ТЕМА 9. Регресійні моделі в прогнозуванні

План

9.1. Поняття кореляційних зв'язків та регресійної залежності
9.2. Лінійний кореляційний та регресійний аналіз двох змінних
9.3. Класична лінійна багатофакторна модель
9.4. Передумови застосування методу найменших квадратів
9.5. Багатофакторна регресія та її оціночні характеристики
9.6. Оцінка якості економетричних моделей
9.7. Прогнозування розвитку економічних процесів

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).

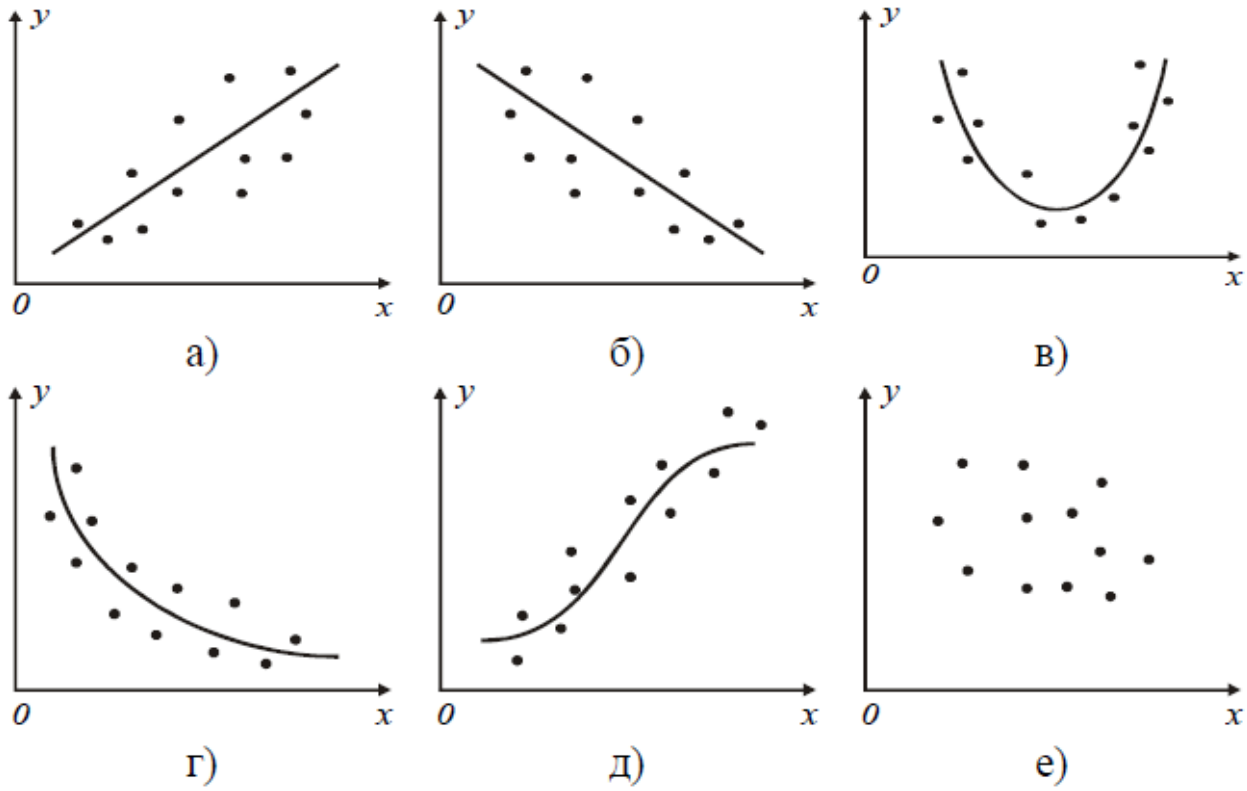


Рис. 9.1. Поле кореляції (діаграма розсіювання)

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).

