ТЕМА 2
Методи побудови загальної лінійної моделі

Частина 2.1. Проста лінійна вибіркова регресія

Побудова простої вибіркової лінійної регресії є одним із найбільш поширених методів формалізації економічних залежностей. В темі 2 частині 2.1 розглядаються підходи до створення лінійних регресій та наведено ряд тестів, що допомагають аналізувати адекватність моделі. Після вивчення даної теми студенти зможуть будувати прості лінійні моделі, оцінювати їх параметри, здійснювати прогнозування на основі моделі, ознайомляться з нелінійними моделями, які застосовуються під час моделювання економічних явищ і процесів.

Основні питання, що розглядаються:

1. Специфікація моделі
2. Лінійна регресія і кореляція: зміст та оцінка параметрів
3. Основні припущення, що лежать в основі методу найменших квадратів
4. Оцінка значущості параметрів лінійної регресії та кореляції
5. Інтервальний прогноз на основі лінійного рівняння регресії
6. Нелінійна регресія

Основні терміни:

Проста регресія; множинна (багатофакторна) регресія; метод найменших квадратів; лінійний коефіцієнт кореляції; вибіркова коваріація; коефіцієнт детермінації; коефіцієнт еластичності; F-критерій Фішера; ступінь вільності *df*; загальна сума квадратів SST; сума квадратів, що пояснює регресію SSR; залишкова сума квадратів SSE; середній квадрат відхилень (дисперсія); ANOVA-аналіз; t-критерій Ст'юдента; інтервальна оцінка прогнозного значення; гетероскедастичність; внутрішньолінійні та внутрішньонелінійні моделі.

2.1.1. СПЕЦИФІКАЦІЯ МОДЕЛІ

Ставлячи за мету дати кількісний опис взаємозв'язків між економічними змінними, економетрика перш за все пов'язана з методами регресії та кореляції.

Строго кажучи, за своїм смисловим навантаженням слово «регресія» не має відношення до сутності стохастичних зв'язків, для опису яких воно використовується. Пояснення цьому терміну можна дати, звернувшись до історії досліджень в області статистичного аналізу зв'язків між ознаками. Одним із перших прикладів досліджень такого роду була робота шведських статистиків, що намагалися за спосте-реженнями пар ознак: x – відхилення від середнього рівня в зрості батька, y – відхилення від середнього рівня в зрості дорослого сина цього батька, – встановити і описати стохастичний взаємозв'язок, що існує між x та y. В процесі дослідження було підтверджено природну гіпотезу про наявність позитивного статистичного зв'язку між зростом батька та сина («у високих батьків в середньому високі сини, і навпаки»), однак водночас була підмічена тенденція *регресії* (відступу, повернення) в рості синів до середнього рівня, а саме: «у дуже високих батьків сини в середньому високі, але вже не такі високі, як батьки, і навпаки: у дуже маленьких за зростом батьків сини в середньому низькорослі, але все-таки вище, ніж їх батьки». Функцію, що описує ця закономірність, автори назвали *функцією регресії,* після чого цей термін і стали використовувати стосовно *будь-якої функції,* що побудована аналогічними методами.

Лінійна регресія знаходить широке застосування в економетриці через чітку економічну інтерпретацію її параметрів.

Залежно від кількості факторів, включених у рівняння регресії, прийнято розрізняти просту (парну) та багатофакторну регресію (множинну).

Проста регресія являє собою модель, де середнє значення залежної (пояснюваної) змінною *y* розглядається, як функція однієї незалежної (пояснювальної) змінної x, тобто це модель виду:

$$\bar{y} = f(x).$$

Множинна (багатофакторна) регресія являє собою модель, де середнє значення залежної (пояснюваної) змінної *y* розглядається як функція декількох незалежних (пояснювальних) змінних x1, x2, тобто це модель виду:

$$\bar{y} = f(x_1, x_2, ..., x_k).$$

Будь-яке економічне дослідження починається зі ***специфікації моделі,*** тобто формулювання виду моделі, виходячи із відповідної теорії зв'язку між змінними. Іншими словами, дослідження починається з теорії, що встановлює взаємозв'язок між явищами.

У першу чергу, з усього кола факторів, що впливають на результуючу ознаку, необхідно виділити найбільш суттєві фактори впливу. Парна регресія достатня, якщо є домінуючий фактор, який і використовується в якості пояснювальної змінної. Припустимо, висувається гіпотеза про те, що величина попиту *y* на товар А знаходиться в зворотній залежності від ціни *x*, тобто $\bar{y}_x = a - b \cdot x$. В цьому випадку треба знати, які інші фактори передбачаються незмінними, в подальшому їх доведеться врахувати в моделі й від простої регресії перейти до множинної.

Рівняння простої регресії характеризує взаємозв'язок між двома змінними, яка проявляється як деяка закономірність лише в середньому за сукупністю спостережень. Так, якщо залежність попиту *y* від ціни *x* характеризується, наприклад, рівнянням *y=5000-2·x*, то це означає, що із зростанням ціни на 1 грошову одиницю попит у середньому зменшується на 2 грошових одиниці. В рівнянні регресії кореляційний по суті зв'язок ознак представляється у вигляді функціонального зв'язку, що виражається відповідною математичною функцією. Практично в кожному окремому випадку величина *y* складається з двох складових:

$$y_j = \bar{y}_{x_j} + \varepsilon_j,$$

де y_j – фактичне значення результативної ознаки;

$\bar{y}_{x_j}$ – теоретичне значення результативної ознаки, що знаходиться, виходячи із відповідної математичної функції зв'язку *y* та *x*, тобто з рівняння регресії;

ε_j – випадкова величина, що характеризує відхилення реального значення результативної ознаки від теоретичного, знайденого за рівнянням регресії.

Випадкова величина ε_j або ***хвилювання (обурення) чи відхилення***, включає вплив неврахованих у моделі факторів, випадкових помилок і особливостей вимірювання. Її присутність у моделі зумовлена трьома джерелами:

1. Специфікація моделі. Форма рівняння попиту, наведена вище, може бути і нелінійною, а наприклад, оберненою функцією чи степеневою тощо. Помилок у моделі буде тим менше, чим ближче будуть теоретичні значення залежної змінної підходити до фактичних значень (спостережень). Можна усунути ці помилки шляхом зміни формули (маніпулювання математичною формулою).

2. Вибірковий характер даних. Оскільки дослідник працює з вибірковими даними і не завжди може досліджувати всю статистичну сукупність, це зумовлює викривлення, пов'язані із *неоднорідністю даних* у вихідній статистичній сукупності. Для отримання позитивного результату зазвичай виключають із сукупності одиниці з аномальними значеннями досліджуваних ознак. І в цьому випадку результати регресії являють собою вибіркові характеристики. Використання *часової інформації* також являє собою вибірку із всієї множини хронологічних дат. Змінивши часовий інтервал, можна отримати інші результати регресії. Ці помилки усуваються за рахунок збільшення обсягу вибірки.

3. Помилки вимірювання. Є найбільш небезпечними помилками, оскільки вони зводять нанівець усі зусилля по кількісній оцінці зв'язку між ознаками. Особлива велика роль

помилок вимірювання при дослідженні на макрорівні. Так, у дослідженнях попиту і споживання в якості пояснювальної змінної широко використовується «дохід на душу населення». Разом з тим статистичне вимірювання величини доходу пов'язано з рядом складнощів і не позбавлено можливих помилок, наприклад у результаті приховування доходів. Інший приклад: на сьогодні органи статистики отримують баланси підприємств, достовірність яких ніхто не підтверджує. Подальше узагальнення такої інформації може містити помилки вимірювання. Досліджуючи, наприклад, у якості результуючої ознаки прибуток підприємства, ми маємо бути впевнені, що підприємства показують у звітності адекватні реальній дійсності величини.

Після визначення сукупності даних, що будуть використовуватися в моделюванні, здійснюється ***підбір форми моделі***. Це можна зробити двома методами:

1) графічний – він є наочним. Основними типами кривих є такі:

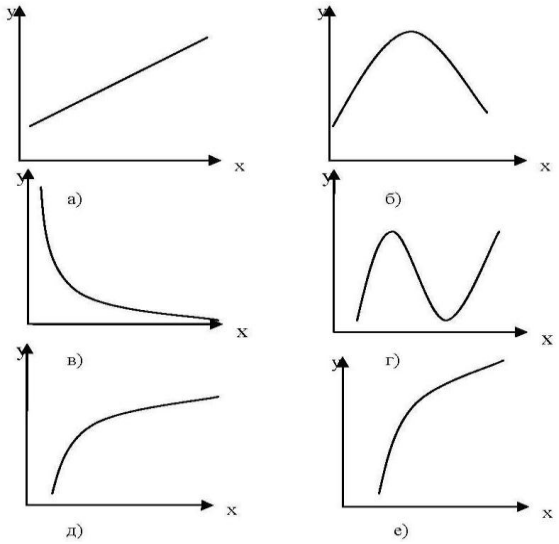

Рис. 2.1. Основні типи кривих, що використовуються при кількісній оцінці зв'язків між двома змінними:

а) $\hat{y}_x = a + b \cdot x$; б) $\hat{y}_x = a + b \cdot x + c \cdot x^2$;

в) $\hat{y}_x = a + b/x$; г) $\hat{y}_x = a + b \cdot x + c \cdot x^2 + d \cdot x^3$;

д) $\hat{y}_x = a \cdot x^b$; е) $\hat{y}_x = a \cdot b^x$

2) аналітичний метод – заснований на вивченні матеріальної природу зв'язку досліджуваних ознак. Нехай, наприклад, вивчається потреба підприємства в електроенергії *y* залежно від обсягу продукції *x,* що випускається. Все споживання електроенергії *y* можна поділити на дві частини:

– не пов'язане із виробництвом продукції *a*;

– безпосередньо зв'язане з обсягом випущеної продукції, пропорційно зростаюче із збільшенням обсягу випуску (b*x). Тоді залежність споживання електроенергії від обсягу продукції можна виразити рівнянням регресії виду

$$\hat{y}_x = a + b \cdot x.$$

Якщо потім розділити обидві частини рівняння на величину обсягу випуску продукції *x*, то отримаємо вираз залежності питомої витрати електроенергії на одиницю продукції $\left(z = \frac{y}{x}\right)$ від обсягу випущеної продукції *x* у вигляді рівняння рівносторонньої гіперболи:

$$\hat{z}_x = b + \frac{a}{x}.$$

При обробці інформації на комп'ютері вибір виду рівняння регресії зазвичай проводиться експериментальним методом, тобто шляхом порівняння величини залишкової дисперсії, розрахованої за різними моделями.

2.1.2. ЛІНІЙНА РЕГРЕСІЯ: ЗМІСТ ТА ОЦІНКА ПАРАМЕТРІВ

Лінійна регресія зводиться до знаходження рівняння виду:

$$\hat{y}_x = a + b \cdot x \text{ або } y = a + b \cdot x + \varepsilon$$

Рівняння виду $\hat{y}_x = a + b \cdot x$ дозволяє по заданих значеннях фактору *x* мати теоретичні значення результативної ознаки підстановкою в нього фактичних значень фактору *x*.

Побудова лінійної моделі зводиться до оцінки її параметрів – a і b. Оцінки параметрів лінійної регресії можуть бути знайдені різними методами.

Один з методів – використання графіка первинних даних (див. рис. 2.2). Через дві точки на графіку проводимо пряму лінію і за графіком визначаємо її параметри – a і b.

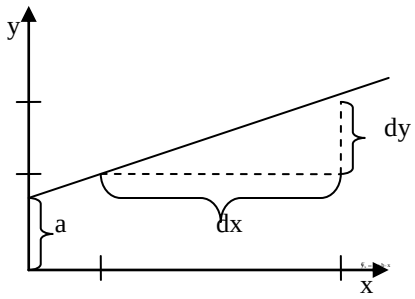

Рис. 2.2. Метод знаходження параметрів лінійної регресії за графіком первинних даних

Класичний підхід до оцінювання параметрів лінійної регресії заснований на методі найменших квадратів.

Метод найменших квадратів (МНК) дозволяє отримати такі оцінки параметрів a і b, при яких сума квадратів відхилень фактичних значень результативної ознаки *y* від розрахункових (теоретичних) $\hat{y}_x$ мінімальна:

$$\sum_i (y_i - \hat{y}_{x_i})^2 \to \min. \qquad (1)$$

Іншими словами, з усієї множини ліній лінія регресії на графіку обирається так, щоб сума квадратів відстані по вертикалі між точками і цією лінією була б мінімальною (як це показано на рис. 2.3): $\varepsilon_i = y_i - \hat{y}_x$; $\sum_i \varepsilon_i^2 \to \min$.

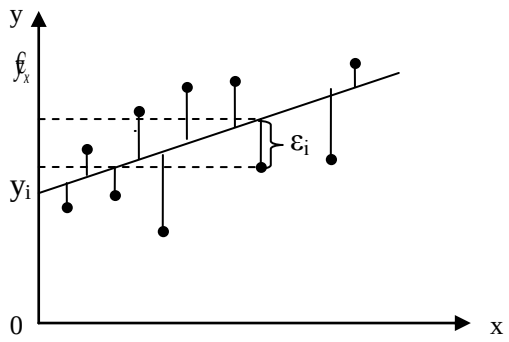

Рис. 2.3. Відхилення фактичних значень залежної змінної від теоретичних

Для того, щоб знайти мінімум функції (1), треба вирахувати часткові похідні за кожним з параметрів a та b і прирівняти їх до нуля.

