МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ АВІАЦІЙНИЙ УНІВЕРСИТЕТ
Факультет економіки та бізнес-адміністрування
Кафедра економічної кібернетики

Густера О.М.

КОНСПЕКТ ЛЕКЦІЙ

з дисципліни «Економетричні моделі в умовах цифрової економіки»

за спеціальністю 051 «Економіка», освітньо-професійною програмою «Цифрова економіка»

Укладачі:
асистент кафедри економічної кібернетики
к.е.н Густера О.М.

Конспект лекцій розглянутий та схвалений
на засіданні кафедри
економічної кібернетики
Протокол № № _ від _____р.

Завідувач кафедри_____Н.О. Іванченко

Рекомендовано Вченою радою ФЕБА (протокол №

Густера О.М. Інтернет речей. Конспект лекцій. – К. :НАУ, 2019. – 80 с.

Містить основні положення з дисципліни «Інтернет речей». Для студентів-магістрів спеціальності 051 «Економіка» спеціалізації «Цифрова економіка».

Зміст

ВСТУП

Термін "*економетрика"* буквально означає «економічне вимірювання». Хоча вимірювання є важливою частиною економетрики, сфера її застосування набагато ширша.

Економетрика – економіко-математична наука, яка на основі соціально-економічних статистичних даних вивчає методику побудови економічних моделей для відображення закономірностей, кількісних зв'язків, динаміки соціально-економічних процесів з метою прогнозування, аналізу взаємовпливу явищ і прийняття оптимальних рішень, що стосуються планування і т.д.

Можна сказати, що економетрика є сплавом економічної теорії, математичної економіки, економічної та математичної статистик.

Яким же чином проводиться економетричний аналіз? Тобто яка методологія? У широкому значенні слова методологія економетрики включає такі етапи:

1. Затвердження теорії або гіпотези.
2. Специфікація (уточнення або вибір) математичної моделі теорії.
3. Специфікація економетричної моделі теорії.
4. Збирання даних.
5. Обчислення (отримання) параметрів економетричної моделі.
6. Перевірка гіпотез.
7. Прогнозування.
8. Використання моделі для прийняття економічних рішень або з політичною метою.

Щоб проілюструвати вищевказані кроки, розглянемо добре відому теорію споживання Кейнса (John Keynes).

1. Затвердження теорії або гіпотези. За Кейнсом, витрати людей, як правило, збільшуються зі збільшенням доходів, але не настільки, щоб перевищити їх. Тобто *гранична схильність до споживання* (marginal propensity to consume, MPC) більша за нуль, але менша за одиницю:

$$0< \text{MPC} <1.$$

2. Специфікація математичної моделі споживання. Можна припустити, що залежність між витратами і прибутками лінійна, тобто може бути виражена такою формулою:

$$Y = \beta_1 + \beta_2 X\ ,\ \ 0 < \beta_2 < 1,$$

де Y – витрати; X – доходи, а β_1 і β_2 – деякі параметри моделі, визначувані на основі даних.

3. Специфікація економетричної моделі споживання. Вищенаведена формула припускає, що існує точний зв'язок між витратами і доходами. Але очевидно, що ця залежність змінюватиметься від людини до людини. Щоб урахувати цю неточність у математичному співвідношенні, його переписують у такому вигляді:

$$Y = \beta_1 + \beta_2 X + u\ ,$$

де u – так званий *збурююча складова*, яка є випадковою (стохастичною) величиною. Ця величина охоплює всі ті чинники, які впливають на споживання, але для простоти не враховані.

4. Збирання даних. Збирання надійних даних є завданням економічної статистики.

5. Обчислення параметрів економетричної моделі. Докладний алгоритм обчислення параметрів β_1 і β_2 буде поданий нижче. Відзначимо тільки, що метод обчислення цих параметрів називається *регресійним аналізом*. На основі вищенаведених даних можна отримати такі значення:

$$\beta_1 = -231{,}8 \quad \text{і} \quad \beta_2 = 0{,}7194 .$$

6. Перевірка гіпотез. Вважаючи, що отримана модель задовільно описує дійсність, необхідно розробити критерії, які дозволять з'ясувати, наскільки добре узгоджуються теоретичні дані з експериментальними.

7. Прогнозування. Якщо отримана модель підтверджує гіпотези, то її можна використовувати для прогнозу майбутніх витрат на основі очікуваних величин прибутків.

8. Використання моделі для прийняття економічних рішень або з політичною метою. Припустимо, що витрати в 4000 підтримуватимуть рівень безробіття в 6,5%. Який рівень прибутків забезпечить такий рівень витрат? Ця величина може бути знайдена з рівняння

$$4000 = -231{,}8 + 0{,}7194X .$$

Звідси $X = 5882$. Тобто, при такому рівні прибутків витрати складуть 4000. Очевидно, що ці дані можуть бути використані з політичною метою. Наприклад, відповідною податковою політикою уряд може контролювати рівень прибутків і тим самим отримати бажаний рівень витрат.

1. РЕГРЕСІЙНИЙ АНАЛІЗ. РЕГРЕСІЙНИЙ АНАЛІЗ ДЛЯ ДВОХ ЗМІННИХ: ОСНОВНІ ІДЕЇ

Як було зазначено у вступі, основним інструментом економетрики є регресійний аналіз. Зупинимося коротко на його суті.

Історичне походження терміна «регресія». Уперше термін "регресія" був введений Френсісом Галтоном[1] . Галтон установив таке: хоча й існує тенденція того, що у високих батьків народжуються високі діти, а в невисоких – невисокі, середній зріст дітей, народжених від батьків певного зросту, має тенденцію зміщуватися, "регресувати" в бік середнього зросту в популяції в цілому. Іншими словами, зріст дітей незвичайно високих або низьких батьків має тенденцію зміщуватися в бік середнього зросту популяції. Друг Галтона Карл Пірсон (Karl Pearson) за результатами зібраних ним даних про зріст у групах сімей підтвердив установлений Галтоном закон про універсальну регресію. Він установив, що середній зріст синів з групи високих батьків був менший, ніж середній зріст їх батьків, а середній зріст синів з групи низьких батьків був більший середнього зросту групи батьків, тобто високі й низькі сини «регресували» в бік середнього зросту чоловіків. Галтон охарактеризував це явище як регресію в бік звичайності.

[1] Galton F. Family Likeness in Stature // Proceedings Royal Society. - L., 1886. - Vol.40. - p.42-72.

Сучасна інтерпретація регресії. Сучасне значення, що вкладається в термін “регресія”, зовсім інше. У достатньо широкому значенні слова можна сказати, що *регресійний аналіз пов’язаний із вивченням залежності однієї змінної, такої, що пояснюється, від однієї або декількох пояснювальних змінних, з метою обчислення і/чи прогнозування середньої величини першої при відомих (фіксованих) значеннях останніх.*

Важливість такого підходу до поняття регресійного аналізу стане зрозумілішою в процесі заглиблення в економетрику.

Раніше ми обговорювали концепцію регресійного аналізу в широкому значенні. Зараз ми звернемо увагу на формальний бік предмета. Зокрема, ця частина присвячена введенню в теорію найпростішої регресійної моделі двох змінних. Її розгляд пов’язаний не стільки з важливістю практичного використання, скільки з поданням основних ідей регресійного аналізу в найпростішій формі, і може бути проілюстрований графічно двовимірними діаграмами. Більше того, як ми побачимо далі, більш загальний множинний регресійний аналіз багато в чому є логічним розвитком двовимірної моделі.

1.1. Гіпотетичний приклад

Як зазначалося раніше, регресійний аналіз займається головним чином обчисленням і/чи прогнозуванням середньої величини залежної змінної при фіксованих або передбачуваних значеннях пояснювальних змінних. Щоб зрозуміти, як це робиться, розглянемо гіпотетичний приклад. Уявимо собі гіпотетичну країну з населенням із 60 сімей. Припустимо, що нас цікавить вивчення зв’язку між тижневими споживацькими витратами Y і тижневим прибутком X після сплати податків. Нехай ми хочемо спрогнозувати середні тижневі споживацькі витрати, знаючи тижневий прибуток сім’ї.

Розділимо ці 60 сімей на 10 груп із приблизно однаковим прибутком і дослідимо споживацькі витрати в кожній групі. Гіпотетичні дані наведені в табл. 1.1.

Таблиця 1.1

Прибуток та витрати сімей за тиждень

Тижневий прибуток сім’ї X, дол.	80	100	120	140	160	180	200	220	240	260
Витрати сім’ї за тиждень Y, дол.	55	65	79	80	102	110	120	135	137	150
	60	70	84	93	107	115	136	137	145	152
	65	74	90	95	110	120	140	140	155	175
	70	80	94	103	116	130	144	152	165	178
	75	85	98	108	118	135	145	157	175	180
	-	88	-	113	125	140	-	160	189	185
	-	-	-	115	-	-	-	162	-	191
Усього	325	462	445	707	678	750	685	1043	966	1211

Табл. 1.1 можна інтерпретувати таким чином. У групі сімей із тижневим доходом у 80 дол. (таких сімей п'ять) витрати на споживацькі товари змінюються від 55 до 75 дол. Аналогічно в групі сімей із тижневим доходом у X=240 дол. (шість сімей) витрати на споживацькі товари змінюються від 137 до 189 дол. Іншими словами, кожна колонка табл. 1.1 подає розподіл споживацьких витрат Y, відповідний фіксованому рівню доходу X. Тобто це дає умовний розподіл Y залежно від даної величини X.

Ураховуючи, що наведені в табл. 1.1 дані є повною групою результатів, ми можемо легко підрахувати умовну ймовірність Y при заданому значенні X, $p(Y \mid X)$. Так, наприклад, для $X=80$ Y набуває одного з п'яти значень: 55 дол., 60 дол., 65 дол., 70 дол. і 75 дол. Отже, при даному X=80 дол. ймовірність отримання будь-якого значення з указаних споживацьких витрат дорівнює 1/5. Умовимося позначати це таким чином: $p(Y=55 \mid X=80)$=1/5. Аналогічно $p(Y$=150$\mid X$=260)=1/7 і т. д. Умовна ймовірність даних з табл. 1.1 наведена в табл. 1.2.

Таблиця 1.2

Умовна ймовірність $p(Y \mid X)$ даних табл. 1.1

X	80	100	120	140	160	180	200	220	240	260
$p(Y \mid X)$, умовні ймовірності	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	-	1/6	-	1/7	1/6	1/6	-	1/7	1/6	1/7
	-	-	-	1/7	-	-	-	1/7	-	1/7
Середнє значення	65	77	89	101	113	125	137	149	161	173

Тепер для кожної умовної вірогідності розподілу Y ми можемо вирахувати середню величину витрат, відому як умовне середнє значення або умовне сподівання, що позначається $E(Y \mid X= X_i)$ (читається як "очікувана величина Y при X, що набуває конкретного значення X_i"), яке скорочено ми записуватимемо як $E(Y \mid X_i)$. Сподівання – середня величина витрат у групі. Для наших гіпотетичних даних ці умовні сподівання можна легко підрахувати шляхом множення відповідних значень Y з табл. 2.1 на їх умовну ймовірність і подальшим підсумовуванням добутків. Як приклад підрахуємо умовне середнє значення або сподівання Y при X = 80:

$$55 \cdot \frac{1}{5} + 60 \cdot \frac{1}{5} + 65 \cdot \frac{1}{5} + 70 \cdot \frac{1}{5} + 75 \cdot \frac{1}{5} = 65.$$

Підраховані таким чином умовні середні значення наведені в нижньому рядку табл. 1.2.

Перш ніж просуватися далі, подивимося на дані табл. 1.1, наведені на рис. 1.1. На рисунку точками зображений розподіл величин *Y*, відповідних різним значенням *X*. З рисунка бачимо, що хоча й існують відмінності за споживацькими витратами в окремих сім'ях, але зрозуміло, що середні споживацькі витрати сім'ї збільшуються зі зростанням доходу сім'ї.

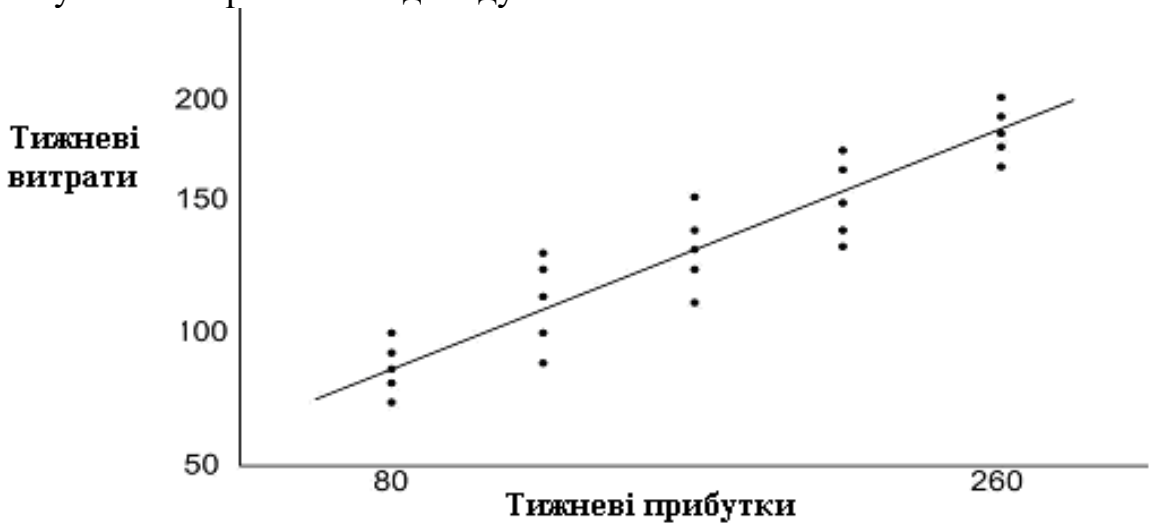


Рис. 1.1. Розподіл величин *Y*, відповідних різним значенням *X*

Це спостереження ще чіткіше простежується, якщо звернутися до зображення на координатній площині точок, що позначають умовні середні значення *Y*.

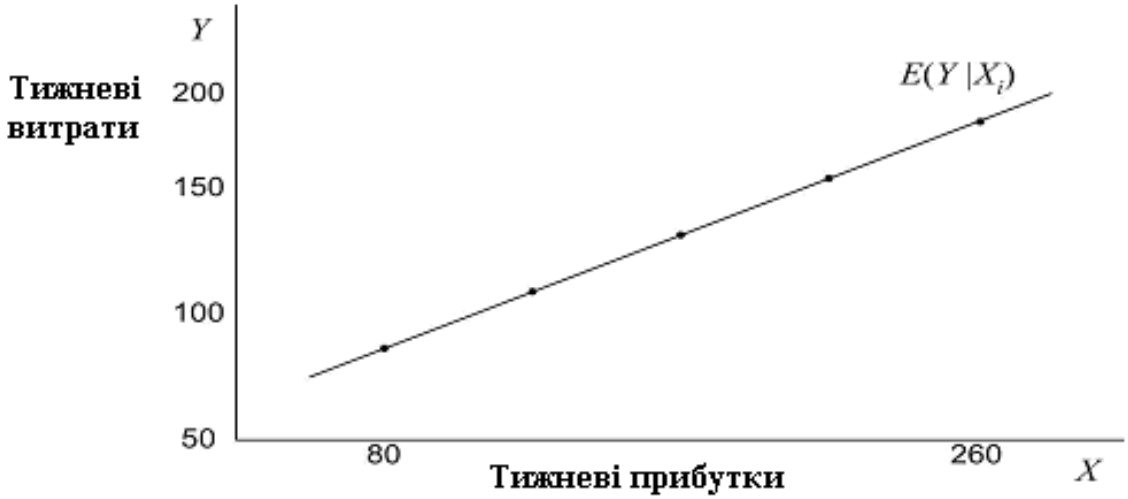


Рис. 1.2. Умовні середні значення *Y*

На рис.1.2 умовні середні значення лежать на прямій лінії з позитивним коефіцієнтом при *X* (При цьому слід пам'ятати, що ми розглядаємо гіпотетичні дані, а отже, умовні середні значення не обов'язково лежатимуть на прямій лінії, вони можуть лежати і на кривій). Ця лінія називається лінією регресії популяції або кривою популяції регресії. Більш просто, це регресія *Y* від *X*.

Геометричне значення таке: регресійна крива популяції – крива точок умовних середніх значень або сподівань залежної змінної від пояснювальної.

Регресійна крива популяції (рис. 1.2) показує, що кожній точці X_i відповідає розподіл значень і певне умовне середнє значення *Y*. Регресійна пряма або крива проходить через умовні середні значення.

1.2. Концепція регресійної функції популяції (PRF population regression function)

З попередніх міркувань (особливо з рис. 1.1 і 1.2) зрозуміло, що кожне умовне середнє $E(Y \mid X_i)$ є функція від X_i. Символічно це можна зобразити у вигляді

$$E(Y \mid X_i) = f(X_i), \qquad (1.2.1)$$

де $f(X_i)$ позначає деяку функцію від пояснювальної змінної X_i. У нашому гіпотетичному прикладі $E(Y \mid X_i)$ є лінійна функція від X_i. Рівняння (1.2.1) відоме як двовимірна **регресійна функція** (**population regression function**) **популяції (PRF),** яка стверджує, що середнє значення від розподілу Y для даного X_i функціонально пов'язане з X_i. Іншими словами, воно показує, як середнє *Y* змінюється зі зміною *X*.

Який вигляд має функція $f(X_i)$? Це питання важливе, оскільки в реальній практичній ситуації ми не маємо у своєму розпорядженні повної сукупності даних для дослідження. Вигляд функції PRF є питання емпіричне, хоча в конкретному випадку теорія може дещо підказати. Наприклад, економіст може констатувати, що споживацькі витрати лінійно пов'язані з доходом сім'ї. Отже, як перше наближення або гіпотезу ми можемо припустити, що PRF $E(Y \mid X_i)$ є лінійна функція від X_i, скажімо вигляду

$$E(Y \mid X_i) = \beta_1 + \beta_2 X_i, \qquad (1.2.2)$$

де β_1 і β_2 – невідомі, але фіксовані параметри під назвою **регресійні коефіцієнти**. Рівняння (1.2.2) відоме як лінійна функція популяції регресії (linear population regression function) або просто лінійна регресія популяції. Є й інші терміни, що використовуються в літературі, – лінійна регресійна модель популяції або лінійне рівняння регресії популяції. Надалі ми використовуватимемо як синоніми терміни "регресія", "рівняння регресії" і "регресійна модель".

У регресійному аналізі нас цікавитиме обчислення PRF у вигляді (1.2.2), тобто обчислення значень невідомих β_1 і β_2 на основі наявних даних спостережень *Y* і *X*.

1.3. Значення терміна “лінійність”

Оскільки ми надалі застосовуватимемо лінійну модель вигляду (1.2.2), важливо визначити точний зміст терміна «лінійний», оскільки можливі дві різні його інтерпретації.

Лінійність за змінною

Перше і можливо найприродніше значення лінійності стосується того, що умовне сподівання Y є лінійна функція від X_i, наприклад вигляду (1.2.2). Геометрично регресійна крива в цьому випадку являє собою пряму лінію. У такій інтерпретації регресійна функція вигляду $E(Y \mid X_i) = \beta_1 + \beta_2 X_i^2$ не є лінійна функція, оскільки змінна X входить у вираз PRF у другому степені.

Лінійність за параметрами

Друга інтерпретація лінійності стосується того, що умовне сподівання Y, $E(Y \mid X_i)$, є лінійна функція за параметрами β_1 і β_2, вона може бути лінійною за змінною або не бути такою. У цій інтерпретації $E(Y \mid X_i) = \beta_1 + \beta_2 X_i^2$ є лінійна регресійна модель, а $E(Y \mid X_i) = \beta_1 + \sqrt{\beta_2} X_i$ не є такою. Остання є нелінійна регресійна моделль (за параметрами). Ми не будемо торкатися подібних випадків.

Надалі вважатимемо, що термін «лінійність» обов’язковий щодо параметрів; він може стосуватися змінних, а може й не стосуватися.

1.4. Стохастичні властивості PRF

З рис. 1.1 бачимо, що зі зростанням доходу сім’ї зростають у середньому і її витрати на споживацькі товари. Проте що можна сказати про витрати на споживання конкретної сім’ї з фіксованим рівнем доходу? З табл. 1.1 і рис. 1.1 бачимо, що зі зростанням доходу конкретної сім’ї рівень її витрат на споживання не обов'язково зростає. Наприклад, з табл. 1.1 бачимо, що в групі сімей із рівнем доходу 100 дол. є одна сім’я, чиї витрати на споживання складають 65 дол., що менше, ніж споживацькі витрати у двох сім’ях з групи з доходом 80 дол. Однак зауважимо, що *середній* рівень споживацьких витрат сімей із тижневим доходом 100 дол. вищий, ніж середній рівень споживацьких витрат у групі сімей із тижневим доходом 80 дол. (77 дол. проти 65 дол.).

Що тоді можна сказати про співвідношення між споживацькими витратами конкретної сім’ї і рівнем її доходу? З рис. 1.1 ми бачимо, що для даного рівня доходу X_i споживацькі витрати сімей розташовуються біля середнього рівня споживацьких витрат групи сімей із тижневим доходом X_i, тобто біля їх умовного сподівання. Отже, ми можемо виразити відхилення індивідуального Y_i від величини його сподівання таким чином:

$$u_i = Y_i - E(Y\left(X_i\right)$$

або (1.4.1)

$$Y_i = E(Y \mid X_i) + u_i,$$

де відхилення u_i – випадкова величина, що набуває як позитивних, так і негативних значень. Величину u_i називають **стохастичним збуренням** або **складовою стохастичної помилки**.

Яку інтерпретацію можна дати (1.4.1)? Можна сказати, що витрати індивідуальної сім'ї з фіксованим рівнем доходу можна подати у вигляді суми двох складових: $E(Y\big(X_i)$, що позначає середній рівень витрат сімей з даним рівнем доходу. Цей доданок відомий як термін **систематичної** або **детермінованої** складової. Другий доданок – u_i, є випадкова величина, так звана **несистематична** компонента. Згодом ми дослідимо природу u_i, а зараз просто відзначимо, що вона містить усі опущені або знехтувані змінні, які можуть впливати на Y, але не включені в регресійну модель.

Якщо $E(Y\big(X_i)$ передбачається лінійною функцією за X_i, як це робилося в (1.2.2), то рівняння (1.4.1) можна переписати у вигляді

$$Y_i = E(Y \mid X_i) + u_i = \beta_1 + \beta_2 X_i + u_i. \qquad (1.4.2)$$

Із рівняння (1.4.2) бачимо, що споживацькі витрати сім'ї лінійно пов'язані з її доходами плюс випадкова складова. Таким чином, індивідуальні витрати на споживацькі товари при X=80 дол. (див. табл. 1.1) можуть бути виражені

$$\begin{aligned} Y_1 &= 55 = \beta_1 + \beta_2(80) + u_1, \\ Y_2 &= 60 = \beta_1 + \beta_2(80) + u_2, \\ Y_3 &= 65 = \beta_1 + \beta_2(80) + u_3, \\ Y_4 &= 70 = \beta_1 + \beta_2(80) + u_4, \\ Y_5 &= 75 = \beta_1 + \beta_2(80) + u_5. \end{aligned} \qquad (1.4.3)$$

Ураховуючи відомі властивості математичного сподівання:

1. Математичне сподівання сталої величини дорівнює самій сталій. Отже, якщо b є константа, то

$$E(b)=b.$$

2. Якщо a і b сталі, а X – випадкова величина, то

$$E(aX+b)=aE(X)+b.$$

Це правило може бути узагальнене. Якщо X_1, X_2 ...X_N – випадкові величини, а a_1, a_2 ..a_N і b – константи, то

$$E(a_1X_1+a_2X_2+...+a_NX_N+b)=a_1E(X_1)+a_2E(X_2)+...+a_NE(X_N)+b.$$

Обчислимо математичне сподівання від обох частин рівності (1.4.1):

$$E(Y_i\big|X_i) = E\Big[E(Y\big|X_i)\Big] + E(u_i\big|X_i) = E(Y\big|X_i) + E(u_i\big|X_i). \qquad (1.4.4)$$

При цьому враховувалося, що математичне сподівання випадкової величини є величина стала. Оскільки в ліву і праву частини рівності (1.4.4) входять однакові величини $E(Y_i \mid X_i)$ і $E(Y\mid X_i)$, то легко одержуємо

$$E(u_i \left(X_i \right) = 0 . \qquad (1.4.5)$$

Таким чином, припущення про те, що лінія регресії проходить через умовні середні *Y*, має на увазі, що математичне сподівання стохастичної складової u_i дорівнює нулю.

Після цього стає зрозуміло, що рівності (1.2.2) і (1.4.2) еквівалентні, якщо для стохастичної складової виконується рівність (1.4.5). Проте стохастичне уточнення (1.4.2) має перевагу, оскільки показує, що окрім доходу сім'ї є й інші змінні, які впливають на споживацькі витрати конкретної сім'ї, і що витрати сім'ї не можуть бути повністю поясненими тільки змінною (змінними), включеними в модель.

1.5. Важливість урахування складової стохастичного збурення

Як було зазначено в попередньому розділі, випадкова складова u_i враховує колективний вплив на *Y* всіх змінних, які не включені в модель. Очевидно постає питання, чому б не ввести в модель ці змінні, або інакше, чому б не доповнити множинну регресійну модель можливою великою кількістю змінних? Причин існує багато.

1. *Неясність теорії.* Теорія (якщо вона взагалі існує), що визначає характер *Y*, може бути (що частіше всього й трапляється) неповною. Ми можемо знати напевно, що тижневий дохід *X* сім'ї впливає на її споживацькі витрати *Y*, проте ми можемо не знати або не бути впевненими у впливі на Y інших змінних. Отже, u_i може бути застосований як заміна для всіх виключених або опущених у моделі змінних.

2. *Неповнота даних*. Навіть якщо нам і відомо, що собою являють деякі змінні, виключені з моделі, і отже, має місце множинна, а не найпростіша регресія, нам може бути недоступна кількісна інформація щодо цих змінних. В емпіричних дослідженнях дуже часто трапляється так, що дані, необхідні нам, виявляються недоступними. Наприклад, ми могли б ввести в модель на додачу до доходу як пояснювальну змінну накопичені заощадження сім'ї. Але, на жаль, інформація про грошові заощадження сім'ї виявляється недоступною. Отже, ми вимушені виключити цю змінну з моделі всупереч тому, що вона значною мірою впливає на пояснення споживацьких витрат сім'ї.

3. *Протиставлення* г*оловних змінних периферійним змінним.* Припустимо, що в нашій моделі "споживання – дохід" окрім доходу X_1 враховуються кількість дітей в сім'ї X_2, стать X_3, віросповідання X_4, рівень освіти X_5, регіон мешкання X_6, які також впливають на рівень споживацьких витрат. Але цілком можливо, що сумарний вплив усіх або деяких із цих змінних може бути таким незначним, що з погляду на їх внесок, і зважаючи на ускладнення моделі, їх не слід включати у модель в явному вигляді. Можна сподіватися, що їх сумарний вплив враховується як випадкова величина u_i .

4. *Властива природі людини випадковість*. Навіть за умови введення в модель усіх змінних, все одно залишається властива індивідуальному значенню *Y*

випадковість, яку, як би ми не хотіли, пояснити неможливо. Стохастична складова *u* може дуже добре відображати цю властиву випадковість.

5. *Слабка довіра до змінних.* Хоча класична регресійна модель припускає точне вимірювання змінних *Y* і *X*, на практиці дані можуть містити помилки. Розглянемо, наприклад, добре відому модель функції споживання Мільтона Фрідмана (Milton Friedman). Згідно з нею постійне споживання (*Yp*) вважається функцією постійного доходу (*Xp*). Але оскільки дані за цими змінними не є доступними безпосередньо, на практиці ми застосовуємо змінні, що їх замінюють, такі як поточне споживання (*Y*) і поточний дохід (*X*), за якими дані доступні. Оскільки дані, доступні за *Y* і *X*, не обов'язково повинні збігатися з даними за змінними *Yp* і *Xp*, то існує проблема похибки вимірювання. Збурююча складова *u* може при цьому являти собою помилку вимірювання. Як ми побачимо далі, наявність похибки вимірювання серйозно позначається на точності визначення коефіцієнтів β.

6. *Дотримання принципу найбільшої простоти.* Згідно з відомою тезою про те, що модель потрібно зберігати в найпростішому вигляді, поки не доведена її неадекватність, є природним бажання зберегти регресійну модель у найпростішому вигляді. Якщо ми можемо пояснити характер *Y* достатньо докладно за допомогою двох або трьох пояснювальних змінних і якщо наша теорія не припускає, що інші змінні повинні бути включені в модель, навіщо їх тоді вводити? Нехай u_i замінює собою всю решту змінних. Звичайно ж, не слід виключати з моделі змінні, що стосуються справи, лише заради збереження простоти моделі.

7. *Неправильний вид функціональної залежності.* Навіть якщо ми маємо теоретично коректні пояснювальні змінні, що описують явище, а також маємо достовірні дані, що стосуються цих змінних, дуже часто ми не знаємо вигляду функціональної залежності між регресандом і регресорами. Функція споживацьких витрат є лінійною по відношенню до доходу чи нелінійною? У двовимірному регресійному аналізі уявлення про вигляд функціональної залежності може бути отримане з графіка. У множинному регресійному аналізі досить нелегко визначити правильний вигляд функціональної залежності, оскільки неможливо візуально скласти уявлення про неї.

З усіх цих причин, що стосуються стохастичного збурення, складова u_i відіграє в регресійному аналізі надзвичайно важливу роль.

1.6. Вибіркова регресійна функція (SRF)

Обмежуючи наше обговорення випадком дослідження популяції *Y*, відповідної деякому фіксованому *X*, ми навмисно уникали розгляду вибіркових даних. Уміщені в табл. 1.1 дані стосуються всієї популяції. Проте настав час звернутися до розгляду вибіркових даних, оскільки на практиці такі випадки зустрічаються найчастіше. Отже, нашою задачею є обчислення PRF на основі вибіркової, тобто неповної інформації.

Наприклад, нам була відома не вся інформація з табл. 1.1, а лише її частина, отримана шляхом випадкової вибірки з цієї таблиці. Отримана таким чином інформація подана в табл. 1.3.

Таблиця 1.3

X	80	100	120	140	160	180	200	220	240	260
Y	70	65	90	95	110	115	120	140	155	150

На відміну від табл. 1.1, ми зараз маємо єдине значення *Y*, відповідне фіксованому значенню *X*; кожне *Y* (при фіксованому *X*) з табл. 1.4 вибране випадковим чином із табл. 1.1.

Питання полягає в такому: чи можемо ми за вибірковими даними табл. 1.1 розрахувати середні тижневі споживацькі витрати за всією популяцією? Іншими словами, чи можемо обчислити PRF за вибірковими даними? Інтуїтивно зрозуміло, що ми не можемо обчислити PRF «точно» через флуктуації за вибіркою. Щоб побачити це більш наочно, припустимо, що ми маємо іншу випадкову вибірку з табл. 1.1, наведену в табл. 1.4.

Таблиця 1.4

X	80	100	120	140	160	180	200	220	240	260
Y	55	88	90	80	118	120	145	135	145	175

Дані з таблиць 1.4 і 1.5 зобразимо на графіку (рис. 1.3). На рис. 1.3 дві прямі вибіркової регресії проведені так, щоб найкращим чином відповідати даним таблиць: SRF1 ґрунтується на даних першої вибірки, а SRF2 – другої. Яка з цих двох ліній є істинна лінія регресії популяції? Якщо не брати до уваги рис. 1.1, де зображена лінія регресії популяції, то не існує способу, що дозволяє напевно стверджувати, яка з ліній являє собою істинну лінію регресії популяції. Показанні на рис. 1.3 лінії називаються **лініями вибіркової регресії (sample regression lines)**. Передбачається, що вони зображають лінію регресії популяції, проте через флуктуації за вибіркою вони в кращому разі зображають апроксимацію. У загальному випадку для *N* вибіркових даних може бути отриманий *N* SRF і, швидше за все, усі вони будуть різними.

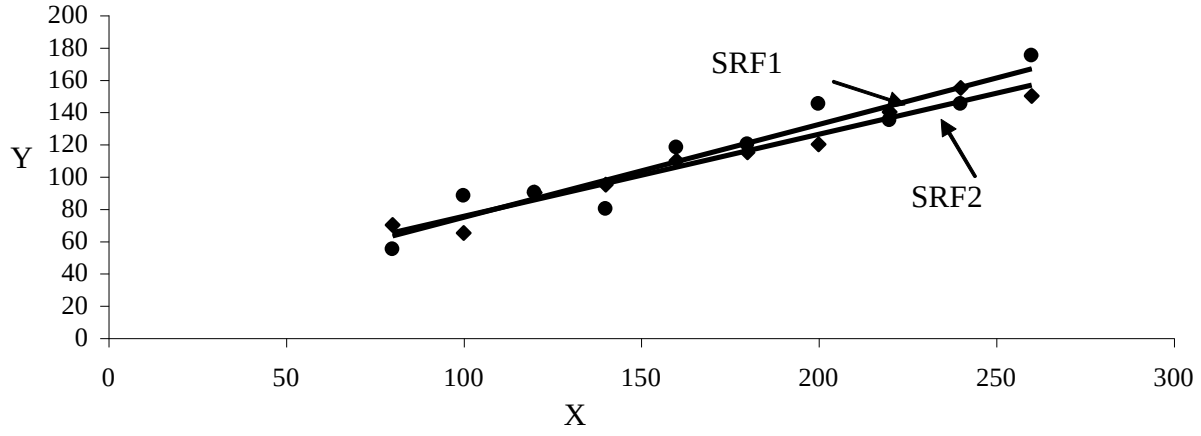


Рис. 1.3. Лінії вибіркової регресії SRF1, SRF2

Тепер за аналогією з поняттям функції регресії популяції PRF можна ввести поняття функції вибіркової регресії SRF. Аналог формули (1.2.2) може бути записаний у вигляді

$$Y_i = \beta_1 + \beta_2 X_i, \tag{1.6.1}$$

де Y_i – математичне сподівання $E(Y \mid X_i)$, обчислене на основі вибіркових даних; β_1, β_2 – коефіцієнти регресії, обчислені на основі вибіркових даних.

Тепер за аналогією з уявленням PRF у двох еквівалентних формах (1.2.2) і (1.4.2) ми можемо подати SRF (1.6.1) у стохастичному вигляді

$$Y_i = \beta_1 + \beta_2 X_i + u_i, \tag{1.6.2}$$

де u_i – залишковий член, отриманий за вибірковими даними.

Підсумовуючи сказане, ми уявляємо собі основну задачу регресійного аналізу в обчисленні PRF

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

на основі SRF

$$Y_i = \beta_1 + \beta_2 X_i + u_i.$$

Частіше на практиці доводиться проводити аналіз на основі вибіркових даних, а не даних за всією популяцією. Але, зважаючи на флуктуації вибіркових даних, отримана на підставі SRF функція PRF є в кращому разі апроксимацією істинної PRF.

Для $X = X_i$ ми маємо одне спостереження $Y = Y_i$. По відношенню до SRF Y_i може бути подано таким чином:

$$Y_i = Y_i + u_i, \tag{1.6.3}$$

а по відношенню до PRF його можна подати у вигляді

$$Y_i = E(Y \mid X_i) + u_i. \tag{1.6.4}$$

Y_i дає в порівнянні з істинним значенням $E(Y\left(X_i\right)$ для X_i завищене значення. Зрозуміло, що подібні заниження і завищення пояснюються флуктуаціями вибіркових даних.

Основне питання полягає в такому: розуміючи, що SRF є не більше ніж апроксимація PRF, чи можемо ми запропонувати правило або метод, що дозволяли б якомога більше наблизитися до PRF? Іншими словами, яким чином побудувати SRF так, щоб β_1, β_2 були б якомога ближчі до β_1, β_2, хоча ми й не знаємо істинні значення β_1, β_2. Відповідь на це питання така: ми запропонуємо метод, що дозволить сконструювати SRF так, щоб вона найбільш точно відображала властивості PRF. Надзвичайно цікавим є той факт, що ми зможемо це зробити навіть не маючи нагоди визначити дійсну PRF.

2. ДВОВИМІРНА РЕГРЕСІЙНА МОДЕЛЬ. ЗАДАЧА ОЦІНКИ

2.1. Метод найменших квадратів

Метод найменших квадратів (МНК) був запропонований Карлом Фрідріхом Гауссом – німецьким математиком. При деяких припущеннях МНК має дуже привабливі статистичні властивості, які роблять його одним із щонайпотужніших і популярних методів регресійного аналізу. Для того щоб зрозуміти цей метод, слід спочатку пояснити принцип найменших квадратів.

Пригадаємо двовимірну PRF:

$$Y_i = \beta_1 + \beta_2 X_i + u_i .$$

Як нами було відзначене раніше, PRF не є об'єктом, який можна отримати прямо. Ми оцінюємо його з вибіркової регресійної функції (SRF):

$$Y_i = \beta_1 + \beta_2 X_i + u_i,$$

$$Y_i = Y_i + u_i ,$$

де $Y_i = \beta_1 + \beta_2 X_i$ – оцінена величина Y_i.

Для визначення SRF зобразимо спочатку (1.6.3) у вигляді

$$u_i = Y_i - Y_i = Y_i - \beta_1 - \beta_2 X_i, \qquad (2.1.1)$$

який підказує, що u_i (залишки, стохастична або випадкова складова) є різницею між дійсною та оціненою величиною *Y*.

Тепер для даних *N* пар спостережень над (Y_i, X_i), визначимо SRF так, щоб вона була розташована якомога ближче до дійсних *Y*. Для цього можна обрати такий критерій: виберемо SRF так, щоб сума залишків $\sum u_i = \sum (Y_i - Y_i)$ була наскільки можливо мала. Хоча інтуїтивно такий критерій здається привабливим, насправді він не дуже вдалий, як це можна бачити з гіпотетичного прикладу, показаного на рис. 2.1.

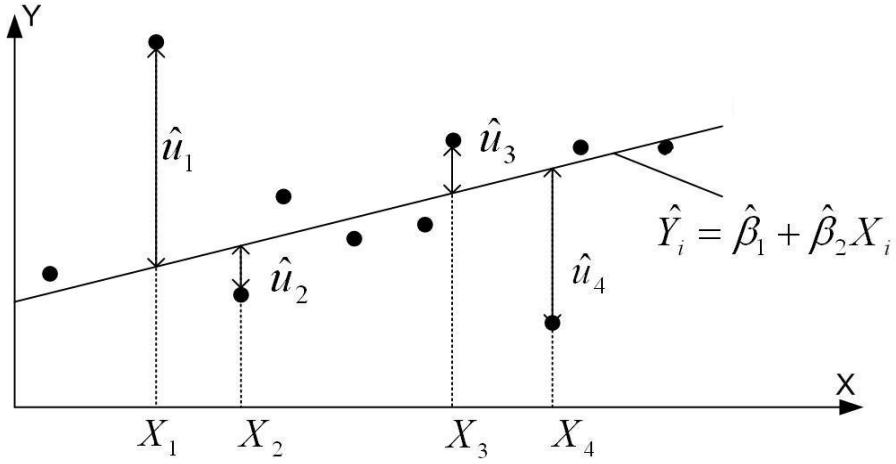


Рис. 2.1. Критерій найменших квадратів

Якщо ми візьмемо критерій мінімізації $\sum u_i$, то згідно з рисунком залишки u_2 і u_3 мають у сумі $(u_1 + u_2 + u_3 + u_4)$ те ж значення, що й u_1 і u_4, хоча перші два залишки набагато ближчі до SRF, ніж два останніх. Іншими словами, усі за-

лишки мають однакові значення безвідносно до того, наскільки вони близько або далеко розташовані від SRF. Щоб переконатися в цьому, хай u_1, u_2, u_3, u_4 на рис. 2.1 набувають, відповідно, значень 10, –2, 2, –10. Сума цих залишків є нуль, хоча u_1 і u_4 розташовані набагато далі, ніж u_2 і u_3. Ми можемо уникнути подібної ситуації, якщо візьмемо критерій найменших квадратів, який стверджує, що SRF можна фіксувати, якщо

$$\sum u_i^2 = \sum (Y_i - Y_i)^2 = \sum (Y_i - \beta_1 - \beta_2 X_i)^2 \qquad (2.1.2)$$

набуває найменшого значення. За рахунок піднесення у квадрат цей метод надає більшого значення таким залишкам як u_1 і u_4 порівняно з u_2 і u_3. Як було відзначено раніше, при критерії мінімуму $\sum u_i$ можлива ситуація, коли сума залишків мала, але u_i широко розкидані навколо лінії SRF. Проте це неможливо при виборі критерію мінімуму суми квадратів залишків, оскільки чим більше u_i (за абсолютною величиною), тим більше $\sum u_i^2$. Наступне обґрунтовування переваги МНК над іншим критерієм полягає в тому, що отримані за МНК оцінки мають деякі привабливі з погляду статистики властивості.

З (2.1.2) очевидно, що

$$\sum u_i^2 = f(\beta_1, \beta_2), \qquad (2.1.3)$$

тобто сума квадратів залишків є функція від оцінок β_1 і β_2. Для будь-якої заданої сукупності даних вибір різних величин для β_1 і β_2 дає різні залишки u_i і, отже, різну величину $\sum u_i^2$. Щоб було зрозуміліше, розглянемо гіпотетичні дані *Y* і *X*, наведені в перших двох колонках табл. 2.1.

Таблиця 2.1

Експериментальне визначення SRF

Y_i	X_i	Y_{1i}	u_{1i}	u_{1i}^2	Y_{2i}	u_{2i}	u_{2i}^2
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
4	1	2.929	1.071	1.147	4	0	0
5	4	7.000	–2.000	4.000	7	–2	4
7	5	8.357	–1.357	1.841	8	–1	1
12	6	9.714	2.286	5.226	9	3	9
Сума: 28	16		0.000	12.214		0	14

Зауваження: $Y_{1i} = 1{,}572 + 1{,}357 X_i$ (тобто $\beta_1 = 1{,}572$, $\beta_2 = 1{,}357$),

$$Y_{2i} = 3 + X_i, \quad u_{1i} = Y_i - Y_{1i}, \quad u_{2i} = Y_i - Y_{2i}.$$

Проведемо два експерименти. В експерименті 1 покладемо $\beta_1 = 1{,}572$ і $\beta_2 = 1{,}357$ (зараз ми не говоримо про те, як ми отримали ці значення, припустимо

ми просто вгадали). Застосовуючи ці значення та величини *X*, подані в колонці *2*, ми легко можемо визначити оцінку Y_i, наведену в колонці 3 як Y_{1i} (індекс 1 позначає перший експеримент). Проведемо другий експеримент, поклавши цього разу $\beta_1 = 3$ і $\beta_2 = 1$ (оцінки величини Y_i при цьому наведені в колонці *6*). Оскільки β - величини у двох експериментах різні, ми маємо два різні значення для суми квадратів залишків (як відомо з таблиці); u_{1i} – залишки за першим експериментом, а u_{2i} – за другим (квадрати цих залишків наведені в колонках *5* і *8*). З наведених у таблиці даних бачимо, що, як і очікувалося з (2.1.3), суми квадратів залишків у двох експериментах різні, оскільки вони базуються на різних β - величинах.

Тепер виникає питання, які β-величини нам слід вибрати? Оскільки β-величини в першому експерименті дають меншу величину $\sum u_i^2$ (=12.214), ніж у другому (=14), ми можемо сказати, що β– величини в першому експерименті «кращі», ніж у другому. Але чи «кращі» вони за всіма можливими значеннями? Якби ми мали необмежений час, то могли б багато разів проводити подібні експерименти, вибираючи різні β-величини кожного разу й порівнюючи результуючу величину $\sum u_i^2$. При цьому, зрозуміло, нам потрібно було б перебрати всі можливі значення β_1 і β_2. Але оскільки час обмежений, то потрібен більш поширений метод, ніж процедура спроб і помилок. На щастя, МНК дозволяє скористатися добре відомою процедурою знаходження мінімуму функції з двома змінними.

Для цього отримаємо вираз для функції $f(\beta_1, \beta_2)$ з (2.1.3):

$$\sum u_i^2 = \sum (Y_i - Y_i)^2 = \sum (Y_i - \beta_1 - \beta_2 X_i)^2 . \qquad (2.1.4)$$

З курсу вищої математики відомо, що функція (2.1.4) набуватиме якнайменшого значення при тих числових значеннях β_1 і β_2, при яких перетворюються в нуль частинні похідні $\frac{\partial\left(\sum u_i^2\right)}{\partial \beta_1}$ і $\frac{\partial\left(\sum u_i^2\right)}{\partial \beta_2}$. Обчислимо їх, пам'ятаючи, що X_i і Y_i не залежать від β-величин:

$$\frac{\partial\left(\sum u_i^2\right)}{\partial \beta_1} = -2\sum (Y_i - \beta_1 - \beta_2 X_i) = -2\sum u_i \qquad (2.1.5)$$

$$\frac{\partial\left(\sum u_i^2\right)}{\partial \beta_2} = -2\sum (Y_i - \beta_1 - \beta_2 X_i) X_i = -2\sum u_i X_i \qquad (2.1.6)$$

Зрівнюючи до нуля ці вирази, одержуємо таку систему двох лінійних алгебраїчних рівнянь щодо коефіцієнтів регресії β_1 і β_2:

$$\begin{cases} n\beta_1 + \beta_2 \sum X_i = \sum Y_i, \\ \beta_1 \sum X_i + \beta_2 \sum X_i^2 = \sum X_i Y_i. \end{cases} \tag{2.1.7}$$

Отримаємо розв'язок цієї системи, застосовуючи, наприклад, формули Крамера:

$$\Delta = \begin{vmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{vmatrix} = n\sum X_i^2 - \left(\sum X_i\right)^2 \neq 0;$$

$$\Delta_1 = \begin{vmatrix} \sum Y_i & \sum X_i \\ \sum X_i Y_i & \sum X_i^2 \end{vmatrix} = \sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i;$$

$$\Delta_2 = \begin{vmatrix} n & \sum Y_i \\ \sum X_i & \sum X_i Y_i \end{vmatrix} = n\sum X_i Y_i - \sum X_i \sum Y_i;$$

$$\beta_1 = \frac{\Delta_1}{\Delta} = \frac{\sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i}{n\sum X_i^2 - \left(\sum X_i\right)^2}; \tag{2.1.8}$$

$$\beta_2 = \frac{\Delta_2}{\Delta} = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - \left(\sum X_i\right)^2}. \tag{2.1.9}$$

Дведемо, що можна отримати більш прості вирази для коефіцієнтів регресії. Для цього введемо нові величини – відхилення від середніх за вибіркою значень

$$x_i = X_i - \bar{X}, \quad y_i = Y_i - \bar{Y}, \tag{2.1.10}$$

де $\bar{X}$ і $\bar{Y}$ позначають середні за вибіркою значення

$$\bar{X} = \frac{1}{n}\sum X_i, \quad \bar{Y} = \frac{1}{n}\sum Y_i.$$

З цих формул відразу випливають прості співвідношення

$$\sum X_i = n\bar{X}, \quad \sum Y_i = n\bar{Y}. \tag{2.1.11}$$

Покажемо, що

$$\sum x_i = 0, \quad \sum y_i = 0. \tag{2.1.12}$$

Дійсно

$$\sum x_i = \sum (X_i - \bar{X}) = \sum X_i - \sum \bar{X} = n\bar{X} - n\bar{X} = 0.$$

Аналогічно доводиться і друга рівність. Окрім цього справедлива формула

$$\sum x_i^2 = \sum X_i^2 - n\bar{X}^2 = \frac{1}{n}\left[n\sum X_i^2 - \left(\sum X_i\right)^2 \right]. \tag{2.1.13}$$

Шляхом елементарних перетворень одержуємо

$$\sum x_i^2 = \sum (X_i - \bar{X})^2 = \sum X_i^2 - 2\sum X_i \bar{X} + \sum \bar{X}^2 =$$
$$= \sum X_i^2 - 2\bar{X}\sum X_i + n\bar{X}^2 = \sum X_i^2 - 2\bar{X}\, n\,\bar{X} + n\bar{X}^2 =$$
$$= \sum X_i^2 - n\bar{X}^2 = \sum X_i^2 - n\left(\frac{1}{n}\sum X_i\right)^2 =$$
$$= \sum X_i^2 - \frac{1}{n}\left(\sum X_i\right)^2 = \frac{1}{n}\left[n\sum X_i^2 - \left(\sum X_i\right)^2\right].$$

Порівнюючи вирази для Δ і (2.1.13), можна помітити, що визначник Δ відрізняється від нуля у випадку, якщо не всі X_i мають однакові значення.

Наведемо ще одне співвідношення

$$\sum x_i y_i = \frac{1}{n}\left[n\sum X_i Y_i - \sum X_i \sum Y_i\right] = \sum x_i Y_i = \sum X_i y_i \qquad (2.1.14)$$

Доведення:

$$\sum x_i y_i = \sum (X_i - \bar{X})\sum (Y_i - \bar{Y}) = \sum X_i Y_i - \sum \bar{X} Y_i - \sum X_i \bar{Y} + \sum \bar{X}\bar{Y} =$$
$$= \sum X_i Y_i - \bar{X}\sum Y_i - \bar{Y}\sum X_i + n\bar{X}\bar{Y} =$$
$$= \sum X_i Y_i - \bar{X} n \bar{Y} - \bar{Y} n \bar{X} + n\bar{X}\bar{Y} =$$
$$= \sum X_i Y_i - n\bar{X}\bar{Y} = \sum X_i Y_i - n\left(\frac{1}{n}\sum Y_i\right)\left(\frac{1}{n}\sum X_i\right) =$$
$$= \sum X_i Y_i - \frac{1}{n}\sum X_i \sum Y_i = \frac{1}{n}\left[n\sum X_i Y_i - \sum X_i \sum Y_i\right].$$

Водночас

$$\sum x_i y_i = \sum (X_i - \bar{X})\sum (Y_i - \bar{Y}) = \sum X_i (Y_i - \bar{Y}) - \sum \bar{X}(Y_i - \bar{Y}) =$$
$$= \sum X_i y_i - \sum \bar{X}(Y_i - \bar{Y}) = \sum X_i y_i.$$

Звернемося тепер до виразу для β_2 (2.1.9). Застосовуючи (2.1.13) і (2.1.14), одержуємо

$$\beta_2 = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{\sum x_i Y_i}{\sum X_i^2 - n\bar{X}^2} = \frac{\sum X_i y_i}{\sum X_i^2 - n\bar{X}^2} \qquad (2.1.15)$$

Спростимо тепер вираз (2.1.8) для β_1. З першого рівняння системи (2.1.7) одержуємо

$$\beta_1 = \frac{1}{n}\sum Y_i - \beta_2 \frac{1}{n}\sum X_i = \bar{Y} - \beta_2 \bar{X}\,. \qquad (2.1.16)$$

У наступному пірозділі ми зупинимося на деяких властивостях оцінок, отриманих за методом найменших квадратів.

2.2. Властивості оцінок за МНК

1. Оцінки за МНК виражаються тільки через величини спостережуваних змінних *X* і *Y*. Унаслідок цього вони можуть бути легко обчислені за формулами (2.1.15) (2.1.16).

2. Ці оцінки є точковими, тобто для даної вибірки кожен оцінювач матиме єдину величину, визначувану спостережуваними значеннями. Надалі ми розглянемо так звані інтервальні оцінки, які дають інтервал можливих значень невідомих параметрів.

Відзначимо деякі властивості лінії SRF.

1. Лінія SRF проходить через середні значення вибірки *X* і *Y*, тобто через точку $(\bar{X},\bar{Y})$. Цей факт безпосередньо випливає з формули (2.1.16), якщо її перетворити до вигляду

$$\bar{Y} = \beta_1 + \beta_2 \bar{X}\,.$$

На рис. 2.2 показано, що лінія SRF проходить через точку $(\bar{X},\bar{Y})$.

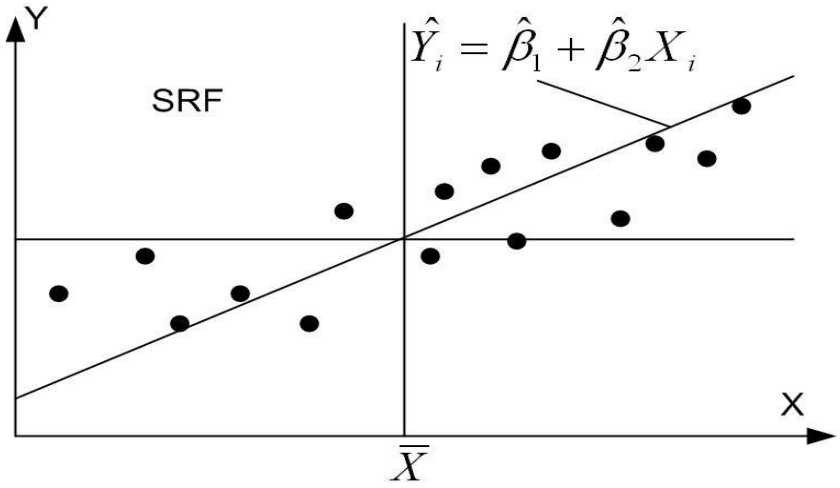


Рис. 2.2. Лінія SRF

2. Середня величина *Y* за наслідками вибірки дорівнює середній величині, отриманій із рівняння регресії $Y_i = \beta_1 + \beta_2 X_i$, тобто

$$\bar{Y} = \bar{Y}. \qquad (2.2.1)$$

Щоб довести це, виконаємо такі перетворення:

$$Y_i = \beta_1 + \beta_2 X_i = (\bar{Y} - \beta_2 \bar{X}) + \beta_2 X_i = \bar{Y} + \beta_2 (X_i - \bar{X})\,.$$

Ми підставили замість β_1 його вираз із (2.1.16). Підсумуємо обидві частини отриманої рівності за всім обсягом вибірки

$$\sum Y_i = \sum \bar{Y} + \beta_2 \sum (X_i - \bar{X}),$$

$$n\bar{Y} = n\bar{Y}\,,\ \ \bar{Y} = \bar{Y}.$$

При цьому ми використовували властивість (2.1.12).

