# ТЕМА 9. Моделі парної регресії та їх економетричний аналіз (парний кореляційно-регресійний аналіз)

## Частина 1. Однофакторні лінійні економетричні моделі (парна лінійна регресія)

*План*

*9.1. Поняття кореляційних зв'язків та регресійної залежності*

*9.2. Лінійний кореляційний та регресійний аналіз двох змінних*

*9.3. Умови застосування метода найменших квадратів (1МНК)*

### *9.1. Поняття кореляційних зв'язків та регресійної залежності*

В природі, суспільстві, економіці явища, процеси, об'єкти знаходяться між собою в причинній залежності.

Зв'язок між двома величинами називається функціональним, якщо будь-якому визначеному значенню величини $x$ (із множини її можливих значень) відповідає одне і тільки одне значення $y$, тобто $y$ можемо представити функцією від $x$:

$$y = f(x) \qquad (9.1)$$

Зв'язок між двома величинами називається стохастичним, якщо після визначення величини $x$ величина $y$ залишається випадковою і може приймати різні значення з обумовленими імовірностями.

При вивченні зв'язку між явищами стохастична залежність частково вказує на відповідну причинну залежність (наприклад, залежність продуктивності праці робітника від стажу його роботи за фахом). Але за наявності стохастичного зв'язку між явищами може і не бути причинної залежності між ними. Це виникає тому, що обидва явища окремо залежать від загальних факторів. Так, зв'язок між фондовіддачею і собівартістю є стохастичним і непричинним, тому що обидва ці показники залежать від фондоозброєності, електроозброєнності і т.д.

Окремими випадками стохастичної форми зв'язку можуть бути кореляційні зв'язки. Дві випадкові величини є кореляційно залежними, якщо математичне очікування однієї з них змінюється в залежності від зміни другої.

**«Кореляція»** походить від англійського «corelation» і означає співвідношення або відповідність між факторами й ознаками. Основоположниками теорії кореляції вважаються англійські статистики Ф. Гальтон і К. Пірсон. Термін «кореляція» застосовується в різних галузях науки і техніки для позначення взаємозалежності, взаємної відповідності.

При виконанні кореляційних розрахунків необхідно розрізняти факторну та результативну ознаку. Факторною називається така ознака, від якої залежить інша ознака, а сама вона є незалежною. Залежна ознака називається результативною.

У процесі формалізації економіко-статистичної моделі факторна ознака позначається через $X_i$, а результативна через $Y_i$, тобто умовно можна сказати, що факторна ознака виражає аргумент, а результативна – функцію. Факторна ознака $X_i$ є незалежною від змінної $Y_i$ так як відсутній зворотний вплив $Y_i$ на $X_i$. У зв'язку з цим чинники $X_i$ часто називають екзогенними (зовнішніми), а змінну $Y_i$ – ендогенною (внутрішньою) змінною моделі. Значення змінних $X_i$ визначаються поза моделлю, тобто задаються як початкові дані.

Факторна ознака або фактор – це технічні, технологічні, природні, кліматичні, економічні, організаційні, соціально-демографічні та інші показники, що проявляють вплив на який-небудь результативний економічний показник: прибуток, собівартість, продуктивність праці та ін. Задача математичного моделювання полягає у виявленні кількісного зв'язку між факторами та результативним економічним показником.

*Фактор, що включається в економетричну модель, повинен відповідати таким вимогам:*

1) мати кількісне вираження;

2) між фактором і результуючим показником повинен бути причинний зв'язок і статистичний зв'язок;

3) між факторами у багатофакторній моделі не повинно бути мультиколінеарності (тісного зв'язку між факторами).

Кореляційний зв'язок між факторами в економіці класифікують за ознаками:

- за типом: прямий і обернений;
- за формою: лінійний і нелінійний;
- за тіснотою зв'язку: слабкий, помірний, помітний, сильний, дуже сильний;
- за участю факторних ознак: парний, множинний.

Розглянемо приклади факторних та результативних ознак.

**Приклад 9.1.** Досліджується кореляційна залежність між рівнем продуктивності праці та рівнем механізації праці. Отже, рівень механізації праці – факторна ознака ($x$), а рівень продуктивності праці – результативна ознака ($y$). Зв'язок між ознаками прямий та лінійний.

**Приклад 9.2.** Досліджується залежність між рівнем продуктивності праці робітника та його віком. У цьому випадку вік робітника – факторна ознака ($x$), а рівень продуктивності праці – результативна ознака ($y$). Зв'язок між ознаками нелінійний: з початку прямий, а потім обернений, а залежність описується квадратичною параболою.

**Приклад 9.3.** Досліджується залежність між собівартістю одиниці продукції та врожайністю. Врожайність сільськогосподарської культури є факторною ознакою ($x$), а собівартість одиниці продукції – результативною ($y$). Зв'язок між ознаками: нелінійний та обернений.

Кількісний вплив факторів на результативний показник вивчається за допомогою регресійного аналізу, який дозволяє встановити вид аналітичної залежності між змінними *x, y* та оцінити параметри економетричної моделі.

Вперше термін "регресія" був введений Френсісом Галтоном. 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.

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.

Під **регресією** розуміється функціональна залежність між пояснюючими змінними і умовним математичним очікуванням (середнім значенням) залежної змінної, яка будується з метою прогнозу (прогнозування) цього середнього значення при фіксованих значеннях перших.

Прикладом можливого застосування регресійного аналізу в економіці може бути дослідження продуктивності праці, собівартості та інших якісних економічних показників від таких факторів як вартість основних засобів, питома вага заробітної плати у витратах на виробництво, рівень спеціалізації, кооперування, плинність та рівень кваліфікації кадрів. Регресійні моделі також широко застосовуються в прогнозуванні.

При виборі форми кореляційної залежності $y = f(x)$ виходять перш за все із економічної природи явищ, простоти функції і вимоги на обмеження числа параметрів. Форму кореляційного зв'язку можна визначити як графічним, так і аналітичним методами.

У випадку парної кореляції вихідні дані $n$ пар $x_i$, $y_i$ $(i = \overline{1,n})$ в прямокутній системі координат утворюють кореляційне поле (рис. 9.1). Розміщення точок на кореляційному полі дозволяє визначати характер залежності: $y = f(x)$ (а, б – лінійна; в – параболічна; г – гіперболічна; д – логістична; е – відсутня).

