7,75	7,625	– 2,025
11	6,4	8	7,875	– 1,475
12	11	8,25	8,125	2,875
13	9	8,4	8,325	0,675
14	6,6	8,35	8,375	– 1,775
15	7			
16	10,8			

Знайдемо середні квартальні оцінки сезонної компоненти:

1 квартал – $\bar{S}_1 = (0{,}575 + 0{,}55 + 0{,}675)/3 = 0{,}6$;

2 квартал – $\bar{S}_2 = (-2{,}075 - 2{,}025 - 1{,}775)/3 = -1{,}958$;

3 квартал – $\bar{S}_3 = (-1{,}25 - 1{,}1 - 1{,}475)/3 = -1{,}275$;

4 квартал – $\bar{S}_4 = (2{,}55 + 2{,}7 + 2{,}875)/3 = 2{,}708$.

Скорегований коефіцієнт: $k = (0{,}6 - 1{,}958 - 1{,}275 + 2{,}708)/4 = 0{,}01875$.

Скореговані значення сезонної компоненти:

$S_1 = \overline{S}_1 - k = 0{,}6 - 0{,}01875 = 0{,}58;$

$S_2 = \overline{S}_2 - k = -1{,}958 - 0{,}01875 = -1{,}98;$

$S_3 = \overline{S}_3 - k = -1{,}275 - 0{,}01875 = -1{,}29;$

$S_4 = \overline{S}_4 - k = 2{,}708 - 0{,}01875 = 2{,}69.$

На другому етапі виключимо сезонні компоненти із початкового часового ряду. Результати обчислень наведено в третьому стовпчику табл. 7.3.

Таблиця 7.3

t	y_t	S_i	$\tau_t = y_t - S_i$	$t\tau_t$	$\hat{\tau}_t$	$\hat{y}_t$
1	6	0,58	5,42	5,42	5,905	6,485
2	4,4	– 1,98	6,38	12,76	6,091	4,111
3	5	– 1,29	6,29	18,87	6,277	4,987
4	9	2,69	6,31	25,24	6,463	9,153
5	7,2	0,58	6,62	33,1	6,649	7,229
6	4,8	– 1,98	6,78	40,88	6,835	4,855
7	6	– 1,29	7,29	51,03	7,021	5,731
8	10	2,69	7,31	58,48	7,207	9,897
9	8	0,58	7,42	66,78	7,393	7,973
10	5,6	– 1,98	7,58	75,8	7,579	5,599
11	6,4	– 1,29	7,69	84,59	7,765	6,475
12	11	2,69	8,31	99,72	7,951	10,641
13	9	0,58	8,42	109,46	8,137	8,717
14	6,6	– 1,98	8,58	120,12	8,323	6,343
15	7	– 1,29	8,29	124,35	8,509	7,219
16	10,8	2,69	8,11	129,76	8,695	11,385
$\sum$	—	—	116,8	1056,16	—	—

На третьому етапі за перетвореним рядом даних $\tau_t = y_t - S_i$ знайдемо статистичні оцінки a_0 і a_1 лінійного тренду $\hat{\tau}(t) = a_0 + a_1 t$, які задовольняють системі нормальних рівнянь (7.8):

$$\begin{cases} a_0 + \bar{t}a_1 = \bar{\tau}, \\ \bar{t}a_0 + \overline{t^2}a_1 = \overline{t\tau}. \end{cases}$$

Згідно із формулами (6.10) маємо ($n = 16$):

$$\bar{t} = \frac{n+1}{2} = \frac{16+1}{2} = 8{,}5, \qquad \overline{t^2} = \frac{(n+1)(2n+1)}{6} = \frac{17 \cdot 33}{6} = 93{,}5.$$

Використовуючи нижній рядок табл. 7.3, отримаємо:

$$\bar{\tau} = \frac{1}{16}\sum_{t=1}^{16} \tau_t = 7{,}3, \qquad \overline{t\tau} = \frac{1}{16}\sum_{t=1}^{n} t\tau_t = 66{,}01, \quad \begin{cases} a_0 + 8{,}5a_1 = 7{,}3, \\ 8{,}5a_0 + \overline{93{,}5}a_1 = 66{,}01. \end{cases}$$

Єдиний розв'язок останньої системи згідно із формулами (7.9):

$$a_1 = \frac{\overline{t\tau} - \bar{t}\cdot\bar{\tau}}{\overline{t^2} - \left(\bar{t}\right)^2} = \frac{66{,}01 - 8{,}5\cdot 7{,}3}{93{,}5 - (8{,}5)^2} = 0{,}186, \quad a_0 = \bar{\tau} - a_1\bar{t} = 7{,}3 - 0{,}186\cdot 8{,}5 = 5{,}719.$$

Тоді рівняння тренду має такий вигляд:

$$\hat{\tau}(t) = 5{,}719 + 0{,}186t.$$

На четвертому етапі обчислимо оцінки значень початкового часового ряду за формулою $\hat{y}_t = \hat{\tau}_t + S_t$, де

$$S_t = \begin{cases} 0{,}58; & t = 1,5,9,13: \\ -1{,}98; & t = 2,6,10,14; \\ -1{,}29; & t = 3,7,11,15; \\ 2{,}69; & t = 4,8,12,16. \end{cases}$$

Результати обчислень $\hat{\tau}_t$ і $\hat{y}_t$ при $t = \overline{1{,}16}$ наведено в двох останніх стовпчиках табл. 7.3. ◎

§8. КОМП'ЮТЕРНА РЕАЛІЗАЦІЯ МЕТОДІВ ЕКОНОМЕТРИКИ

Система Excel, що є складовою пакета програм Microsoft Office, є найпоширенішою серед табличних процесорів. Вона має потужні можливості в обчисленні даних у вигляді таблиць, у виконанні бухгалтерських розрахунків, аналізі даних, статистичних обчисленнях. Вона також дає змогу ілюструвати розрахунки графіками та діаграмами.

Розглянемо відповідні можливості цієї системи при розв'язуванні основних типів задач економетрії.

Парна лінійна регресія

Задача 8.1. На основі статистичних даних фактора x і показника Y із задачі 2.1. потрібно:

1) знайти статистичні оцінки параметрів лінійного рівняння регресії;

2) точкову оцінку та довірчий інтервал дисперсії збурень із надійністю $\gamma = 0{,}9$*;*

3) для рівня значущості $\alpha = 0{,}05$ *перевірити значущість коефіцієнтів регресії* α_0 *та* α_1*;*

4) знайти довірчі інтервали коефіцієнтів регресії з надійністю $\gamma = 0{,}95$*;*

5) знайти вибіркові коефіцієнт детермінації, коефіцієнт кореляції, а також інші показники якості лінійної регресії (MAPE, MPE);

6) *знайти та побудувати довірчу зону функції регресії з надійністю* $\gamma = 0{,}95$*;*

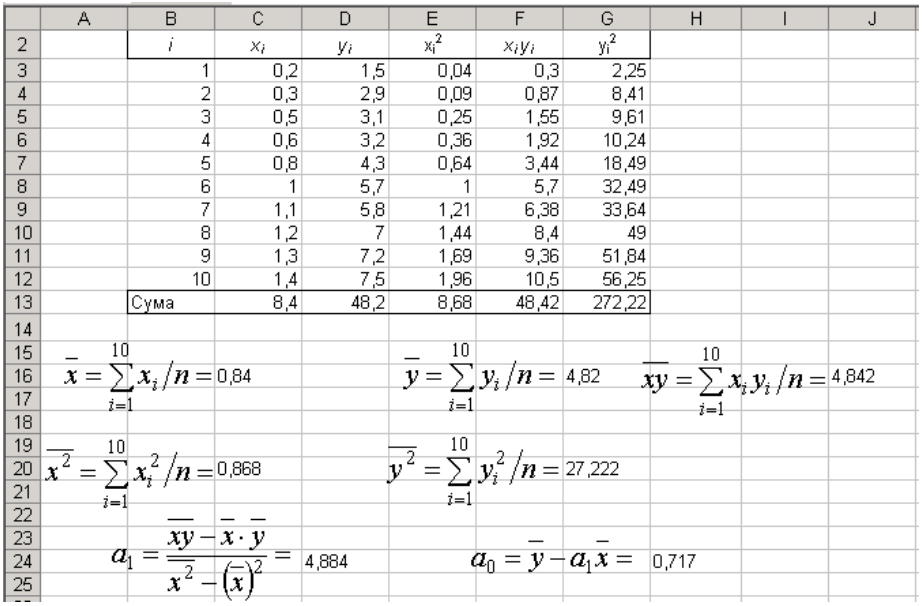
7) на рівні значущості $\alpha = 0{,}05$ *перевірити виконання передумови 2 за тестом Глейзера.*

1) Для роботи використовується пакет ***Excel.*** Блок вихідних даних формується, наприклад, в перших двох стовпцях (***B3:C12***). За блоком вихідних даних іде блок проміжних розрахунків (***D3:G12***).

Для знаходження добутку $x_1 \cdot y_1$ у комірку ***F3*** вводиться формула ***=C3*D3***. Далі копіюємо одержану формулу в інші комірки стовпця ***F***. Для ***копіювання формули*** необхідно: відмітити мишкою комірку ***F3***, натиснути праву клавішу мишки та вибрати з меню команду ***Копіювати.*** Потім відмічаємо блок копіювання (***F4:F12***): ставимо покажчик миші на комірку ***F4,*** натискуємо ліву клавішу миші і, тримаючи її, рухаємо покажчик до комірки ***F12*** включно. Можна скопіювати іншим способом: ставимо курсор на комірку ***F3***, а покажчик миші наводимо на маленький квадратик в правому

нижньому куті комірки. Після перетворення покажчика миші в хрестик натискуємо ліву клавішу миші і відмічаємо діапазон комірок копіювання до ***F12*** включно. Формули копіюються в помічений блок. Аналогічним чином обчислюється значення x_i^2 та y_i^2 $(i=\overline{1,10})$.

Умова і необхідні обчислення показані на наступних рисунках.



	i	x_i	y_i	x_i^2	x_iy_i	y_i^2
	1	0,2	1,5	0,04	0,3	2,25
	2	0,3	2,9	0,09	0,87	8,41
	3	0,5	3,1	0,25	1,55	9,61
	4	0,6	3,2	0,36	1,92	10,24
	5	0,8	4,3	0,64	3,44	18,49
	6	1	5,7	1	5,7	32,49
	7	1,1	5,8	1,21	6,38	33,64
	8	1,2	7	1,44	8,4	49
	9	1,3	7,2	1,69	9,36	51,84
	10	1,4	7,5	1,96	10,5	56,25
	Сума	8,4	48,2	8,68	48,42	272,22

$\bar{x}=\sum_{i=1}^{10}x_i/n=0{,}84$ $\quad \bar{y}=\sum_{i=1}^{10}y_i/n=4{,}82$ $\quad \overline{xy}=\sum_{i=1}^{10}x_iy_i/n=4{,}842$

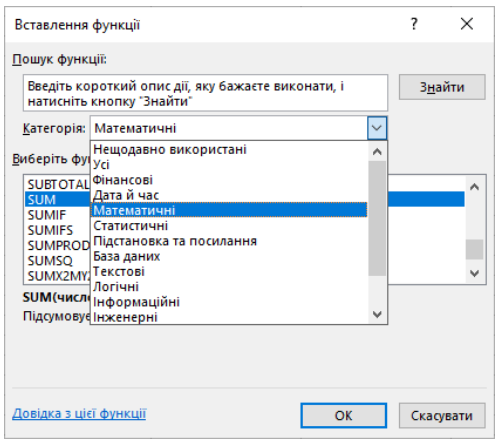
$\overline{x^2}=\sum_{i=1}^{10}x_i^2/n=0{,}868$ $\quad \overline{y^2}=\sum_{i=1}^{10}y_i^2/n=27{,}222$

$a_1=\dfrac{\overline{xy}-\bar{x}\cdot\bar{y}}{\overline{x^2}-(\bar{x})^2}=4{,}884$ $\quad a_0=\bar{y}-a_1\bar{x}=0{,}717$

В режимі формул даний рисунок має вигляд:

	A	B	C	D	E	F	G	H	I	J
1										
2		i	x_i	y_i	x_i^2	x_iy_i	y_i^2			
3		1	0,2	1,5	=C3^2	=C3*D3	=D3^2			
4		2	0,3	2,9	=C4^2	=C4*D4	=D4^2			
5		3	0,5	3,1	=C5^2	=C5*D5	=D5^2			
6		4	0,6	3,2	=C6^2	=C6*D6	=D6^2			
7		5	0,8	4,3	=C7^2	=C7*D7	=D7^2			
8		6	1	5,7	=C8^2	=C8*D8	=D8^2			
9		7	1,1	5,8	=C9^2	=C9*D9	=D9^2			
10		8	1,2	7	=C10^2	=C10*D10	=D10^2			
11		9	1,3	7,2	=C11^2	=C11*D11	=D11^2			
12		10	1,4	7,5	=C12^2	=C12*D12	=D12^2			
13		Сума	=SUM(C3:C12)	=SUM(D3:D12)	=SUM(E3:E12)	=SUM(F3:F12)	=SUM(G3:G12)			
14										
15		$\bar{x}=\sum_{i=1}^{10}x_i/n=$	=C13/B12			$\bar{y}=\sum_{i=1}^{10}y_i/n=$	=D13/B12	$\overline{xy}=\sum_{i=1}^{10}x_iy_i/n=$	=F13/B12	
16										
17		$\overline{x^2}=\sum_{i=1}^{10}x_i^2/n=$	=E13/B12			$\overline{y^2}=\sum_{i=1}^{10}y_i^2/n=$	=G13/B12			
18										
19		$a_1=\dfrac{\overline{xy}-\bar{x}\cdot\bar{y}}{\overline{x^2}-(\bar{x})^2}=$	=(J16-C16*G16)/(C20-C16^2)			$a_0=\bar{y}-a_1\bar{x}=$	=G16-D24*C16			
20										
21										
22										
23										
24										
25										

Для ***визначення сум*** стовпців використовуємо кнопку автосумування на панелі інструментів $\sum$ або вбудовану функцію ***SUM (діапазон комірок).*** Після встановлення курсору на комірку ***C13*** натискаємо клавіші ***Shift+F3*** або кнопку f_x на панелі інструментів. Відкривається вікно ***Вставлення функції*** процесора ***Excel***. У категорії активізуємо позицію *Математичні,* в функції — ***SUM***, і натискаємо клавішу ***Enter*** або ***OK***.