Якщо ми позначимо $\sum \varepsilon_i^2$ через S, тоді:

$$S = \sum (y_i - \hat{y}_x)^2 = \sum (y - a - b \cdot x)^2$$

$$\frac{dS}{da} = -2\sum y + 2 \cdot n \cdot a + 2 \cdot b \sum x = 0 \qquad (2)$$

$$\frac{dS}{db} = -2\sum y \cdot x + 2 \cdot a \sum x + 2 \cdot b \sum x^2 = 0$$

Перетворюючи формулу (2), отримуємо таку систему нормальних рівнянь для оцінки параметрів a і b:

$$\begin{cases} n \cdot a + b \sum x = \sum y \\ a \sum x + b \sum x^2 = \sum y \cdot x. \end{cases} \quad (3)$$

Розв'язуючи систему нормальних рівнянь (3) або методом послідовного виключення змінних, або методом визначників, знайдемо шукані оцінки параметрів a і b. Можна скористатися такими формулами для a і b:

$$a = \bar{y} - b \cdot \bar{x}. \quad (4)$$

Формула (4) отримана з першого рівняння системи (3), якщо всі його члени розділити на n:

$$b = \frac{\text{cov}(x, y)}{\sigma_x^2} = \frac{\overline{yx} - \bar{y} \cdot \bar{x}}{\overline{x^2} - \bar{x}^2}. \quad (5)$$

Формула (5) отримується також при розв'язанні системи (3) методом визначників, якщо всі елементи розрахунку поділити на n^2.

Параметр *b* називається коефіцієнтом регресії. Його величина показує середню зміну результату із зміною фактору на одиницю. Наприклад, $\hat{y}_x = 20 + 4x$, то із зростанням величини *x* на 1, *y* буде збільшуватись у середньому на 4. Знак при коефіцієнті регресії вказує на напрямок зв'язку, якщо він більше 0, то зв'язок прямий, якщо менше – зворотний.

Приклад. По групі підприємств, що випускають один і той же вид продукції, розглядається функція витрат $y = a + b \cdot x + e$. Необхідна для розрахунку оцінок параметра a і b інформація представлена в таблиці:

Номер підприємства	**Випуск продукції, тис. од. x**	**Витрати на виробництво, тис. грн y**	**y·x**	**x^2**	**y^2**	$\hat{y}_x$
1	1	30	30	1	900	31,1
2	2	70	140	4	4900	67,9

Закінчення таблиці

3	4	150	600	16	22 500	141,6
4	3	100	300	9	10 000	104,7
5	5	170	850	25	28 900	178,4
6	3	100	300	9	10 000	104,7
7	4	150	600	16	22 500	141,6
Разом	22	770	2 820	80	99 700	770,0

Система нормальних рівнянь буде мати вигляд:

$$\begin{cases} 7 \cdot a + b \cdot 22 = 770 \\ a \cdot 22 + b \cdot 80 = 2\,820. \end{cases}$$

Розв'язавши її, отримаємо:

$$a = -5{,}79$$
$$b = 36{,}84.$$

Запишемо рівняння регресії:

$$\hat{y}_x = -5{,}79 + 36{,}84 \cdot x$$

Підставивши в рівняння x знайдемо теоретичні значення y (останній стовпчик). В даному випадку величина параметра a не має ніякого економічного змісту. В нашому прикладі маємо:

$$\bar{x} = 3{,}14; \sigma_x = 1{,}25; V_x = 39{,}8\,\%$$
$$\bar{y} = 110; \sigma_y = 46{,}29; V_y = 42{,}1\,\%$$

Те, що $a<0$, відповідає випередженню зміни результату над зміною фактору $V_y > V_x$.

Парне лінійне рівняння регресії може бути записане в матричній формі:

$$Y = BX + E,$$

де *Y* – вектор-стовпчик розмірності *(n*x*1)* фактичних значень результативної ознаки;

B – вектор-стовпчик розмірності *(2x1)* параметрів моделі, що підлягають оцінці, тобто коефіцієнта регресії *b* і вільного члена (параметра *a*);

X = (x₀, x₁) – матриця розмірності *(nx2)* значень факторів. При цьому $x_0 = 1$ і зв'язане з наявністю в рівнянні регресії вільного члена *(a)*, а x_1 – фактичні значення включеного в рівняння регресії фактору.

E – вектор-стовпчик випадкової величини e_i розмірності (nx1).

Матриця вихідних даних має вигляд:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \therefore \\ y_n \end{pmatrix}, \qquad X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{pmatrix}$$

Оцінюючи параметри лінійного рівняння регресії, знайдемо вектор B і далі вектор випадкової компоненти E:

$$B = \begin{pmatrix} a \\ b \end{pmatrix}, \qquad E = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

У матричній формі застосування МНК записується так:

$$S = (Y - XB)^T (Y - XB) \to \min.$$

Диференціюючи S по вектору B і прирівнюючи перші часткові похідні по B до нуля, отримаємо:

$$\frac{\partial S}{\partial B} = -2X^T Y + 2X^T XB = 0$$

Звідси отримаємо $(X^T X)B = X^T Y$.

Відповідно оцінка вектора B складе:

$$B = (X^T X)^{-1} X^T Y$$

Приклад. Застосуємо до попереднього прикладу матричний метод визначення МНК-оцінок, який зводиться до такого:

1) за правилом добутку матриць:

$$X^T X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 3 & 5 & 3 & 4 \end{pmatrix} \begin{pmatrix} 11 \\ 12 \\ 14 \\ 13 \\ 15 \\ 13 \\ 14 \end{pmatrix} = \begin{pmatrix} 7 & 22 \\ 22 & 80 \end{pmatrix}$$

2) знайдемо зворотну матрицю:

$$(X^T X)^{-1} = \frac{1}{7 \cdot 80 - (22)^2} \begin{pmatrix} 80 & -22 \\ -22 & 7 \end{pmatrix} = \begin{pmatrix} 1{,}05263 & -0{,}28947 \\ -0{,}28947 & 0{,}09211 \end{pmatrix};$$

$$X^T Y = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 3 & 5 & 3 & 4 \end{pmatrix} \begin{pmatrix} 30 \\ 70 \\ 150 \\ 100 \\ 170 \\ 100 \\ 150 \end{pmatrix} = \begin{pmatrix} 770 \\ 2\,820 \end{pmatrix};$$

Вектор оцінок параметрів регресії дорівнює:

$$B = \begin{pmatrix} 1{,}05263 & -0{,}28947 \\ -0{,}28947 & 0{,}09211 \end{pmatrix} \begin{pmatrix} 770 \\ 2820 \end{pmatrix} = \begin{pmatrix} -5{,}79 \\ 36{,}84 \end{pmatrix}.$$

Рівняння регресії завжди доповнюється показником тісноти зв'язку. При використанні лінійної регресії в якості такого показника виступає ***лінійний коефіцієнт кореляції*** r_{xy}**.** Існують різні модифікації цієї формули:

$$r_{xy} = b\frac{\sigma_x}{\sigma_y} = \frac{\mathrm{cov}(x,y)}{\sigma_x\sigma_y} = \frac{\overline{yx} - \bar{y}\cdot\bar{x}}{\sigma_x\sigma_y}$$

Величина $\mathrm{cov}(x,y)$, що знаходиться в чисельнику, визначається співвідношенням:

$$\mathrm{cov}(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

і називається ***вибірковою коваріацією*** змінних x та y, так що, формально,

$$\mathrm{cov}(x,x) = \mathrm{var}(x),\ \ \mathrm{cov}(y,y) = \mathrm{var}(y).$$

Якщо вказана тенденція виражена на діаграмі розсіювання доволі ясно, то значення r_{xy} за абсолютною величиною (за модулем) близькі до одиниці (тобто значення r_{xy} близькі до + 1 або до – 1). Якщо ж наявність лінійної тенденції зв'язку виявляється на діаграмі розсіювання важко, то тоді значення r_{xy} близькі до нуля.

Якщо взяти наші дані і розрахувати лінійний коефіцієнт кореляції то він буде дорівнювати 0,991, що означає наявність дуже тісного зв'язку витрат на виробництво від величини обсягу випущеної продукції.

Графічно можна визначити характер кореляції між залежною і незалежною змінними (див. рис. 2.4.).

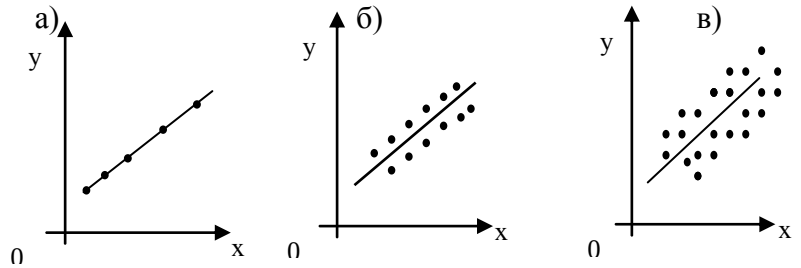

Рис. 2.4. Типи кореляції: а) повна кореляція; б) сильна кореляція; в) слабка кореляція

Для оцінки якості підбору лінійної функції розраховується квадрат лінійного коефіцієнту кореляції r^2_{xy}, що називається коефіцієнтом детермінації.

Коефіцієнт детермінації характеризує частку дисперсії результа-тивної ознаки *y*, що пояснюється регресією. Відповідно величина $1\text{-}r^2_{xy}$ характеризує частку дисперсії *y*, викликану впливом інших не врахованих у моделі факторів.

Величина коефіцієнта детермінації є одним з критеріїв оцінки якості лінійної моделі.

Зв'язок коефіцієнта кореляції з параметром b описується таким чином:

$$r_{xy} = b_{y/x}\frac{\sigma_x}{\sigma_y}.$$

Його величина виступає в якості стандартизованого коефіцієнту регресії і характеризує середню в сігмах (σ_y) зміну результату із зміною фактору на одну σ_x.

Лінійний коефіцієнт кореляції як вимір тісноти лінійного зв'язку ознак логічно пов'язаний не тільки з коефіцієнтом регресії b, але й з коефіцієнтом еластичності.

$$E_{y/x} = b_{y/x}\frac{\bar{x}}{\bar{y}}$$

Таким чином вимірювачем *тісноти зв'язку* виступає коефіцієнт кореляції, а *коефіцієнт регресії і* ***коефіцієнт еластичності*** – це показники *сили зв'язку* коефіцієнт регресії – абсолютна міра; коефіцієнт еластичності – відносна міра.

Парна регресія застосовується в економічних дослідженнях при знаходженні:

– функції споживання: $C=K\cdot y+L$, де С – споживання, К та L – параметри функції, у – дохід;

– балансове рівняння: $y=C+I-r$, де I – розмір інвестицій; r – заощадження (це і попереднє рівняння можуть розглядатися в системі).

2.1.3. ОСНОВНІ ПРИПУЩЕННЯ, ЩО ЛЕЖАТЬ В ОСНОВІ МЕТОДУ НАЙМЕНШИХ КВАДРАТІВ

Мета регресійного аналізу полягає не тільки у визначенні невідомих параметрів вибіркової лінійної моделі, а насамперед, у висновках, які ми можемо зробити щодо дійсних значень параметрів узагальненої моделі. Для цього ми маємо не тільки точно визначити функціональну форму моделі, а й зробити певні припущення щодо випадкової величини та зв'язку між випадковою величиною *e* та незалежною змінною *x*.

Припущення 1. Математичне сподівання випадкової величини *e* дорівнює нулеві. Це означає, що усереднений чи очікуваний вплив цих помилок на *y* дорівнює нулеві.

Припущення 2. Відсутність автокореляції між випадковими величинами *e*. Воно стверджує, що випадкові величини незалежні одна від одної, тобто будь-яке і-те значення випадкової величини *e* не впливає на будь-яке j-те значення цієї величини.

Припущення 3. Гомоскедастичність, або однакова дисперсія випадкової величини *e*. Це припущення вимагає, щоб усі випадкові величини, незалежно від номера спостереження, мали однакову дисперсію. ***Гетероскедастичність*** **–** умовна дисперсія розподілу *y* зростає із збільшенням значень *x* (рис. 2.5).

Припущення 4. Незалежність між значеннями випадкової величини *e* і значеннями змінної *x*, або нульова коваріація між *e* та *x*. Передбачається відсутність зв'язку між цими двома величинами.

Припущення 5. Регресійну модель визначено (специфіковано) правильно (відсутність похибки) (форма).

Припущення 6. Випадкова величина розподілена нормально з математичним сподіванням 0 (оскільки сума помилок = 0, тому і середнє = 0) та сталою дисперсією (відсутність гетероскедастичності).