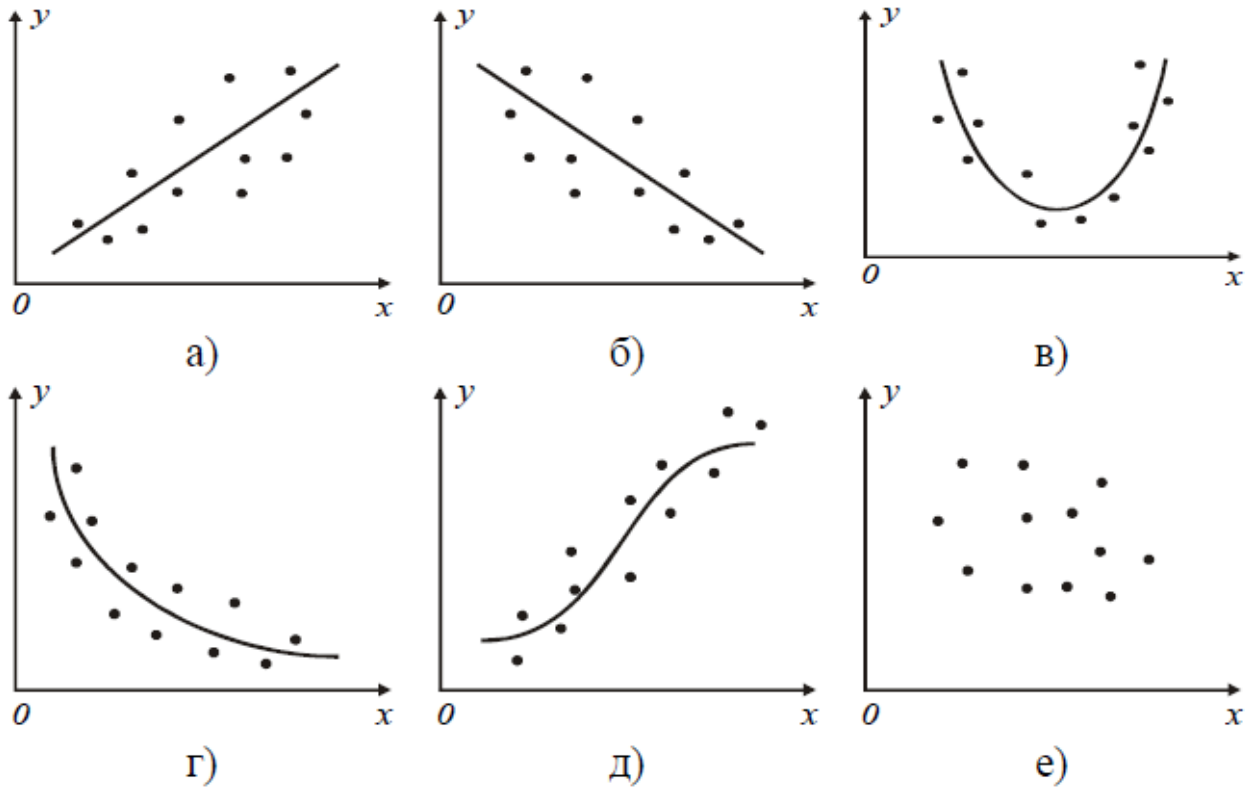



Рис. 9.1. Поле кореляції (діаграма розсіювання)

Якщо $x_i$ – різні $(i=\overline{1,n})$, то точки кореляційного поля з'єднують в послідовності зростання абсциси і одержують так звану емпіричну лінію регресії (рис. 9.2).

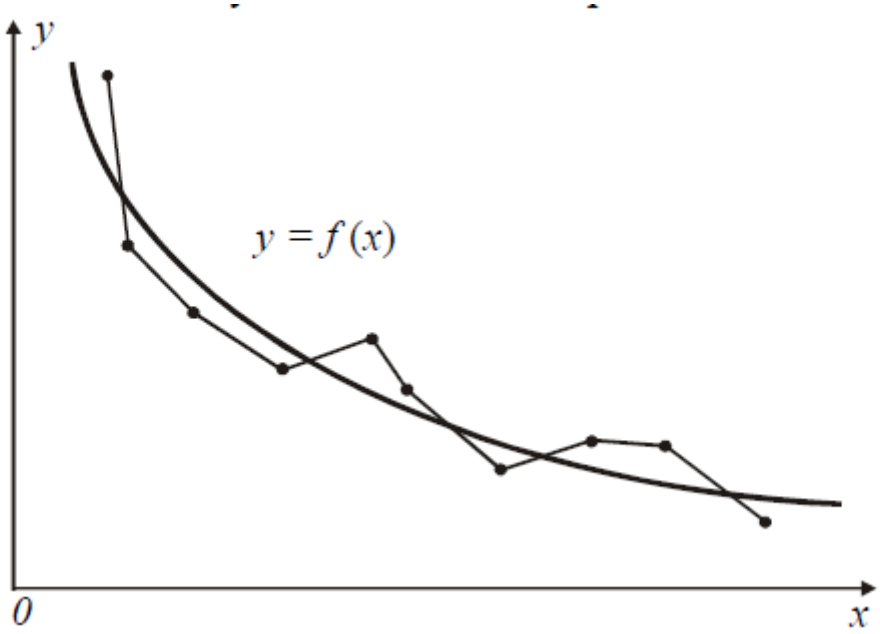



Рис. 9.2. Емпірична та теоретична лінії регресії

Графік функції $y = f(x)$ називають теоретичною лінією регресії. Щоб обрати ту чи іншу форму кореляційної залежності, слід зіставити кореляційне поле або

емпіричну лінію регресії з графіками відомих функцій. Для більш точного встановлення форми зв'язку вихідні дані обробляють на *ЕОМ* по програмах кореляційного аналізу. При цьому аналізують кілька функцій $y = f(x)$ і беруть ту, для якої кореляційне відношення $\eta$ або коефіцієнт парної кореляції $r$ найбільші (або середня похибка апроксимації $\varepsilon$ найменша).

### *9.2. Лінійний кореляційний та регресійний аналіз двох змінних*

Економетричні моделі в залежності від обсягу вибірки статистичних даних поділяються на узагальнені та вибіркові.