Діалогове вікно вбудованих функцій

У відкритому вікні відмічаємо діапазон комірок ***C3:C12*** і натискуємо клавішу ***Enter*** або ***OK.*** Введена формула копіюється в необхідні комірки ***13***-го рядка. Середні значення ***x***, ***y*** обчислюються в комірках ***C16***, ***G16.*** Ці ж значення можна обчислити з використанням вбудованої статистичної функції ***AVERAGE*** (***діапазон комірок***).

До комірок ***D24, H24*** вводяться формули для визначення оцінок параметрів відповідно a_1 і a_0.

2) Для знаходження точкової оцінки та довірчого інтервалу дисперсії збурень необхідно виконати наступні обчислення:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
34		i	x_i	y_i	x_i^2	y_i^2	$\hat{y}_i$	$y_i-\hat{y}_i$	$(y_i-\hat{y}_i)^2$	$\frac{\hat{y}_i-y_i}{y_i}$	$\left\|\frac{\hat{y}_i-y_i}{y_i}\right\|$	$S_{\hat{y}_i}$	$\hat{y}_i-t(\gamma,n-2)S_{\hat{y}_i}$	$\hat{y}_i+t(\gamma,n-2)S_{\hat{y}_i}$	
35		1	0,2	1,5	0,04	2,25	1,6938	-0,1938	0,0376	0,1292	0,1292	0,225444	1,1739	2,2137	
36		2	0,3	2,9	0,09	8,41	2,1822	0,7178	0,5152	-0,2475	0,2475	0,200848	1,7190	2,6454	
37		3	0,5	3,1	0,25	9,61	3,1590	-0,0590	0,0035	0,0190	0,0190	0,157167	2,7966	3,5214	
38		4	0,6	3,2	0,36	10,24	3,6474	-0,4474	0,2002	0,1398	0,1398	0,139814	3,3250	3,9698	
39		5	0,8	4,3	0,64	18,49	4,6242	-0,3242	0,1051	0,0754	0,0754	0,120715	4,3458	4,9026	
40		6	1	5,7	1	32,49	5,6010	0,0990	0,0098	-0,0174	0,0174	0,129247	5,3030	5,8990	
41		7	1,1	5,8	1,21	33,64	6,0894	-0,2894	0,0838	0,0499	0,0499	0,142957	5,7597	6,4191	
42		8	1,2	7	1,44	49	6,5778	0,4222	0,1783	-0,0603	0,0603	0,161076	6,2064	6,9492	
43		9	1,3	7,2	1,69	51,84	7,0662	0,1338	0,0179	-0,0186	0,0186	0,182296	6,6458	7,4866	
44		10	1,4	7,5	1,96	56,25	7,5546	-0,0546	0,003	0,0073	0,0073	0,205657	7,0804	8,0288	
45		Сума	8,4	48,2	8,68	272,2		0,0044	1,1544	0,0768	0,7644				
46															
47		$S_u^2=\frac{1}{n-2}\sum_{i=1}^{10}(y_i-\hat{y}_i)^2=$				0,144	$S_u=$	0,37987		$\sigma_x^2=\overline{x^2}-(\overline{x})^2=$		0,1624	$\sigma_y^2=\overline{y^2}-(\overline{y})^2=$		3,9896
48															
49		Ліва і права межі довірчого інтервалу для σ_u^2													
50		$\frac{(n-2)S_u^2}{\chi_2^2}=$		0,0744		$\frac{(n-2)S_u^2}{\chi_1^2}=$		0,42245		$S_{a_1}=\sqrt{\frac{S_u^2}{n\sigma_x^2}}=$		0,2981	$S_{a_0}=\sqrt{\frac{S_u^2\overline{x^2}}{n\sigma_x^2}}=$		0,2777
51															
52															
53															
54		$\left\|\frac{a_0}{S_{a_0}}\right\|=$		2,582	>	$t_{кр.}=$	2,306	і	$\left\|\frac{a_1}{S_{a_1}}\right\|=$		16,385	>	$t_{кр.}=$	2,3060	
55															
56															
57															
58		$a_0-t_0(\gamma,k)S_{a_0}<\alpha_0<a_0+t_0(\gamma,k)S_{a_0}$								$a_1-t_1(\gamma,k)S_{a_1}<\alpha_1<a_1+t_1(\gamma,k)S_{a_1}$					
59		0,08	<	α_0	<	1,3574				4,196615	<	α_1	<	5,5714	
60															
61										$\sum_{i=1}^{10}(y_i-\hat{y}_i)^2$					
62		Отже, значення коефійієнта детермінації						$R^2=1-$		$n\sigma_y^2$	$=$	0,9711	Коефіцієнт кореляції	$r=\pm\sqrt{R^2}=$	0,9854
63															
64															
65		$MAPE=\frac{1}{n}\sum_{i=1}^{10}\left\|\frac{\hat{y}_i-y_i}{y_i}\right\|\cdot 100\%=$				7,644	<10%			$MPE=\frac{1}{n}\sum_{i=1}^{10}\frac{\hat{y}_i-y_i}{y_i}\cdot 100\%=$				0,768	<5%
66															

У режимі формул дана таблиця має вигляд:

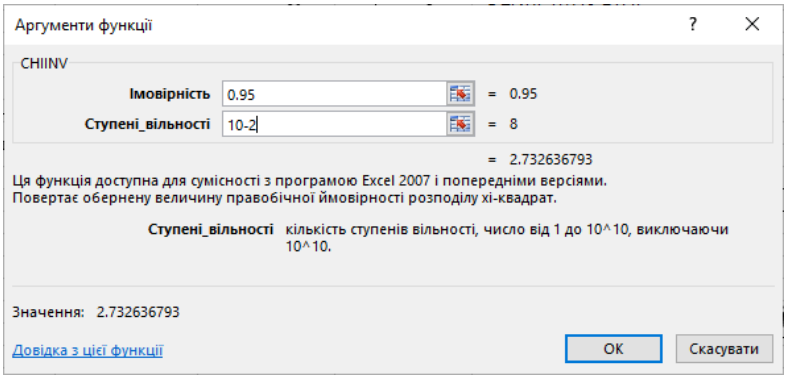
	A	B	C	D	E	F	G	H
33								
34		i	x_i	y_i	x_i^2	y_i^2	$\hat{y}_i$	$y_i - \hat{y}_i$
35		1	=C3	=D3	=C35^2	=D35^2	=$H$24+$D$24*C35	=D35-G35
36		2	=C4	=D4	=C36^2	=D36^2	=$H$24+$D$24*C36	=D36-G36
37		3	=C5	=D5	=C37^2	=D37^2	=$H$24+$D$24*C37	=D37-G37
38		4	=C6	=D6	=C38^2	=D38^2	=$H$24+$D$24*C38	=D38-G38
39		5	=C7	=D7	=C39^2	=D39^2	=$H$24+$D$24*C39	=D39-G39
40		6	=C8	=D8	=C40^2	=D40^2	=$H$24+$D$24*C40	=D40-G40
41		7	=C9	=D9	=C41^2	=D41^2	=$H$24+$D$24*C41	=D41-G41
42		8	=C10	=D10	=C42^2	=D42^2	=$H$24+$D$24*C42	=D42-G42
43		9	=C11	=D11	=C43^2	=D43^2	=$H$24+$D$24*C43	=D43-G43
44		10	=C12	=D12	=C44^2	=D44^2	=$H$24+$D$24*C44	=D44-G44
45		Сума	=SUM(C35:C44)	=SUM(D35:D44)	=SUM(E35:E44)			=SUM(F35:F44)
46								
47				$S_u^2 = \frac{1}{n-2}\sum_{i=1}^{10}(y_i - \hat{y}_i)^2 =$		=1/(10-2)*J45	$S_u =$	=SQRT(F47)
48								
49	Ліва і права межі довірчого інтервалу для σ_u^2							
50								
51				$\frac{(n-2)S_u^2}{\chi_2^2} =$	=(10-2)*F47/CHIINV(0.05,8)		$\frac{(n-2)S_u^2}{\chi_1^2} =$	=(10-2)*F47/CHIINV(0.95,8)
52								
53								
54		$\left\|\frac{a_0}{S_{a_0}}\right\| =$						$\left\|\frac{a_1}{S_{a_1}}\right\| =$
55			=H24/O50	>		=TINV(0.05,8)	i	
56								
57				$a_0 - t_0(\gamma,k)S_{a_0} < \alpha_0 < a_0 + t_0(\gamma,k)S_{a_0}$				
58								
59		=H24-F55*O50	<	α_0	<	=H24+F55*O50		
60								
61								
62	Отже, значення коефіцієнта детермінації							
63								
64								
65				$MAPE = \frac{1}{n}\sum_{i=1}^{10}\left\|\frac{\hat{y}_i - y_i}{y_i}\right\| \cdot 100\% =$		=1/10*L45*100	<10%
66								

Продовження розрахункової таблиці:

	I	J	K	L	M	N	O
33							
34	$(y_i - \hat{y}_i)^2$	$\frac{\hat{y}_i - y_i}{y_i}$	$\frac{\hat{y}_i - y_i}{y_i}$	$S_{\hat{y}_i}$	$\hat{y}_i - t(\gamma, n-2)S_{\hat{y}_i}$	$\hat{y}_i + t(\gamma, n-2)S_{\hat{y}_i}$	
35	=H35^2	=(G35-D35)/D35	=ABS(J35)	=$H$47*SQRT((1+(C35-0.84)^2/$L$47)/10)	=G35-$F$55*L35	=G35+$F$55*L35	
36	=H36^2	=(G36-D36)/D36	=ABS(J36)	=$H$47*SQRT((1+(C36-0.84)^2/$L$47)/10)	=G36-$F$55*L36	=G36+$F$55*L36	
37	=H37^2	=(G37-D37)/D37	=ABS(J37)	=$H$47*SQRT((1+(C37-0.84)^2/$L$47)/10)	=G37-$F$55*L37	=G37+$F$55*L37	
38	=H38^2	=(G38-D38)/D38	=ABS(J38)	=$H$47*SQRT((1+(C38-0.84)^2/$L$47)/10)	=G38-$F$55*L38	=G38+$F$55*L38	
39	=H39^2	=(G39-D39)/D39	=ABS(J39)	=$H$47*SQRT((1+(C39-0.84)^2/$L$47)/10)	=G39-$F$55*L39	=G39+$F$55*L39	
40	=H40^2	=(G40-D40)/D40	=ABS(J40)	=$H$47*SQRT((1+(C40-0.84)^2/$L$47)/10)	=G40-$F$55*L40	=G40+$F$55*L40	
41	=H41^2	=(G41-D41)/D41	=ABS(J41)	=$H$47*SQRT((1+(C41-0.84)^2/$L$47)/10)	=G41-$F$55*L41	=G41+$F$55*L41	
42	=H42^2	=(G42-D42)/D42	=ABS(J42)	=$H$47*SQRT((1+(C42-0.84)^2/$L$47)/10)	=G42-$F$55*L42	=G42+$F$55*L42	
43	=H43^2	=(G43-D43)/D43	=ABS(J43)	=$H$47*SQRT((1+(C43-0.84)^2/$L$47)/10)	=G43-$F$55*L43	=G43+$F$55*L43	
44	=H44^2	=(G44-D44)/D44	=ABS(J44)	=$H$47*SQRT((1+(C44-0.84)^2/$L$47)/10)	=G44-$F$55*L44	=G44+$F$55*L44	
45	=SUM(H35:H44)	=SUM(I35:I44)	=SUM(J35:J44)				
46							
47		$\sigma_x^2 = \overline{x^2} - (\bar{x})^2 =$		=E45/10-(C45/10)^2	$\sigma_y^2 = \overline{y^2} - (\bar{y})^2 =$	=H45/10-(D45/10)^2	
48							
49		$S_{a_1} = \sqrt{\frac{S_u^2}{n\sigma_x^2}} =$		=SQRT(F47/(10*L47))	$S_{a_0} = \sqrt{\frac{S_u^2 \overline{x^2}}{n\sigma_x^2}} =$	=SQRT(F47*C20/(10*L47))	
50							
51							
52							
53							
54							
55	=D24/L50	>	$t_{кр.} =$	=TINV(0.05,8)			
56							
57		$a_1 - t_1(\gamma, k)S_{a_1} < \alpha_1 < a_1 + t_1(\gamma, k)S_{a_1}$					
58							
59	=D24-F55*L50	<	α_1	<	=D24+F55*L50		
60							
61		$R^2 = 1 - \frac{\sum_{i=1}^{10}(y_i - \hat{y}_i)^2}{n\sigma_y^2} =$					
62			=1-J45/(10*O47)	Коефіцієнт кореляції		$r = \pm\sqrt{R^2} =$	=SQRT(K62)
63							
64							
65				$MPE = \frac{1}{n}\sum_{i=1}^{10}\frac{\hat{y}_i - y_i}{y_i} \cdot 100\% =$		=1/10*K45*100	<5%
66							
67							