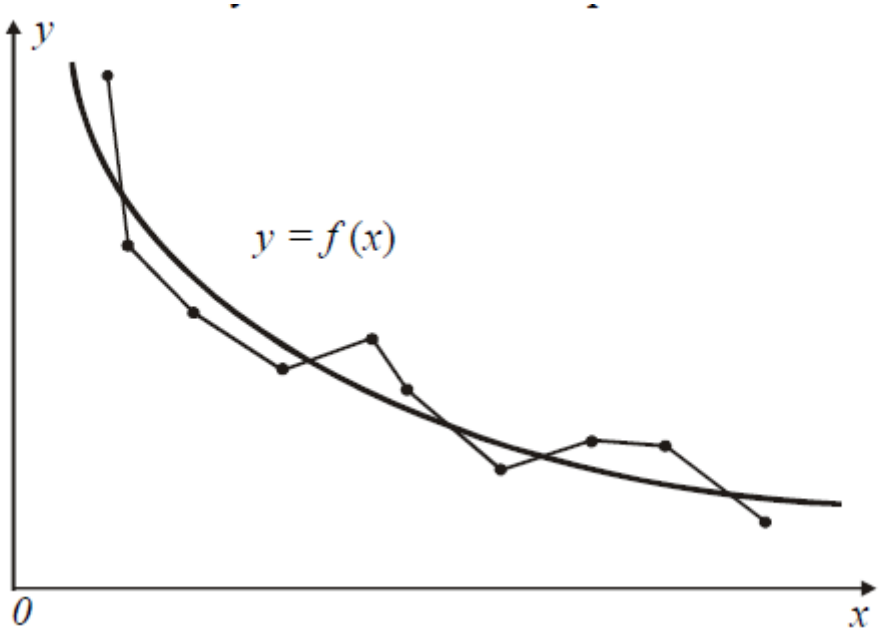


Рис. 9.2. Емпірична та теоретична лінії регресії

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 η або коефіцієнт парної кореляції r найбільші (або середня похибка апроксимації ε 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 β_0, β_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.