Узагальненою вважається регресійна модель, побудована по статистичних даних генеральної сукупності і має вигляд: $y = \beta_0 + \beta_1 x + u$, де $\beta_0$, $\beta_1$ – параметри моделі, $u$ – випадкова величина (відхилення).

Вибіркова модель будується по статистичних даних вибіркової сукупності. У загальному вигляді вибіркова регресійна модель між факторною ознакою $X = \{x_1, x_2, \ldots, x_n\}$ та результативною ознакою $Y = \{y_1, y_2, \ldots, y_n\}$ з урахуванням фактору випадкових величин (помилок) $U = \{u_1, u_2, \ldots, u_n\}$ записується у вигляді:

$$y = a_0 + a_1 x + u \qquad (9.2)$$

де $a_0$ , $a_1$ – невідомі параметри економетричної моделі;

$u$ – випадкова величина (відхилення).

Тут і надалі з метою уникнення неоднозначності великими літерами $X$ ,$Y$ ,$U$ ми позначаємо дискретні (векторні) величини, а малими $x$, $y$, $u$ - неперервні.

**Причини обов'язкової присутності в регресійних моделях випадкової змінної (відхилення) *u* такі:**

1. Невключення до моделі всіх пояснюючих змінних. Будь-яка регресійна модель є спрощенням реальної ситуації. Остання завжди являє собою взаємодію різних чинників, багато з яких в моделі не враховуються, що обумовлює відхилення реальних значень залежної змінної від її модельних значень. Проблема полягає ще й в тому, що наперед не відомо, які фактори при певних умовах дійсно є визначальними, а якими можна нехтувати.

2. Неправильний вибір функціональної форми моделі. Через недостатню вивченість процесу чи явища, що моделюється, може бути невірно підібрана аналітична функція, якою проводиться моделювання.

3. Агрегація змінних. У багатьох моделях розглядаються залежності між чинниками, які представляють складну комбінацію інших, простіших змінних. Наприклад, чинник сукупний попит є складною композицією індивідуальних попитів, які впливають на результативний показник. Це може виявитись причиною відхилення реальних значень від модельних.

4. Помилка вимірювань. Якою б якісною не була модель, помилки вимірювань змінних вплинуть на невідповідність модельних значень емпіричним даним, що також відобразиться на величині випадкового члена (відхилення).

5. Обмеженість статистичних даних. Часто будуються моделі, що виражаються безперервними функціями. Але для цього використовується набір даних, що мають дискретну структуру. Ця невідповідність знаходить свій вираз у випадковому відхилені.

6. Непередбачуваність людського чинника. Ця причина може «зіпсувати» найякіснішу модель. Дійсно, при правильному виборі форми моделі, скрупульозному підборі пояснюючих змінних все одно неможливо спрогнозувати поведінку кожного індивідуума.

Таким чином, відхилення (випадкова величина) є віддзеркаленням впливу всіх описаних вище причин. До того ж, цей перелік може бути доповненим.

Метод математичної статистики, який вивчає кореляційні зв'язки між явищами, називається ***кореляційним аналізом***. Кореляційний аналіз представляє собою інструмент, який дозволяє кількісно оцінити зв'язки між великою кількістю взаємодіючих економічних явищ, при цьому деякі з них невідомі. Застосування кореляційного аналізу дає можливість перевірити різні економічні гіпотези про наявність і силу зв'язку між двома явищами або явищем та групою явищ, а також гіпотезу про форму зв'язку.

Задачею регресійного аналізу є обчислення невідомих параметрів $a_0$, $a_1$ рівняння регресії $\hat{y} = a_0 + a_1 x$. При цьому необхідно досягти «найкращої» апроксимації. Найчастіше при цьому використовують метод найменших квадратів, що передбачає мінімізацію виразу:

$$Q(a_0, a_1) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} u_i^{\,2} \to \min \qquad (9.3)$$

де $y_i, \hat{y}_i$ фактичні (емпіричні) та розрахункові (теоретичні) значення результативної ознаки.

На рисунку 9.3 пряма є теоретичною лінією регресії.

Розглянемо геометричну інтерпретацію комбінації цих двох складових (рис.15.1.1). Показники $x_1, x_2, .., x_n$ – це гіпотетичні значення пояснювальної змінної. Якщо би співвідношення між $y$ та $x$ були однаковими, то відповідні значення $y$ були би представлені точками $B_1, B_2, .., B_n$ на одній прямій. Наявність випадкового члена збурення приводить до того, що насправді значення $y$ отримують іншим. Відзначимо на графіку реальні значення $y$ при відповідних значення $x$ з допомогою точок $A_1$, $A_2$, ..., $A_n$.

Із множини прямих необхідно вибрати «найкращу» з точки зору мінімізації суми квадратів відхилень $u_i$ : $u_i = y_i - \hat{y}_i = y_i - a_0 - a_1 \cdot x_i; i = \overline{1,n}$. Відхилення або помилки $u_i$ ще іноді називають залишками. Теоретичну лінію регресії необхідно проводити таким чином, щоб сума квадратів відхилень була мінімальною.