3. Середня величина залишків u_i є нуль, тобто $\bar{u} = 0$. Ця властивість виходить безпосередньо з (2.1.5) і (2.1.7). Використовуючи цю властивість, можна перетворити рівняння SRF

$$Y_i = \beta_1 + \beta_2 X_i + u_i \qquad (2.2.2)$$

до іншого вигляду, в який будуть входити лише змінні у відхиленнях. Для цього підсумуємо останню рівність за всією вибіркою

$$\sum Y_i = n\beta_1 + \beta_2 \sum X_i + \sum u_i = n\beta_1 + \beta_2 \sum X_i .$$

Поділивши останню рівність на *N*, одержуємо

$$\bar{Y} = \beta_1 + \beta_2 \bar{X} .$$

Віднімемо цю рівність від (2.2.2):

$$Y_i - \bar{Y} = \beta_2 (X_i - \bar{X}) + u_i$$

або

$$y_i = \beta_2 x_i + u_i . \qquad (2.2.3)$$

Рівняння (2.2.3) називається рівнянням у відхиленнях. Зауважимо, що β_1 в це рівняння не входить. Рівняння SRF у відхиленнях можна подати у вигляді

$$y_i = \beta_2 x_i . \qquad (2.2.3)$$

4. Залишки u_i не корелюються з x_i, тобто $\sum x_i u_i = 0$. Це твердження виходить безпосередньо з (2.1.6) і (2.1.7).

5. Залишки u_i не корелюються з Y_i, тобто $\sum Y_i u_i = 0$. Дійсно

$$\sum Y_i u_i = \sum (\beta_1 + \beta_2 X_i) u_i = \beta_1 \sum u_i + \beta_2 \sum X_i u_i = 0 .$$

6. Залишки u_i не корелюються з y_i, тобто $\sum y_i u_i = 0$. Підставляючи замість y_i його значення з (2.2.3), одержуємо

$$\sum y_i u_i = \beta_2 \sum x_i u_i = \beta_2 \sum (X_i - \bar{X}) u_i = \beta_2 \sum X_i u_i - \beta_2 \bar{X} \sum u_i = 0 .$$

Припущення, що застосовуються в моделі

Як згадувалося раніше, нереально провести повний перепис популяції з метою підрахувати середні значення і вивести функцію популяції регресії. На практиці дослідник обмежується випадковою вибіркою будинків і визначає необхідні параметри з метою отримання вибіркової функції регресії. Табл. 2.2 містить дані за вибіркою з 14 будинків (*T=14*), проданих у районі Сан-Дієго в 1990 р. На графіку (рис. 2.3), побудованому за цими даними, зображені точки (X_t, Y_t). Цей графік називається діаграмою розсіювання даних вибірки. Вона схожа на графік (рис.1.2), але там ми зображали (X_t, Y_t) для популяції в цілому, а графік (рис.2.3) ґрунтується лише на даних вибірки.

Таблиця 2.2

Дійсна й оцінена ціна будинку і його житлова площа у кв. футах

t	*SQFT*	*Дійсна ціна PRICE*	Оцінена середня ціна
1	1065	1999,9	200,386
2	1254	228	226,657
3	1300	235	233,051
4	1577	285	271,554
5	1600	239	274,751
6	1750	293	295,601
7	1800	285	302,551
8	1870	365	312,281
9	1935	295	321,316
10	1948	290	323,123
11	2254	385	265,657
12	2600	505	413,751
13	2800	425	441,551
14	3000	415	469,351

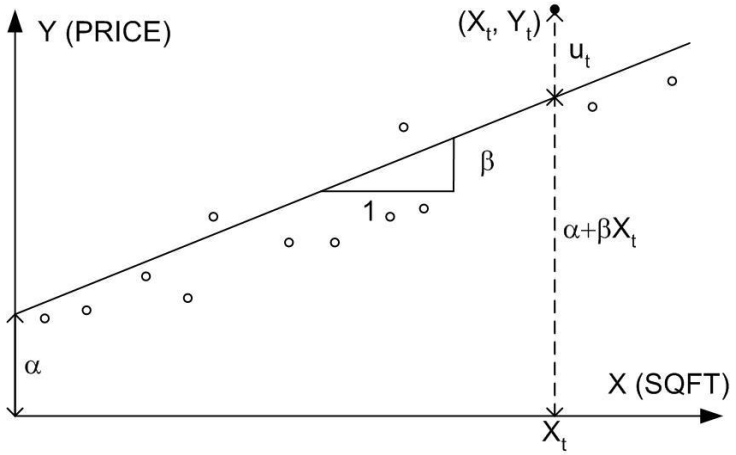


Рис. 2.3. Діаграма вибірки розкиду ціни будинку залежно від його площі

Фундаментальне припущення, на якому основана найпростіша регресійна модель, полягає в тому, що середні значення $(\bar{X},\bar{Y})$ лежать на прямій лінії (позначеній $\alpha+\beta\cdot SQFT$), яка зображає функцію регресії популяції і є умовним середнім значенням (або сподіванням) ціни (*PRICE*) для заданої площі *SQFT*. Нижче наводиться загальне визначення найпростішої лінійної регресійної моделі.

Припущення 2.1 – лінійність моделі:

$$Y_t=\alpha+\beta X_t+u_t,$$

де X_t і Y_t є *t*-ті спостереження (*t* змінюється від 1 до *T*) незалежної і залежної змінних відповідно; α і β – невідомі параметри, які необхідно визначити; u_t – невідома складова, яка передбачається випадковою величиною з певними власти-

востями. α і β мають назву коефіцієнтів регресії. Індекс t можна визначити як тимчасовий чинник у спостереженнях або типове спостереження з таблиці.

Припустимо, що нам відомі величини α і β. Побудуємо пряму $Y = \alpha + \beta X$ на цій діаграмі. Вона зображає функцію регресії популяції. Відхилення за вертикаллю дійсного значення ціни (Y_t) від лінії регресії ($Y = \alpha + \beta X$) є випадкова похибка u_t. Кутовий коефіцієнт прямої (β) є також $\frac{\Delta Y}{\Delta X}$. Він позначає зростання Y при одиничному зростанні X.

β може бути інтерпретований як граничний ефект X на Y. Так, якщо $\beta = 0{,}065$, то кожне збільшення житлової площі на 1 кв. фут призведе до зростання ціни будинку в середньому на 0,065 тис. дол. (зверніть увагу на важливість вибору одиниць вимірювання) або 65 дол. Більш імовірно, що зростання житлової площі на 100 кв. футів призведе до зростання очікуваної середньої ціни на 6 500 дол. Хоча α відповідає середньому значенню Y при X=0, його не можна інтерпретувати як вартість ділянки, на якій стоїть будинок. Пояснення цьому твердженню полягає в тому, що α містить вплив невключених у модель змінних, які, проте, впливають на залежну змінну.

При формулюванні найпростішого лінійного співвідношення між PRICE і SQFT ми ігноруємо той факт, що ціна будинку залежить також від інших характеристик, таких як, наприклад, величина ділянки землі й кількість ванних. Таким чином, ми припускаємо, що ефекти від них абсорбуються залишковим членом u_t. Залишковий член u_t, є, по суті, комбінацією чотирьох різних ефектів:

1. Відповідає за ефект дії змінних, не включених у модель.
2. Поглинає ефекти нелінійності співвідношення між Y і X. Так, якби істинною була модель $Y_t = \alpha + \beta X_t + \gamma X_t^2 + u_t$, а ми б застосовували модель 2.1, то ефект від X_t^2 був би включений в u_t.
3. Містить похибку у вимірюванні X і Y.
4. Включає властиві непередбачувані випадкові ефекти.

Хоча досліджувана нами модель дуже проста і, звичайно, не реалістична, з її допомогою можна легко зрозуміти різні концепції в економетриці. Далі ми розширимо модель на випадок, коли в неї входять більше ніж одна пояснювальна змінна.

Наступною нашою задачею після формулювання моделі є отримання “найкращих” оцінок для α і β. Після того як це буде зроблено, ми перевіримо властивості цих оцінок, а також припущення 2.1. Потім перевіримо гіпотези, що належать до них, і застосуємо оцінену пряму для проведення умовного прогнозу ціни будинку для даного значення X. Але перш ніж ми все це виконаємо, нам необхідно зробити відносно u_t і X_t додаткові припущення. Нижче наведено повний список припущень.

Припущення найпростішої лінійної регресійної моделі

1. Регресійна модель лінійна по відношенню до невідомих коефіцієнтів α і β, тобто $Y_t = \alpha + \beta X_t + u_t$ для *t=1, 2, ..., T.*

2. Залишкова складова u_t – випадкова величина, що має нульове середнє значення, тобто $E(u_t) = 0$.

3. Не всі із спостережуваних значень *X* мають однакову величину, принаймні одне з них відрізняється від інших.

4. X_t задані й невипадкові, тобто вони не корелюються з u_t, отже, $\text{Cov}(X_t, u_t) = E(X_t, u_t) - E(X_t)E(u_t) = 0$.

5. u_t має сталу дисперсію для всіх *t*, тобто $\text{Var}(u_t) = E(u_t^2) = \sigma^2$.

6. u_t і u_s розподілені незалежно для $t \neq s$, звідси $\text{Cov}(u_t, u_s) = E(u_t u_s) = 0$.

7. u_t розподілена нормально, тобто $u_t \sim N(0, \sigma^2)$. Це означає, що для даного X_t $Y_t \sim N(\alpha + \beta X_t, \sigma^2)$.

Розглянемо ці припущення більш детально.

Припущення 2.2 – середні значення залишкових членів нульові

Кожна u_i – випадкова величина з $E(u_i) = 0$.

На рис. 2.3. бачимо, що деякі з точок спостереження лежать вище за лінію $Y = \alpha + \beta X$, а деякі нижче. Це означає, що частина складових u_t позитивні, а інші негативні. Оскільки $Y = \alpha + \beta X$ – лінія середніх значень, доцільно припустити, що ці випадкові відхилення взаємно знищуються в середньому за всією популяцією. Отже, припущення про те, що u_t – випадкові величини з нульовим математичним сподіванням, достатньо реалістичне.

Припущення 2.3 – не всі значення X однакові

Не всі X_i мають однакові значення. Принаймні є два різних значення. Іншими словами, дисперсія вибірки $\frac{1}{T-1}\sum(X_i - \bar{X})^2$ відмінна від нуля.

Це припущення є дуже важливим, оскільки інакше модель не може бути оцінена. На інтуїтивному рівні, якщо X_i не змінюється, то неможливо пояснити, чому змінюється Y_i. Як приклад, припустимо, що Y_t – споживацькі витрати сім'ї в *t*–му місяці, а X_t – дохід сім'ї в тому ж місяці. Звичайно дохід сім'ї від місяця до місяця трохи змінюється, а споживацькі витрати можуть значно варіювати в різні місяці. Якщо дохід не змінюється, то не можна пояснити, чому змінюються витрати на споживання. Це, однак, не означає, що дохід сім'ї не впливає на її споживацькі витрати. Якщо наступного року зарплата підвищиться, то підвищаться й середні витрати. Рис. 2.4 графічно ілюструє припущення 2.3.

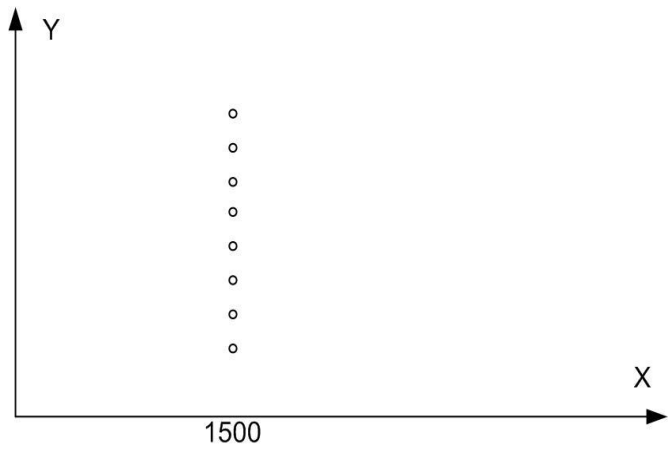


Рис. 2.4. Приклад, у якому величини *X* не змінюються

У разі застосування моделі, що описує вартість будинку залежно від його площі, припустимо, що була зібрана інформація про вартість будинків площею тільки 1500 кв. футів. Діаграма розкиду даних вибірки зображена на рис.2.4. Зрозуміло, що за цією діаграмою неможливо провести адекватну оцінку лінії регресії популяції.

Припущення 2.4 – значення X задані й невипадкові

$E(u_t) = 0$, оскільки X_t задані і, отже, невипадкові, то з цього випливає $\text{Cov}(X_t, u_t) = E(X_t, u_t) - E(X_t)E(u_t) = X_t E(u_t) - X_t E(u_t) = 0$.

З $E(X_t u_t) = 0$ випливає, що коваріація популяції між X_t і u_t дорівнює нулю. Отже, *X* і *u* не корельовані. Як ми побачимо пізніше, це припущення є основоположним для того, щоб метод оцінювання α і β мав деякі бажані властивості. На інтуїтивному рівні, якщо *X* і *u* корельовані, то зі зміною *X* повинна також змінюватися *u*. У цьому випадку очікуване значення *Y* не буде дорівнювати $\alpha + \beta X$.

Припущення 2.5 – гомоскедастичність

Усі випадкові величини u_t мають однаковий розподіл дисперсій σ^2, так що $\text{Var}(u_t) = E(u_t^2) = \sigma^2$. Ця властивість називається гомоскедастичністю (рівнорозкиданістю).

Припущення 2.6 – серійна незалежність

Усі u_t розподілені незалежно, так що $\text{Cov}(u_s, u_t) = E(u_t u_s) = 0$ для всіх $t \neq s$. Ця властивість має назву серійної незалежності.

Виконання цих двох гіпотез приводить до того, що залишкові члени незалежно й однаково розподілені. Згідно з рис.1.2. для даного *X* є розкид значень *Y*, який задає умовний розподіл. Залишок u_t – відхилення від умовного середнього значення $\alpha + \beta X$. Припущення 2.5 має на увазі, що розподіл випадкової величини u_t має ту ж дисперсію (σ^2), що й у u_s для різних *X* спостережень S. Рис. 2.3 – приклад гетероскедастичності (нерівних розкидів), у якому дисперсія непостійна

за спостереженнями. Припущення 2.6 говорить про те, що u_t і u_s незалежні і, отже, не корельовані. Зокрема, послідовні складові не корельовані. Рис.2.5 – приклад серійної кореляції, коли це припущення порушується.

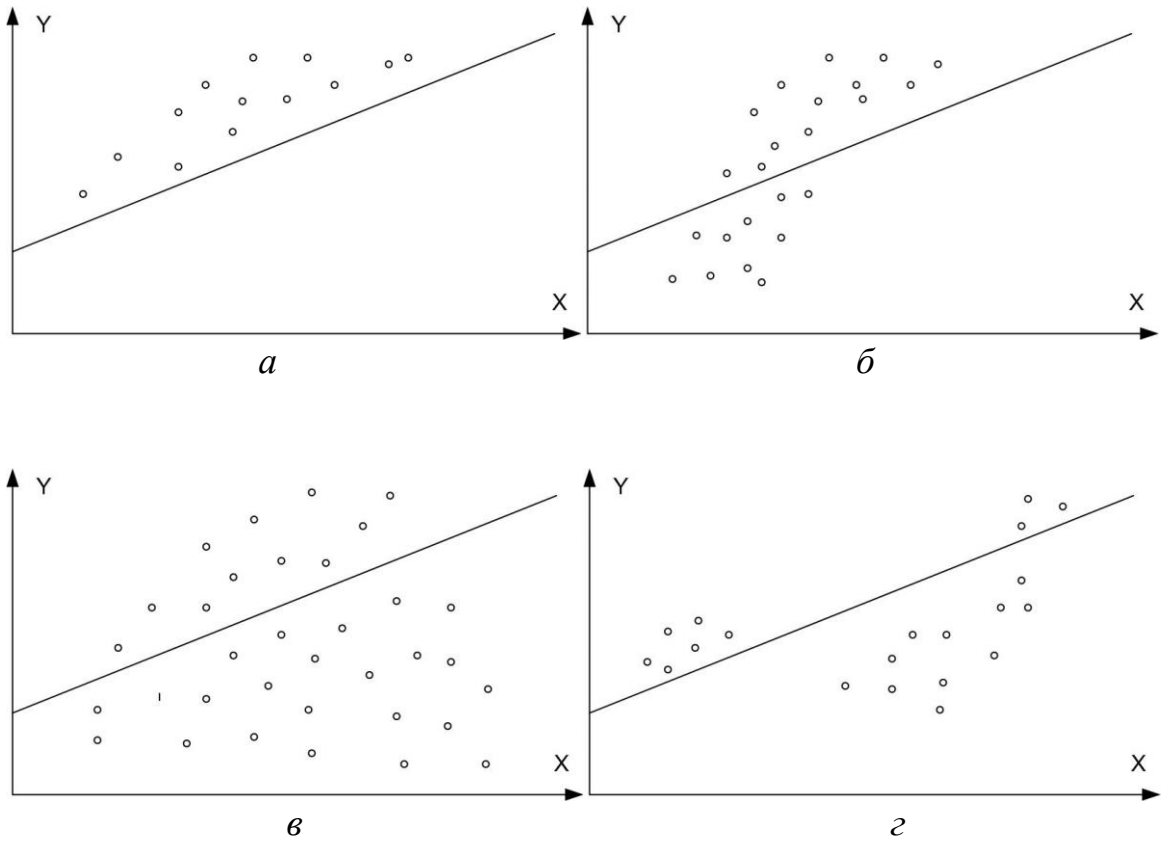


Рис. 2.5. Приклади порушення припущень: *а* – порушення 2.2; *б* – порушення 2.1; *в* – порушення 2.5; *г* – порушення 2.6

Припущення 2.7 – нормальність відхилень

Кожне u_t розподілено нормально згідно із законом $N(0,\sigma^2)$, з чого виходить, що умовна густина розподілу *Y* для заданого значення *X* задається законом $N(\alpha+\beta X,\sigma^2)$.

Таким чином, залишкові складові $u_1, u_2, u_3, ..., u_T$ вважаються розподіленими незалежно і за нормальним законом з нульовим середнім значенням і загальною дисперсією σ^2. Це припущення дуже важливе при висуванні та перевірці гіпотез.

Класична лінійна регресійна модель:
припущення, що лежать в основі методу найменших квадратів

Якби нашою метою було тільки отримання оцінок β_1 і β_2, то все вищезгадане вирішувало б питання. Але ми пам'ятаємо, що в регресійному аналізі нашою метою є не лише отримання β_1 і β_2, а й висновок про істинні величини β_1 і β_2.

Наприклад, нам хотілося б знати, наскільки близькі β_1 і β_2, до їх аналогів за всією популяцією або наскільки близьке Y_i до істинного $E(Y / X_i)$.

Стандартна класична лінійна модель регресії (CLRM)

Дана модель лежить в основі більшості моделей економетрики і заснована на 10 припущеннях. Розглянемо їх на прикладі моделі з двома змінними, а пізніше поширимо на модель із великою кількістю змінних.

Припущення 1 – лінійність регресійної моделі

Регресійна модель лінійна за параметрами, як це бачимо з виразу

$$Y_i = \beta_1 + \beta_2 X_i + u_i .$$

Ми вже обговорювали цю модель. Оскільки лінійність моделі за параметрами – початкова точка CLRM, ми будемо завжди це припускати й надалі. Також пам'ятатимемо, що регресант *Y* і регресор *X* й самі можуть бути нелінійними.

Припущення 2 – величини X фіксовані в повторних вибірках

Значення, що набуваються регресором *X*, вважаються фіксованими в повторних вибірках. Безумовно, вважатимемо, що *X* не стохастична.

Це припущення неявно було присутнє в обговоренні PRF раніше. Дуже важливо розуміти значення «фіксованість» у повторних вибірках. Його можна пояснити на прикладі табл. 2.1. Розглянемо змінну *Y*, відповідну рівням доходу. Зберігаючи величину доходу фіксованою, наприклад на рівні 80 дол., ми випадково вибрали сім'ю з тижневими витратами, скажімо, у 60 дол. Зберігаючи *X=80*, ми випадково вибрали іншу сім'ю з *Y=75*. У кожному з цих випадків (повторні вибірки) значення *X* були фіксовані. Ми можемо повторити цей процес для всіх *X* з таблиці. Це означає, що регресійний аналіз є умовним регресійним аналізом, тобто з умовою заданості величин регресора *X*.

Припущення 3 – рівність нулю середньої величини збурення u_i

Для даного значення величини *X* середня величина або математичне сподівання випадкової збурюючої складової u_i дорівнює нулю:

$$E(u_i / X_i) = 0$$

Це припущення стверджує, що середня величина u_i, відповідна даному X_i, є нуль.

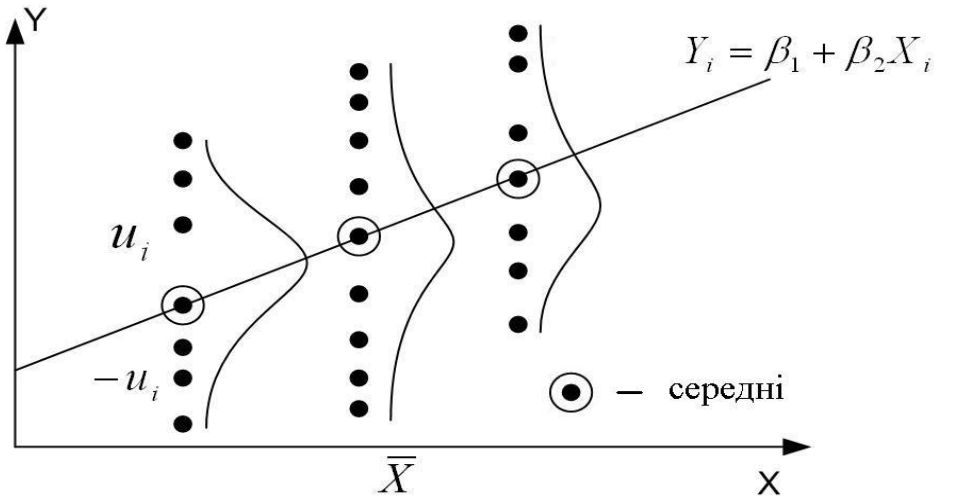


Рис. 2.6 Геометрична інтерпретація припущення 3

Геометричне значення цього припущення показане на рис. 2.6. Як бачимо, при фіксованому значенні *X* відповідні значення *Y* розташовані довкола середніх значень, коли одні зі значень *Y* лежать вище середніх, а інші – нижче середніх. Відстані, вище і нижче за середні, є не що інше як u_i. За формулою (2.2.1), середня величина цих відхилень для фіксованого *X* повинна дорівнювати нулю.

Припущення 4 – гомоскедастичність або рівність дисперсій u_i

Для даного *X* дисперсії u_i однакові для всіх спостережень:

$$D(u_i / X_i) = E[u_i - E(u_i) / X_i]^2 = E(u_i^2 / X_i) = \sigma^2. \qquad (2.2.4)$$

Рівність (2.2.4) стверджує, що дисперсія u_i для кожного X_i є деяка позитивна величина, що дорівнює σ^2. Іншими словами, (2.2.4) означає, що величини *Y*, відповідні різним значенням *X*, мають однакову дисперсію.

На рис. 2.7 показана ситуація, коли ця властивість не виконується, тобто коли має місце рівність

$$D(u_i \mid X_i) = \sigma_i^2.$$

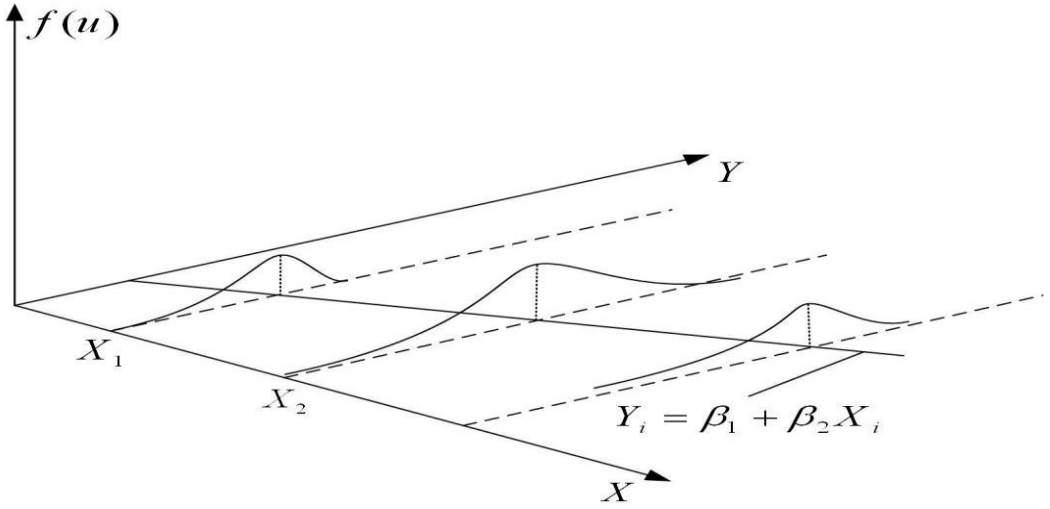


Рис. 2.7. Гомоскедастичність

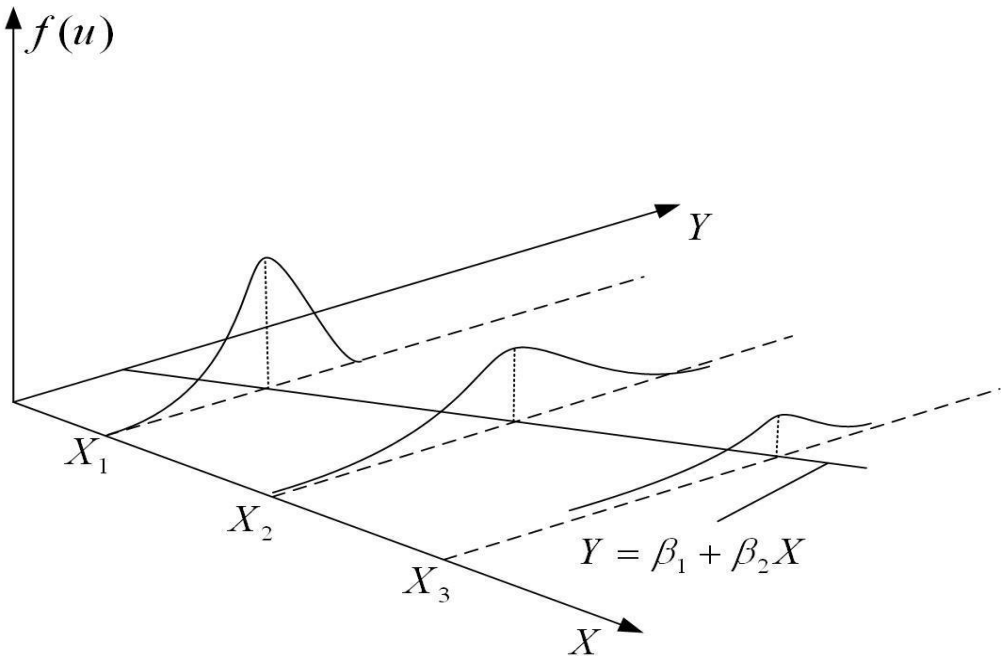


Рис. 2.8. Гетероскедастичність

Щоб краще зрозуміти різницю між двома цими ситуаціями, повернемося до нашого прикладу. Графіки 2.7 і 2.8 показують, що зі зростанням доходу зростають і середні витрати на споживання. Але на рис. 2.7 дисперсія (розкид) витрат однакова для всіх рівнів доходів, а на рис. 2.8 дисперсія витрат збільшується зі зростанням *X*, тобто багаті сім'ї в середньому купують більше, ніж бідні, але в них і більший розкид у витратах, ніж у бідних.

На рис. 2.8 показано, що

$$D(u_1 \mid X_1) < D(u_2 \mid X_2) < \ldots$$

Припущення 5 – відсутність автокореляції між збуреннями

Для будь-яких двох даних величин *X*, X_i і X_j $(i \neq j)$, кореляція між u_i і u_j дорівнює нулю:

$$\operatorname{cov}(u_i, u_j \mid X_i, X_j) = E[u_i - E(u_i) \mid X_i][u_j - E(u_j) \mid X_j] =$$
$$= E(u_i \mid X_i)(u_j \mid X_j) = 0.$$

Припущення 6 – нульова коваріація між u_i і X_i

Коваріація між u_i і X_i дорівнює нулю, тобто

$$\operatorname{cov}(u_i, X_i) = E[u_i - E(u_i)][X_i - E(X_i)] = E[u_i(X_i - E(X_i))] =$$
$$= E(u_i X_i) - E(X_i)E(u_i) = E(u_i X_i) = 0.$$

Це припущення стверджує, що збурення u і пояснювальна змінна *X* не корельовані. Зміст цього полягає в такому: коли ми виразили PRF у вигляді

$$Y_i = \beta_1 + \beta_2 X_i + u_i,$$

то припустили, що *X* і *u* поодинці впливають на *Y*. Але якщо *X* і *Y* корельовані, то неможливо визначити їх індивідуальний вплив на *Y*.

Припущення 6 автоматично виконується, якщо X – невипадкова величина і справедлива гіпотеза 3:

$$E(u_i) = 0:\ \text{cov}(u_i, X_i) = E[X_i - E(X_i)][u_i - E(u_i)] = (X_i - E(X_i))E[u_i - E(u_i)] = 0.$$

Припущення 7 – кількість спостережень N повинна бути більшою, ніж кількість параметрів у моделі, тобто більшою, ніж кількість пояснювальних змінних

У нашому прикладі уявимо, що в табл. 1.1 є лише одна пара значень Y і X.

З цього єдиного спостереження неможливо визначити невідомі параметри β_1 і β_2 моделі. Нам необхідні принаймні дві пари значень для знаходження параметрів β_1 і β_2.

Припущення 8 – мінливість пояснювальних змінних X

Значення змінної X у даній вибірці не повинні бути однаковими. У математичних термінах $D(x)$ повинна бути позитивним числом.

Це припущення не таке нешкідливе, як здається. Розглянемо рівняння

$$\beta_2 = \frac{\sum x_i y_i}{\sum x_i^2}.$$

Якщо всі значення X збігаються, то $X_i = \bar{X}$ і знаменник дробу обертається в нуль, що робить неможливим знаходження параметрів регресії. У нашому випадку зрозуміло, що при малій мінливості X не можна пояснити велику мінливість Y. Потрібно пам'ятати, що мінливість обох змінних X і Y є основоположним моментом регресійного аналізу. Коротше кажучи, змінні повинні змінюватися.

Припущення 9 – регресійна модель вибирається коректно

Інакше кажучи, відсутня помилка зсуву в моделі, що застосовується в емпіричному аналізі.

Ми вже зазначали, що класична регресійна модель передбачається вибраною коректно. Економічні дослідження починаються з вибору економетричної моделі, що лежить в основі предмета дослідження. При цьому повинні бути з'ясовані такі питання: 1) які змінні повинні бути включені в модель? 2) який вигляд функціональної залежності? 3) які припущення імовірності робляться відносно Y_i і X_i і u_i, що входять у модель?

Ці питання важливі , оскільки, як ми побачимо, не включення суттєвих змінних у модель, або неправильний вибір функціональної залежності, або неправильні стохастичні припущення роблять дуже сумнівними отримані при економетричному аналізі результати.

Припустимо, що ми вибрали такі дві моделі для встановлення залежності рівнів інфляції і безробіття:

$$Y_i = \alpha_1 + \alpha_2 X_i + u_i, \qquad (2.2.5)$$

$$Y_i = \beta_1 + \beta_2\left(\frac{1}{X_i}\right) + u_i, \qquad (2.2.6)$$

де Y_i – рівень інфляції, а X_i – рівень безробіття.

Модель (2.2.5) лінійна і за параметрами, і за змінними, а (2.2.6) нелінійна за параметром X_i. На рис. 2.9 зображені ці моделі. Якщо модель (2.2.6) коректна, то модель (2.2.5) дає нам неправильний прогноз: між точками *A* і *B*, для будь-яких значень X_i модель (2.2.5) дає завищені середні значення *Y*, тоді як зліва від *A* або справа від *Y* вона дає занижені середні значення *Y*.

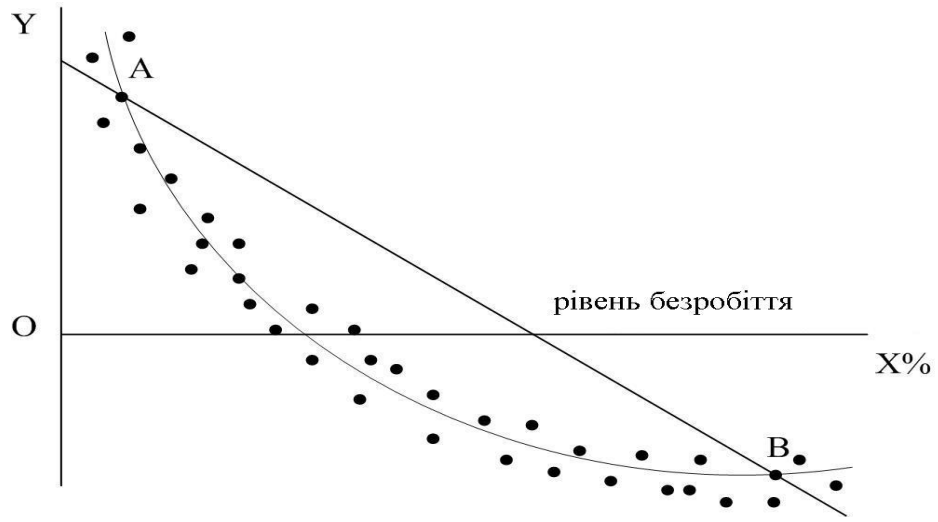


Рис. 2.9. Дві моделі залежності рівнів інфляції та безробіття

Припущення 10 – відсутність точної мультиколінеарності
Тобто між пояснювальними змінними відсутня точна мультиколінеарність.

2.3. Точність або стандартна похибка оцінювачів за МНК

З рівнянь (2.1.15), (2.1.16) зрозуміло, що оцінювачі β_1 і β_2 за МНК – функція від даних вибірки. Але оскільки дані можуть змінюватися від вибірки до вибірки, то змінюватимуться й значення оцінювачів. Отже, необхідна міра «достовірності» або точності обчислення оцінювачів β_1 і β_2. У статистиці точність оцінювача вимірюється його стандартною похибкою. Як зазначено раніше при прийнятих гіпотезах оцінювачі за МНК мають такі значення дисперсій і стандартної похибки:

$$D(\beta_2) = \frac{\sigma^2}{\sum x_i^2}; \tag{2.3.1}$$

$$\sigma(\beta_2) = \frac{\sigma}{\sqrt{\sum x_i^2}}, \tag{2.3.2}$$

$$D(\beta_1) = \frac{\sum X_i^2}{n\sum x_i^2}\sigma^2; \tag{2.3.3}$$

$$\sigma(\beta_1) = \sqrt{\frac{\sum X_i^2}{n\sum x_i^2}}\sigma; \tag{2.3.4}$$

де σ^2 – дисперсія u_i. Усі величини, що входять у (2.3.1)–(2.3.4), за виключенням σ^2, можна підрахувати за даними вибірки. Для самої ж величини σ^2 справедлива формула

$$\sigma^2 = \frac{\sum u_i^2}{n-2}, \tag{2.3.5}$$

де σ^2 – оцінка за МНК істинного значення σ^2. Величина (*N–2*) має назву кількості степенів вільності (df, number degrees freedom). $\sum u_i^2$ – сума квадратів залишків RSS (residual sum squares).

Оскільки $\sum u_i^2$ відома, то σ^2 може бути легко обчислений. Сама величина $\sum u_i^2$ може бути обчислена за формулою

$$\sum u_i^2 = \sum (Y_i - \beta_1 - \beta_2 X_i)^2$$

або

$$\sum u_i^2 = \sum y_i^2 - \beta_2^2 \sum x_i^2 . \tag{2.3.6}$$

Зауважимо, що формула (2.3.6) простіша в застосуванні порівняно з (2.1.2), тому що не вимагає обчислення u_i для кожного *i*, хоча саме таке обчислення виявляється корисним.

Оскільки

$$\beta_2 = \frac{\sum x_i y_i}{\sum x_i^2},$$

з (2.3.6) можна отримати ще один вираз для підрахунку $\sum u_i^2$:

$$\sum u_i^2 = \sum y_i^2 - \frac{\left(\sum x_i y_i\right)^2}{\sum x_i^2}. \tag{2.3.7}$$

Позитивний квадратний корінь з σ^2

$$\sigma = \sqrt{\frac{\sum u_i^2}{n-2}} \tag{2.3.8}$$

називається стандартною похибкою оцінювача. Вона є стандартним відхиленням величин *Y* від оціненої лінії регресії і часто застосовується як сумарна міра “якості підгонки” оціненої лінії регресії.

Відзначимо такі властивості дисперсії (а отже, і стандартної похибки) коефіцієнтів β_1 і β_2.

1. Дисперсія коефіцієнта β_2 прямо пропорційна σ^2, але обернено пропорційна $\sum x_i^2$. Тобто для даного σ^2, чим більша область зміни величин *X*, тим менша дисперсія β_2 і, отже, вища точність, з якою може бути оцінений β_2. Коротше кажучи, при значній зміні величин *X* коефіцієнт β_2 може бути виміряний більш точно, ніж при незначній зміні величин *X*. Крім того, для даної величини $\sum x_i^2$ чим більше σ^2, тим більша дисперсія β_2. Зауважимо, що зі збільшенням обсягу вибірки *N*, кількість членів у $\sum x_i^2$ зростатиме. Тому зі зростанням *N* точність оцінки коефіцієнта β_2 збільшуватиметься.

2. Дисперсія коефіцієнта β_1 прямо пропорційна σ^2 і $\sum X_i^2$, але обернено пропорційна $\sum x_i^2$ і розміру вибірки *N*.

3. Оскільки β_1 і β_2 є оцінками, вони будуть не тільки змінюватися від вибірки до вибірки, але й будуть взаємно залежними в межах однієї вибірки. Ця залежність вимірюється коваріацією між ними. Пізніше буде показано, що

$$\text{cov}(\beta_1, \beta_2) = -\bar{X}D(\beta_2) = -\bar{X}\left(\frac{\sigma^2}{\sum x_i^2}\right). \qquad (2.3.9)$$

Оскільки $D(\beta_2)$ завжди позитивна, коваріація між β_1 і β_2 залежить від знака $\bar{X}$. Якщо $\bar{X} > 0$, тоді, як бачимо з формули (2.3.9), коваріація буде негативна. Так, якщо кутовий коефіцієнт β_2 оцінений із завищенням (тобто лінія регресії підіймається із завищеною крутістю) коефіцієнт перетину β_1 матиме занижене значення (перетин матиме менше значення).

2.4. Властивості оцінювачів за МНК: теорія Гаусса-Маркова

Як було відзначено раніше, оцінювачі, отримані за МНК при зроблених припущеннях CLRM, мають деякі ідеальні або оптимальні властивості. Вони зазначені в добре відомій теоремі Гаусса-Маркова. Для того щоб зрозуміти її значення, необхідно розглянути властивість найкращого лінійного незміщеного оцінювача (best linear unbiasedness property an estimator). Оцінювач, скажімо β_2 за МНК, вважається найкращим лінійним незміщеним оцінювачем BLUE (best linear unbiased estimator) β_2, якщо він має такі властивості:

1. Він лінійний, тобто являє собою лінійну функцію випадкової змінної, таку як залежна змінна *Y* в регресійній моделі.

2. Він є незміщеним оцінювачем, тобто $E(\beta_2) = \beta_2$.

3. Він має найменшу дисперсію в класі всіх лінійних незміщених оцінювачів; незміщений оцінювач з найменшою дисперсією відомий як ефективний оцінювач.

Можна довести, що оцінювачі, отримані за МНК, мають властивості найкращого лінійного незміщеного оцінювача BLUE. Це є висновком відомої теоре-

ми Гаусса-Маркова, яка може бути сформульована таким чином: при прийнятих гіпотезах класичної регресійної лінійної моделі отримані за методом найменших квадратів оцінювачі в класі лінійних незміщених оцінювачів мають найменшу дисперсію, тобто вони є найкращими лінійними незміщеними оцінювачами.

Дана теорема дуже важлива при регресійному аналізі, оскільки стосується як теорії, так і практики.

Пояснимо значення теореми за допомогою рис. 2.10.

На рис. 2.10, *а* показаний розподіл за вибірками оцінювача β_2, отриманого за МНК, у вибірках, що повторюються. Для зручності ми припустимо, що розподіл β_2 розташований симетрично. Як бачимо з рисунка, математичне сподівання β_2 дорівнює істинному значенню β_2, тобто $E(\beta_2) = \beta_2$. Це і є значення, яке ми вкладаємо в термін "незміщена оцінка". На рис. 2.10, *б* показаний розподіл оцінювача β_2^*, отриманого за альтернативним методом. Для зручності ми припустили, що β_2^*, як і β_2, має властивість незміщеності. Припустимо, що і β_2 і β_2^* є лінійними оцінювачами, тобто вони є лінійними функціями від *Y*. Який із оцінювачів – β_2 або β_2^* – слід вибрати?

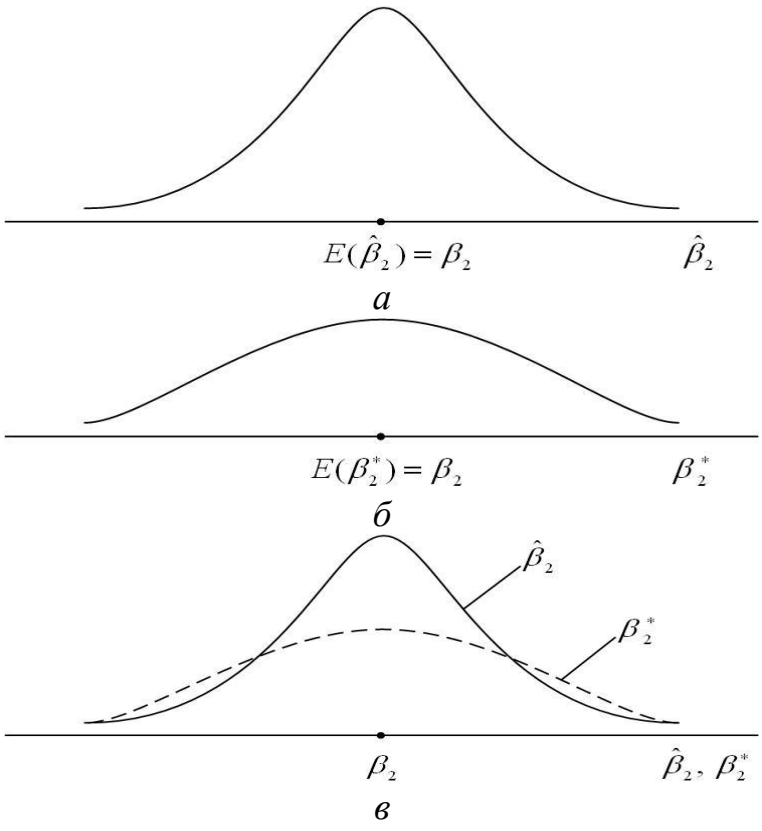


Рис. 2.10. Розподіл за вибіркою за МНК β_2 і альтернативного оцінювача β_2^*: *а* – розподіл β_2; *б* – розподіл β_2^*; *в* – розподіл β_2 і β_2^*

Щоб відповісти на це запитання накладемо два рисунки, як показано на рис. 2.8, *в*. Із рис. 2.10, *в* бачимо, що хоча обидва розподіли незміщені, для β_2^* ми маємо більш розмитий розподіл біля середнього значення в порівнянні з розподілом β_2. Іншими словами, дисперсія β_2^* більша, ніж дисперсія β_2. Зрозуміло, що з двох даних оцінювачів, які мають властивості лінійності і незміщеності, слід вибрати оцінювач із меншою дисперсією, оскільки він ближче до β_2, ніж альтернативний оцінювач. Отже, завжди слід вибирати найкращий лінійний незміщений оцінювач (BLUE).

Розглянуті нами статистичні властивості відомі як властивості кінцевих вибірок. Ці властивості зберігаються незалежно від розміру вибірки, за даними якої отримані оцінювачі. Пізніше ми матимемо нагоду розглянути асимптотичні властивості, тобто властивості, які зберігаються тільки у випадку, коли вибірка дуже велика (нескінченна).

2.5. Коефіцієнт детермінації r^2: міра «якості підгонки»

Звернемося зараз до розгляду питання якості підгонки лінії регресії до множини даних, тобто дослідимо, наскільки «добре» лінія вибіркової регресії підходить до цих даних. Із рис.1.1 бачимо, що якби всі спостереження знаходилися на лінії регресії, ми отримали б "точну підгонку", але на практиці це окремий випадок. У загальному ж випадку будуть як позитивні відхилення u_i, так і негативні. Ми прагнемо, щоб ці залишки були наскільки можливо малі. Коефіцієнт детермінації r^2 (випадок двох змінних) або R^2 (множинна регресія) являє собою сумарну міру якості підгонки лінії регресії до даних спостереження.

Перш ніж з'ясовувати, як підраховується r^2, розглянемо евристичне пояснення r^2 за допомогою графічних діаграм, відомих як діаграма Венна (Venn) (рис. 2.11).

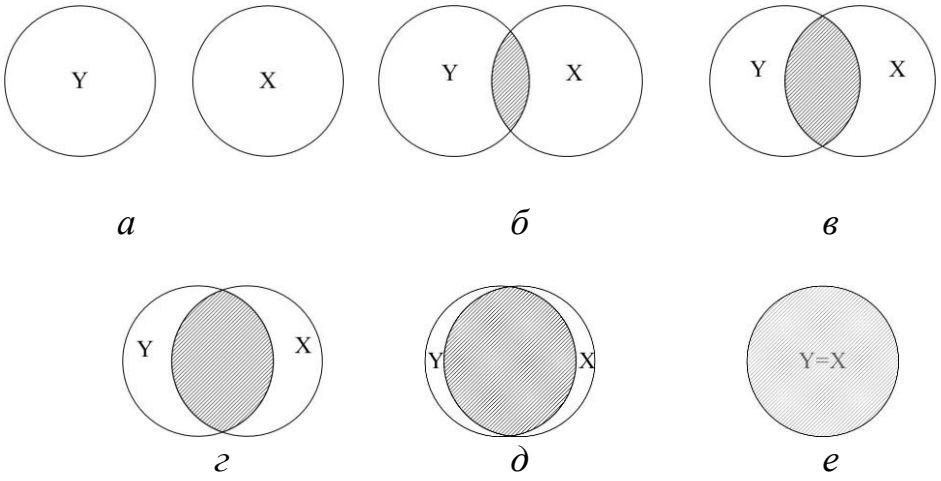


Рис. 2.11. Діаграма для пояснення r^2: *а* – r^2 дорівнює 0; *б* – r^2 близький до 0; *в* – r^2 близький до 0,5; *г* – r^2 більш ніж 0,5; *д* – r^2 близький до 1; *е* – r^2 дорівнює 1

На цій діаграмі коло *Y* зображає дисперсію залежної змінної *Y*, а коло *X* – дисперсію пояснювальної змінної *X*. Перетин двох кіл (заштрихована область) являє собою область, у якій дисперсія *Y* пояснюється дисперсією в *X* (скажімо, за регресією МНК). Чим більша область перетину, тим більше дисперсія *Y* пояснюється за допомогою *X*. Коефіцієнт детермінації r^2 зображає числову міру області перетину. На рис. 2.9 бачимо, що при русі зліва направо область перекриття збільшується, тобто послідовно зростає частина варіації *Y*, з'ясована за допомогою *X*, – r^2 зростає. Коли перекриття немає, r^2, очевидно, дорівнює нулю, а коли відбувається повне перекриття, то $r^2 = 1$, оскільки 100% дисперсії *Y* пояснюється за допомогою *X*. Як незабаром переконаємося r^2 лежить між 0 і 1.

Для обчислення коефіцієнта r^2 зробимо так. Пригадаємо, що

$$Y_i = Y_i + u_i$$

або у формі відхилень

$$y_i = y_i + u_i . \qquad (2.5.1)$$

Підносячи обидві частини цієї рівності у квадрат і підсумовуючи, отримаємо

$$\sum y_i^2 = \sum y_i^2 + \sum u_i^2 + 2\sum y_i x_i = \sum y_i^2 + \sum u_i^2 = \beta_2^2 \sum x_i^2 + \sum u_i^2 , \qquad (2.5.2)$$

оскільки $\sum y_i u_i = 0$ і $y_i = \beta_2 x_i$.

Різні суми квадратів, що входять у (2.5.2), можуть бути описані таким чином: $\sum y_i^2 = \sum (Y_i - \bar{Y})^2$ – загальна дисперсія величини *Y* відносно середньої величини за вибіркою, що називається загальною сумою квадратів TSS (total sum squares); $\sum y_i^2 = \sum (Y_i - \bar{Y})^2 = \sum (Y_i - \bar{Y})^2 = \beta_2^2 \sum x_i^2$ – дисперсія оціненої величини *Y* щодо її середнього значення ($\bar{Y} = \bar{Y}$) або з'ясована сума квадратів з рівняння регресії ESS (explained sum squares); $\sum u_i^2$ – залишкова або нез'ясована дисперсія величини *Y* щодо лінії регресії або просто залишкова сума квадратів RSS (residual sum squares). Таким чином, з (2.5.2) одержуємо рівність

$$\text{TSS=ESS+RSS.} \qquad (2.5.3)$$

Вона показує, що загальна варіація спостережуваних величин *Y* щодо їх середнього значення може бути розбита на дві частини, одна відповідає лінії регресії, а інша – випадковим відхиленням, оскільки не всі спостережувані *Y* лежать на лінії регресії. На рис. 2.10 це розбиття пояснене геометрично.

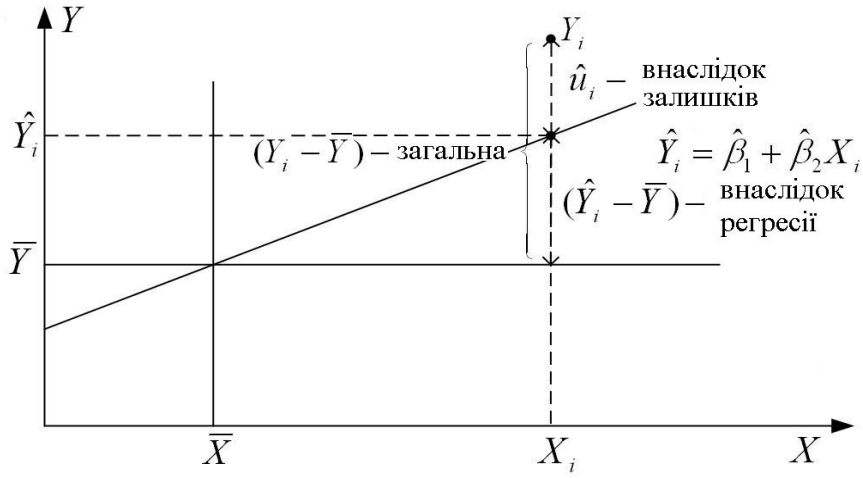


Рис. 2.12. Розбиття варіації Y_i на дві компоненти

Розділивши обидві частини (2.5.3) на TSS, одержуємо

$$1 = \frac{ESS}{TSS} + \frac{RSS}{TSS} = \frac{\sum(Y_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} + \frac{\sum u_i^2}{\sum(Y_i - \bar{Y})^2}. \qquad (2.5.4)$$

Визначимо тепер коефіцієнт детермінації r^2 таким способом:

$$r^2 = \frac{ESS}{TSS} = \frac{\sum(Y_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} \qquad (2.5.5)$$

або в альтернативному вигляді

$$r^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum u_i^2}{\sum(Y_i - \bar{Y})^2}. \qquad (2.5.5a)$$

Визначена таким чином величина r^2, відома як коефіцієнт детермінації, і є мірою якості підгонки лінії регресії, що широко застосовується.

Відзначимо такі дві властивості r^2:

1. Коефіцієнт r^2 не негативний (випливає з виразу (2.5.5)).

2. r^2 має межі $0 \le r^2 \le 1$. При значенні $r^2 = 1$ ми маємо випадок точної підгонки, тобто $Y_i = Y_i$ для кожного *i*. Водночас випадок $r^2 = 0$ означає відсутність зв'язку між регресантом і регресором (тобто $\beta_2 = 0$, $Y_i = Y_i$ для всіх *i*). У цьому випадку, як бачимо з рівняння $Y_i = Y_i + \beta_2(X_i - \bar{X})$, $Y_i = Y_i$, тобто кращим прогнозом для будь-якої величини *Y* є її середнє значення. При цьому лінія регресії – паралель осі *X*.