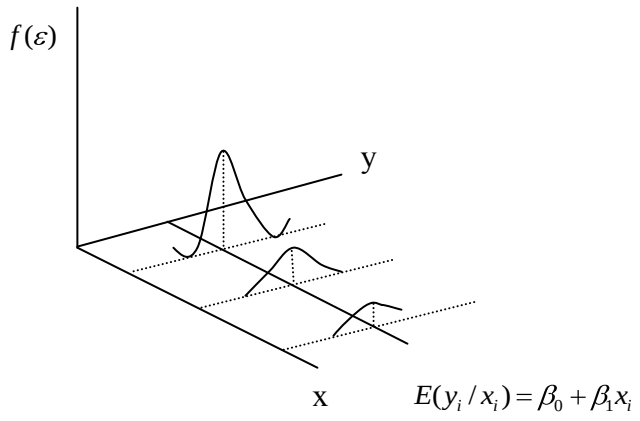

Рис. 2.5. Гетероскедастичність, або нерівна дисперсія

2.1.4. ОЦІНКА ЗНАЧУЩОСТІ ПАРАМЕТРІВ ЛІНІЙНОЇ РЕГРЕСІЇ ТА КОРЕЛЯЦІЇ

Після того, як рівняння лінійної регресії знайдено, проводиться оцінка значущості, як рівняння в цілому, так і окремих його параметрів.

Оцінка значущості рівняння регресії в цілому здійснюється за допомогою ***F-критерію Фішера***. При цьому висувається нульова гіпотеза, що коефіцієнт регресії дорівнює нулеві, тобто: H_0: b=0, тобто фактор *x* не спричиняє впливу на результат *y*.

Безпосередньо розрахунку F-критерію передує аналіз дисперсії. Центральне місце в ньому займає розклад загальної суми квадратів відхилень змінної y від середнього значення $\bar{y}$ на дві частини – «пояснену» та «залишкову» (непояснену):

$$\Sigma(y-\bar{y})^2 = \Sigma(\hat{y}_x-\bar{y})^2 + \Sigma(y-\hat{y}_x)^2$$

загальна сума квадратів = сума квадратів відхилень, + залишкова сума квадратів

відхилень — пояснена — відхилень регресією

SST — *SSR* — *SSE*

Кожна сума квадратів пов'язана з числом, яке називають її «ступенем вільності». Це число показує, скільки незалежних елементів інформації, що утворилися з елементів y_1, y_2 тощо потрібно для розрахунку даної суми квадратів.

У статистиці ***кількістю ступенів вільності df*** певної величини часто називають різницю між кількістю різних дослідів і кількістю констант, встановлених у результаті цих дослідів, незалежно один від одного.

Для SST (загальної суми квадратів) – потрібно *(n-1)* незалежних чисел, тому що з чисел (y_1- $\bar{y}$); (y_2- $\bar{y}$); (y_n- $\bar{y}$) незалежні тільки *(n-1)* завдяки властивості: $\Sigma(y-\bar{y})=0$.

Приклад. Маємо ряд значень 1, 2, 3, 4, 5. Середнє з них дорівнює 3, і тоді n відхилень від середнього складе: – 2, – 1, 0, 1, 2. Оскільки сума їх дорівнює 0, то вільно варіюють лише 4 відхилення, а п'яте може бути визначене, якщо 4 попередніх відомі.

SSR (сума квадратів, що пояснює регресію) розраховується простим шляхом:

$$\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2=\sum_{i=1}^{n}(b_0+b_1x_i-(b_0+b_1\bar{x}))^2=b_1^2\sum_{i=1}^{n}(x_i-\bar{x})^2;$$

$$\sum_{i=1}^{n}(\hat{y}_x-\bar{y})^2=b_1{}^2\sum_{i=1}^{n}(x_i-\bar{x})^2\ .$$

Отже, суму квадратів, що пояснює просту лінійну регресію, можна утворити, використовуючи тільки одну одиницю незалежної інформації, а саме b_1 (тільки для однофакторної регресії). Звідси SSR має *один ступінь вільності.*

SSE (сума квадратів помилок або залишкова сума квадратів) має *(n-2)* ступенів вільності. Ця сума базується на кількості ступенів вільності, яка дорівнює різниці між кількістю спостережень і кількістю параметрів, що оцінюються. У разі простої лінійної регресії оцінюються два параметри b_0 та b_1.

У разі простої лінійної регресії ступені вільності, як і суми квадратів можна розкласти таким чином:

$$n-1=1+(n-2)$$

Розділивши кожну суму квадратів на кількість ступенів вільності, отримаємо ***середній квадрат відхилень*** або ***дисперсію*** *на один ступінь вільності*.

$$D_{загальн}=\frac{\sum(y-\bar{y})^2}{n-1};$$

$$MSR=\frac{\sum(\hat{y}_x-\bar{y})^2}{1};$$

$$MSE=\frac{\sum(y-\hat{y}_x)^2}{n-2}.$$

Визначення дисперсії на один ступінь вільності приводить дисперсії до порівняного виду. Співставляючи факторну та залишкову дисперсії в розрахунку на один ступінь вільності, отримаємо величину F-критерію:

$$F=\frac{MSR}{MSE}.$$

Якщо нульова гіпотеза справедлива H_0, то факторна і залишкова дисперсії не відрізняються одна від одної. Якщо H_0 несправедлива, то факторна дисперсія перевищує залишкову в декілька разів.

Англійський статистик Снедекор розробив таблиці критичних значень F-відношень при різних рівнях значущості нульової гіпотези і різній кількості ступенів вільності. Табличне значення F-критерію – це максимальна величина відношення дисперсій, яка може мати місце при випадковому розходженні їх для даного рівня ймовірності наявності нульової гіпотези.

Обчислене значення F-відношення визнається достовірним (відмінним від одиниці), якщо воно більше табличного. В цьому випадку нульова гіпотеза про відсутність зв'язку ознак відхиляється і робиться висновок про суттєвість цього зв'язку:

$$F_{факт} > F_{табл}, \ H_0 \text{ відхиляється.}$$

Якщо ж величина F виявиться менше табличної, то імовірність нульової гіпотези вище заданого рівня (наприклад, 0,05) і вона не може бути відхилена без ризику зробити неправильний висновок про наявність зв'язку. В цьому випадку рівняння регресії виявляється статистично не значущим.

Величина F-критерію зв'язана з коефіцієнтом детермінації R^2. Факторну суму відхилень можна представити як:

$$\sum(\hat{y}_x - \bar{y})^2 = r^2 \cdot \sigma_y^2 \cdot n,$$

а залишкову суму квадратів – як:

$$\sum(\hat{y}_x - \bar{y})^2 = (1 - r^2) \cdot \sigma_y^2 \cdot n,$$

Тоді значення F-критерію можна виразити таким чином:

$$F = \frac{r^2}{1 - r^2} \cdot (n - 2).$$

У нашому прикладі F = 278; Fкр. = 6,61 (при імовірності 99 %).

Таким чином, можна зробити висновок, що зв'язок між змінними в регресійному рівнянні суттєвий.

Суми квадратів пов'язані з певним джерелом варіації, а також із ступенями вільності і середніми квадратами. Зведемо їх усіх у таблицю, яка називається таблицею дисперсійного аналізу (***ANOVA-аналіз***).

Таблиця

ANOVA-аналіз

Джерело варіації	**Кількість ступенів вільності**	**Сума квадратів**	**Середні квадрати**	**F-відношення**	
				Фактичне	**Табличне при α = 0,05**
Зумовлене регресією (модель)	1	$SSR = \sum_{i=1}^{n}(\hat{y} - \bar{y})^2$	$MSR = \sum_{i=1}^{n}(\hat{y} - \bar{y})^2 / 1$	278	6.61
Не пояс-нюване за допомогою регресії (помилка)	n-2	$SSE = \sum_{i=1}^{n}(y - \hat{y})^2$	$MSE = \sum_{i=1}^{n}(y - \hat{y})^2 / (n-2)$	1	–
Загальне	n-1	$SST = \sum_{i=1}^{n}(y - \bar{y})^2$			

У лінійній регресії зазвичай оцінюється значущість не тільки рівняння в цілому, але й окремих його параметрів. З цією метою за кожним з параметрів визначається його стандартна похибка: m_b та m_a.

Стандартна похибка коефіцієнту регресії параметра m_b розра-ховується за формулою:

$$m_b = \sqrt{\frac{\sum (y - \hat{y}_x)^2 / (n-2)}{\sum (x - \bar{x})^2}} = \sqrt{\frac{MSE}{\sum (x - \bar{x})^2}},$$

MSE – залишкова дисперсія на один ступінь свободи.

Відношення коефіцієнту регресії до його стандартної похибки дає *t*-статистику, яка підкоряється статистиці Ст'юдента при *(n-2)* ступенях вільності. Ця статистика застосовується для перевірки статистичної значущості коефіцієнтів регресії і для розрахунку його довірчих інтервалів.

Для оцінки значущості коефіцієнту регресії його величину порівнюють з його стандартною похибкою, тобто визначають фактичне значення ***t-критерію Ст'юдента***: $t_b = \frac{b}{m_b}$, яке потім порівнюють з табличним значенням при певному рівні значущості α і числа ступенів свободи *(n-2)*.

Між оцінкою параметра регресії та коренем з коефіцієнта Фішера існує зв'язок:

$$t_b = \sqrt{F};$$

$$t_b^2 = \frac{b^2}{m_b^2} = b^2 / \frac{\sum (y - \hat{y}_x)^2 / (n-2)}{\sum (x - \bar{x})^2} = \frac{b^2 \sum (x - \bar{x})^2}{\sum (y - \hat{y}_x)^2 / (n-2)} =$$

$$= \frac{\sum (\hat{y}_x - \bar{y})^2}{\frac{\sum (y - \hat{y}_x)^2}{(n-2)}} = \frac{MSR}{MSE} = F.$$

Якщо фактичне значення *t*-критерію перевищує табличне, гіпотезу про несуттєвість коефіцієнтів регресії можна відхилити. Якщо менше – то вони незначущі.

Довірчий інтервал для коефіцієнту регресії визначається як: $b \pm tm_b$.

Стандартна похибка параметра *a* визначається за формулою:

$$m_a = \sqrt{\frac{\sum (y - \hat{y}_x)^2}{n-2} \cdot \frac{\sum x^2}{n \sum (x - \bar{x})^2}} = \sqrt{MSE \cdot \frac{\sum x^2}{n \sum (x - \bar{x})^2}}.$$

Процедура оцінювання значущості даного параметра не відрізняється від розглянутої вище для коефіцієнта регресії: розраховується *t*-критерій:

$$t_a = \frac{a}{m_a},$$

його величина порівнюється з табличним значенням при df = n-2 ступенях вільності.

Значущість лінійного *коефіцієнта кореляції* перевіряється на основі величини похибки коефіцієнта кореляції m_r:

$$m_r = \sqrt{\frac{1 - r^2}{n-2}}.$$

Фактичне значення t-критерію Ст'юдента визначається як:

$$t_r = \frac{r}{\sqrt{1 - r^2}} \cdot \sqrt{n-2}.$$

Оскільки t^2=F, то $F = \frac{r}{\sqrt{1 - r^2}} \cdot \sqrt{n-2}$, відповідно $t_r^2 = t_b^2$.

Принцип оцінки коефіцієнта Ст'юдента аналогічний до критерію Фішера.

2.1.5. ІНТЕРВАЛЬНИЙ ПРОГНОЗ НА ОСНОВІ ЛІНІЙНОГО РІВНЯННЯ РЕГРЕСІЇ

У прогнозних розрахунках за рівнянням регресії визначається передбачуване y_p (прогнозне) як точковий прогноз $\hat{y}_x$ при $x_p = x_к$, тобто шляхом підстановки в лінійне рівняння регресії значення *x*. Однак точковий прогноз явно нереальний, тому він доповнюється розрахунком стандартної похибки $\hat{y}_x$, тобто $m_{\hat{y}_x}$, і відповідно ми отримуємо ***інтервальну оцінку прогнозного значення*** *y**:

$$\hat{y}_x - t \cdot m_{\hat{y}_x} \le y^* \le \hat{y}_x + t \cdot m_{\hat{y}_x}$$

Для того, щоб зрозуміти, як будується формула для визначення величини стандартної похибки $\hat{y}_x$, підставимо в рівняння лінійної регресії вираження параметра *a*:

$$a = y - bx$$

тоді рівняння регресії матиме вигляд:

$$\hat{y}_x = \bar{y} - b\bar{x} + bx = \bar{y} + b(x - \bar{x}).$$

Звідси випливає, що стандартна похибка $m_{\hat{y}_x}$, залежить від похибки $\bar{y}$ і похибки коефіцієнта регресії b, тобто

$$m_{\hat{y}_x}^2 = m_{\bar{y}}^2 + m_b^2 (x - \bar{x})^2.$$

З теорії вибірки відомо, що $m_{\bar{y}}^2 = \frac{\sigma^2}{n}$. Використовуючи в якості оцінки дисперсії σ^2 залишкову дисперсію на один ступінь вільності MSE, отримуємо формулу розрахунку похибки середнього значення змінної *y*:

$$m_{\bar{y}}^2 = \frac{MSE}{n}.$$

Помилка коефіцієнта регресії, як вже було показано, визначається формулою:

$$m_b{}^2 = \frac{MSE}{\sum (x - \bar{x})^2}.$$

Вважаючи, що прогнозне значення фактору $x_p = x_к$, отримаємо таку формулу для розрахунку стандартної похибки передбачуваного за лінією регресії значення, тобто $m_{\hat{y}_x}$:

$$m_{\hat{y}_x}^2 = \frac{MSE}{n} + \frac{MSE}{\sum (x - \bar{x})^2} (x_k - \bar{x})^2 = MSE \cdot \left(\frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum (x - \bar{x})^2} \right).$$

Відповідно $m_{\hat{y}_x}$ має вираз:

$$m_{\hat{y}_x} = \sqrt{MSE \cdot (\frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum (x - \bar{x})^2})}.$$

Ця формула характеризує похибку положення лінії регресії. Величина стандартної похибки $m_{\hat{y}_x}$ досягає мінімуму при $x_к = \bar{x}$ і зростає по мірі того, як «віддаляється» від $\bar{x}$ у будь-якому напрямку.