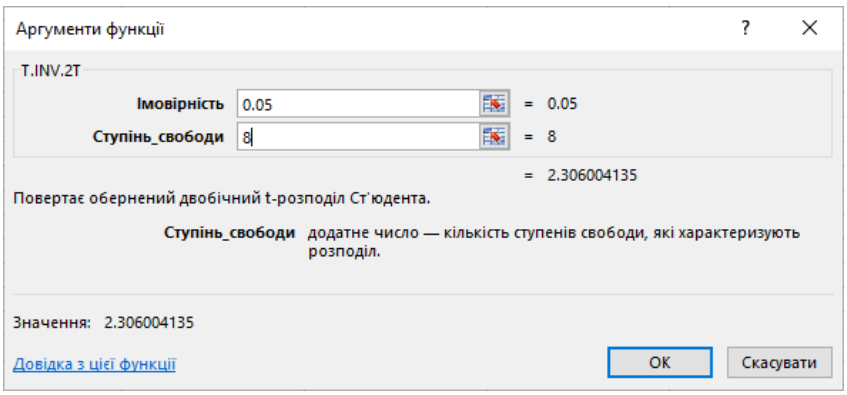
Для обчислення значення $\hat{y}_i$ $(i=\overline{1,10})$ у комірку ***G35*** вводимо формулу $a_0 + a_1 x_1$ (=***$H$24+$D$24*C35***) з абсолютним (не змінним, для цього використали знак ***$***) посиланнями координат-параметрів $\boldsymbol{a_0}$ та $\boldsymbol{a_1}$ і відносним посиланням координати $\boldsymbol{x_1}$ (а саме ***C35***). Одержану формулу у комірці ***G35*** копіюємо у блок ***G35:G44***.

Незміщену точкову оцінку S_u^2 невідомої дисперсії збурень σ_u^2 обчислено в комірці ***F47***, а її ліва і права межі наведено в комірках ***D51*** та ***H51***. При цьому величини χ_1^2 та χ_2^2 при значеннях $k=10-2=8$, $p=(1-0{,}9)/2=0{,}05$ і $p=(1+0{,}9)/2=0{,}95$ знаходимо за допомогою вбудованої статистичної функції ***XCHIINV***.



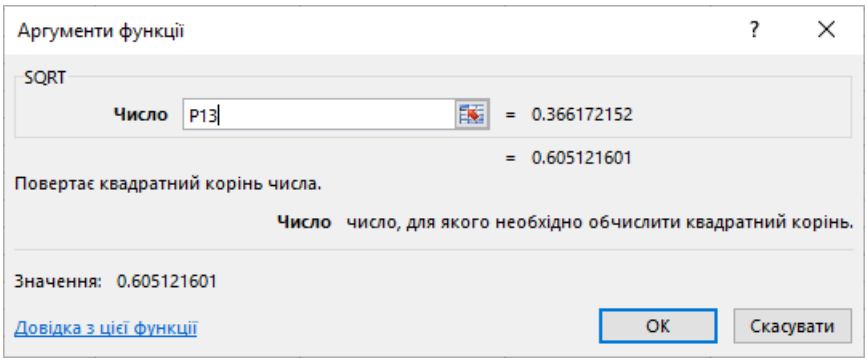
Діалогове вікно знаходження ***t****-статистики*

3) Для перевірки значущості коефіцієнтів регресії α_0 та α_1 скористаємося ***t***-статистикою. Значення S_{a_0} та S_{a_1} знайдено в комірках ***O50*** та ***L50*** відповідно, а спостережені значення критерію обчислені в комірках ***C55*** та ***I55***. В комірках ***F55*** та ***L55*** знайдено критичну точку для двосторонньої критичної області $t_{кр.}=t_{двост.}(\alpha,k)$ при значеннях $\alpha=0{,}05$, $k=10-2=8$ за допомогою вбудованої статистичної функції ***TINV***.



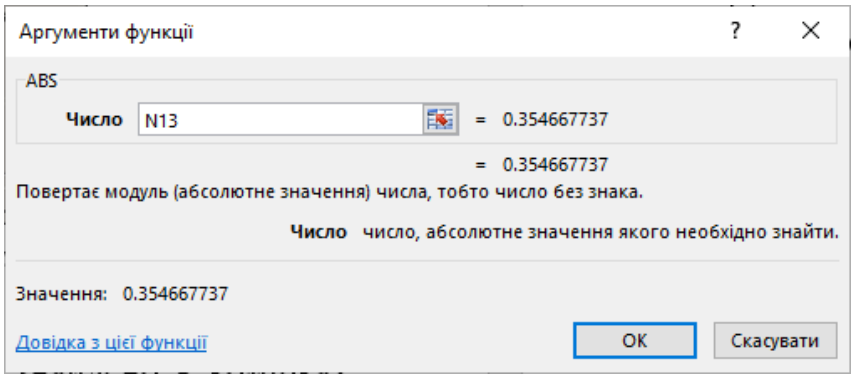
Діалогове вікно знаходження ***t****-статистики*

Для обчислення кореня квадратного в комірках ***H47***, ***L50***, ***O50*** використовують вбудовану математичну функцію ***SQRT***:



Діалогове вікно знаходження квадратного кореня з числа

Для обчислення абсолютної величини в діапазоні ***K35:K44*** використовують вбудовану функцію ***ABS***.



Діалогове вікно знаходження абсолютної величини з числа

4) Ліві та праві межі довірчих інтервалів з надійністю $\gamma = 0{,}95$ для невідомих параметрів регресії a_0 та a_1 знайдено у комірках ***B59***, ***F59***, ***I59*** та ***M59***.

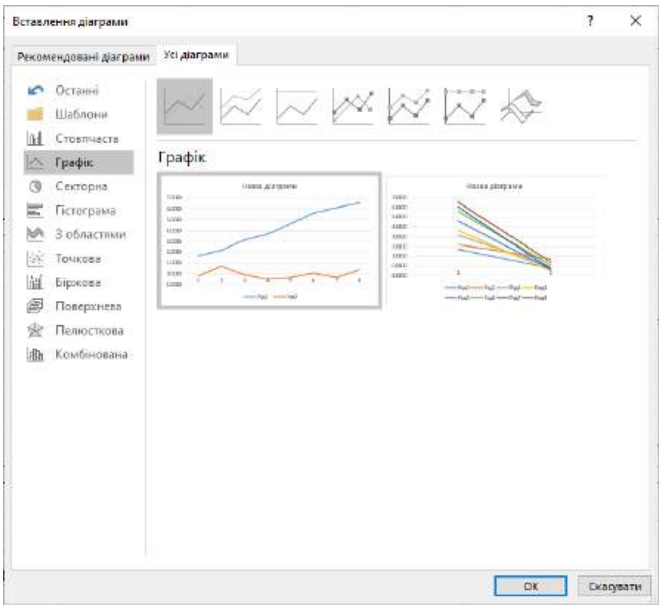
5) Значення вибіркових коефіцієнтів детермінації та кореляції наведено в комірках ***K62***, ***O62***. Нагадаємо, що $a_1 = 4{,}884 > 0$, то і $r > 0$. Показники якості лінійної регресії *MAPE* та *MPE* знайдено в комірках ***G65*** та ***N65***. Для цього спочатку в комірках ***J35*** та ***K35*** були набрані відповідні формули, які для наступних комірок були скопійовані вище описаним методом.

Для побудови довірчої зони для функції регресії спочатку в діапазонах ***L35:L44***, ***M35:M44*** та ***N35:N44*** знаходимо значення величин $S_{\hat{y}_i}$, $\hat{y}_i - t(\gamma, n-2)S_{\hat{y}_i}$, $\hat{y}_i + t(\gamma, n-2)S_{\hat{y}_i}$ $(i = \overline{1{,}10})$ відповідно.

6) Для наочного уявлення одержаних розрахунків ***будуємо графіки***: лінії регресії**,** довірчу зону $\boldsymbol{Y_{min}}$ та $\boldsymbol{Y_{max}}$.

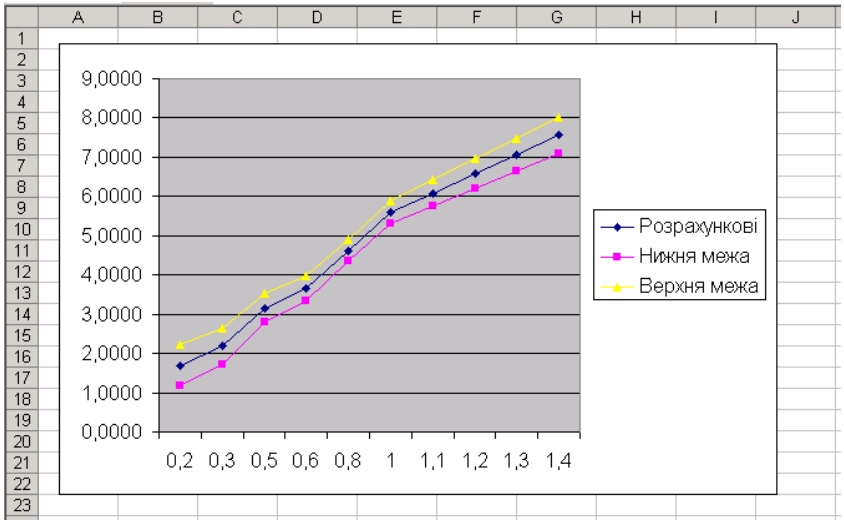
Порядок побудови графіків

1. Тримаючи натиснутою клавішу ***Ctrl***, відмічаємо лівою клавішею миші необхідні для побудови графіків блоки комірок із числовими даними: ***D35:D44; M35:M44; N35:N44***. При переході до іншого блоку комірок ліву клавішу миші відпускаємо.



Діалогове вікно типів діаграм

2. На панелі інструментів в меню ***Вставлення*** вибираємо закладку ***Діаграми*** і натискаємо ліву клавішу мишки. Відкривається діалогове вікно ***Вставлення діаграми.*** У відкритому вікні вибираємо тип діаграм ***Точкова*** або ***Графік***, також вибираємо його вид. Далі тиснемо ***ОК*** і на робочому аркуші з'являється графік.



Для редагування графіка (або його частин) необхідно навести на нього курсор і натиснути 2 рази на ліву клавішу мишки.