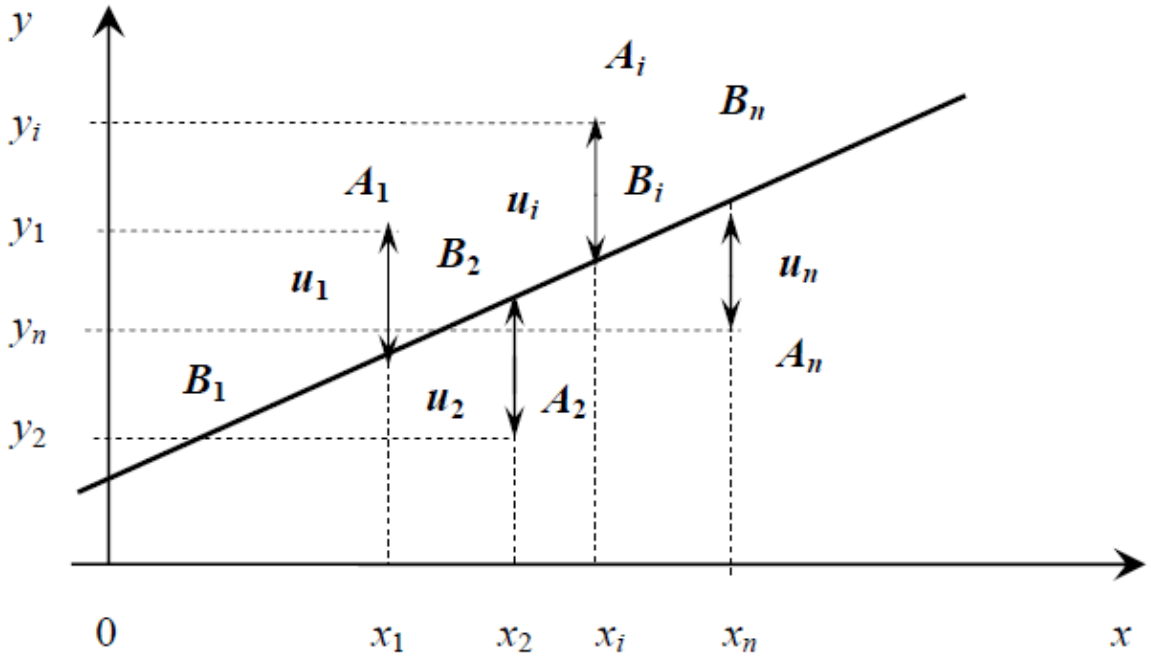



Рис. 9.3. Фактична залежність між y та x

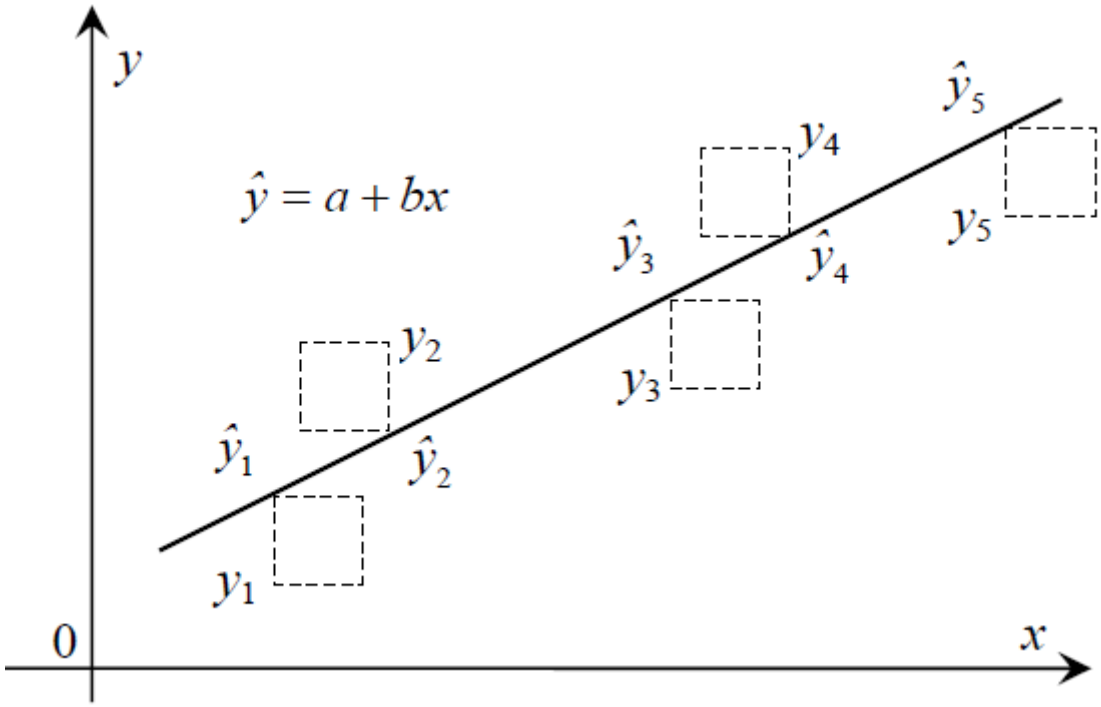



Рис. 9.4. Графічна інтерпретація методу найменших квадратів

У цьому і полягає метод найменших квадратів: невідомі параметри $a_0$ та $a_1$ визначаються таким чином, щоб мінімізувати $\sum_{i=1}^{n} u_{i^2}$. Мінімум функції (9.2) досягається за умови, коли перші похідні дорівнюють нулю. Тому підставивши в

вираз (9.2), взявши частинні похідні $\frac{\partial Q}{\partial a_0}$ і $\frac{\partial Q}{\partial a_1}$, після елементарних перетворень одержимо систему нормальних рівнянь:

$$\begin{cases} na_0 + a_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i \end{cases} \quad (9.4)$$

де $n$ – кількість спостережень або довжина вибірки.

Шляхом розв'язання системи нормальних рівнянь на основі метода найменших квадратів оцінюються параметри лінійної економетричної моделі $a_0$ та $a_1$:

$$a_0 = \frac{\sum_{i=1}^{n} y_i \cdot \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} y_i x_i \cdot \sum_{i=1}^{n} x_i}{n \cdot \sum_{i=1}^{n} x_i^2 - \left( \sum_{i=1}^{n} x_i \right)^2} \quad (9.5)$$

$$a_1 = \frac{n \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{n \cdot \sum_{i=1}^{n} x_i^2 - \left( \sum_{i=1}^{n} x_i \right)^2} \quad (9.6)$$

Параметри $a_0$ , $a_1$ мають таку економічну інтерпретацію або зміст: параметр $a_0$ характеризує деяке середнє значення результативного показника $y$ , а параметр $a_1$ показує, як в середньому зміниться $y$ при зміні $x$ на одну одиницю.

Постійна величина $a_0$ визначає точку перетину прямої регресії з віссю ординат і є середнім значенням у точці $x_0$=0. Зрозуміло, що економічна інтерпретація $a_0$ не тільки утруднена, а й взагалі неможлива. Величина $a_0$ у рівнянні регресії лише виконує функцію вирівнювання і має розмірність $y$. При цьому слід відзначити, що завдяки постійній $a_0$ функція регресії непомилкова. Рівняння регресії інтерпретується тільки в області скупчення точок і, як наслідок, тільки між найменшим і найбільшим значенням змінної $x$, яка спостерігається. Більш практичний інтерес представляє економічний зміст величин $a_1$ та $y\hat{}$.