Приклад. Для даних з нашого прикладу $m_{\hat{y}_x}$ складе:

$$m_{\hat{y}_x} = \sqrt{53 \cdot (\frac{1}{7} + \frac{(x_k - 3{,}143)^2}{10{,}857})}.$$

При $x_k = \bar{x}$

$$m_{\hat{y}_x} = \sqrt{53 \div 7} = 2{,}75.$$

При $x_k = 4$

$$m_{\hat{y}_x} = \sqrt{53 \cdot (\frac{1}{7} + \frac{(4 - 3{,}143)^2}{10{,}857})} = 3{,}34.$$

Для прогнозованого значення $\hat{y}_x$ 95 %-ві довірчі інтервали при $x_k = 4$ визначаються виразом:

$$\hat{y}_{x_k} \pm t_\alpha \cdot m_{\hat{y}_x},$$

тобто $\hat{y}_{x_k} \pm 2{,}57 \cdot 3{,}34$, *або* $\hat{y}_{x_k} \pm 8{,}58$.

При $x_k = 4$ прогнозне значення складе:

$$y_p = -5{,}79 + 36{,}84 \cdot 4 = 141{,}57,$$

яке являє собою точковий прогноз.

Прогноз ліній регресії в інтервалі складе:

$$132{,}99 \le \hat{y}_{x_k} \le 150{,}15.$$

2.1.6. НЕЛІНІЙНА РЕГРЕСІЯ

Якщо між економічними явищами існують нелінійні співвідно-шення, то вони виражаються за допомогою відповідних нелінійних функцій: наприклад, рівносторонньої гіперболи:

$$y = a + \frac{b}{x} + \varepsilon$$

параболи другої степені:

$$y=a+b*x+c*x^2+\varepsilon \text{ та ін.}$$

Розрізняють два класи нелінійних функцій:

А) регресії, нелінійні відносно включених в аналіз пояснювальних змінних, але лінійні за оцінюваними параметрами;

Б) регресії, нелінійні за оцінюваними параметрами.

Прикладом регресії групи А) можуть служити такі функції:

- Поліноми різних степенів: $y=a+b*x+c*x^2+\varepsilon$.
- Рівнобічна гіпербола: $y = a + \frac{b}{x} + \varepsilon$.

До нелінійних регресій за оцінюваними параметрами відносяться функції:

- Степенева: $y=a\cdot x^b\cdot\varepsilon$.
- Показникова $y=a\cdot b^x\cdot\varepsilon$.
- Експоненційна $y=e^{a+bx}\cdot\varepsilon$

Нелінійна регресія за включеними змінними не має ніяких складностей для оцінки її параметрів. Вони визначаються МНК, бо ці функції лінійні за параметрами. Так, у параболі 2-го степеня:

$$y = a_0 + a_1 x + a_2 x^2 + \varepsilon,$$

замінивши $x = x_1$; $x^2 = x_2$, отримуємо двофакторну регресію:

$$y = a_0 + a_1 x_1 + a_2 x_2 + \varepsilon.$$

Відповідно, для поліномів 3-го порядку – трифакторну, 4-го – чотирифакторну тощо.

Застосування МНК для оцінки параметрів параболи другої степені призводить до такої системи нормальних рівнянь:

$$\begin{cases} \sum y = n\cdot a + b\cdot\sum x + c\cdot\sum x^2, \\ \sum y\cdot x = a\cdot\sum x + b\cdot\sum x^2 + c\cdot\sum x^3, \\ \sum y\cdot x^2 = a\cdot\sum x^2 + b\cdot\sum x^3 + c\cdot\sum x^4. \end{cases}$$

Розв’язати її відносно параметрів a, b, c можна методом визначників:

$$a = \frac{\Delta a}{\Delta};\ b = \frac{\Delta b}{\Delta};\ c = \frac{\Delta c}{\Delta},$$

У класі нелінійних функцій, параметри яких без особливих ускладнень оцінюються МНК, добре відома рівнобічна гіпербола:

$$\hat{y}_x = a + \frac{b}{x}.$$

Вона використовується для характеристики зв’язку питомої ваги витрат сировини, матеріалів, палива з обсягом продукції, що випускається, часу обороту товарів з величиною товарооборота не тільки на мікрорівні, але й на макрорівні. Класичним її прикладом є *крива Філіпса*, що характеризує нелінійне співвідношення між нормою безробіття x і процентом росту заробітної плати y.

$$y = a + \frac{b}{x} + \varepsilon$$

Якщо в рівнянні рівносторонньої гіперболи замінити $1/x$ на z, отримуємо звичайне лінійне рівняння, оцінка параметрів якого може бути дана МНК. Система нормальних рівнянь має вигляд:

$$\begin{cases} \sum y = na + b\sum \frac{1}{x}, \\ \sum \frac{y}{x} = a\sum \frac{1}{x} + b\sum \frac{1}{x^2}. \end{cases}$$

Інша справа стоїть з регресією, нелінійною за оцінюваними параметрами. Даний клас моделей поділяється на внутрішні лінійні та внутрішні нелінійні. Якщо ***нелінійна модель внутрішньолінійна,*** то за допомогою відповідних перетворень вона може бути зведена до лінійного виду. Якщо ж ***нелінійна модель внутрішньонелінійна,*** то вона не може бути зведена до лінійної функції. Наприклад, в економетричних дослідженнях при вивченні еластичності попиту від ціни широко використовується степенева функція:

$$y = ax^b\varepsilon$$

y – попит (кількість);

x – ціна;

ε – випадкова похибка.

Дана модель нелінійна відносно оцінюваних параметрів, тому що включає параметри a та b неадитивно. Однак її можна вважати внутрішньолінійною, бо логарифмування даного рівняння за основою e призводить його до лінійного виду:

$$\ln y = \ln a + b \cdot \ln x + \ln \varepsilon$$

Якщо ж модель представити у вигляді $y=a+x^b+e$, то вона стає внутрішньонелінійною, бо її неможливо перетворити в лінійний вид.

Приклади внутрішньонелінійних функцій:

$$y = a + bx^c + \varepsilon$$

$$y = a \cdot (1 - \frac{1}{1 - x^b}) + \varepsilon.$$

Якщо модель внутрішньо нелінійна за параметрами, то для оцінки параметрів використовуються ітераційні процедури, успішність яких залежить від виду рівнянь та особливостей ітераційної процедури. Однак більше поширення отримали моделі, що можуть бути зведені до лінійних в економіці.

ПИТАННЯ ДЛЯ САМОКОНТРОЛЮ

1. Що являє собою проста лінійна регресія? Яке походження терміна «регресія»?
2. У чому відмінність простої регресії від множинної (багато-факторної)?
3. Дайте тлумачення складовим регресійного рівняння.
4. Назвіть основні типи кривих, що використовуються в моделюванні.
5. Які основні методи знаходження параметрів простої лінійної регресії?
6. Які існують способи знаходження параметрів лінійної регресії методом найменших квадратів?
7. Які показники щільності зв'язку між змінними вам відомі?
8. Як вимірюється еластичність між змінними?
9. Що показує коефіцієнт детермінації?
10. Назвіть основні припущення методу найменших квадратів.
11. Які основні тести слід провести для перевірки моделі на адекватність та значущість параметрів регресії?
12. Які види дисперсій обчислюються при аналізі моделі?
13. Як здійснюється точковий та інтервальний прогноз по моделі?
14. Які види нелінійних регресії можна виділити?
15. Які особливості знаходження параметрів нелінійних регресій?

ПРАКТИЧНІ ЗАВДАННЯ

1. Зобразіть графічно степеневу функцію $y = \alpha x^{\beta}$ та надайте їй характеристику, якщо:

а) β – не ціле число і $\beta>1$;

б) β – не ціле число і $0<\beta<1$;

в) $\beta<0$;

г) β=1.

2. Знайдіть перетворення даних, яке зводить дану модель в лінійну:

а) $y = \frac{1}{\alpha + \beta \cdot e^{x}}$;

б) $y = 1 + \alpha e^{\beta x}$.

3. Доведіть, що $\sum(y - \bar{y})^2 = \sum(\hat{y}_x - \bar{y})^2 + \sum(y - \hat{y}_x)^2$.

4. Визначте, які кошти можна виділити на переобладнання, якщо метою переобладнання є підвищення рентабельності. Модель має вигляд:

$$y = -ax^2 + bx + c,\ a>0;\ b>0;\ c>0,$$

де y – рентабельність; x – кошти на переобладнання.

5. Для наведених нижче спостережень обчисліть коефіцієнти кореляції та детермінації, коефіцієнт еластичності. Зробіть висновки щодо тісноти взаємозв'язку між змінними y та x.

Вихідні дані до задачі 5

Y	5	13	12	17	12	22	15	22
X	90	25	42	50	36	35	12	60

6. Наявна інформація про випуск продукції y 10-ти підприємств галузі та розмір їх робочого капіталу x:

Вихідні дані до задачі 6

№ з/п	**Випуск, тис. грн, *y***	**Робочий капітал, тис. грн, *x***
1	48	179
2	50	210
3	35	167
4	68	223
5	89	247
6	77	250

Закінчення таблиці

7	46	140
8	56	175
9	72	234
10	73	199

Необхідно:

1) Побудувати парну лінійну регресію.

2) Знайти суми квадратів відхилень *SST, SSR, SSE* та побудувати таблицю ANOVA-аналізу.

3) Перевірити модель на адекватність за тестом Фішера та на значущість параметрів регресії за тестом Ст'юдента.

4) Побудувати інтервали довіри до параметрів регресії з рівнем надійності 95 %.

5) Визначити, як зросте випуск при збільшенні робочого капіталу на 10 %.

6) Визначити, як зросте випуск при збільшенні робочого капіталу на 23 тис. грн.

ТЕСТИ

1. Найбільш наглядним видом вибору рівняння парної регресії є:
   а) аналітичний;
   б) графічний;
   в) експериментальний (табличний);
   г) немає вірної відповіді.
2. Розраховувати параметри парної лінійної регресії має сенс, якщо в нас є:
   а) не менше 5 спостережень;
   б) не менше 7 спостережень;
   в) не менше 10 спостережень;

г) понад 100 спостережень.

3. Сутність методу найменших квадратів полягає в:
   а) мінімізації суми залишкових величин;
   б) мінімізації дисперсії результативної ознаки;
   в) мінімізації суми квадратів залишкових величин;
   г) мінімізації суми квадратів відхилень фактичних і середнього значення результативної ознаки.
4. Коефіцієнт рівняння простої лінійної регресії:
   а) показує середню зміну результату із зміною фактору на одну одиницю;
   б) оцінює статистичну значущість рівняння регресії;
   в) показує, на скільки відсотків зміниться в середньому результат, якщо фактор зміниться на 1 %.
5. На підставі спостережень за 50 сім'ями побудовано рівняння регресії $\hat{y} = 284{,}56 + 0{,}627 \cdot x$, де y – споживання, x – дохід. Чи відповідають знаки і значення коефіцієнтів регресії теоретичним уявленням?
   а) так;
   б) ні;
   в) нічого певного сказати не можна.
6. Суть коефіцієнта детермінації R^2 полягає в наступному:
   а) оцінює якість моделі з відносних відхилень за кожним спостереженням;
   б) характеризує долю дисперсії результативної ознаки y, що пояснюється регресією, в загальній дисперсії результативної ознаки;
   в) характеризує долю дисперсії y, що викликана впливом не врахованих у моделі факторів.
7. Якість моделі з відносних відхилень за кожним спостережнням оцінює:
   а) коефіцієнт детермінації;
   б) F-критерій Фішера;
   в) коефіцієнт кореляції.
8. Значущість рівняння регресії в цілому оцінює:
   а) F-критерій Фішера;
   б) критерій Ст'юдента;
   в) коефіцієнт детермінації.
9. Класичний метод до оцінки параметрів регресії побудований на:
   а) методі найменших квадратів;
   б) методі максимальної подібності;
   в) кроковому регресійному аналізі.
10. Залишкова сума квадратів дорівнює нулеві:
   а) коли правильно підібрана регресійна модель;
   б) коли між ознаками існує точний функціональний зв'язок;
   в) ніколи.
11. Пояснена (факторна) сума квадратів відхилень у лінійній парній моделі має число ступенів вільності, що дорівнює:
   а) *n-1*;
   б) *1*;
   в) *n-2*.
12. Залишкова сума квадратів відхилень у лінійній простій регресії має число ступенів вільності, що дорівнює:
   а) *n-1*;
   б) *1*;
   в) *n-2*.
13. Загальна сума квадратів відхилень у лінійній парній регресії має число ступенів вільності, що дорівнює:
   а) *n-1*;

б) *1*;
в) *n-2*.