7) Для перевірити виконання передумови 2 на рівні значущості $\alpha = 0{,}05$ за тестом Гейзера потрібно виконати наступні обчислення:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1									$\delta=2$		$\delta=3$		$\delta=1/2$		$\delta=1/3$	
2	i	x_i	y_i	x_i^2	$\hat{y}_i$	$y_i-\hat{y}_i$	$\lvert e_i\rvert$	$x_i\lvert e_i\rvert$	$\lvert\hat{e}_i\rvert$	$(\lvert e_i\rvert-\lvert\hat{e}_i\rvert)^2$	$\lvert\hat{e}_i\rvert$	$(\lvert e_i\rvert-\lvert\hat{e}_i\rvert)^2$	$\lvert\hat{e}_i\rvert$	$(\lvert e_i\rvert-\lvert\hat{e}_i\rvert)^2$	$\lvert\hat{e}_i\rvert$	$(\lvert e_i\rvert-\lvert\hat{e}_i\rvert)^2$
3	1	0,2	1,5	0,04	1,6938	-0,1938	0,1938	0,0388	0,4302	0,0559	0,4364	0,0589	0,3508	0,0246	0,3239	0,0169
4	2	0,3	2,9	0,09	2,1822	0,7178	0,7178	0,2153	0,4204	0,0884	0,4327	0,0813	0,3311	0,1495	0,3074	0,1684
5	3	0,5	3,1	0,25	3,1590	-0,0590	0,059	0,0295	0,3892	0,1091	0,4136	0,1258	0,3000	0,0581	0,2832	0,0502
6	4	0,6	3,2	0,36	3,6474	-0,4474	0,4474	0,2684	0,3678	0,0063	0,3959	0,0027	0,2869	0,0258	0,2735	0,0303
7	5	0,8	4,3	0,64	4,6242	-0,3242	0,3242	0,2594	0,3131	0,0001	0,3381	0,0002	0,2635	0,0037	0,2569	0,0045
8	6	1	5,7	1	5,6010	0,0990	0,099	0,099	0,2429	0,0207	0,2429	0,0207	0,2429	0,0207	0,2429	0,0207
9	7	1,1	5,8	1,21	6,0894	-0,2894	0,2894	0,3183	0,2019	0,0077	0,1783	0,0123	0,2334	0,0031	0,2366	0,0028
10	8	1,2	7	1,44	6,5778	0,4222	0,4222	0,5066	0,1571	0,0703	0,1009	0,1033	0,2243	0,0392	0,2307	0,0367
11	9	1,3	7,2	1,69	7,0662	0,1338	0,1338	0,1739	0,1083	0,0007	0,0094	0,0155	0,2156	0,0067	0,2251	0,0083
12	10	1,4	7,5	1,96	7,5546	-0,0546	0,0546	0,0764	0,0556	0,0000	-0,0974	0,0231	0,2072	0,0233	0,2197	0,0273
13	Сума	8,4	48,2	8,68		0,0044	2,7412	1,9858		0,3591		0,4436		0,3547		0,3662
14																
15	$b_1=\dfrac{\overline{x\lvert e\rvert}-\overline{x}\,\overline{\lvert e\rvert}}{\overline{x^2}-(\overline{x})^2}=$		-0,1951		$b_0=\overline{\lvert e\rvert}-b_1\overline{x}=$		0,4380		$\sigma_x^2=\overline{x^2}-(\overline{x})^2=$		0,1624					
16																
17																
18	$\delta=2$	$S_\varepsilon^2=\dfrac{1}{n-2}\sum_{i=1}^{n}(\lvert e_i\rvert-\lvert\hat{e}_i\rvert)^2=$				0,0449	$S_{b_1}=\sqrt{\dfrac{S_\varepsilon^2}{n\sigma_x^2}}=$		0,1663	$\left\lvert\dfrac{b_1}{S_{b_1}}\right\rvert=$	1,1735	$<t_{\text{двостор}}(0{,}05;8)=$	2,3060			
19																
20																
21	$\delta=3$	$S_\varepsilon^2=\dfrac{1}{n-2}\sum_{i=1}^{n}(\lvert e_i\rvert-\lvert\hat{e}_i\rvert)^2=$				0,0555	$S_{b_1}=\sqrt{\dfrac{S_\varepsilon^2}{n\sigma_x^2}}=$		0,1848	$\left\lvert\dfrac{b_1}{S_{b_1}}\right\rvert=$	1,0558	$<t_{\text{двостор}}(0{,}05;8)=$	2,3060			
22																
23																
24	$\delta=0{,}5$	$S_\varepsilon^2=\dfrac{1}{n-2}\sum_{i=1}^{n}(\lvert e_i\rvert-\lvert\hat{e}_i\rvert)^2=$				0,0443	$S_{b_1}=\sqrt{\dfrac{S_\varepsilon^2}{n\sigma_x^2}}=$		0,1652	$\left\lvert\dfrac{b_1}{S_{b_1}}\right\rvert=$	1,1808	$<t_{\text{двостор}}(0{,}05;8)=$	2,3060			
25																
26																
27	$\delta=0{,}3333$	$S_\varepsilon^2=\dfrac{1}{n-2}\sum_{i=1}^{n}(\lvert e_i\rvert-\lvert\hat{e}_i\rvert)^2=$				0,0458	$S_{b_1}=\sqrt{\dfrac{S_\varepsilon^2}{n\sigma_x^2}}=$		0,1679	$\left\lvert\dfrac{b_1}{S_{b_1}}\right\rvert=$	1,1621	$<t_{\text{двостор}}(0{,}05;8)=$	2,3060			
28																

В режимі формул дана таблиця має вигляд:

	A	B	C	D	E	F	G	H	I
1									$\delta = 2$
2	i	x_i	y_i	x_i^2	$\hat{y}_i$	$y_i - \hat{y}_i$	$\lvert u_i \rvert$	$x_i \lvert u_i \rvert$	$\lvert \hat{u}_i \rvert$
3	1	0.2	1.5	=B3^2	=0.717+4.884*B3	=C3-E3	=ABS(F3)	=B3*ABS(F3)	=$I$15+$C$15*B3^$B$18
4	2	0.3	2.9	=B4^2	=0.717+4.884*B4	=C4-E4	=ABS(F4)	=B4*ABS(F4)	=$I$15+$C$15*B4^$B$18
5	3	0.5	3.1	=B5^2	=0.717+4.884*B5	=C5-E5	=ABS(F5)	=B5*ABS(F5)	=$I$15+$C$15*B5^$B$18
6	4	0.6	3.2	=B6^2	=0.717+4.884*B6	=C6-E6	=ABS(F6)	=B6*ABS(F6)	=$I$15+$C$15*B6^$B$18
7	5	0.8	4.3	=B7^2	=0.717+4.884*B7	=C7-E7	=ABS(F7)	=B7*ABS(F7)	=$I$15+$C$15*B7^$B$18
8	6	1	5.7	=B8^2	=0.717+4.884*B8	=C8-E8	=ABS(F8)	=B8*ABS(F8)	=$I$15+$C$15*B8^$B$18
9	7	1.1	5.8	=B9^2	=0.717+4.884*B9	=C9-E9	=ABS(F9)	=B9*ABS(F9)	=$I$15+$C$15*B9^$B$18
10	8	1.2	7	=B10^2	=0.717+4.884*B10	=C10-E10	=ABS(F10)	=B10*ABS(F10)	=$I$15+$C$15*B10^$B$18
11	9	1.3	7.2	=B11^2	=0.717+4.884*B11	=C11-E11	=ABS(F11)	=B11*ABS(F11)	=$I$15+$C$15*B11^$B$18
12	10	1.4	7.5	=B12^2	=0.717+4.884*B12	=C12-E12	=ABS(F12)	=B12*ABS(F12)	=$I$15+$C$15*B12^$B$18
13	Сума	=SUM(B3:	=SUM(C3:C12)	=SUM(D3:D12)		=SUM(F3:F12)	=SUM(G3:G12)	=SUM(H3:H12)	=S
14									
15	$b_1 = \frac{\overline{x\lvert u \rvert} - \bar{x}\,\overline{\lvert u \rvert}}{\overline{x^2} - (\bar{x})^2}$ =	=(H13/10-B13/10*G13/10)/M15						$b_0 = \overline{\lvert u \rvert} - b_1\bar{x}$ =	=G13/10-C15*B13/10
16									
17								$S_{b_1} = \sqrt{\frac{S_\varepsilon^2}{n\sigma_x^2}}$ =	
18	$\delta = 2$			$S_\varepsilon^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(\lvert u_i \rvert - \lvert \hat{u}_i \rvert\right)^2$	=J13/(10-2)			$S_{b_1} = \sqrt{\frac{S_\varepsilon^2}{n\sigma_x^2}}$ =	=SQRT(F18/(10*M15))
19									
20								$S_{b_1} = \sqrt{\frac{S_\varepsilon^2}{n\sigma_x^2}}$ =	
21	$\delta = 3$			$S_\varepsilon^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(\lvert u_i \rvert - \lvert \hat{u}_i \rvert\right)^2$	=L13/(10-2)			$S_{b_1} = \sqrt{\frac{S_\varepsilon^2}{n\sigma_x^2}}$ =	=SQRT(F21/(10*M15))
22									
23								$S_{b_1} = \sqrt{\frac{S_\varepsilon^2}{n\sigma_x^2}}$ =	
24	$\delta = 0.5$			$S_\varepsilon^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(\lvert u_i \rvert - \lvert \hat{u}_i \rvert\right)^2$	=N13/(10-2)			$S_{b_1} = \sqrt{\frac{S_\varepsilon^2}{n\sigma_x^2}}$ =	=SQRT(F24/(10*M15))
25									
26								$S_{b_1} = \sqrt{\frac{S_\varepsilon^2}{n\sigma_x^2}}$ =	
27	$\delta =$	=1/3		$S_\varepsilon^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(\lvert u_i \rvert - \lvert \hat{u}_i \rvert\right)^2$	=P13/(10-2)			$S_{b_1} = \sqrt{\frac{S_\varepsilon^2}{n\sigma_x^2}}$ =	=SQRT(F27/(10*M15))
28									

Продовження розрахункової таблиці:

	J	K	L	M	N	O	P
1	2	$\delta = 3$		$\delta = 1/2$		$\delta = 1/3$	
2	$(\lvert u_i\rvert - \lvert\hat{u}_i\rvert)^2$	$\lvert\hat{u}_i\rvert$	$(\lvert u_i\rvert - \lvert\hat{u}_i\rvert)^2$	$\lvert\hat{u}_i\rvert$	$(\lvert u_i\rvert - \lvert\hat{u}_i\rvert)^2$	$\lvert\hat{u}_i\rvert$	$(\lvert u_i\rvert - \lvert\hat{u}_i\rvert)^2$
3	=(G3-I3)^2	=$I$15+$C$15*B3^$B$21	=(G3-K3)^2	=$I$15+$C$15*B3^$B$24	=(G3-M3)^2	=$I$15+$C$15*B3^$B$27	=(G3-O3)^2
4	=(G4-I4)^2	=$I$15+$C$15*B4^$B$21	=(G4-K4)^2	=$I$15+$C$15*B4^$B$24	=(G4-M4)^2	=$I$15+$C$15*B4^$B$27	=(G4-O4)^2
5	=(G5-I5)^2	=$I$15+$C$15*B5^$B$21	=(G5-K5)^2	=$I$15+$C$15*B5^$B$24	=(G5-M5)^2	=$I$15+$C$15*B5^$B$27	=(G5-O5)^2
6	=(G6-I6)^2	=$I$15+$C$15*B6^$B$21	=(G6-K6)^2	=$I$15+$C$15*B6^$B$24	=(G6-M6)^2	=$I$15+$C$15*B6^$B$27	=(G6-O6)^2
7	=(G7-I7)^2	=$I$15+$C$15*B7^$B$21	=(G7-K7)^2	=$I$15+$C$15*B7^$B$24	=(G7-M7)^2	=$I$15+$C$15*B7^$B$27	=(G7-O7)^2
8	=(G8-I8)^2	=$I$15+$C$15*B8^$B$21	=(G8-K8)^2	=$I$15+$C$15*B8^$B$24	=(G8-M8)^2	=$I$15+$C$15*B8^$B$27	=(G8-O8)^2
9	=(G9-I9)^2	=$I$15+$C$15*B9^$B$21	=(G9-K9)^2	=$I$15+$C$15*B9^$B$24	=(G9-M9)^2	=$I$15+$C$15*B9^$B$27	=(G9-O9)^2
10	=(G10-I10)^2	=$I$15+$C$15*B10^$B$21	=(G10-K10)^2	=$I$15+$C$15*B10^$B$24	=(G10-M10)^2	=$I$15+$C$15*B10^$B$27	=(G10-O10)^2
11	=(G11-I11)^2	=$I$15+$C$15*B11^$B$21	=(G11-K11)^2	=$I$15+$C$15*B11^$B$24	=(G11-M11)^2	=$I$15+$C$15*B11^$B$27	=(G11-O11)^2
12	=(G12-I12)^2	=$I$15+$C$15*B12^$B$21	=(G12-K12)^2	=$I$15+$C$15*B12^$B$24	=(G12-M12)^2	=$I$15+$C$15*B12^$B$27	=(G12-O12)^2
13	=SUM(J3:J12)		=SUM(L3:L12)		=SUM(N3:N12)		=SUM(P3:P12)
14							
15	$\sigma_x^2 = \overline{x^2} - (\overline{x})^2$ = =D13/10-(B13/10)^2						
16							
17	$\left\lvert \frac{b_1}{S_{b_1}} \right\rvert$ = =ABS(C15/I18)		$< t_{\text{додатк}}(0,05;8)$ =	=TINV(0.05,10-2)			
18							
19							
20	$\left\lvert \frac{b_1}{S_{b_1}} \right\rvert$ = =ABS(C15/I21)		$< t_{\text{додатк}}(0,05;8)$ =	=TINV(0.05,10-2)			
21							
22							
23	$\left\lvert \frac{b_1}{S_{b_1}} \right\rvert$ = =ABS(C15/I24)		$< t_{\text{додатк}}(0,05;8)$ =	=TINV(0.05,10-2)			
24							
25							
26	$\left\lvert \frac{b_1}{S_{b_1}} \right\rvert$ = =ABS(C15/I27)		$< t_{\text{додатк}}(0,05;8)$ =	=TINV(0.05,10-2)			
27							
28							

В комірках ***C15***, ***G1***5 знайдено МНК-оцінки b_1 та b_0. Точкову оцінку S_ε^2 невідомої дисперсії збурень σ_ε^2 знайдено в комірці ***F18***, попередньо обчисливши величини $\hat{u}_i$ $(i=\overline{1,10},\ \delta=2)$ в діапазоні ***I3:I12.*** Для цього у комірку ***I3*** вводимо формулу $b_0+b_1x_1^\delta$ ***(=$G$15+$C$15*B3^$B$18)*** з абсолютним посиланнями координат-параметрів b_0, b_1 та δ і відносним посиланням координати x_1 (а саме ***B3***). Одержану формулу у комірці ***I3*** копіюємо у блок ***I3:I12***. Для визначення значущості β_1 порівнюємо знайдені в комірках ***K18*** та ***O18*** величину $\left|\frac{b_1}{S_{b_1}}\right|$ та ***t***-статистику відповідно.