Відповідно до рівняння $a_1$ визначає середню величину зміни результативного показника при зміні пояснювальної змінної $x$ на одну одиницю. Знак $a_1$ визначає напрямок цієї зміни, а розмірність цього коефіцієнта є відношенням розмірності залежної змінної до розмірності пояснювальної змінної.

**Приклад 9.4.** Нехай залежність денного виробітку робітника від рівня механізації праці описується рівнянням регресії: $y$ = 2,142 + 0,051$x$. У цьому рівнянні параметр $a_0$ є середнім денним виробітком при виконанні операції вручну, а $a_1$ – перевищення середнього виробітку при механізованому виконанні операції. А тому

параметр $a_1$ (коефіцієнт нахилу) показує, що при підвищенні рівня механізації на 1% денний виробіток зростає в середньому на 0,051 одиниць.

Отже, при моделюванні та аналізі багатьох соціально-економічних явищ та процесів виникає задача виявлення та оцінки зв'язку між ними, одне з яких є незалежною змінною ( $x$ ), чи фактором, а інше ( $y$ ) – залежною або результативною ознакою. Форма зв'язку між змінними $x$ та $y$ встановлюється шляхом логічного аналізу їх природи та зовнішнього вигляду кореляційного поля та емпіричної лінії регресії, а тіснота зв'язку – величиною коефіцієнта кореляції.

**Тіснота (щільність)** зв'язку між змінними $x$ та $y$ оцінюється **коефіцієнтом парної кореляції або коефіцієнтом кореляції Пірсона** $r_{xy}$ (якщо зв'язок лінійний) і **кореляційним відношенням** $\eta_{xy}$ (якщо зв'язок нелінійний).

Коефіцієнт кореляції являє собою ступінь асоціативності між двома змінними. Для обчислення коефіцієнта кореляції пропонуються різні формули.

Розглянемо деякі з них:

$$r_{xy} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \cdot \sigma_y} \tag{9.7}$$

де $\overline{xy}$ – середнє значення добутку змінної $x$ та змінної $y$ ; $\bar{x}$, $\bar{y}$ – середнє значення змінних $x$ та $y$:

$$\overline{xy} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \cdot y_i , \qquad \bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i , \qquad \bar{y} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i \tag{9.8}$$

$$r_{xy} = \frac{Cov(x,y)}{\sqrt{Var(x) \cdot Var(y)}} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{n \cdot \sigma_x \cdot \sigma_y} \tag{9.9}$$

$$r_{xy} = \frac{n \cdot \sum_{i=1}^{n} x_i \cdot y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{\sqrt{\left[ n \cdot \sum_{i=1}^{n} x_i^{\,2} - \left( \sum_{i=1}^{n} x_i \right)^2 \right] \cdot \left[ n \cdot \sum_{i=1}^{n} y_i^{\,2} - \left( \sum_{i=1}^{n} y_i \right)^2 \right]}} \tag{9.10}$$

де $n$ – довжина вибірки або кількість спостережень;

$Cov(x, y)$ – коефіцієнт коваріації між змінними $x$ та $y$ ;

$$Cov(x,y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{n-1} \tag{9.11}$$

$Var(x)$ – дисперсія змінної $x$:

$$Var(x) = \sigma_x^{\,2} = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{9.12}$$

$Var(y)$ – дисперсія змінної $y$:

$$Var(y) = \sigma_y^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad (9.13)$$

Визначена таким чином величина має назву коефіцієнта кореляції за вибіркою.

***Властивості коефіцієнта кореляції:***

1. Він може бути позитивним або негативним, знак *r* залежить від знаку чисельника (9.10), що є мірою коваріації за вибіркою двох змінних.

2. Коефіцієнт кореляції змінюється в інтервалі:

$$-1 \le r_{xy} \le 1$$

3. За своєю природою він симетричний, тобто коефіцієнт кореляції між $x_i$ і $y_i$ ($r_{xy}$) той же, що й між $y_i$ і $x_i$ ($r_{yx}$). Тому, іноді скорочено будемо його позначати просто *r*.

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 $r_{xy}$ між *X** і *Y** той 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. 9.5 *з*).

6. Коефіцієнт кореляції є мірою 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 $r = 0$.

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

При коефіцієнті кореляції рівному 0, між *y* та *x* не існує кореляційного зв'язку. Якщо коефіцієнт кореляції знаходиться в інтервалі $-1 \le r_{xy} \le 1$ або $0 \le r_{xy} \le 1$ 0, між *y* та *x* існує обернена або пряма кореляційна залежність.

*За щільністю зв'язку можна виділити:*

а) слабкий зв'язок, якщо $r_{xy} \le 0{,}3$;

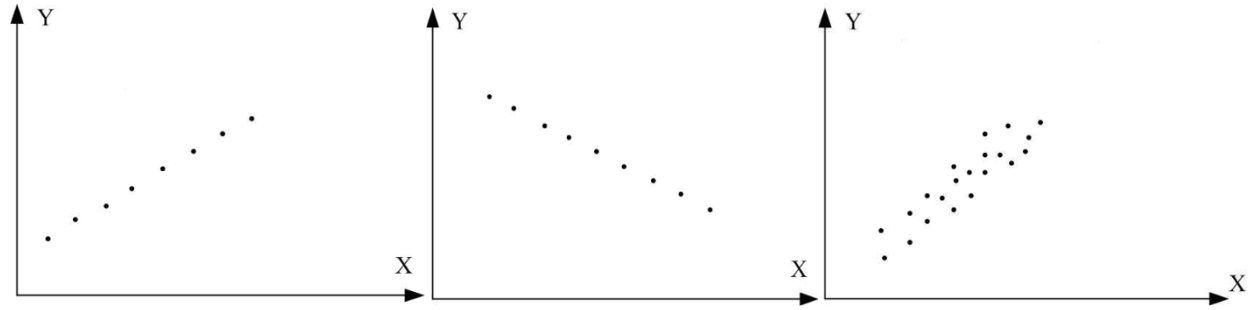
б) середній зв'язок, якщо $r_{xy}$ = 0,31-0,65;

в) сильний зв'язок, якщо $r_{xy}$ = 0,66-0,95.