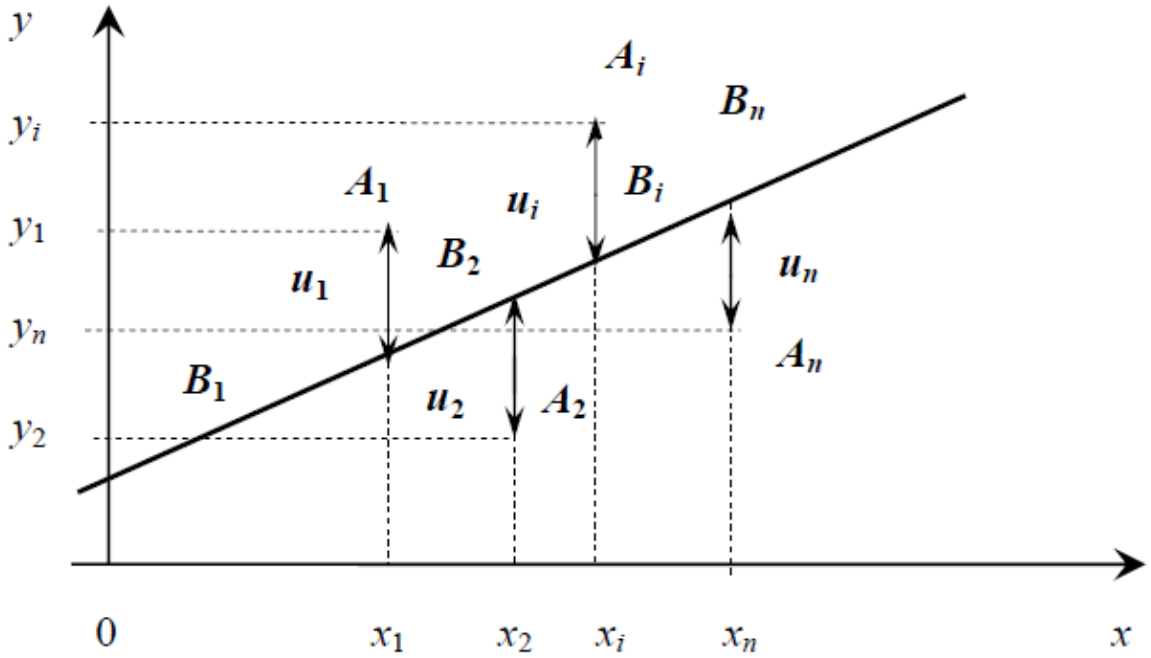


Рис. 9.3. Фактична залежність між y та x

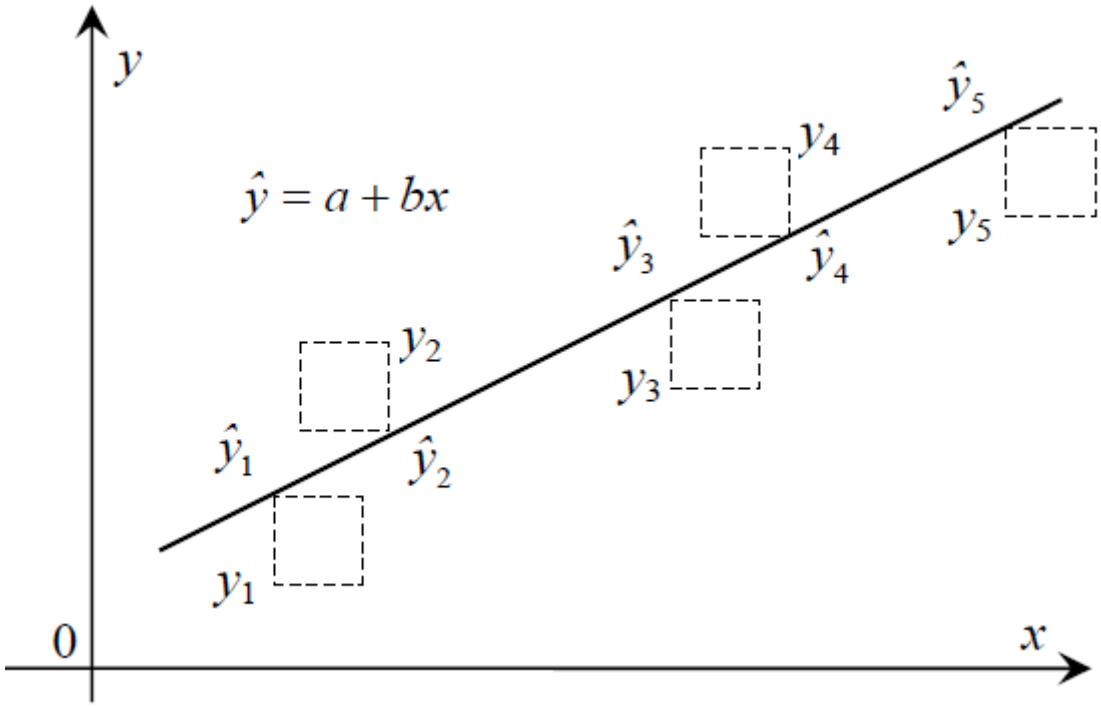


Рис. 9.4. Графічна інтерпретація методу найменших квадратів

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 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 + 0,051x. 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** η_{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 $r_{xy} \le 0{,}3$;

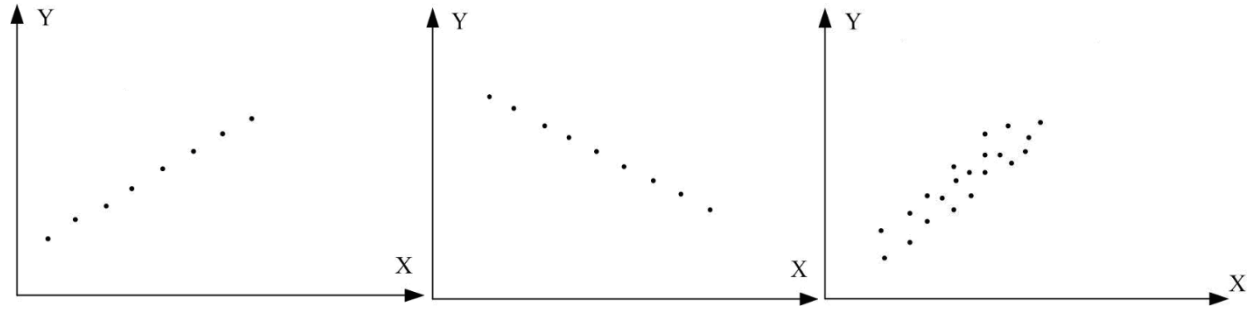
б) середній зв'язок, якщо r_{xy} = 0,31-0,65;