За аналогією проводимо обчислення для значень $\delta=3$, $\delta=1/2$, $\delta=1/3$. Виявляється, що у всіх випадках коефіцієнт регресії β_1 незначущий, тобто для статистичних даних задачі передумова 2 виконується.

Матрична форма МНК для оцінки параметрів множинної регресії

Допустимо, що між показником y і факторами x_1, x_2 існує лінійна залежність $\hat{y}=X'a=a_0+a_1x_1+a_2x_2$, де $X'=(1,x_1,x_2)$, $a'=(a_0,a_1,a_2)$.

Оцінки параметрів вектора a шукатимемо за формулою $a=(X'X)^{-1}X'Y$.

Порядок знаходження оцінок параметрів регресії:

1. Знаходимо транспоновану матрицю X' в блоці по відношенню до матриці X в блоці, використовуючи в категорії ***Підстановка та посилання*** вбудовану функцію ***TRANSPOSE***.

2. Знаходимо добуток матриць $X'X$ в виділеному блоці, використовуючи вбудовану математичну функцію ***MMULT*** (блок даних першої матриці; блок даних другої матриці).

3. Обернену матрицю $(X'X)^{-1}$ знаходимо в іншому виділеному блоці, використовуючи вбудовану математичну функцію ***MINVERSE***.

4. Добуток матриць $X'Y$ знаходимо, використовуючи вбудовану математичну функцію ***MMULT***, виділивши перед тим масив, в якому буде знайдений добуток матриць.

5. Оцінки вектора знаходимо в виділеному для цього блоці, використовуючи вбудовану математичну функцію ***MMULT*** (блок даних матриці $(X'X)^{-1}$; блок даних матриці $X'Y$).

6. Для перевірки значущості параметрів регресії розрахуємо *t*-статистику кожного із параметрів за формулою

$$|t_{спост.}| = \frac{|a_i|}{S_{a_i}} > t_{кр.},$$

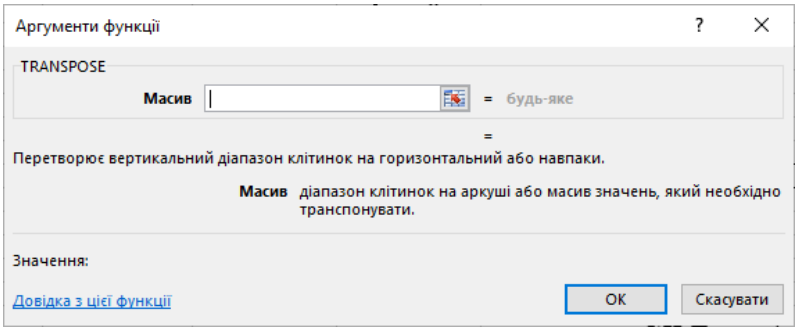
де $S_{a_i} = S_e\sqrt{[(X'X)^{-1}]_{ii}}$, $i = \overline{0,2}$, S_e — середньоквадратичне відхилення статистичних даних від розрахункових.

Порядок знаходження транспонованої матриці в ET Excel

1. Відмічаємо діапазон комірок, де має знаходитись транспонована матриця.

2. Відкриваємо діалогове вікно ***Вставлення функції***, вибираємо функцію ***TRANSPOSE*** у полі категорій ***Підстановка та посилання*** і натискуємо на кнопку ***ОК*** для переходу в наступне діалогове вікно.

3. У другому діалоговому вікні відмічаємо діапазон комірок, у яких знаходяться елементи вихідної матриці і натискуємо клавішу ***ОК***.



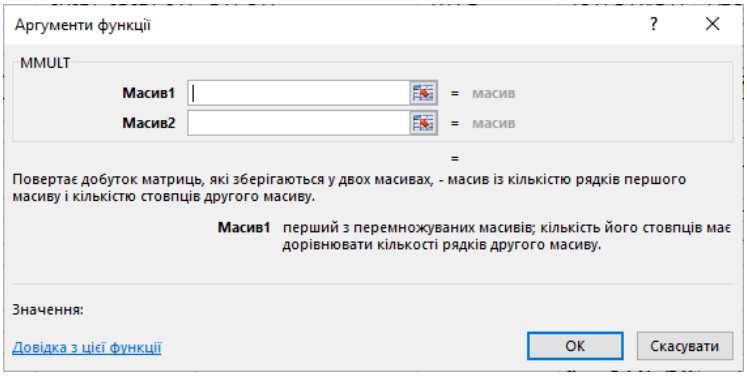
4. Натискуємо клавішу ***F2***, а потім одночасно клавіші ***Ctrl+Shift+Enter.***

Порядок знаходження добутку двох матриць в ET Excel

1. Відмічаємо діапазон комірок, де має знаходитись матриця, яка є результатом добутку двох матриць. розмір цієї матриці визначається розмірами матриць, які перемножуються. Якщо, наприклад, розмір матриці X' – 3×15, розмір матриці X – 15×3, тоді розмір матриці $X'X$ буде 3×3. (якщо $C = AB$, де розмір матриць: A – $(m\times k)$, B – $(k\times n)$, то розмір матриці C – $(m\times n)$).

2. Відкриваємо діалогове вікно ***Вставлення функції*** (натискаємо на ***f_x***), вибираємо функцію ***MMULT*** у полі категорії ***Математичні*** і натискаємо на кнопку ***ОК*** для переходу у наступне діалогове вікно.

3. У діалоговому вікні ***MMULT*** відмічаємо: у першому рядку діапазон комірок, в яких знаходяться елементи першої матриці, у другому рядку діапазон комірок, в яких знаходяться елементи другої матриці. Натискаємо клавішу ***ОК***.



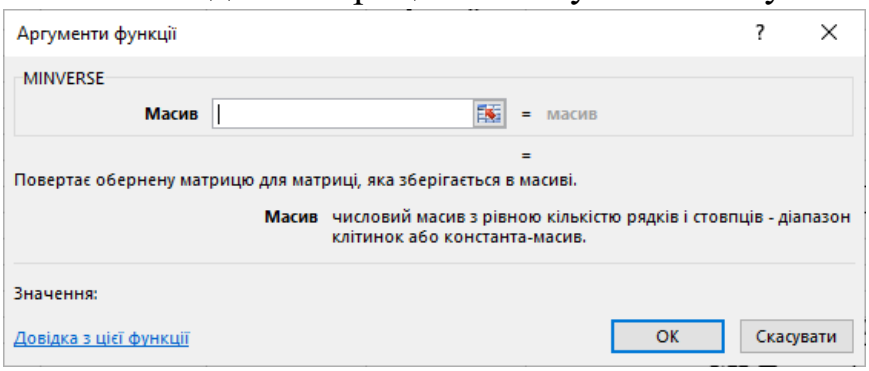
4. Натискаємо клавішу F_2 і одночасно клавіші ***Ctrl+Shift+Enter***.

Порядок знаходження оберненої матриці в ET Excel

1. Відмічаємо діапазон комірок, де має знаходитись обернена матриця.

2. Відкриваємо діалогове вікно ***Вставлення функції*** (натискуємо на кнопку f_x на панелі інструментів) Вибираємо функцію ***MINVERSE*** у полі категорій ***Математичні*** і натискаємо на кнопку *ОК* для переходу у діалогове вікно ***MINVERSE***.

3. У діалоговому вікно ***MINVERSE*** вказуємо діапазон комірок, у яких знаходяться елементи вихідної матриці. Натискуємо кнопку *ОК*.



4. Натискаємо клавішу F_2 і одночасно клавіші ***Ctrl+Shift+Enter***.

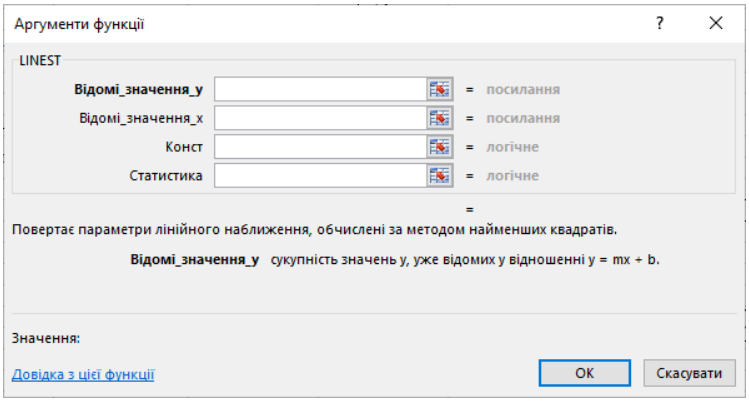
Оцінки параметрів регресії можна знайти, використовуючи вбудовану статистичну функцію *LINEST*.

Опишемо порядок знаходження оцінок параметрів регресії з використанням функції ***LINEST***:

1. Відмічаємо блок, де мають знаходитись розрахункові дані: ширина блоку дорівнює числу оцінюваних параметрів, а висота дорівнює п’яти рядкам.

2. Відкриваємо діалогове вікно ***Вставлення функції***, вибираємо функцію ***LINEST*** у полі категорії ***СТАТИСТИЧНІ*** і натискаємо на кнопку ***OK.***

3. У другому діалоговому вікні вводимо: в перший рядок (в перше поле) блок даних результативного показника, вказуючи діапазон комірок або ім'я блоку даних; у другий рядок — блок даних факторів або ім'я цього блоку; в третій рядок вводиться слово ***TRUE***, якщо a_0 не дорівнює нулю, і слово ***FALSE***, якщо a_0 дорівнює нулю; в четвертий рядок вводиться слово ***TRUE***, якщо необхідно знайти не лише параметри лінії регресії, а й додаткову регресійну статистику. Якщо необхідно знайти лише параметри лінії регресії, то вводимо слово ***FALSE*** і натискаємо на кнопку ***OK*** для отримання розрахункових даних.



4. Для того щоб у блоці розрахункових даних було видно не лише значення першої комірки, натискуємо клавішу ***F2,*** потім ***Ctrl+Shift+Enter***.

Таблиця розрахункових значень додаткової регресійної статистики має вигляд:

Розміщення значень додаткової регресійної статистики

a_2	a_1	a_0
σ_{a_2}	σ_{a_1}	σ_{a_0}
R^2	S	#Н/Д
F_{r_1}	k	#Н/Д
$\sum(\hat{y}-\bar{y})^2$	$\sum(\hat{y}_i-y_i)^2$	#Н/Д

Опишемо розрахункові дані:

У **першому** рядку справа наліво знаходяться оцінки параметрів множинної лінійної регресії відповідно a_0, a_1, a_2.

У **другому** рядку справа наліво знаходяться середні квадратичні відхилення оцінок параметрів $\sigma_{a_0}, \sigma_{a_1}, \sigma_{a_2}$.

У **третьому** рядку в першій комірці знаходиться коефіцієнт детермінації, а в другій комірці — середнє квадратичне відхилення показника.

У **четвертому** рядку в першій комірці знаходиться розрахункове значення *F*-статистики, в другій комірці знаходиться *k* — число ступенів вільності.

У **п'ятому** рядку в першій комірці знаходиться сума квадратів відхилень розрахункових значень показника від його середнього значення, в другій комірці — залишкова сума квадратів.