14. Для оцінки значущості коефіцієнтів регресії обчислюють:
а) F-критерій Фішера;
б) t-критерій Ст'юдента;
в) коефіцієнта детермінації R^2.

15. Яке рівняння регресії неможна звести до лінійного виду:
а) $\hat{y}_x = a + b \cdot \ln x$;
б) $\hat{y}_x = a \cdot x^b$;
в) $\hat{y}_x = a + b \cdot x^c$.

16. Яке рівняння є степеневим:
а) $\hat{y}_x = a + b \cdot \ln x$;
б) $\hat{y}_x = a \cdot x^b$;
в) $\hat{y}_x = a + b \cdot x^c$.

17. Параметр *b* у степеневій моделі є:
а) коефіцієнтом детермінації;
б) коефіцієнтом еластичності;
в) коефіцієнтом кореляції.

18. Коефіцієнт кореляції r_{xy} може приймати значення:
а) від – 1 до 1;
б) від 0 до 1;
в) будь-які.

19. Яке з наступних рівнянь нелінійне за оцінюваними параметрами:
а) $y = a + b \cdot x + \varepsilon$;
б) $y = a + b \cdot \ln x + \varepsilon$;
в) $y = a \cdot x^b \cdot \varepsilon$.

20. Лінійна регресія:
а) лінія, що відображає зв'язок між незалежною і залежною змінними;
б) інша назва простої регресії;
в) лінія, яка завжди має нахил, що дорівнює 1;
г) графік значень незалежної і залежної змінних;
д) лінія, яка завжди має нахил, що дорівнює 0.

21. Нахил – це:
а) точка, де лінія регресії перетинає вісь *y*;
б) вимірює придатність лінії регресії;
в) вимірює зв'язок між залежною і незалежною змінними;
г) завжди дорівнює 1;
д) інша назва коефіцієнта детермінації.

22. Перетин:
а) точка, де лінія регресії перетинає вісь *y*;
б) вимірює придатність лінії регресії;
в) вимірює зв'язок між залежною і незалежною змінними;
г) завжди дорівнює 1;
д) завжди дорівнює 0.

23. Що з наведеного не є припущенням моделі лінійної регресії:
а) або x_i є сталими числами, або вони є статистично незалежними від випадкових величин ε_i;
б) дисперсія випадкової величини ε_i є сталою;
в) математичне сподівання випадкової величини ε_i дорівнює нулеві;
г) дисперсія випадкової величини дорівнює 0;

д) випадкові величини є статистично незалежними одна від одної.

24. Коефіцієнт детермінації:
   а) точка, де лінія регресії перетинає вісь *y*;
   б) вимірює придатність лінії регресії;
   в) вимірює зв'язок між незалежною і залежною змінними;
   г) завжди дорівнює 1;
   д) завжди дорівнює 0.

25. Стандартна помилка оцінювання:
   а) точка, де лінія регресії перетинає вісь *y*;
   б) вимірює придатність лінії регресії;
   в) вимірює зв'язок між незалежною і залежною змінними;
   г) завжди дорівнює 1;
   д) інша назва коефіцієнт детермінації.

26. Коефіцієнт детермінації вимірює:
   а) варіацію незалежної змінної;
   б) нахил лінії регресії;
   в) перетин лінії регресії;
   г) загальну варіацію залежної змінної, що пояснюється регресією;
   д) завжди дорівнює 1.

27. SST є:
   а) $\sum(y_i - \bar{y})^2$;
   б) $\sum(\hat{y}_i - \bar{y})^2$;
   в) $\sum(y_i - \hat{y}_i)^2$;
   г) SSR-SST;
   д) SSE-SSR.

28. Коваріація між *x* та *y*:
   а) $\sum(x_i - \bar{x})^2$;
   б) $\sum(y_i - \bar{y})^2$;
   в) $(1/n)\sum(x_i - \bar{x})(y_i - \bar{y})$;
   г) $r_{xy}\sigma_x\sigma_y$;
   д) в і г.

29. Якщо ми хочемо, використовуючи регресійний аналіз, виміряти зв'язок між досвідом роботи і заробітною платою, то:
   а) незалежною змінною має бути заробітна плата;
   б) незалежною змінною має бути досвід роботи;
   в) залежною змінною має бути досвід роботи;
   г) залежною змінною має бути досвід роботи;
   д) б і в.

30. У регресії: *y=0,34+1,2x* нахил дорівнює:
   а) *x*;
   б) *y;*
   в) 0,34;
   г) 1,2;
   д) 1,2/0,34.

31. У регресії: *y=0,34+1,2x* перетин дорівнює:
   а) *x*;
   б) *y;*
   в) 0,34;
   г) 1,2;

д) 1,2/0,34.

32. З урахуванням співвідношення між заробітною платою (в гривнях) – *y* і освітою (в роках) – *x*, *y=12,201+525x*, особа, яка навчалася додатково один рік, може очікувати на таку додаткову оплату:

а) 12,201;
б) 525;
в) 24,402;
г) 1,050;
д) 12,201+525.

33. З урахуванням співвідношення між заробітною платою (в гривнях) – *y* і освітою (в роках) – *x*, *y=12,201+525x*, особа, яка навчалася додатково нуль років, може очікувати на таку додаткову оплату:

а) 12,201;
б) 525;
в) 24,402;
г) 1,050;
д) 12,201+525.

34. Якщо регресія має $R^2=0,8$, то регресійна лінія:

а) пояснює 80 % варіації змінної *x*;
б) пояснює 80 % варіації змінної *y*;
в) матиме нахил 0,8;
г) матиме перетин 0,8;
д) не пояснює зв'язку між *x* та *y*.

35. Якщо нахил регресії становить 2,4 і дисперсія нахилу 0,8, то величина t, що її використовуються для перевірки H_0:$\beta_1=0$, становитиме:

а) 0,8/2,4;
б) 2,4/√0,8;
в) (2,4-1)/ √0,8;
г) 2,4/0,8;
д) (2,4-1)/0,8.

36. За інших рівних умов, чим більша оцінка середньоквадратичного відхилення нахилу, тим:

а) більша *t*-величина нахилу;
б) менша *t*-величина нахилу;
в) більша *t*-величина перетину;
г) менша *t*-величина перетину;
д) більший коефіцієнт нахилу.

37. Яке з поданих тверджень є правильним:

а) *SSE+SSE>SST*;
б) $R^2= -0,5$;
в) $R^2=1,83$;
г) $\sigma_\varepsilon = -0,35$;
д) $t= -2,3$.

38. Для регресії з n спостережень інтервал довіри 1-α % для перетину буде:

а) $b_0 \pm t_{(\alpha,n-2)} \cdot \sigma^2_{b_0}$;
б) $b_0 \pm t_{(\alpha,n-2)} \cdot \sigma^2_{b_1}$;
в) $b_0 \pm t_{(\alpha/2,n-2)} \cdot \sigma_{b_0}$;
г) $b_0 \pm t_{(\alpha/2,n-2)} \cdot \sigma^2_{b_1}$;
д) $b_0 \pm t_{(\alpha/2,n-2)} \cdot \sigma_{b_\varepsilon}$.

39. Припустимо, що залежність витрат від доходу описується функцією: $y=b_0+b_1x$. Середнє значення $y=2$, середнє значення $x=6$, *а* $b_1=3$. Тоді коефіцієнт еластичності витрат від доходу дорівнює:
а) 8;
б) 1;
в) 9;
г) 4.

40. Припустимо, залежність витрат від доходу описується функцією: $\ln y=b_0 + b_1 \ln x$. Середнє значення $y=15$, середнє значення $x=7$, а $b_1=4$. Тоді коефіцієнт еластичності витрат від доходу дорівнює:
а) 38;
б) 1/7;
в) 9;
г) 4.

Частина 2.2. Множинна регресія

Складність економічного життя вимагає використовувати у моделюванні не один фактор впливу на певну економічну величину, а декілька. При цьому ускладнюється й сам аналіз моделі та розрахунок її параметрів. У частині 2.2 подано статистико-математичний апарат, за допомогою якого створюються стохастичні багатофакторні моделі (множинні регресії), здійснюється їх тестування за окремими критеріями. Вивчення цієї теми допоможе студенту-досліднику навчитися створювати економічно й статистично обґрунтовані багатофакторні моделі, знаходити за ними точкові та інтервальні прогнози.

Основні питання, що розглядаються:
1. Відбір факторів для побудови множинної регресії та вибір форми її рівняння
2. Оцінка параметрів рівняння множинної регресії
3. Множинна кореляція
4. Оцінка надійності результатів регресії та кореляції
5. Побудова інтервалів довіри для параметрів регресії та прогнозних значень *y*

Основні терміни:

Специфікація; інтерпретованість параметрів; метод усіх можливих регресій; метод виключень; кроковий регресійний метод; регресія на головні компоненти; ступеневий регресійний аналіз; PRESS-регресія; метод визначників; стандартизовані коефіцієнти регресії; множинна кореляція; скоригований коефіцієнт детермінації; частковий коефіцієнт кореляції; частковий F-критерій.

2.2.1. ВІДБІР ФАКТОРІВ ДЛЯ ПОБУДОВИ МНОЖИННОЇ РЕГРЕСІЇ ТА ВИБІР ФОРМИ ЇЇ РІВНЯННЯ

Проста регресія може дати гарний результат при моделюванні, якщо вплив інших факторів на об’єкт дослідження можна ігнорувати. Наприклад, при побудові моделі споживання того чи іншого товару залежно від рівня доходу дослідник передбачає, що в кожній групі доходу однаковий вплив на споживання таких факторів, як ціна товару, розмір сім’ї та її склад. Разом з тим дослідник ніколи не може бути впевнений у справедливості даного припущення. Для того, щоб мати правильне уявлення про вплив доходу на споживання, необхідно вивчити їх кореляцію при незмінному рівні інших факторів. Для дослідника-фізика, хіміка, біолога цей процес провести легше, оскільки вони часто можуть контролювати величини факторів, їх поведінку. Дослідник-економіст цього робити не може, тому одним з шляхів подолання таких труднощів є можливість виявити вплив інших факторів, вводячи їх у модель, тобто побудувати рівняння множинної регресії

$$y = a + b_1 \cdot x_1 + b_2 \cdot x_2 + ... + b_p x_p + \varepsilon .$$

Множинна регресія широко використовується в розв'язанні проблем попиту, доходності акцій, при вивченні функцій витрат, у макроекономічних розрахунках та цілого ряду інших питань. Основна мета множинної регресії – побудувати модель з великою кількістю факторів, визначивши при цьому вплив кожного з них окремо, а також сукупну дію на результативний показник.

Побудова моделі починається із ***специфікації***. Вона передбачає вирішення двох проблем – відбору факторів та вибору виду рівняння регресії.

Фактори, що включаються в множинну регресію повинні відповідати таким вимогам:

1) бути кількісно вимірюваними. Якщо необхідно включити в модель фактор, що має якісний вимір, то треба надати йому кількісної оцінки (наприклад, у вигляді балів, *dummy*-змінних, рангів);

2) не повинні корелювати між собою і тим більше знаходитися у функціональній залежності. Включення до моделі факторів з високою інтеркореляцією, коли коефіцієнт кореляції між фактором x_1 та y_1 менше, ніж коефіцієнт кореляції між факторами x_1 та x_2, це може призвести до небажаних наслідків – система нормальних рівнянь може виявитися погано зумовленою і мати наслідком нестійкість та ненадійність оцінок коефіцієнтів регресії. Якщо між факторами існує висока кореляція, то неможливо визначити правильно їх вплив окремо і параметри регресії будуть *неінтерпретовані.*

Незважаючи на те, що теоретично регресійна модель дозволяє врахувати будь-яку кількість факторів, практично в цьому немає необхідності. Відбір факторів відбувається на основі якісного теоретико-економічного аналізу. Однак теоретичний аналіз часто не дозволяє однозначно відповісти на питання про кількісний взаємозв'язок розгляданих ознак і доцільності включення фактору в модель. Тому відбір факторів зазвичай проводять у дві стадії:

– на першій відбираються фактори, виходячи із суті проблеми;

– на другій на основі матриці показників кореляції і визначення *t*-статистики для параметрів регресії обираються придатні для побудови моделі параметри.

Серед методів, які дозволяють відібрати найкращі фактори, виділяють такі, як ***метод усіх можливих регресій, метод виключень, крокový регресійний метод, регресія на головні компоненти, ступеневий регресійний аналіз, PRESS-регресія тощо.***

Для вибору регресії існує два протилежних критерії:

1) якщо ми хочемо зробити модель корисною для прогнозу, маємо включити якомога більше факторів для того, щоб визначення величин, які прогнозуються, було надійнішим.