Хоча r^2 можна обчислити безпосередньо за формулами (2.5.5), (2.5.5a), простіше скористатися такими формулами:

$$r^2 = \frac{ESS}{TSS} = \frac{\sum y_i^2}{\sum y_i^2} = \frac{\beta_2^2 \sum x_i^2}{\sum y_i^2} = \beta_2^2 \left(\frac{\sum x_i^2}{\sum y_i^2} \right). \qquad (2.5.6)$$

Розділивши чисельник і знаменник (2.5.6) на розмір вибірки N (або $N–1$, якщо розмір вибірки малий), одержуємо

$$r^2 = \beta_2^2 \left(\frac{S_x^2}{S_y^2} \right), \tag{2.5.7}$$

де S_y^2 і S_x^2 – вибіркові дисперсії (sample variances) за Y і X відповідно.

Оскільки $\beta_2 = \frac{\sum x_i y_i}{\sum x_i^2}$, рівняння (2.5.6) можна зобразити у вигляді

$$r^2 = \frac{\left(\sum x_i y_i\right)^2}{\sum x_i^2 \sum y_i^2}. \tag{2.5.8}$$

Застосовуючи вирази для r^2, ми можемо подати ESS і RSS таким чином:

$$ESS = r^2 TSS = r^2 \sum y_i^2, \tag{2.5.9}$$

$$RSS = TSS - ESS = TSS\left(1 - \frac{ESS}{TSS}\right) = (1 - r^2)\sum y_i^2. \tag{2.5.10}$$

Отже, ми можемо записати

$$TSS = ESS + RSS,$$

$$\sum y_i^2 = r^2 \sum y_i^2 + (1 - r^2)\sum y_i^2. \tag{2.5.11}$$

Коефіцієнт кореляції, що являє собою ступінь асоціативності між двома змінними, кількісно близько пов’язаний з r^2, але концептуально вони дуже різні. Коефіцієнт кореляції можна визначити за формулою

$$r = \pm\sqrt{r^2} \tag{2.5.12}$$

або

$$r = \frac{\sum x_i y_y}{\sqrt{\sum x_i^2 \sum y_i^2}} = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{\sqrt{\left[n\sum X_i^2 - \left(\sum X_i\right)^2\right]\left[n\sum Y_i^2 - \left(\sum Y_i\right)^2\right]}}. \tag{2.5.13}$$

Визначена таким чином величина має назву коефіцієнта кореляції за вибіркою.

Ось деякі його властивості:

1. Він може бути позитивним або негативним, знак r залежить від знака чисельника (2.5.13), що є мірою коваріації за вибіркою двох змінних.

2. Він лежить між –1 і 1, тобто $-1 \le r \le 1$.

3. За своєю природою він симетричний, тобто коефіцієнт кореляції між X і Y (r_{XY}) той же, що й між Y і X (r_{YX}).

4. Він незалежний по відношенню до вибору початку системи координат і масштабу вздовж осей координат, тобто, якщо ми визначимо $X_i^* = aX_i + C$ і

$Y_i^* = bY_i + d$, де $a > 0$, $b > 0$, a, b і d – константи, то r між X^* і Y^* те ж, що й між початковими змінними X і Y.

5. Якщо X і Y статистично незалежні, коефіцієнт кореляції між ними дорівнює нулю, але якщо $r = 0$, це не означає, що дві змінні незалежні. Іншими словами, нульовий коефіцієнт кореляції не обов'язково означає незалежність (рис. 2.13*з*).

6. Коефіцієнт кореляції є міра тільки лінійної асоціативності або лінійної залежності; він незастосовний для опису нелінійної залежності. Так, на рис. 2.13, *з* $Y=X^2$ є точна залежність, хоча $r = 0$.

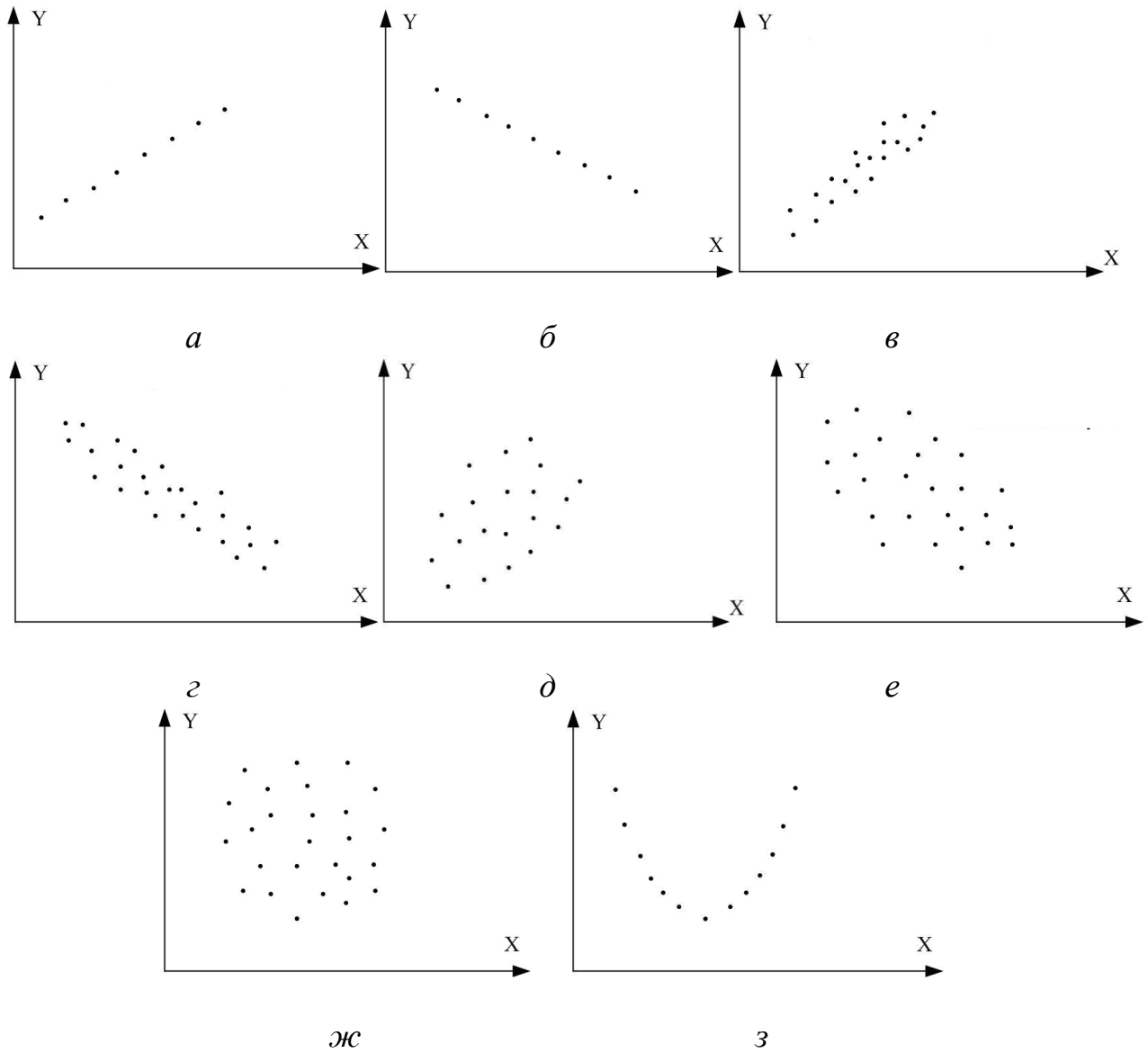


Рис. 2.13. Кореляційний коефіцієнт для різних випадків вибірок:
а – r=1; *б* – r=-1; *в* – r близький до 1; *г* – r близький до –1;
д – r додатній, близький до 0; *е* – r від'ємний, близький до 0; *ж* – r=0; *з* – r=0

7. Хоча r є мірою лінійної асоціативності між двома змінними, це необов'язково означає який-небудь причинно-наслідковий зв'язок, як було відзначено раніше.

У контексті регресії r^2 більш інформативний, ніж r, оскільки r^2 вказує на частку варіації в залежній змінній, що з'ясовується пояснювальною змінною.

Насамкінець зауважимо, що коефіцієнт детермінації r^2 може бути обчислений як квадрат коефіцієнта кореляції між змінними Y_i і Y_i за такою формулою:

$$r^2 = \frac{\left[\sum (Y_i - \bar{Y})(Y_i - \bar{Y})\right]^2}{\sum (Y_i - \bar{Y})^2 \sum (Y_i - \bar{Y})^2}$$

або

$$r^2 = \frac{\left(\sum y_i y_i\right)^2}{\sum y_i^2 \sum y_i^2}.$$

2.6. Числовий приклад

Проілюструємо теорію економічного аналізу на прикладі функції споживання Кейнса. Пригадаємо, за Кейнсом, "фундаментальним психологічним законом є те, що чоловіки (жінки) налаштовані, як правило, в середньому, збільшувати обсяг споживаних благ у міру зростання свого доходу, але в меншій мірі, ніж збільшується дохід, тобто гранична схильність до споживання (MPC) більше нуля, але менше одиниці". Хоча Кейнс не вказує точний вид функціональної залежності між споживанням і доходом, для простоти припустимо, що співвідношення лінійне

$$Y_i = \beta_1 + \beta_2 X_i + u_i.$$

Для перевірки теорії Кейнса скористаємося даними вибірки, наведеними в табл. 1.3. Оцінка лінії регресії (рис. 2.14), отже, має вигляд

$$Y_i = 24{,}4545 + 0{,}5091 X_i.$$

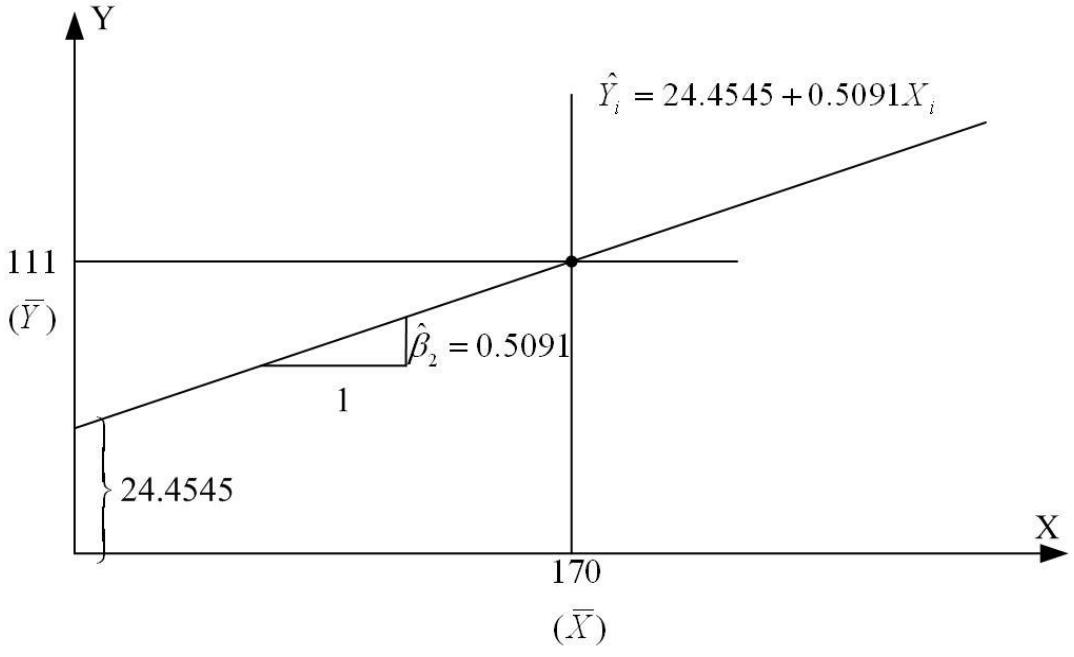


Рис. 2.14. Лінія вибіркової регресії

$$\beta_1 = 24,4545,\ \ D(\beta_1) = 41,370,\ \ \sigma(\beta_1) = 6,4138,$$
$$\beta_2 = 0,5091,\ \ D(\beta_2) = 0,0013,\ \ \sigma(\beta_2) = 0,0357,$$
$$\text{cov}(\beta_1, \beta_2) = -0,2172,\ \ \sigma^2 = 42,1591,$$
$$r^2 = 0,9621,\ \ r = 0,9809,\ \ df = 8. \qquad (2.6.1)$$

Оцінена лінія регресії має такий вигляд:

$$Y_i = 24,4545 + 0,5091X_i. \qquad (2.6.2)$$

Відповідно до висловленого вище отримані результати можна інтерпретувати таким чином. Кожна точка на лінії регресії являє собою оцінку очікуваної або середньої величини *Y*, відповідної вибраному значенню *X*; тобто Y_i – оцінка $E(Y \mid X_i)$. Величина $\beta_2 = 0,5091$, визначаюча кутовий коефіцієнт лінії регресії показує, що для вибірки з областю зміни *X* доходу за місяць між 80 дол. і 260 дол. зі зростанням *X* на 1 дол. оцінка збільшення середніх витрат сім'ї на споживання складає близько 51 цента. Величина $\beta_1 = 24,4545$, визначаюча точку перетину лінії регресії з віссю *Y*, означає середні граничні витрати сім'ї, що має нульовий рівень доходів. Звичайно, це суто механічна інтерпретація коефіцієнта β_1. У регресійному аналізі подібна інтерпретація не завжди відповідає значенню задачі, хоча в нашому прикладі на користь подібного трактування можна сказати, що сім'я без доходу (через безробіття, скорочення виробництва та под.) може підтримувати деякий мінімальний рівень споживання або за рахунок позичання грошей, або використовуючи заощадження. Але в загальному випадку при інтерпретації значення коефіцієнта β_1 потрібно керуватися здоровим глуздом, оскільки часто область зміни *X* може не включати нуль як одну зі спостережуваних величин.

Можливо, краще інтерпретувати β_1 як середню величину впливу на *Y* всіх змінних, не включених явно в модель. Величина $r^2 = 0,9621$ означає, що близько 96% дисперсії тижневих витрат на споживання пояснюється за рахунок доходу. Оскільки r^2 може набути найбільшого значення 1, то можна сказати, що якість лінії регресії дуже добра. Коефіцієнт кореляції $r = 0,9809$ показує, що дві змінні, споживацькі витрати і дохід, високопозитивно корельовані.

2.7. Ілюстративні приклади

Споживання кави в США в 1970–1980 рр.

Розглянемо дані, наведені в табл. 2.3.

Таблиця 2.3

Споживання кави в США (*Y*) та середня роздрібна ціна на каву (X*)* у 1970–1980 рр.

Рік	Y (кількість чашок за день, випитих однією людиною)	X (ціна за фунт кави, дол.)
1970	2,57	0,77
1971	2,50	0,74
1972	2,35	0,72
1973	2,30	0,73
1974	2,25	0,76
1975	2,20	0,75
1976	2,11	1,08
1977	1,94	1,81
1978	1,97	1,39
1979	2,06	1,20
1980	2,02	1,17

З мікроекономіки відомо, що попит на товар в основному залежить від ціни цього товару, ціни товарів, що конкурують з ним або замінюють його, а також доходу покупця. Для того щоб включити всі ці змінні у функцію попиту, нам знадобилася б множинна регресійна модель. До цього ми поки не готові. Тому побудуємо модель функції попиту, що включає залежність попиту лише від ціни товару. Решта величин, що мають відношення до цієї моделі, вважатимемо фіксованими. Тоді, застосовуючи розглянуту раніше двовимірну регресійну модель, отримаємо такі результати:

$$\begin{gathered} Y_t = 2{,}6911 - 0{,}4795 X_t, \\ D(\beta_1) = 0{,}0148, \ \ \sigma(\beta_1) = 0{,}1216, \\ D(\beta_2) = 0{,}0129, \ \ \sigma(\beta_2) = 0{,}0114, \ \ \sigma^2 = 0{,}01656, \\ r^2 = 0{,}6628, \ \ r = -0{,}8141. \end{gathered} \qquad (2.7.1)$$

Інтерпретація отриманих результатів така. Якщо середня роздрібна ціна за фунт кави виросте на 1 дол., то середнє споживання чашок кави в день одна людина зменшить на 0,5 чашки. Якби ціна кави дорівнювала нулю, то середнє споживання кави однією людиною в день склало б близько 2,7 чашок. Звичайно, як було зазначено раніше, дуже часто ми не можемо дати фізичної інтерпретації коефіцієнту β_1. Зверніть увагу, що навіть при нульовій ціні на каву люди не стали б надмірно її споживати через негативний вплив кофеїну на здоров'я. Величина r^2 означає, що близько 66% дисперсії щоденного споживання кави пояснюється дисперсією її роздрібної ціни.

Наскільки ця модель є реалістичною? Зауваживши, що вона не включає всіх змінних, які стосуються її, ми не можемо сказати, що отримали остаточну функцію попиту на каву. Пізніше ми розглянемо більш реальну модель попиту на каву.

Функція споживання Кейнса для США в період 1980–1991 рр.

Таблиця 2.4

Дані по США *Y* (особисті витрати на споживання) і *X* (валовий внутрішній продукт) (ціни 1987 р.)

Рік	*Y*, млн. дол.	*X*, млн. дол.
1980	2,447,1	3,776,3
1981	2,476,9	3,843,1
1982	2,503,7	3,760,3
1983	2,619,4	3,906,6
1984	2,746,1	4,148,5
1985	2,865,8	4,279,8
1986	2,969,1	4,404,5
1987	3,052,2	4,539,5
1988	3,162,4	4,718,6
1989	3,223,3	4,838,0
1990	3,260,4	4,877,5
1991	3,240,8	4,821,0

На основі даних табл. 2.4 може бути проведена така оцінка за МНК для *Y* і *X*:

$$Y_t = -231{,}80 + 0{,}71943 X_t,$$
$$\sigma(\beta_1) = 0{,}9453, \quad \sigma(\beta_2) = 0{,}02175, \qquad (2.7.2)$$
$$r^2 = 0{,}9909, \; r = 0{,}9954.$$

Із цих результатів випливає, що в період 1980–1991 рр. середні споживацькі витрати зростали на 72 центи з одного долара в прирості ВВП, тобто гранична схильність до споживання (MPC) дорівнювала приблизно 72 центам. Якщо інтерпретувати буквально величину 232, що визначає перетин лінії регресії з віссю *Y*, то можна сказати, що при нульовому ВВП витрати на споживання склали б 232 млн дол. Ще раз вкажемо на відсутність економічного значення в подібному трактуванні, оскільки в нашому випадку інтервал зміни *X* не містить цього значення. Величина $r^2 = 0.9909$ означає, що ВВП пояснює близько 99% дисперсії середніх споживацьких витрат. Це дуже високе значення коефіцієнта детермінації. Однак постає питання, чи дійсно найпростіша модель функції споживання підходить для пояснення сукупних витрат на споживання в США. Виявляється, що іноді дуже проста (двовимірна) регресійна модель може дати корисну інформацію. Оцінки MPC для США, проведені на основі складних моделей, також показують, що MPC дорівнює приблизно 0,7.

3. ІНТЕРВАЛЬНІ ОЦІНКИ І ПЕРЕВІРКА ГІПОТЕЗ

Оцінювання та перевірка гіпотез – дві головні гілки класичної статистики. Теорія оцінок (оцінювання) складається з двох частин: точкові оцінки й інтервальні оцінки. Точкові оцінки вже обговорювалися в попередніх розділах. У даному розділі будуть розглянуті інтервальні оцінки, а потім перевірка гіпотез, яка безпосередньо пов'язана з інтервальними оцінками.

3.1. Інтервальні оцінки: основні ідеї

Для демонстрації підходу розглянемо гіпотетичний приклад «сімейний дохід – витрати на споживацькі товари», наведений раніше. Рівняння (2.6.2) показує, що гранична схильність до споживання (MPC, marginal propensity to consume) β_2 дорівнює 0,5091, має точкове значення невідомого MPC β_2. Наскільки ця оцінка достовірна? Як зазначено в розд. 2, унаслідок вибіркових флуктуацій точкове значення швидше за все відрізняється від істинного β_2, хоча при повторних вибірках її математичне сподівання дорівнює β_2, оскільки $E(\beta_2) = \beta_2$. У статистиці достовірність точкового значення вимірюється її стандартною помилкою. Отже, замість того, щоб покладатися лише на точкове значення, ми можемо побудувати інтервал навкруги точки оцінювання, скажімо у дві або три стандартні помилки вліво і вправо від точки, так що цей інтервал містить, скажімо, з імовірністю 95% істинне значення параметра. У цьому полягає ідея інтервальної оцінки.

Припустимо, що ми хочемо знайти наскільки "близько" розташована β_2 до β_2. Для цього спробуємо знайти такі два позитивних числа δ і α, останнє знаходиться між 0 і 1, що випадковий інтервал $(\beta_2 - \delta,\ \beta_2 + \delta)$ містить істинне значення β_2 з імовірністю $1-\alpha$. У математичних позначеннях це записується як

$$\Pr(\beta_2 - \delta \le \beta_2 \le \beta_2 + \delta) = 1 - \alpha . \qquad (3.1.1)$$

Такий інтервал, якщо він, звичайно, існує, має назву довірчого інтервалу; $1-\alpha$ називається довірчим коефіцієнтом, а α $(0 < \alpha < 1)$ відомий як рівень (ступінь) значимості (level significance). Кінці довірчого інтервалу називаються межами довіри (confident limits): $(\beta_2 - \delta)$ – нижня межа довіри і $(\beta_2 + \delta)$ – верхня межа довіри. На практиці α і $(1-\alpha)$ часто виражаються у відсотках.

Рівняння (3.1.1) показує, що на противагу точковій оцінці інтервальна оцінка є інтервалом, побудованим таким чином, що він містить з певною імовірністю $(1-\alpha)$ істинне значення параметра. Наприклад, якщо $\alpha = 0{,}05$ або 5%, (3.1.1) свідчить таке: імовірність того, що випадковий інтервал у (3.1.1) містить істинне значення β_2, дорівнює 0,95 або 95%. Інтервальна оцінка дає, таким чином, інтервал, у якому може лежати істинне значення β_2.

Важливо знати такі аспекти інтервальної оцінки:

1. Рівняння (3.1.1) не свідчить про те, що імовірність того, що β_2 лежить між фіксованими межами, дорівнює $1-\alpha$, оскільки β_2, хоч воно і невідоме, не фіксоване число, не можна сказати, лежить воно усередині інтервалу чи ні. Формула (3.1.1) стверджує, що застосовуючи описаний вище метод, імовірність побудови інтервалу, який містить β_2, дорівнює $1-\alpha$.

2. Інтервал (3.1.1) – випадковий, тобто він змінюватиметься (варіюватиметься) від однієї вибірки до іншої, оскільки в його основі лежить β_2, що є випадковим.

3. Оскільки довірчий інтервал є випадковим, твердження імовірності, що стосується нього, слід розуміти в значенні тривалого повторення вибірок. Більш точно (3.1.1) стверджує: якщо в повторних вибірках довірчі інтервали, подібні побудованому, отримані в результаті великої кількості вибірок на базі ймовірності $1-\alpha$, то в середньому такі інтервали будуть містити в $1-\alpha$ випадках істинне значення параметра β_2.

4. Як зазначено в пункті 2, інтервал (3.1.1.) випадковий, поки β_2 невідомий. Але якщо ми взяли конкретну вибірку і отримали конкретне числове значення β_2, то інтервал (3.1.1) більше не випадковий, а фіксований. У цьому випадку ми не можемо стверджувати (3.1.1) про імовірність, тобто ми не можемо сказати, що з імовірністю $1-\alpha$ фіксований інтервал містить істинне значення β_2. При цьому β_2 або лежить усередині інтервалу, або поза ним. Отже, імовірність цієї події дорівнює 1 або 0. Так, для нашого гіпотетичного прикладу “споживання-дохід” якщо 95%-й довірчий інтервал був отриманий як $(0,4268 \le \beta_2 \le 0,5914)$, ми не можемо сказати, що з імовірністю 95% цей інтервал містить істинне β_2. Імовірність становить 1 або 0.

Як будується довірчий інтервал? З огляду на вищезазначене можна сподіватися, що якщо розподіл імовірності оцінювачів відомий, то можна побудувати довірчий інтервал (3.1.1). Ми знаємо, що при припущенні про нормальність закону розподілу u_i отримані за МНК оцінки β_1 і β_2 також нормально розподілені, а оцінка за МНК σ^2 пов’язана з розподілом χ^2. Покажемо, що процедура побудови довірчого інтервалу – нескладна задача.

3.2 Довірчі інтервали для регресійних коефіцієнтів β_1 і β_2

Як було відзначено раніше, при припущенні про нормальний закон розподілу залишків u_i отримані за МНК оцінки β_1 і β_2 самі розподілені за нормальним законом з відомими математичними сподіваннями і дисперсіями:

$$\beta_1 \sim N(\beta_1, \sigma^2_{\beta_1}), \quad \sigma^2_{\beta_1} = \frac{\sum X_i^2}{n\sum x_i^2}\sigma^2,$$

$$\beta_2 \sim N(\beta_2, \sigma^2_{\beta_2}),\ \ \sigma^2_{\beta_2} = \frac{\sigma^2}{n\sum x_i^2}.$$

Отже, наприклад, змінна

$$Z = \frac{\beta_2 - \beta_2}{\sigma(\beta_2)} = \frac{(\beta_2 - \beta_2)\sqrt{\sum x_i^2}}{\sigma} \qquad (3.2.1)$$

є стандартизована нормальна змінна. Якщо σ^2 відома, то важливою властивістю нормально розподіленої величини зі сподіванням μ і дисперсією σ є те, що площа під густиною розподілу між $\mu \pm \sigma$ становить 68%, між $\mu \pm 2\sigma$ – близько 95%, а між $\mu \pm 3\sigma$ – близько 99,7%.

На практиці σ^2 відома рідко і замінюється її незміщеною оцінкою σ^2. Якщо замінити в (3.2.1) σ на σ, то можна отримати

$$t = \frac{\beta_2 - \beta_2}{\sigma(\beta_2)} = \frac{(\beta_2 - \beta_2)\sqrt{\sum x_i^2}}{\sigma},\ \ \sigma^2 = \frac{\sum u_i^2}{n-2}, \qquad (3.2.2)$$

де $\sigma(\beta_2)$ – стандартна помилка оцінювача β_2. Можна показати, що визначена таким чином змінна *t* розподілена за законом розподілу Ст'юдента з *N–2* степенями вільності. Слід звернути увагу на різницю між (3.2.1) і (3.2.2). Отже, замість того щоб застосовувати нормальний розподіл, ми можемо застосовувати розподіл Ст'юдента для побудови довірчого інтервалу величини β_2:

$$\Pr(-t_{\alpha/2} \le t \le t_{\alpha/2}) = 1 - \alpha, \qquad (3.2.3)$$

де *t* визначається за формулою (3.2.2), а $t_{\alpha/2}$ є величиною розподілу Ст'юдента для $\alpha/2$ рівня значимості і *N–2* степеня вільності. Вона часто називається критичною величиною при $\alpha/2$ рівні значимості. Підстановка (3.2.2) у (3.2.3) дає рівність

$$\Pr\left(-t_{\alpha/2} \le \frac{\beta_2 - \beta_2}{\sigma(\beta_2)} \le t_{\alpha/2}\right) = 1 - \alpha. \qquad (3.2.4)$$

Перетворюючи (3.2.4), одержуємо

$$\Pr\left[\beta_2 - t_{\alpha/2}\sigma(\beta_2) \le \beta_2 \le \beta_2 + t_{\alpha/2}\sigma(\beta_2)\right] = 1 - \alpha. \qquad (3.2.5)$$

Ця рівність являє собою $100(1-\alpha)\%$-й довірчий інтервал для β_2, який може бути записаний більш компактно у вигляді

$$\beta_2 \pm t_{\alpha/2}\sigma(\beta_2). \qquad (3.2.6)$$

За аналогією з цим, застосовуючи (3.2.1) і (3.2.2), ми можемо записати довірчий інтервал для β_1:

$$\Pr\left[\beta_1 - t_{\alpha/2}\sigma(\beta_1) \le \beta_1 \le \beta_1 + t_{\alpha/2}\sigma(\beta_1)\right] = 1 - \alpha \qquad (3.2.7)$$

або більш компактно

$$\beta_1 \pm t_{\alpha/2}\sigma(\beta_1). \qquad (3.2.8)$$

Відзначимо важливу межу довірчих інтервалів (3.2.6) і (3.2.8). В обох випадках довжина довірчого інтервалу пропорційна стандартній помилці оцінювачів. Тобто чим більша стандартна помилка, тим більша довжина довірчого інтервалу. Інакше кажучи, чим більша стандартна помилка оцінювача, тим більша невизначеність оцінки істинного значення параметра. Так, стандартна помилка оцінювача часто описується як міра точності оцінювача, тобто наскільки точно оцінювач вимірює дійсний параметр генеральної сукупності.

Повертаючись до нашого ілюстрованого прикладу моделі "споживання-дохід", нагадаємо, що ми знайшли $\beta_2 = 0,5091$, $\sigma(\beta_2) = 0,0357$ і *df*=8. Якщо ми покладемо $\alpha = 5\%$, тобто 95%-й довірчий коефіцієнт, тоді за таблицею розподілу Ст'юдента знаходимо критичне $t_{\alpha/2} = t_{0,025} = 2,306$. Підставляючи цю величину в (3.2.5), можна перевірити, що 95%-й довірчий інтервал для β_2 буде

$$0,4268 \le \beta_2 \le 0,5914. \qquad (3.2.9)$$

Або, застосовуючи (3.2.6),

$$0,5091 \pm 2,306(0,0357),$$

тобто

$$0,5091 \pm 0,0823. \qquad (3.2.10)$$

Інтерпретація цього довірчого інтервалу така: для даного 95%-го довірчого коефіцієнта при довготривалій вибірці в 95 зі 100 випадків інтервали типу (0,4268; 0,5914) міститимуть істинне β_2. Але, як ми попереджали раніше, зауважимо, що не можна говорити про імовірність у 95% того, що специфічний інтервал (0,4268; 0,5914) містить істинне β_2, оскільки цей інтервал фіксований; отже, β_2 або лежить у ньому, або ні. Таким чином, імовірність знаходження β_2 у фіксованому інтервалі дорівнює або 1, або 0.

Згідно з (3.2.7) можна перевірити, що 95%-й довірчий інтервал для β_1 в нашому прикладі буде

$$9,6643 \le \beta_1 \le 39,2448, \qquad (3.2.11)$$

або, використовуючи (3.2.8), ми знаходимо

$$24,4545 \pm 2,306 \times 6,4138,$$

тобто

$$24,4545 \pm 14,7902. \qquad (3.2.12)$$

Ще раз нагадаємо про правильну інтерпретацію цього довірчого інтервалу.

3.3. Довірчий інтервал для σ^2

Як було зазначено в розд. 2, при припущенні про нормальність розподілу u_i змінна

$$\chi^2 = (n-2)\frac{\sigma^2}{\sigma^2} \qquad (3.3.1)$$

розподіляється за законом розподілу χ^2 з (*N–2*) степенями вільності. Отже, для побудови довірчого інтервалу для змінної σ^2 ми можемо застосовувати χ^2 закон розподілу

$$\Pr(\chi^2_{1-\alpha/2} \le \chi^2 \le \chi^2_{\alpha/2}) = 1-\alpha, \qquad (3.3.2)$$

де χ^2 – визначена за формулою (3.3.1) змінна, що стоїть посередині нерівності, а $\chi^2_{1-\alpha/2}$ і $\chi^2_{\alpha/2}$ – дві величини χ^2 (критичні величини χ^2), отримані з таблиць розподілу згідно із законом χ^2 з (*N–2*) степенями вільності, причому такими, що відсікають 100(α/2)% хвостових областей розподілу χ^2, як показано на рис. 3.1.

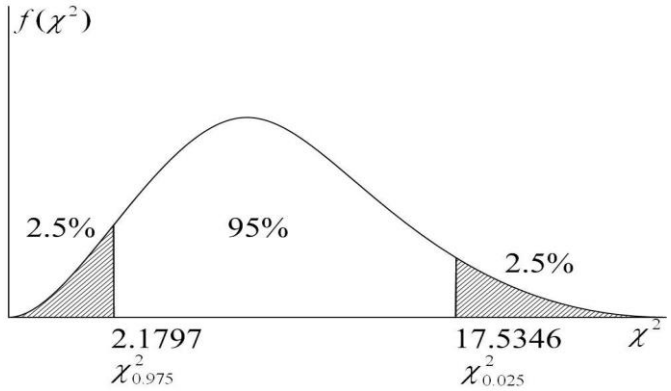


Рис. 3.1. 95%-й довірчий інтервал для χ^2 розподілу з *df* = 8

Підставляючи в (3.3.2) χ^2 з рівності (3.3.1) і перерозподіляючи члени, одержуємо

$$\Pr\left((n-2)\frac{\sigma^2}{\chi^2_{\alpha/2}}\chi^2_{1-\alpha/2} \le \sigma^2 \le (n-2)\frac{\sigma^2}{\chi^2_{1-\alpha/2}} \right) = 1-\alpha, \qquad (3.3.3)$$

який дає 100(1–a)%-й довірчий інтервал для σ^2.

Як ілюстрацію розглянемо досліджений раніше приклад (2.6.1)-(2.6.2):

$$\beta_1 = 24,4545,\ D(\beta_1) = 41,370,\ \sigma(\beta_1) = 6,4138,$$
$$\beta_2 = 0,5091,\ D(\beta_2) = 0,0013,\ \sigma(\beta_2) = 0,0357,$$
$$\mathrm{cov}(\beta_1, \beta_2) = -0,2172,\ \sigma^2 = 42,1591,$$
$$r^2 = 0,9621,\ r = 0,9809,\ df = 8,$$
$$Y_i = 24,4545 + 0,5091X_i.$$

Тут ми маємо $\sigma^2 = 42,1591$ і $df = 8$. Якщо вибрати α=5%, то таблиця χ^2 розподілу для $df = 8$ дає такі критичні величини: $\chi^2_{0,025} = 17,5346$ і $\chi^2_{0,975} = 2,1797$. Вони показують, що імовірність величини χ^2 бути не менше 17,5346 складає

2,5% і не перевищувати 2,1797 є 97,5%. Отже, обмежений цими двома значеннями інтервал є 95%-й довірчий інтервал для χ^2, як показано на рис. 3.1.

Підставляючи дані нашого прикладу в (3.3.3), можна перевірити справедливість такого 95%-го довірчого інтервалу для σ^2:

$$8\frac{42,1591}{17,5346} \le \sigma^2 \le 8\frac{42,1591}{2,1797},$$

$$19,2347 \le \sigma^2 \le 154,7336.$$

Інтерпретація цього інтервалу така: якщо ми встановимо 95%-ві довірчі межі і будемо повторювати багато разів цю процедуру, то в 95 випадках зі ста σ^2 лежатиме всередині цих інтервалів.

3.4. Перевірка гіпотез: загальні зауваження

Обговоривши проблеми точкової і інтервальної оцінок, ми перейдемо до розділу про перевірку гіпотез. Зазначимо коротко деякі загальні аспекти цієї проблеми.

Проблема перевірки статистичних гіпотез може бути сформульована таким чином: узгоджується дане спостереження або висновок з деякою сформульованою гіпотезою чи ні? Використовуваний термін “узгоджуватися” слід розуміти в значенні «достатньої близькості» до величини, про яку йдеться в гіпотезі. Отже, ми не відкидаємо цю гіпотезу. Так, якщо деяка теорія або досвід приводить нас до переконання, що істинне значення кутового коефіцієнта β_2 в прикладі “споживання – дохід”, є одиниця, чи є $\beta_2 = 0,5091$, отримане з розрахунків за даними з вибірки в табл. 1.3. Якщо це так, ми не відкидаємо гіпотезу, інакше ми можемо її відкинути.

Мовою статистики зазначена гіпотеза має назву нульової гіпотези і позначається символом H_0. Нульова гіпотеза перевіряється альтернативною гіпотезою H_1 (відомою також як maintained hypothesis), яка може стверджувати, наприклад, що істинне β_2 не дорівнює одиниці. Альтернативна гіпотеза може бути простою або складною (composite). Наприклад, H_1: $\beta_2 = 1,5$ є проста гіпотеза, а H_1: $\beta_2 \ne 1,5$ – складна гіпотеза.

Теорія перевірки гіпотез займається розвитком методів або процедур для вирішення питання стосовно нульової гіпотези. Існує два взаємно доповнювальних підходи для розробки (devising) таких правил: довірчий інтервал і перевірка значимості. Згідно з цими підходами, дана змінна (статистика або оцінювач) має деякий розподіл імовірності й перевірка гіпотез включає висловлювання про величини параметрів таких розподілів. Наприклад, ми знаємо, що при припущенні про нормальний розподіл β_2 має математичне сподівання β_2 і дисперсію, визначувану (2.3.3). Якщо ми висуваємо гіпотезу, що $\beta_2 = 1$, ми робимо припущення про один з параметрів нормального розподілу, а саме, про його математичне сподівання. Більшість із тих статистичних гіпотез, що розглядатимуться нами далі, будуть подібного типу – висловлювання припущення про один або більше пара-

метрів розподілів, таких як нормальний, F, t або χ^2. Як це відбувається на практиці розглянемо у двох наступних підрозділах.

3.5. Перевірка гіпотез: підхід на основі довірчого інтервалу

Для показу підходу на основі довірчого інтервалу ще раз звернемося до розглянутого прикладу "споживання – дохід". Як ми знаємо, оцінка МРС $\beta_2 = 0,5091$. Припустимо ми постулювали, що

$$H_0 : \beta_2 = 0,3,$$
$$H_1 : \beta_2 \neq 0,3,$$

тобто за нульовою гіпотезою істинне значення β_2 дорівнює 0,3, а за альтернативною – менше або більше цього значення. Нульова гіпотеза – це проста гіпотеза, а альтернативна гіпотеза – складна, вона ще відома як двостороння гіпотеза. Дуже часто така двостороння гіпотеза відображає той факт, що в нас немає вагомих теоретичних аргументів, які говорили б, у якій бік повинна переміщатися альтернативна гіпотеза в порівнянні з нульовою.

Чи відповідає отримане β_2 гіпотезі H_0? Щоб відповісти на це запитання, звернемося до довірчого інтервалу (3.2.9):

$$0,4268 \leq \beta_2 \leq 0,5915.$$

Ми знаємо, що серед множини інтервалів, подібних (3.2.9), β_2 міститиметься з імовірністю 95%. Отже, послідовність інтервалів визначає межі, в яких істинне β_2 може потрапляти з довірчим коефіцієнтом, скажімо, 95%. Таким чином, довірчий інтервал являє собою множину відповідних нульовій гіпотезі інтервалів. Отже, якщо β_2 з гіпотези H_0 потрапляє всередину інтервалу з 100(1–а)%-ю імовірністю, то ми не відкидаємо нульову гіпотезу; якщо ж він лежить поза інтервалом, ми можемо її відкинути. Ця область показана на рис. 3.2.

Правило прийняття рішення: побудуйте 100(1–а)%-й довірчий інтервал для β_2. Якщо β_2 з нульової гіпотези H_0 потрапляє всередину інтервалу, не відкидайте H_0, якщо β_2 лежить поза інтервалом, відкидайте H_0.

Згідно з цим правилом для нашого гіпотетичного прикладу при $H_0 : \beta_2 = 0,3$ ми бачимо, що з 95%-ю імовірністю β_2 лежить поза довірчим інтервалом. Отже, ми можемо відкинути гіпотезу про те, що з 95%-ю імовірністю істинне значення МРС β_2 дорівнює 0,3. Якби нульова гіпотеза була справедлива, імовірність отримання нами величини МРС, такої як 0,5091, була б не більше 5%, тобто дуже незначною.

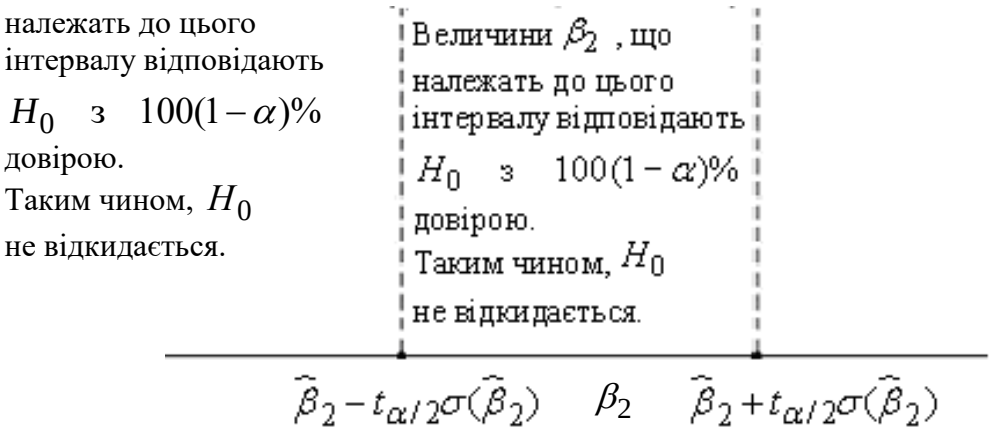


Рис. 3.2. 100(1–α)%-й довірчий інтервал для

У статистиці, коли ми відкидаємо нульову гіпотезу, говорять, що наш висновок статистично значущий.

Односторонній тест

Іноді буває сильне апріорне або теоретичне сподівання (або досвід, заснований на раніше проведених експериментальних роботах), що альтернативна гіпотеза є односторонньою, а не двосторонньою, як було описано раніше. Так, для нашого прикладу моделі "споживання – дохід" хтось може постулювати, що

$$H_0 : \beta_2 \le 0{,}3 \text{ і } H_1 : \beta_2 > 0{,}3.$$

Можливо економічна теорія або попереднє емпіричне дослідження вважає, що МРС більше 0,3. Хоча процедура перевірки цієї гіпотези може бути легко виведена з (3.2.5)

$$\Pr\left[\beta_2 - t_{\alpha/2}\sigma(\beta_2) \le \beta_2 \le \beta_2 + t_{\alpha/2}\sigma(\beta_2)\right] = 1-\alpha .$$

Дійсну механіку краще пояснити термінами підходу перевірки на значимість, який описується нижче.

3.6. Перевірка гіпотез: підхід, оснований на перевірці значимості

Перевірка значимості коефіцієнта регресії: *t*–тест.

Альтернативний по відношенню до методу довірчого інтервалу, але доповнюючий його метод перевірки статистичних гіпотез, є підхід перевірки на значимість, що розроблявся незалежно Р.А.Фішером (R.A.Fisher) і спільно Нейманом і Пірсоном (Neyman and Pearson). У загальному значенні перевірка на значимість є процедура, за допомогою якої результати вибірки використовуються для перевірки істинності або помилковості нульової гіпотези. Основна мета, що лежить в основі цього, полягає в перевірці статистики (оцінювача) і розподілу вибірки за умови виконання нульової гіпотези. Рішення про те, прийняти або відкинути H_0, приймається на основі величини перевірки статистики, виконаної на даних, що є в розпорядженні .

Як ілюстрацію пригадаємо, що при припущенні про нормальність розподіл змінної

$$t = \frac{\beta_2 - \beta_2}{\sigma(\beta_2)} = \frac{(\beta_2 - \beta_2)\sqrt{\sum x_i^2}}{\sigma}, \quad \sigma^2 = \frac{\sum u_i^2}{n-2} \tag{3.6.1}$$

підпорядковується розподілу Ст'юдента з (*N–2*) степенями вільності. Якщо істинне значення β_2 визначене нульовою гіпотезою, змінна *t* може бути легко обчислена за наявними даними вибірки, і, отже, може сприяти перевірці статистики. А оскільки цей тест статистики відповідає *t*–розподілу, можна записати такий довірчий інтервал:

$$\Pr\left[-t_{\alpha/2} \le \frac{\beta_2 - \beta_2^*}{\sigma(\beta_2)} \le t_{\alpha/2}\right] = 1-\alpha, \tag{3.6.2}$$

де β_2^* – величина β_2 за визначенням H_0, а $-t_{\alpha/2}$ і $t_{\alpha/2}$ – величини (критичні величини), отримані з таблиці розподілу Ст'юдента для $\alpha/2$ рівня значимості й *N–2* степенів вільності.

Довірчий інтервал, перетворюваний (3.2.7) до вигляду

$$\Pr\left[\beta_2^* - t_{\alpha/2}\sigma(\beta_2) \le \beta_2 \le \beta_2^* + t_{\alpha/2}\sigma(\beta_2)\right] = 1-\alpha, \tag{3.6.3}$$

дає інтервал, куди β_2 потрапляє з імовірністю 1–α при даному $\beta_2 = \beta_2^*$. При перевірці гіпотез 100(1–α)%-й довірчий інтервал (3.2.8) відомий як область прийнятності (region acceptance), а область поза цим довірчим інтервалом називається критичною областю або областю відмови гіпотези. Як було відзначено раніше, самі кінці довірчого інтервалу носять назву критичних величин.

Внутрішній зв'язок між довірчим інтервалом і тестом на перевірку значимості гіпотези можна чіткіше побачити, якщо порівняти (3.6.4) і (3.6.5)

$$\Pr\left[\beta_2 - t_{\alpha/2}\sigma(\beta_2) \le \beta_2 \le \beta_2 + t_{\alpha/2}\sigma(\beta_2)\right] = 1-\alpha; \tag{3.6.4}$$

$$\Pr\left[\beta_2^* - t_{\alpha/2}\sigma(\beta_2) \le \beta_2 \le \beta_2^* + t_{\alpha/2}\sigma(\beta_2)\right] = 1-\alpha. \tag{3.6.5}$$

У процедурі побудови довірчого інтервалу ми намагаємося встановити область або інтервал, у який з певною імовірністю потрапляє істинне значення β_2, тоді як у тесті на перевірку значимості ми надаємо величині β_2 деякого значення і намагаємося визначити, чи лежить підраховане значення β_2 у достатній близькості від прийнятої величини для β_2.

Звернемося знову до нашого прикладу "споживання – дохід". Ми знаємо, що $\beta_2 = 0{,}5091$, $\sigma(\beta_2) = 0{,}0357$ і $df = 8$. Якщо ми покладемо α=5% $t_{\alpha/2} = 2{,}306$ і вважатимемо $H_0 : \beta_2 = 0{,}3$ і $H_1 : \beta_2 \ne 0{,}3$, то з (3.2.8) одержимо

$$\Pr(0{,}2177 \le \beta_2 \le 0{,}3823) = 0{,}95, \tag{3.6.6}$$

як показано на рис. 3.3.

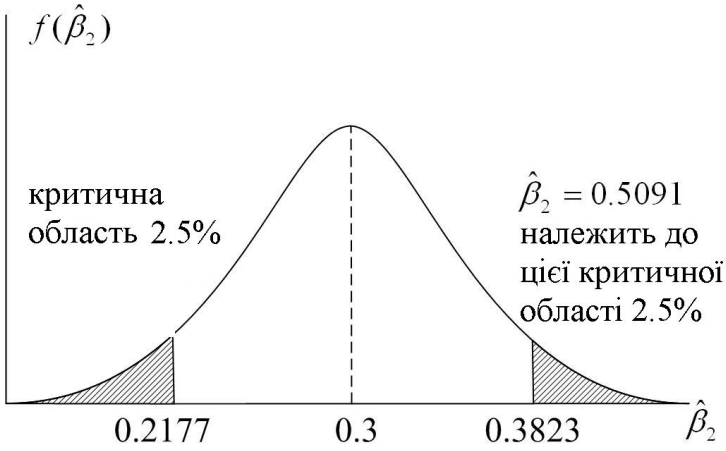


Рис. 3.3. 95%-й довірчий інтервал для β_2 при гіпотезі, що $\beta_2 = 0{,}3$

На практиці немає необхідності обчислювати (3.6.3) явно. Можна обчислити величину $t = \dfrac{\beta_2 - \beta_2}{\sigma(\beta_2)}$, що стоїть посередині нерівності (3.6.2), і подивитися, чи лежить вона між критичними значеннями *t* чи ні. Для нашого випадку

$$t=5{,}86 \qquad (3.6.7)$$

бачимо, що вона лежить у критичній області, зображеній на рис. 3.4.

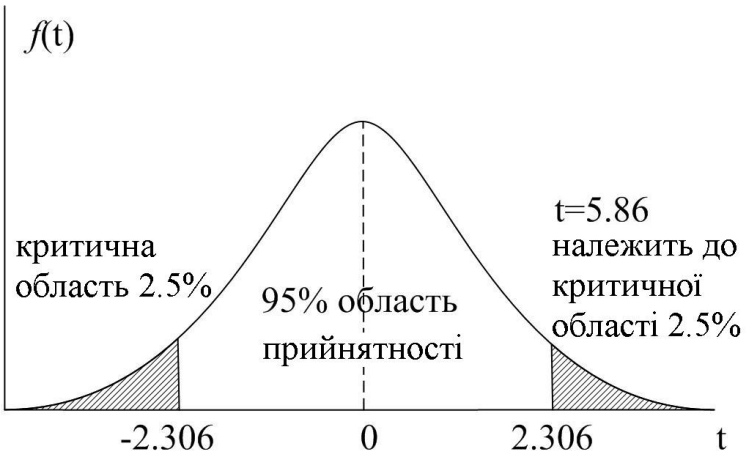


Рис. 3.4. 95%-й довірчий інтервал для *t* (df = 8)

Зауважимо, що якщо β_2 дорівнює значенню, взятому за гіпотезою, то величина *t* у (3.3.10) буде дорівнювати 0. У міру того як β_2 віддалятиметься від значення, взятого за гіпотезою, |*t*| зростатиме. Отже, "велика величина" |*t*| служить аргументом проти гіпотези. Звичайно, ми завжди можемо використовувати таблицю розподілу Ст'юдента для визначення того, чи є конкретна величина *t* великою або малою; відповідь, як ми знаємо, залежить від кількості степенів вільності й від імовірності припустимої помилки. Із таблиці *t*–розподілу можна побачити, що для будь-якої даної величини *df* імовірність отримання все більшої величини |*t*| стає все меншою. Так, для *df*=20 імовірність отримання |t|=1,725 і більше дорівнює 0,10 або 10%, але для тієї ж величини *df* імовірність отримання |t|=3,552 і більше дорівнює тільки 0,002 або 0,2%.

Оскільки ми застосовуємо *t*–розподіл, попередня процедура перевірки має відповідну назву *t*–тесту. На мові перевірки значимості про статистику говорять, що вона є статистично значимою, якщо величина статистики лежить у критичній області. У такому випадку нульова гіпотеза відкидається. За тих же умов про тест говорять як про статистично не значимий, якщо величина тесту статистики лежить в області прийнятності. У цьому випадку нульова гіпотеза не відкидається. У нашому прикладі *t*–тест є значимий і, отже, ми відкидаємо нульову гіпотезу.

Перш ніж закінчити обговорення теми перевірки гіпотез, зазначимо, що описана процедура перевірки відома як двостороння процедура перевірки значимості, у якій ми розглядаємо два хвости розподілу імовірності в області відкидання гіпотези і відкидаємо гіпотезу, якщо вона лежить у будь-якому з цих хвостів. Але це сталося через те, що H_1 – двостороння складна гіпотеза; $\beta_2 \neq 0{,}3$ означає, що β_2 більше 0,3 або менше 0,3. Але припустимо, що наш попередній досвід підказує, що МРС повинен бути більше ніж 0,3. У цьому випадку ми маємо: $H_0 : \beta_2 \leq 0{,}3$ і $H_1 : \beta_2 > 0{,}3$. Хоча H_1 залишається складною гіпотезою, зараз вона одностороння. Для перевірки цієї гіпотези ми використовуємо односторонній тест (правосторонній хвіст), як показано на рис. 3.5.

Процедура перевірки залишається тією ж, що й раніше, за винятком того, що верхня критична межа тепер відповідає $t_\alpha = t_{0{,}05}$, тобто рівню 5%. Як показано на рис. 3.5, нам не потрібно розглядати ліву частину кривої розподілу в такому випадку. Розгляд односторонньої або двосторонньої області залежить від того, як формулюється альтернативна гіпотеза, що, у свою чергу, залежить від раніше набутого досвіду.