Нелінійна регресія

Задача 8.2. *На основі статистичних даних факторів K і L та показника Y із задачі 5.2 потрібно побудувати: 1) виробничу функцію Кобба-Дугласа виду* $\hat{Y} = a_0 K^{a_1} L^{a_2}$; *2) виробничу функцію Кобба-Дугласа при* $a_0 = 1$. *При побудові використати лише вбудовані функції Excel.*

○ Для побудови виробничої функції використаємо перетворені вхідні дані:

№	Y^*	K^*	L^*
1	13,6058	10,3281	7,7842
2	13,7852	10,6629	7,7522
3	13,7901	10,6697	7,7500
4	13,7146	10,7416	7,7300
5	13,7909	10,6566	7,6119
6	14,0029	10,7411	7,5171
7	14,1931	11,0385	7,4909
8	14,4061	11,2103	7,4815
9	14,5385	11,5216	7,4879
10	14,5777	11,5700	7,5139
11	14,5369	11,3434	7,4600

Для побудови лінійної регресійної моделі скористаємося функцією ***LINEST***, а також покажемо використання пакету аналізу Дані – Аналіз даних – Регресія.

Результат побудови лінійної регресійної моделі за функцією ***LINEST*** в Excel показано на рисунку:

-0,81969	0,681764	12,84634
0,294541	0,09391	3,126084
0,97058	0,071861	#Н/Д
131,9629	8	#Н/Д
1,362914	0,041312	#Н/Д

Результат побудови лінійної регресійної моделі за пакетом аналізу Дані – Аналіз даних – Регресія показано на рисунку:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,98518
R Square	0,97058
Adjusted R Square	0,963225
Standard Error	0,071861
Observations	11

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	1,3629141	0,681457	131,9629	7,49132E-07
Residual	8	0,041312	0,005164		
Total	10	1,4042261			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95,0%*	*Upper 95,0%*
Intercept	12,84633702	3,126084023	4,109402	0,003393	5,637574337	20,0551	5,63757434	20,0550997
X Variable 1	0,681763844	0,093910419	7,259725	8,72E-05	0,465206029	0,8983217	0,46520603	0,89832166
X Variable 2	-0,81969082	0,294540677	-2,78295	0,023816	-1,49890284	-0,140479	-1,49890284	-0,1404788

Вибіркове рівняння множинної регресії має наступний вигляд:

$$\hat{Y}^* = 12,8463 + 0,6818K^* - 0,8197L^*.$$

Перейдемо до початкових змінних $(a_0 = e^{a_0^*} = e^{12,8463} = 379396,5)$ і отримаємо виробничу функцію:

$$\hat{Y} = 379396,5 \cdot K^{0,6818} \cdot L^{-0,8197}.$$

Коефіцієнт a_2 є від'ємним, це означає, що із збільшенням трудових ресурсів обсяг продукції переробної галузі абсолютно знижується.

Коефіцієнт a_0 називають коефіцієнтом технічного прогресу. Часто дослідники коефіцієнт a_0 приймають рівним одиниці.

Побудуємо лінійну регресійну модель при умові $a_0 = 1$ і порівняємо з першою моделлю. Для побудови використаємо пакет аналізу Дані – Аналіз даних – Регресія. Результат побудови показано на рисунку:

SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0,999971				
R Square	0,999941				
Adjusted R Square	0,888824				
Standard Error	0,119498				
Observations	11				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	2183,72516	1091,863	76462,43	7,48784E-18
Residual	9	0,12851754	0,01428		
Total	11	2183,85367			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95,0%*	*Upper 95,0%*
Intercept	0	#Н/Д	#Н/Д	#Н/Д	#Н/Д	#Н/Д	#Н/Д	#Н/Д
X Variable 1	1,03110111	0,066361916	15,53754	8,31E-08	0,880980028	1,1812222	0,88098003	1,181222195
X Variable 2	0,36736064	0,095711166	3,838221	0,003977	0,150846942	0,5838743	0,15084694	0,583874341

Вибіркове рівняння множинної регресії матиме наступний вигляд:

$$\hat{Y}^* = 1,0311K^* + 0,3674L^*.$$

Перейдемо до початкових змінних, отримаємо виробничу функцію:

$$\hat{Y} = K^{1,0311} \cdot L^{0,3674}.$$

Порівняємо графік початкових даних та графіки значень, отриманих за двома побудованими виробничими функціями Кобба-Дугласа (рис. 8.1)

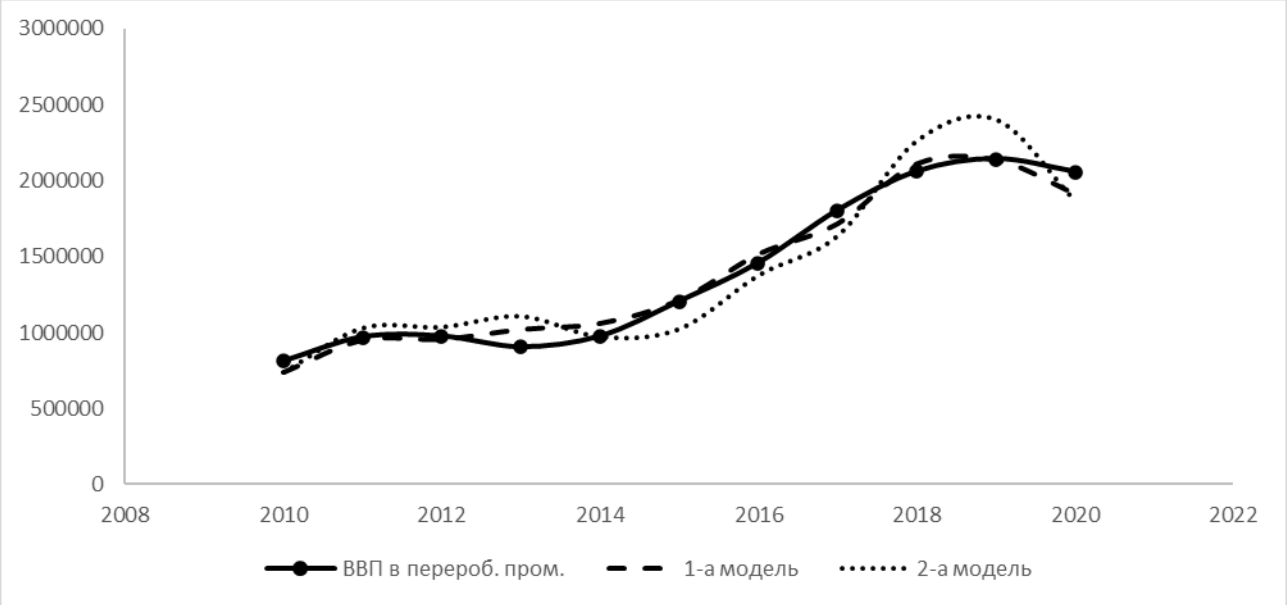


Рисунок 8.1. Графіки початкових даних та графіки значень, отриманих за двома побудованими виробничими функціями

Перша модель краще апроксимує початкові дані. На що вказує й нормований R^2: 0,963 – для першої моделі і 0,889 – для другої. ◎

Задача 8.3. *За статистичними даними [14] податкових надходжень (y, млн. грн.) і рівня податкового навантаження за 1998-2020 рр. побудувати криву Лаффера виду (5.4).*

Рік	Податкові надходження, млн. грн	Рівень податкового навантаження
1998	21848	0,21
1999	25130	0,19
2000	32318	0,18
2001	36716	0,18
2002	45393	0,2
2003	54321	0,2
2004	63162	0,18
2005	98065	0,22
2006	125743	0,23
2007	161264	0,22
2008	227165	0,24
2009	208073	0,23
2010	234448	0,22
2011	334692	0,353
2012	445525	0,36
2013	353968	0,35
2014	367511	0,329
2015	507636	0,343
2016	650782	0,32
2017	828159	0,332
2018	986349	0,336
2019	1070322	0,37
2020	1136687	0,404

○ Для знаходження коефіцієнтів системи (5.5) складемо розрахункову таблицю.

№п/п	x_i	$y_i^*=\ln y_i$	x_i^2	x_i^3	x_i^4	$x_i y_i^*$	$x_i^2 y_i^*$
1	0,21	9,992	0,044	0,009	0,002	2,098	0,441
2	0,19	10,132	0,036	0,007	0,001	1,925	0,366
3	0,18	10,383	0,032	0,006	0,001	1,869	0,336

4	0,18	10,511	0,032	0,006	0,001	1,892	0,341
5	0,2	10,723	0,040	0,008	0,002	2,145	0,429
6	0,2	10,903	0,040	0,008	0,002	2,181	0,436
7	0,18	11,053	0,032	0,006	0,001	1,990	0,358
8	0,22	11,493	0,048	0,011	0,002	2,529	0,556
9	0,23	11,742	0,053	0,012	0,003	2,701	0,621
10	0,22	11,991	0,048	0,011	0,002	2,638	0,580
11	0,24	12,333	0,058	0,014	0,003	2,960	0,710
12	0,23	12,246	0,053	0,012	0,003	2,816	0,648
13	0,22	12,365	0,048	0,011	0,002	2,720	0,598
14	0,353	12,721	0,125	0,044	0,016	4,491	1,585
15	0,36	13,007	0,130	0,047	0,017	4,683	1,686
16	0,35	12,777	0,123	0,043	0,015	4,472	1,565
17	0,329	12,815	0,108	0,036	0,012	4,216	1,387
18	0,343	13,138	0,118	0,040	0,014	4,506	1,546
19	0,32	13,386	0,102	0,033	0,010	4,283	1,371
20	0,332	13,627	0,110	0,037	0,012	4,524	1,502
21	0,336	13,802	0,113	0,038	0,013	4,637	1,558
22	0,37	13,883	0,137	0,051	0,019	5,137	1,901
23	0,404	13,944	0,163	0,066	0,027	5,633	2,276
Сума	6,197	278,966	1,794	0,553	0,179	77,045	22,797

$\overline{x} = 6{,}197 / 23 = 0{,}269$, $\overline{y^*} = 278{,}966 / 23 = 12{,}129$,

$\overline{x^2} = 1{,}794 / 23 = 0{,}078$, $\overline{x^3} = 0{,}553 / 23 = 0{,}024$, $\overline{x^4} = 0{,}179 / 23 = 0{,}008$,

$\overline{xy^*} = 77{,}045 / 23 = 3{,}350$, $\overline{x^2 y^*} = 22{,}797 / 23 = 0{,}991$.

Коефіцієнти a_0^*, a_3, a_1 знаходимо як розв'язок такої системи лінійних рівнянь

$$\begin{cases} a_0^* + 0{,}269a_3 + 0{,}078a_1 = 12{,}129, \\ 0{,}269a_0^* + 0{,}078a_3 + 0{,}024a_1 = 3{,}350, \\ 0{,}078a_0^* + 0{,}024a_3 + 0{,}008a_1 = 0{,}991. \end{cases}$$

Систему рівнянь розв'яжемо матричним способом.

$$A = \begin{pmatrix} 1 & 0{,}269 & 0{,}078 \\ 0{,}269 & 0{,}078 & 0{,}024 \\ 0{,}078 & 0{,}024 & 0{,}008 \end{pmatrix}, \quad B = \begin{pmatrix} 12{,}129 \\ 3{,}350 \\ 0{,}991 \end{pmatrix}, \quad X = \begin{pmatrix} a_0^* \\ a_3 \\ a_1 \end{pmatrix}.$$

З допомогою команди MINVERSE(масив) обчислимо A^{-1}:

$$A^{-1} = \begin{pmatrix} 482,628 & -3652,681 & 6441,871 \\ -3652,681 & 27908,871 & 49569,650 \\ 6441,871 & 49569,650 & 88628,345 \end{pmatrix},$$

а з допомогою команди ***MMULT***(масив1;масив2) добуток $A^{-1}B$:

$$A^{-1}B = \begin{pmatrix} 482,628 & -3652,681 & 6441,871 \\ -3652,681 & 27908,871 & 49569,650 \\ 6441,871 & 49569,650 & 88628,345 \end{pmatrix} \begin{pmatrix} 12,129 \\ 3,350 \\ 0,991 \end{pmatrix} = \begin{pmatrix} 3,064 \\ 53,536 \\ -68,701 \end{pmatrix}.$$

Отже, отримали $a_0^* = 3,064$, $a_3 = 53,536$, $a_1 = -68,701$.

Дальше обчислюємо наступні оцінки. З рівності $-2a_1a_2 = a_3$ маємо $a_2 = \dfrac{53,536}{-2 \cdot (-68,701)} = 0,390$. З рівності $\ln a_0 + a_1 a_2^2 = a_0^*$ маємо $\ln a_0 = -68,701 \cdot 0,390^2 + 3,064 = 13,494$. Звідси $a_0 = e^{13,494} = 724713,6$.