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

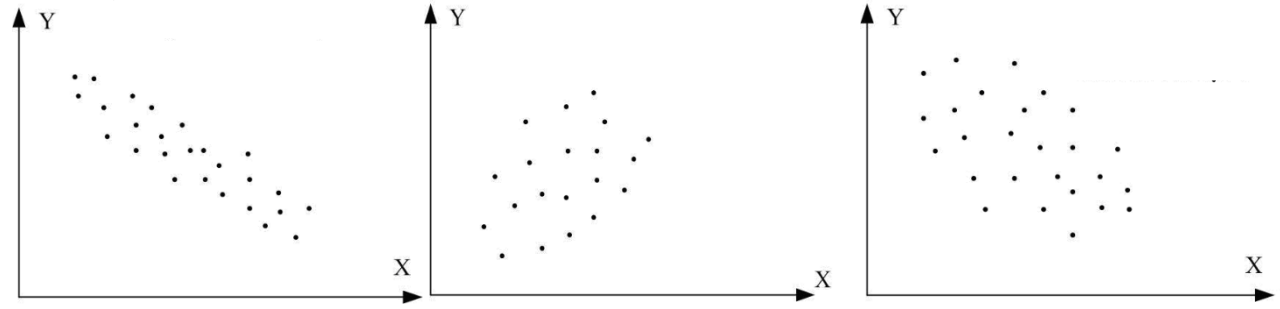
За значенням коефіцієнта кореляції можна зробити такі висновки:

- 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.

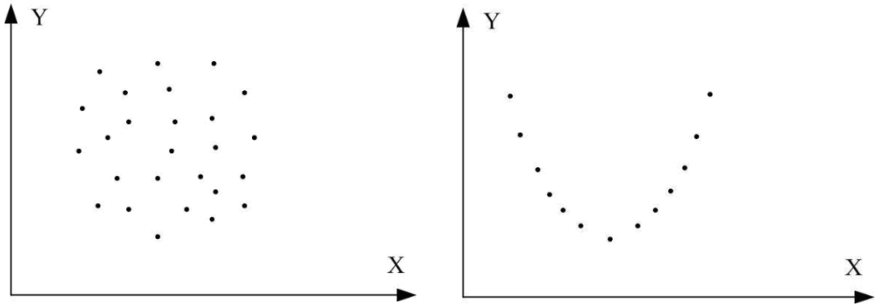
Відзначимо, що знак коефіцієнта кореляції r вказує на напрям зв'язку між ознаками x y, в той час як $|r_{xy}|$ характеризує щільність зв'язку.



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

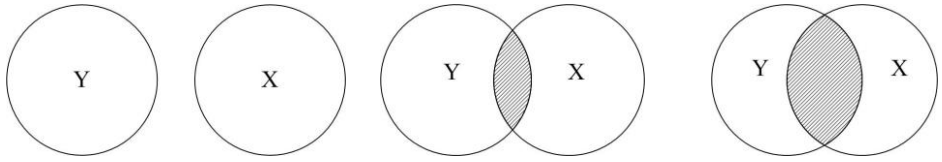
Рис. 9.5. Кореляційний коефіцієнт для різних випадків вибірок

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

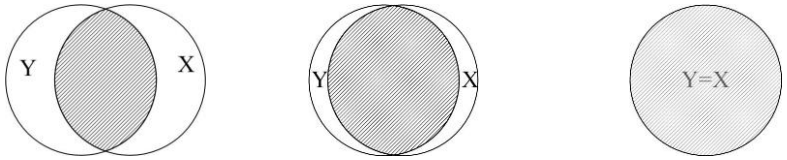
Звернемося зараз до розгляду питання якості підгонки лінії регресії до множини даних, тобто дослідимо, наскільки «добре» лінія вибіркової регресії підходить до цих даних. Із рис.9.3 видно, що якби всі спостереження знаходилися на лінії регресії, ми отримали б «точну підгонку», але на практиці це окремий 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. На рис. 9.5 бачимо, що при русі зліва направо область перекриття збільшується, тобто послідовно зростає частина варіації *Y*, з'ясована за допомогою *X*, - r^2 зростає. Коли перекриття немає, r^2, очевидно, дорівнює нулю, а коли відбувається повне перекриття, то $r^2=1$, оскільки 100% дисперсії *Y* пояснюється за допомогою *X*. 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).