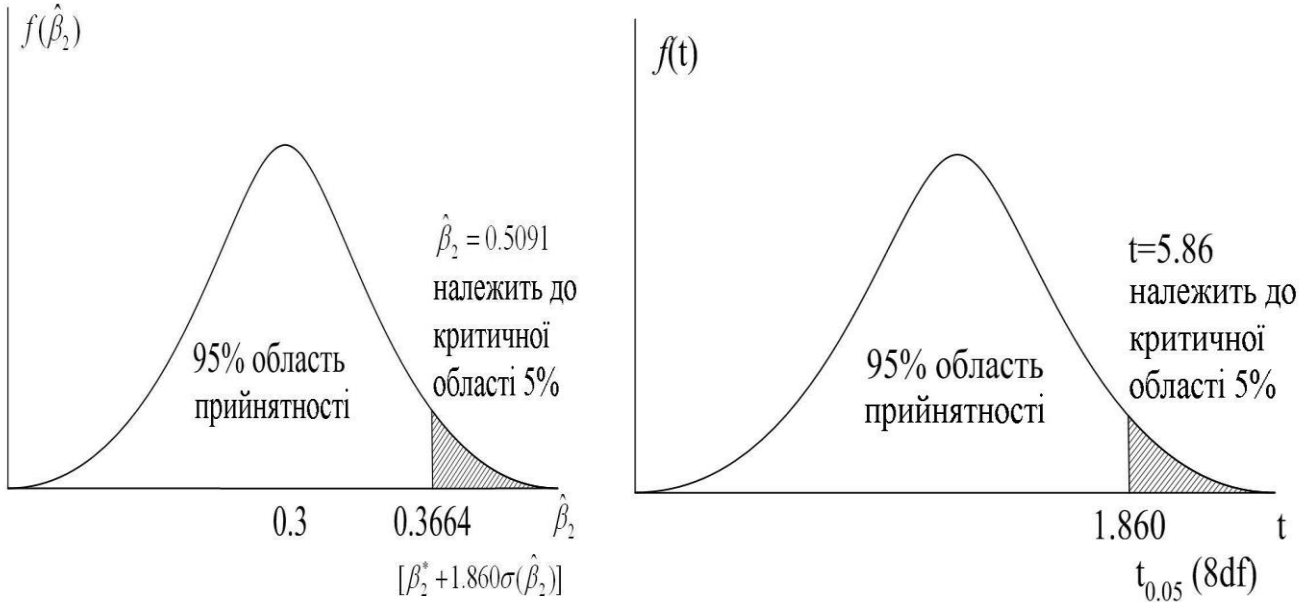


Рис. 3.5. Односторонній тест значимості

Ми можемо підвести підсумки стосовно *t*–тесту перевірки значимості в табл. 3.1.

Таблиця 3.1

T–тест значимості: прийняття рішення

Тип гіпотези	H_0: нульова гіпотеза	H_1: альтернативна гіпотеза	Прийняття рішення: відкинути H_0, якщо
Двостороння	$\beta_2 = \beta_2^*$	$\beta_2 \neq \beta_2^*$	$\lvert t \rvert > t_{\alpha/2}, df$
Правостороння	$\beta_2 \leq \beta_2^*$	$\beta_2 > \beta_2^*$	$t > t_\alpha, df$
Лівостороння	$\beta_2 \geq \beta_2^*$	$\beta_2 < \beta_2^*$	$t < -t_\alpha, df$

3.7. Перевірка значимості σ^2: хі-квадрат тест

Як інший приклад застосування методології тесту значимості розглянемо таку змінну

$$\chi^2 = (n-2)\frac{\sigma^2}{\sigma^2}, \qquad (3.7.1)$$

яка, як було зазначено раніше, розподіляється за законом розподілу χ^2 з (*N–2*) степенями вільності. Для нашого прикладу $\sigma^2 = 42{,}1591$ і *df = 8*. Якщо ми постулюємо, що H_0: $\sigma^2 = 85$; $H_1: \sigma^2 \neq 85$, то рівняння (3.4.1) – перевірка статистики для H_0. Підставляючи в (3.7.1) відповідні значення вхідних величин, одержуємо

$$\chi^2 = 8 \cdot \frac{42{,}1591}{85} = 3{,}97 .$$

Таблиця 3.2

Сумарна таблиця тесту χ^2

Тип гіпотези	H_0: нульова гіпотеза	H_1: альтернативна гіпотеза	Прийняття рішення: відкинути H_0, якщо
Двостороння	$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$\frac{df(\sigma^2)}{\sigma_0^2} > \chi^2_{\alpha/2}, df$ або $\frac{df(\sigma^2)}{\sigma_0^2} < \chi^2_{(1-\alpha/2)}, df$
Правостороння	$\sigma^2 = \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$\frac{df(\sigma^2)}{\sigma_0^2} > \chi^2_\alpha, df$
Лівостороння	$\sigma^2 = \sigma_0^2$	$\sigma^2 < \sigma_0^2$	$\frac{df(\sigma^2)}{\sigma_0^2} < \chi^2_{1-\alpha}, df$
σ_0^2 є величина σ^2 за нульовою гіпотезою			

Якщо ми покладемо α=5%, критичні значення для χ^2 будуть 2,1797 і 17,5346. Оскільки підрахована величина χ^2 лежить між цими межами, дані підтримують нульову гіпотезу і ми її не відкидаємо (див. рис. 3.1). Ця процедура перевірки носить назву χ^2–тесту значимості. У табл. 3.2 наводяться узагальнення χ^2–тесту значимості.

3.8. Регресійний аналіз і аналіз дисперсії

У цьому параграфі ми звернемося до регресійного аналізу з погляду аналізу дисперсії.

Раніше нами була доведена така рівність:

$$\sum y_i^2 = \sum y_i^2 + \sum u_i^2 = \beta_2^2 \sum x_i^2 + \sum u_i^2 , \qquad (3.8.1)$$

тобто TSS=ESS+RSS, яке розкладає загальну суму квадратів (TSS) на два доданки: пояснена сума квадратів (ESS) і сума квадратів залишків (RSS). Вивчення цих доданків у TSS відоме під терміном (ANOVA, analysis variance) аналізу дисперсії з погляду регресії.

З кожною сумою квадратів пов'язані кількість її степенів вільності *df*, кількість незалежних спостережень, на яких вона заснована. TSS має (*N–1*) степінь вільності, оскільки ми втрачаємо один степінь при підрахунку середньої величини вибірки $\bar{Y}$. RSS має (*N–2*) степені вільності (оскільки $\sum u_i = 0$, $\sum u_i X_i = 0$). ESS має всього один степінь вільності, з огляду на те, що $ESS = \beta_2^2 \sum x_i^2$ є функція тільки від β_2, оскільки $\sum x_i^2$ відома. Відзначимо, що вказана кількість степенів вільності справедлива тільки для випадку рівняння регресії з двома змінними.

Помістимо перераховані суми квадратів і відповідні їм степені вільності в табл. 3.3, що є стандартним видом таблиці ANOVA.

Таблиця 3.3

ANOVA таблиця для регресійної моделі з двома змінними

Джерела дисперсії	SS	*df*	MSS
Унаслідок дисперсії (ESS)	$\sum y_i^2 = \beta_2^2 \sum x_i^2$	1	$\beta_2^2 \sum x_i^2$
Унаслідок залишків (RSS)	$\sum u_i^2$	*N–2*	$\frac{\sum u_i^2}{n-2} = \sigma^2$
TSS	$\sum y_i^2$	*N–1*	
SS – сума квадратів (sum squares) MSS – середня сума квадратів (mean sum squares)			

Розглянемо таку змінну:

$$F = \frac{\text{середня сума квадратів ESS}}{\text{середня сума квадратів RSS}} = \frac{MSS\ of\ ESS}{MSS\ of\ RSS} = \frac{\beta_2^2 \sum x_i^2}{\sum u_i^2/(n-2)} = \frac{\beta_2^2 \sum x_i^2}{\sigma^2}. \quad (3.8.2)$$

Якщо ми припустимо, що збурення u_i розподілені нормально і $H_0 : \beta_2 = 0$, то можна показати, що змінна F з (3.8.2) задовольняє умови такої теореми: якщо Z_1 і Z_2 незалежні змінні з k_1 і k_2, відповідно, степенями вільності, що розподіляється за законом розподілу χ^2, тоді змінна

$$F = \frac{Z_1/k_1}{Z_2/k_2} \sim F_{k_1,k_2}$$

розподіляється за законом F-розподілу з *k1* і *k2* степенями вільності, де *k1* називають чисельником степеня вільності, а *k2* – знаменником.

Отже, змінна F з (3.8.2) розподіляється за законом F –розподілу з 1 і (*N–2*) степенями вільності.

Таким чином, можна показати, що

$$E\left(\beta_2^2 \sum x_i^2\right) = \sigma^2 + \beta_2^2 \sum x_i^2 \quad (3.8.3)$$

і

$$E\left(\frac{\sum u_i^2}{n-2}\right) = E(\sigma^2) = \sigma^2. \quad (3.8.4)$$

Отже, якщо β_2 дорівнює 0, обидва рівняння (3.8.3) і (3.8.4) дають нам оцінку істинного значення σ^2. У цьому випадку пояснювальна змінна X не впливає лінійно на Y, і зміна Y пояснюється тільки за рахунок випадкового u_i. Водночас, якщо β_2 не дорівнює нулю, (3.8.3) і (3.8.4) будуть різними і частину дисперсії в Y можна пояснити за рахунок X. Отже, коефіцієнт F у (3.8.1) являє собою тест нульової гіпотези $H_0 : \beta_2 = 0$. Оскільки всі величини, що входять у вираз (3.8.1), отримані з вибірки, коефіцієнт дозволяє перевірити гіпотезу про те, що $\beta_2 = 0$. Усе, що необхідно для цього зробити, це підрахувати F і порівняти його з критичною величиною F, отриманою з таблиць розподілу густини F з вибраним рівнем значимості, або отримати p-величину, обчислену за статистикою F.

Для ілюстрації звернемося до нашого прикладу “споживання – дохід” (табл. 3.4).

Таблиця 3.4

ANOVA-таблиця для прикладу “споживання – дохід”

Джерело дисперсії	SS	*df*	MSS	*F*-відношення
Унаслідок регресії (*ESS*)	8552,73	1	8552,73	*F*=85552,73/42,15
Унаслідок залишків (*RSS*)	337,27	8	42,159	9=
TSS	8890,00	9		=202,87

Із таблиці бачимо, що обчислена величина F = 202,87. Величина p, відповідна цій статистиці, з 1 і 8 степенями вільності не може бути отримана з таблиці розподілу F, але, використовуючи електронні комп'ютерні таблиці, можна показати, що ця величина є 0,0000001, тобто дуже малою. Якщо ви при перевірці гіпотез застосуєте підхід за рівнем значимості α=0,01 або 1%, то побачите, що обчислений коефіцієнт F = 202,87 є значимим для цього рівня. Отже, якщо ми відкинемо нульову гіпотезу про те, що $\beta_2 = 0$, то імовірність виникнення помилки 1 типу (відкидається правильна гіпотеза) дуже мала. Отже, з великою імовірністю ми можемо зробити висновок, що дохід X впливає на витрати і споживання Y.

Пригадаємо теорему про те, що квадрат величини t з k степенями вільності дорівнює F з 1 степенем вільності чисельника і k степенями вільності знаменника, тобто $t_k^2 = F_{1,k}$. Для нашого прикладу моделі "споживання – дохід", якщо ми покладемо $H_0 : \beta_2 = 0$, то з формули для t (3.3.2) легко отримати

$$t = \frac{\beta_2 - \beta_2}{\sigma(\beta_2)} = \frac{\beta_2}{\sigma(\beta_2)} = \frac{0,5091}{0,0357} = 14,24.$$

Це значення змінної t має 8 степенів вільності. При тій же нульовій гіпотезі F = 202,87 має 1 і 8 степенів вільності. Отже, з точністю до округлення маємо 14,242=202,87.

Таким чином, t- і F-тести дають нам два альтернативних, але взаємодоповнюючих шляхи перевірки нульової гіпотези про те, що $\beta_2 = 0$. Якщо це так, то чому не обмежитися t-тестом і не перевіряти F-тест? Виявляється, що для моделі з двома змінними це можна припустити. Але для моделі множинної регресії ми побачимо, що F тест має деякі цікаві додатки, що робить його дуже корисним і потужним методом перевірки статистичних гіпотез.

3.9. Застосування регресійного аналізу: проблема прогнозу

На основі вибіркових даних табл. 3.2 нами було отримане таке рівняння вибіркової регресії:

$$Y_i = 24.4545 + 0.5091 X_i, \qquad (3.9.1)$$

де Y_i – оцінювач істинного $E(Y_i)$, відповідного даному X. Яка користь може бути отримана з цієї історичної регресії (historical regression)? Рівняння можна застосовувати для *прогнозу* або *передбачення* майбутніх споживацьких витрат Y, відповідних деякому даному рівню доходу X. Можливі такі два види прогнозу:

1. Прогноз умовної середньої величини Y, відповідний вибраному X, скажімо X_0, тобто точки на самій лінії регресії популяції.
2. Прогноз індивідуальної величини Y, відповідної X_0.

Ми називатимемо ці два прогнози середнім прогнозом й індивідуальним прогнозом.

Середній прогноз

Для більшої чіткості припустимо, що X_0=100 і ми хочемо спрогнозувати $E(Y/X_0$=100). Зрозуміло, що рівняння історичної регресії (2.6.2) дає оцінку середнього прогнозу таким чином:

$$Y_0 = \beta_1 + \beta_2 X_0 = 24,4545 + 0,5091(100) = 75,3645, \qquad (3.9.2)$$

де Y_0 – оцінка $E(Y/X_0)$. Можна показати, що цей точковий прогноз – краща лінійна незміщена оцінка (best linear unbiased estimator, BLUE).

Оскільки Y_0 є оцінкою, напевно числове її значення відрізняється від істинного. Різниця між двома цими величинами дасть деяке уявлення про помилку прогнозу. Для отримання цієї помилки нам необхідно знайти розподіл вибірки Y_0. Y_0, з формули (3.6.1), – нормально розподілена величина із середнім ($\beta_1+\beta_2 X_0$) і її дисперсія задається такою формулою:

$$D(\overleftarrow{Y}_0) = \sigma^2\left(\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum x_i^2}\right). \qquad (3.9.3)$$

Позначимо невідому σ^2 її незміщеною оцінкою σ^2. Тоді змінна

$$t = \frac{Y_0 - (\beta_1 + \beta_2 X_0)}{\sigma(Y_0)} \qquad (3.9.4)$$

розподілена за законом розподілу Ст'юдента з *N–2* степенями вільності. Отже, цей закон розподілу можна застосовувати для побудови довірчих інтервалів істинного $E(Y_0|X_0)$ і перевірки гіпотез звичайним способом:

$$\Pr[\beta_1 + \beta_2 X_0 - t_{\alpha/2}\sigma(Y_0) \le \beta_1 + \beta_2 X_0 \le \beta_1 + \beta_2 X_0 + t_{\alpha/2}\sigma(Y_0)] = 1-\alpha, \qquad (3.9.5)$$

де $\sigma(Y_0)$ отримано з (3.6.2):

$$\sigma(Y_0) = \sigma\sqrt{\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum x_i^2}}.$$

Для наших даних одержуємо

$$D(Y_0) = 42,1591\left(\frac{1}{10} + \frac{(100-170)^2}{33000}\right) = 10,4759$$

і

$$\sigma(Y_0) = 3,2366.$$

Отже, 95%-й довірчий інтервал для істинного $E(Y_0|X_0)=\beta_1+\beta_2 X_0$ задається формулою

$$67,9010 \le E(Y_0 \mid X = 100) \le 82,8381. \qquad (3.9.6)$$

Таким чином, для $X_0=100$ у вибірках, що повторюються, 95 зі 100 інтервалів, подібних (3.6.5), міститимуть істинну середню величину; найкращим точковим оцінювачем буде 75,3645.

Якщо ми отримаємо 95%-ві довірчі інтервали, подібні (3.9.7), для кожного значення X, наведених у табл. 1.3, ми отримаємо довірчу область (confidence band) для функції регресії популяції, показаної на рис. 3.6.

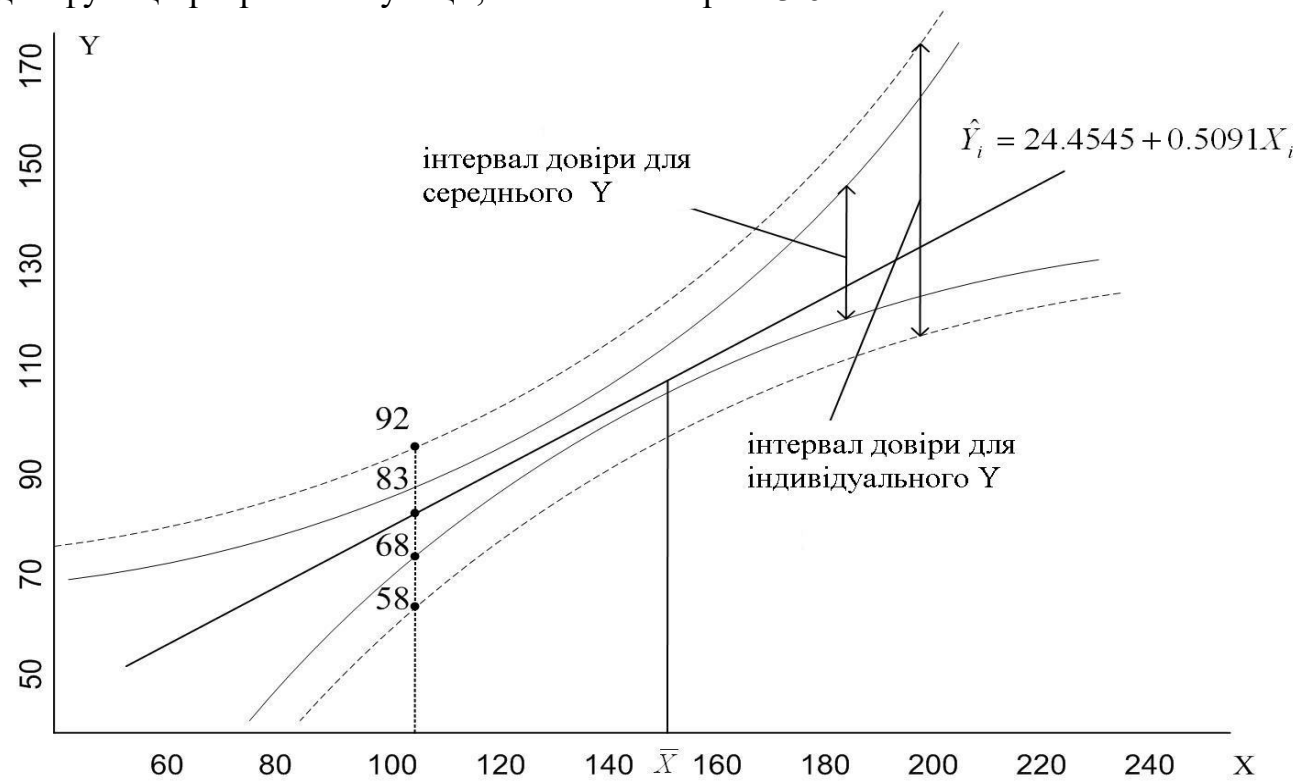


Рис. 3.6. Довірчі інтервали (області) для середньої Y й індивідуальної величини Y

Індивідуальний прогноз

Якщо нас цікавить індивідуальний прогноз величини Y для Y_0, відповідної заданій величині X_0, то краща лінійна незміщена оцінка Y_0 також подається формулою (3.6.1), але її дисперсія має вигляд

$$D(Y_0 - Y_0) = E[Y_0 - Y_0]^2 = \sigma^2\left(1 + \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum x_i^2}\right). \qquad (3.9.7)$$

Можна показати, що Y_0 також підкоряється нормальному закону розподілу з математичним сподіванням і дисперсією, заданими формулами (3.6.1) і (3.6.6) відповідно. Підставляючи σ^2 замість невідомої σ^2 можна показати, що

$$t = \frac{Y_0 - Y_0}{\sigma(Y_0 - Y_0)}, \quad \sigma(Y_0 - Y_0) = \sqrt{D(Y_0 - Y_0)} = \sigma\sqrt{1 + \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum x_i^2}}$$

також підкоряється розподілу Ст'юдента. Отже, цей закон розподілу може бути застосований для висновку про істинне значення Y_0. Продовжуючи дослідження моделі "споживання – дохід", ми знаходимо, що прогнозована точка Y_0=75,3645 така ж, як і Y_0, і її дисперсія є

$$D(Y_0 - Y_0) = 42{,}159\left(1 + \frac{1}{10} + \frac{(100-170)^2}{33000}\right) = 52{,}6349,$$

$$\sigma(Y_0 - Y_0) = \sqrt{52{,}6349} = 7{,}2550.$$

Отже, 95%-й довірчий інтервал для Y_0, відповідного X_0=100, визначається таким чином:

$$(58{,}6345 \le Y_0 \mid X_0 = 100 \le 92{,}0945). \qquad (3.9.8)$$

Порівнюючи цей інтервал з (3.6.5), ми бачимо, що довірчий інтервал для індивідуального Y_0 ширший, ніж за тих же умов довірчий інтервал для середнього значення $E(Y|X_0)$. Обчисливши подібні довірчі інтервали для різних значень X з табл. 1.3, ми отримаємо 95%-ву довірчу область для індивідуальних значень Y при даних значеннях X. Ця довірча область зображена на рис.3.6, як і довірча область для Y_0.

Звернемо увагу на важливу властивість довірчих областей, зображених на рис.3.6. Ширина цих областей найменша при $X_0 = \overline{X}$ (див. вирази для дисперсій і доданок у них $\frac{(X_0 - \overline{X})^2}{\sum x_i^2}$). Ширина областей довіри збільшується в міру віддалення X_0 від $\overline{X}$. Така поведінка свідчить про те, що здатність прогнозу зменшується в міру віддалення X_0 від $\overline{X}$.

Отже, потрібно дуже обережно екстраполювати лінії історичної регресії для прогнозу $E(Y|X_0)$ або Y_0 для заданого X_0, що знаходиться на значній відстані від $\overline{X}$ – середнього значення за вибіркою.

3.10. Форма звіту за результатами регресійного аналізу

Існують різні форми звіту за результатами регресійного аналізу, але тут ми застосуємо ту, що належить до нашого ілюстрованого прикладу:

$$\begin{array}{lll} Y_i = 24{,}4545 & + \quad 0{,}5091X_i, & \\ \sigma = (6{,}4138) & (0{,}0357) & r^2 = 0{,}9621 \\ t = (3{,}8128) & (14{,}2405) & df = 8 \quad 5/11/1) \\ p = (0{,}002571) & (0{,}000000289) & F_{1,8} = 202{,}87. \end{array} \qquad (3.10.1)$$

У виразах (3.10.1) числа в дужках у другому рядку являють собою стандартні помилки коефіцієнтів регресії, числа в дужках у третьому рядку – оцінки величин t:

$$t = \frac{\beta_2 - \beta_2}{\sigma(\beta_2)}, \qquad t = \frac{\beta_1 - \beta_1}{\sigma(\beta_1)}$$

у припущенні нульової гіпотези про те, що β_1=0 і β_2=0 (3,8128 = 24,4545/6,4138; 14,2405 = 0,5091/0,0357), а числа у третьому рядку – обчислена ймовірність. Так, для 8 *df* імовірність отримання величини $t \ge 3{,}8128$ дорівнює 0,0026, а ймовірність отримання $t \ge 14{,}2405$ – близько 0,0000003.

Шляхом подання імовірності *p* оцінюваних коефіцієнтів ми можемо відразу бачити точний рівень значимості кожної оцінюваної величини *t*. Так, при нульовій гіпотезі H_0: $\beta_1 = 0$, точна імовірність (тобто величина *p*) отримання величини $t \geq 3{,}8128$ приблизно дорівнює 0,0026. Отже, якщо ми відкидаємо цю нульову гіпотезу, імовірність того, що ми допускаємо помилку 1 становить 26 випадків із 10 000, що насправді дуже мало. У практичних задачах можна сказати, що істинне значення $\beta_1 \neq 0$. Ще більше це стосується схильності до покупки β_2.

Як ми раніше згадували $F_{1,k} = t_k^2$. При нульовій гіпотезі $\beta_2 = 0$ величина *F*=202,87 (1 *df* чисельник, 8 *df* знаменник), а *t*=14,24 (8 *df*); відповідно до теорії (14,24)2=202,87.

3.11. Обчислення результатів регресійного аналізу

У вступі нами була описана загальна схема економетричного моделювання:

1. Економічна теорія.
2. Математична модель теорії.
3. Економетрична модель теорії.
4. Дані.
5. Проведення оцінки параметрів економетричної моделі.
6. Перевірка гіпотез.
7. Прогнозування.
8. Використання моделі для прийняття рішень або з політичною метою.

Зараз, після подання результатів регресійного аналізу нашої моделі "споживання – дохід" у вигляді (3.11.1), природно задати питання про адекватність цієї моделі. Наскільки «добре» вона підігнана до наявних даних? Для відповіді на це запитання нам потрібні деякі критерії.

По-перше, чи відповідають знаки оцінених коефіцієнтів теоретичним прогнозам або наявному досвіду? Апріорі β_2, схильність до покупки, повинна бути позитивною. По-друге, якщо згідно з теорією взаємозв'язок повинен бути не тільки позитивним, але й статистично значимим, чи виконується це? Як ми зазначили в розд. 3, MPC – схильність до покупки – є не тільки позитивною величиною, а й статистично значимо відрізняється від нуля. Ці ж зауваження стосуються й коефіцієнта β_1. По-третє, наскільки добре регресійна модель пояснює зміну споживацьких витрат? Для відповіді на це запитання можна застосувати r^2. У нашому випадку $r^2 \approx 0{,}96$ дуже високий, враховуючи, що максимальне значення r^2 є 1.

Таким чином, вибрана нами модель для пояснення характеру споживацьких витрат є цілком прийнятною. Але перш ніж підвести межу, цікаво з'ясувати, чи задовольнить модель припущення CNLRM (classical normal linear regression model) класичної лінійної моделі з нормальним законом розподілу. Ми не звертатимемо увагу на різні припущення, оскільки модель є дуже простою. Але є одна гіпотеза, яку ми хотіли б перевірити, а саме – гіпотеза про нормальний розподіл випадкової складової u_i.

Пригадаємо, що використані раніше *t*- і *F*-тести припускали, що u_i розподілена за нормальним законом розподілу. Інакше процедура перевірки не буде дійсна для малих або кінцевих вибірок.

Тест на нормальність

Хоча в літературі обговорюється ряд тестів перевірки на нормальність, обмежимося розглядом двох: 1) тест χ^2 якості підгонки і 2) тест Jarque-Bera. Обидва ці тести використовують залишки u_i і розподіл імовірності χ^2.

χ^2 тест якості підгонки. Цей тест проводять таким чином. Спочатку ми отримуємо рівняння регресії, а також залишки u_i, підраховуємо стандартне відхилення u_i ($D(u_i) = \sum(u_i - \bar{u})^2/(n-1) = \sum u_i^2/(n-1)$, оскільки $\bar{u} = 0$). Потім упорядковуємо залишки і розміщуємо їх у різних групах (у нашому прикладі ми розміщуємо їх у шести групах), відповідних величині відхилення від нуля. Для нашого прикладу ми одержуємо такі дані (табл. 3.5).

Таблиця 3.5

Залишки для проведення χ^2-тесту

Спостережувані залишки (*Oi*)	0,0	2,0	3,0	4,0	1,0	0,0
Очікувані залишки (*Ei*)	0,2	1,4	3,4	3,4	1,4	0,2
$(Oi–Ei)^2/E_i$ Sum=0,92	0,2	0,26	0,05	0,10	0,11	0,2
$(O_i–E_i)^2/E_i$ Sum=0,92	0,2	0,26	0,05	0,10	0,11	0,2
$O_i = u_i$, де u_i залишки за МНК						

У табл. 3.5 рядок, позначений як спостережувані залишки, дає частоту розподілу залишків для встановленого стандартного відхилення нижче і вище за нуль. У нашому прикладі немає залишків у два стандартних відхилення нижче нуля, 3 залишки між 1 і 2 стандартних відхилення нижче за нуль, 3 залишки між 0 і 1 стандартним відхиленням нижче за нуль, 4 залишки між 0 і 1 стандартним відхиленням вище за нуль, 1 залишок між 1 і 2 стандартними відхиленнями вище за нуль і жодного залишку, більшого за 2 стандартні відхилення вище за нуль.

Дані, поміщені в рядок очікуваних залишків, дають частоту розподілу залишків на основі передбачуваного закону розподілу ймовірності, у нашому випадку – нормального закону розподілу. У третьому рядку обчислюємо різницю між спостережуваними й очікуваними частотами, підносимо цю різницю у квадрат, ділимо результат на очікувану частоту і підсумовуємо. Алгебраїчний запис буде такий:

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}, \qquad (3.11.1)$$

де O_i – спостережувана частота в класі або інтервалі *i*, *Ei* – очікувана частота в класі *i* на основі гіпотетичного розподілу, скажімо, нормального. Якщо різниця між спостережуваними й очікуваними частотами "мала", то можна припустити, що відхилення u_i невеликі.

4. РОЗВИТОК ДВОВИМІРНОЇ ЛІНІЙНОЇ МОДЕЛІ РЕГРЕСІЇ

Деякі аспекти лінійного регресійного аналізу можна без особливих зусиль використовувати на базі описаної вище лінійної моделі регресії. Перш за все розглянемо випадок регресії, яка проходить через початок координат, тобто ситуацію, коли коефіцієнт β_1 відсутній у моделі. Потім ми розглянемо питання про одиниці вимірювання, тобто як вимірюються змінні Y і X і чи впливають одиниці вимірювання на результати регресії. Нарешті, ми розглянемо питання про вид функції в лінійній регресійній моделі. Дотепер ми розглядали моделі, лінійні як за параметрами, так і за змінними. Але пригадаємо, що теорія припускає лінійність моделі лише за параметрами; змінні можуть входити в модель як лінійно, так і нелінійно. Розглядаючи моделі, не обов'язково лінійні за змінними, ми покажемо, що вони застосовні до багатьох цікавих практичних задач.

Оскільки ідеї, що розглядаються в цьому розділі, легкі для сприйняття, їх поширення на багатовимірний регресійний аналіз буде цілком зрозумілим продовженням.

4.1. Регресія, що проходить через початок координат

Бувають випадки, коли PRF з двома змінними зображають у вигляді

$$Y_i = \beta_2 X_i + u_i . \qquad (4.1.1)$$

У цій моделі параметр β_1, визначаючий точку перетину з віссю ординат, відсутній або дорівнює нулю. Така модель називається регресією, що проходить через початок координат.

Як ілюстрацію розглянемо модель оцінки капітальних активів (Capital Asset Pricing Model) CAPM сучасної теорії портфеля, яка може бути подана у формі "ризик-премія":

$$(ER_i - r_f) = \beta_i (ER_m - r_f), \qquad (4.1.2)$$

де ER_i – очікувана норма прибутку за цінними паперами *i*;

ER_m – очікувана норма прибутку на ринку портфеля, як наприклад фондовий індекс;

r_f – безризикова норма прибутку;

β_i – коефіцієнт, що являє собою міру систематичного ризику, тобто ризику, який не може бути виключений за допомогою диверсифікації. Це така міра впливу ринку на норму прибутку *i*-го цінного паперу. При $\beta_i > 1$ говорять, що цінні папери нестійкі або агресивні, а при $\beta_i < 1$ – що вони захищені.

Якщо ринки капіталів працюють ефективно, то згідно з CAPM очікувана премія за ризик за *i*-ми цінними паперами $ER_i - r_f$ дорівнює добутку відповідного β-коефіцієнта на премію за ризик на ринку $ER_m - r_f$. На рис.4.1 зображена лінія ринку цінних паперів.

Рівняння (4.1.2.) часто записують у вигляді

$$R_i - r_f = \beta_i (R_m - r_f) + u_i \qquad (4.1.3)$$

або

$$R_i - r_f = \alpha_i + \beta_i (R_m - r_f) + u_i . \qquad (4.1.4)$$

Остання модель носить назву моделі ринку. Якщо CAPM виконується, то очікується, що α_i повинен дорівнювати нулю.

Відзначимо, що в наведених рівняннях (4.1.3) і (4.1.4) залежна змінна Y є $(R_i - r_f)$, а пояснююча змінна X є β_i, а не $(R_m - r_f)$. Отже, для складання рівняння регресії (4.1.4) необхідно спочатку оцінити β_i, які звичайно виводяться з рівняння характеристичної лінії.

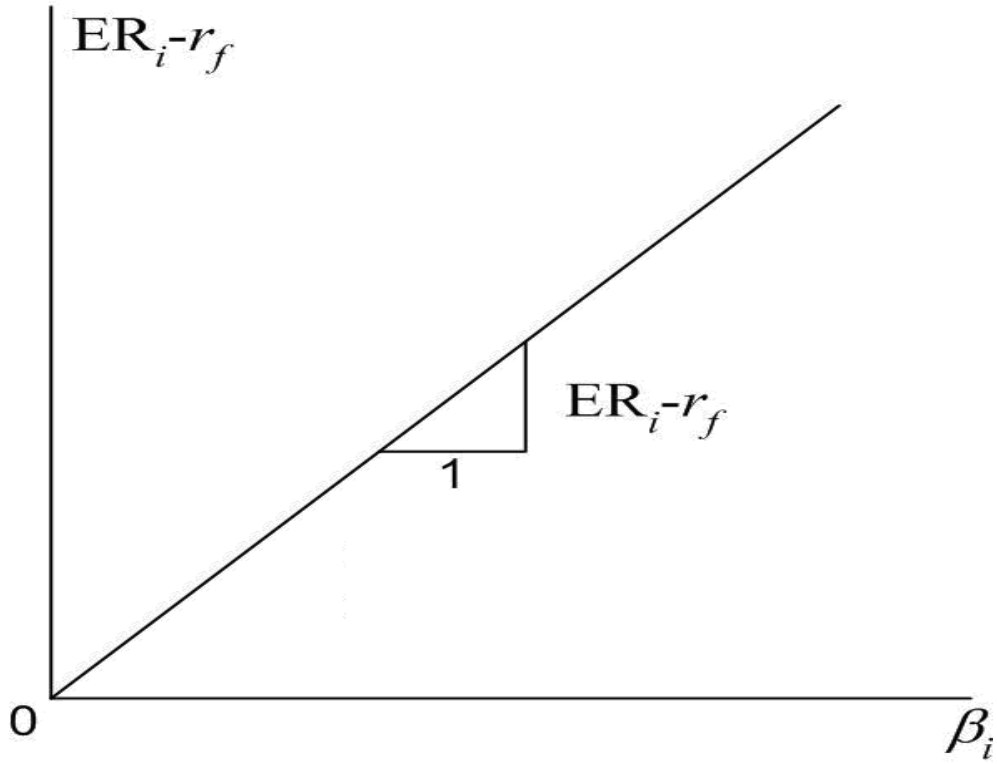


Рис.4.1. Лінія ринку цінних паперів

Як нам провести оцінку моделі, подібної (4.1.1), і які при цьому виникають особливості? Щоб відповісти на це питання, запишемо спочатку SRF моделі (4.1.1):

$$Y_i = \beta_2 X_i + u_i . \qquad (4.1.5)$$

Застосуємо метод найменших квадратів до (4.1.5), отримаємо таку формулу для β_2 і його дисперсії:

$$\sum u_i^2 = \sum (Y_i - \beta_2 X_i)^2 , \; \frac{d}{d\beta_2} \sum u_i^2 = -2 \sum X_i (Y_i - \beta_2 X_i) = -2 \sum X_i Y_i + 2\beta_2 \sum X_i^2 = 0 ,$$

звідси

$$\beta_2 = \frac{\sum X_i Y_i}{\sum X_i^2}. \qquad (4.1.6)$$

Підставимо в (4.1.6) (4.1.5), одержимо

$$\beta_2 = \frac{\sum X_i(\beta_2 X_i + u_i)}{\sum X_i^2} = \frac{\beta_2 \sum X_i^2 + \sum X_i u_i}{\sum X_i^2} = \beta_2 + \frac{\sum X_i u_i}{\sum X_i^2}.$$

Враховуючи незміщеність оцінки $E(\beta_2) = \beta_2$, одержимо

$$E(\beta_2 - \beta_2) = E\left(\frac{\sum X_i u_i}{\sum X_i^2}\right)^2.$$

Піднесемо у квадрат вираз, узятий у дужки в правій частині вищенаведеного рівняння. Оскільки X_i передбачаються нестохастичними, використовуючи властивість гомоскедастичності і некорельованості відхилень u_i, одержуємо

$$D(\beta_2) = E(\beta_2 - \beta_2)^2 = E\left(\frac{X_1^2 u_1^2 + X_2^2 u_2^2 + \ldots + 2X_1 X_2 u_1 u_2 + \ldots + 2X_{n-1} X_n u_{n-1} u_n}{\left(\sum X_i^2\right)^2}\right) =$$

$$= \frac{1}{\left(\sum X_i^2\right)^2} E\left(\sum X_i^2 u_i^2 + 2\sum_{i=1}^{n-1} u_i u_{i+1}\right) = \frac{1}{\left(\sum X_i^2\right)^2}\left\{\sum X_i^2 E\left(u_i^2\right) + 2\sum_{i=1}^{n-1} E\left(u_i u_{i+1}\right)\right\} =$$

$$= \frac{\sigma^2 \sum X_i^2}{\left(\sum X_i^2\right)^2} = \frac{\sigma^2}{\sum X_i^2},$$

тобто

$$\text{var}(\beta_2) = \frac{\sigma^2}{\sum X_i^2}, \qquad (4.1.7)$$

де оцінкою σ^2 служить

$$\sigma^2 = \frac{\sum u_i^2}{n-1}. \qquad (4.1.8)$$

Зауважимо, що при виведенні рівняння для визначення β_2, легко отримується рівність

$$\sum X_i u_i = 0.$$

Подібна рівність була виконана і для моделі, що містить β_1. Нагадаємо, що для моделі, яка містить β_1, крім того виконувалася рівність $\sum u_i = 0$. З вищезазначеного зрозуміло, що для регресії, яка проходить через початок координат, сума залишків $\sum u_i$ не обов'язково дорівнює нулю.

Пригадаємо також, що $Y_i = Y_i + u_i$. Підсумовуючи обидві частини рівності за всією вибіркою і розділивши потім на її обсяг, одержуємо

$$\bar{Y} = \bar{Y} + \bar{u}.$$

Оскільки для моделі, що проходить через початок координат $\sum u_i \neq 0$, а отже і $\bar{u} \neq 0$, то

$$\bar{Y} \neq \bar{Y}.$$

Таким чином, у цьому випадку середнє фактичних значень Y не дорівнює середнім значенням за регресією. У цьому відмінність від властивостей моделі, що містить β_1.

Цікаво порівняти формули (4.1.6)–(4.1.8) з отриманими для моделі з β_1, а саме з формулами (2.1.9), (2.3.1) і (2.3.5).

У моделі з $\beta_1 = 0$ є декілька моментів, про які слід згадати. R^2, коефіцієнт детермінації, у випадку з $\beta_1 = 0$ може бути негативним. Тому, обчислений звичайним способом, він може не відповідати своєму значенню для регресії, що проходить через початок координат.

R^2 для регресії, що проходить через початок координат

Як ми тільки що відзначили, звичайний коефіцієнт детермінації R^2, розглянутий нами раніше для моделі, що містить коефіцієнт β_1, не придатний для регресії, що проходить через початок координат. Але можна обчислити так званий raw R^2 (сирий) для такої моделі, якщо застосувати формулу

$$rawR^2 = \frac{\sum (X_i Y_i)^2}{\sum X_i^2 \sum Y_i^2}. \qquad (4.1.9)$$

Хоча цей raw R^2 задовольняє умову $0 < R^2 < 1$, його не можна порівнювати безпосередньо зі звичайним R^2. Тому деякі автори не наводять його для моделі з $\beta_1 = 0$.

Ураховуючи ці особливості, необхідно бути дуже обережним при використанні моделі з $\beta_1 = 0$. Якщо немає дуже вагомих підстав, краще використовувати звичайну модель, у якій наявний коефіцієнт β_1. Це має подвійну перевагу. По-перше, якщо β_1 включений у модель, але виявиться статистично не значимим (тобто таким, що статистично дорівнює нулю), то для практичних застосувань ми маємо регресію, що проходить через початок координат. Друге, і більш важливе зауваження, якщо в моделі насправді наявний β_1, але ми не включили його і використовуємо модель з $\beta_1 = 0$, то ми допускаємо помилку специфікації моделі.

Ілюстрований приклад: характеристична лінія теорії портфеля

У табл. 4.1 наведені дані за річною нормою прибутку (у %) Afuture Fund, взаємного фонду, завданням інвестиційної політики якого є отримання максимального приросту капіталу, а також ринку портфеля, визначуваного індексом Фішера, за період 1971–1980 рр.

Ми вже згадували, що характеристична лінія в інвестиційному аналізі може бути записана у вигляді

$$Y_i = \alpha_i + \beta_i X_i + u_i, \qquad (4.1.10)$$

де Y_i – річна норма прибутку (у %) для Afuture Fund;

X_i – річна норма прибутку (у %) на ринку портфеля;

β_i – кутовий коефіцієнт, відомий як бета-коефіцієнт в теорії портфеля;

α_i – вільний член у рівнянні регресії.

У літературі не існує єдиної думки про значення α_i. Деякі емпіричні результати говорять про те, що він позитивний і статистично значимий, інші ж свідчать, що він не відрізняється суттєво від нуля. В останньому випадку ми можемо записати рівняння у вигляді

$$Y_i = \beta_i X_i + u_i. \qquad (4.1.11)$$

Це рівняння регресії, що проходить через початок координат. Якщо ми застосуємо рівняння (4.1.11), то отримаємо такі результати:

$$\begin{aligned} Y_i &= 1{,}08999 X_i \\ &\quad (0{,}1916) \\ \text{t} &= (5{,}6884) \qquad raw R^2 = 0{,}7825. \end{aligned} \qquad (4.1.12)$$

Результати свідчать, що β_i суттєво більше нуля. Інтерпретація отриманих результатів така: при зростанні норми прибутку на ринку портфеля на 1%, норма прибутку для Afuture Fund в середньому зростає на 1,09%.

Таблиця 4.1.

Річна норма прибутку для Afuture Fund та за індексом Фішера (ринок портфеля) за 1971–1980 рр.

Рік	Норма прибутку Afuture Fund, Y	Норма прибутку за індексом Фішера, X
1971	67,5	19,5
1972	19,2	8,5
1973	–35,2	–29,3
1974	–42,0	–26,5
1975	63,7	61,9
1976	19,3	45,5
1977	3,6	9,5
1978	20,0	14,0
1979	40,3	35,3
1980	37,5	31,0

Чи можна бути впевненим в тому, що модель (4.1.11), а не (4.1.10) є правильною, особливо, якщо врахувати той факт, що не існує сильної попередньої довіри до гіпотези, яка стверджує, що коефіцієнт α_i дорівнює нулю? Це можна перевірити, якщо використовувати модель регресії (4.1.10). За цією моделлю ми одержуємо такі результати:

$$\begin{array}{lll} Y_i = 1,2645 + 1,0714 X_i & & \\ \quad (7,6838) \quad (0,2387) & & \\ t = (0,1643) \quad (4,4880) & R^2 = 0,7157 & (4.1.13) \\ p = (0,873566) \quad (0,002034) & & \end{array}$$

Із цих результатів випливає, що ми не можемо відкидати гіпотезу про рівність нулю коефіцієнта β_1. Це виправдовує використання моделі (4.1.1) – регресії, що проходить через початок координат. Порівняння результатів регресійного аналізу (4.1.12) і (4.1.13) показує, що між ними немає істотної різниці. Слід відзначити лише меншу величину стандартної похибки для коефіцієнта β_i. Це підтримує існуючу думку про більш точне обчислення коефіцієнта β для моделі, що проходить через початок координат. Використовуючи ці дані, можна перевірити, що 95%-й довірчий інтервал для (4.1.12) є (0,6566; 1,5232), а для моделі (4.1.13) цей інтервал є (0,52093; 1,622). Другий довірчий інтервал вужчий, ніж перший, що свідчить про більшу точність визначення коефіцієнта β_1 за моделлю (4.1.12).

4.2. Масштабування й одиниці вимірювання

Для того щоб зрозуміти суть питання, розглянемо дані, наведені в табл. 4.2.

Таблиця 4.2

Валові внутрішні приватні інвестиції (GPDI) і валовий національний продукт (GNP) у цінах 1972 р. у доларах США, 1974–1983 рр.

Рік	GPDI, млрд дол	GPDI, млн дол	GNP, млрд. дол	GNP, млн. дол
1974	195,5	195500	1246,3	1246300
1975	154,8	154800	1231,6	1231600
1976	184,5	184500	1298,2	1298200
1977	214,2	214200	1369,7	1369700
1978	236,7	136700	1438,6	1438600
1979	236,3	136300	1479,4	1479400
1980	208,5	208500	1475,0	1475000
1981	230,9	230900	1512,2	1512200
1982	194,3	194300	1480,0	1480000
1983	221,0	221000	1534,7	1534700

Припустимо, що в регресії GPDI за GNP один дослідник використовує дані, що обчислюються в мільярдах, а інший – у мільйонах доларів. Чи будуть результати регресійного аналізу однаковими в обох випадках? Якщо ні, то який результат слід використовувати? Інакше, чи впливають одиниці, в яких вимірюються Y і X, на результати регресійного аналізу?

Щоб відповісти на це запитання, зробимо так. Хай

$$Y_i = \beta_1 + \beta_2 X_i + u_i, \tag{4.2.1}$$

де Y_i – GPDI, X_i – GNP.

Визначимо

$$Y_i^* = w_1 Y_i, \tag{4.2.2}$$

$$X_i^* = w_2 X_i, \tag{4.2.3}$$

де w_1 і w_2 – константи, що називаються масштабними чинниками. w_1 і w_2 можуть збігатися, а можуть бути різними.

З рівнянь (4.2.2) і (4.2.3) зрозуміло, що Y_i^* і X_i^* змінюють шкалу вимірювань Y_i і X_i. Так, якщо Y_i і X_i вимірюються в мільярдах доларів, а ми хочемо перейти до вимірювання в мільйонах доларів, то

$$Y_i^* = 1000 Y_i,$$

$$X_i^* = 1000 X_i.$$

У цьому випадку $w_1 = w_2 = 1000$.

Тепер розглянемо регресію, застосовуючи змінні Y_i^* і X_i^*:

$$Y_i^* = \beta_1^* + \beta_2^* X_i + u_i^*, \tag{4.2.4}$$

де $Y_i^* = w_1 Y_i$, $X_i^* = w_2 X_i$ і $u_i^* = w_1 u_i$.

Ми можемо отримати зв'язок між парами:

$$\beta_1 \text{ і } \beta_1^*,$$

$$\beta_2 \text{ і } \beta_2^*,$$

$$\text{var}(\beta_1) \text{ і } \text{var}(\beta_1^*),$$

$$\text{var}(\beta_2) \text{ і } \text{var}(\beta_2^*),$$

$$\sigma^2 \text{ і } \sigma^{*2},$$

$$r_{xy}^2 \text{ і } r_{x^*y^*}^2.$$

За методом найменших квадратів ми маємо

$$\beta_1 = \bar{Y} - \bar{X}\beta_2, \tag{4.2.5}$$

$$\beta_2 = \frac{\sum x_i y_i}{\sum x_i^2}, \tag{4.2.6}$$

$$\text{var}(\beta_1) = \frac{\sum X_i^2}{n\sum x_i^2}\sigma^2, \tag{4.2.7}$$

$$\text{var}(\beta_2) = \frac{\sigma^2}{\sum x_i^2}, \tag{4.2.8}$$

$$\sigma^2 = \frac{\sum u_i^2}{n-2}. \tag{4.2.9}$$

Застосовуючи МНК до (6.2.4), одержуємо

$$\beta_1^* = \bar{Y}^* - \bar{X}^*\beta_2^*, \tag{4.2.10}$$

$$\beta_2^* = \frac{\sum x_i^* y_i^*}{\sum x_i^{*2}}, \tag{4.2.11}$$

$$\text{var}(\beta_1^*) = \frac{\sum X_i^{*2}}{n\sum x_i^{*2}}\sigma^{*2}, \tag{4.2.12}$$

$$\text{var}(\beta_2^*) = \frac{\sigma^{*2}}{\sum x_i^{*2}}, \tag{4.2.13}$$

$$\sigma^{*2} = \frac{\sum u_i^{*2}}{n-2}. \tag{4.2.14}$$

З цієї рівності легко отримати співвідношення між двома наборами оцінок параметрів. Усе, що для цього потрібно, так це застосувати рівності: $Y_i^* = w_1 Y_i$ (або $y_i^* = w_1 y_i$); $X_i^* = w_2 X_i$ (або $x_i^* = w_2 x_i$); $u_i^* = w_1 u_i$; $\bar{Y}_i^* = w_1 \bar{Y}_i$; $\bar{X}_i^* = w_2 \bar{X}_i$. Застосовуючи ці співвідношення, нескладно отримати рівності, що цікавлять нас:

$$\beta_2^* = \left(\frac{w_1}{w_2} \right) \beta_2, \tag{4.2.15}$$

$$\beta_1^* = w_1 \beta_1, \tag{4.2.16}$$

$$\sigma^{*2} = w_1^2 \sigma^2, \tag{4.2.17}$$

$$\text{var}(\beta_1^*) = w_1^2 \, \text{var}(\beta_1), \tag{4.2.18}$$

$$\text{var}(\beta_2^*) = \left(\frac{w_1}{w_2} \right)^2 \text{var}(\beta_2), \tag{4.2.19}$$

$$r_{xy}^2 = r_{x^* y^*}^2. \tag{4.2.20}$$

Із цих формул зрозуміло, як за наслідками регресійного аналізу в одних одиницях вимірювання перейти до інших одиниць вимірювання при заданих значеннях масштабного чинника.

Числовий приклад. Співвідношення між GPDI і GNP в США

Наведемо результати регресійного аналізу за даними табл. 4.2.
GPDI і GNP вимірюються в мільярдах доларів:

$$Y_i = -37,0015205 + 0,17395 X_i$$

$$(76,261127) \quad (0,05406) \qquad R^2 = 0,5641$$

GPDI і GNP вимірюються в мільйонах доларів:

$$Y_i = -37,0015205 + 0,17395 X_i$$

$$(76261,1278) \quad (0,05406) \qquad R^2 = 0,5641$$

GPDI – в мільярдах доларів, а GNP – мільйонах доларів:

$$Y_i = -37,0015205 + 0,00017395 X_i$$

$$(76,261127) \quad (0,00005406) \qquad R^2 = 0,5641$$

GPDI – в мільйонах доларів, а GNP – мільярдах доларів:

$$Y_i = -37001,5205 + 173,95 X_i$$

$$(76261,127) \quad (54,06) \qquad R^2 = 0,5641$$

4.3. Функціональний вид регресійної моделі

Як ми відзначали раніше, розглядувані нами регресійні моделі повинні бути лінійні за параметрами. При цьому вони можуть і не бути лінійними за змінними. У наступних розділах буде приділена увага широко застосовуваним регресійним моделям, які можуть бути нелінійними за змінними, але лінійними за параметрами. Зокрема, розглядатимуться такі моделі:

1) лінійно-логарифмічна;
2) напівлогарифмічна;
3) зворотні.

4.4. Вимірювання еластичності. Лінійно-логарифмічна модель

Розглянемо експоненціальну регресійну модель

$$Y_i = \beta_1 X_i^{\beta_2} e^{u_i}, \qquad (4.4.1)$$

яка після логарифмування обох частин може бути подана у вигляді

$$\ln Y_i = \ln \beta_1 + \beta_2 \ln X_i + u_i. \qquad (4.4.2)$$

Якщо ми запишемо (4.4.2) у вигляді

$$\ln Y_i = \alpha + \beta_2 \ln X_i + u_i, \qquad (4.4.3)$$

де $\alpha = \ln \beta_1$, то побачимо, що ця модель лінійна за параметрами α і β_2, а також за логарифмами змінних X і Y. Отже, для знаходження α і β_2 може бути застосований МНК. Така модель ще називається “подвійною логарифмічною” або “логарифмічною лінійною”.

Якщо припущення класичної лінійної регресійної моделі виконуються, то параметри (4.4.3) можуть бути оцінені за МНК із рівняння

$$Y_i^* = \alpha + \beta_2 X_i^* + u_i, \qquad (4.4.4)$$

де $Y_i^* = \ln Y_i$, $X_i^* = \ln X_i$. Отримані за МНК оцінки α і β_2 будуть найкращими незміщеними лінійними оцінками для α і β_2, відповідно.

Перевагою цієї моделі є те, що кутовий коефіцієнт β_2 є мірою еластичності *Y* по відношенню до *X*, тобто визначає відсоток зміни *Y* для даного (малого) відсотка зміни *X*. Так, якщо *Y* зображує попит на товар, а *X* – ціну одиниці товару, то β_2 вимірює величину еластичності попиту за ціною, параметр ln*Y*, що становить в економіці значний інтерес.

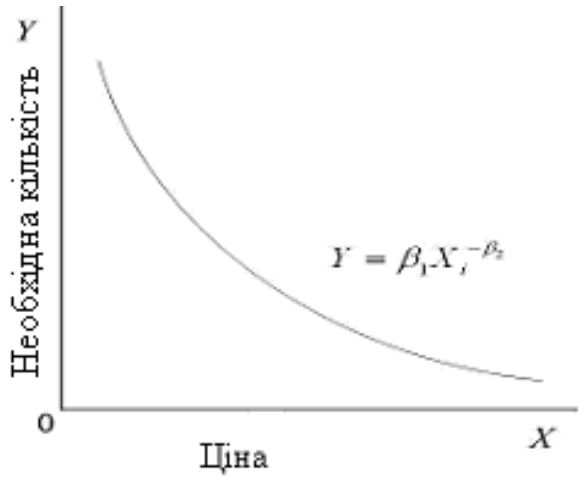


Рис. 4.2. Експоненціальна модель

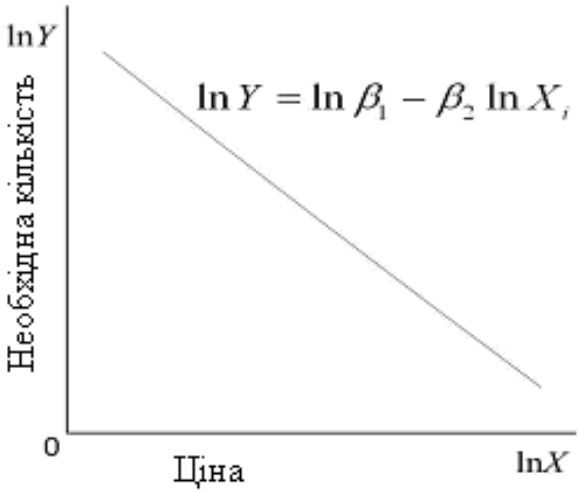


Рис. 4.3. Лінійно-логарифмічна модель

Якщо співвідношення між величиною попиту і ціною таке, як зображено на рис. 4.2, то подвійне логарифмічне перетворення даватиме оцінку еластичності ціни ($-\beta_2$).