2) оскільки отримання інформації з послідовним контролем при великій кількості змінних *x* потребує великих витрат, слід прагнути, щоб модель включала якомога менше факторів *x*.

Компромісом між цими крайнощами є те, що називають вибором «найкращого рівняння регресії». Для реалізації такого вибору немає єдиної статистичної процедури, тому в кожному окремому випадку застосовуються різні методи. Розглянемо їх більш детально.

А) Метод усіх можливих регресій.

Це історично перший метод побудови регресії. Він дуже громіздкий і може бути найкраще реалізований на ЕОМ. Метод потребує побудови кожного з усіх можливих регресійних рівнянь, які обов'язково включають член b_0. Оскільки для кожного фактора x_i є дві можливості – бути включеним або невключеним у регресію, то всього буде 2^p рівнянь (усі можливі одно-, двох-, ... *p*-факторні моделі, де p – кількість факторів x_i, $i = \overline{1, p}$). Кожне з рівнянь потім оцінюється за допомогою трьох критеріїв: R^2, *MSE* та C_p – статистики (критерію Малоуза).

Б) Метод виключень

Цей метод економічніший, ніж метод усіх регресій. Загальний алгоритм складається з 5 етапів. На першому етапі розраховується регресійне рівняння, що включає всі фактори моделі. На другому етапі обчислюється величина часткового F-критерію. Найменше значення F-критерію позначається як F_l і порівнюється з критичним значенням. За умови

підтвердження нуль-гіпотези H_0 фактор виключається з рівняння, в іншому випадку – залишається. Аналогічно проводяться всі подальші дії аж доти, доки найменше значення F-критерію не буде більшим за його критичне значення. Іноді замість F-критерію використовується *t*-критерій, який є коренем квадратним від значення F-критерію часткового.

В) Крроковий регресійний метод

Крроковий регресійний метод діє в зворотному порядку порівняно з методом виключень. Фактори по черзі включаються в модель доти, поки вона не стане задовільною. Порядок включення вибирається за допомогою коефіцієнту кореляції фактору *x* з фактором *y* як міри важливості факторів, які ще не включені в модель. Алгоритм такий: обирається фактор, що має найбільший коефіцієнт кореляції із залежною змінною, будується регресійна модель з однією незалежною змінною. Після цього перевіряється, чи буде значущим F-критерій. Якщо ні, то приймаємо, що $y = \bar{y}$. І припиняємо процес. В іншому випадку включаємо другий за значущістю фактор і т.п.

З іншими методами пропонується ознайомитися самостійно.

Як і в парній регресії, використовуються різні форми рівнянь множинної регресії: лінійні та нелінійні.

Найчастіше використовуються лінійні та степеневі функції (функція Кобба-Дугласа). В даному випадку зберігається той самий принцип, що й для однофакторної регресії – існують нелінійні рівняння, які можна звести до форми лінійних та такі, що не можна. Зведення степеневої функції до лінійної форми відбувається через процес логарифмування.

Приклад: Візьмемо виробничу функцію $Y = A \cdot K^{\alpha} \cdot L^{\beta}$. Лінеаризація цього рівняння матиме такий вигляд: $\lg Y = A + \alpha \lg K + \beta \lg L$.

2.2.2. ОЦІНКА ПАРАМЕТРІВ РІВНЯННЯ МНОЖИННОЇ РЕГРЕСІЇ

Параметри рівняння множинної регресії оцінюються, як і в парній регресії, методом найменших квадратів. При його застосуванні будується система нормальних рівнянь, розв'язання якої і дозволяє отримати оцінки параметрів регресії.

Так, для рівняння виду

$$y = a + b_1 \cdot x_1 + b_2 \cdot x_2 + ... + b_p x_p + \varepsilon$$

система нормальних рівнянь становитиме:

$$\begin{cases} \sum y = n \cdot a + b_1 \cdot \sum x_1 + b_2 \cdot \sum x_2 + ... + b_p \cdot \sum x_p \\ \sum y \cdot x_1 = a \cdot \sum x_1 + b_1 \cdot \sum x_1^{\,2} + b_2 \cdot \sum x_1 \cdot x_2 + ... + b_p \cdot \sum x_p \cdot x_1 \\ ..... \\ \sum y \cdot x_p = a \cdot \sum x_p + b_1 \cdot \sum x_1 \cdot x_p + b_2 \cdot \sum x_2 \cdot x_p + ... + b_p \cdot \sum x^2{}_p \end{cases}$$

Її розв'язок найкраще здійснюється матричним методом, ***методом визначників***:

$$a = \frac{\Delta a}{\Delta}, \quad b_1 = \frac{\Delta b_1}{\Delta}, ..., \quad b_p = \frac{\Delta b_p}{\Delta},$$

де Δ – визначник системи;

Δa, Δb_1, ... Δb_p – часткові визначники.

При цьому:

$$\Delta = \begin{vmatrix} n & \sum x_1 & \sum x_2 & \dots & \sum x_p, \\ \sum x_1 & \sum x_1^2 & \sum x_2 x_1 & \dots & \sum x_p x_1, \\ \sum x_2 & \sum x_1 x_2 & \sum x_2^2 & \dots & \sum x_p x_2, \\ \dots & \dots & \dots & \dots & \dots \\ \sum x_p & \sum x_1 x_p & \sum x_2 x_p & \dots & \sum x_p^2, \end{vmatrix}$$

а Δa, Δb_1, ..., Δb_p отримуються шляхом заміни відповідного стовпчика матриці визначника системи даними лівої частини системи.

Рівняння множинної лінійної регресії у матричній формі має вигляд:

$$Y=XB+e,$$

де $Y=\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix}$, $X=\begin{pmatrix} 1 & x_{11} & \dots & x_{k1} \\ 1 & x_{12} & \dots & x_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \dots & x_{kn} \end{pmatrix}$; $B=\begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{pmatrix}$; $E=\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$

Процедура оцінки параметрів та сама, що й у парній регресії, тобто, для знаходження вектору B отримуємо:

$$B=(X^TX)^{-1}X^TY$$

Приклад.

Наявні такі дані по 10 підприємствах концерну щодо прибутку (y – млн грн), продукції на одного робітника (x_1 – одиниць) та частки продукції, виробленої на експорт (x_2 – %), що наведені в таблиці.

№	y	x_1	x_2	y^2	x_1^2	x_2^2	yx_1	yx_2	x_1x_2
1	2	11	3	4	121	9	22	6	33
2	1	10	2	1	100	4	10	2	20
3	3	12	4	9	144	16	36	12	48
4	8	18	10	64	324	100	144	80	180
5	7	15	11	49	225	121	105	77	165
6	5	13	6	25	169	36	65	30	78
7	4	13	5	16	169	25	52	20	65
8	6	15	7	36	225	49	90	42	105
9	7	16	10	49	256	100	112	70	160
10	7	17	12	49	289	144	119	84	204
Разом	50	140	70	302	2022	604	755	423	1 058

Система нормальних рівнянь має вигляд:

$$\begin{cases} 10a+140b_1+70b_2=50 \\ 140a+2022b_1+1058b_2=755 \\ 70a+1058b_1+604b_2=423. \end{cases}$$

Розв’язуємо цю систему методом визначників:

$$\Delta=9\,840,\ \Delta_a=-4\,760,\ \Delta b_1=5\,760,\ \Delta b_2=2\,360,$$

звідки:

$$a=-4{,}874:\ b_1=0{,}585:\ b_2=0{,}240.$$

Рівняння регресії виглядає таким чином:

$$y=-4{,}874+0{,}585x_1+0{,}240x_2+\varepsilon.$$

У матричному вигляді оцінка параметрів регресії складає:

$$X^TX=\begin{pmatrix} 1 & 1 & \dots & 1 \\ 11 & 10 & \dots & 17 \\ 3 & 2 & \dots & 12 \end{pmatrix}\begin{pmatrix} 1 & 11 & 3 \\ 1 & 10 & 2 \\ \dots & \dots & \dots \\ 1 & 17 & 12 \end{pmatrix}=\begin{pmatrix} 10 & 140 & 70 \\ 140 & 2022 & 1058 \\ 70 & 1058 & 604 \end{pmatrix}.$$

$$X^TY=\begin{pmatrix} 1 & 1 & \dots & 1 \\ 11 & 10 & \dots & 17 \\ 3 & 2 & \dots & 12 \end{pmatrix}\begin{pmatrix} 2 \\ 1 \\ \dots \\ 7 \end{pmatrix}=\begin{pmatrix} 50 \\ 755 \\ 423 \end{pmatrix}.$$

Обернена матриця визначається як:

$$A^{-1}=(X^TX)^{-1},\ A^{-1}=\frac{1}{|A|}\overline{A},$$

де $|A|$ – визначник матриці $X^T X$; $\overline{A}$ – матриця, приєднана до матриці $X^T X$, елементи якої одержуються як $a_{ij} = (-1)^{i+j}\left|M_{ij}\right|$; M_{ij} – матриця, отримана з А шляхом викреслення і-го рядку та j-го стовпця.

У нашому прикладі $|A| = 9\,840$.

$$A^{-1} = \frac{1}{9840}\begin{pmatrix} 101924 & -10\,500 & 6\,580 \\ -10\,500 & 1140 & -780 \\ 6\,580 & -780 & 620 \end{pmatrix}.$$

Потім, помноживши цю матрицю на вектор $X^T Y$, отримаємо оцінки параметрів регресії:

$$b = \frac{1}{9840}\begin{pmatrix} -47\,960 \\ 5\,760 \\ 2\,360 \end{pmatrix} = \begin{pmatrix} -4{,}874 \\ 0{,}585 \\ 0{,}240 \end{pmatrix},$$

що співпадає з отриманими раніше оцінками.

$$a = -4{,}874;\ \ b_1 = 0{,}585;\ \ b_2 = 240.$$

Можливий інший підхід до визначення параметрів множинної регресії, коли на основі матриці парних коефіцієнтів кореляції будується рівняння регресії у стандартизованому масштабі:

$$t_y = \beta_1 t_{x1} + \beta_2 t_{x2} + ... + \beta_p t_{xp} + \varepsilon,$$

де $t_y, t_{x1}, ..., t_{xp}$ – стандартизовані змінні: $t_y = \frac{y - \bar{y}}{\sigma_y}$, $t_{x1} = \frac{x_i - \bar{x}_i}{\sigma_{x_j}}$, для яких середнє квадратичне відхилення дорівнює одиниці: $\sigma_{t_y} = \sigma_{t_x} = 1$; β – стандартизовані коефіцієнти регресії.

Застосувавши МНК до рівняння множинної регресії у стандартизованому масштабі, після відповідних перетворень отримаємо систему нормальних рівнянь вигляду:

$$\begin{cases} r_{yx_1} = \beta_1 + \beta_2 r_{x_2x_1} + \beta_3 r_{x_3x_1} + ... + \beta_p r_{x_px_1}, \\ r_{yx_2} = \beta_1 r_{x_2x_1} + \beta_2 + \beta_3 r_{x_3x_2} + ... + \beta_p r_{x_px_2}, \\ ..... \\ r_{yx_p} = \beta_1 r_{x_px_1} + \beta_2 r_{x_px_2} + \beta_3 r_{x_px_3} + ... + \beta_p. \end{cases}$$

Розв'язуємо її методом визначників, знаходимо параметри – стандартизовані коефіцієнти регресії (β-коефіцієнти).

Стандартизовані коефіцієнти регресії показують, на скільки σ_y зміниться в середньому результат, якщо відповідний фактор x_i зміниться на одну σ_y за незмінного середнього рівня інших факторів. У силу того, що всі змінні задані як центровані і нормовані, стандартизовані коефіцієнти регресії β_p можна порівняти між собою. Порівнюючи їх між собою, можна ранжувати фактори за силою їх впливу на результат. В цьому основна перевага стандартизованих коефіцієнтів регресії на відміну від коефіцієнтів «чистої» регресії, які не можна порівняти між собою.

2.2.3. МНОЖИННА КОРЕЛЯЦІЯ

Практична значущість рівняння множинної регресії оцінюється за допомогою показника множинної кореляції та його квадрату – коефіцієнту детермінації.

Показник ***множинної кореляції*** характеризує тісноту сукупного впливу факторів *x* на *y*. Незважаючи на форму зв'язку, показник множинної кореляції може бути знайдений як індекс множинної кореляції:

$$R_{yx_1x_2...x_p} = \sqrt{1 - \frac{\sigma^2_{зал}}{\sigma^2_y}},$$

де $\sigma^2_{зал}$ – залишкова дисперсія для рівняння $y = f(x_1, x_2, ..., x_p)$;

σ^2_y – загальна дисперсія результативної ознаки.

Границі зміни цього показника – від 0 до 1.

Розрахунок індексу множинної кореляції передбачає визначення рівняння множинної регресії і на його основі залишкової дисперсії:

$$\sigma^2_{зал} = \frac{\sum(y - \hat{y}_{x_1x_2})^2}{n}$$

$$\sigma^2_{y} = \frac{\sum(y - \bar{y})^2}{n}$$

Низьке значення коефіцієнта множинної кореляції означає, що в регресійну модель не включені суттєві фактори, з одного боку, а з іншого боку, форма зв'язку, що розглядається, не відображає реальні відношення між змінними, включеними до моделі. Тому вимагаються подальші дослідження щодо якості моделі і збільшення її практичної значущості.