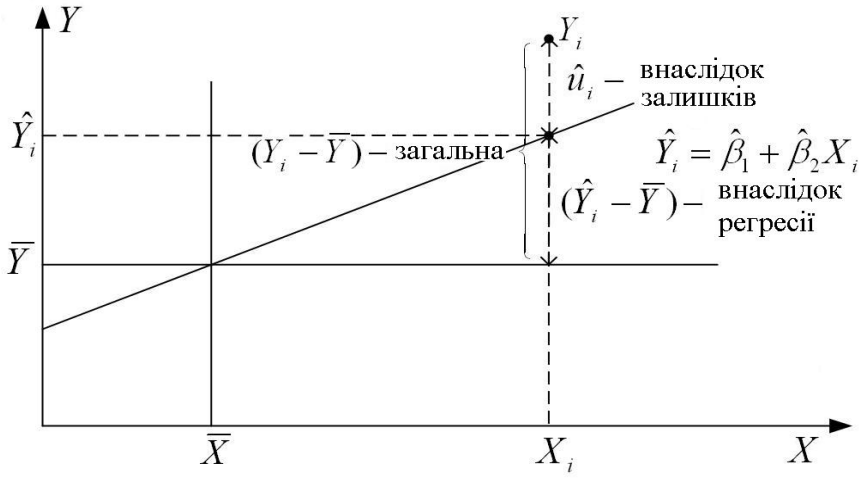


Рис. 9.6. Розбиття варіації Y_i на дві компоненти

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

$$r = \pm\sqrt{r^2} \qquad (9.22\ б)$$

або

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

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

Оскільки

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

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

можна записати вираз для коефіцієнта кореляції через параметр a_1 :

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

Запишемо формули для розрахунку параметрів економетричної моделі:

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

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

Необхідно відмітити, що при лінійній формі зв'язку коефіцієнт кореляції r_{xy} є оцінкою точності апроксимації, тобто адекватності моделі і дорівнює кореляційному відношенню η_{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) MS – середня сума квадратів (mean sum squares)			

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, %:

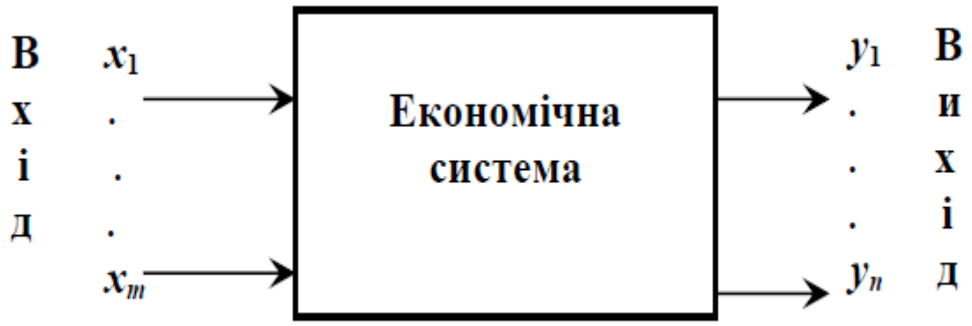
$$\varepsilon = \frac{100}{n}\cdot\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \qquad (9.27)$$

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. Класична лінійна багатофакторна модель

Більшість економічних показників формується під впливом багатьох різноманітних факторів. Їх виявлення та оцінювання ступеня цього впливу складає основу множинного регресійного аналізу.

Розглянемо схематичну інтерпретацію багатофакторної моделі економічної системи з вхідними факторами та вихідними показниками:



Припустимо, що деяка змінна y залежить від множини незалежних змінних $x_1, x_2, \ldots, x_n$. Тоді у випадку лінійної форми взаємозв'язку економетрична модель матиме вид:

$$y = b_0 + b_1 x_1 + \ldots + b_m x_m + u, \qquad (16.1)$$

де y – залежна змінна; $x_1, \ldots, x_m$ – незалежні змінні; $b_0, b_1, \ldots, b_m$ – параметри моделі, для яких потрібно буде знайти оцінки; u – збурення або залишок.

Тоді оціночне рівняння для окресленої моделі буде:

$$\hat{y} = a_0 + a_1 x_1 + \ldots + a_m x_m, \qquad (16.2)$$

де $\left\{ a_j,\ j = \overline{0;m} \right\}$ – оцінки невідомих параметрів $\left\{ b_j,\ j = \overline{0;m} \right\}$.

Нехай задано сукупність спостережень за залежною змінною $y = \left\{ y_i, i = \overline{1;n} \right\}$ і незалежною змінною $x_j = \left\{ x_{ij}, i = \overline{1;n} \right\}$, $j = \overline{1;m}$.