Отже, ми отримали криву Лаффера у вигляді $\hat{y} = 724713,6 \cdot e^{-68,701(x-0,390)^2}$.

Графік кривої Лаффера і статистичних даних зображено на рисунку 8.2.

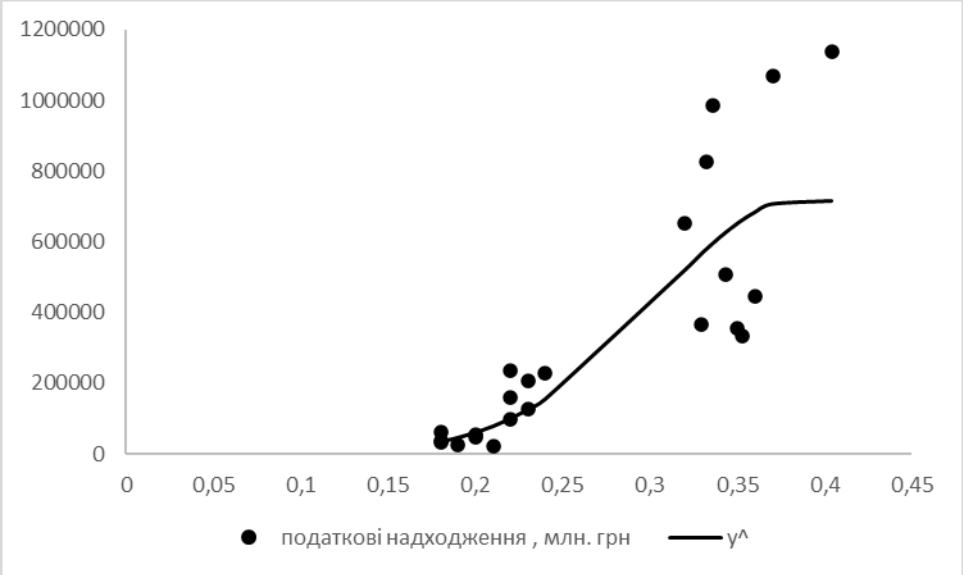


Рисунок 8.2. Графіки статистичних даних та графік кривої Лаффера

ДОДАТКИ

Таблиця 1

Значення функції Лапласа $\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_0^x e^{-\frac{t^2}{2}}dt$

x	Соті долі x									
	0	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
0,0	0,00000	00399	00798	01197	01595	01994	02392	02790	03188	03586
0,1	03983	04380	04776	05172	05567	05962	06356	06749	07142	07535
0,2	07926	08317	08706	09095	09483	09871	10257	10642	11026	11409
0,3	11791	12172	12552	12930	13307	13683	14058	14431	14803	15173
0,4	15542	15910	16276	16640	17003	17364	17724	18082	18439	18793
0,5	19146	19497	19847	20194	20540	20884	21226	21566	21904	22240
0,6	22575	22907	23237	23565	23891	24215	24537	24857	25175	25490
0,7	25804	26115	26424	26730	27035	27337	27637	27935	28230	28524
0,8	28814	29103	29389	29673	29955	30234	30511	30785	31057	31327
0,9	31594	31859	32121	32381	32639	32894	33147	33398	33646	33891
1,0	34134	34375	34614	34850	35083	35314	35543	35769	35993	36214
1,1	36433	36650	36864	37076	37285	37493	37698	37900	38100	38298
1,2	38493	38686	38877	39065	39251	39435	39617	39796	39973	40147
1,3	40320	40490	40658	40824	40988	41149	41308	41466	41621	41774
1,4	41924	42073	42220	42364	42507	42647	42786	42922	43056	43189
1,5	43319	43448	43574	43699	43822	43943	44062	44179	44295	44408
1,6	44520	44630	44738	44845	44950	45053	45154	45254	44352	45449
1,7	45543	45637	45728	45818	45907	45994	46080	46164	46246	46327
1,8	46407	46485	46562	46638	46712	46784	46856	46926	46995	47062
1,9	47128	47193	47257	47320	47381	47441	47500	47558	47615	47670
2,0	47725	47778	47831	47882	47932	47982	48030	48077	48124	48169
2,1	48214	48257	48300	48341	48382	48422	48461	48500	48537	48574
2,2	48610	48645	48679	48713	48745	48778	48809	48840	48870	48899
2,3	48928	48956	48983	49010	49036	49061	49086	49111	49134	49158
2,4	49180	49202	49224	49245	49266	49286	49305	49324	49343	49361
2,5	49379	49396	49413	49430	49446	49461	49477	49492	49506	49520
2,6	49534	49547	49560	49573	49585	49598	49609	49621	49632	49643
2,7	49653	49664	49674	49683	49693	49702	49711	49720	49728	49736
2,8	49744	49752	49760	49767	49774	49781	49788	49795	49801	49807
2,9	49813	49819	49825	49831	49835	49841	49846	49851	49856	49861
3,0	49865	49869	49874	49878	49882	49886	49889	49893	49897	49900
3,1	49903	49906	49910	49913	49916	49918	49921	49924	49926	49929
3,2	49931	49934	49936	49938	49940	49942	49944	49946	49948	49950
3,3	49952	49953	49955	49957	49958	49960	49961	49962	49964	49965
3,4	49966	49968	49969	49970	49971	49972	49973	49974	49975	49976
3,5	49977	49978	49978	49979	49980	49981	48881	49982	49983	49983
3,6	49984	49985	40085	49986	49986	49987	49987	49988	49988	49989
3,7	49989	49990	49990	49990	49991	49991	49992	49992	49992	49992
3,8	49993	49993	49993	49994	49994	49994	49994	49995	49995	49995
3,9	49995	49995	49996	49996	49996	49996	49996	49996	49997	49997

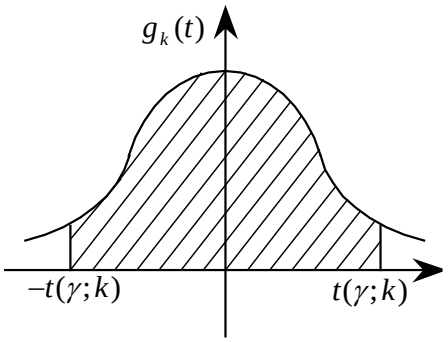
x	Десяті долі x				
	0	*2*	*4*	*6*	*8*
4,	0,4999683	4999867	4999946	4999979	4999992
5,	4999997				

Таблиця 2

Значення $t = t(\gamma; k)$,
що задовільняють рівнянню

$$P\left(|T| < t\right) = 2\int_0^t g_k(t)dt = \gamma,$$

де $g_k(t)$ — густина **розподілу Ст'юдента (*t*-розподілу)**, k — число ступенів вільності



k	γ				
	0,9	*0,95*	*0,98*	*0,99*	*0,999*
1	6,314	12,706	31,821	63,657	636,619
2	2,920	4,303	6,965	9,925	31,599
3	2,353	3,182	4,541	5,841	12,924
4	2,132	2,776	3,747	4,604	8,610
5	2,015	2,571	3,365	4,032	6,869
6	1,943	2,447	3,143	3,707	5,969
7	1,895	2,365	2,998	3,499	5,408
8	1,860	2,306	2,896	3,355	5,041
9	1,833	2,262	2,821	3,250	4,781
10	1,812	2,228	2,764	3,169	4,587
11	1,796	2,201	2,718	3,106	4,437
12	1,782	2,179	2,681	3,055	4,318
13	1,771	2,160	2,650	3,012	4,221
14	1,761	2,145	2,624	2,977	4,140
15	1,753	2,131	2,602	2,947	4,073
16	1,746	2,120	2,583	2,921	4,015
17	1,740	2,110	2,567	2,898	3,965
18	1,734	2,101	2,552	2,878	3,922
19	1,729	2,093	2,539	2,861	3,883
20	1,725	2,086	2,528	2,845	3,850
25	1,708	2,060	2,485	2,785	3,725
30	1,697	2,042	2,457	2,750	3,646
40	1,684	2,021	2,423	2,704	3,551
50	1,676	2,009	2,403	2,678	3,496
60	1,671	2,000	2,390	2,660	3,460
70	1,667	1,994	2,381	2,648	3,435
80	1,664	1,990	2,374	2,639	3,416
90	1,662	1,987	2,368	2,632	3,402
100	1,660	1,984	2,364	2,626	3,390
120	1,658	1,980	2,358	2,617	3,373
∞	1,645	1,960	2,326	2,576	3,291

Таблиця 3

Критичні точки розподілу Ст'юдента (*t*-розподілу)

Для двосторонньої критичної області критична точка $t_{двост.кр}(\alpha;k)=t_\alpha$ є коренем рівняння $\int_0^{t_\alpha} g_k(t)dt=(1-\alpha)/2;$ для односторонньої (правосторонньої) критичної області точка $t_{правост.кр}(\alpha;k)=t_{2\alpha}$ є коренем рівняння $\int_{t_{2\alpha}}^{\infty} g_k(t)dt=(1-2\alpha)/2,$ де $g_k(t)$ — густина розподілу Ст'юдента, k — число ступенів вільності. Для лівосторонньої критичної області $t_{лівост.кр}(\alpha;k)=-t_{2\alpha}.$

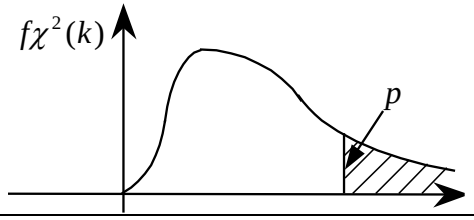
Число ступенів вільності k	Рівень значущості α (**двостороння** критична область)					
	0,10	*0,05*	*0,02*	*0,01*	*0,002*	*0,001*
1	6,314	12,71	31,82	63,66	318,3	637,0
2	2,920	4,303	6,965	9,925	22,33	31,60
3	2,353	3,182	4,541	5,841	10,22	12,94
4	2,132	2,776	3,747	4,604	7,173	8,610
5	2,015	2,571	3,365	4,032	5,893	6,859
6	1,943	2,447	3,143	3,707	5,208	5,959
7	1,895	2,365	2,998	3,499	4,785	5,405
8	1,860	2,306	2,896	3,355	4,501	5,041
9	1,833	2,262	2,821	3,250	4,297	4,781
10	1,812	2,228	2,764	3,169	4,144	4,587
11	1,796	2,201	2,718	3,106	4,025	4,437
12	1,782	2,179	2,681	3,055	3,930	4,318
13	1,771	2,160	2,650	3,012	3,852	4,221
14	1,761	2,145	2,624	2,977	3,787	4,140
15	1,753	2,131	2,602	2,947	3,733	4,073
16	1,746	2,120	2,583	2,921	3,686	4,015
17	1,740	2,110	2,567	2,898	3,646	3,965
18	1,734	2,101	2,552	2,878	3,611	3,922
19	1,729	2,093	2,539	2,861	3,579	3,883
20	1,725	2,086	2,528	2,845	3,562	3,850
21	1,721	2,080	2,518	2,831	3,527	3,819
22	1,717	2,074	2,508	2,819	3,505	3,792
23	1,714	2,069	2,500	2,807	3,485	3,767
24	1,711	2,064	2,492	2,797	3,467	3,745
25	1,708	2,060	2,485	2,787	3,450	3,725
26	1,706	2,056	2,479	2,779	3,435	3,707
27	1,703	2,052	2,473	2,771	3,421	3,690
28	1,701	2,048	2,467	2,763	3,408	3,674
29	1,699	2,045	2,462	2,756	3,396	3,659
30	1,697	2,042	2,457	2,750	3,385	3,646
40	1,684	2,021	2,423	2,704	3,307	3,551
60	1,671	2,000	2,390	2,660	3,232	3,460
120	1,658	1,981	2,362	2,624	3,172	3,374
∞	1,645	1,960	2,326	2,576	3,090	3,291
Число ступенів вільності k	*0,05*	*0,025*	*0,01*	*0,005*	*0,001*	*0,0005*
	Рівень значущості α (**одностороння** критична область)					