Коефіцієнт детермінації R^2

Одна з проблем, пов'язана з використанням R^2 для оцінки моделей (а саме – його збільшення при включенні в модель додаткових змінних), розв'язується шляхом коригування коефіцієнту на зменшення кількості ступенів вільності в результаті появи в моделі додаткових змінних.

Коефіцієнт, що отримується в результаті, називається ***скоригованим коефіцієнтом детермінації*** і позначається $\overline{R}^2$. Визначається він таким чином:

$$\overline{R}^2 = 1 - \left[\frac{n-1}{n-p}(1-R^2)\right],$$

де n – кількість спостережень;

p – кількість незалежних змінних.

Як видно з формули, при додаванні змінних $\overline{R}^2$ буде збільшуватися тільки в тому випадку, якщо зростання R^2 буде «переважати» збільшення кількості змінних. Таким чином, скоригований коефіцієнт детермінації можна використовувати в якості критерію для прийняття рішення про включення або невключення в модель додаткових змінних.

У статистичних пакетах прикладних програм в процедурі множинної регресії наводиться або скоригований коефіцієнт (індекс) множинної кореляції (детермінації) або і звичайний, і скоригований (нормований). Можливості пакету *Data Analysis* в програмному додатку *Excel* дозволяють розрахувати як звичайний, так і скоригований коефіцієнт детермінації.

Часткові коефіцієнти кореляції

Крім загального коефіцієнту кореляції можуть розраховуватися також часткові коефіцієнти кореляції, які є критеріями відбору факторів: доцільність включення того чи іншого фактору в модель доводиться величиною показника часткової кореляції.

Часткові коефіцієнти кореляції характеризують тісноту зв'язку між результатом і відповідним фактором при усуненні впливу інших факторів, включених в рівняння регресії.

Показники часткової кореляції являють собою відношення скорочення залишкової дисперсії за рахунок додаткового включення в аналіз нового фактору до залишкової дисперсії, що мала місце до введення його в модель.

Приклад.

Припустимо, що залежність об'єму продукції y від витрат праці x_1 характеризується рівнянням:

$$\hat{y}_{x_1} = 27{,}5 + 3{,}5x_1;\ r_{yx_1} = 0{,}58.$$

Підставивши в це рівняння фактичні значення x_1, знайдемо теоретичні величини обсягу продукції $\hat{y}_{x_1}$ і відповідну величину залишкової дисперсії MSE:

$$MSE_{yx_1} = \frac{\sum (y_i - \hat{y}_{x_1})^2}{n}.$$

Включивши в рівняння регресії додатковий фактор x_2 – технічну оснащеність виробництва, отримаємо рівняння регресії вигляду:

$$\hat{y}_{x_1x_2} = 20{,}2 + 2{,}8x_1 + 0{,}2x_2.$$

Для цього рівняння залишкова дисперсія менша. Припустимо, що $MSE_{yx_1x_2} = 3{,}7$, *а* $MSE_{yx_1} = 6$. Чим більше число факторів, що включено в модель, тим менша величина залишкової дисперсії.

Скорочення залишкової дисперсії за рахунок додаткового включення фактора x_2 складає:

$$MSE_{yx_1} - MSE_{yx_1x_2} = 2{,}3.$$

Чим більша частка цього скорочення в залишковій варіації до введення додаткового фактору, тобто в MSE_{yx_1}, тим більш тісний зв'язок між y та x_2 при постійній дії фактора x_1. Квадратний корінь з цієї величини є індексом (коефіцієнтом) часткової кореляції, що показує в «чистому» вигляді тісноту зв'язку y з x_2.

Таким чином, чистий вплив фактора x_2 на результативну ознаку y можна знайти як:

$$r_{yx_2 \cdot x_1} = \sqrt{\frac{MSE_{yx_1} - MSE_{yx_1x_2}}{MSE_{yx_1}}}$$

Знак «крапка» у виразі часткового коефіцієнту кореляції $r_{yx2 \bullet x1}$ означає елімінування тієї змінної (змінних), яка стоїть після знаку «крапка».

Аналогічно визначається і чистий вплив на результат y фактора x_1:

$$r_{yx_1 \cdot x_2} = \sqrt{\frac{MSE_{yx_2} - MSE_{yx_1x_2}}{MSE_{yx_2}}}$$

Якщо припустити, що $MSE_{yx_2} = 5$, то часткові показники кореляції для рівняння $\hat{y}_{x_1x_2} = 20{,}2 + 2{,}8x_1 + 0{,}2x_2$ складуть:

$$r_{yx_1 \cdot x_2} = \sqrt{\frac{5 - 3{,}7}{5}} = 0{,}51$$

та

$$r_{yx_2 \cdot x_1} = \sqrt{\frac{6 - 3{,}7}{6}} = 0{,}619.$$

Порівнюючи отримані результати, бачимо, що більш сильний вплив на обсяг продукції спричиняє технічна оснащеність підприємств.

Якщо виразити залишкову дисперсію через показник детермінації $MSE_{зал} = \sigma_y^2(1 - r^2)$, то формула коефіцієнта часткової кореляції матиме вигляд:

$$r_{yx_1 \cdot x_2} = \sqrt{\frac{MSE_{yx_2} - MSE_{yx_1x_2}}{MSE_{yx_2}}} = \sqrt{1 - \frac{MSE_{yx_1x_2}}{MSE_{yx_2}}} = \sqrt{1 - \frac{1 - R^2_{yx_1x_2}}{1 - r^2_{yx_2}}}.$$

Відповідно,

$$r_{yx_2 \cdot x_1} = \sqrt{1 - \frac{1 - R^2_{yx_1x_2}}{1 - r^2_{yx_1}}}.$$

Розглянуті показники часткової кореляції називаються коефіцієнтами часткової кореляції першого порядку, бо вони фіксують тісноту зв'язку двох змінних при закріпленні (елімінуванні впливу) одного фактору.

Якщо розглядається регресія з числом факторів *p*, то можливі часткові коефіцієнті кореляції не тільки першого, але й другого, третього, ..., (*p-1*)-го порядку, тобто вплив фактору x_1 можна оцінити за різних умов незалежності дії інших факторів.

У загальному вигляді при наявності *p* факторів для рівняння:

$$Y=a+b_1*x_1+b_2*x_2+\ldots+b_p*x_p+e$$

коефіцієнт часткової кореляції, що вимірює вплив на у фактору x_i при незмінному рівні інших факторів, можна визначити за формулою:

$$r_{yx_i\cdot x_1x_2\ldots x_{i-1}x_{i+1}\ldots x_p}=\sqrt{1-\frac{1-R^2_{yx_1x_2\ldots x_i\ldots x_p}}{1-R^2_{yx_1\cdot x_2\ldots x_{i-1}x_{i+1}\ldots x_p}}},$$

де $R^2_{yx_1x_2\ldots x_i\ldots x_p}$ – множинний коефіцієнт детермінації всього комплексу *p* факторів з результатом;

$R^2_{yx_1\cdot x_2\ldots x_{i-1}x_{i+1}\ldots x_p}$ = той же показник детермінації, але без введення в модель фактору x_i.

2.2.4. ОЦІНКА НАДІЙНОСТІ РЕЗУЛЬТАТІВ РЕГРЕСІЇ ТА КОРЕЛЯЦІЇ

Перш, ніж знайти значущість рівняння множинної регресії в цілому, необхідно скласти ANOVA-таблицю:

Джерела варіації	**Кількість ступенів вільності, *df***	**Сума квадратів, SS**	**Дисперсія на один ступінь вільності, MS**	**Фактичне значення розподілу Фішера**	**Критичне значення критерію Фішера**
За рахунок регресії	p	$SSR=\sum_{i=1}^{n}(\hat{y}-\bar{y})^2$	$MSR=\sum_{i=1}^{n}(\hat{y}-\bar{y})^2/p$		
Залишкова	n- p-1	$SSE=\sum_{i=1}^{n}(y-\hat{y})^2$	$MSR=\frac{\sum_{i-1}^{n}(y-\bar{y})^2}{(n-p-1)}$		
Загальна	n-1	$SST=\sum_{i=1}^{n}(y-\bar{y})^2$			

Значущість рівняння множинної регресії в цілому, так само, як і в парній регресії, оцінюється за допомогою F-критерію Фішера:

$$F=\frac{MSR}{MSE}=\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2/p}{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2/(n-p-1)}=\frac{R^2}{1-R^2}\cdot\frac{n-p-1}{p},$$

де MSR – факторна сума квадратів на один ступінь вільності;

R^2 – коефіцієнт множинної детермінації;

n – кількість спостережень;

p – кількість факторів, які увійшли в модель;

MSE – залишкова сума квадратів на один ступінь вільності.

Критичне значення F-критерію відображено в таблиці F-розподілу Фішера за такими параметрами: $F_{p,\ (n-p-1)}$. Якщо модель двофакторна і для її побудови ми використали 10 спостережень, то критичне значення коефіцієнту Фішера становитиме: $F_{2,\ 7}$ = 4,74.

Мірою для оцінки включення фактору в модель служить ***частковий F-критерій***, тобто F_{xi}.

Частковий F-критерій побудований на порівнянні приросту факторної дисперсії, зумовленого впливом додатково включеного фактору, з залишковою дисперсією на один ступінь вільності за регресійною моделлю в цілому. Припустимо, що оцінюється

значущість впливу x_1, як додатково включеного в модель фактору. Використовуємо таку формулу:

$$F_{x_1} = \frac{R^2_{yx_1x_2...xp} - R^2_{yx_2...x_p}}{1 - R^2_{yx_1x_2...x_p}} \cdot \frac{n-p-1}{1},$$

де $R^2_{yx_1x_2...x_p}$ – коефіцієнт множинної детермінації для моделі з повним набором факторів;

$R^2_{yx_2...x_p}$ – той же показник, але без включення в модель фактору x_1;

n – число спостережень;

p – число факторів у моделі.

Якщо оцінюємо значущість впливу фактору x_p після включення в модель факторів x_1, x_2, ..., x_{p-1}, то формула часткового F-критерію матиме вигляд:

$$F_{x_p} = \frac{R^2_{yx_1x_2...x_p} - R^2_{yx_1x_2...x_{p-1}}}{1 - R^2_{yx_1x_2...x_p}}$$

У загальному вигляді для фактору x_i частковий F-критерій визначиться як:

$$F_{xi} = \frac{R^2_{yx_1...x_i...x_p} - R^2_{yx_1...x_{i-1}x_{i+1}...x_p}}{1 - R^2_{yx_1...x_i...x_p}} \cdot \frac{n-p-1}{1}.$$

У чисельнику цих формул показано приріст частки поясненої варіації y за рахунок додаткового включення в модель відповідного фактору:

$$R^2_{yx_1x_2...x_p} - R^2_{yx_2...x_p} - \textit{за рахунок } x_1;$$

$$R^2_{yx_1x_2...x_p} - R^2_{yx_1x_2...x_{p-1}} - \textit{за рахунок } x_p;$$

$$R^2_{yx_1...x_i...x_p} - R^2_{yx_1...x_{i-1}x_{i+1}...x_p} - \textit{ за рахунок } x_i.$$

Фактичне значення F-критерію порівнюється з табличним, і якщо воно перевищує табличне, то наявність відповідного фактору в моделі статистично виправдано і коефіцієнт b_i при кожному факторі x_i є статистично значущим. Якщо ні, то включення відповідного фактору є недоцільним.

Оцінка значущості коефіцієнтів чистої регресії по t-критерію Ст'юдента може бути проведена без розрахунку часткових F-критеріїв. У цьому випадку, як і в парній регресії, для кожного фактору використовується формула:

$$t_{b_i} = \frac{b_i}{m_{b_i}}$$

де b_i – коефіцієнт чистої регресії при факторі x_i;

m_{bi} – середня квадратична помилка коефіцієнту регресії b_i.

Для рівняння множинної регресії

$$\hat{y} = a + b_1x_1 + b_2x_2 + ...b_px_p$$

середня квадратична помилка коефіцієнта регресії може бути визначена за такою формулою:

$$m_{b_i} = \frac{\sigma_y\sqrt{1 - R^2_{yx_1...x_p}}}{\sigma_{x_i}\sqrt{1 - R_{x_1x_i...x_p}}} \cdot \frac{1}{\sqrt{n-p-1}},$$

де σ_y – середнє квадратичне відхилення для признаку y;

$R^2_{yx_1...x_p}$ – коефіцієнт детермінації для рівняння множинної регресії;

σ_{x_i} – середнє квадратичне відхилення для признаку x_i;

$R^2_{x_1x_i...x_p}$ – коефіцієнт детермінації для залежності фактора x_i зі всіма іншими факторами рівняння множинної регресії;

$n-p-1$ – число ступенів вільності для залишкової суми квадратів відхилень.

Як бачимо, для того, щоб скористатися цією формулою, необхідні матриця міжфакторної кореляції і розрахунок за нею відповідних коефіцієнтів детермінації $R^2_{xix1\ldots xp}$.