Як і у випадку парного регресійного аналізу, коефіцієнти регресії повинні розглядатися як випадкові змінні, випадковість компонентів яких зумовлена наявністю в моделях випадкового члена. Кожний коефіцієнт регресії обчислюється як функція значень y та незалежних змінних у вибірці, а y в свою чергу визначається незалежними змінними і вільним членом. Далі виберемо значення коефіцієнтів регресії таким чином, щоб сума квадратів відхилень фактичних даних від теоретичних була мінімальною. Цю вимогу можна представити таким чином:

$$F(a_0, a_1, \ldots, a_m) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 \to \min, \qquad (16.3)$$

де e_i є оцінкою для u_i та залишком в i-спостереженні, тобто різниця між фактичним значенням у спостереженні та розрахованим значенням $\hat{y}$ за рівнянням (16.2).

Таким чином, наша задача зводиться до мінімізації функції (16.3). Необхідною умовою цього є перетворення в нуль перших частинних похідних цієї функції стосовно кожної змінної $\{a_j,\ j = \overline{1; m}\}$.

Оскільки

$$F(a_0, a_1, \ldots, a_m) = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_{i1} - a_2 x_{i2} - \ldots - a_m x_{im})^2, \qquad (16.4)$$

то отримаємо таку систему нормальних рівнянь:

$$\begin{cases} \dfrac{\partial F}{\partial a_0} = -2\sum_{i=1}^{n} (y_i - a_0 - a_1 x_{i1} - \ldots - a_m x_{im}) = 0, \\ \dfrac{\partial F}{\partial a_1} = -2\sum_{i=1}^{n} (y_i - a_0 - a_1 x_{i1} - \ldots - a_m x_{im}) x_{i1} = 0, \\ \vdots \\ \dfrac{\partial F}{\partial a_m} = -2\sum_{i=1}^{n} (y_i - a_0 - a_1 x_{i1} - \ldots - a_m x_{im}) x_{im} = 0. \end{cases} \qquad (16.5)$$

Після виконання відповідних перетворень (16.5) матиме вигляд:

$$\begin{cases} na_0 + a_1\sum_{i=1}^{n} x_{i1} + a_2\sum_{i=1}^{n} x_{i2} + \ldots + a_m\sum_{i=1}^{n} x_{im} = \sum_{i=1}^{n} y_i, \\ a_0\sum_{i=1}^{n} x_{i1} + a_1\sum_{i=1}^{n} x_{i1}^2 + a_2\sum_{i=1}^{n} x_{i2}x_{i1} + \ldots + a_m\sum_{i=1}^{n} x_{im}x_{i1} = \sum_{i=1}^{n} y_i x_{i1}, \\ \vdots \\ a_0\sum_{i=1}^{n} x_{im} + a_1\sum_{i=1}^{n} x_{i1}x_{im} + a_2\sum_{i=1}^{n} x_{i2}x_{im} + \ldots + a_m\sum_{i=1}^{n} x_{im}^2 = \sum_{i=1}^{n} y_i x_{im}. \end{cases} \quad (16.6)$$

Розв'язавши систему рівнянь (16.6), отримаємо множину оцінок $\{b_j,\ j=\overline{0;m}\}$ для відповідних параметрів регресії $\{\beta_j;\ j=\overline{0;m}\}$. Побудуємо систему (16.6) для випадку m=2, тобто знайдемо ефект впливу двох факторів x_1 і x_2 на y.

Отримаємо:

$$\begin{cases} na_0 + a_1\sum_{i=1}^{n} x_{i1} + a_2\sum_{i=1}^{n} x_{i2} = \sum_{i=1}^{n} y_i, \\ a_0\sum_{i=1}^{n} x_{i1} + a_1\sum_{i=1}^{n} x_{i1}^2 + a_2\sum_{i=1}^{n} x_{i1}x_{i2} = \sum_{i=1}^{n} y_i x_{i1}, \ . \\ a_0\sum_{i=1}^{n} x_{i2} + a_1\sum_{i=1}^{n} x_{i1}x_{i2} + a_2\sum_{i=1}^{n} x_{i2}^2 = \sum_{i=1}^{n} y_i x_{i2}. \end{cases} \quad (16.7)$$

Таким чином, нами отримано систему рівнянь (16.7) з трьома невідомими величинами: a_0, a_1, a_2. ~~З першого рівняння маємо:~~

Властивості методу найменших квадратів. Властивості МНК для випадку множинної регресії збігаються з його властивостями для парної регресії.

1. Багатофакторна регресійна модель правильна для середніх точок $\bar{y}, \bar{x}_1, \bar{x}_2, \ldots, \bar{x}_m$.