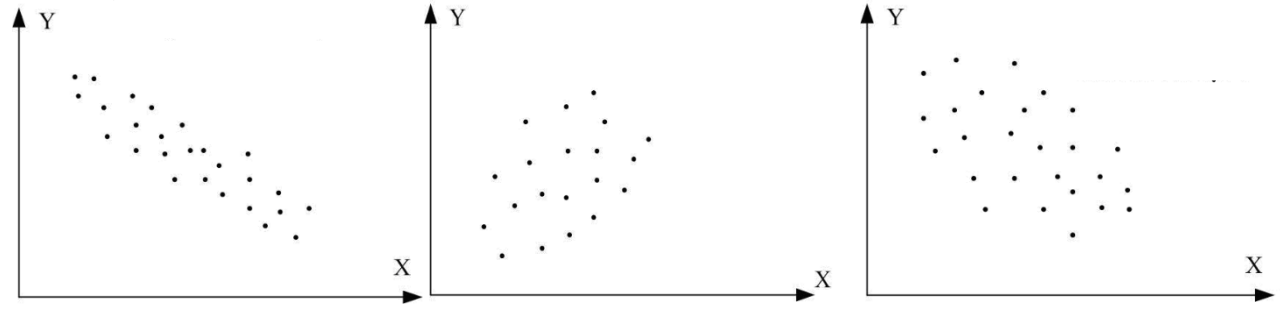
За значенням коефіцієнта кореляції можна зробити такі висновки:

- якщо $r_{xy}$ набуває значення, яке близьке до -1, то між факторами існує щільний зворотний (обернений) зв'язок;
- якщо $r_{xy} = 0$, то зв'язок відсутній;
- якщо $r_{xy}$ близьке до +1, то між факторами існує щільний прямий зв'язок;
- якщо $|r_{xy}| = 1$, то між досліджуваними показниками існує функціональний зв'язок.

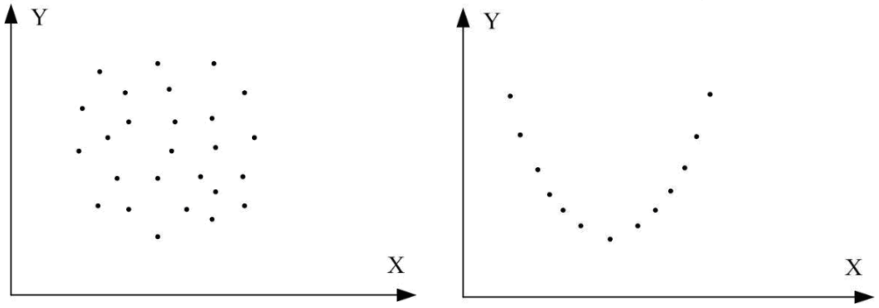
Відзначимо, що знак коефіцієнта кореляції $r$ вказує на напрям зв'язку між ознаками $x$ $y$, в той час як $|r_{xy}|$ характеризує щільність зв'язку.




*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

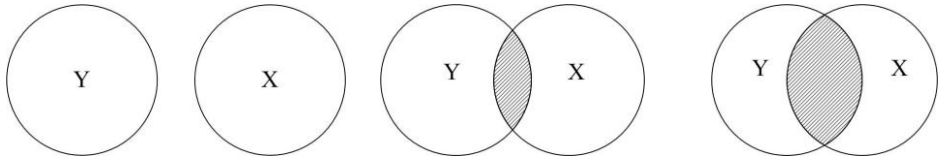
Рис. 9.5. Кореляційний коефіцієнт для різних випадків вибірок

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

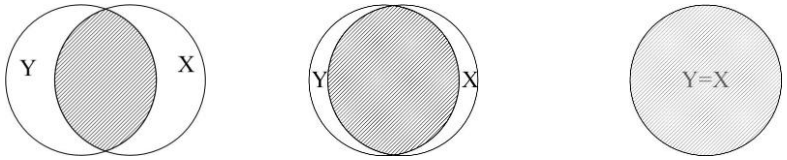
Звернемося зараз до розгляду питання якості підгонки лінії регресії до множини даних, тобто дослідимо, наскільки «добре» лінія вибіркової регресії підходить до цих даних. Із рис.9.2 видно, що якби всі спостереження знаходилися на лінії регресії, ми отримали б «точну підгонку», але на практиці це окремий 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. 9.6).

На рис. 9.6 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. З рис. 9.6 видно, 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*. Отже, $r^2$ лежить між 0 і 1.




*а)* $r^2=0$ *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

Рис. 9.6. Діаграма для пояснення $r^2$

Для визначення варіації результативного показника під впливом факторів обчислюють **коефіцієнт детермінації $r^2_{yx}$.** Припустимо, що $r^2_{yx}$=0,8; тоді можна сказати, що 80% варіації результативного показника відбувається під впливом фактору *x*, а решта 20% приходиться на інші фактори та випадкові величини.

При виявленні зв'язку між варіацією факторної ознаки *(x)* і варіацією результативної ознаки *(y)* використовують такі *дисперсії:*

1) дисперсія, яка вимірює загальну варіацію за рахунок дії всіх факторів, або *загальна дисперсія*:

$$\sigma^2_{загальна} = \frac{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}{n} \qquad (9.14)$$

2) *пояснювальна дисперсія*, яка вимірює варіацію результативної ознаки *y* за рахунок дії факторної ознаки *x* або дисперсія, що пояснює регресію:

$$\sigma^2_{пояснювальна} = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2}{n} \quad (9.15)$$

3) *залишкова (непояснювальна) дисперсія*, яка характеризує варіацію ознаки *y* за рахунок всіх факторів, крім *x* (тобто при виключенні *x*) або дисперсія помилок:

$$\sigma^2_{непояснювальна} = \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n} \quad (9.16)$$

тоді по правилу додавання дисперсій:

$$\sigma^2_{загальна} = \sigma^2_{пояснювальна} + \sigma^2_{непояснювальна} \quad (9.17)$$

або

$$\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2 + \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \quad (9.18)$$

де $\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$ – загальна сума квадратів, яка позначається через *TSS (total sum squares)*; вона відображає дисперсію величини $y_i$ (емпіричне або фактичне значення) відносно її середнього значення;

$\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2$ – сума квадратів, що пояснює регресію та позначається через *ESS (explained sum squares);* відображає дисперсію оціненої (теоретичної) величини $\hat{y}_i$ відносно середнього значення $y_i$;

$\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$ – сума квадратів помилок, яка позначається через *RSS (residual sum squares)*; відображає залишкову або нез'ясовану дисперсію величини $y_i$ щодо лінії регресії $\hat{y}_i$ або просто залишкова сума квадратів.