Слід зазначити два особливі моменти лінійної логарифмічної моделі: ця модель припускає, що коефіцієнт еластичності між Y і X (β_2) залишається постійним на всьому проміжку зміни X. Цю властивість можна перевірити, оскільки еластичність Y по X обчислюється за формулою

$$\frac{dY}{dX} \cdot \frac{X}{Y} = \frac{d \ln Y}{d \ln X}.$$

Якщо підставити в неї $Y = \beta_1 X^{\beta_2}$, то отримаємо

$$\frac{dY}{dX} \cdot \frac{X}{Y} = \beta_2.$$

Ця властивість пояснює, чому дана модель називається моделлю з постійною еластичністю. Іншими словами, зміна ln$Y$ при одиничній зміні ln$X$, тобто еластичність β_2, залишається незмінною незалежно від точки ln$X$, в якій проводиться вимірювання (рис.4.3). Іншою особливістю моделі є те, що, хоча α і β_2 є незміщеними оцінками α і β_2, β_1 (параметр, що входить у початкову модель $\beta_1 = \exp(\alpha)$) є зміщеною оцінкою. У більшості практичних задач, проте, цей член має другорядне значення і немає необхідності в отриманні незміщеної оцінки.

Ілюстративний приклад. Попит на каву

Провівши обчислення за даними, наведеними в таблиці, одержуємо такий результат:

$$\begin{array}{rrll} \ln Y_t = 0,7774 & - & 0,2530 \ln X_t & R^2 = 0,7448 \\ (0,0152) & & (0,0494) & F_{1,9} = 26,27 \\ t = (51,0045) & & (-5,1251) & \end{array} \qquad (4.4.5)$$

де *Yt* – попит на каву (кількість випитих чашок кави за день), а *Xt* – ціна на каву в доларах за фунт кави.

З цих результатів ми бачимо, що коефіцієнт еластичності ціни дорівнює –0,25. Це значить, що при зростанні на 1% реальної ціни за фунт кави, попит на неї зменшиться в середньому на 0,25%. Оскільки еластичність ціни менше 1 за абсолютною величиною, ми можемо сказати, що попит на каву нееластичний за ціною.

Цікавим є наступне питання. Чи можемо ми визначити, яка з моделей є кращою, порівнюючи результати лінійної та лінійно-логарифмічної функції попиту? Чи можемо ми сказати, що (4.4.5) краще, ніж (3.7.1), оскільки її R^2 вище (0,7448 проти 0,6628)? На жаль, ми не можемо зробити такий висновок, оскільки в тому випадку, коли залежна змінна у двох моделях неоднакова (тут *Y* і ln*Y*), коефіцієнти детермінації безпосередньо порівнювати не можна. Ми не можемо порівнювати два кутових коефіцієнти ще й з іншої причини. Для моделі (3.7.1) кутовий коефіцієнт дає ефект одиничної зміни ціни на каву, скажемо на 1 дол. за фунт, на постійну абсолютну величину зменшення споживання, що складає 0,4795 чашки кави за день. З іншого боку, коефіцієнт –0,2530, отриманий в (4.4.5), дає постійне процентне зменшення споживання кави в результаті 1%-го зростання ціни на каву за фунт, тобто дає еластичність ціни.

Чи можна порівняти результати двох використаних моделей? Одним із шляхів порівняння двох моделей є наближене визначення еластичності ціни для моделі (2.7.1). Це може бути зроблено таким чином.

Еластичність *E* змінної *Y* по відношенню до іншої змінної *X* визначається за формулою:

$$E = \frac{\%\text{зміна Y}}{\%\text{зміна X}} = \frac{(\Delta Y / Y)\cdot 100}{(\Delta X / X)\cdot 100} = \frac{\Delta Y}{\Delta X}\cdot\frac{X}{Y} = \frac{dY}{dX}\cdot\frac{X}{Y}. \qquad (4.4.6)$$

Для лінійної моделі (2.7.1) оцінка кутового коефіцієнта β_2 дорівнює –0,4795. Як показано в (4.4.6), для знаходження еластичності необхідно помножити кутовий коефіцієнт β_2 на відношення *X/Y* (ціни до кількості). Але які величини *X* і *Y* повинні ми вибрати? Звичайно, ми можемо підрахувати еластичність для кожної з 11 пар значень *X* і *Y*. На практиці, проте, еластичність підраховується в середніх значеннях *X* і *Y*. Ми одержуємо оцінку середньої еластичності. Для нашого прикладу $\overline{Y} = 2,206$ чашки і $\overline{X} = 1,011$ дол. Використовуючи ці величини і $\beta_2 = -0,4795$, одержуємо з (4.4.6) середню еластичність –0,2197 або наближено –0,22. Цей результат отриманий із лінійної моделі й відрізняється від коефіцієнта еластичності –0,25, отриманого в лінійній логарифмічній моделі. Зазначимо, що цей коефіцієнт еластичності є постійним у всіх точках *X*, тоді як у лінійній моделі коефіцієнт еластичності змінюється від точки до точки.

4.5. Напівлогарифмічні моделі. Визначення темпів зростання. Log-lin модель

Економісти, бізнесмени, урядовці часто стикаються з питанням визначення темпів зростання різних економічних показників, таких як населення, валовий національний продукт, грошова маса, зайнятість, продуктивність, торговий дефіцит і под.

У табл. 4.3 наведені дані про реальний валовий внутрішній продукт США за період 1972–1991 рр.

Таблиця 4.3

Валовий внутрішній продукт у поточних доларах і доларах 1987 р.

Рік	ВВП, в поточ. млрд. дол.	ВВП, млрд. дол. 1978 р.
1972	1207,0	3107,1
1973	1349,6	3268,6
1974	1458,6	3248,1
1975	1585,9	3221,7
1976	1768,4	3380,8
1977	1974,1	3533,3
1978	2232,7	3703,5
1979	2488,6	3796,8
1980	2708,0	3776,3
1981	3030,6	3843,1
1982	3149,6	3760,3
1983	3405,0	3906,6
1984	3777,2	4148,5
Продовження табл. 4.3		
1985	4038,7	4279,8
1986	4268,6	4404,5
1987	4539,9	4539,9
1988	4900,4	4718,6
1989	5250,8	4838,0
1990	5522,2	4877,5
1991	5677,5	4821,0

Припустимо, ми хочемо визначити за цей період темпи зростання внутрішнього валового продукту. Позначимо через *Yt* реальний валовий внутрішній продукт у момент часу *t*, а Y_0 – початкове (в 1972 р.) значення цієї змінної. Застосуємо добре відому формулу складних відсотків

$$Y_t = Y_0(1+r)^t, \tag{4.5.1}$$

де r – складний відсоток приросту Y. Логарифмуючи обидві частини рівняння (4.5.1), одержуємо

$$\ln Y_t = \ln Y_0 + t\ln(1+r). \tag{4.5.2}$$

Позначимо тепер

$$\beta_1 = \ln Y_0, \tag{4.5.3}$$

$$\beta_2 = \ln(1+r). \tag{4.5.4}$$

Рівність (4.5.2) можна переписати у вигляді

$$\ln Y_t = \beta_1 + \beta_2 t\,. \qquad (4.5.5)$$

Додаючи в (4.5.5) випадкову складову, одержуємо

$$\ln Y_t = \beta_1 + \beta_2 t + u_t\,. \qquad (4.5.6)$$

Ця модель схожа на звичайну модель лінійної регресії. Єдина різниця полягає в тому, що регресант входить у рівняння як логарифм *Y*, а регресор, час *t*, набуває значення 1, 2, 3 і т.д.

Модель вигляду (4.5.6) називається напівлогарифмічною моделлю, оскільки одна зі змінних (у даному випадку регресант) входить у неї у вигляді логарифма. Подібну модель називають також log-lin моделлю. Пізніше ми розглянемо модель, у якій регресант входить лінійно, а регресор – у вигляді логарифма. Така модель називається lin-log-моделлю.

Перш ніж звернутися до результатів регресії, дослідимо властивості моделі (4.5.5). У цій моделі кутовий коефіцієнт визначає постійну відносну зміну *Y* для даної абсолютної зміни регресора (у даному випадку змінну *t*), тобто

$$\beta_2 = \frac{\text{відносна зміна регресанта}}{\text{відносна зміна регресора}}\,. \qquad (4.5.7)$$

Якщо помножити відносну зміну *Y* на 100, то (4.5.7) дасть нам процентну зміну або темп зростання *Y* для абсолютної зміни регресора *X*.

Модель типу (4.5.5) є особливо корисною в разі, коли змінна *X* – час, як у нашому випадку з валовим національним продуктом. У такому випадку модель визначає постійний (β_2) темп зростання (якщо $\beta_2 > 0$) або зменшення ($\beta_2 < 0$) змінної *Y*. Тому ця модель ще називається моделлю постійного зростання.

Повертаючись, наприклад, до реального валового внутрішнього продукту, ми одержуємо такі результати:

$$\begin{array}{rcll} \ln Y_t = 8{,}0139 & + & 0{,}02469t & \\ (0{,}0114) & & (0{,}00956) & R^2 = 0{,}9738 \\ t = (700{,}54) & & (25{,}8643) & \\ p = (0{,}0000) & & (0{,}0000) & \end{array} \qquad (4.5.8)$$

Інтерпретація результатів. За період 1972–1991 рр. реальний валовий внутрішній продукт мав річний темп приросту 2,469%. Оскільки $\ln Y_0 = 8{,}0139$ то $Y_0 = e^{8{,}0139} \approx 3022{,}7$, тобто на початку 1972 р. оцінка реального внутрішнього валового продукту складала приблизно 3 023 млрд дол. Лінія регресії показана на рис. 4.4.

Кутовий коефіцієнт 0,02469, отриманий в (4.5.8), або в загальному випадку коефіцієнт β_2 в моделі зростання (4.5.5) дає миттєвий (для даного моменту часу) темп зростання, а не складний (за даний період часу) відсоток приросту. Але останній можна легко отримати з рівняння (4.5.4). Оскільки $\beta_2 = \ln(1+r)$, то $r = \exp(\beta_2) - 1$. Підставляючи сюди $\beta_2 = 0{,}02469$, отримуємо таке значення оцінки складного відсотка: $\hat{r} = 0{,}024997$ або близько 2,499%. Таким чином, за дослі-

джуваний період складний відсоток темпів зростання валового внутрішнього продукту складає приблизно 2,499% на рік. Цей темп зростання трохи вищий, ніж миттєвий темп зростання 2,469%.

Модель лінійного тренда. Замість моделі (4.5.6) іноді застосовується модель

$$Y_t = \beta_1 + \beta_2 t + u_t, \qquad (4.5.9)$$

тобто замість використання в моделі регресії $\ln Y$ ми використовуємо величину Y. Як регресор знову виступає час t. Така модель називається моделлю лінійного тренда, а час t – змінною тренда. Під трендом ми розуміємо безперервний приріст або спад у характері змінної. Якщо кутовий коефіцієнт в (4.5.9) позитивний, то ми говоримо про тенденцію до підвищення Y, а при негативному значенні – до пониження Y.

Для нашого прикладу про реальний ВВП модель (4.5.9) дає такі результати:

$$\begin{array}{lll} Y_t = 2933,0538 + 97,6806t & & \\ \quad (50,5913) \quad (4,2233) & R^2 = 0,9674 & (4.5.10) \\ t = (57,9754) \quad (23,1291) & & \\ p = (0,0000) \quad (0,0000) & & \end{array}$$

Інтерпретація результатів. За період 1972–1991 рр. в середньому реальний ВВП зростав в абсолютному (не у відносному) значенні приблизно на 97,68 млрд дол. Таким чином, за цей період реальний ВВП мав тенденцію до зростання.

Вибір між моделями зростання (4.5.8) і лінійною моделлю тренда (4.5.10) залежить від того, що нас цікавить – абсолютні чи відносні зміни реального ВВП. Відзначимо, що здебільшого відносні зміни важливіші. Ще раз відзначимо, що ми не можемо порівнювати коефіцієнти детермінації моделей (4.5.8) і (4.5.10), оскільки регресанти в них різні.

Lin-Log модель. Припустимо, що у нас є дані, наведені в табл. 6.3, в якій Y – ВНП, а X – грошова маса (М2= засоби грошового обігу + внески + депозити до запитання + дорожні чеки + інші чекові внески +...). Припустимо далі, що нас цікавить, наскільки (в абсолютному значенні) зростає ВНП, якщо грошова маса зростає, скажімо, на 1%.

На відміну від тільки що розглянутої моделі зростання, в якій нас цікавив відсоток зростання Y для абсолютної одиничної зміни X, тепер нас цікавить абсолютна зміна Y для заданого відсотка зміни X. Модель, що дозволяє виконати це завдання, така:

$$Y_i = \beta_1 + \beta_2 \ln X_i + u_i. \qquad (4.5.11)$$

Дану модель називають lin-log-моделлю. Пояснимо значення коефіцієнта β_2. Як завжди

$$\beta_2 = \frac{\text{зміна Y}}{\text{зміна lnX}} = \frac{\text{зміна Y}}{\text{відносна зміна X}}.$$

Друга рівність базується на тому, що зміна логарифма величини дорівнює відносній зміні цієї величини.

У математичних символах це може бути записано у вигляді

$$\beta_2 = \frac{\Delta Y}{(\Delta X / X)}. \qquad (4.5.12)$$

Рівняння (4.5.12) можна записати і в еквівалентному вигляді

$$\Delta Y = \beta_2 \left(\frac{\Delta X}{X} \right). \qquad (4.5.13)$$

Це рівняння говорить про те, що абсолютна зміна *Y* (тобто ΔY) дорівнює добутку β_2 й відносної зміни *X*. Якщо останнє (тобто $\Delta X/X$) помножити на 100, то (4.5.13) дасть абсолютну зміну *Y* для процентної зміни *X*. Так, якщо $\Delta X/X$ зміниться на 0,01 (або 1%), то абсолютна зміна *Y* буде дорівнювати $0{,}011\,\beta_2$. Таким чином, якщо при розв'язанні задачі регресії знайдено β_2=500, то абсолютна зміна *Y* дорівнює 0,01×500 або 5. Отже, при використовуванні моделі регресії (4.5.11) після знаходження за МНК кутового коефіцієнта β_2 для визначення абсолютної зміни *Y* при зміні *X* на 1% необхідно β_2 помножити на 0,01.

Таблиця 4.4

ВНП і грошова маса в США, 1973–1987 рр.

Рік	ВНП, млрд дол.	Грошова маса, млрд дол.
1973	1359,3	861,0
1974	1472,8	908,5
1975	1598,4	1023,2
1976	1782,8	1163,7
1977	1990,5	13286,7
1978	2249,7	1389,0
1979	2508,2	1500,2
1980	2723,0	1633,1
1981	3052,6	1795,5
1982	3166,0	1954,5
1983	3405,7	2185,2
1984	3772,2	2363,6
1985	4014,4	2562,6
1986	4240,3	2807,7

1987	4526,7	2901,0

Використовуючи дані, наведені в табл. 4.4, можна отримати такі результати:

$$\begin{array}{lll} Y_t = -16329,21 + 2584,785X_i & & \\ \quad (696,599) \quad (94,041) & R^2 = 0,9831 & (4.5.14) \\ t = (-23.441) \quad (27.486) & & \\ p = (0.0000) \quad (0.0000) & & \end{array}$$

Інтерпретація результатів. **Значення кутового коефіцієнта 2584,785 означає, що за даний період збільшення грошової маси на 1% призводило, в середньому, до зростання ВНП приблизно на 25,85 млрд дол. (2585×0,01=25,85).**

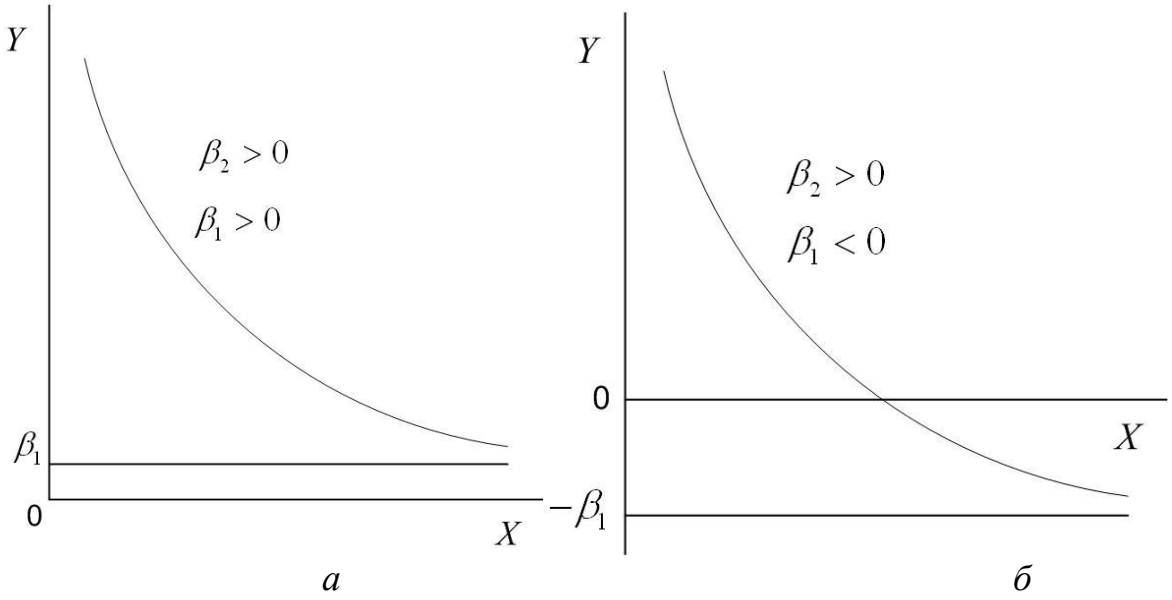
4.6. Обернені моделі

Під оберненими моделями розуміють моделі такого вигляду:

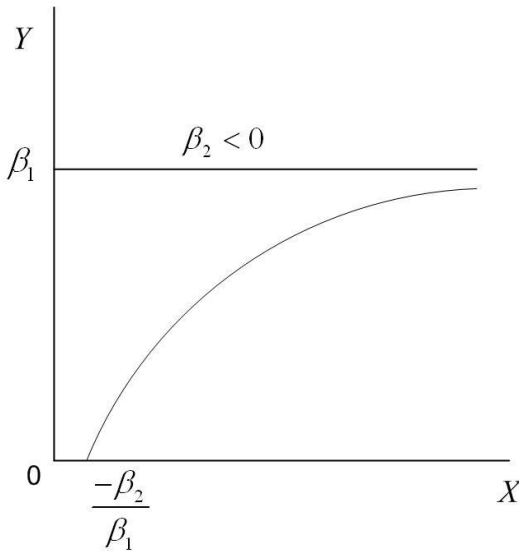
$$Y_i = \beta_1 + \beta_2 \left(\frac{1}{X_i} \right) + u_i . \qquad (4.6.1)$$

Ця модель, як і вищерозглянуті, є лінійною щодо регресора X, а також лінійна щодо коефіцієнтів регресії β_1 і β_2.

У оберненої моделі (4.6.1) є одна відмінна риса. При необмеженому зростанні змінної X складова β_2 / X_i прямує до нуля, а регресант Y_i асимптотично наближається до величини β_1. Таким чином, модель (4.6.1) містить асимптотичний параметр, до якого прямує залежна змінна Y при необмеженому зростанні змінної X.



а *б*



в

Рис. 4.4 Приклади графіків обернених моделей: *а* – моделі (4.6.1);
б – кривої Філіпса; *в* – кривої Енгеля

На рис. 4.4 зображені характерні криві оберненої моделі залежно від знаків коефіцієнтів регресії.

Одним із важливих додатків моделі (рис. 4.4, *б*) є відома з макроекономіки крива Філіпса. На підставі даних про темпи зростання заробітної плати (*Y*) і процента зміни рівня безробіття (*X*) для Великобританії в період з 1861 по 1957 рр. Філіпс отримав криву вигляду рис. 4.4, *б*, яка детально зображена на рис.4.5.

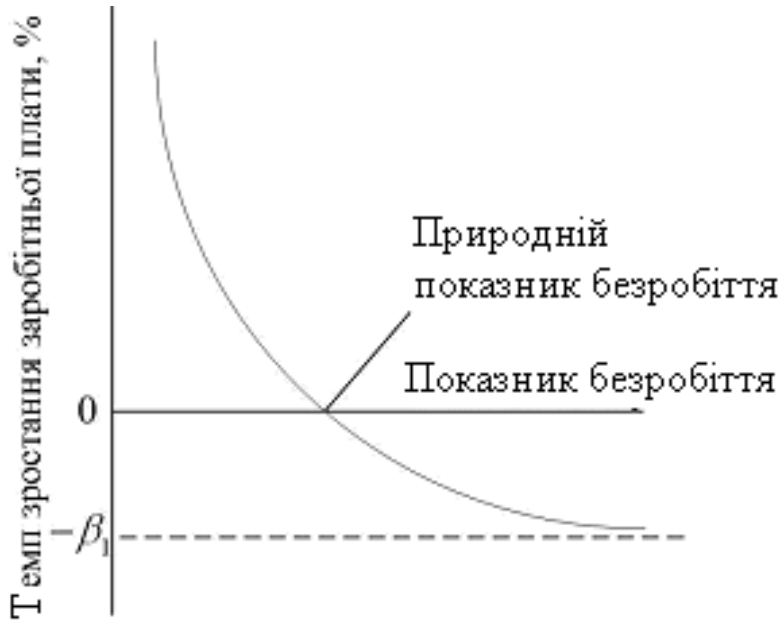


Рис. 4.5. Крива Філіпса

З рис. 4.5 бачимо, що є асиметрія в залежності темпів зростання заробітної плати від рівня безробіття: заробітна плата зростає швидше при рівні безробіття, меншому величини UN (natural rate unemployment), який називається природним рівнем безробіття. При рівні безробіття, більшому за цю величину, зменшення те-

мпів зростання зарплати стає більш плавним. При необмеженому збільшенні рівня безробіття темпи зростання зарплати наближаються до асимптотичного значення.

Важливим додатком залежності вигляду (4.6.1) є крива витрат Енгеля, що характеризує зв'язок витрат споживача на товар із його загальними витратами або доходом. Якщо позначити через Y витрати на товар, а через X – дохід споживача, то певні товари матимуть такі властивості. Існує деякий критичний або пороговий рівень доходу, нижче якого товар не купується. На рис. 4.4, *в* цей рівень має значення $-\beta_2 / \beta_1$. Існує рівень насичення товару, вище цього рівня споживання товару не відбувається, яким би високим не був рівень доходу. Цей рівень у моделі (4.6.1) визначається асимптотою $Y = \beta_1$. Для таких товарів найбільш відповідною є модель, подана на рис. 4.4, *в*.

Ілюстрований приклад. Крива Філіпса для Великобританії, 1950–1966 рр.

У табл. 4.5 наведені дані про щорічні темпи зростання заробітної плати Y і рівня безробіття X у Великобританії в період 1950–1966 рр.

Таблиця 4.5

Темпи зростання зарплати й рівня безробіття у Великобританії, 1950–1966 рр.

Рік	Темп зростання зарплати, %	Рівень безробіття, %
1950	1,8	1,4
1951	8,5	1,1
1952	8,4	1,5
1953	4,5	1,5
1954	4,3	1,2
1955	6,4	1
1956	8	1,1
1957	5	1,3
1958	3,6	1,8
1959	2,6	1,9
1960	2,6	1,5
1961	4,2	1,4
1962	3,6	1,8
1963	3,7	2,1
1964	4,8	1,5
1965	4,3	1,3
1966	4,6	1,4

Застосування моделі (4.6.1) дає такі результати:

$$Y_i = -1,42818 \quad + \quad 8,7243442\left(\frac{1}{X_i}\right)$$

$$\begin{array}{lll} & (2,067478) & (2,8477792) \quad R^2 = 0,3849 \\ t = & (0,690782) & (3,0635606) \quad R_{1,15} = 9,3854 \\ p = & (0,500253) & (0,0078824) \quad p = (0,00788) \end{array}$$

Згідно з цими результатами граничний рівень зниження зарплати за рік дорівнює –1,43%. Тобто, якщо X необмежено зростає, зниження заробітної плати буде не більше ніж 1,43% на рік.

Як підсумок наведемо табл. 4.6, що містить основні результати розглянутих нами моделей.

Таблиця 4.6

Основні формули для нелінійних за змінними моделей

Модель	Вигляд	Кутовий коефіцієнт $\frac{dY}{dX}$	Еластичність $\frac{dY}{dX}\frac{X}{Y}$
Лінійна	$Y = \beta_1 + \beta_2 X$	β_2	$\beta_2\left(\frac{X}{Y}\right)$
Log-log	$\ln Y = \beta_1 + \beta_2 \ln X$	$\beta_2\left(\frac{Y}{X}\right)$	β_2
Log-lin	$\ln Y = \beta_1 + \beta_2 X$	$\beta_2 Y$	$\beta_2 X$
Lin-log	$Y = \beta_1 + \beta_2 \ln X$	β_2 / X	β_2 / Y
Обернена	$Y = \beta_1 + \beta_2 / X$	$-\beta_2 / X^2$	$-\beta_2 (XY)$

Якщо вираз для коефіцієнта еластичності залежить від змінних, то на практиці звичайно використовуються їх середні значення.

4.7. Зауваження щодо стохастичної складової

Розглянемо модель регресії, у якій порівняно з (4.4.1) відсутня стохастична складова:

$$Y_i = \beta_1 X^{\beta_2}. \qquad (4.7.1)$$

З метою проведення регресійного аналізу цю модель можна подати в трьох різних видах:

$$Y_i = \beta_1 X_i^{\beta_2} u_i; \qquad (4.7.2)$$

$$Y_i = \beta_1 X_i^{\beta_2} \exp(u_i); \tag{4.7.3}$$

$$Y_i = \beta_1 X_i^{\beta_2} + u_i. \tag{4.7.4}$$

Після логарифмування цих рівнянь одержуємо

$$\ln Y_i = \alpha + \beta_2 \ln X_i + \ln u_i; \tag{4.7.5}$$

$$\ln Y_i = \alpha + \beta_2 \ln X_i + u_i;$$

$$\ln Y_i = \ln\left(\beta_1 X_i^{\beta_2} + u_i\right),$$

де $\alpha = \ln \beta_1$.

Модель типу (4.7.2) приводиться до лінійної (за параметрами) моделі регресії в тому розумінні, що шляхом відповідного перетворення (логарифмування) вона може бути приведена до лінійної щодо параметрів α і β_2 моделі. Зазначимо, що вона нелінійна щодо параметра β_1.

Хоча моделі (4.7.2) і (4.7.3) є лінійними моделями регресії і їх оцінка може проводитися за МНК, слід велику увагу приділити властивостям стохастичної складової, що входить у модель . Пригадаємо, що властивість якнайкращої лінійної незміщеної оцінки за МНК вимагає, щоб стохастична складова u_i мала математичне сподівання, яке дорівнює нулю, постійну дисперсію та нульову автокореляцію. При перевірці гіпотез ми також припускаємо, що u_i розподіляється за нормальним законом розподілу з математичним сподіванням і дисперсією, згаданими вище, тобто припускаємо, що $u_i \sim N(0, \sigma^2)$.

Звернемося до моделі (4.4.2). Логарифмічне перетворення приводить її до вигляду (4.4.5). Для використання класичної лінійної моделі регресії необхідно зробити припущення про те, що

$$\ln u_i \sim N(0, \sigma^2). \tag{4.7.6}$$

Отже, коли ми проводимо регресію за моделлю (4.7.5), то $\ln u_i$ повинен бути розподілений за нормальним законом розподілу з нульовим математичним сподівання і постійною дисперсією. У такому разі u_i в (4.7.2) розподілений за логарифмічно-нормальним законом розподілу з математичним сподіванням $\exp(\sigma^2/2)$ і дисперсією $\exp(\sigma^2)\left(\exp(\sigma^2)-1\right)$.

5. МНОЖИННИЙ РЕГРЕСІЙНИЙ АНАЛІЗ. ЗАДАЧА ОЦІНЮВАННЯ

Вивчена раніше регресійна модель із двома змінними часто виявляється на практиці неадекватною. Наприклад, у нашому випадку моделі "споживання – дохід" передбачалося, що дохід *X* впливає на витрати *Y*. Але економічна теорія рідко буває настільки простою, оскільки окрім доходу чимала кількість інших змінних може впливати на витрати й споживання. Очевидним прикладом є заощадження покупця. Іншим прикладом є залежність попиту на товар не тільки від ціни цього товару, але й від ціни конкуруючих товарів, доходу покупця, його соціального статусу і т.д. Отже, ми повинні поширити нашу просту модель з двома змінними на випадок з великою кількістю змінних. Додавання додаткових змінних приводить нас до обговорення множинної регресійної моделі, тобто такої моделі, у якій залежна змінна, або регресант *Y* залежить від двох або більше змінних.

Найпростішою з таких моделей є модель із трьома змінними – однією пояснюваною і двома пояснювальними. Перейдемо до вивчення цієї моделі. При цьому ми вважатимемо, що множинна регресійна модель лінійна за параметрами, хоча може бути нелінійною за змінними.

5.1. Модель із трьома змінними. Позначення і гіпотези

Узагальнюючи функцію PRF з двома змінними, ми можемо записати її для трьох змінних таким чином:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i, \qquad (5.1.1)$$

де *Y* – залежна змінна; X_2 і X_3 – пояснювальні змінні, або регресори; *u* – стохастичний збурююча складова; *i* – спостереження; для випадку тимчасових рядів замість індексу *i* використовується індекс *t*.

У (5.1.1) β_1 – складова, що визначає точку перетину. Як завжди, він має середню дію на *Y* всіх змінних, виключених із моделі, хоча його механічна інтерпретація є середня величина *Y* при X_2 і X_3, що дорівнюють нулю. Коефіцієнти β_1 і β_2 зазвичай називають частинними регресійними коефіцієнтами.

Ми продовжуватимемо оперувати в рамках схеми класичної лінійної регресійної моделі (CLRM). Зокрема, ми припускаємо, що:

- середня величина u_i дорівнює нулю:

$$E(u_i / X_{2i}, X_{3i}) = 0 \text{ для всіх } i; \qquad (5.1.2)$$

- серійна кореляція відсутня:

$$\operatorname{cov}(u_i, u_j) = 0, \qquad i \neq j; \qquad (5.1.3)$$

- має місце гомоскадастичність:

$$D(u_i) = \sigma^2; \qquad (5.1.4)$$

- коваріація u_i і кожної змінної *X* дорівнюють нулю:

$$\operatorname{cov}(u_i X_{2i}) = \operatorname{cov}(u_i X_{3i}) = 0; \qquad (5.1.5)$$

- модель коректно специфікована; (5.1.6)

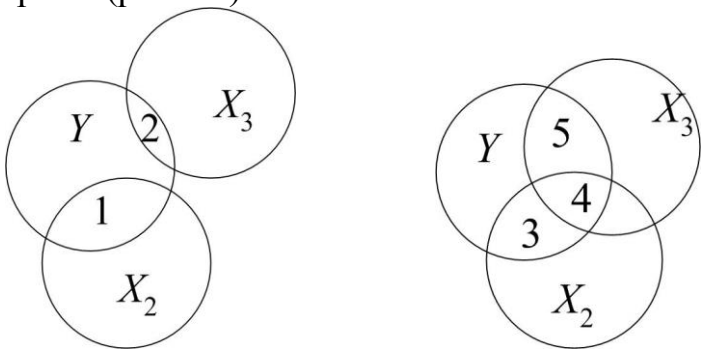
- відсутня точна колінеарність між змінними X, тобто

відсутній точний лінійний зв'язок між X_2 і X_3. (5.1.7)

Крім того, ми припускаємо, що множинна регресійна модель лінійна за параметрами, величини регресорів фіксовані в повторних вибірках і регресори змінюються достатньою мірою.

Логічна обґрунтованість гіпотез (5.1.2) – (5.1.6) та ж, що була наведена в розд. 3.2. Гіпотеза (5.1.7) про відсутність точного лінійного зв'язку між X_2 і X_3, відома як гіпотеза про відсутність колінеарності або мультиколінеарності, якщо є можливість існування більше ніж одного лінійного співвідношення, є новою і потребує деякого пояснення.

Відсутність колінеарності означає, що ніяка з пояснювальних змінних не може бути зображена у вигляді лінійної комбінації решти змінних. Це можна пояснити на прикладі діаграми (рис.5.1).



а *б*

Рис. 5.1. Діаграма, що демонструє відсутність (*а*) і наявність колінеарності (*б*)

На рис.5.1 коло Y зображає варіацію залежної змінної Y, а кола X_2 і X_3 зображають, відповідно, варіації регресорів X_2 і X_3. На рис.5.1, *а* область 1 зображає варіацію в Y, пояснену за рахунок X_2, а область 2 – варіацію в Y, пояснену за рахунок X_3. Області 3 і 4 (рис.5.1, *б*) зображають варіацію в Y, пояснену через X_2, а області 4 і 5 – пояснену через X_3. Але, оскільки область 4 є загальною для X_2 і X_3, ми не можемо назвати *a priori*, яка частина 4 належить X_2, а яка – X_3. Загальна область 4 представляє випадок колінеарності. Гіпотеза відсутності колінеарності вимагає відсутності перекриття між X_2 і X_3, тобто загальна область 4 повинна бути нульовою. Іншими словами, ми хочемо мати справу із ситуацією, зображеною на рис.5.1, *а*.

Формально відсутність колінеарності означає, що немає таких чисел λ_2 і λ_3, які одночасно не є нулями, що

$$\lambda_2 X_{2i} + \lambda_3 X_{3i} = 0. \quad (5.1.8)$$

Якщо ж подібне лінійне співвідношення виконується, тоді говорять, що X_2 і X_3 колінеарні або лінійно залежні. Якщо ж (5.1.8) виконується тільки при $\lambda_2 = 0$ і $\lambda_3 = 0$, то говорять, що X_2 і X_3 лінійно незалежні.

Так, якщо

$$X_{2i} = -4X_{3i} \text{ або } X_{2i} + 4X_{3i} = 0, \tag{5.1.9}$$

то дві змінні лінійно залежні, і якщо обидві вони включені в регресійну модель, ми матимемо точну колінеарність або точну лінійну залежність між двома змінними.

Припустимо тепер, що $X_{3i} = X_{2i}^2$. Чи буде при цьому порушена гіпотеза відсутності колінеарності? Ні, оскільки зв'язок між регресорами нелінійний і не порушується вимога про відсутність лінійного зв'язку між регресорами.

Пояснимо обґрунтованість гіпотези відсутності колінеарності на прикладі. Припустимо, що в (5.1.1) Y, X_2 і X_3 представляють витрати на споживацькі товари, дохід і заощадження покупця відповідно. Постулюючи лінійний зв'язок між споживацькими витратами, доходом і заощадженнями, економічна теорія припускає, що дохід і заощадження мають деякий незалежний вплив на витрати. Якщо це не так, то не має сенсу включати їх у модель по окремості. На противагу цьому, якщо існує точна лінійна залежність між доходом і накопиченням, то ми маємо всього одну незалежну змінну, а не дві, і немає способу оцінити роздільний вплив на витрати доходу і заощаджень. Для більшої виразності покладемо $X_{3i} = 2X_{2i}$ в регресійній моделі «витрати-прибуток-заощадження». Тоді (5.1.1) можна подати у вигляді

$$\begin{aligned} Y_i &= \beta_1 + \beta_2 X_{2i} + \beta_3 (2X_{2i}) + u_i = \\ &= \beta_1 + (\beta_2 + 2\beta_3) X_{2i} + u_i = \beta_1 + \alpha X_{2i} + u_i, \end{aligned} \tag{5.1.10}$$

де $\alpha = \beta_2 + 2\beta_3$. Тобто ми фактично маємо регресійну модель із двома змінними, а не з трьома. Крім того, якщо ми застосовуємо (5.1.10) і отримаємо α, то не зможемо оцінити роздільний вплив X_2 (=β_2) і X_3 (=β_3) на Y.

Підводячи підсумки зазначимо, що гіпотеза про відсутність мультиколінеарності вимагає, щоб у функцію PRF включали тільки ті змінні, які не є лінійними функціями деяких інших змінних моделі.

5.2. Інтерпретація рівняння множинної регресії

Із гіпотез класичної моделі регресії витікає, що якщо узяти математичне сподівання від обох частин рівності (5.1.1), то отримаємо

$$E(Y_i \mid X_{2i}, X_{3i}) = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i}. \tag{5.2.1}$$

Рівність (5.2.1) дає математичне сподівання Y при фіксованих значеннях змінних X_2 і X_3. Отже, як і у випадку з двома змінними, множинний регресійний аналіз є регресійним аналізом за умови фіксованості пояснювальних змінних, у результаті ми отримуємо середнє значення Y для фіксованих змінних X.

5.3. Значення частинних коефіцієнтів регресії

Значення частинних коефіцієнтів полягає в такому: β_2 є мірою зміни середньої величини Y, $E(Y \mid X_2, X_3)$ при одиничній зміні X_2 за умови того, що X_3 не змінюється. Іншими словами, β_2 дає степінь нахилу $E(Y \mid X_2, X_3)$ по відношенню до X_2, коли X_3 не змінюється. Зазначимо також, що β_1 і β_2 є частинними похідними $E(Y \mid X_2, X_3)$ по X_2 і X_3 відповідно. Аналогічно β_3 є мірою зміни середньої величини $У$ при одиничній зміні X_3, коли X_2 не змінюється. Тобто він дає «прямий», або «нетто», ефект одиничної зміни X_3 на середню величину Y за вирахуванням впливу X_2.

Яке точне значення терміна «залишаючи постійною»? Щоб зрозуміти це, припустимо, що Y характеризує випуск продукції, а X_2 і X_3 – трудовитрати й капітал відповідно. Припустимо далі, що для виробництва Y потрібні і X_2, і X_3, і що пропорція, у якій вони можуть бути використані у виробництві Y, може варіюватися. Тепер припустимо, що ми збільшили на одиницю трудовитрати, очевидно, це приведе до деякого збільшення випуску продукції. Ми не можемо пояснювати результуючу зміну випуску продукції виключно збільшенням витрат X_2, бо в такому разі збільшиться внесок X_2 в Y, оскільки X_2 одержує кредит за частину зміни в Y внаслідок збільшення капіталу. Отже, щоб оцінити «істинний внесок» X_2 в зміну Y, ми повинні контролювати вплив від X_2.

Процедура подібного контролю така. Припустимо, що ми хочемо контролювати лінійний вплив капіталу X_3 на випуск продукції, при зміні трудовитрат X_2 на одиницю. Здійснюємо це в три етапи.

1. Регресуємо Y тільки за X_3 таким чином:

$$Y_i = b_1 + b_{13} X_{3i} + u_{1i}. \qquad (5.3.1)$$

Рівняння (5.3.1) являє собою рівняння двовимірної регресії. При цьому індекс 1 стосується змінної Y, а u_{1i} позначає залишкову складову за наслідками вибірки.

2. Регресуємо X_2 тільки за змінною X_3:

$$X_{2i} = b_2 + b_{23} X_{3i} + \hat{u}_{2i}, \qquad (5.3.2)$$

де $\hat{u}_{2i}$ – залишкова складова. Маємо

$$\hat{u}_{1i} = Y_i - b_1 - b_{13} X_{3i} = Y_i - \hat{Y}_i; \qquad (5.3.3)$$

$$\hat{u}_{2i} = X_{2i} - b_2 - b_{23} X_{3i} = X_{2i} - \hat{X}_{2i}, \qquad (5.3.4)$$

де $\hat{Y}_i$ і $\hat{X}_{2i}$ – оцінені величини з регресійних моделей (5.3.1) і (5.3.2) відповідно.

Яке значення залишків $\hat{u}_{1i}$ і $\hat{u}_{2i}$? Складова $\hat{u}_{1i}$ позначає величину Y_i після вилучення лінійного впливу на неї X_3. Аналогічно $\hat{u}_{2i}$ позначає величину X_{2i} після вилучення впливу на неї X_3. Отже, $\hat{u}_{1i}$ і $\hat{u}_{2i}$ – «звільнені від впливу X_3 величини» Y_i і X_{2i}.

3. Отже, якщо ми проведемо регресійний аналіз $\hat{u}_{1i}$ і $\hat{u}_{2i}$ у вигляді

$$\hat{u}_{1i} = a_0 + a_1\hat{u}_{2i} + \hat{u}_{3i}, \qquad (5.3.5)$$

де $\hat{u}_{3i}$ – залишкова складова по вибірці, то a_1 повинен дати оцінку «істинного, або звільненого, ефекту» від одиничної зміни X_2 на Y або істинний нахил Y по відношенню до X_2, тобто оцінку β_2.

За МНК a_1 можна подати у вигляді

$$a_1 = \frac{\sum(\hat{u}_{1i} - \bar{\hat{u}}_1)(\hat{u}_{2i} - \bar{\hat{u}}_2)}{\sum(\hat{u}_{2i} - \bar{\hat{u}}_2)^2} = \frac{\sum\hat{u}_{1i}\hat{u}_{2i}}{\sum\hat{u}_{2i}^2}.$$

Оскільки $\bar{\hat{u}}_1 = \bar{\hat{u}}_2 = 0$, то рівняння (5.3.1) і (5.3.2) можна переписати у вигляді

$$y_i = b_{13}x_{3i} + \hat{u}_{1i};$$

$$x_{2i} = b_{23}x_{3i} + \hat{u}_{2i},$$

де малі букви використовуються для позначення величин у формі відхилень. Отримаємо з цих рівнянь $\hat{u}_{1i}$ і $\hat{u}_{2i}$:

$$\hat{u}_{1i} = y_i - b_{13}x_{3i};$$

$$\hat{u}_{2i} = x_{2i} - b_{23}x_{3i}$$

і підставимо їх у вирази для a_1:

$$a_1 = \frac{\sum\hat{u}_{1i}\hat{u}_{2i}}{\sum\hat{u}_{2i}^2} = \frac{\sum(y_i - b_{13}x_{3i})(x_{2i} - b_{23}x_{3i})}{\sum(x_{2i} - b_{23}x_{3i})^2} =$$

$$= \frac{\sum y_i x_{2i} - b_{23}\sum y_i x_{3i} - b_{13}\sum x_{2i}x_{3i} + b_{13}b_{23}\sum x_{3i}^2}{\sum x_{2i}^2 + b_{23}^2\sum x_{3i}^2 - 2b_{23}\sum x_{2i}x_{3i}}.$$

Враховуючи, що

$$b_{23} = \frac{\sum x_{2i}x_{3i}}{\sum x_{3i}^2},\ b_{13} = \frac{\sum y_i x_{3i}}{\sum x_{2i}^2},$$

одержуємо

$$a_1 = \frac{\left(\sum y_i x_{2i}\right)\left(\sum x_{3i}^2\right) - \left(\sum y_i x_{3i}\right)\left(\sum x_{2i}x_{3i}\right)}{\left(\sum x_{2i}^2\right)\left(\sum x_{3i}^2\right) - \left(\sum x_{2i}x_{3i}\right)^2}.$$

Як буде показано в наступному параграфі, цей вираз збігається з оцінкою β_2 за МНК. Таким чином, на практиці немає необхідності здійснювати всі розрахунки, оскільки a_1 обчислюється за достатньо простими формулами.

5.4. Оцінка частинних коефіцієнтів регресії за МНК

Для знаходження за МНК оцінок для параметрів запишемо SRF, що відповідає PRF з (5.1.1), у вигляді

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{u}_i, \tag{5.4.1}$$

де $\hat{u}_i$ – залишкова складова, відповідна стохастичній збурюючій складовій в u_i.

Як було відзначено раніше, процедура МНК полягає у виборі величин невідомих параметрів таким чином, щоб сума квадратів залишків (RSS) $\sum \hat{u}_i^2$ була якомога меншою. Із (5.4.1) одержуємо

$$\sum \hat{u}_i^2 = \sum \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i}\right)^2. \tag{5.4.2}$$

Відповідно до процедури МНК знаходимо частинні похідні:

$$\frac{\partial \sum \hat{u}_i^2}{\partial \hat{\beta}_1} = -2\sum \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i}\right) = -2\sum \hat{u}_i;$$

$$\frac{\partial \sum \hat{u}_i^2}{\partial \hat{\beta}_2} = -2\sum \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i}\right) X_{2i} = -2\sum \hat{u}_i X_{2i};$$

$$\frac{\partial \sum \hat{u}_i^2}{\partial \hat{\beta}_3} = -2\sum \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i}\right) X_{3i} = -2\sum \hat{u}_i X_{3i}.$$

Прирівнюючи до нуля ці вирази, одержуємо систему алгебраїчних рівнянь щодо коефіцієнтів регресії:

$$\overline{Y} = \hat{\beta}_1 + \hat{\beta}_2 \overline{X}_2 + \hat{\beta}_3 \overline{X}_3; \tag{5.4.3}$$

$$\sum Y_i X_{2i} = \hat{\beta}_1 \sum X_{2i} + \hat{\beta}_3 \sum X_{2i}^2 + \hat{\beta}_3 \sum X_{2i} X_{3i}; \tag{5.4.4}$$

$$\sum Y_i X_{3i} = \hat{\beta}_1 \sum X_{3i} + \hat{\beta}_3 \sum X_{2i} X_{3i} + \hat{\beta}_3 \sum X_{3i}^2. \tag{5.4.5}$$

Із рівняння (5.4.3) одержуємо величину для $\hat{\beta}_1$:

$$\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X}_2 - \hat{\beta}_3 \overline{X}_3. \tag{5.4.6}$$

Для $\hat{\beta}_2$ і $\hat{\beta}_3$ можна отримати, розв'язуючи (5.4.3)–(5.4.5), такі вирази:

$$\hat{\beta}_2 = \frac{\left(\sum y_i x_{2i}\right)\left(\sum x_{3i}^2\right) - \left(\sum y_i x_{3i}\right)\left(\sum x_{2i} x_{3i}\right)}{\left(\sum x_{2i}^2\right)\left(\sum x_{3i}^2\right) - \left(\sum x_{2i} x_{3i}\right)^2}; \tag{5.4.7}$$

$$\hat{\beta}_3 = \frac{\left(\sum y_i x_{3i}\right)\left(\sum x_{2i}^2\right) - \left(\sum y_i x_{2i}\right)\left(\sum x_{2i} x_{3i}\right)}{\left(\sum x_{2i}^2\right)\left(\sum x_{3i}^2\right) - \left(\sum x_{2i} x_{3i}\right)^2}. \qquad (5.4.8)$$

При цьому малі букви використовуються для позначення значень змінних у відхиленнях.

Примітки:

1. *Рівняння (5.4.7) і (5.4.8) симетричні в тому значенні, що одне може бути отримане з іншого заміною X_2 на X_3 і навпаки.*
2. *Знаменники в обох формулах однакові.*
3. *Тривимірний випадок є природним узагальненням двовимірної моделі.*

На закінчення розглянемо випадок, коли модель містить *k* параметрів регресії:

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + ... + \hat{\beta}_k X_{ki} + \hat{u}_i .$$

Відповідно до МНК нам необхідно мінімізувати суму квадратів відхилень

$$\sum \hat{u}_i^2 = \sum \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i} - ... - \hat{\beta}_k X_{ki}\right)^2 .$$

За вже відомою нам процедурою одержуємо

$$\frac{\partial \sum \hat{u}_i^2}{\partial \hat{\beta}_1} = -2\sum \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i} - ... - \hat{\beta}_k X_{ki}\right) = -2\sum \hat{u}_i ;$$

$$\frac{\partial \sum \hat{u}_i^2}{\partial \hat{\beta}_2} = -2\sum \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i} - ... - \hat{\beta}_k X_{ki}\right) X_{2i} = -2\sum \hat{u}_i X_{2i}$$

;

$$\frac{\partial \sum \hat{u}_i^2}{\partial \hat{\beta}_3} = -2\sum \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i} - ... - \hat{\beta}_k X_{ki}\right) X_{3i} = -2\sum \hat{u}_i X_{3i}$$

;

.....

$$\frac{\partial \sum \hat{u}_i^2}{\partial \hat{\beta}_k} = -2\sum \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i} - ... - \hat{\beta}_k X_{ki}\right) X_{ki} = -2\sum \hat{u}_i X_{ki}$$

.

Отримуємо таку систему лінійних алгебраїчних рівнянь:

$$\sum Y_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum X_{2i} + \hat{\beta}_3 \sum X_{3i} + ... + \hat{\beta}_k \sum X_{ki} ;$$

$$\sum Y_i X_{2i} = \hat{\beta}_1 \sum X_{2i} + \hat{\beta}_2 \sum X_{2i}^2 + \hat{\beta}_3 \sum X_{2i} X_{3i} + ... + \hat{\beta}_k \sum X_{2i} X_{ki}$$

;

$$\sum Y_i X_{3i} = \hat{\beta}_1 \sum X_{3i} + \hat{\beta}_2 \sum X_{3i} X_{2i} + \hat{\beta}_3 \sum X_{3i}^2 + ... + \hat{\beta}_k \sum X_{3i} X_{ki} ;$$

.....

$$\sum Y_i X_{3i} = \hat{\beta}_1 \sum X_{3i} + \hat{\beta}_2 \sum X_{3i} X_{2i} + \hat{\beta}_3 \sum X_{3i}^2 + ... + \hat{\beta}_k \sum X_{3i} X_{ki}$$

.

Переходячи до малих букв, цю систему можна переписати в такому вигляді:

$$\sum y_i x_{2i} = \hat{\beta}_2 \sum x_{2i}^2 + \hat{\beta}_3 \sum x_{2i} x_{3i} + ... + \hat{\beta}_k \sum x_{2i} x_{ki} ;$$

$$\sum y_i x_{3i} = \hat{\beta}_2 \sum x_{3i} x_{2i} + \hat{\beta}_3 \sum x_{3i}^2 + ... + \hat{\beta}_k \sum x_{3i} x_{ki} ;$$

.....

$$\sum y_i x_{ki} = \hat{\beta}_2 \sum x_{ki} x_{2i} + \hat{\beta}_3 \sum x_{ki} x_{3i} + ... + \hat{\beta}_k \sum x_{ki}^2 .$$

Відзначимо також, що регресійна модель з k змінними задовольняє таким рівнянням:

$$\sum \hat{u}_i = \sum \hat{u}_i X_{2i} = \sum \hat{u}_i X_{3i} = ... = \sum \hat{u}_i X_{ki} = 0 .$$

Дисперсія та стандартна похибка оцінювачів за МНК

Для отриманих значень частинних коефіцієнтів регресії можна вивести формули для дисперсії і стандартної похибки подібно до того, як це було зроблено для двовимірної моделі. Ці величини потрібні для двох основних завдань: встановлення довірчих інтервалів і перевірка статистичних гіпотез. Відповідні формули мають такий вигляд:

$$D(\hat{\beta}_1) = \left[\frac{1}{n} + \frac{\bar{X}_2^2 \sum x_{3i}^2 + \bar{X}_3^2 \sum x_{2i}^2 - 2\bar{X}_2 \bar{X}_3 \sum x_{2i} x_{3i}}{\sum x_{2i}^2 \sum x_{3i}^2 - \left(\sum x_{2i} x_{3i} \right)^2} \right] \sigma^2 ; \qquad (5.4.9)$$

$$\sigma_{\hat{\beta}_1} = \sqrt{D(\hat{\beta}_1)} ; \qquad (5.4.10)$$

$$D(\hat{\beta}_2) = \frac{\sum x_{3i}^2}{\sum x_{2i}^2 \sum x_{3i}^2 - \left(\sum x_{2i} x_{3i} \right)^2} \sigma^2 , \qquad (5.4.11)$$

або

$$D(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2 \left(1 - r_{23}^2 \right)} , \qquad (5.4.12)$$

$$\sigma_{\hat{\beta}_2} = \sqrt{D(\hat{\beta}_2)}; \tag{5.4.13}$$

$$D(\hat{\beta}_3) = \frac{\sum x_{2i}^2}{\sum x_{2i}^2 \sum x_{3i}^2 - \left(\sum x_{2i} x_{3i}\right)^2} \sigma^2, \tag{5.4.14}$$

або

$$D(\hat{\beta}_3) = \frac{\sigma^2}{\sum x_{3i}^2 \left(1 - r_{23}^2\right)}; \tag{5.4.15}$$

$$\sigma_{\hat{\beta}_3} = \sqrt{D(\hat{\beta}_3)}; \tag{5.4.16}$$

$$\operatorname{cov}(\hat{\beta}_2, \hat{\beta}_3) = \frac{-r_{23}\sigma^2}{\left(1 - r_{23}^2\right)\sqrt{\sum x_{2i}^2}\sqrt{\sum x_{3i}^2}}, \tag{5.4.17}$$

де $R^2{}_3$ – коефіцієнт кореляції між X_2 і X_3, що визначається за формулою

$$r_{23}^2 = \frac{\left(\sum x_{2i} x_{3i}\right)^2}{\sum x_{2i}^2 \sum x_{3i}^2}.$$

У всіх цих формулах σ^2 є гомоскедастична дисперсія збурень u_i. Можна показати, що незміщена оцінка для σ^2 визначається за формулою

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-3}. \tag{5.4.18}$$

Відзначимо аналогію між (5.4.18) і формулою для двовимірного випадку

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-2}.$$

Тепер кількість степенів вільності дорівнює (N–3), оскільки при оцінюванні $\sum \hat{u}_i^2$ ми повинні спочатку оцінити β_1, β_2 і β_3, які «поглинають» три степені вільності.