Тобто для моделі

$$y_i = a_0 + a_1 x_{1i} + \ldots + a_m x_{1m} + e_i$$

має місце $a_0 = \bar{y} - a_1 \bar{x}_1 - \ldots - a_m \bar{x}_m$.

2. Середнє значення оцінки дорівнює середньому значенню фактичних даних: $\overline{\hat{y}} = \bar{y}$.

3. Сума помилок дорівнює нулю: $\sum_{i=1}^{n} e_i = \bar{e} = 0$.

4. Помилки e_i некорельовані з $x_{1i}, x_{2i}, \ldots, x_{mi}$, тобто має місце:

$$\sum_{i=1}^{n} e_i x_{1i} = \sum_{i=1}^{n} e_i x_{2i} = \ldots = \sum_{i=1}^{n} e_i x_{mi} = 0.$$

5. Помилки e_i некорельовані з $\hat{y}_i$, тобто має місце:

$$\sum_{i=1}^{n} e_i \hat{y}_i = 0.$$

6. Якщо правильні припущення класичної лінійної регресійної моделі, то МНК – оцінки є не тільки лінійними, без відхилень оцінками, але мають найменшу дисперсію.

Вираження для $a_1, a_2, \ldots, a_m$ стає досить складним, тому доцільно це зробити з допомогою матричної алгебри.

Розглянемо узагальнення моделі множинної регресії для m пояснювальних змінних з допомогою математичного апарату матричної алгебри.

Припустимо, що економетрична модель (16.1) у матричній формі має вигляд:

$$Y = XB + U, \qquad (16.10)$$

де Y – вектор значень залежної змінної; X – матриця значень незалежних змінних $(n \times m)$; B – вектор параметрів моделі; U – вектор залишків моделі.

Для постійної величини a_0 в (16.2) введемо фіктивну змінну $x_0 = 1$. Тоді (16.2) набуде вигляду:

$$\hat{y} = a_0 x_0 + a_1 x_1 + \ldots + a_m x_m. \qquad (16.11)$$

Результати досліджень $\{y_1, y_2, \ldots, y_n\}$ запишемо за допомогою вектора-стовпця Y, значення змінних $\{x_0, x_1, \ldots, x_m\}$ у вигляді матриці X розмірності $n \times (m+1)$, а решту складових у вигляді вектор-стовпців:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}; X = \begin{bmatrix} x_{10} & x_{11} & \cdots & x_{1m} \\ x_{20} & x_{21} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n0} & x_{n1} & \cdots & x_{nm} \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1m} \\ 1 & x_{21} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{bmatrix}; A = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix}; \hat{Y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix}; e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix},$$

де вектори A та e є операторами оцінювання відповідно векторів B і U.

Отже, економетрична модель у матричній формі матиме вигляд:

$$Y = XA + e. \qquad (16.12)$$

Для знаходження оцінок параметрів використаємо МНК.

Рівняння (16.12) представимо у вигляді: $e = Y - XA$. Далі запишемо суму квадратів залишків таким чином:

$$e(A) = \sum_{i=1}^{n} e_i^2 = e'e = \left(Y - \hat{Y}\right)'\left(Y - \hat{Y}\right) = (Y - XA)'(Y - XA) = \qquad (16.13)$$

$$= Y'Y - 2A'X'Y + A'X'XA \to min,$$

де символ штрих ($'$) означає операцію транспонування; $\hat{Y} = X \cdot A$.

Знайдемо частинну похідну окресленого виразу за компонентами вектора A і прирівняємо до нуля:

$$\frac{\partial(A)}{\partial A} = -2X'Y + 2X'XA = 0. \qquad (16.14)$$

Звідси, отримаємо систему рівнянь у матричній формі, якій повинен задовольняти вектор A при дотриманні вимоги (16.12):

$$X'XA = X'Y. \qquad (16.15)$$

Якщо матриця $X'X$ зворотна, тобто існує обернена їй матриця $(X'X)^{-1}$, то отримаємо розв'язком системи нормальних рівнянь вектор-стовпець шуканих оцінок параметрів регресії:

$$A = (X'X)^{-1}X'Y. \qquad (16.16)$$

На відміну від простої моделі регресії алгоритм визначення параметрів багатофакторної моделі є більш складним та трудомістким і містить у собі ряд послідовних етапів:

1. *Постановка задачі та апріорне дослідження економічної проблеми.*

Відповідно до мети дослідження на основі знань економічної теорії конкретизуються явища, процеси, залежності між якими потрібно оцінити. Тут насамперед необхідно чітко визначити економічні явища, встановити об'єкти та періоди дослідження. На початковому етапі повинні бути сформульовані та економічно обґрунтовані можливі гіпотези про залежність економічних явищ.