Вирази (9.17), (9.18) запишемо у скороченому вигляді:

$$TSS=ESS+RSS \quad (9.19)$$

Формула (9.19) 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. 9.7 це розбиття пояснене геометрично.

Таким чином, ми розклали загальну дисперсію на дві частини: дисперсію, яка пояснює регресію, та дисперсію помилок (або дисперсію випадкової величини).

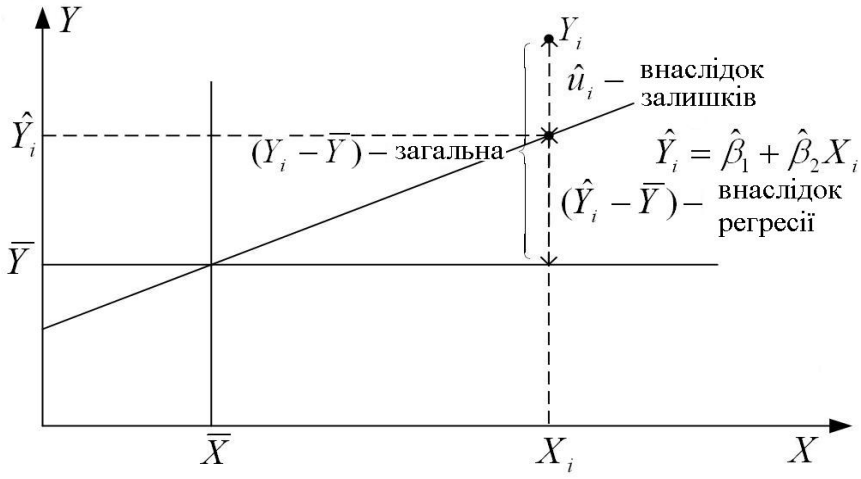



Рис. 9.7. Розбиття варіації $Y_i$ на дві компоненти

Поділивши обидві частини виразу (9.19) на $TSS=б^2_{загальна}$, отримаємо:

$$\frac{TSS}{TSS} = \frac{ESS}{TSS} + \frac{RSS}{TSS}$$

або (9.20)

$$1 = \frac{\sigma^2_{пояснювальна}}{\sigma^2_{загальна}} + \frac{\sigma^2_{непояснювальна}}{\sigma^2_{загальна}}$$

Із виразу (9.20) випливає, що перша частина $\frac{\sigma^2_{пояснювальна}}{\sigma^2_{загальна}}$ є складовою дисперсії, яку можна пояснити через лінію регресії. Друга частина $\frac{\sigma^2_{непояснювальна}}{\sigma^2_{загальна}}$ є питомою вагою помилок у загальній дисперсії, тобто часткою дисперсії, яку не можна пояснити через регресійний зв'язок.

Частина дисперсії, що пояснюється регресію, називається **коефіцієнтом детермінації** і позначається $r^2$. Коефіцієнт детермінації використовується як критерій адекватності моделі, бо є мірою пояснювальної сили незалежності змінної *x*.

Таким чином, коефіцієнт детермінації:

$$R^2 = \frac{\sigma^2_{пояснювальна}}{\sigma^2_{загальна}} \quad (9.21)$$

або

$$R^2 = \frac{ESS}{TSS} \quad (9.22)$$

Або в альтернативному вигляді:

$$r^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum u_i^2}{\sum (Y_i - \bar{Y})^2} \quad (9.22\ а)$$

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

***Властивості коефіцієнту детермінації*** $r^2$***:***

1. Коефіцієнт $r^2$ завжди додатній (випливає з виразів (9.21)-(9.22)).

2. $r^2$ має межі $0 \le r^2 \le 1$. При значенні $r^2 = 1$ ми маємо випадок точної підгонки, тобто $\hat{y}_i = y_i$ для кожного *i*. Водночас випадок $r^2=0$ означає відсутність зв'язку між регресантом і регресором (тобто параметр перед $x$ - $a_1$ для всіх *i*). В цьому випадку кращим прогнозом для будь-якої величини *Y* є її середнє значення. При цьому лінія регресії - паралель осі *X*.

Коефіцієнт кореляції *r* кількісно близько пов'язаний з коефіцієнтом детермінації $r^2$, але концептуально вони дуже різні. Коефіцієнт кореляції можна визначити за формулою

$$r = \pm\sqrt{r^2} \quad (9.22\ б)$$

або

$$r = \frac{\sum x_i y_y}{\sqrt{\sum x_i^2 \sum y_i^2}} = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{\sqrt{\left[n\sum X_i^2 - \left(\sum X_i\right)^2\right]\left[n\sum Y_i^2 - \left(\sum Y_i\right)^2\right]}}. \quad (9.22\ в)$$

Між коефіцієнтом кореляції і нахилом $a_1$ та середнім квадратичним відхиленням $\sigma_x$, $\sigma_y$ існує певний зв'язок. Це дає можливість розрахувати параметри вибіркового рівняння регресії $y = a_0 + a_1 x + u$ через ці величини.

Оскільки

$$r_{xy} = \frac{Cov(x,y)}{\sqrt{Var(x) \cdot Var(y)}} \quad (9.9)$$

$$a_1 = \frac{Cov(x,y)}{\sqrt{Var(x)}} = \frac{Cov(x,y)}{\sigma_x^2} \quad (9.23)$$

можна записати вираз для коефіцієнта кореляції через параметр $a_1$ :

$$r_{xy} = \left(\frac{Cov(x,y)}{\sigma_x^2}\right) \cdot \left(\frac{\sigma_x}{\sigma_y}\right) = a_1 \frac{\sigma_x}{\sigma_y} \quad (9.24)$$

Запишемо формули для розрахунку параметрів економетричної моделі:

$$a_1 = \frac{Cov(x,y)}{Var(x)} = \frac{Cov(x,y)}{\sigma_x^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\cdot(y_i - \bar{y})}{\sigma_x^2} = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\sigma_x^2} = r_{xy}\cdot\frac{\sigma_y}{\sigma_x} \qquad (9.25)$$

$$a_0 = \bar{y} - r_{xy}\frac{\sigma_y}{\sigma_x}\cdot\bar{x} = \bar{y} - a_1\cdot\bar{x} \qquad (9.26)$$

Необхідно відмітити, що при лінійній формі зв'язку коефіцієнт кореляції $r_{xy}$ є оцінкою точності апроксимації, тобто адекватності моделі і дорівнює кореляційному відношенню $\eta_{xy}$.