Разом з тим, якщо врахувати, що

$$b_i = \frac{\sigma_y}{\sigma_{x_i}} \cdot \sqrt{\frac{R^2_{yx_1\ldots x_p} - R^2_{yx_1\ldots x_{i-1}x_{i+1}\ldots x_p}}{1 - R^2_{x_1x_i\ldots x_p}}},$$

то можна переконатися, що

$$t_{b_i} = \frac{b_i}{m_{b_i}} = \sqrt{F_{x_i}}.$$

2.2.5. ПОБУДОВА ІНТЕРВАЛІВ ДОВІРИ ДЛЯ ПАРАМЕТРІВ РЕГРЕСІЇ ТА ПРОГНОЗНИХ ЗНАЧЕНЬ *Y*

Для того, щоб визначити, як же знайдені оцінки параметрів багатофакторної регресії пов'язані з параметрами узагальненої регресії, потрібно побудувати інтервали довіри для параметрів.

Процедура побудови інтервалів довіри є аналогічною процедурі тестування нуль-гіпотези для параметрів за t-тестом Ст'юдента, який ми розглянули вище.

Спочатку необхідно розрахувати стандартну похибку m_b кожного параметра та оцінити параметри щодо інтервалів довіри $b_i \pm t m_b$ для $t_{\alpha/2}$ з $(n\text{-}k)$ ступенями вільності. Відповідно, інтервали довіри будуть розраховуватися як:

$$b_i - t_{\alpha/2} m_{bi} \le b_i \le b_i + t_{\alpha/2} m_{bi}$$

Якщо побудована регресійна модель адекватна за F-критерієм Фішера, її можна використовувати для прогнозу залежної змінної. Припустимо, що нам відомі значення факторів в $(n+t)$ період, тоді ми можемо отримати прогнозне значення y, якщо підставимо в рівняння багатофакторної регресії прогнозні значення x.

Теорія прогнозування дає змогу отримати точкові та інтервальні прогнози. Точкові прогнози – на основі підстановки, інтервальні – на основі побудови інтервалів довіри.

Для того, щоб отримати інтервальний прогноз математичного сподівання залежної змінної, розглянемо, чому дорівнює дисперсія цієї величини. Можна показати, що

$$var(E(y_j/x_j) = var\,(\tilde{y}_j/x_j) = \sigma^2_e x'_j (X'X)^{-1} x_j,$$

де σ^2_e – дисперсія випадкової величини ε;

x'_j – вектор значень з p факторів y період j. Виходячи з цього, інтервали довіри для 100 (1-α) рівня довіри математичного сподівання y дорівнюватимуть:

$$\hat{y}_i - t_{\alpha/2}\sqrt{\hat{\sigma}^2 x_j (X'X)^{-1} x_j} \le E(y/x_j) \le \hat{y}_j + t_{\alpha/2}\sqrt{\hat{\sigma}^2 x_j (X'X)^{-1} x_{j.}}$$

Для прогнозного значення y формула для дисперсії буде такою:

$$\mathrm{var}(y_j / x_j) = \sigma^2_\varepsilon (1 + x'_j (X'X)^{-1} x_j).$$

Звідси дещо змінюється формула для обчислення інтервалу довіри, а саме:

$$\hat{y}_i - t_{\alpha/2}\sqrt{\hat{\sigma}^2 (1 + x_j (X'X)^{-1} x_j)} \le y_j \le \hat{y}_j + t_{\alpha/2}\sqrt{\hat{\sigma}^2 (1 + x_j (X'X)^{-1} x_j)}.$$

ПИТАННЯ ДЛЯ САМОКОНТРОЛЮ

1. Назвіть основні вимоги до побудови багатофакторних моделей.
2. Охарактеризуйте методи відбору факторів для множинної регресії.
3. У чому сутність методу визначників?
4. Розкрийте поняття «множинна кореляція», «частковий коефіцієнт кореляції». Що показує коефіцієнт детермінації у багатофакторній регресії?
5. Які особливості проведення тестів Фішера та Ст'юдента для множинної моделі?
6. Як знаходяться інтервали довіри для прогнозного значення в багатофакторній регресії?

ПРАКТИЧНІ ЗАВДАННЯ

1. Визначте кількість ступенів вільності *df* для сум квадратів відхилень *SST, SSR, SSE* лінійної множинної регресії, побудованої на 15 спостереженнях, кількість факторів у моделі $p = 3$.

2. Визначити, яку розмірність мають матриці X, $X'X$, $(X'X)^{-1}$, e, $e'e$, якщо матриця спостережень побудована для 30 спостережень з 5 факторами.

3. Визначити, чи адекватна модель, якщо

$$\sum_{i=1}^{n} (y_i - \hat{y})^2 = 150 \qquad \sum_{i=1}^{n} (y_i - \bar{y})^2 = 220 \cdot$$

Вихідні дані: для побудови моделі використано 20 спостережень, містить 4 фактори.

1. Знайдіть параметри регресії за даними, що наведено у таблиці, двома способами – за допомогою системи нормальних рівнянь та матричним.

Вихідні дані до задачі 4

№ з/п	y	x_1	x_2
1	20	4,8	79
2	22	4,3	96
3	24	3,9	85
4	18	2,6	74
5	29	6,6	93
6	19	3,3	67
7	25	5,8	88
8	27	6,1	89
9	27	6,2	76
10	21	4,9	72

Необхідно:

1) перевірити модель на адекватність за допомогою тесту Фішера;

2) перевірити значущість параметрів регресії за допомогою тесту Ст'юдента;

3) обчислити коефіцієнти еластичності;

4) обчислити часткові коефіцієнти кореляції та зробити висновок щодо доцільності включення в модель факторів x_1 та x_2;

5) спрогнозувати точкове значення y, якщо x_1 прийме значення 8,0, а x_2 значення 73. Знайти інтервали довіри для прогнозного значення.

ТЕСТИ

1. Найбільш трудомістким методом відбору факторів множинної регресії є:
   а) метод усіх можливих регресій;
   б) метод виключень;
   в) кроковий регресійний метод.
2. Додавання в рівняння множинної регресії нової пояснювальної змінної:
   а) зменшує значення коефіцієнта детермінації;
   б) збільшує значення коефіцієнта детермінації;
   в) не спричиняє ніякого впливу на коефіцієнт детермінації.
3. Із збільшенням числа пояснювальних змінних скоригований коефіцієнт детермінації:
   а) збільшується;
   б) зменшується;
   в) не змінюється.
4. Число ступенів вільності для залишкової суми квадратів у лінійній моделі множинної регресії дорівнює:
   а) $n-1$;
   б) m;
   в) $n-m-1$.
5. Число ступенів вільності для загальної суми квадратів у лінійній моделі множинної регресії дорівнює:

а) *n-1*;
б) *m*;
в) *n-m-1*.

6. Число ступенів вільності для факторної суми квадратів у лінійній моделі множинної регресії дорівнює:
   а) *n-1*;
   б) *m*;
   в) *n-m-1*.
7. Скоригований коефіцієнт детермінації:
   а) менше звичайного коефіцієнта детермінації;
   б) більше звичайного коефіцієнта детермінації;
   в) менше або дорівнює звичайному коефіцієнта детермінації.
8. Множинний коефіцієнт кореляції $Ryx_1x_2 = 0,9$. Визначте, який відсоток дисперсії залежної змінної y пояснюється впливом факторів x_1 та x_2:
   а) 90 %;
   б) 81 %;
   в) 19 %.
9. Для побудови моделі лінійної множинної регресії виду $\hat{y} = a + b_1x_1 + b_2x_2$ Необхідна кількість спостережень має бути не менше:
   а) 2;
   б) 7;
   в) 14.
10. Стандартизовані коефіцієнти регресії β_i :
    а) дозволяють ранжувати фактори за силою їх впливу на результат;
    б) оцінювати статистичну значущість факторів;
    в) є коефіцієнтами еластичності.
11. Часткові коефіцієнти кореляції:
    а) характеризують тісноту зв'язку набору факторів, що розглядаються, з досліджуваною ознакою;
    б) містять коригування на кількість ступенів вільності і не допускають збільшення тісноти зв'язку;
    в) характеризують тісноту зв'язку між результатом і відповідним фактором при елімінуванні інших факторів, включених у рівняння регресії.
12. Частковий F-критерій:
    а) оцінює значущість рівняння регресії в цілому;
    б) слугує мірою для оцінки включення фактору в модель;
    в) ранжує фактори за силою їх впливу на результат.
13. У багатофакторній регресії:
    а) більш ніж одна залежна змінна і тільки одна незалежна змінна;
    б) більш ніж одна незалежна змінна і тільки одна залежна змінна;
    в) більш ніж одна залежна змінна і більш ніж одна незалежна змінна;
    г) тільки одна залежна змінна і тільки одна незалежна змінна;
    д) більш ніж дві залежні змінні і більш ніж одна незалежна змінна.
14. При геометричній інтерпретації регресійної моделі з двома незалежними змінними ми будуємо:
    а) пряму лінію, щоб показати зв'язок між залежною змінною та незалежними змінними;
    б) трикутник, щоб показати зв'язок між залежною змінною та незалежними змінними;
    в) площину, щоб показати зв'язок між залежною змінною та незалежними змінними;
    г) коло, щоб показати зв'язок між залежною змінною та незалежними змінними;
    д) еліпс, щоб показати зв'язок між залежною змінною та незалежними змінними.
15. У множинній регресії кожен параметр показує:
    а) загальний вплив усіх незалежних змінних на залежну змінну;

б) вплив незалежної змінної на залежну за умови, що всі інші незалежні змінні залишаються незмінними;

в) де площина регресії перетинає вісь *y*;

г) як частковий, так і загальний вплив незалежних змінних;

д) точку, що дорівнює значенню перетину.

16. Щоб перевірити значущість окремого параметра, використовують:

а) F-тест;

б) t-тест;

в) χ^2-тест;

г) біноміальний розподіл;

д) експоненційний розподіл.

17. Для перевірки значущості одночасно всіх параметрів використовується:

а) F-тест;

б) t-тест;

в) χ^2-тест;

г) біноміальний розподіл;

д) експоненційний розподіл.

18. За інших рівних умов, якщо ми збільшуємо кількість незалежних змінних у регресії:

а) R^2 збільшується;

б) R^2 зменшується;

в) R^2 може або збільшитись, або зменшитись;

г) немає ніякого ефекту на R^2;

д) імовірність мультиколінеарності зменшується.

19. Зв'язок між R^2 та оціненим $\overline{R}^2$ є:

а) $\overline{R}^2 = R^2$;

б) $\overline{R}^2 = R^2(n-1)/(n-p-1)$;

в) $\overline{R}^2 = [1-(1- R^2)](n-1)/(n-p-1)$;

д) $\overline{R}^2 = 1-R^2$;

е) $\overline{R}^2 = 1+R^2$;

20. Для регресії з n спостережень та p незалежних змінних зв'язок між R^2 та F є:

а) $F_{p, n-p-1} = [(n-p-1)/p] [R^2/ (1- R^2)]$;

б) $F_{p, n-p-1} = R^2/(1- R^2)$;

в) $F_{p, n-p-1} = [(n/p] [R^2/ (1- R^2)]$;

г) $F_{p, n-p-1} = [(n-p-1)/p] R^2$;

д) $F_{p, n-p-1} = R^2/ (1+ R^2)$.

21. Ступені вільності для знаходження критичного значення *t*-статистики в регресії, що складається з 35 спостережень та 3 незалежних змінних, такі:

а) 35;

б) 3;

в) 32;

г) 33;

д) 31.

22. Ступені вільності чисельника F-статистики в регресії, що складається з 50 спостережень та 4 незалежних змінних, такі:

а) 50;

б) 4;

в) 3;

г) 46;

д) 45.

23. Ступені вільності знаменника F-статистики, що складається з 50 спостережень та 4 незалежних змінних, такі:
а) 50;
б) 4;
в) 3;
г) 46;
д) 45.
24. Однією з проблем, що може виникнути в багатофакторній регресії і ніколи не буває в простій регресії, є:
а) кореляція між величинами помилок;
б) нерівна дисперсія помилок;
в) кореляція між помилками і незалежними змінними;
г) кореляція між незалежними змінними;
д) помилка без нульової середньої.
25. Інтервали довіри для прогнозного значення *y* знаходяться з метою:
а) виявлення меж коливання реального прогнозного значення з певним рівнем імовірності;
б) ускладнення результатів моделювання і прогнозування;
в) отримання більшої кількості даних для забезпечення вибору найкращого результату;
г) усі відповіді вірні;
д) немає вірної відповіді.
26. ANOVA-аналіз проводиться для:
а) виявлення сум квадратів помилок;
б) розрахунку дисперсій;
в) обчислення F-критерію Фішера;
г) аналізу моделі на адекватність;
д) усі відповіді вірні.
27. Ступені вільності відображають:
а) різницю між кількістю різних дослідів і кількістю констант, встановлених у результаті цих дослідів, незалежно один від одного;
б) суму різних дослідів і кількості констант, встановлених у результаті цих дослідів, незалежно один від одного;
в) суму квадратів, що пояснює регресію;
г) суму квадратів, що не пояснюється регресією;
д) немає вірної відповіді.