Для обчислення $\sum \hat{u}_i^2$ з (5.4.18) можна застосувати більш просту для обчислень формулу

$$\sum \hat{u}_i^2 = \sum y_i^2 - \hat{\beta}_2 \sum y_i x_{2i} - \hat{\beta}_3 \sum y_i x_{3i}, \tag{5.4.19}$$

що є тривимірним аналогом формули (2.3.6).

Дійсно, пригадаємо, що

$$\hat{u}_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i},$$

або у формі відхилень від середніх величин

$$\hat{u}_i = y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i}.$$

Тепер можемо виконати такі прості перетворення:

$$\sum \hat{u}_i^2 = \sum \hat{u}_i \hat{u}_i = \sum \hat{u}_i \left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i}\right) = \sum \hat{u}_i y_i .$$

При цьому нами застосована рівність

$$\sum \hat{u}_i x_{2i} = \sum \hat{u}_i x_{3i} = 0.$$

Звідси маємо

$$\sum \hat{u}_i^2 = \sum \hat{u}_i y_i = \sum y_i \left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i}\right) =$$
$$= \sum y_i^2 - \hat{\beta}_2 \sum y_i x_{2i} - \hat{\beta}_3 \sum y_i x_{3i},$$

що доводить справедливість (5.4.19)

Властивості оцінювачів за МНК

Властивості оцінювачів за МНК для множинної регресійної моделі аналогічні властивостям для двовимірного випадку. А саме:

1. Тривимірна поверхня (площина) регресії проходить через середні $\overline{Y}$, $\overline{X}_2$ і $\overline{X}_3$. Це очевидно з (5.4.3). Ця властивість допускає узагальнення на випадок лінійної регресійної моделі з *k* змінними (один регресант і (*k*–1) регресорів):

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + u_i . \qquad (5.4.20)$$

Для цієї моделі виконується

$$\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X}_2 + \hat{\beta}_3 \overline{X}_3 + ... + \hat{\beta}_k \overline{X}_k . \qquad (5.4.21)$$

2. Середня величина оцінки $Y_i (= \hat{Y}_i)$ дорівнює середній величині дійсних Y_i, що легко доводиться:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} = (\overline{Y} - \hat{\beta}_2 \overline{X}_2 - \hat{\beta}_3 \overline{X}_3) + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} =$$
$$= \overline{Y} + \hat{\beta}_2 \left(X_{2i} - \overline{X}_2\right) + \hat{\beta}_3 \left(X_{3i} - \overline{X}_3\right) = \overline{Y} + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i}. \qquad (5.4.22)$$

Підсумовуючи обидві частини (5.4.22) за об’ємом усієї вибірки і поділяючи на *N*, одержуємо $\overline{\hat{Y}} = \overline{Y}$. При цьому використовується відома властивість $\sum x_{2i} = \sum x_{3i} = 0$. Зауважимо також, що з (5.4.22) випливає рівність

$$\hat{y}_i = \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i}, \qquad (5.4.23)$$

де $\hat{y}_i = \left(\hat{Y}_i - \bar{Y}\right)$.

Отже, SRF (5.4.1) може бути поданий у формі відхилень

$$y_i = \hat{y}_i + \hat{u}_i = \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \hat{u}_i . \qquad (5.4.24)$$

3. $\sum \hat{u}_i = \bar{\hat{u}} = 0$. Ця рівність є наслідком МНК.
4. Залишки $\hat{u}_i$ некорельовані з X_{2i} і X_{3i}, тобто $\sum \hat{u}_i X_{2i} = \sum \hat{u}_i X_{3i} = 0$.
5. Залишки $\hat{u}_i$ некорельовані з $\hat{Y}_i$, тобто $\sum \hat{u}_i \hat{Y}_i = 0$. Цю властивість легко довести, якщо обидві частини рівності $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i}$ помножити на $\hat{u}_i$ і підсумувати за об'ємом вибірки з урахуванням властивості 4.
6. Із (5.4.12) і (5.4.15) бачимо, що зі зростанням до 1 кореляційного коефіцієнта R^2_3 між X_2 і X_3 дисперсії $\hat{\beta}_2$ і $\hat{\beta}_3$ при фіксованих значеннях σ^2 і $\sum x_{2i}^2$, $\sum x_{3i}^2$ також зростають. Коли R^2_3=1 (точна колінеарність), ці дисперсії стають необмеженими. Тобто зі зростанням R^2_3 значення β_2 і β_3 стають усе більш невизначеними.
7. Із (5.4.12) і (5.4.15) зрозуміло, що для даних величин R^2_3, $\sum x_{2i}^2$, $\sum x_{3i}^2$ дисперсії коефіцієнтів $\hat{\beta}_2$ і $\hat{\beta}_3$ прямо пропорційні σ^2, тобто зі зростанням σ^2 вони зростають. Водночас для даних величин σ_2 і R^2_3 дисперсія $\hat{\beta}_2$ обернено пропорційна $\sum x_{2i}^2$, тобто чим більше змінюється за вибіркою X_2, тим менша дисперсія $\hat{\beta}_2$ і, отже, тим точніше можна оцінити β_2. Аналогічний висновок можна зробити про дисперсію $\hat{\beta}_3$.
8. При прийнятих гіпотезах класичної лінійної регресійної моделі можна показати, що оцінки за МНК частинних коефіцієнтів регресії не тільки лінійні й незміщені, але й мають якнайменшу дисперсію в класі лінійних незміщених оцінок, тобто вони мають властивість BLUE.

5.5. Коефіцієнт детермінації R^2 і коефіцієнт кореляції множинної регресійної моделі

У випадку моделі з двома змінними ми бачимо, що коефіцієнт

$$r^2 = 1 - \frac{\sum \hat{u}_i^2}{\sum \left(Y_i - \bar{Y}\right)^2} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

виміряв якість підгонки рівняння регресії, точніше, він дав співвідношення або відсоток у загальній варіації Y, пояснений за рахунок пояснювальної змінної X. Даний підхід може бути узагальненим на випадок моделей, які містять більше двох змінних. Так, у моделі з трьома змінними нас цікавить, яка частина у варіації $У$ пояснюється за рахунок змінних X_2 і X_3. У цьому випадку коефіцієнт познача-

ється R^2 і називається коефіцієнтом детермінації множинної регресії. Концептуально він наближений до R^2.

Для виведення R^2 можна скористатися процедурою виведення R^2, описаною в підрозд. 2.5. Пригадаємо, що

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{u}_i = \hat{Y}_i + \hat{u}_i . \qquad (5.5.1)$$

Переходячи до малих букв, цю рівність можна записати у вигляді

$$y_i = \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \hat{u}_i = \hat{y}_i + \hat{u}_i . \qquad (5.5.2)$$

Підносячи (5.5.2) до квадрата й підсумовуючи, одержуємо

$$\sum y_i^2 = \sum \hat{y}_i^2 + \sum \hat{u}_i^2 + 2\sum \hat{u}_i \hat{y}_i = \sum \hat{y}_i^2 + \sum \hat{u}_i^2 . \qquad (5.5.3)$$

Ця рівність показує, що загальна сума квадратів (TSS) дорівнює поясненій сумі квадратів (ESS) + сума квадратів залишків (RSS). Підставляючи замість $\sum \hat{u}_i^2$ вираз (5.4.19), одержуємо

$$\sum y_i^2 = \sum \hat{y}_i^2 + \sum y_i^2 - \hat{\beta}_2 \sum y_i x_{2i} - \hat{\beta}_3 \sum y_i x_{3i} .$$

Звідси одержуємо вираз для ESS пояснюючої суми квадратів:

$$ESS = \sum \hat{y}_i^2 = \hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i} . \qquad (5.5.4)$$

За визначенням маємо

$$R^2 = \frac{ESS}{TSS} = \frac{\sum \hat{y}_i^2}{\sum y_i^2} = \frac{\hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i}}{\sum y_i^2} . \qquad (5.5.5)$$

Зазначимо, що можна отримати й інший вираз для R^2, якщо застосувати (5.5.3) і розділити обидві частини на $\sum y_i^2$. Одержимо

$$R^2 = 1 - \frac{\sum \hat{u}_i^2}{\sum y_i^2} .$$

R^2, як і R^2, лежить між 0 і 1. Якщо R^2=1, то лінія регресії на 100% пояснює варіацію в Y. Якщо ж R^2=0, модель нічого не пояснює у варіації Y. Проте R^2 зазвичай лежить між цими граничними величинами. Вважається, що підгонка моделі тим краща, чим більше R^2 наближається до 1.

Пригадаємо, що в моделі з двома змінними величина r позначала степінь лінійного зв'язку між двома змінними. У моделі з трьома й більше змінними аналогом r є коефіцієнт множинної кореляції, що позначається R. Він позначає степінь асоціативності між Y і всіма пояснювальними змінними одночасно. На відміну від

r, який може бути і негативним, R набуває завжди позитивних значень. На практиці, проте, R відіграє незначну роль, більш важлива величина R^2.

Зазначимо зв'язок між R^2 і дисперсіями частинних коефіцієнтів регресії в моделі множинної регресії з k змінними (5.4.20):

$$D(\hat{\beta}_j) = \frac{\sigma^2}{\sum x_j^2}\left(\frac{1}{1-R_j^2}\right), \qquad (5.5.6)$$

де $\hat{\beta}_j$ – частинний коефіцієнт регресії при регресорі X_j, а R_j^2 – R^2 в регресії X_j по тих (k–2) змінних, що залишилися регресорами (у регресійній моделі з k змінними є (k–1) регресор). Ця рівність є узагальнення формул (5.4.12), (5.4.15) для моделі з трьома змінними (один регресант і два регресори).

Приклад. Крива Філіпса для США, 1970–1982 рр.

З метою демонстрації викладеного вище підходу розглянемо таку модель:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t, \qquad (5.5.7)$$

де Y_t – дійсний рівень інфляції на час t, %; X_{2t} – рівень безробіття на час t, %; X_{3t} – очікуваний або прогнозований рівень безробіття на час t, %. Ця модель відома як крива очікування-зростання Філіпса.

Відповідно до теорії макроекономіки β_2 – негативна величина, а β_3 – позитивна. Теорія переконує нас, що $\beta_3 = 1$.

З метою перевірки теорії розглянемо наведені в табл.5.1 дані.

Таблиця 5.1

Дійсний рівень інфляції Y, рівень безробіття X_2, очікуваний рівень інфляції X_3 в США, 1970–1982 рр.

Рік	Y*, %	X_2, %	X_3, %
1970	5,92	4,9	4,78
1971	4,30	5,9	3,84
1972	3,30	5,6	3,13
1973	6,23	4,9	3,44
1974	10,97	5,6	6,84
1975	9,14	8,5	9,47
1976	5,77	7,7	6,51
1977	6,45	7,1	5,92
1978	7,60	6,1	6,08
1979	11,47	5,8	8,09
1980	13,46	7,1	10,01
1981	10,24	7,6	10,81
1982	5,99	9,7	8,00

На підставі даних (табл.5.1) одержуємо за МНК такі результати:

$$\hat{Y}_t = 7{,}1933 - 1{,}3925 X_{2t} + 1{,}4700 X_{3t}$$
$$(1{,}5948) \quad (0{,}3050) \quad (0{,}1758)$$
$$R^2 = 0{,}8766. \qquad (5.5.8)$$

У дужках наведені стандартні похибки коефіцієнтів регресії.

Інтерпретація результатів. Якщо для заданого періоду фіксувати X_2 і X_3 на нульовому рівні, то середній рівень інфляції складе 7,19%. Проте, як уже не раз наголошувалося, така інтерпретація занадто технічна. Часто вона не має економічного сенсу. Значення $\hat{\beta}_2 = -1{,}3925$ позначає, що при фіксованому рівні очікуваної інфляції X_3 рівень дійсної інфляції зменшиться в середньому на 1,4% при збільшенні рівня безробіття на 1%. Аналогічно, при збереженні сталого рівня безробіття збільшення на 1% очікуваного рівня інфляції приводитиме в середньому до збільшення на 1,47% дійсного рівня інфляції. Величина R^2=0.88 вказує на те, що модель (5.5.8) пояснює приблизно 88% варіації дійсного рівня інфляції за рахунок включених в модель змінних, що свідчить про високу надійність цієї моделі. Відповідно до наших умов коефіцієнти регресії мають правильні знаки.

5.6. Проста регресія в контексті множинної регресії

Припущення (5.1.6) класичної лінійної регресії стверджує, що модель регресії коректно специфікована, тобто помилка зсуву внаслідок неправильної специфікації відсутня. Наведені в попередніх розділах відомості дозволяють пояснити це твердження.

Припустимо, що (5.5.7) є "істинна" модель, що пояснює рівень дійсної інфляції за допомогою рівня безробіття й рівня очікуваної інфляції. Водночас припустимо, що хтось застосовує таку модель регресії:

$$Y_t = b_1 + b_{12} X_{2t} + \hat{u}_{1t}, \qquad (5.6.1)$$

де Y_t – дійсний на момент часу t рівень інфляції; X_{2t} – рівень безробіття на той же момент часу; $\hat{u}_{1t}$ – залишковий складова. Кутовий коефіцієнт b_{12} визначає зміну середнього рівня інфляції, викликану одиничною зміною рівня безробіття.

Оскільки «істинною моделлю» є (5.6.1), та рівність (5.6.1) має помилку специфікації, яка полягає у відсутності в моделі змінної X_3 очікуваного рівня інфляції.

Ми знаємо, що $\hat{\beta}_2$ у множинній регресії (5.5.7) є незміщеною оцінкою β_2, тобто $E(\hat{\beta}_2) = \beta_2$. Чи може коефіцієнт b_{12} простої регресії Y по X_2 також бути незміщеною оцінкою β_2? Тобто чи виконуватиметься $E(b_{12}) = \beta_2$? У термінах нашого прикладу це питання можна сформулювати таким чином. Чи буде коефіцієнт рівня безробіття в (5.6.1) давати незміщену оцінку істинного впливу на рівень інфляції, якщо ми знаємо, що в рівняння не включена змінна X_3 (очікуваний рівень інфляції)? У загальному випадку відповідь звучить так: b_{12} не буде незміщеною оцінкою β_2. Крім того, var(b_{12}) може бути зміщеною оцінкою var($\hat{\beta}_2$). Можна показати, що насправді виконується рівність

$$b_{12} = \beta_2 + \beta_3 b_{32} + похибка\ , \qquad (5.6.2)$$

де b_{32} – кутовий коефіцієнт у регресії X_3 за X_2, тобто

$$X_{3t} = b_2 + b_{32} X_{2t} + \hat{u}_{2t}\,. \qquad (5.6.3)$$

Із (5.6.2) можна отримати рівність

$$E(b_{12}) = \beta_2 + \beta_3 b_{32}\,. \qquad (5.6.4)$$

За даними вибірки b_{23} обчислюється за формулою

$$b_{32} = \frac{\sum x_{3i} x_{2i}}{\sum x_{2i}^2}\,.$$

Як бачимо з рівняння (5.6.3), у випадку $\beta_3 b_{32} \neq 0$ коефіцієнт b_{12} є зміщеною оцінкою β_3. Якщо $\beta_3 b_{32} > 0$, то зсув буде у бік завищення, а у випадку $\beta_3 b_{32} < 0$ – у бік заниження.

Із цього випливає, що згідно з (5.6.2), коефіцієнт простої регресії b_{12} враховує не тільки «прямий», або «нетто», вплив X_2 на Y (при фіксованому значенні впливу X_3), але й непрямий вплив на Y через невключену змінну X_3. Коротше кажучи, b_{13} визначає «загальний ефект» (прямий + непрямий) X_2 на Y, тоді як $\hat{\beta}_2$ позначає тільки прямий ефект впливу X_2 на Y, оскільки вплив X_3 зберігається фіксованим.

Зі сказаного вище робимо висновок: загальний ефект X_2 на Y (=b_{12}) складається з прямого (=β_2) та непрямого (=$\beta_3 b_{32}$) ефекту X_2 на Y.

У визначеннях нашого прикладу це звучить таким чином. Загальний вплив одиничної зміни рівня безробіття на дійсний рівень інфляції складається з прямого впливу (при фіксованому рівні очікуваної інфляції) і непрямого ефекту через дію безробіття на очікуваний рівень інфляції. Цей результат може бути пояснений за допомогою діаграми (рис.5.2).

Рис. 5.2. Прямий і непрямий ефекти X_2 на Y

Проілюструємо теоретичне міркування на прикладі кривої Філіпса.

Використовуючи дані табл.5.1, одержуємо для моделі (5.6.1) такі результати:

$$\hat{Y}_t = 6{,}1272 + 0{,}2448 X_{2t}$$
$$\textbf{(4,2853)} \quad \textbf{(0,6304)} \qquad (5.6.5)$$
$$t = (1{,}4298) \quad (0{,}3885) \qquad R^2 = 0{,}135.$$

Несподіваним в (5.6.5) є те, що b_{12}=0.244*8* не тільки має позитивний знак (позитивний кутовий коефіцієнт кривої Філіпса), але також мало відрізняється від нуля. Водночас у (5.6.2) ми бачимо, що $\hat{\beta}_2 = -1{,}3925$ не тільки має очікуваний негативний знак, але й статистично значно відрізняється від нуля, що пояснюється непрямим ефектом доданку $\beta_3 b_{32}$ з (5.6.4). Із (5.5.8) ми знаємо, що $\hat{\beta}_3 = 1{,}4700$. Для знаходження b_{32} виконаємо регресію (5.6.2) за наявними даними:

$$X_{3t} = 0{,}7252 + 1{,}1138 X_{2t}$$
$$(2{,}7267) \quad (0{,}4011) \qquad (5.6.6)$$
$$t = (-0{,}2659) \quad (2{,}7769), \qquad r = 0{,}4120.$$

Значення b_{32}=1,1138 позначає, що збільшення X_2 на одиницю приводить до зростання (в середньому) X_3 приблизно на 1,11 одиниць. Але якщо на цю величину зросте X_3, то її дія на Y буде $\hat{\beta}_3 b_{32} = 1{,}4700 \times 1{,}1138 = 1{,}6373$. Отже, з (5.6.2) ми маємо остаточно

$$\hat{\beta}_2 + \hat{\beta}_3 b_{32} = -1,3925 + 1.6373 = 0,2448 = b_{12}.$$

Із наведених міркувань можна зробити такий висновок. Якщо з якоїсь причини ухвалити рішення про застосування моделі тривимірної регресії, не слід звертатися до найпростішої двовимірної регресії. У більш загальному випадку це звучить так. Якщо ви віддали перевагу конкретній моделі регресії і вважаєте її «істинною», то не слід модифікувати її шляхом виключення з моделі якої-небудь змінної. Якщо ви знехтуєте цією рекомендацією, то отримаєте зміщені оцінки параметрів. Крім того, ви можете отримати хибне значення $\hat{\sigma}^2$ і некоректні довірчі інтервали для параметрів регресії. Звернемо увагу, що стандартна похибка коефіцієнта $\hat{\beta}_2$ в моделі (5.5.8) набагато менша (у порівнянні з величиною $\hat{\beta}_2$), ніж у моделі (5.6.5). Тому довірчі інтервали й перевірка гіпотез на підставі (5.5.8) мають більший ступінь довіри, ніж за моделлю (5.6.5).

5.7. R^2 і скорегований R^2

Коефіцієнт детермінації R^2 є неспадна функція від кількості пояснювальних змінних або регресорів у моделі. При збільшенні кількості регресорів R^2 майже неминуче зростає і ніколи не спадає. Інакше кажучи, додавання змінної X до моделі не зменшить R^2. Щоб переконатися в цьому, пригадаємо визначення коефіцієнта детермінації:

$$R^2 = 1 - \frac{\sum \hat{u}_i^2}{\sum y_i^2}. \qquad (5.7.1)$$

Величина $\sum y_i^2$ не залежить від кількості змінних X у моделі, оскільки це просто $\sum (Y_i - \bar{Y})^2$. Величина RSS $\sum \hat{u}_i^2$, проте, залежить від кількості присутніх у моделі регресорів. Зрозуміло, що при зростанні кількості змінних X величина $\sum \hat{u}_i^2$ повинна спадати (принаймні не зростати). Отже, R^2 зростатиме. Враховуючи це, при порівнянні двох моделей регресії з однаковою залежною змінною, але різною кількістю змінних X, потрібно бути дуже обережними з наданням переваги моделі з більш високим R^2.

Порівнюючи два коефіцієнти детермінації R^2, потрібно обов'язково враховувати кількість регресорів X, присутніх у моделі. Це можна зробити, якщо скористатися визначенням альтернативного коефіцієнта детермінації, обчислюваного за такою формулою:

$$\bar{R}^2 = 1 - \frac{\sum \hat{u}_i^2 / (n-k)}{\sum y_i^2 / (n-1)}, \qquad (5.7.2)$$

де k – кількість параметрів у моделі (для моделі з трьома змінними k=3). Визначений таким чином коефіцієнт детермінації називається скоректованим R^2 і позна-

чається $\overline{R}^2$. Термін «скорегований» позначає скоректованість за кількістю степенів вільності, пов'язаних із сумами квадратів, що входять в (5.7.2). $\sum \hat{u}_i^2$ має (N–k) степенів вільності. У разі моделі з трьома змінними ми знаємо, що $\sum \hat{u}_i^2$ має (N–3) степенів вільності.

Рівняння (5.7.2) можна також переписати у вигляді

$$\overline{R}^2 = 1 - \frac{\hat{\sigma}^2}{S_y^2}, \tag{5.7.3}$$

де $\hat{\sigma}^2$ – дисперсія залишків, а S_y^2 – вибіркова дисперсія Y: $S_y^2 = \dfrac{\sum (Y_i - \overline{Y})^2}{n-1}$.

Легко показати, що R^2 і $\overline{R}^2$ пов'язані між собою співвідношенням

$$\overline{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k}. \tag{5.7.4}$$

Цю рівність можна отримати, якщо підставити (5.7.1) у (5.7.2). Із (5.7.4) бачимо, що для k>1 $\overline{R}^2 < R^2$, а це означає, що при збільшенні змінних X $\overline{R}^2$ зростає меншою мірою, ніж R^2. Крім того, $\overline{R}^2$ може набувати й негативних значень, тоді як R^2 завжди позитивний. У прикладних задачах у випадках, коли $\overline{R}^2$ виявляється негативним, його вважають таким, що дорівнює нулю. Для розглянутого нами прикладу кривої Філіпса маємо $R^2 = 0{,}8766$, а $\overline{R}^2 = 0{,}8519$.

Порівняння величин R^2

У першу чергу дуже важливо відзначити, що при порівнянні моделей на основі значень коефіцієнтів детермінації, як нескоректованих, так і скоректованих, повинні бути однаковими об'єми вибірки N і одними й тими ж залежні змінні. Пояснювальні змінні можуть бути будь-якого вигляду. Так, для моделей

$$\ln Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i; \tag{5.7.5}$$

$$Y_i = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + u_i \tag{5.7.6}$$

не можна порівнювати підраховані R^2. Причина полягає в тому, що за визначенням R^2 визначає частину дисперсії залежної змінної, поясненої за рахунок пояснювальних змінних, отже, в (5.7.5) R^2 виміряє частину в дисперсії $\ln Y_i$, пояснену за рахунок X_2 і X_3, а в (5.7.6) це частина в дисперсії Y_i, пояснена тими ж змінними. Зрозуміло, що це різні речі. Раніше ми відзначали, що змінювання $\ln Y_i$ дає відносне змінювання Y, тоді як змінювання Y є абсолютним змінюванням. Отже,

$\operatorname{var}\hat{Y}_i / \operatorname{var} Y_i$ не одне й те ж, що $\operatorname{var}(\ln\hat{Y}_i)/\operatorname{var}(\ln Y_i)$. Тому коефіцієнти детермінації (5.7.5) і (5.7.6) не можна порівнювати.

Якщо ми звернемося до функції попиту на каву (2.7.1):

$$\hat{Y}_t = 2,6911 - 0,4795X_t;$$

$$D(\hat{\beta}_1) = 0,0148,\ \sigma(\hat{\beta}_1) = 0,1216;$$

$$D(\hat{\beta}_2) = 0,0129,\ \sigma(\hat{\beta}_2) = 0,0114,\ \hat{\sigma}^2 = 0,01656;$$

$$r^2 = 0,6628,\ r = -0,8141,$$

що являє собою лінійну модель, і функції попиту (6.4.5):

$$\ln Y_t = 0,7774 - 0,2530\ln X_t \qquad R^2=0,7448$$
$$(0,0152)\quad (0,0494) \qquad F_{1,9}=26,27$$
$$t = (51,0045)\quad (-5,1251),$$

що представляє Log-Lin – модель, то порівнювати їх коефіцієнти детермінації безпосередньо не можна. Як же все-таки порівнювати величини R^2 для моделей вигляду (3.7.1) і (6.4.5)? Покажемо це на прикладі функції попиту на каву.

Для порівняння коефіцієнтів детермінації, отриманих із моделей із різним видом залежної змінної, як, наприклад, у моделях (3.7.1) і (6.4.5), можна застосувати два способи.

1. За відомим значенням $\ln\hat{Y}_t$ у моделі (6.4.5) знаходимо $\hat{Y}_t$, а потім підраховуємо R^2 між $\hat{Y}_t$ і Y_t за формулою (2.5.14):

$$R^2 = \frac{\left(\sum y_i \hat{y}_i\right)^2}{\sum y_i^2 \sum \hat{y}_i^2}.$$

Підрахований таким чином R^2 можна порівнювати з коефіцієнтом детермінації з моделі (6.4.5).

2. Підраховуємо за відомими $\hat{Y}_t$ з моделі (3.7.1) $\ln\hat{Y}_t$ і $\ln Y_t$, обчислюємо R^2 між ними. Цей коефіцієнт детермінації можна порівнювати з R^2 із моделі (6.4.5).

Припустимо, що ми спочатку вирішили порівнювати величину R^2 лінійної моделі (3.7.1) з величиною R^2 подвійної логарифмічної моделі (6.4.5). Використаємо значення $\hat{Y}_i$ моделі (3.7.1) і знайдемо за ними $\ln\hat{Y}_t$, а потім за фактичними значеннями Y_i знайдемо $\ln Y_t$. За отриманими значеннями $\ln Y_t$ і $\ln\hat{Y}_t$ ми можемо підрахувати R^2, наприклад, за формулою

$$R^2 = \frac{\left(\sum y_i \hat{y}_i\right)^2}{\sum y_i^2 \sum \hat{y}_i^2}.$$

Використовуючи дані, наведені в (5) і (6) стовпцях табл. 2.1, можна підрахувати за цією формулою величину R^2. Отримуємо $R^2 = 0,6779$. Цю величину R^2 вже можна порівняти зі значенням $R^2 = 0,7448$ з подвійної логарифмічної моделі. Порівняння цих величин виявляється на користь логарифмічної моделі.

Якщо ж ми хочемо порівняти величину R^2 із подвійної логарифмічної моделі з R^2 із лінійної моделі, то необхідно за значеннями $\ln \hat{Y}_t$ з моделі (6.4.5) обчислити $\hat{Y}_i = \exp(\ln \hat{Y}_i)$, а потім за цими значеннями і Y_i обчислити R^2. Використовуючи дані (4) і (1) стовпців табл. 2.1, знаходимо $R^2 = 0.7187$. Це значення R^2 вже можна порівнювати зі значенням коефіцієнта детермінації з лінійної моделі $R^2 = 0.6628$. Як і раніше, подвійна логарифмічна модель має високий коефіцієнт детермінації.

На закінчення зробимо таке зауваження. Іноді дослідник прагне збільшити $\overline{R}^2$, тобто вибрати ту модель, яка має найвищий $\overline{R}^2$. Це не правильно, оскільки в регресійному аналізі нашою метою є не отримання за будь-яку ціну високого $\overline{R}^2$, а отримання надійних оцінок істинних коефіцієнтів популяцій регресії, що дають можливість зробити статистичні висновки. В емпіричному аналізі часто виникає ситуація, коли $\overline{R}^2$ має високі значення, а при цьому деякі коефіцієнти регресії виявляються статистично незначущими або ж мають знак, протилежний очікуваному. Отже, дослідник повинен більше зосереджуватися на логічному або теоретичному зв'язку між пояснюваним і пояснювальними змінними та їх статистичній значущості. Якщо при цьому ми отримаємо високе значення $\overline{R}^2$, то тим краще. Однак, якщо $\overline{R}^2$ малий, то це не означає, що наша модель обов'язково погана.

5.8. Частинні коефіцієнти кореляції

Раніше ми розглядали коефіцієнт кореляції *R*, що дає міру степеня лінійної асоційованості між двома змінними. Для моделі регресії з трьома змінними можемо підрахувати вже три коефіцієнти кореляції: R_{12} (кореляція між Y і X_2), R_{13} (між Y і X_3) і R^2_3 (між X_2 і X_3). Відзначимо, що індекс 1 ми відносимо до Y, а 2 і 3 до X_2 і X_3 відповідно. Ці кореляційні коефіцієнти називаються простими кореляційними коефіцієнтами, або коефіцієнтами нульового порядку. Їх можна підраховувати за відомою формулою

$$r_{xy} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2} \sqrt{\sum (Y_i - \overline{Y})^2}} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}.$$

Розглянемо тепер, чи дійсно, скажімо, R_{12} дає "істинний" степінь лінійної асоційованості між Y і X_2, коли третя змінна X_3 асоційована з ними. Припустимо, що істинна модель регресії задається рівністю

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i .$$

Ми не включатимемо в модель змінну X_{3i} і побудуємо регресію Y тільки за X_2 з кутовим коефіцієнтом b_{12}. Чи буде цей коефіцієнт дорівнювати "істинному" коефіцієнту β_2. У загальному випадку R_{12} не відображає істинний степінь асоційованості між Y і X_2 у присутності X_3. Як ми пізніше покажемо, він дає спотворене уявлення про природу асоціативності між Y і X_2. Отже, нам потрібен такий коефіцієнт кореляції, який не залежав би від впливу X_3 ні на X_2, ні на Y. Такий кореляційний коефіцієнт можна отримати, і він називається частинним коефіцієнтом кореляції. Концептуально він аналогічний частинному коефіцієнту регресії. Частинні коефіцієнти кореляції вводяться таким чином:

- $r_{12.3}$ – частинний коефіцієнт кореляції між Y і X_2 при постійному значенні X_3;
- $r_{13.2}$ – частинний коефіцієнт кореляції між Y і X_3 при постійному значенні X_2;
- $r_{23.1}$ – частинний коефіцієнт кореляції між X_2 і X_3 при постійному значенні Y.

Один із шляхів визначення частинних коефіцієнтів кореляції полягає в такому. Регресуємо Y за X_3 таким чином:

$$Y_i = b_1 + b_{13} X_{3i} + \hat{u}_{1i} .$$

Проводимо регресію X_2 за X_3:

$$X_{2i} = b_2 + b_{23} X_{3i} + \hat{u}_{2i} .$$

Звернемо увагу на те, що залишки $\hat{u}_{1i}$ є величиною Y_i після «звільнення» її від впливу змінної X_3. Аналогічно $\hat{u}_{2i}$ – величина X_{2i} після «звільнення» її від впливу змінної X_3. Тепер ми можемо регресувати $\hat{u}_{1i}$ за $\hat{u}_{2i}$ і тим самим знайти простий коефіцієнт кореляції $r_{12.3}$, оскільки значення X_3 залишається постійним. Одержуємо

$$r_{12.3} = r_{\hat{u}_1 \hat{u}_2} \frac{\sum (\hat{u}_{1i} - \bar{\hat{u}}_1)(\hat{u}_{2i} - \bar{\hat{u}}_2)}{\sqrt{\sum (\hat{u}_{1i} - \bar{\hat{u}}_1)^2} \sqrt{\sum (\hat{u}_{2i} - \bar{\hat{u}}_2)^2}} = \frac{\sum \hat{u}_{1i} \hat{u}_{2i}}{\sqrt{\sum \hat{u}_{1i}^2 \sum \hat{u}_{2i}^2}} . \qquad (5.8.1)$$

При цьому ми врахували, що $\bar{\hat{u}}_1 = \bar{\hat{u}}_2 = 0$.

Із вищезазначеного зрозуміло, що частинний коефіцієнт кореляції між Y і X_2 при постійному $X3$ є простий (нульового порядку) коефіцієнт кореляції між залишками від регресії Y за X_3 і X_2 за X_3 відповідно. Аналогічним чином інтерпретуються коефіцієнти кореляції $r_{13.2}$ і $r_{23.1}$.

На практиці, проте, немає необхідності проводити цю процедуру, що складається з трьох етапів. Для обчислення частинних коефіцієнтів кореляції можна застосувати формули, що виражають їх через прості коефіцієнти кореляції:

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}; \qquad (5.8.2)$$

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{(1-r_{12}^2)(1-r_{23}^2)}}; \qquad (5.8.3)$$

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{(1-r_{13}^2)(1-r_{12}^2)}}. \qquad (5.8.4)$$

Визначені таким чином величини називаються коефіцієнтами кореляції першого порядку. Під порядком ми маємо на увазі кількість індексів, що стоять після крапки. Так $r_{12.34}$ – другого порядку, а $r_{12.345}$ – третього. Як наголошувалося раніше, r_{12} і r_{13} – прості або нульового порядку. Інтерпретація, скажімо, $r_{12.34}$ така, що існує коефіцієнт кореляції між Y і X_2 при постійних значеннях X_3 і X_4.

Інтерпретація простого і частинного коефіцієнтів кореляції

У разі моделі з двома змінними R мав чітке визначення, він визначав степінь лінійної зв'язаності між залежною змінною Y і єдиною пояснювальною змінною X. Якщо ж ми виходимо за рамки моделі з двома змінними, нам потрібно приділити увагу простому коефіцієнту кореляції. З рівності (5.8.2), наприклад, бачимо, що:

1. Якщо навіть $r_{12} = 0$, то $r_{12.3}$ не буде нульовим за умови, що не дорівнюють нулю r_{13} і r_{23} (або обидва вони дорівнюють нулю).
2. Якщо $r_{12} = 0$, а r_{13} і r_{23} ненульові і мають однакові знаки, то $r_{12.3}$ буде негативним, якщо ж вони мають протилежні знаки, він буде позитивним. Приклад допоможе краще зрозуміти це. Нехай Y позначає урожай сільгоспкультури, X_2 – опади, а X_3 – температуру. Припустимо, що $r_{12} = 0$, тобто немає зв'язку між урожаєм і опадами. Припустимо далі, що $r_{13} > 0$ і $r_{23} < 0$. Тоді, як бачимо з (5.8.2), $r_{12.3} > 0$. Тобто при незмінній температурі існує позитивний зв'язок між урожаєм і опадами. Цей, здавалося б, парадоксальний результат, відповідає дійсності, оскільки температура X_3 впливає і на урожай Y і на опади X_2. Для знаходження істинного співвідношення між урожаєм і опадами необхідно виключити вплив температури. На цьому прикладі показано, як можна помилятися, використовуючи простий коефіцієнт кореляції.
3. Значення $r_{12.3}$ і r_{12} не обов'язково повинні мати однакові знаки. Це зауваження стосується й інших коефіцієнтів кореляції.
4. Для двовимірної регресії, як нам відомо, величина R^2 лежить між 0 і 1. Ця властивість залишається справедливою і для квадратів частинних коефіцієнтів кореляції. Спираючись на цей факт, можна перевірити справедливість властивості

$$0 \le r_{12}^2 + r_{13}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23} \le 1. \qquad (5.8.5)$$

5. Припустимо, що $r_{13} = r_{23} = 0$. Чи означає це, що й $r_{12} = 0$? Відповідь очевидна з (5.8.5). Некорельованість Y і X_3, а також X_2 і X_3, не спричиняє корельованості Y і X_2.

Побіжно відзначимо, що вираз $r_{12.3}^2$ можна назвати коефіцієнтом частинної детермінації й інтерпретувати його як частину варіації Y, пояснену не за рахунок змінної X_3, а через включену в модель змінну X_2. Таке трактування $r_{12.3}^2$ стає зрозумілим, якщо застосувати формулу

$$r_{12.3}^2 = \frac{R^2 - r_{13}^2}{1 - r_{13}^2}.$$

Відзначимо також співвідношення між R^2, простими коефіцієнтами кореляції та частинними кореляційними коефіцієнтами:

$$R^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}; \qquad (5.8.6)$$

$$R^2 = r_{12}^2 + (1 - r_{12}^2)r_{13.2}^2; \qquad (5.8.7)$$

$$R^2 = r_{13}^2 + (1 - r_{13}^2)r_{12.3}^2. \qquad (5.8.8)$$

Відзначимо на закінчення таке. Раніше мова йшла про те, що R^2 не спадає при включенні в модель додаткових пояснювальних змінних. Це зрозуміло з рівності (5.8.7), яка також показує, що частина варіації Y, поясненої спільно змінними X_2 і X_3, має 2 складники: частина, пояснена тільки X_2 (це r_{12}^2), і частина, непояснена X_2 (це $1 - r_{12}^2$), помножена на пояснену змінною X_3 частину при постійному значенні X_2. Оскільки $r_{13.2}^2 > 0$, то $R^2 > r_{12}^2$. У крайньому випадку при $r_{13.2}^2 = 0$ маємо $R^2 = r_{12}^2$.

5.9. Виробнича функція Коба – Дугласа

У попередньому розділі ми показали, як шляхом відповідного перетворення можна перейти від нелінійного за змінними рівняння до лінійного з подальшою реалізацією схеми регресійного аналізу. Цей прийом може бути застосований і в моделях з кількістю змінних, більшою за двох. Покажемо це на прикладі відомої виробничої функції Коба – Дугласа.

Функція Коба – Дугласа може бути подана в статистичній формі таким чином:

$$Y_i = \beta_1 X_{2i}^{\beta_2} X_{3i}^{\beta_3} e^{u_i}, \qquad (5.9.1)$$

де Y_i – вироблена продукція; X_{2i} – трудовитрати; X_{3i} – капіталозатрати, u_i – стохастичний збурюючий складова.

Рівняння (5.9.1) є нелінійним стосовно змінних, що входять до нього. Після логарифмування (5.9.1) ми одержимо

$$\ln Y_i = \beta_0 + \beta_2 \ln X_{2i} + \beta_3 \ln X_{3i} + u_i, \qquad (5.9.2)$$

де $\beta_0 = \ln \beta_1$.

Така форма рівняння вже є лінійною щодо параметрів β_0, β_2, β_3. Зазначимо також, що це рівняння залишається нелінійним стосовно змінних Y, X_2, X_3, але лінійним за змінними $\ln Y$, $\ln X_2$, $\ln X_3$. Ця модель називається «логарифмічною лінійною моделлю». Її двовимірний аналог був нами досліджений на прикладі (6.4.3).

Функція Коба – Дугласа має такі особливості:

1. Коефіцієнт β_2 є частинним коефіцієнтом еластичності виробленої продукції щодо трудовитрат, тобто він є мірою процентної зміни вироблюваної продукції при 1%-му збільшенні, скажімо, трудовитрат і збереженні при цьому на постійному рівні капіталозатрат.
2. Коефіцієнт β_3 також є частинним коефіцієнтом еластичності виробленої продукції щодо капіталозатрат при постійному рівні трудовитрат.
3. Сума $\beta_2 + \beta_3$ дає інформацію про обсяг збільшення виробництва, тобто являє собою відповідь виробництва на пропорційну зміну витрат. Якщо ця сума дорівнює 1, то ми маємо одиничний обсяг збільшення виробництва, тобто подвоєння витрат викликає подвоєння продукції, потроєння витрат – потрійне збільшення виробництва й т.д. Якщо сума менше 1, ми маємо спадний обсяг виробництва, тобто подвоєння витрат дає менший ніж у два рази приріст виробництва. Нарешті, якщо сума більше 1, ми маємо випадок зростаючого обсягу – подвоєння витрат приводить до збільшення виробництва більше ніж у два рази.

Побіжно відзначимо, що якщо використовується логарифмічна лінійна модель з довільною кількістю змінних, то коефіцієнт при кожній змінній X є мірою частинної еластичності залежної змінної Y стосовно X. Для моделі з k змінними

$$\ln Y_i = \beta_0 + \beta_2 \ln X_{2i} + \beta_3 \ln X_{3i} + \ldots + \beta_k \ln X_{ki} + u_i. \qquad (5.9.3)$$

Кожний коефіцієнт регресії є еластичністю Y щодо відповідної змінної.

Застосуємо виробничу функцію для аналізу агросектору Тайваню за 1958–1972 рр. (табл. 5.2).

Вважаючи, що модель (5.9.2) задовольняє припущення лінійної регресійної моделі, одержуємо

$$\ln \hat{Y}_i = -3{,}3384 + 1{,}4988 \ln X_{2i} + 0{,}4899 \ln X_{3i} \qquad (5.9.4)$$

$$(2{,}4495)\quad (0{,}5398)\qquad (0{,}1020)$$
$$t = (-1{,}3629)\quad (2{,}7765)\qquad (4{,}8005)$$
$$R^2 = 0{,}8890\,, \qquad \mathrm{DF}=12;$$
$$\bar{R}^2 = 0{,}8705\,.$$

Таблиця 5.2

Показники агросектору Тайваню в 1958–1972 рр.

Рік	Валовий продукт Y, млн дол.	Трудодні X_2, млн	Капіталозатрати X_3, млн дол.
1958	16607,7	275,5	17803,7
1959	17511,3	274,4	18096,8
1960	20171,2	269,7	18271,8
1961	20932,9	267,0	19167,3
1962	20406,0	267,8	19647,6
1963	20831,6	275,0	20803,5
1964	24806,3	283,0	22076,6
1965	26465,8	300,7	23445,2
1966	27403,0	307,5	24939,0
1967	28628,7	303,7	26713,7
1968	29904,5	304,7	29957,8
1969	27508,2	298,6	31585,9
1970	29035,5	295,5	33474,5
1971	19281,5	299,0	34821,8
1972	31535,8	288,1	41794,3

Таким чином, отримуємо, що в агросекторі Тайваню за період 1958–1972 рр. коефіцієнт еластичності продукції по трудо- і капітало-витратам складав 1,4988 і 0,4899 відповідно. Іншими словами, за досліджуваний період при постійних капіталозатратах збільшення на 1% трудовитрат приводило в середньому до 1,5% зростання виробництва продукції. Аналогічно, при постійних трудовитратах збільшення на 1% капіталозатрат викликало приблизно 0,5% приросту продукції. Складаючи коефіцієнти еластичності виробництва продукції, одержуємо 1,9887, що дає міру обсягу приросту виробництва. Як ми бачимо, за досліджуваний період агросектор Тайваню мав зростаючий обсяг приросту виробництва.

Ми також бачимо, що побудована модель має достатньо високий коефіцієнт детермінації $R^2 = 0{,}889$, тобто близько 89% варіації логарифма продукції, яка випускається, пояснюється за рахунок варіації логарифмів трудовитрат і капіталозатрат.

5.10. Поліноміальная модель регресії

Розглянемо так званий клас поліноміальних моделей регресії, що широко застосовуються в дослідженнях, пов’язаних із функцією виробництва. Беручи до

розгляду ці моделі, ми ще більше розширюємо область застосування класичної лінійної моделі регресії.

Для кращого розуміння суті розглянемо рис. 5.3, який демонструє зв'язок короткострокових граничних витрат *Y* на виробництво продукції і обсяг виробництва *X*. Схематично побудована крива витрат має U-подібний вигляд, що свідчить про нелінійність зв'язку граничних витрат і обсягу виробництва. У зв'язку з цим виникає питання про вибір економетричної моделі, яка правильно б відображала первинне зниження та подальше зростання граничних витрат.

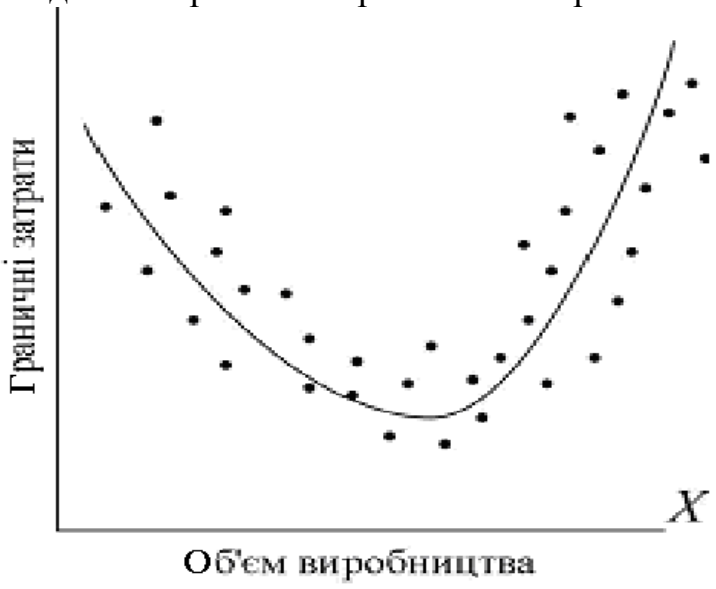


Рис. 5.3. Крива граничних витрат

Зображена на рис. 5.3 крива формою нагадує параболу. Рівняння параболи має вигляд

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 . \qquad (5.10.1)$$

Стохастичний аналог цього рівняння

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i . \qquad (5.10.2)$$

Дане рівняння має назву поліноміальної регресії 2-го порядку.

У загальному випадку поліноміальна регресія *k*-го порядку може бути записана у вигляді

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \ldots + \beta_k X_i^k + u_i . \qquad (5.10.3)$$

Відзначимо, що цей тип регресії має тільки одну пояснювальну змінну, яка входить в рівняння з різними степенями. Унаслідок цього ми маємо множинну регресію.

Чи має ця модель які-небудь особливості в порівнянні з раніше розглянутими? Оскільки коефіцієнти регресії входять в моделі (5.10.2), (5.10.3) лінійно, до них може бути застосований звичайний метод найменших квадратів. Чи виникає при цьому проблема мультиколінеарності? Чи можуть бути висококорельованими

різні степені X? Якщо пригадати, що X_2, X_3 і так далі є лінійними функціями, то мультиколінеарність поліноміальній регресії не загрожує.

Для прикладу поліноміальної регресії звернемося до даних із виробництва товару та загальних витрат на його виробництво, наведених у табл. 5.3.

Таблиця 5.3

Продукція, X	Загальні витрати, Y
1	193
2	226
3	240
4	244
5	257
6	260
7	274
8	297
9	350
10	420

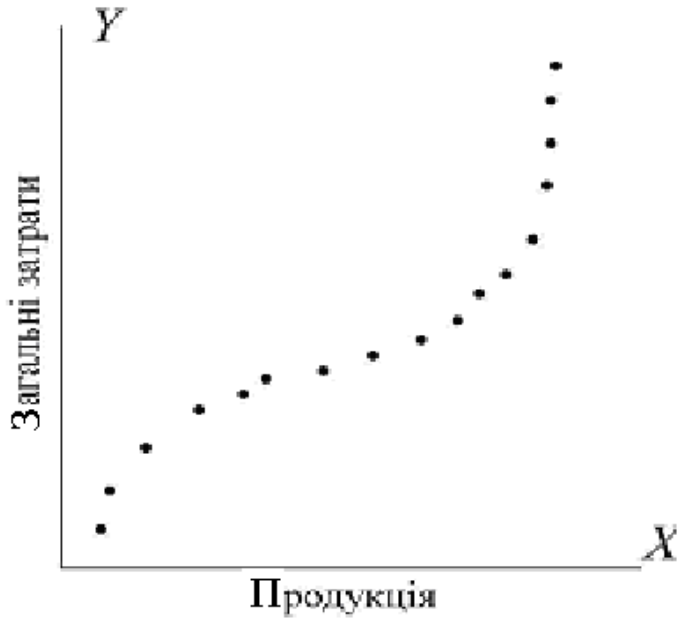


Рис. 5.4

Щоб визначити тип регресії, який відповідає даним табл.5.3, слід зобразити ці дані на площині. З поданого графіка (рис.5.4) бачимо, що співвідношення між випуском продукції та загальними витратами на виробництво має вигляд, подібний до букви S. Для опису залежності, яка нас цікавить, зручною є поліноміальна регресія третьої степені

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + u_i,$$

де Y – загальні витрати, а X – продукція.

Застосувавши цю модель для наведених у таблиці даних, отримаємо рівняння

$$\hat{Y}_i = 141,7667 + 63,4776 X_i - 12,96715 X_i^2 + 0,9396 X_i^2$$
$$(6,3753) \quad (4,7786) \quad (0,9857) \quad (0,0591)$$
$$R^2 = 0,9983.$$

У дужках вказані стандартні похибки коефіцієнтів регресії.

Висновки

1. Термін лінійна модель множинної регресії стосується поняття лінійності за коефіцієнтами регресії.
2. Хоча модель регресії з трьома змінними багато в чому є продовженням двовимірної моделі, з'являються й нові поняття, такі як "частинні коефіцієнти регресії", "частинні коефіцієнти кореляції", "множинний коефіцієнт кореля-

ції”, “скорегований і нескорегований коефіцієнти детермінації”, “мультиколінеарність”.

3. Хоча R^2 і $\bar{R}^2$ є сукупною мірою якості підгонки регресії до початкових даних, їх значущість не слід перебільшувати. Критичними є такі чинники, як відповідність знаків коефіцієнтів моделі нашим апріорним уявленням, а також їх статистична значущість.
4. Наведені результати для моделі з трьома змінними можна узагальнити на випадок лінійної моделі регресії, що містить будь-яку кількість регресорів. Однак при цьому алгебраїчні вирази стають дуже громіздкими. Цей недолік долається шляхом переходу до матричних операцій.

6. ПРИПУЩЕННЯ НОРМАЛЬНОСТІ РОЗПОДІЛУ ЗАЛИШКІВ

Якби нашою єдиною метою було отримання точних оцінок параметрів моделі регресії за МНК, який не вимагає ніяких припущень про закон розподілу залишків, то отриманих результатів було б достатньо. Проте якщо метою дослідження є й отримання статистичних висновків, то нам необхідно зробити припущення про закон розподілу u_i.

З причин, детально розглянутих раніше, ми знову припускаємо, що u_i підпадають під нормальний закон розподілу з нульовим середнім і постійною дисперсією $u_i \sim N(0, \sigma^2)$ і σ^2. При цьому припущенні виявляється, що знайдені за МНК частинні коефіцієнти регресії є найкращими лінійними незміщеними оцінками. Крім того, оцінки $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$ самі підкоряються нормальному закону розподілу із середніми значеннями β_1, β_2, β_3 і дисперсіями, визначуваними за формулами (5.4.9), (5.4.11) і (5.4.14). Величина $\dfrac{(n-3)\hat{\sigma}^2}{\sigma^2}$ підкоряється закону розподілу χ^2 з (n–3) степенями вільності, а три оцінки коефіцієнтів регресії розподілено незалежно від $\hat{\sigma}^2$. На підставі вищезазначеного можна показати, що замінюючи σ^2 на її незміщену оцінку $\hat{\sigma}^2$ у формулах для стандартних похибок, кожна з величин

$$t = \frac{\hat{\beta}_1 - \beta_1}{\text{se}(\hat{\beta}_1)}; \tag{6.1}$$

$$t = \frac{\hat{\beta}_2 - \beta_2}{\text{se}(\hat{\beta}_2)}; \tag{6.2}$$

$$t = \frac{\hat{\beta}_3 - \beta_3}{\text{se}(\hat{\beta}_3)} \tag{6.3}$$

розподілена за законом t-розподілу з (n–3) степенями вільності.

Відзначимо, що кількість степенів вільності тепер складає (n–3), оскільки для обчислення $\sum \hat{u}_i^2$ і, отже, $\hat{\sigma}^2$ нам спочатку необхідно визначити три частинні коефіцієнти регресії. Унаслідок цього на суму квадратів залишків (RSS) накладаються три обмеження (за логікою в моделі з чотирма змінними кількість степенів вільності дорівнює n–4 і т.д.).

Таким чином, t-розподіл може бути застосований для побудови довірчих інтервалів, а також для перевірки статистичних гіпотез про значення істинних коефіцієнтів регресії. Аналогічно, розподіл χ^2 може бути використаний для перевірки гіпотез про істинне значення σ^2. Для демонстрації описаного механізму звернемося до ілюстрованого прикладу.

Зв'язок між особистими витратами й особистими доходами в США в 1956 –1970 рр.

Припустимо, що ми хочемо досліджувати зміну особистих витрат в США за декілька минулих років. Скористаємося такою простою моделлю:

$$E(Y \mid X_2, X_3) = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i}, \qquad (6.4)$$

де Y – особисті витрати;

X_2 – особистий дохід (після сплати податків);

X_3 – час, що вимірюється в роках.

Рівняння (6.4) постулювало лінійний зв'язок між особистими доходами, доходами і часом або змінною тренда. У більшості випадків множинного регресійного аналізу, що включає дані тимчасових рядів, у модель вводять на додаток до інших пояснювальних змінних змінну тренда, що пояснюється такими причинами.