**Регресійний аналіз і аналіз дисперсії**

Звернемося до регресійного аналізу з погляду аналізу дисперсії.

Раніше нами була доведена така рівність:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (9.18)$$

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 (табл. 9.1).

Таблиця 9.1

*ANOVA*-таблиця для регресійної моделі

| *Джерело варіації* | *Ступені свободи, df* | *Сума квадратів відхилень,* SS | *Середні суми квадратів відхилень, MS* |
|---|---|---|---|
| Регресії | $k_1$=*m-1* | $ESS = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ | $MSE = \frac{ESS}{k_1}$ |
| Залишків | $k_2$=*n-m* | $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ | $MSR = \frac{RSS}{k_2}$ |
| Загальної змінної | *n-1* | $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$ | $MST = \frac{TSS}{n-1}$ |
| SS – сума квадратів (sum squares)<br>MS – середня сума квадратів (mean sum squares)<br>*n* – кількість спостережень,<br>*m* – кількість незалежних змінних $x_i$. | | | |

З кожною сумою квадратів пов'язані кількість її степенів вільності (свободи) *df* - це кількість незалежних спостережень, на яких вона заснована. Іншими словами це числа, що показують скільки незалежних елементів інформації зі змінних $y_i$ потрібно для розрахунку відповідної суми квадратів.

Після побудови моделі обчислюється також середня відносна похибка апроксимації, %:

$$\varepsilon = \frac{100}{n} \cdot \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \qquad (9.27)$$

Середня похибка апроксимації показує в процентах середнє для всіх значення відхилення результативного показника від розрахункових значень. Модель можна вважати адекватною, якщо середня похибка апроксимації буде знаходитись у межах 12-15%.

### *9.3. Умови застосування методу найменших квадратів (1МНК)*

Запишемо вибіркову економетричну модель у матричній формі:

$$Y = XA + U \qquad (9.28)$$

де $Y$ – вектор значень залежної змінної;

$X$ – матриця незалежних змінних розміром $n \times m$ ( $n$ – кількість спостережень, $m$ – кількість незалежних змінних);

$A$ – вектор оцінок параметрів моделі;

$U$ – вектор залишків (похибок).

**Застосування однокрокового метода найменших квадратів (1МНК) для оцінки параметрів моделі передбачає такі умови:**

*1) математичне сподівання залишків дорівнює нулю*, тобто:

$$M(U)=0 \qquad (9.29)$$

а залишки мають нормальний розподіл.

*2) значення $U_i$ вектору залишків $U$ незалежні між собою і мають постійну дисперсію:*

$$M(UU')=\sigma^2 \times I, \qquad (9.30)$$

де $I$ – одинична матриця,

3*) незалежні змінні моделі не пов'язані із залишками:*

$$M(x' U) = 0 \qquad (9.31)$$

4) незалежні змінні моделі утворюють лінійно незалежну систему векторів або, іншими словами, незалежні змінні не повинні бути мультиколінеарними, тобто $|x' \cdot x| \neq 0$:

$$Var\left(x_k' \cdot x_j\right) = 0, \quad k \neq j; \qquad Var\left(x_k' \cdot x_j\right) = 1, \quad k = j; \qquad (9.32)$$

де $x_k$ – *k-й* вектор матриці пояснювальних змінних; $x_j$ – *j-й* вектор цієї матриці пояснювальних змінних $x$, $k = \overline{1,m}$, $j = \overline{1,m}$

**Перша умова** є очевидною. Адже коли математичне сподівання залишків не дорівнює нулю, то це означає, що існує систематичний вплив на залежну змінну, а

до модельної специфікації не введено всіх основних незалежних змінних. Якщо ця умова не виконується, то має місце помилка специфікації.

В економетричних моделях з вільним членом за рахунок його значення можна скоригувати рівняння так, щоб математичне сподівання залишків дорівнювало нулю. Отже, для таких моделей перша умова практично виконується завжди.

**Друга умова** передбачає наявність сталої дисперсії залишків. Цю властивість називають **гомоскедастичністю.** Проте вона може використовуватись лише тоді, коли залишки є помилками вимірювання. Якщо залишки акумулюють загальний вплив змінних, які не враховані в моделі, то дисперсія залишків не може бути сталою величиною. Вона змінюється для окремих груп спостережень. У цьому випадку має місце явище **гетероскедастичності**, яке впливає на методи оцінювання параметрів.

**Третя умова** передбачає незалежність між залишками $U_i$ та пояснювальними змінними $y_i$, яка порушується насамперед тоді, коли економетрична модель будується на базі одночасових структурних рівнянь або має лагові змінні. Тоді для оцінювання параметрів моделі використовується, як правило, дво- або трикровий метод найменших квадратів.

**Четверта умова** означає, що всі пояснювальні змінні ($x_j$) , які входять до економетричної моделі, мають бути незалежними між собою. Проте очевидно, що в економіці дуже важко сформувати такий масив незалежних пояснювальних змінних, які були б зовсім не пов’язані між собою. Тоді кожного разу необхідно з’ясувати, чи не впливатиме залежність пояснювальних змінних на оцінку параметрів моделі.

Це явище називаються мультиколінеарністю змінних. Воно призводить до ненадійності оцінки параметрів моделі, робить їх чутливими до вибраної специфікації моделі та до конкретного вибору даних. Знижується рівень довіри до результатів верифікації моделей з допомогою однокрокового метода найменших квадратів (1МНК). Отже, це явище з усіх точок зору є дуже небажаним. Але воно досить поширене. Існують методи для виявлення мультиколінеарності і способи її врахування з допомогою специфікації моделі чи спеціальних методів оцінювання параметрів (методу Ейткена).