Таблиця 4

Значення

$$P\left(\chi^2(k) > \chi^2(p;k)\right) = p,$$

де k — число ступенів вільності



k	p							
	0,999	*0,99*	*0,95*	*0,90*	*0,10*	*0,05*	*0,01*	*0,001*
1	$0{,}157\cdot10^{-5}$	0,0002	0,004	0,02	2,71	3,84	6,63	10,83
2	0,002	0,02	0,10	0,21	4,61	5,99	9,21	13,82
3	0,02	0,12	0,35	0,58	6,25	7,82	11,34	16,27
4	0,09	0,30	0,71	1,06	7,78	9,49	13,28	18,47
5	0,21	0,55	1,15	1,61	9,24	11,07	15,08	20,51
6	0,38	0,87	1,64	2,20	10,65	12,59	16,81	22,46
7	0,60	1,24	2,17	2,83	12,02	14,06	18,48	24,32
8	0,86	1,65	2,73	3,49	13,36	15,51	20,09	26,12
9	1,15	2,09	3,33	4,17	14,68	16,92	21,67	27,88
10	1,48	2,56	3,94	4,87	15,99	18,31	23,21	29,59
11	1,83	3,05	4,58	5,58	17,28	19,68	24,72	31,26
12	2,21	3,57	5,23	6,30	18,55	21,03	26,22	32,91
13	2,62	4,11	5,89	7,04	19,81	22,36	27,69	34,53
14	3,04	4,66	6,57	7,79	21,06	23,69	29,14	36,12
15	3,48	5,23	7,26	8,55	22,31	25,00	30,58	37,70
16	3,94	5,81	7,96	9,31	23,54	26,30	32,00	39,25
17	4,42	6,41	8,67	10,09	24,77	27,59	33,41	40,79
18	4,90	7,02	9,39	10,86	25,99	28,87	34,81	42,31
19	5,41	7,63	10,12	11,65	27,20	30,14	36,19	43,82
20	5,92	8,26	10,85	12,44	28,41	31,41	37,57	45,31
21	6,45	8,90	11,59	13,24	29,62	32,67	38,93	46,80
22	6,98	9,54	12,34	14,04	30,81	33,92	40,29	48,27
23	7,53	10,20	13,20	14,85	32,01	35,17	41,64	49,73
24	8,08	10,86	13,85	15,66	33,19	36,42	43,98	51,18
25	8,65	11,52	14,61	16,47	34,38	37,65	44,31	52,62
26	9,22	12,20	15,37	17,29	35,56	38,89	45,64	54,05
27	9,80	12,88	16,15	18,11	36,74	40,11	46,96	55,48
28	10,39	13,56	16,93	18,94	37,92	41,34	48,28	56,89
29	10,99	14,26	17,71	19,77	39,09	42,56	49,59	58,30
30	11,59	14,95	18,49	20,60	40,26	43,77	50,89	59,70
40	17,92	22,16	26,51	29,05	51,81	55,76	63,69	73,40
50	24,67	29,71	34,76	37,69	63,17	67,51	76,15	86,66
100	61,92	70,07	77,93	82,36	118,50	124,34	135,81	149,45

Таблиця 5

Критичні точки $F_{кр}(\alpha; k_1, k_2)$ розподілу Фішера-Снедокора,

що задовільняють рівнянню $P\left[F > F_{кр}(\alpha; k_1, k_2) \right] = \alpha$ при $\alpha = 0{,}05$

k_2	k_1										
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*12*	*24*	∞
1	161,45	199,50	215,71	224,58	230,16	233,99	237	238,88	243,91	249,05	254,32
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,41	19,45	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,84	8,74	8,64	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	5,91	5,77	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,68	4,53	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,00	3,84	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,57	3,41	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,28	3,12	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,07	2,90	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	2,91	2,74	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,79	2,61	2,40
12	4,75	3,88	3,49	3,26	3,11	3,00	2,91	2,85	2,69	2,50	2,30
13	4,67	3,80	3,41	3,18	3,02	2,92	2,83	2,77	2,60	2,42	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,53	2,35	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,48	2,29	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,42	2,24	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,38	2,19	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,34	2,15	1,92
19	4,38	3.52	3,13	2,90	2,74	2,63	2,54	2,48	2,31	2,11	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,28	2,08	1,84
25	4,24	3,38	2,99	2,76	2,60	2.49	2,40	2,34	2.16	1,96	1,71
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,09	1,89	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,00	1,79	1,51
60	4,00	3,15	2,76	2,52	2,37	2,25	2,17	2,10	1,92	1,70	1,39
120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,83	1,61	1,25
∞	3,84	2,99	2,60	2,37	2,21	2,09	2,01	1,94	1,75	1,52	1,00

Таблиця 6

Критичні значення статистик $d_в$ і $d_н$ критерія Дарбіна-Уотсона при $\alpha = 0{,}05$

m	1		2		3		4		5		6	
	$d_н$	$d_в$	$d_н$	$d_в$	$d_н$	$d_в$	$d_н$	$d_в$	$d_н$	$d_в$	$d_н$	$d_в$
6	0,61	0,40										
7	0,70	1,36	0,47	1,90								
8	0,76	1,33	0,56	1,78	0,37	2,29						
9	0,82	1,32	0,63	1,70	0,46	2,13	0,30	2,59				
10	0,88	1,32	0,70	1,64	0,53	2,02	0,38	2,41	0,24	2,81		
11	0,93	1,32	0,76	1,60	0,60	1,93	0,44	2,28	0,32	2,65	0,12	2,89
12	0,97	1,33	0,81	1,58	0,66	1,86	0,51	2,18	0,38	2,51	0,16	2,67
13	1,01	1,34	0,86	1,56	0,72	1,82	0,57	2,09	0,45	2,39	0,21	2,49
14	1,05	1,35	0,91	1,55	0,77	1,78	0,63	2,03	0,51	2,30	0,26	2,35
15	1,08	1,36	0,95	1,54	0,81	1,75	0,69	1,96	0,56	2,22	0,30	2,24
16	1,11	1,37	0,98	1,54	0,86	1,73	0,73	1,94	0,62	2,16	0,35	2,15
17	1,13	1,38	1,02	1,54	0,90	1,71	0,78	1,90	0,66	2,10	0,39	2,08
18	1,16	1,39	1,05	1,54	0,93	1,70	0,82	1,87	0,71	2,06	0,44	2,02
19	1,18	1,40	1,07	1,54	0,97	1,69	0,96	1,85	0,75	2,02	0,48	1,96
20	1,20	1,41	1,10	1,54	1,00	1,68	0,89	1,83	0,79	1,99	0,52	1,92
21	1,22	1,42	1,13	1,54	1,03	1,67	0,93	1,81	0,83	1,96	0,55	1,88
22	1,24	1,43	1,15	1,54	1,05	1,66	0,96	1,80	0,86	1,94	0,57	1,85
23	1,26	1,44	1,17	1,54	1,08	1,66	0,99	1,79	0,90	1,92	0,62	1,82
24	1,27	1,45	1,19	1,55	1,10	1,66	1,01	1,78	0,93	1,90	0,65	1,79
25	1,29	1,45	1,21	1,55	1,12	1,65	1,04	1,77	0,95	1,89	0,68	1,78
26	1,30	1,46	1,23	1,55	1,14	1,65	1,06	1,76	0,98	1,87	0,71	1,76
27	1,32	1,47	1,24	1,56	1,16	1,65	1,08	1,75	0,10	1,86	0,74	1,74
28	1,33	1,78	1,26	1,56	1,18	1,65	1,10	1,75	1,03	1,85	0,76	1,73
29	1,34	1,48	1,27	1,56	1,20	1,65	1,12	1,74	1,05	1,84	0,79	1,72
30	1,35	1,49	1,28	1,57	1,21	1,65	1,14	1,74	1,07	1,83	0,81	1,71
35	1,40	1,52	1,34	1,58	1,28	1,65	1,22	1,73	1,16	1,80	0,91	1,67
40	1,44	1,54	1,39	1,60	1,34	1,66	1,29	1,72	1,23	1,79	1,00	1,65
45	1,43	1,57	1,43	1,62	1,38	1,67	1,34	1,72	1,29	1,78	1,07	1,64
50	1,50	1,59	1,46	1,63	1,42	1,67	1,38	1,72	1,34	1,77	1,12	1,64
55	1,53	1,60	1,49	1,64	1,45	1,68	1,41	1,72	1,37	1,77	1,17	1,64
60	1,55	1,62	1,51	1,65	1,48	1,69	1,44	1,73	1,41	1,77	1,21	1,64
65	1,57	1,63	1,54	1,66	1,50	1,70	1,47	1,73	1,44	1,77	1,25	1,64
70	1,58	1,64	1,55	1,67	1,53	1,70	1,49	1,74	1,46	1,77	1,28	1,65
75	1,60	1,65	1,57	1,68	1,54	1,71	1,52	1,79	1,49	1,77	1,31	1,65
80	1,61	1,66	1,59	1,69	1,56	1,72	1,53	1,74	1,51	1,77	1,34	1,65
95	1,65	1,69	1,62	1,71	1,60	1,73	1,58	1,76	1,56	1,78	1,39	1,66
100	1,65	1,69	1,53	1,72	1,61	1,74	1,59	1,76	1,57	1,78	1,42	1,67

Список використаних джерел

1. Грубер Й. Економетрія: Навч. посіб. для студ. екон. спец.: Пер. з рос. Т. 1. Вступ до множинної регресії та економетрії. К. : Нічлава, 1998. 381 с.
2. Грубер Й. Економетрія: Навч. посіб. для студ. екон. спец.: Пер. з рос. Т. 2. Економетричні прогнозні та оптимізаційні моделі. Київ : «Нічлава», 1999. 296 с.
3. Економетрика : навч. посіб. / О. Є. Лугінін та ін. Херсон : ОЛДІ-ПЛЮС, 2016. 320 с.
4. Єрьоменко В. О., Шинкарик М. І. Теорія імовірностей : навч. посіб. Тернопіль : Економічна думка, 2002. 176 с.
5. Єрьоменко В. О., Шинкарик М. І. Математична статистика : навч. посіб. Тернопіль : Економічна думка, 2003. 247 с.
6. Здрок В. В., Логоцький Т. Я. Економетрія: підручник. К. : Знання. 2010. 541 с.
7. Іващук О. Т. Економетричні методи та моделі. Тернопіль : Економічна думка, 2002. 348 с.
8. Козьменко О. В., Кузьменко О. В. Економіко-математичні методи та моделі (економетрика) : навч. посіб. Суми : Університетська книга. 2018. 406 с.
9. Лещинський О. Л., Рязанцева В. В., Юнькова О. О. Економетрія: навчальний посібник. К. : МАУП, 2003. 208 с.
10. Лук'яненко І. Г., Краснікова Л. І. Економетрика: підручник. К. : Товариство «Знання», 1998. 494 с.
11. Моделі сталого розвитку: колективна монографія / за ред. Мартинюк О. М. Тернопіль : Підручники і посібники, 2022. 400 с.
12. Назаренко О. М. Основи економетрики : підручник. К. : Центр навчальної літератури, 2005. 392 с.
13. Наконечний С. І., Терещенко Т. О., Романюк Т. П. Економетрія : підручник. 3-є вид., допов. та перероб. К. : КНЕУ, 2005. 520 с.
14. Статистичний щорічник України за 2011-2020 роки / Державна служба статистики України [Електронний ресурс]. Режим доступу: www.ukrstat.gov.ua.
15. Толбатов Ю. А. Економетрика: Підручник для студентів економічних спеціальностей вищих навчальних закладів. К. : Четверта хвиля, 1997. 320 с.

16. Johnston J. Econometric Methods: 2nd Edition. New York: McGraw-Hill. 1972. 437 p.
17. Watsham Terry J., Parramore K. Quantitative Methods in Finance International. Thomson Business Press, 1997. 395 p.
18. Wooldridge J. M. Introductory Econometrics. A Modern Approach: 5th Edition. South-Western, 2013. 910 p.

ЗМІСТ

Вступ 3
§ 1. Поняття про економетричні моделі 5
§ 2. Класична нормальна лінійна модель парної регресії 17
§ 3. Перевірка виконання передумов класичної нормальної лінійної моделі парної регресії 48
§ 4. Множинний регресійний аналіз 70
§ 5. Нелінійні економетричні моделі 101
§ 6. Статистичне оцінювання і тести в узагальнених регресійних моделях 111
§ 7. Часові ряди 120
§ 8. Комп'ютерна реалізація методів економетрики 135
Додатки 156
Список використаних джерел 162

ДЛЯ НОТАТОК

ДЛЯ НОТАТОК

ДЛЯ НОТАТОК

Навчальне видання

Єрьоменко *Валерій Олександрович*
Алілуйко *Андрій Миколайович*
Березька *Катерина Миколаївна*
Мартинюк *Олеся Миронівна*

ЕКОНОМЕТРИКА

Навчальний посібник

Формат 60×84/8. 9,77 ум. др. арк., 8,34 обл.-вид. арк. Тираж 300.
Видавець, виготовлювач і розповсюджувач видавничої продукції
Редакція газети «Підручники і посібники»
46000, м. Тернопіль, вул. Поліська, 6а. Тел.: (0352) 43-15-15; 43-10-31
Збут: pip.ternopil@ukr.net Редакція: editoria@i.ua
Інтернет-магазин: www.pp-books.com.ua
Свідоцтво про внесення суб'єкта видавничої справи
до Державного реєстру видавців, виготовлювачів і розповсюджувачів видавничої продукції
серія ДК № 5143 від 05.07.2016 р.