1. Нас може цікавити, як залежна змінна поводиться зі зміною часу. Наприклад, на графіках часто демонструється, скажімо, поведінка ВНП, зайнятості, безробіття, вартості акцій тощо за деякі послідовні інтервали часу. Вид цих графіків дозволяє визначити тенденцію зміни величини за часом, її зростання, спадання або відсутність певної тенденції. У подібному аналізі нас можуть не цікавити причини, що стоять за тенденцією зростання або спадання досліджуваної величини, а метою може бути просто опис даних, змінюваних у часі.

2. У багатьох випадках змінна тренда є заміною для основної змінної, що впливає на Y. Але ця основна змінна не дозволяє отримати дані, що описують її, або їх отримання пов'язане зі значними труднощами. Наприклад, у моделі виробництва подібною змінною є технологія. Ми чудово розуміємо її вплив, проте незрозуміло, як його виміряти. Отже, можна припустити, що технологія є деяка функція часу, що вимірюється хронологічно. У деяких ситуаціях можна вважати, що впливаюча на Y основна змінна настільки тісно пов'язана з часом, що зручніше ввести в модель сам час, а не цю змінну. Наприклад, у моделі (6.4) час X_3 може представляти населення. Витрати на споживання зростають зі зростанням населення, а саме населення може бути достатньо добре описане лінійною функцією часу.

4. Ще однією підставою для введення змінної тренда є бажання уникнути псевдокореляції. Дані, що стосуються економічних тимчасових рядів, таких як PCE і PDI, часто змінюються в одному напрямі, відображаючи тенденцію зростання або спадання. Отже, якщо побудувати модель регресії PCE по відношенню до PDI, то можна отримати велике значення R^2, що може не відображати істинного зв'язку між PCE і PDI, а просто бути наслідком загального тренда. Щоб виключити псевдоасоційованість між економічними тимчасовими рядами, можна застосувати кілька прийомів. В одному випадку, припускаючи, що тимчасові ряди демонструють лінійну тенденцію, можна ввести в модель час або саму тенденцію, як це зроблено в рівнянні (6.4). Унаслідок чого β_2 в рівнянні (6.4) відображає істинний зв'язок між PCE і PDI. В іншому випадку можна здійснити процедуру детренда Y (PCE) і X_2 (PDI), а потім провести регресію за отриманими після детренда змінними Y і X_2, припускаючи наявність лінійного етапу, як це було виконано в розд. 5. Спочатку проводимо регресію Y за X_3 (час) і одержуємо залишки від цієї регресії, скажімо $\hat{u}_{1t}$. Потім проводимо регресію X_2 за X_3 і одержуємо залишки з цієї регресії, наприклад $\hat{u}_{2t}$. Нарешті, проводимо регресію $\hat{u}_{1t}$ за $\hat{u}_{2t}$, які не піддаються впливу (лінійному) часу. Кутовий коефіцієнт у цій регресії відображатиме істинний зв'язок між Y і X_2 і повинен, отже, дорівнювати коефіцієнту β_2. Зрозуміло, що перший метод більш придатний, оскільки він менш трудомісткий.

Для перевірки моделі (6.4) скористаємося даними, наведеними в табл. 6.1

Таблиця 6.1

Витрати на споживання і особистий дохід у США за 1956–1970 рр.

PCE, Y, млрд дол.	PDI, X_2, млрд дол.	Час, X_3
281,4	309,3	1956=1
288,1	316,1	1957=2
290,0	318,8	1958=3
307,3	333,0	1959=4
316,1	340,3	1960=5
322,5	350,5	1961=6
338,4	367,2	1962=7
353,3	381,2	1963=8
373,7	408,1	1964=9
397,7	434,8	1965=10
418,1	458,9	1966=11
430,1	477,5	1967=12
452,7	499,0	1968=13
468,1	513,5	1969=14
476,9	533,2	1970=15

Рівняння регресії, отримане за даними табл.6.1, має вигляд

$$\hat{Y}_t = 53{,}1603 + 0{,}7266 X_{2i} + 2{,}7363 X_{3i} \quad (6.5)$$

$$(13{,}0261)\ \ (0{,}0487)\ \ \ \ (0{,}8486)$$

$$t = (4{,}0811)\ \ (14{,}9060)\ \ \ (3{,}2246)$$

$$p = (0{,}0008) \quad (0{,}000) \qquad (0{,}0036)$$

$$\text{DF}=12, \quad R^2=0{,}9988, \quad \bar{R}^2 = 0{,}9986, \quad F_{2,12}=5128{,}88$$

Інтерпретація рівняння (6.5). Якщо обидві змінні X_2 і X_3 перетворюються в нуль, то середня величина витрат на споживання, що, можливо, відображає вплив не включених у модель чинників, приблизно дорівнює 53,16 млрд. дол. Як уже неодноразово наголошувалося величина коефіцієнта β_1 часто не має істотного економічного змісту. Частинний коефіцієнт регресії 0,7266 означає, що при фіксованому значенні решти змінних (X_3 у даному випадку), збільшення особистого доходу на 1 дол. спричиняє зростання особистих витрат на споживання на 73 центи. Аналогічно, якщо не змінювати значення X_2, то за рік особисті витрати на споживання зростають приблизно на 2,7 млрд дол. Значення R^2=0,9988 показує високу степінь (близько 99.9%) пояснення двома змінними дисперсії за даний період. Скорегований $\bar{R}^2 = 0{,}9986$ коефіцієнт детермінації демонструє, що після обчислення кількості степенів вільності змінні X_2 і X_3 продовжують пояснювати близько 99,8% дисперсії Y.

7. ПЕРЕВІРКА ГІПОТЕЗ МНОЖИННОЇ РЕГРЕСІЇ. ЗАГАЛЬНІ ЗАУВАЖЕННЯ

Перевірка гіпотез множинної регресії має декілька нових форм, відмінних від найпростішої двовимірної моделі регресії. До них належать:

1. Перевірка гіпотез про частинні коефіцієнти регресії.
2. Перевірка загальної значущості оціненої множинної моделі регресії, тобто чи можуть перетворюватися в нуль одночасно всі частинні кутові коефіцієнти регресії.
3. Перевірка того, що два або більше кутових коефіцієнти дорівнюють один одному.
4. Перевірка того, що частинний коефіцієнт регресії задовольняє певне обмеження.
5. Перевірка стабільності оціненої моделі регресії за часом.
6. Перевірка функціонального виду моделі регресії.

Оскільки перераховані питання зустрічаються на практиці достатньо часто, розглянемо кожне з них окремо.

7.1. Перевірка гіпотези про частинний коефіцієнт регресії

Якщо ми приймаємо гіпотезу про те, що $u_i \sim N(0, \sigma^2)$, тоді, як було відзначено раніше, ми можемо скористатися *t*-тестом для перевірки гіпотези щодо будь-якого частинного коефіцієнта регресії. Для ілюстрації механізму розглянемо наш числовий приклад. Припустимо, ми постулювали

$$H_0 : \beta_2 = 0 \text{ і } H_1 : \beta_2 \neq 0 .$$

Нульова гіпотеза твердить, що зберігаючи постійним значення X_3, особистий дохід не має лінійного впливу на особисті витрати на споживання. Для пере-

вірки нульової гіпотези використовуємо *t*-тест, визначений рівністю (6.1.4). Відповідно до нашої методики ми перевіряємо, чи перевершує значення *t*-тесту критичне значення *t*-величини при вибраному нами рівні значущості. Якщо це так, то ми відкидаємо нульову гіпотезу, а в противному випадку – залишаємо. Для нашого прикладу, застосовуючи (6.1.4) і враховуючи, що згідно з нульовою гіпотезою $\beta_2 = 0$, одержуємо

$$t = \frac{0,7266}{0,0487} = 14,92. \qquad (7.1.1)$$

Якщо ми візьмемо α=0,05, $t_{\alpha/2}$=2,179 для DF=12. Оскільки підрахована величина значно перевершує критичне значення $t_{кр}$=2,179, ми можемо відкинути нульову гіпотезу і стверджувати, що $\hat{\beta}_2$ статистично значиме, тобто значно відрізняється від нуля. Більше того, як показано у (6.1.5), отримана за (7.1.1) відповідна *p*–величина вкрай мала. На рис. 7.1. описана вище процедура проілюстрована графічно.

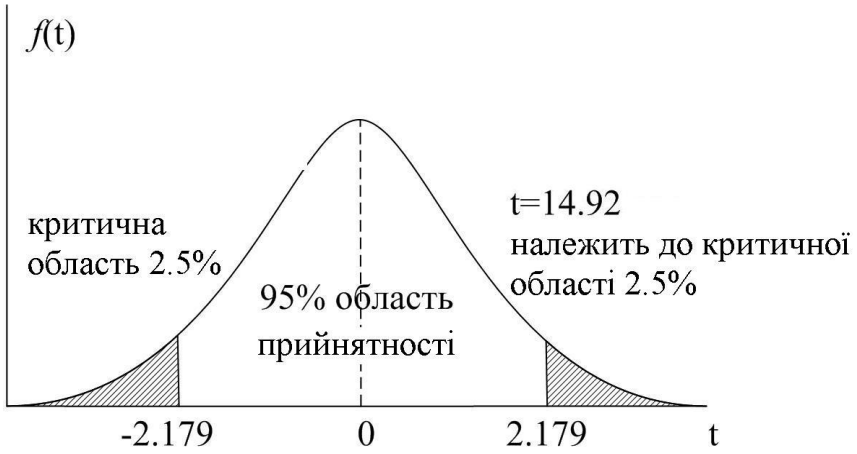


Рис. 7.1.

Раніше ми встановили тісний зв'язок між перевіркою гіпотез і оціненими довірчими інтервалами. Для нашого прикладу 95%-й довірчий інтервал для β_2 визначається нерівністю

$$\hat{\beta}_2 - t_{\alpha/2}\,\mathrm{se}(\hat{\beta}_2) \le \beta_2 \le \hat{\beta}_2 + t_{\alpha/2}\,\mathrm{se}(\hat{\beta}_2).$$

Одержуємо

$$0,6205 \le \beta_2 \le 0,8327 \qquad (7.1.2)$$

Таким чином, β_2 лежить між 0,6205 і 0,8327 з 95% вірогідністю. Так, якщо 100 вибірок обсягом 15 одиниць дозволяють побудувати 100 довірчих інтервалів вигляду $\hat{\beta}_2 \pm t_{\alpha/2}\,\mathrm{se}(\hat{\beta}_2)$, то 95 із них повинні містити істинне значення параметра β_2. Оскільки значення β_2, постулюємо нульовою гіпотезою, не потрапляє в інтервал (7.1.2), ми приходимо до висновку про відхилення нульової гіпотези. Це очевидно, враховуючи зв'язок між *t*-тестом і довірчими інтервалами.

Дотримуючись аналогічної процедури, ми можемо перевірити й інші гіпотези, що стосуються параметрів моделі (6.4), якщо скористатися результатами регресійного аналізу (6.5). Наприклад, ми можемо відкинути гіпотези про рівність нулю решти коефіцієнтів регресії з α=0,05.

7.2. Перевірка вибіркової регресії на загальну значущість

У попередньому підрозділі ми зосередили свою увагу на перевірці на значущість окремих оцінених частинних коефіцієнтів регресії, тобто перевірялася гіпотеза, що окремо взятий коефіцієнт регресії дорівнює нулю. Зараз ми розглянемо гіпотезу

$$H_0 : \beta_2 = \beta_3 = 0. \qquad (7.2.1)$$

Дана нульова гіпотеза припускає, що коефіцієнти β_2 і β_3 одночасно дорівнювати нулю. Перевірка подібної гіпотези називається перевіркою на загальну значущість оціненої лінії регресії, тобто перевіркою на зв'язок Y і змінних X_2 і X_3.

Чи можна замінити перевірку гіпотези (7.2.1) перевіркою на значущість кожного окремого коефіцієнта β_2 і β_3? Відповідь на це питання негативна з таких причин.

При перевірці на індивідуальну значущість отриманих частинних коефіцієнтів регресії ми неявно припускали, що кожен тест на значущість ґрунтується на незалежній вибірці. Так, при перевірці на значущість коефіцієнта $\hat{\beta}_2$ при нульовій гіпотезі $\beta_2 = 0$ неявно передбачається, що тестування ґрунтується на даних іншої вибірки стосовно вибірки, що використовується для перевірки гіпотези про $\beta_3 = 0$. Але для перевірки спільної гіпотези (7.2.1) при використанні даних табл. 6.1 ми порушуємо припущення, на якому ґрунтується процедура тестування. Ми можемо розглядати це питання й інакше. Нерівність (7.1.2) задає 95%-й довірчий інтервал для коефіцієнта β_2. Але якщо ми використовуємо ті ж вибіркові дані для знаходження 95%-го довірчого інтервалу для коефіцієнта β_3, то ми не можемо стверджувати, що обидва коефіцієнти β_2 і β_3 лежать усередині відповідних довірчих інтервалів із вірогідністю $(1-\alpha)\times(1-\alpha)$=0,90.

Іншими словами, хоча твердження

$$\Pr\left[\hat{\beta}_2 - t_{\alpha/2}\,\mathrm{se}(\hat{\beta}_2) \le \beta_2 \le \hat{\beta}_2 + t_{\alpha/2}\,\mathrm{se}(\hat{\beta}_2)\right] = 1-\alpha;$$

$$\Pr\left[\hat{\beta}_3 - t_{\alpha/2}\,\mathrm{se}(\hat{\beta}_3) \le \beta_3 \le \hat{\beta}_3 + t_{\alpha/2}\,\mathrm{se}(\hat{\beta}_3)\right] = 1-\alpha$$

індивідуально справедливі, однак несправедливе твердження про те, що вірогідність одночасного потрапляння коефіцієнтів β_2 і β_3 в інтервали $\hat{\beta}_2 \pm t_{\alpha/2}\,\mathrm{se}(\hat{\beta}_2)$, $\hat{\beta}_3 \pm t_{\alpha/2}\,\mathrm{se}(\hat{\beta}_3)$ є $(1-\alpha)^2$, оскільки ці інтервали можуть не бути незалежними при використовуванні однакових даннях при їх знаходженні.

Перевірка на загальну значущість множинної регресії на підставі аналізу дисперсії. F–тестування

Ми вже говорили, що *t*-тест не дозволяє судити про загальну значущість множинної регресії. Для цього застосовується підхід, який базується на аналізі дисперсії (ANOVA), розглянутий нами раніше для моделі з двома змінними. Даний підхід застосовується і в разі множинної регресії.

Пригадаємо, що

$$\sum y_i^2 = \hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i} + \sum \hat{u}_i^2 ; \qquad (7.2.2)$$

TSS=ESS+RSS.

Як уже раніше наголошувалося, TSS має (n–1) степенів вільності, RSS – (n–3), а ESS – 2 степені. Повторюючи знайому нам процедуру аналізу дисперсії, складаємо таку ANOVA-таблицю:

Таблиця 7.1

ANOVA-таблиця для регресії з трьома змінними

Джерело дисперсії	Сума квадратів	DF	Середня сума квадратів
За регресією (ESS)	$\hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i}$	2	$\left(\hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i}\right)/2$
За залишками (RSS)	$\sum \hat{u}_i^2$	N–3	$\hat{\sigma}^2 = \dfrac{\sum \hat{u}_i^2}{n-3}$
Загальна	$\sum y_i^2$	N–1	

Можна показати, що при припущенні нормальності розподілу залишків u_i і нульовій гіпотезі $\beta_2 = \beta_3 = 0$ змінна

$$F = \frac{\left(\hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i}\right)/2}{\sum \hat{u}_i^2 /(n-3)} = \frac{ESS/df}{RSS/df} \qquad (7.2.3)$$

підкоряється закону F-розподілу з 2 і (n–3) степенями вільності.

При припущенні, що $u_i \sim N(0, \sigma^2)$, справедлива рівність

$$E\left(\frac{\sum \hat{u}_i^2}{n-3}\right) = E\left(\hat{\sigma}^2\right) = \sigma^2 . \qquad (7.2.4)$$

Додатково припустивши, що $\beta_2 = \beta_3 = 0$, можна також показати, що

$$E\left(\frac{\hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i}}{2}\right) = \sigma^2 . \qquad (7.2.5)$$

Отже, якщо нульова гіпотеза виконується, то обидві рівності (7.2.4) і (7.2.5) дають ідентичні оцінки для σ^2. Оскільки через зв'язок, існуючий між Y, X_2 і X_3, єдиним джерелом дисперсії Y є випадкова складова u_i. Якщо ж нульова гіпотеза несправедлива, тобто X_2 і X_3 безумовно впливають на y, еквівалентність між

(7.2.4) і (7.2.5) виконуватися не буде. У такому випадку величина ESS буде порівняно більша, ніж RSS, враховуючи кількість їх степенів вільності. Отже, величина *F*, визначувана рівністю (7.2.3), може служити тестом на перевірку нульової гіпотези про те, що кутові коефіцієнти β_i одночасно перетворюються в нуль. Якщо підрахована за (7.2.3) величина *F* більша, ніж визначене за таблицею критичне для вибраного рівня значущості значення, то ми відкидаємо гіпотезу H_0, в іншому випадку – не відкидаємо. Альтернативою служить використання *p*–величини. Якщо отримане з використанням *F* значення *p*–величини достатньо мале, ми можемо відхиляти гіпотезу H_0.

Наведемо дані аналізу дисперсії для нашого прикладу (табл. 7.2).

Таблиця 7.2

Джерело дисперсії	Сума квадратів	DF	Середнє значення
За регресією	65965.1003	2	32982,5502
За залишками	77.1690	12	6,4308
Усього	66042.2693	14	

За даними з табл.7.2 знаходимо

$$F = \frac{32982,5502}{6,4308} = 5128,8181. \qquad (7.2.6)$$

Якщо ми візьмемо α=0.05, то критичне значення *F* для 2 і 12 степенів вільності буде $F_{0.05}(2, 12)$=3,88529. Очевидно що підрахована величина *F* значуща і ми можемо, отже, відхиляти нульову гіпотезу. Якщо вибрати рівень значущості α=0.01, то $F_{0.01}(2, 12)$=6.9266. Підрахована нами величина *F* залишається значно перевершуючою це критичне значення. Ми, як і раніше, відхиляємо нульову гіпотезу. Побіжно зазначимо, що p–величина, відповідна значенню *F*, украй мала (2.545×10^{-18}).

Описану процедуру F–тестування можна узагальнити на випадок множинної регресії з *k* змінними.

Хай задана розглянута модель з *k* змінними

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + ... + \hat{\beta}_k X_{ki} + \hat{u}_i .$$

Для перевірки гіпотези

$$H_0 : \beta_2 = \beta_3 = ... = \beta_k = 0$$

і альтернативної гіпотези

H_1: не всі кутові коефіцієнти одночасно дорівнюють нулю

підрахуємо

$$F = \frac{ESS/df}{RSS/df} = \frac{ESS/(k-1)}{RSS/(n-k)}. \qquad (7.2.7)$$

Якщо $F>F_\alpha(k–1, n–k)$, відхиляємо гіпотезу H_0; у протилежному випадку – не відхиляємо. Тут $F_\alpha(k–1, n–k)$ – критичне значення величини F для заданого рівня значущості α, $(k–1)$ – кількість степенів вільності в чисельнику і $(n–k)$ – кількість степенів вільності в знаменнику. За альтернативним підходом, якщо відповідна F величина p достатньо мала, відхиляємо гіпотезу H_0.

Зв'язок між R^2 і F

Існує внутрішній зв'язок між коефіцієнтом детермінації R^2 і F-тестуванням, що використовується в аналізі дисперсії. Припускаючи нормальний закон розподілу збурень u_i і приймаючи нульову гіпотезу про те, що $\beta_2 = \beta_3 = 0$, ми бачимо, що

$$F = \frac{ESS/2}{RSS/(n-3)} \qquad (7.2.8)$$

розподілена за законом F-розподілу з 2 і $(n–3)$ степенями вільності.

Для загального випадку моделі з k змінними, припускаючи нормальний закон розподілу залишків і приймаючи нульову гіпотезу

$$H_0 : \beta_2 = \beta_3 = ... = \beta_k = 0 \qquad (7.2.9)$$

отримуємо, що

$$F = \frac{ESS/(k-1)}{RSS/(n-k)} \qquad (7.2.10)$$

розподілене за законом F-розподілу з $(k–1)$ і $(n–k)$ степенями вільності. Перетворивши (7.2.10), отримаємо

$$F = \frac{n-k}{k-1}\frac{ESS}{RSS} = \frac{n-k}{k-1}\frac{ESS}{TSS-ESS} = \frac{n-k}{k-1}\frac{ESS/TSS}{1-ESS/TSS} = \\ = \frac{n-k}{k-1}\frac{R^2}{1-R^2} = \frac{R^2/(k-1)}{(1-R^2)(n-k)}. \qquad (7.2.11)$$

При цьому ми спиралися на той факт, що $R^2 = ESS/TSS$. Із рівняння (7.2.11) бачимо, що величини F і R^2 пов'язані безпосередньо. Якщо R^2=0, то й F=0. Чим більше R^2, тим більше F. У межі при $R^2 \to 1,\ F \to \infty$.

Для випадку моделі з трьома змінними (7.2.11) набуває вигляду

$$F = \frac{R^2/2}{(1-R^2)(n-3)}. \qquad (7.2.12)$$

Враховуючи тісний зв'язок між F і R^2, ANOVA-таблиця може бути перетворена до вигляду табл. 7.3.

Таблиця 7.3

ANOVA-таблиця в термінах R^2

Джерело дисперсії	SS	DF	MSS
За регресією	$R^2\left(\sum y_i^2\right)$	2	$R^2\left(\sum y_i^2\right)/2$
За залишками	$\left(1-R^2\right)\sum y_i^2$	N–3	$\left(1-R^2\right)\sum y_i^2/(n-3)$
Загальна	$\sum y_i^2$	N–1	

Наведемо правило перевірки на загальну значущість множинної регресії в термінах R^2.

Правило ухвалення рішення. Хай задана модель регресії з k змінними

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + ... + \hat{\beta}_k X_{ki} + \hat{u}_i .$$

Для перевірки гіпотези

$$H_0 : \beta_2 = \beta_3 = ... = \beta_k = 0$$

і альтернативної їй гіпотези

H_1: не всі кутові коефіцієнти одночасно дорівнюють нулю

підраховуємо

$$F = \frac{R^2/(k-1)}{(1-R^2)(n-k)} . \qquad (7.2.13)$$

Якщо $F>F_\alpha(k–1, n–k)$, відхиляємо гіпотезу H_0, у протилежному випадку можна прийняти H_0. Можна скористатися підходом, що базується на обчисленні p-величини. Якщо отримана відповідно до (7.2.13) p-величина достатньо мала, ми відхиляємо гіпотезу H_0.

Додатковий або граничний внесок пояснювальної змінної

Повернемося до ілюстрованого прикладу. З (7.2.2) ми знаємо, що коефіцієнти при X_2 (дохід) і при X_3 (тренд) значущо відрізняються від нуля на основі роздільних t-тестів. Ми також бачили, що на підставі F-тестів (7.2.7) і (7.2.13) сама лінія регресії також значуща. Припустимо тепер, що ми вводимо X_2 і X_3 послідовно, тобто ми спочатку проводимо регресію Y за X_2 і оцінюємо її значущість, а потім додаємо до моделі X_3 і визначаємо, наскільки її внесок виявляється істотним (зрозуміло, що порядок, у якому змінні вводяться в модель, може бути змінений на протилежний). Під внеском ми розуміємо таке: Чи приводить додавання змінної до моделі до зростання ESS (і R^2) значущо в порівнянні з RSS. Цей внесок можна назвати додатковим або граничним внеском пояснювальної змінної.

Дослідження питання про додатковий внесок змінної має важливе практичне значення. У більшості емпіричних досліджень може бути не зовсім зрозуміло, чи варто додавати до моделі змінну *X* крім інших змінних, включених в модель. Не слід включати в модель змінну, внесок якої в ESS незначний, і виключати з моделі змінну, що має в ESS істотний внесок. Як же визначити, наскільки істотно змінна *X* зменшує RSS? Розвиток техніки аналізу дисперсій дозволяє легко відповісти на це питання.

Припустимо, що ми спочатку проводимо регресію *Y* (особисті витрати на споживання) за X_2 (особистий дохід) і одержуємо таку регресію:

$$\hat{Y}_i = 12,762 + 0,8812 X_{2i}$$
$$(4,6818) \quad (0,0114)$$
$$t = (2,7259) \quad (77,2982) \qquad (7.2.14)$$
$$R^2 = 0,9978, \qquad \bar{R}^2 = 0,9977$$

Згідно з нульовою гіпотезою $\beta_{12} = 0$, а отже t=77.2982. Очевидно, що β_{12} значущо відрізняється від нуля як при 5%-му рівні значущості, так і при 1%-му рівні. Таким чином, X_2 значущо впливає на *Y*. Таблиця ANOVA регресії (7.2.14) подана нижче.

Таблиця 7.4

ANOVA-таблиця для регресії (7.2.14)

Джерело дисперсії	SS	DF	MSS
ESS	65898,2353	1	65898,2353
RSS	144,0340	13	11,0800
Загальна	66042,2693	14	

Припускаючи нормальність закону розподілу збурень u_i і нульову гіпотезу $\beta_{12} = 0$, ми маємо, що F=5947.494 згідно із законом *F*-розподілу з 1 і 13 степенями вільності.

Очевидно, що знайдене значення *F* свідчить про значущість β_{12} при загальноприйнятих рівнях значущості. Таким чином, так як і раніше, ми можемо відхиляти гіпотезу про $\beta_{12} = 0$. Побіжно зазначимо, що T^2=(77.2982)^2=5947.494. Цей результат очікуваний, оскільки ми знаємо, що при однаковій нульовій гіпотезі й однаковому рівні значущості квадрат величини t з (n–2) степенями вільності рівний величині *F* з 1 і (n–2) степеням вільності.

Припустимо, що після отримання регресії (7.2.14) ми ви рішили додати до моделі змінну *X3* (тренд) і отримали множинну регресію (7.2.2). При цьому слід отримати відповіді на такі питання. Який додатковий внесок у модель змінної X_3 в порівнянні з моделлю (7.2.14)? Чи є отримана в порівнянні з моделлю (7.2.14), добавка статистично значущою? Який критерій для включення в модель додаткових змінних? Відповіді на ці питання можна отримати на підставі аналізу дисперсій. Для цього розглянемо табл. 7.5.

Таблиця 7.5

ANOVA-таблиця для оцінки додаткового внеску змінної.

Джерело дисперсії	SS	DF	MSS
ESS унаслідок X_2	$Q_1 = \hat{\beta}_{12}^2 \sum x_{2i}^2$	1	$Q_1/1$
ESS унаслідок X_3	$Q_2=Q_3-Q_1$	1	$Q_2/1$
ESS унаслідок X_2 і X_3	$Q_3 = \hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i}$	2	$Q_3/2$
RSS	$Q_4=Q_5-Q_3$	$N-3$	$Q_4/(N-3)$
Загальна	$Q_5 = \sum y_i^2$	$N-1$	

Для даного прикладу одержуємо такі результати (табл. 7.6).

Таблиця 7.6

Джерело дисперсії	SS	DF	MSS
ESS унаслідок X_2	Q_1=65898,2353	1	65898,2353
ESS унаслідок X_3	Q_2=66,8647	1	66,8647
ESS унаслідок X_2 і X_3	Q_3=65965,100	2	32892,55
RSS	Q_4=77,1693	12	6,4302
Загальна	Q_5=66042,2693	14	

Для отримання оцінки додаткового внеску від X_3 до вже врахованої змінної X_2 підраховуємо

$$F' = \frac{Q_2 \,/\, DF}{Q_4 \,/\, DF} = \frac{(\mathrm{ESS_н} - \mathrm{ESS_c})/\mathrm{m}}{\mathrm{RSS_н}\,/(n-k)} = \frac{Q_2\,/1}{Q_4\,/(15-3)}, \qquad (7.2.15)$$

де $\mathrm{ESS_н}$ – оцінена сума квадратів нової моделі (тобто після додавання нових регресорів); $\mathrm{ESS_c}$ – оцінена сума квадратів старої моделі (Q_1); $\mathrm{RSS_н}$ – сума квадратів залишків (Q_4); m – кількість нових регресорів; k – кількість параметрів у новій моделі.

Для нашого прикладу одержуємо

$$F=10{,}3973. \qquad (7.2.16)$$

Тепер при звичайних припущеннях про нормальність u_i і нульовій гіпотезі $\beta_3 = 0$ можна показати, що підрахована за (7.2.15) величина F підкоряється закону F-розподілу з 1 і 12 степенями вільності.

Побіжно відзначимо, що величину F з (7.2.15) можна подати тільки через значення R^2, як це було зроблено в (7.2.13). Одержуємо вираз

$$F = \frac{(R_\mathrm{н}^2 - R_\mathrm{c}^2)\,/\,DF}{(1-R_\mathrm{н}^2)\,/\,DF} = \frac{(R_\mathrm{н}^2 - R_\mathrm{c}^2)\,/\,m}{(1-R_\mathrm{н}^2)\,/(n-k)}. \qquad (7.2.17)$$

Для нашого прикладу, застосовуючи (7.2.2) і (7.2.14), отримуємо

$$R_\mathrm{н}^2 = 0{,}9988,\ R_\mathrm{c}^2 = 0{,}9978.$$

Отже,

$$F=10{,}973. \qquad (7.2.18)$$

Таким чином, на підставі будь-якого з *F*-тестів, ми можемо відкинути нульову гіпотезу і зробити висновок, що додавання в модель X_3 значущо підвищує ESS і, отже, R^2. Висновок: змінну тренда X_3 слід додати в модель.

Пригадаємо, що в (7.2.2) ми отримали t=3,246 для коефіцієнта при X_3 при гіпотезі $H_0 : \beta_3 = 0$. Зазначимо, що t^2=10,973=F. Це співвідношення очевидне, якщо врахувати зв'язок між F і t^2, згадуваний раніше.

Описана процедура *F*-теста дає формальний метод ухвалення рішення про те, чи варто включати додаткову змінну в модель регресії. Часто дослідники стикаються із задачею вибору найкращої з моделей, що мають однакову пояснювану змінну, але різні пояснювальні. Багато хто як критерій вибору використовує величину $\overline{R}^2$, тобто вибирають модель із найбільшим $\overline{R}^2$. Отже, включення змінної спричиняє зростання $\overline{R}^2$, вона зберігається в моделі, хоча й не зменшує RSS значущо в статистичному значенні. У такому разі виникає питання: коли $\overline{R}^2$ зростає? Можна показати, що $\overline{R}^2$ зростатиме, якщо величина t для коефіцієнта при новій доданій в модель змінній за абсолютною величиною більше 1, де t – статистика, підрахована при нульовій гіпотезі про рівність нулю відповідного коефіцієнта. Наведений критерій може бути сформульований й інакше. $\overline{R}^2$ зростатиме з додаванням нової пояснювальної змінної, якщо F (=t^2) цієї змінної більше 1.

За цим критерієм змінна тренда X_3 з t=3.2246 або F=10.3973 повинна приводити до зростання $\overline{R}^2$, що насправді й відбувається. З її одаванням $\overline{R}^2$ зростає з 0,9977 до 0,9986.

Чи можна поширити наведений критерій на групу коефіцієнтів, що додаються? Відповідь очевидна з (7.2.17). Якщо додавання до моделі групи змінних дає величину F більше 1, то $\overline{R}^2$ при цьому зростає.

7.3. Перевірка на рівність двох коефіцієнтів регресії

Припустимо, що в множинній регресії

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i \qquad (7.3.1)$$

ми хочемо перевірити гіпотези

$$H_0 : \beta_3 = \beta_4 \text{ або } \beta_3 - \beta_4 = 0,$$
$$H_1 : \beta_3 \neq \beta_4 \text{ або } \beta_3 - \beta_4 \neq 0 \qquad (7.3.2)$$

про рівність двох кутових коефіцієнтів.

Подібна гіпотеза має важливе практичне значення. Наприклад, хай у моделі (7.3.1) представлена функція попиту на товар. У ній Y – величина попиту на товар, X_2 – ціна товару, X_3 – дохід покупця, X_4 – статки покупця. Нульова гіпотеза в цьому випадку означає, що коефіцієнти з доходу й статків збігаються. Або якщо Y_i і змінні X виражені в логарифмічному вигляді

$$\ln Y_i = \beta_1 + \beta_2 \ln X_{2i} + \beta_3 \ln X_{3i} + \beta_4 \ln X_{4i} + u_i,$$

то нульова гіпотеза (7.3.2) припускає, що еластичності за доходом і статками однакові.

Як перевірити цю гіпотезу? Можна показати, що при виконанні звичайних класичних припущень величина

$$t = \frac{(\hat{\beta}_3 - \hat{\beta}_4) - (\beta_3 - \beta_4)}{\text{se}(\hat{\beta}_3 - \hat{\beta}_4)} \qquad (7.3.3)$$

випливає із закону *t*-розподілу з (*n*–4) степенями вільності, оскільки (7.3.1) являє собою модель із 4 змінними. У загальному випадку моделі з *k* параметрами кількість степенів вільності дорівнює (*n*–*k*). Величина $\text{se}(\hat{\beta}_3 - \hat{\beta}_4)$, що входить в (7.3.3) обчислюється за формулою

$$\text{se}(\hat{\beta}_3 - \hat{\beta}_4) = \sqrt{D(\hat{\beta}_3) + D(\hat{\beta}_4) - 2\,\text{cov}(\hat{\beta}_3, \hat{\beta}_4)}\,. \qquad (7.3.4)$$

З урахуванням нульової гіпотези можна подати (7.3.3) у вигляді

$$t = \frac{(\hat{\beta}_3 - \hat{\beta}_4) - (\beta_3 - \beta_4)}{\sqrt{D(\hat{\beta}_3) + D(\hat{\beta}_4) - 2\,\text{cov}(\hat{\beta}_3, \hat{\beta}_4)}}\,. \qquad (7.3.5)$$

Процедура використання *t*-статистики за (7.3.5) нічим не відрізняється від уже нам відомої. Якщо визначена за (7.3.5) величина *t* перевищує критичне значення для вибраного рівня значущості і відповідної кількості степенів вільності, то ми відхиляємо нульову гіпотезу; в протилежному випадку гіпотеза не відхиляється. Можна застосувати після визначення *t* за (7.3.5) і підхід на підставі p-величини.

Розглянемо приклад, до якого ми зверталися раніше, кубічної функції вартості:

$$\hat{Y}_i = 141{,}7667 + 63{,}4777 X_i - 12{,}9615 X_i^2 + 0{,}9396 X_i^3$$
$$\text{Se}=(6{,}3753) \quad (4{,}7786) \quad (0{,}9857) \quad (0{,}0591)$$
$$\text{cov}(\hat{\beta}_3, \hat{\beta}_4) = -0{,}0576, R^2=0{,}9983,$$

де *Y* – загальні витрати; *X* – обсяг продукції, що випускається.

Припустимо, ми хочемо перевірити гіпотезу про те, що коефіцієнти при X_2 і X_3 у функції вартості збігаються, тобто $\beta_3 = \beta_4$ або $\beta_3 - \beta_4 = 0$. За наведеними даними підрахуємо значення *t*:

$$t = -13.3130.$$

Легко перевірити, що для DF=6 отримане значення *t* перевершує критичне значення навіть при α=0,002 (або 0,2%). Отже, ми можемо відхиляти гіпотезу про рівність у кубічній функції вартості коефіцієнтів при X_2 і X_3.

7.4. Перевірка лінійних обмежень

Бувають випадки, коли економічна теорія припускає, що коефіцієнти в рівнянні регресії задовольняють деяким, лінійним обмеженням. Розглянемо, наприклад, продуктивну функцію Коба–Дугласа

$$Y_i = \beta_1 X_{2i}^{\beta_2} X_{3i}^{\beta_3} e^{u_i}, \tag{7.4.1}$$

де Y_i – вироблена продукція; X_2 – трудовитрати; X_3 – капіталозатрати; u_i – стохастичний збурюючий складова. Після переходу до логарифмічних змінних, одержуємо

$$\ln Y_i = \beta_0 + \beta_2 \ln X_{2i} + \beta_3 \ln X_{3i} + u_i, \tag{7.4.2}$$

де $\beta_0 = \ln \beta_1$.

Економічна теорія припускає, що

$$\beta_2 + \beta_3 = 1. \tag{7.4.3}$$

Рівність (7.4.3) являє собою приклад лінійного обмеження.

Для перевірки постійного зростання доходу, тобто справедливості обмеження (7.4.3) можна застосовувати два підходи.

Перший підхід ґрунтується на *t*-тесті. При цьому спочатку проводиться оцінка рівняння (7.4.2) без урахування обмеження (7.4.3). Називатимемо цю регресію регресією без обмежень. Отримавши оцінки $\hat{\beta}_2$ і $\hat{\beta}_3$ за МНК, проведемо перевірку виконання гіпотези (7.4.3) на підставі *t*-тесту (див. 7.3.5):

$$t = \frac{(\hat{\beta}_3 - \hat{\beta}_4) - 1}{\sqrt{D(\hat{\beta}_3) + D(\hat{\beta}_4) - 2\operatorname{cov}(\hat{\beta}_3, \hat{\beta}_4)}}. \tag{7.4.4}$$

Згідно з процедурою перевірки гіпотези порівняємо (7.4.4) з критичним значенням *t*-розподілу для (n–k) степенів вільності (у нашому випадку n–3) і вибраного рівня значущості α, якщо t з (7.4.4) перевершує критичне значення, то гіпотеза відкидається, у протилежному випадку – не відкидається.

Описана процедура має ту особливість, що спочатку проводиться регресія, а перевірка лінійного обмеження (7.4.3) здійснюється на підставі результатів регресії без обмежень. Можливий інший підхід, коли лінійне обмеження в (7.4.3) враховується із самого початку. У нашому випадку подамо (7.4.3) у вигляді

$$\beta_2 = 1 - \beta_3 \tag{7.4.5}$$

або

$$\beta_3 = 1 - \beta_2. \tag{7.4.6}$$

Застосовуючи одне з цих рівнянь, можна виключити один з коефіцієнтів регресії в рівнянні (7.4.2). Якщо застосувати (7.4.5), то можна подати виробничу функцію (7.4.2) у вигляді

$$\ln Y_i = \beta_0 + (1-\beta_3)\ln X_{2i} + \beta_3 \ln X_{3i} + u_i =$$
$$= \beta_0 + \ln X_{2i} + \beta_3(\ln X_{3i} - \ln X_{2i}) + u_i$$

або

$$\ln Y_i - \ln X_{2i} = \beta_0 + \beta_3(\ln X_{3i} - \ln X_{2i}) + u_i; \qquad (7.4.7)$$

$$\ln\left(\frac{Y_i}{X_{2i}}\right) = \beta_0 + \beta_3 \ln\left(\frac{X_{3i}}{X_{2i}}\right) + u_i, \qquad (7.4.8)$$

де $\frac{Y_i}{X_{2i}}$ – відношення обсягу продукції до трудовитрат, а $\frac{X_{3i}}{X_{2i}}$ – коефіцієнт відношення капіталу до трудовитрат.

Звернемо увагу на те, як змінилося початкове рівняння (7.4.2). Якщо ми проведемо оцінку β_3 з (7.4.8), то ми легко можемо знайти β_2 за рівністю (7.4.5). Зайве говорити, що подібна процедура гарантує дорівнюваність одиниці суми коефіцієнтів регресії. Описана процедура називається регресією з обмеженнями. Цей метод може бути поширений на моделі, що містять будь-яку кількість пояснювальних змінних і більш ніж одну умову типу (7.4.3).

Як ми можемо порівняти результати регресії без обмежень з обмеженими? Іншими словами, як визначити чи виконується лінійне обмеження (7.4.3)? На це питання можна відповісти після проведення *F*-тесту.

Введемо деякі позначення:

- $\sum \hat{u}_{UR}^2$ – сума квадратів залишків регресії (7.4.2) без обмежень;
- $\sum \hat{u}_R^2$ – сума квадратів залишків регресії з обмеженнями (7.4.8);
- m – кількість лінійних обмежень (у нашому випадку m=1);
- n – кількість спостережень.

Обчислимо

$$F = \frac{\left(\sum \hat{u}_R^2 - \sum \hat{u}_{UR}^2\right)/m}{\sum \hat{u}_{UR}^2/(n-k)}. \qquad (7.4.9)$$

Можна показати, що обчислена величина підкоряється закону *F*-розподілу з m і (n–k) степенями вільності.

Наведений вище *F*-тест можна виразити і в R^2. У такому випадку маємо вираз

$$F = \frac{\left(R_{UR}^2 - R_R^2\right)/m}{\left(1 - R_{UR}^2\right)/(n-k)}. \qquad (7.4.10)$$

Тут R_{UR}^2 і R_R^2 позначають коефіцієнти детермінації регресії без обмежень (7.4.2) і з обмеженнями (7.4.8) відповідно. Побіжно зазначимо, що

$$R^2_{UR} \geq R^2_R,\ \sum \hat{u}^2_{URi} \leq \sum \hat{u}^2_{Ri}\,.$$

Нагадаємо також, що оскільки в регресії без обмежень і з обмеженнями залежні змінні не збігаються, то не можна безпосередньо порівнювати коефіцієнти детермінації R^2_{UR} і R^2_R.

Приклад. Виробнича функція Коба–Дугласа для агросектору Тайваню в 1958–1972 рр.

Проілюструємо викладений вище підхід прикладом виробничої функції для агросектора Тайваню в 1958–1972 рр. (див. табл. 5.3).

Результати регресії за моделлю без обмежень подані формулами (7.10.4). Тепер припустимо, що ми накладаємо обмеження $\beta_2 + \beta_3 = 1$ і застосовуємо модель (7.4.8). У результаті одержуємо

$$\ln\left(\frac{\hat{Y}_i}{X_{2i}}\right) = 1{,}7086 + 0{,}61298\ln\left(\frac{X_{3i}}{X_{2i}}\right) \qquad (7.4.11)$$

$$(0{,}4159)\ \ (0{,}0933)$$

$$R^2 = 0{,}7685, \qquad \bar{R}^2 = 0{,}7507\,.$$

Відзначимо, що величини R^2 в (7.10.4) і (7.4.11) не можна порівнювати, оскільки праві частини рівнянь мають різний вигляд. Для їх порівняння потрібно виконати процедуру, описану в розд. 5 . Якщо застосувати даний прийом, то для моделі (7.4.13) ми одержуємо R^2=0,8489, що виявляється менше, ніж R^2=0,8890 в моделі (7.12.4) без обмежень.

Отже, з (7.10.4) ми отримали $R^2_{UR} = 0{,}8890$, а з (7.4.13) – $R^2_R = 0{,}8489$. Тепер ми можемо застосувати *F*-тест (7.4.10):

$$F = \frac{\left(R^2_{UR} - R^2_R\right)/m}{\left(1 - R^2_{UR}\right)/(n-k)} = 4{,}3587\,. \qquad (7.4.12)$$

Дана величина *F* підкоряється закону *F*-розподілу з 1 і 12 степенями вільності. Із таблиць розподілу Фішера знаходимо, що $F_{0.05}(1,\ 12)$=4,75, але $F_{0.1}(1,\ 12)$=3,18. Таким чином, значення *F*=4,3587 не є значущим при α=0,05, але є таким при α=0,1. Якщо при дослідженні ми встановимо α=0,05, то гіпотезу $\beta_2 + \beta_3 = 1$ ми відхилити не зможемо, а це означає, що період дослідження $\hat{\beta}_2 + \hat{\beta}_3 = 1{,}9887$ не відрізняється значущо від 1. Цей приклад показує, наскільки важливо проводити формальну перевірку гіпотези, а не покладатися тільки на величину оцінених коефіцієнтів регресії. Даний випадок також нагадує, що ми повинні встановлювати рівень значущості до початку перевірки гіпотези, а не після її проведення. Як згадувалося раніше, доцільно наводити значення p-величини. У даному прикладі *p*=0,0588. Звідси бачимо, що *F*=4,35*87* значуща для $\alpha \approx 0{,}06$.

Узагальнений F-тест

Формули (7.4.9), (7.4.10) представляють загальний метод перевірки гіпотез щодо параметрів моделі регресії з k змінними

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i . \qquad (7.4.13)$$

F-тест (7.2.16) або t-тест (8.3.3) є не що інше, як спеціальний випадок (7.4.10). Наприклад, такі гіпотези, як

$$H_0 : \beta_2 = \beta_3 ; \qquad (7.4.14)$$

$$H_0 : \beta_3 + \beta_4 + \beta_5 = 3 , \qquad (7.4.15)$$

що накладають декілька лінійних зв'язків на параметри моделі з k параметрами, або гіпотеза

$$H_0 : \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0 , \qquad (7.4.16)$$

стверджуюча, що деякі регресори, відсутні в моделі, можуть бути перевірені за допомогою F-тесту (7.4.10).

Загальна стратегія застосування F-тесту така. Є велика модель, модель без зв'язків (7.4.13), і є менша модель, модель зі зв'язками, яка виходить з більшої шляхом виключення з неї декількох змінних (випадок (7.4.16), або накладанням декількох зв'язків на коефіцієнти великої моделі (випадки (7.4.14) і (7.4.15)).

Виконуємо регресію моделей без обмеження і з обмеженнями й одержуємо коефіцієнти детермінації R^2_{UR} і R^2_R відповідно. Кількість степенів вільності в моделі без обмежень (n–k), а в моделі з обмеженнями m – кількість лінійних обмежень,1 у (7.4.16) або (7.4.17) і 4 у (7.4.18). Потім ми за (7.4.10) підраховуємо F і застосовуємо правило: якщо підраховане F більше $F_\alpha (m,\ n-k)$, яке є критичним F для вибраного рівня значущості, ми відкидаємо нульову гіпотезу, у протилежному випадку – не відкидаємо.

Приклад. Попит на курятину в США, 1960–1982.

Розглянемо таку модель попиту на курятину:

$$\ln Y_t = \beta_1 + \beta_2 \ln X_{2t} + \beta_3 \ln X_{3t} + \beta_4 \ln X_{4t} + \beta_5 \ln X_{5t} + u_t ,$$

де Y – попит на курятину на душу населення; X_2 – реальний дохід, на душу населення, X_3 – роздрібна ціна на курятину за фунт, X_4 – роздрібна ціна на свинину за фунт, X_5 – роздрібна ціна на яловичину за фунт.

У цій моделі β_1, β_2, β_3, β_4 і β_5 є еластичностями за доходом, власною ціною, крос-ціною (свинина) і крос-ціною (яловичина) відповідно. Згідно з економічною теорією

$\beta_2 > 0$;

$\beta_3 < 0$;

$\beta_4 > 0$, якщо курятина й свинина є конкуруючими продуктами;

$\beta_4 < 0$, якщо курятина й свинина є взаємодоповнюючими продуктами;

$\beta_4 = 0$, якщо курятина й свинина є незв'язаними продуктами;

$\beta_5 > 0$, якщо курятина й яловичина є конкуруючими продуктами;

$\beta_5 < 0$, якщо курятина й яловичина є взаємодоповнюючими продуктами;

$\beta_5 = 0$, якщо курятина й яловичина є незв'язаними продуктами.

Припустимо, хтось вважає, що курятина й свинина, курятина й яловичина є незв'язаними продуктами в тому значенні, що на попит курятини не впливають ціни на свинину й яловичину. У прийнятих позначеннях це означає

$$H_0 : \beta_4 = \beta_5 = 0. \tag{7.4.17}$$

З урахуванням цього регресія з обмеженнями набуває вигляду

$$\ln Y_t = \beta_1 + \beta_2 \ln X_{2t} + \beta_3 \ln X_{3t} + u_t. \tag{7.4.18}$$

Зрозуміло, що за нашою термінологію рівняння (7.4.19) є регресія без обмежень.

Якщо скористатися конкретними даними, то можна отримати такі результати.

Регресія без обмежень:

$$\ln \hat{Y}_t = 2{,}1898 + 0{,}3425 \ln X_{2t} - 0{,}5046 \ln X_{3t} + 0{,}1485 \ln X_{4t} + 0{,}0911 \ln X_{3t}$$
$$\textbf{(0,1557)} \quad \textbf{(0,0833)} \qquad \textbf{(0,1109)} \qquad \textbf{(0,0997)} \qquad \textbf{(0,1007)} \tag{7.4.19}$$
$$R^2_{UR} = 0{,}9823.$$

Регресія з обмеженнями:

$$\ln \hat{Y}_t = 2{,}0328 + 0{,}4515 \ln X_{2t} - 0{,}3772 \ln X_{3t} \frac{n!}{r!(n-r)!}$$
$$(0{,}1162) \quad (0{,}0247) \qquad (0{,}0635) \tag{7.4.20}$$
$$R^2_{UR} = 0{,}9801.$$

У дужках указані значення стандартних похибок коефіцієнтів регресії. Зазначимо, що R^2 в (7.4.19) і (7.4.20) можна порівнювати, оскільки залежна змінна в обох моделях однакова.

Для перевірки гіпотези (7.4.17) необхідно підрахувати значення F за формулою (7.4.10):

$$F = \frac{\left(R^2_{UR} - R^2_R\right)/m}{\left(1 - R^2_{UR}\right)/(n-k)}.$$

У нашому випадку m=2, оскільки є два обмеження. Кількість степенів вільності знаменника дорівнює 18. З урахуванням цього, одержуємо

$$F = 1{,}1224. \tag{7.4.21}$$

При 5%-му рівні значущості це значення не виявляється статистично значимим, оскільки $F_{0.05}(2, 18)=3,55$. Відповідне значення $p=0,3472$. Отже, ми не маємо підстав відхиляти нульову гіпотезу – попит на курятину не залежить від ціни на свинину та яловичину.

Відзначимо, що отримана функція попиту на курятину відповідає нашим апріорним очікуванням, оскільки еластичність за доходом позитивна. Оцінена еластичність за ціною виявляється за абсолютною величиною статистично менше 1, тобто попит на курятину нееластичний за ціною. Еластичність за доходом, хоча й позитивна, так само виявляється статистично менше 1.

7.5. Перевірка структурної стабільності моделей регресії

У табл. 7.7 наведені дані про особисті заощадження й особистий дохід у Великобританії в період 1946–1963 рр.

Таблиця 7.7

Рік	Заощадження, млрд дол.	Дохід, млрд дол.
1946	0,36	8,8
1947	0,21	9,4
1948	0,08	10,0
1949	0,20	10,6
1950	0,10	11,0
1951	0,12	11,9
1952	0,41	12,7
1953	0,50	13,5
1954	0,43	14,3
1955	0,59	15,5
1956	0,90	16,7
1957	0,95	17,7
1958	0,82	18,6
1959	1,04	19,7
1960	1,53	21,1
1961	1,94	22,8
1962	1,75	23,9
1963	1,99	25,2

Припустимо, ми хочемо визначити, як особисті заощадження пов’язані з особистим доходом, тобто ми хотіли б отримати оцінку функції заощаджень. Якщо звернутися до даних з табл. 7.7, то можна відзначити, що характер заощаджень по відношенню до доходу в період 1946–1954 рр., післявоєнний (реконструктивний) період, відрізняється від періоду 1955–1963 рр., який називатимемо постреконструктивним. Інакше кажучи, функція заощаджень зазнає структурних змін між цими двома періодами, у тому розумінні, що її параметри змінюються.

Для перевірки цього факту припустимо, що функція заощаджень для двох періодів має такий вигляд:

- *реконструктивний період*:

$$Y_t = \alpha_1 + \alpha_2 X_t + u_{1t}, \; t=1, 2,\ldots n_1; \tag{7.5.1}$$

- *постреконструктивний*:

$$Y_t = \beta_1 + \beta_2 X_t + u_{2t}, \; t=1, 2,\ldots n_2, \tag{7.5.2}$$

де Y – особисті заощадження; X – особистий дохід; u_1 і u_2 – збурюючі складові, а n_1, n_2 – кількість спостережень у перший і другий періоди відповідно. Кількість спостережень в періодах може збігатися або ні.

Структурні зміни функції заощаджень означають, що параметри регресій можуть відрізнятися. Зрозуміло, якщо не спостерігаються структурні зміни, то всі дані можна об'єднати й оцінити на їх підставі одну функцію заощаджень

$$Y_t = \lambda_1 + \lambda_2 X_t + u_t. \tag{7.5.3}$$

Для з'ясування, чи існують насправді структурні зміни функції заощаджень застосовують тест Чоу.

Цей тест ґрунтується на таких двох припущеннях:

- $u_{1t} \sim N(0, \sigma^2)$, $u_{2t} \sim N(0, \sigma^2)$, тобто обидва стохастичні складоваи розподілені нормально і мають однакові дисперсії;
- u_{1t} і u_{2t} розподілені незалежно.

Тест Чоу складається з чотирьох етапів:

1. Об'єднуємо спостереження n_1 і n_2, знаходимо оцінку (7.5.3) і підраховуємо суму квадратів залишків RSS, назвемо її s_1 з $(n_1+n_2–k)$ степенями вільності. Тут k – кількість оцінених параметрів у рівнянні регресії (у нашому випадку2);
2. Виконуємо регресії (7.5.1) і (8.82) і одержуємо їх RSS, скажімо, s_2 і s_3 з $(n_1–k)$ і $(n_2–k)$ кількістю степенів вільності, відповідно. Складаємо ці RSS і знаходимо $s_4=s_2+s_3$ з $(n_1+n_2–2k)$ степенями вільності;
3. Одержуємо $s_5=s_1+s_4$.

Можна показати, що при виконанні згаданих припущень

$$F = \frac{S_5 / k}{S_4 /(n_1 + n_2 - 2k)} \tag{7.5.4}$$

розподілено за законом F-розподілу з k і $(n_1+n_2–2k)$ степенями вільності. Якщо значення F більше критичного значення для вибраного рівня значення α, то ми відхиляємо гіпотезу про те, що регресії (7.5.1) і (7.5.2) збігаються, тобто гіпотезу про структурну стабільність. Можна застосувати альтернативний підхід. Якщо відповідна F p-величина мала, то ми відхиляємо гіпотезу про структурну стабільність.

Звернемося до нашого прикладу, врахувавши, що $n_1=n_2=9$.

$$\hat{Y}_t = -1,0821 + 0,1178 X_t$$
$$(0,1452)\ \ (0,0088) \qquad R^2=0,9185$$
$$t = (-7,4548)\ \ (13,4316) \qquad s_1=0,5722 \quad DF=16.$$

Реконструктивний період:

$$\hat{Y}_t = -0,2622 + 0,0470 X_t$$
$$(0,3054)\ \ (0,0266) \qquad R^2=0,3092$$
$$t = (-0,8719)\ \ (1,7700) \qquad s_2=0,1396 \quad DF=7.$$

Постреконструктивний період:

$$\hat{Y}_t = -1,7502 + 0,1504 X_t$$
$$(0,3576)\ \ (0,0175) \qquad R^2=0,9131$$
$$t = (-4,8948)\ \ (8,5749) \qquad s_3=0,1931 \quad DF=7.$$
$$s_4=0,3327, \quad s_5=0,2395, \quad F=5,04.$$

Якщо вибрати 5%-й рівень значущості, то можна визначити, що $F_{2,14}=3,74$. Оскільки значення величини $F=5,04$ більше критичного значення F, то ми маємо підставу відхилити гіпотезу про структурну стабільність функції заощаджень за два періоди. Обчислена величина $p=0,0224$ виявляється достатньо малою.

7.6. Перевірка функціонального виду регресії. Вибір між лінійною моделлю регресії і лінійно-логарифмічною моделлю

Вибір між лінійною моделлю регресії і лінійно-логарифмічною моделлю є постійним питанням емпіричного аналізу. Для вибору між цими моделями можна застосувати MWD-тест.

Розглянемо дві такі гіпотези:

H_0: лінійна модель: Y – лінійна функція регресорів X.

H_1: лінійна логарифмічна модель: $\ln Y$ – лінійна функція регресорsв $\ln X$.

Етапи MWD-тестування:

1. Проводимо оцінку лінійної моделі й отримаємо оцінену величину Y, назвемо її Y_f, тобто $\hat{Y}$.
2. Проводимо оцінку лінійної логарифмічної моделі й одержуємо оцінену величину $\ln Y$, назвемо її $\ln_f$ (тобто $\ln \hat{Y}$).
3. Одержуємо $Z_1=(\ln Y_f - \ln_f)$.
4. Виконуємо регресію Y за X і Z_1, отриманими на етапі 3. Відхиляємо гіпотезу H_0, якщо коефіцієнт при Z_1 статистично значимий за звичайним t-тестом.
5. Одержуємо $Z_2=\exp(\ln Y_f - Y_f)$.
6. Виконуємо регресію Y за $\ln X$ і Z_2. Відхиляємо гіпотезу H_1, якщо коефіцієнт при Z_2 статистично значущий за звичайним t-тестом.

8. ПРОГНОЗУВАННЯ В РАЗІ МНОЖИННОЇ РЕГРЕСІЇ

У розд.5 ми показали, як оцінена модель двовимірної регресії може бути застосована для середнього прогнозу, тобто прогнозування для функції популяції регресії (PRF), а також для індивідуального прогнозу, тобто прогнозування індивідуального значення величини Y для фіксованого значення регресора X_0.

Оцінене рівняння множинної регресії також може бути використане для прогнозу. Його методика є простим поширенням двовимірного випадку й ґрунтується на тих же формулах за винятком формул для оцінок дисперсії і стандартних похибок прогнозованих величин.

Проілюструємо це на прикладі середнього й індивідуального прогнозів для моделі особистих витрат на споживання в США в 1956–1970 рр.:

$$\begin{gathered}\hat{Y}_i = 53{,}103 + 0{,}7266X_{2i} + 2{,}7363X_{3i} \\ (13{,}0261)\ \ (0{,}0487)\quad (0{,}8486) \\ R^2 = 0{,}9988,\end{gathered} \tag{8.1}$$

де Y – особисті витрати на споживання; X_2 – особистий дохід; X_3 – змінна тренда.

Як нам відомо $\hat{Y}_i$ є оцінка величини $E(Y \mid X_2, X_3)$ істинного середнього значення Y для заданих значень X_2 і X_3.

Припустимо, що для 1971 р. маємо X_2=567 і X_3=16. Підставляючи ці значення в (8.1), одержуємо

$$E(\hat{Y} \mid X_2 = 567,\ X_3 = 16) = 508{,}9297. \tag{8.2}$$

Отже, середні особисті витрати на споживання для 1971 р. складають приблизно 509 млрд дол. Із причин, указаних раніше, 509 млрд дол. також є індивідуальним прогнозом для 1971 р., тобто Y_{1971}. Проте дисперсії для середнього $\hat{Y}_{1971}$ й індивідуального Y_{1971} прогнозів різні. Для їх обчислення найзручніше використовувати матричний апарат. Можна показати, що

$$\text{var}(\hat{Y}_{1971} \mid X_2,\ X_3) = 3.6580 \text{ і } \text{se}(\hat{Y}_{1971} \mid X_2,\ X_3) = 1{,}9126; \tag{8.3}$$

$$\text{var}(Y_{1971} \mid X_2,\ X_3) = 10.0887 \text{ і } \text{se}(Y_{1971} \mid X_2,\ X_3) = 3{,}1763. \tag{8.4}$$

При виконанні припущень класичної моделі регресії ми тепер можемо побудувати 10(1–α)-довірчий інтервал для середнього прогнозу:

$$\hat{Y}_{1971} - t_{\alpha/2}\,\text{se}\left(\hat{Y}_{1971}\right) \le E(Y_{1971}) \le \hat{Y}_{1971} + t_{\alpha/2}\,\text{se}\left(\hat{Y}_{1971}\right).$$

Зрозуміло, що подібну процедуру можна виконати й для будь-яких інших значень X_2 і X_3.

100(1–α)-довірчий інтервал для індивідуального прогнозу Y_{1971} складає

$$\hat{Y}_{1971} - t_{\alpha/2}\,\text{se}\left(Y_{1971}\right) \le Y_{1971} \le \hat{Y}_{1971} + t_{\alpha/2}\,\text{se}\left(Y_{1971}\right).$$

Для нашого прикладу ми одержуємо довірчий інтервал для середнього прогнозу

$$504,7518 \le E(Y_{1971}) \le 13,0868.$$

Довірчий інтервал для індивідуального прогнозу має вигляд

$$501,9988 \le Y_{1971} \le 515,8412.$$

Кількість степенів вільності для $t_{\alpha/2}$ є (n–3) для моделі з трьома змінними і (n–k) для моделі з k змінними.

9. МНОЖИННА РЕГРЕСІЯ. МАТРИЧНИЙ МЕТОД

9.1. Лінійна модель регресії з k змінними

Поширимо дво- і тривимірні моделі лінійної регресії на випадок моделі з k змінними у функції PRF, що містить залежну змінну Y і (k–1) пояснювальну змінну. Відповідну функцію PRF можна подати у вигляді

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + u_i, \qquad (9.1.1)$$

де β_1 – коефіцієнт, що визначає значення Y при нульових значеннях вхідних змінних; β_2, β_3,… β_k – частинні кутові коефіцієнти, u – стохастичний збурюючий складова, i – спостереження, n – розмір вибірки. Рівняння (9.1.1) можна інтерпретувати звичайним способом, а саме: воно дає середнє або очікуване значення величини Y при фіксованих значеннях X_2, X_3,…X_k, тобто $E(Y \mid X_{2i}, X_{3i},...X_{ki})$.

Це рівняння є скороченим записом таких рівнянь:

$$\begin{gathered} Y_1 = \beta_1 + \beta_2 X_{21} + \beta_3 X_{31} + ... + \beta_k X_{k1} + u_1, \\ Y_2 = \beta_1 + \beta_2 X_{22} + \beta_3 X_{32} + ... + \beta_k X_{k2} + u_2, \\ ..... \\ Y_n = \beta_1 + \beta_2 X_{2n} + \beta_3 X_{3n} + ... + \beta_k X_{kn} + u_n. \end{gathered} \qquad (9.1.2)$$

Подамо цю систему рівнянь у матричному вигляді

$$\begin{bmatrix} Y_1 \\ Y_2 \\ : \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{21} & X_{31} & ... & X_{k1} \\ 1 & X_{22} & X_{23} & ... & X_{k2} \\ : & : & : & . & : \\ 1 & X_{2n} & X_{3n} & ... & X_{kn} \end{bmatrix} . \begin{bmatrix} \beta_1 \\ \beta_2 \\ : \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ : \\ u_n \end{bmatrix} \qquad (9.1.3)$$

або

$$\mathbf{Y}_{n\times 1} = \mathbf{X}_{n\times k}\boldsymbol{\beta}_{k\times 1} + \mathbf{u}_{n\times 1}, \qquad (9.1.4)$$

де $\mathbf{Y}$ – вектор-стовпець спостережень залежної змінної Y, розміром $n\times 1$; $\mathbf{X}$ – матриця спостережень розміром $n\times k$, перший стовпець якої складається з одиниць, а

наступні – дані змінних від X_2 до X_k; **β** – вектор-стовпець незалежних параметрів $\beta_1,\ \beta_2,\ldots \beta_k$ розміром $k\times1$; **u** – вектор-стовпець n збурень u_i розміром $n\times1$.

У випадках, коли не виникає плутанини щодо розмірів або порядків матриці **X** і векторів **Y, β** й **u** рівняння (9.1.4) може бути записане в простому вигляді

$$\mathbf{Y} = \mathbf{X\beta} + \mathbf{u}. \qquad (9.1.5)$$

Для ілюстрації матричного методу розглянемо модель «доходи-витрати на споживання» з двома змінними, вивчену нами раніше (розд. 3):

$$Y_i = \beta_1 + \beta_2 X_i + u_i,$$

де Y_i – витрати на споживання, а X_i – дохід. Використовуючи дані табл. 3.2, отримаємо

$$\begin{bmatrix}70\\65\\90\\95\\110\\115\\120\\140\\155\\150\end{bmatrix} = \begin{bmatrix}1 & 80\\1 & 100\\1 & 120\\1 & 140\\1 & 160\\1 & 180\\1 & 200\\1 & 220\\1 & 240\\1 & 260\end{bmatrix}\begin{bmatrix}\beta_1\\\beta_2\end{bmatrix} + \begin{bmatrix}u_1\\u_2\\u_3\\u_4\\u_5\\u_6\\u_7\\u_8\\u_9\\u_{10}\end{bmatrix}. \qquad (9.1.6)$$

Так само, як і у випадках моделей з двома або трьома змінними, нашою метою є проведення оцінювання регресії (9.1.1), складання висновків на підставі наявних даних. Для оцінювання ми можемо, як і раніше, використовувати МНК.

9.2. Припущення класичної лінійної моделі регресії в матричній формі

Припущення, на яких базується класична лінійна модель регресії, наведені в табл. 9.1. Вони подані як у скалярних, так і в матричних позначеннях.

Таблиця 9.1

Припущення класичної лінійної моделі регресії

Скалярне позначення	Матричне позначення
1. $E(u_i) = 0,\ \forall i$	$E(\mathbf{U})=0$, де **U** і 0 є $n\times1$ вектори-стовпці, 0 – нульовий вектор
2. $E(u_i u_j) = \begin{cases}0, & i \neq j\\ \sigma^2, & i = j\end{cases}$	$E(\mathbf{uu}') = \sigma^2\mathbf{I}$, де **I** – $n\times n$ одинична матриця

3. X_2, X_3,…, X_k нестохастичні й фіксовані	$n \times k$ матриця **X** є нестохастичною, тобто вона складається із сукупності фіксованих чисел
4. Не існує точного лінійного зв'язку між змінними X, тобто немає мультиколінеарності	Ранг матриці **X** дорівнює k, де k – кількість стовпців в **X**, при чому $k<n$
5. При перевірці гіпотез $u_i \sim N(0, \sigma^2)$	Вектор **u** має багатовимірний нормальний розподіл, тобто $\mathbf{u} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$

Припущення 1 (табл. 9.1) означає, що математичне сподівання збурюючого вектора **u**, тобто кожної його компоненти, дорівнює нулю. Більш точно $E(\mathbf{u}) = 0$ означає

$$E(\mathbf{u}) = E\begin{pmatrix} u_1 \\ u_2 \\ : \\ u_n \end{pmatrix} = \begin{bmatrix} E(u_1) \\ E(u_2) \\ : \\ E(u_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ : \\ 0 \end{bmatrix}. \qquad (9.2.1)$$

Припущення 2 являє собою компактний вираз двох припущень, зображених у скалярному вигляді рівняннями (3.2.5) і (3.2.2). Щоб переконатися в цьому, можна записати

$$E(\mathbf{u}\mathbf{u}') = E\begin{pmatrix} u_1 \\ u_2 \\ : \\ u_n \end{pmatrix}\begin{bmatrix} u_1 & u_2 & \dots & u_n \end{bmatrix},$$

де $\mathbf{u}'$ – транспонований вектор-стовпець **u**, тобто вектор-рядок.

Виконавши перемножування матриць, отримаємо

$$E(\mathbf{u}\mathbf{u}') = E\begin{pmatrix} u_1^2 & u_1u_2 & \dots & u_1u_n \\ u_2u_1 & u_2^2 & \dots & u_2u_n \\ \dots & \dots & . & \dots \\ u_nu_1 & u_nu_2 & \dots & u_n^2 \end{pmatrix}$$

або

$$E(\mathbf{u}\mathbf{u}') = \begin{bmatrix} E(u_1^2) & E(u_1u_2) & \dots & E(u_1u_n) \\ E(u_2u_1) & E(u_2^2) & \dots & E(u_2u_n) \\ \dots & \dots & . & \dots \\ E(u_nu_1) & E(u_nu_2) & \dots & E(u_n^2) \end{bmatrix}. \qquad (9.2.2)$$

Враховуючи гіпотезу про гомоскедастичність і відсутність серійної кореляції, можна подати(9.2.2) у вигляді

$$E(\mathbf{u}\mathbf{u}') = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \dots & \dots & . & \dots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix} = \sigma^2\mathbf{I}, \qquad (9.2.3)$$

де $\mathbf{I}$ – $n\times n$ одинична матриця.

Матриця (9.2.2) називається матрицею варіації збурень u_i; елементи головної діагоналі цієї матриці позначають дисперсії, а решта елементів – коваріації. Зазначимо, що ця матриця симетрична.

Припущення 3 стверджує, що матриця $\mathbf{X}$ нестохастична, тобто її елементи – фіксовані числа.

Припущення 4 стверджує, що ранг матриці дорівнює *k*, тобто дорівнює кількості стовпців матриці. Це означає, що стовпці матриці лінійно незалежні, тобто не існує точного лінійного зв'язку між змінними *X*. Іншими словами, відсутня мультиколінеарність. У скалярних позначеннях це означає, що не існує множини чисел λ_1, λ_2,… λ_k, з яких не всі дорівнюють нулю, тобто

$$\lambda_1X_{1i} + \lambda_2X_{2i} + \dots + \lambda_kX_{ki} = 0, \qquad (9.2.4)$$

де $X_{i1} = 1$ для всіх *i*. У матричних позначеннях (9.2.4) може бути подане у вигляді

$$\boldsymbol{\lambda}'\mathbf{X} = 0, \qquad (9.2.5)$$

де $\boldsymbol{\lambda}'$ – $1\times k$-вектор-рядок, а $\mathbf{X}$ – $k\times 1$-вектор-стовпець.

Якщо лінійне співвідношення вигляду (9.2.4) існує, то говорять, що змінні колінеарні. Якщо ж ця рівність можлива тільки при $\lambda_1 = \lambda_2 = \dots = \lambda_k = 0$, то говорять, що змінні *X* лінійно незалежні.

9.3. Оцінювання за МНК

Щоб отримати оцінку вектора $\boldsymbol{\beta}$ запишемо функцію SRF (вибіркову функцію регресії) з *k* змінними в матричному вигляді:

$$\mathbf{Y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\mathbf{u}}. \qquad (9.3.1)$$

Так само, як і у разі дво- й тривимірних моделей, МНК для *k*-вимірної моделі полягає в мінімізації

$$\hat{\mathbf{u}}'\hat{\mathbf{u}} = \sum \hat{u}_i^2 \,. \qquad (9.3.2)$$

Із (9.3.1) ми одержуємо

$$\hat{\mathbf{u}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} \,. \qquad (9.3.3)$$

Отже,

$$\hat{\mathbf{u}}'\hat{\mathbf{u}} = \mathbf{Y}'\mathbf{Y} - 2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \,. \qquad (9.3.4)$$

Тут ми скористалися властивостями транспонування матриць, а саме $(\mathbf{X}\hat{\boldsymbol{\beta}})' = \hat{\boldsymbol{\beta}}'\mathbf{X}'$. Крім того, оскільки $\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}$ є скаляр, то він не змінюється при транспонуванні $\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} = \mathbf{Y}'\mathbf{X}\hat{\boldsymbol{\beta}}'$.

У скалярних позначеннях МНК полягає в оцінюванні $\beta_1,\ \beta_2, \ldots \beta_k$ таким чином, щоб $\hat{\mathbf{u}}'\hat{\mathbf{u}} = \sum \hat{u}_i^2$ була якомога малою величиною. Це досягається шляхом диференціювання (9.3.4) за $\hat{\beta}_1,\ \hat{\beta}_2, \ldots \hat{\beta}_k$ і прирівнювання частинних похідних до нуля. Ця процедура приводить до системи *k* лінійних алгебраїчних рівнянь з *k* невідомими. Можна показати, що ця система має вигляд

$$\left(\mathbf{X}'\mathbf{X}\right)\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y} \,. \qquad (9.3.5)$$

Відзначимо деякі властивості матриці $\mathbf{X}'\mathbf{X}$:

1) вона дає ряд суми квадратів і попарних добутків змінних *X*. Елементи головної діагоналі являють собою ряд сум квадратів, а елементи зовні головної діагоналі – суму попарних добутків;
2) матриця симетрична, оскільки сума добутків X_{ij} і X_{ji};
3) матриця має порядок *k*×*k*.

У (9.3.5) відомими величинами є $\mathbf{X}'\mathbf{X}$ і $\mathbf{X}'\mathbf{Y}$, а невідомою – $\hat{\boldsymbol{\beta}}$. Розв'язуючи рівняння (9.3.5), знаходимо

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y} \,. \qquad (9.3.6)$$

Рівняння (9.3.6) відображає фундаментальний результат теорії МНК у матричній формі. Воно показує, що оцінка вектора $\hat{\boldsymbol{\beta}}$ може бути проведена за наявними даними.

Ілюстрований приклад

Як ілюстрацію викладеного вище матричного підходу використаємо розглянутий нами раніше приклад "споживання – дохід", дані щодо якого наведені в (9.1.6). Для випадку моделі з двома змінними маємо

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix},\ \mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & \ldots & 1 \\ X_1 & X_2 & X_3 & \ldots & X_n \end{bmatrix} \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ 1 & X_3 \\ \vdots & \vdots \\ 1 & X_3 \end{bmatrix} = \begin{bmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix}.$$

Використовуючи дані, наведені в (9.1.6), одержуємо

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 10 & 1700 \\ 1700 & 322000 \end{bmatrix},\ \mathbf{X}'\mathbf{Y} = \begin{bmatrix} 1110 \\ 205500 \end{bmatrix}.$$

Застосовуючи правила для знаходження оберненої матриці, можна переконатися, що

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} \frac{161}{165} & -\frac{17}{3300} \\ -\frac{17}{3300} & \frac{1}{33000} \end{bmatrix}.$$

Ми бачимо, що результати оцінювання, отримані матричним методом, збігаються з отриманими нами раніше.

Матриця варіації $\hat{\boldsymbol{\beta}}$

Матричний метод дає нам можливість легко отримати не тільки дисперсію (варіацію) для кожного елемента $\hat{\boldsymbol{\beta}}$, але й коваріацію між будь-якими двома елементами, скажімо $\hat{\beta}_i$ і $\hat{\beta}_j$. Значення цих величин необхідні нам для отримання статистичних висновків.

За визначенням матриця варіації $\hat{\boldsymbol{\beta}}$ є

$$\text{var}-\text{cov}(\hat{\boldsymbol{\beta}}) = E\left\{\left[\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})\right]\left[\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})\right]'\right\}.$$

У явному вигляді цю матрицю можна подати так:

$$\text{var}-\text{cov}(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} \text{var}(\beta_1) & \text{cov}(\beta_1, \beta_2) & \ldots & \text{cov}(\beta_1, \beta_k) \\ \text{cov}(\beta_2, \beta_1) & \text{var}(\beta_2) & \ldots & \text{cov}(\beta_2, \beta_k) \\ \vdots & \vdots & . & \vdots \\ \text{cov}(\beta_k, \beta_1) & \text{cov}(\beta_k, \beta_2) & \ldots & \text{var}(\beta_k) \end{bmatrix}. \qquad (9.3.7)$$

Можна показати, що матриця варіації $\hat{\boldsymbol{\beta}}$ може бути отримана за такою формулою:

$$\text{var}-\text{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}, \qquad (9.3.8)$$

де σ^2 – гомоскедастична дисперсія залишків u_i.

У лінійних моделях регресії з двома й трьома змінними незміщена оцінка для σ^2 обчислювалася, відповідно, за формулами $\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-2}$ і $\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-3}$. У разі моделі з *k* змінними формула має вигляд

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-k} = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n-k}, \qquad (9.3.9)$$

де (*n*–*k*) – кількість степенів вільності.

Хоча теоретично $\hat{\mathbf{u}}'\hat{\mathbf{u}}$ може бути підрахована за оцінками залишків, на практиці її зручніше одержувати безпосередньо таким чином.

Пригадавши, що

$$\sum \hat{u}_i^2 = \sum y_i^2 - \hat{\beta}_2 \sum x_i^2,$$

у разі трьох змінних

$$\sum \hat{u}_i^2 = \sum y_i^2 - \hat{\beta}_2 \sum y_i x_{2i} - \hat{\beta}_3 \sum y_i x_{3i},$$

та поширюючи цей принцип на випадок моделі з *k* змінними, одержуємо формулу

$$\sum \hat{u}_i^2 = \sum y_i^2 - \hat{\beta}_2 \sum y_i x_{2i} - \ldots - \hat{\beta}_k \sum y_i x_{ki}, \qquad (9.3.10)$$

або в матричних позначеннях

$$TSS = \sum y_i^2 = \mathbf{y}'\mathbf{y} - n\bar{Y}^2; \qquad (9.3.11)$$

$$ESS = \hat{\beta}_2 \sum y_i x_{2i} + \ldots + \hat{\beta}_k \sum y_i x_{ki} = \hat{\boldsymbol{\beta}}'\mathbf{x}'\mathbf{y} - n\bar{Y}^2, \qquad (9.3.12)$$

де складова $n\bar{Y}^2$ відомий як кореляція для середнього.

Отже,

$$\hat{\mathbf{u}}'\hat{\mathbf{u}} = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{x}'\mathbf{y}. \qquad (9.3.13)$$

Оскільки $\hat{\mathbf{u}}'\hat{\mathbf{u}}$ отримана, $\hat{\sigma}^2$ може бути легко підрахована за (9.3.9), а потім можна підрахувати матрицю варіацій (9.3.8).

Для нашого прикладу $\hat{\mathbf{u}}'\hat{\mathbf{u}} = 337.273$. Отже, $\hat{\sigma}^2 = 42.1591$.

Властивості вектора β

Для випадків моделей із двома й трьома змінними ми знаємо, що оцінки параметрів регресії за МНК є лінійними й незміщеними і в класі лінійних незміщених оцінок вони мають мінімальну дисперсію. Коротше кажучи, оцінки за МНК є якнайкращими лінійними незміщеними оцінками. Ця властивість поширюється й на вектор $\hat{\boldsymbol{\beta}}$, тобто $\hat{\boldsymbol{\beta}}$ – лінійний (кожна з його компонент є лінійна функція від

y, залежної змінної). $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, тобто математичне сподівання кожної компоненти вектора $\hat{\boldsymbol{\beta}}$ дорівнює відповідному елементу істинного вектора $\boldsymbol{\beta}$ і в класі всіх лінійних незміщених оцінок $\boldsymbol{\beta}$ МНК-оцінка $\hat{\boldsymbol{\beta}}$ має мінімальну дисперсію.

9.4. Коефіцієнт детермінації R^2 у матричному позначенні

Коефіцієнт детермінації R^2 визначається так:

$$R^2 = \frac{ESS}{TSS}.$$

У разі двох змінних

$$R^2 = \frac{\hat{\beta}_2 \sum x_i^2}{\sum y_i^2},$$

а в разі трьох змінних

$$R^2 = \frac{\hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i}}{\sum y_i^2}.$$

Узагальнюючи на випадок *k* змінних, отримуємо

$$R^2 = \frac{\hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i} + ... + \hat{\beta}_k \sum y_i x_{ki}}{\sum y_i^2}. \qquad (9.4.1)$$

Рівність (9.4.1) можна переписати як

$$R^2 = \frac{\hat{\boldsymbol{\beta}}\mathbf{x}'\mathbf{y} - n\bar{Y}^2}{\mathbf{y}'\mathbf{y} - n\bar{Y}^2}. \qquad (9.4.2)$$

Для нашого прикладу

$$\hat{\boldsymbol{\beta}}\mathbf{x}'\mathbf{y} = 131762,7272,\ \mathbf{y}'\mathbf{y} = 132100,\ n\bar{Y}^2 = 123210.$$

Підставляючи ці значення в (9.4.2), одержуємо R^2=0,962062.

9.5. Кореляційна матриця

Раніше зупинялися на понятті кореляційних коефіцієнтів нульового порядку r_{12}, r_{13}, r_{23} і на понятті частинних або першого порядку кореляційних коефіцієнтів $r_{12.3}$, $r_{13.2}$, $r_{23.1}$ та їх взаємозв'язку. Для випадку моделі з *k* змінними ми матимемо всього $\frac{k(k-1)}{2}$ кореляційних коефіцієнти нульового порядку. Їх можна подати у вигляді матриці, що називається кореляційною матрицею **R:**

$$\mathbf{R} = \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1k} \\ r_{21} & r_{22} & \dots & r_{2k} \\ \vdots & \vdots & . & \vdots \\ r_{k1} & r_{k2} & \dots & r_{kk} \end{bmatrix} = \begin{bmatrix} 1 & r_{12} & \dots & r_{1k} \\ r_{21} & 1 & \dots & r_{2k} \\ \vdots & \vdots & . & \vdots \\ r_{k1} & r_{k2} & \dots & 1 \end{bmatrix}. \qquad (9.5.1)$$

Тут індекс 1, як і раніше, позначає залежну змінну Y (r_{12} позначає кореляційний коефіцієнт між Y і X_2 і т.д.). При цьому ми використовували той факт, що кореляційний коефіцієнт по відношенню до самого себе дорівнює 1.

Із кореляційної матриці **R** можна отримати кореляційні коефіцієнти першого порядку й більш високих порядків.

9.6. Перевірка гіпотез про індивідуальні коефіцієнти регресії в матричному позначенні

Якщо нашою метою є проведення статистичних висновків, то нам необхідно припустити, що збурення u_i підкоряються деякому закону розподілу. Як ми раніше вже згадували, у регресійному аналізі ми звичайно приймаємо, що кожний u_i підкоряється нормальному закону розподілу з нульовим математичним сподіванням і постійною дисперсією σ^2. У матричних позначеннях ми запишемо

$$\mathbf{u} \sim N(\mathbf{0}, \sigma^2\mathbf{I}). \qquad (9.6.1)$$

Припускаючи нормальність розподілу вектора **u** у разі дво- й тривимірних моделей лінійної регресії, ми знаємо, що отримані за МНК оцінки $\hat{\beta}_i$ також нормально розподілені. Узагальнюючи цей результат на випадок k змінних, можна показати, що

$$\hat{\boldsymbol{\beta}} \sim N\left[\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right],$$

тобто кожен елемент вектора $\hat{\boldsymbol{\beta}}$ розподілений нормально з математичним сподіванням, яке дорівнює відповідному елементу істинного $\boldsymbol{\beta}$, а дисперсія задається добутком σ^2 на відповідний діагональний елемент оберненої матриці $(\mathbf{X}'\mathbf{X})^{-1}$.

Оскільки на практиці σ^2 невідома, то використовують її оцінку $\hat{\sigma}^2$. У такому разі шляхом звичайного переходу до t-розподілу приходимо до висновку про те, що кожний елемент вектора $\boldsymbol{\beta}$ підкоряється закону t-розподілу з (n–k) степенями вільності. У математичних позначеннях

$$t = \frac{\hat{\beta}_i - \beta_i}{\text{se}(\hat{\beta}_i)} \qquad (9.6.2)$$

з (n–k) степенями вільності, де $\hat{\beta}_i$ – будь-яка компонента вектора $\boldsymbol{\beta}$. Отже, t-розподіл може бути застосований для перевірки гіпотез про істинне значення β_i,

а також встановлення для β_i довірчого інтервалу. Сам механізм застосування обговорювався нами раніше.

9.7. Загальна перевірка регресії на значущість. Аналіз дисперсії в матричному позначенні

У розд. 8 була розглянута техніка ANOVA загальної перевірки на значущість оціненої регресії, тобто перевірки нульової гіпотези про те, що істинні кутові коефіцієнти одночасно перетворюються в нуль, і оцінювання додаткового внеску пояснювальної змінної. Цю техніку можна поширити на випадок k змінних. Пригадаємо, що механізм полягає в розкладанні TSS на дві складові: ESS і RSS. Матричні вирази для цих трьох сум квадратів уже наведені в (9.3.10) – (9.3.12). Асоційовані з ними кількості степенів вільності відповідно дорівнюють (n–1), (k–1) і (n–k). За аналогією з табл. 8.2 ми можемо скласти табл. 9.2.

Таблиця 9.2

Матричне формулювання ANOVA-таблиці
для лінійної моделі регресії з k змінними.

Джерело дисперсії	SS	DF	MSS
Унаслідок регресії	$\hat{\boldsymbol{\beta}}\mathbf{x}'\mathbf{y} - n\bar{Y}^2$	k–1	$\dfrac{\hat{\boldsymbol{\beta}}\mathbf{x}'\mathbf{y} - n\bar{Y}^2}{k-1}$
Унаслідок залишків	$\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{x}'\mathbf{y}$	n–k	$\dfrac{\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{x}'\mathbf{y}}{n-k}$
Загальна	$\mathbf{y}'\mathbf{y} - n\bar{Y}^2$	n–1	

Припускаючи, що збурення розподілені нормально й нульова гіпотеза є $\beta_2 = \beta_3 = ... = \beta_k = 0$, як і в розд. 8, можна показати, що

$$F = \frac{\left(\hat{\boldsymbol{\beta}}\mathbf{x}'\mathbf{y} - n\bar{Y}^2\right)/(k-1)}{\left(\mathbf{y}'\mathbf{y} - n\bar{Y}^2\right)/(n-k)}. \qquad (9.7.1)$$

розподілено за законом F-розподілу з (k–1) і (n–k) степенями вільності.

У розд. 8 ми бачили, що при зроблених вище припущеннях існує тісний зв'язок між F і R^2, а саме

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

Отже, табл. 9.2 може бути перетворена до вигляду табл. 9.3.

Таблиця 9.3

ANOVA-таблиця для k змінних в термінах R^2

Джерело дисперсії	SS	DF	MSS
Унаслідок регресії	$R^2\left(\mathbf{y}'\mathbf{y} - n\bar{Y}^2\right)$	k–1	$\dfrac{R^2\left(\mathbf{y}'\mathbf{y} - n\bar{Y}^2\right)}{k-1}$
За залишками	$\left(1-R^2\right)\left(\mathbf{y}'\mathbf{y} - n\bar{Y}^2\right)$	n–k	$\dfrac{\left(1-R^2\right)\left(\mathbf{y}'\mathbf{y} - n\bar{Y}^2\right)}{n-k}$
Загальна	$\mathbf{y}'\mathbf{y} - n\bar{Y}^2$	n–1	

Однією з переваг табл. 9.3 в порівнянні з табл. 9.2 є те, що весь аналіз може бути виконаний у термінах R^2; немає потреби розглядати складова $\mathbf{y}'\mathbf{y} - n\bar{Y}^2$, оскільки він випадає з виразу для F.

9.8. Перевірка лінійних обмежень. Загальний F-тест у матричних позначеннях

У розд. 8 ми описали загальний F-тест для перевірки справедливості лінійних обмежень, що накладаються на один або більше параметрів лінійної регресії з k змінними. Відповідний тест визначається рівнянням (7.4.9). Матричний аналог цього рівняння можна отримати дуже легко. Хай

- $\hat{\mathbf{u}}_R$ – вектор залишків регресії з обмеженнями;
- $\hat{\mathbf{u}}_{UR}$ – вектор залишків регресії без обмежень.

Тоді

- $\hat{\mathbf{u}}'_R\hat{\mathbf{u}}_R = \sum \hat{u}^2_{Ri} = RSS$ з регресії з обмеженнями;
- $\hat{\mathbf{u}}'_{UR}\hat{\mathbf{u}}_{UR} = \sum \hat{u}^2_{URi} = RSS$ з регресії без урахування обмежень;
- m – кількість лінійних обмежень;
- k– кількість параметрів у регресії без обмежень;
- n – кількість спостережень.

Матричний аналог формули (7.4.9) має вигляд

$$F = \frac{\left(\hat{\mathbf{u}}'_R\hat{\mathbf{u}}_R - \hat{\mathbf{u}}'_{UR}\hat{\mathbf{u}}_{UR}\right)/m}{\left(\hat{\mathbf{u}}'_{UR}\hat{\mathbf{u}}_{UR}\right)/(n-k)}. \qquad (9.8.1)$$

У цій формулі F підкоряється закону F-розподілу з (m, n–k) степенями вільності. Як завжди, якщо підрахована величина F більша критичного значення, ми можемо відкинути обмеження на регресію, у протилежному випадку ми їх не відкидаємо.

9.9. Прогнозування в множинній регресії. Матричне формулювання

У розд. 8 ми обговорювали, використовуючи скалярні позначення, як можна застосувати множинну регресію для середнього й індивідуального прогнозів залежної змінної *Y* при заданих значеннях регресорів *X*. У цьому розділі ми покажемо, як виразити ці прогнозовані значення в матричному вигляді, та наведемо формули для дисперсій і стандартних похибок прогнозованих величин.

Середній прогноз

Хай

$$\mathbf{X}_0 = \begin{bmatrix} 1 & X_{02} & X_{03} & \dots & X_{0k} \end{bmatrix}' \tag{9.9.1}$$

зображає вектор величин *X*, для яких ми хочемо отримати прогноз $\hat{Y}_0$, тобто середній прогноз *Y*.

Оцінка множинної регресії в скалярному вигляді

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \dots + \hat{\beta}_k X_{ki}. \tag{9.9.2}$$

У матричному вигляді ця рівність може бути записана більш компактно:

$$\hat{Y}_i = \mathbf{X}_i' \hat{\boldsymbol{\beta}}, \tag{9.9.3}$$

де $\mathbf{X}_i' = \begin{bmatrix} 1 & X_{2i} & X_{3i} & \dots & X_{ki} \end{bmatrix}$, $\hat{\boldsymbol{\beta}}' = \begin{bmatrix} \hat{\beta}_1 & \hat{\beta}_2 & \dots & \hat{\beta}_k \end{bmatrix}$.

Зрозуміло, що рівняння (9.9.2) і (9.9.3) являють собою прогноз для даного значення $\mathbf{X}_i'$.

Якщо замінити в (9.9.2) (9.9.3) $\mathbf{X}_i'$ на $\mathbf{X}_0'$, то ми отримаємо

$$\left(\hat{Y}_i \mid \mathbf{X}_0'\right) = \mathbf{X}_0' \boldsymbol{\beta}. \tag{9.9.4}$$

Зазначимо, що (9.9.4) дає незміщену оцінку прогнозу $E\left(\hat{Y}_i \mid \mathbf{X}_0'\right)$, оскільки $E\left(\mathbf{X}_0' \boldsymbol{\beta}\right) = \mathbf{X}_0' \boldsymbol{\beta}$.

Індивідуальний прогноз

Як ми пам'ятаємо з розд. 5 і 8, індивідуальний прогноз *Y*, тобто Y_0, також задається формулою (9.9.3) або (9.9.4). Тобто

$$\left(\hat{Y}_0 \mid \mathbf{X}_0'\right) = \mathbf{X}_0' \boldsymbol{\beta}. \tag{9.9.5}$$

Як ілюстративний приклад розглянемо випадок з (8.1.1). У матричному формулюванні середній і індивідуальний прогнози визначаються таким чином:

$$\mathbf{X}_0 = \mathbf{X}_{1971} = \begin{bmatrix} 1 \\ 567 \\ 16 \end{bmatrix} \text{ і } \hat{\boldsymbol{\beta}} = \begin{bmatrix} 53.1603 \\ 0.7266 \\ 2.7363 \end{bmatrix}.$$

Отже,

$$\left(\hat{Y}_{1971} \mid \mathbf{X}'_{1971}\right) = 508,9297; \tag{9.9.6}$$

$$\left(Y_{1971} \mid \mathbf{X}'_{1971}\right) = 508,9297. \tag{9.9.7}$$

Дисперсія для середнього прогнозу

Формула для оцінки дисперсії $\left(\hat{Y}_0 \mid \mathbf{X}'_0\right)$ має такий вигляд

$$\text{var}\left(\hat{Y}_0 \mid \mathbf{X}'_0\right) = \sigma^2 \mathbf{X}'_0 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}_0, \tag{9.9.8}$$

де σ^2 – дисперсія залишків u_i, $\mathbf{X}'_0$ – вектор заданих величин X, у яких ми хочемо отримати прогноз, а $(\mathbf{X}'\mathbf{X})$ – матриця, визначена в (9.3.9), тобто матриця, що застосовується для знаходження коефіцієнтів регресії. Замінюючи σ^2 на її незміщену оцінку $\hat{\sigma}^2$, можемо подати (9.9.8) у вигляді

$$\text{var}\left(\hat{Y}_0 \mid \mathbf{X}'_0\right) = \hat{\sigma}^2 \mathbf{X}'_0 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}_0. \tag{9.9.9}$$

Для ілюстративного прикладу з розд. 8 отримуємо такі значення:

$$\hat{\sigma}^2 = 6.4308,\ (\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 26,3858 & -0,0982 & 1,6532 \\ -0,0982 & 0,0004 & -0,0063 \\ 1,6532 & -0,0063 & 0,1120 \end{bmatrix}.$$

Використовуючи ці дані, за формулою (9.9.9) одержуємо

$$\text{var}\left(\hat{Y}_{1971} \mid \mathbf{X}'_{1971}\right) = 3,6580 \tag{9.9.10}$$

і

$$\text{se}\left(\hat{Y}_{1971} \mid \mathbf{X}'_{1971}\right) = \sqrt{3,6580} = 1,9126. \tag{9.9.11}$$

Тепер, застосовуючи підхід із розд. 5 і 8, можна знайти $100(1-\alpha)$%-й довірчий інтервал для середнього прогнозу при заданому $\mathbf{X}_0$:

$$\hat{Y}_0 - t_{\alpha/2}\sqrt{\hat{\sigma}^2 \mathbf{X}'_0 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}_0} \le E(Y \mid \mathbf{X}_0) \le \hat{Y}_0 + t_{\alpha/2}\sqrt{\hat{\sigma}^2 \mathbf{X}'_0 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}_0}. \tag{9.9.12}$$

Дисперсія для індивідуального прогнозу

Формула для обчислення дисперсії для індивідуального прогнозу має такий вигляд:

$$\text{var}(Y_0 \mid \mathbf{X}_0) = \hat{\sigma}^2 \left[1 + \mathbf{X}'_0 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}_0\right], \tag{9.9.13}$$

де $\text{var}(Y_0 \mid \mathbf{X}_0) = E(Y_0 - \hat{Y}_0 \mid \mathbf{X}_0)^2$.

Використовуючи наші дані, одержуємо

$$\text{var}\left(Y_{1971} \mid \mathbf{X}_{1971}\right) = 10,0887 \tag{9.9.14}$$

і

$$\text{se}\left(Y_{1971} \mid \mathbf{X}_{1971}\right) = \sqrt{10,0887} = 3,1763. \tag{9.9.15}$$

Якщо ми хочемо визначити 100(1–α)%-й довірчий інтервал для індивідуального прогнозу, то можемо застосувати формулу (9.9.12), у якій стандартна похибка прогнозу визначається з (9.9.13). Очевидно, що стандартна похибка для індивідуального прогнозу більша, ніж для середнього.

9.10. Ілюстративний приклад у матричних позначеннях

Підводячи підсумки використання матричного апарату, розглянемо числовий приклад для моделі з трьома змінними. Пригадаємо приклад для регресії сукупних особистих витрат на споживання за сукупним особистим доходом і часом на період 1956–1970 рр. Наголошувалося, що змінна тренда *t* може зображувати серед іншого сукупне населення: сукупні витрати на споживання повинні збільшуватися зі зростанням населення. Одним зі шляхів для ізолювання впливу зростання населення є перехід до сукупних витрат і сукупного доходу на душу населення. Регресія сукупних витрат на душу населення залежно від сукупного доходу на душу населення дасть співвідношення між витратами і доходом незалежно від зміни населення. Змінна тренда може залишатися в моделі для обліку впливу на витрати інших чинників (наприклад, технології). Отже, модель регресії може бути зображена у вигляді

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{u}_i, \tag{9.10.1}$$

де Y – витрати на душу населення; X_2 – дохід на душу населення; X_3 – час. У табл. 9.4 подані необхідні для цієї моделі дані.

Таблиця 9.4

Витрати на споживання на душу населення (PPCE) і дохід на душу населення (PPDI) в США за 1956–1970 рр.

PPCE, Y	PPDI, X_2	Час, X_3
1673	1839	1 (1956)
1688	1844	2
1666	1831	3
1735	1881	4
1749	1883	5
1756	1910	6
1815	1969	7
1867	2016	8
1948	2126	9
2048	2239	10
2128	2336	11
2165	2404	12

2257	2487	13
2316	2535	14
2324	2595	15 (1970)

У матричних позначеннях наша задача може бути зображена в такому вигляді:

$$\begin{bmatrix}1673\\1688\\1666\\1735\\1749\\1756\\1815\\1867\\1948\\2048\\2128\\2165\\2257\\2316\\2324\end{bmatrix} = \begin{bmatrix}1 & 1839 & 1\\1 & 1844 & 2\\1 & 1831 & 3\\1 & 1881 & 4\\1 & 1883 & 5\\1 & 1910 & 6\\1 & 1969 & 7\\1 & 2016 & 8\\1 & 2126 & 9\\1 & 2239 & 10\\1 & 2336 & 11\\1 & 2404 & 12\\1 & 2487 & 13\\1 & 2535 & 14\\1 & 2595 & 15\end{bmatrix} \begin{bmatrix}\hat{\beta}_1\\\hat{\beta}_2\\\hat{\beta}_3\end{bmatrix} + \begin{bmatrix}\hat{u}_1\\\hat{u}_2\\\hat{u}_3\\\hat{u}_4\\\hat{u}_5\\\hat{u}_6\\\hat{u}_7\\\hat{u}_8\\\hat{u}_9\\\hat{u}_{10}\\\hat{u}_{11}\\\hat{u}_{12}\\\hat{u}_{13}\\\hat{u}_{14}\\\hat{u}_{15}\end{bmatrix} \qquad (9.10.2)$$

За наведеними даними можна отримати значення величин:

$$\bar{Y} = 1942{,}333,\ \bar{X}_2 = 2126{,}333,\ \bar{X}_3 = 8{,}0,\ \sum\left(Y_i - \bar{Y}\right)^2 = 830121{,}333;$$

$$\sum\left(X_{2i} - \bar{X}_2\right)^2 = 1103111{,}333,\ \sum\left(X_{3i} - \bar{X}_3\right)^2 = 280{,}0;$$

$$\mathbf{X'X} = \begin{bmatrix} n & \sum X_{2i} & \sum X_{3i} \\ \sum X_{2i} & \sum X_{2i}^2 & \sum X_{2i}X_{3i} \\ \sum X_{3i} & \sum X_{2i}X_{3i} & \sum X_{3i}^2 \end{bmatrix} = \begin{bmatrix} 15 & 31895 & 120 \\ 31895 & 68922.513 & 272144 \\ 120 & 272144 & 1240 \end{bmatrix}$$

;

$$\mathbf{X'Y} = \begin{bmatrix} 29135 \\ 62905{,}821 \\ 247934 \end{bmatrix},\ \hat{\boldsymbol{\beta}} = \begin{bmatrix} 300{,}28625 \\ 0{,}74198 \\ 8{,}04356 \end{bmatrix}. \qquad (9.10.3)$$

Сума квадратів залишків може бути підрахована за формулою

$$\sum \hat{u}_i^2 = \hat{\mathbf{u}}'\hat{\mathbf{u}} = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{x}'\mathbf{y} = 1976,85574. \qquad (9.10.4$$

Звідси можна отримати

$$\hat{\sigma}^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{12} = 164,73797. \qquad (9.10.5)$$

Матрицю варіацій можна обчислити за такою формулою

$$\text{var}-\text{cov}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1} = \begin{bmatrix} 6133,650 & -3,70794 & 220,20634 \\ -3,70794 & 0,00226 & -0,13705 \\ 220,20634 & -0,13705 & 8,90155 \end{bmatrix}. \qquad (9.10.6)$$

Діагональні елементи цієї матриці дають значення дисперсій коефіцієнтів $\hat{\beta}_1$, $\hat{\beta}_2$ і $\hat{\beta}_3$, а корінь квадратний із них дає значення стандартної похибки.

За наведеними даними можна легко перевірити, що

$$\text{ESS}=828144,47786; \qquad (9.10.7)$$

$$\text{TSS}=830121,333. \qquad (9.10.8)$$

Отже,

$$R^2=0,99761. \qquad (9.10.9)$$

Застосовуючи формулу (7.8.4) можна визначити наведений коефіцієнт детермінації

$$\bar{R}^2 = 0,99722. \qquad (9.10.10)$$

Підсумовуючи отримані результати, одержуємо

$$\begin{gathered} \hat{Y}_i = 300,28625 + 0,74198X_{2i} + 8,04356X_{3i}, \\ \textbf{(78,31763)} \quad \textbf{(0,04753)} \quad \textbf{(2,98354)} \\ t = \ (3,83421) \quad (15,61077) \quad (2,69598) \\ R^2=0,99761, \quad \bar{R}^2 = 0,99722, \quad \text{DF}=12. \end{gathered} \qquad (9.10.11)$$

Інтерпретація результатів (9.10.14). Якщо обидві величини X_2 і X_3 набувають фіксованих нульових значень, то середня величина витрат на душу населення складає приблизно 300 дол. $\hat{\beta}_1$ повинна братися з великою обережністю. Частинний коефіцієнт регресії $\hat{\beta}_2 = 0,74198$ означає, що не змінюючи інші змінні, зростання доходів приводить до збільшення витрат на споживання на душу населення на 0,74 дол. Коротше кажучи, оцінка граничної схильності до споживання складає приблизно 74%. Аналогічно, не змінюючи інших змінних, середні витрати на споживання зростають за рік приблизно на 8 дол. за досліджуваний період. Величина $R^2=0,99761$ показує, що взяті дві пояснювальні змінні дозволяють врахувати більше 99% дисперсії витрат на споживання на душу населення в США за

даний період. Хоча $\bar{R}^2 = 0{,}99722$ трохи менше, ніж R^2, проте цей коефіцієнт залишається дуже високим.

Переходячи до аналізу статистичної значущості коефіцієнтів регресії, ми відзначаємо з (9.10.14), що кожний окремий коефіцієнт регресії статистично значимий, скажімо, при 5%-му рівні значущості (з таблиць ми бачимо, що критичне значення t для DF=12 є 2.179). Кожна з підрахованих t-величин більша, ніж це значення. Отже, ми можемо відхиляти нульові гіпотези про нульові значення величин істинних коефіцієнтів регресії.

Як уже було відзначено раніше, ми не можемо застосувати результати t-тесту для перевірки гіпотези про те, що $\beta_2 = \beta_3 = 0$, оскільки процедура t-тесту припускає, що при проведенні тесту кожного разу проводиться незалежна вибірка. Якщо ж одна й та сама вибірка використовується для перевірки гіпотез одночасно для β_2 і β_3, то ймовірно, що оцінки $\hat{\beta}_2$ і $\hat{\beta}_3$ від'ємно корельовані (коваріація між ними складає –0,13705). Тому ми не можемо застосовувати t-тест для перевірки гіпотези про те, що $\beta_2 = \beta_3 = 0$.

Для перевірки цієї гіпотези може бути застосований і F-тест, розглянутий нами в розд. 8. Для використовування F-тесту нам необхідні дані ANOVA-таблиці.

Таблиця 9.5

ANOVA-таблиця для даних з таблиці 9.4.

Джерело дисперсії	SS	DF	MSS
Унаслідок X_2 і X_3	828144,47786	2	414072,3893
Унаслідок залишків	1976,85574	12	164,73797
Загальна	830121,33360	14	

За результатами цієї таблиці звичайним способом одержуємо

$$F = 2513{,}52. \qquad (9.10.12)$$

Підрахована величина значно перевершує критичне значення F-розподілу з 2 і 12 степенями вільності. Отже, ми можемо відкинути гіпотезу про те, що $\beta_2 = \beta_3 = 0$, тобто, що витрати на споживання на душу населення не пов'язані лінійно з доходом і трендом.

Раніше ми розглядали застосування регресійної моделі для побудови як середнього, так й індивідуального прогнозів. Припустимо, що для 1971 р. дохід на душу населення складав 2 610 дол., і ми хочемо спрогнозувати відповідні йому витрати на споживання. У такому разі середній і індивідуальний прогнози витрат на душу населення

$$(Y_{1971} \mid X_{2\,1971},\ X_3 = 16) = \mathbf{X}'_{1971}\hat{\boldsymbol{\beta}} = 2365{,}55. \qquad (9.10.13)$$

Як відомо, дисперсія величин $\hat{Y}_{1971}$ і Y_{1971} визначається за формулами

$$\mathrm{var}\left(\hat{Y}_{1971} \mid \mathbf{X}'_{1971}\right) = \hat{\sigma}^2 \mathbf{X}'_{1971}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}_{1971} = 48,6426; \tag{9.10.14}$$

$$\mathrm{var}\left(Y_{1971} \mid \mathbf{X}'_{1971}\right) = \hat{\sigma}^2\left[1 + \mathbf{X}'_{1971}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}_{1971}\right] = 213,3806. \tag{9.10.15}$$

Стандартні похибки наведених величин мають такі значення:

$$\mathrm{se}\left(\hat{Y} \mid \mathbf{X}'_{1971}\right) = 6.9744,\ \mathrm{se}\left(Y \mid \mathbf{X}'_{1971}\right) = 14,6076 \tag{9.10.16}$$

Довірчі інтервали для прогнозованих величин при 5%-му рівні значущості визначаються за формулами

$$\begin{gathered}\hat{Y}_{1971} - t_{\alpha/2}\,\mathrm{se}(\hat{Y} \mid \mathbf{X}'_{1971}) \le E(Y \mid \mathbf{X}'_{1971}) \le \hat{Y}_{1971} + t_{\alpha/2}\,\mathrm{se}(\hat{Y} \mid \mathbf{X}'_{1971});\\ \hat{Y}_{1971} - t_{\alpha/2}\,\mathrm{se}(Y \mid \mathbf{X}'_{1971}) \le Y \mid \mathbf{X}'_{1971} \le \hat{Y}_{1971} + t_{\alpha/2}\,\mathrm{se}(Y \mid \mathbf{X}'_{1971}).\end{gathered} \tag{9.10.17}$$

Підставляючи в (9.10.20) значення вхідних величин, одержуємо такі довірчі інтервали:

$$\begin{gathered}2350,354 \le E(Y \mid \mathbf{X}'_{1971}) \le 2380,746;\\ 2333,723 \le Y \mid \mathbf{X}'_{1971} \le 2397,377.\end{gathered} \tag{9.10.18}$$

Раніше нами було введене поняття кореляційної матриці **R**. Для нашого випадку кореляційна матриця має вигляд

$$\mathbf{R} = \begin{array}{c} \\ Y \\ X_2 \\ X_3 \end{array}\begin{array}{c} \begin{array}{ccc} Y & X_2 & X_3 \end{array} \\ \begin{bmatrix} 1 & 0,9980 & 0,9743 \\ 0,9980 & 1 & 0,9664 \\ 0,9743 & 0,9664 & 1 \end{bmatrix} \end{array}. \tag{9.10.19}$$

Висновок

Основною метою розділу є ознайомлення з матричним підходом до класичної лінійної моделі регресії. Хоча була введена досить незначна кількість нових понять, проте матричні позначення дають компактний метод опису лінійної моделі регресії, що містить довільну кількість змінних.