2. *Формування множини факторів і їх логічний аналіз.*

Проводиться ретроспективний аналіз відносно вибору найбільш характерних факторів, під дією яких формуються результативні показники або заданий економічний процес. При визначенні найбільш сприятливого числа змінних в регресійній моделі насамперед орієнтуються на розуміння професійно-теоретичного характеру процесу дослідження.

3. *Формування інформаційної бази даних.*

Для побудови моделі вхідна інформація може бути сформована у чотирьох видах:

- динамічні (часові) ряди;
- варіаційні ряди;
- просторова інформація, тобто інформація про функціонування декількох об'єктів в одному часовому періоді;
- змінна – таблична форма, тобто інформація про роботу декількох об'єктів за різні періоди.

Спостере-ження	Продуктивність праці, тис.грн. /люд.-год.	Фондомісткість продукції, тис.грн. / тис.грн.	Коефіцієнт плинності робочої сили, %	Рівень втрат робочого часу, %
1	60	30	13,0	15,0
2	61	35	12,5	14,3
3	58	33	12,0	12,0
4	59	34	11,0	12,8
5	62	36	10,0	13,0
6	63	38	9,0	12,5
7	65	40	8,5	11,0
8	60	41	8,2	11,5
9	68	45	8,0	10,0
10	69	45	5,5	9,0
11	70	46	5,0	8,0
12	72	48	4,7	7,5

Обсяг вибірки залежить від числа факторів, які входять до моделі з урахуванням вільного члена. Так для отримання статистично значимої моделі необхідно, щоб мінімальний обсяг вибірки становив:

$$n_{min} = 5(m + k),$$

де m – число факторів, які входять до моделі; k – число вільних членів у рівнянні.

4. *Специфікація функції регресії.*

На цьому етапі дослідження проводиться конкретний опис гіпотези про форму зв'язку (лінійна або нелінійна, проста або множинна і т.д.). З цією метою використовуються різні критерії для перевірки обґрунтованості гіпотетичного виду залежності. Крім цього перевіряються умови кореляційно-регресійного аналізу.

5. *Оцінювання параметрів регресійної моделі.*

З допомогою відповідного математичного апарату визначаються числові значення параметрів регресії та обчислюється ряд статистичних показників, які характеризують точність регресійного аналізу.

6. *Вибір головних факторів.*

Окреслений етап є основою для побудови багатофакторної моделі. На цьому етапі формується множина всеможливих факторів. Як результат, така модель містить велике число факторів. Вона, по-перше, незручна при проведенні економетричного аналізу, а по-друге, буде нестійкою.

Разом з тим, включення до моделі малого числа факторів приводить до порушення принципу адекватності процесів дослідження, що в свою чергу приводить до помилок при прийнятті рішень. Тому виникає необхідність у раціональному виборі певної кількості найбільш важливих і впливових факторів. При цьому проводиться аналіз факторів на мультиколінеарність.

Процес відбору факторів для моделі містить процедуру, яка складається із таких послідовних кроків:

- аналіз факторів на мультиколінеарність та її усунення;
- аналіз тісноти взаємозв'язку незалежних факторів із залежною змінною;
- аналіз бета-коефіцієнтів;
- перевірка коефіцієнтів регресії на статистичну значимість;
- аналіз факторів на керованість;
- побудова нової регресійної моделі без виключених факторів;

• дослідження доцільності виключення факторів із моделі з допомогою коефіцієнта детермінації.

Для реалізації шостого етапу доцільно використати метод покрокової регресії.

7. *Перевірка адекватності моделі.*

Цей етап аналізу містить:

• оцінку значимості коефіцієнта детермінації;

• перевірку якості підбору теоретичного виду рівняння;

• обчислення спеціальних показників, які використовуються для характеристики впливу окремих факторів на результативний показник.

8. *Економіко-математичний аналіз отриманих результатів та їх економічна інтерпретація.*

Результати регресійного аналізу порівнюються з гіпотезами, сформульованими на першому етапі дослідження, і оцінюється їх правдоподібність з економічної точки зору.

9. *Побудова прогнозних сценаріїв.*

Отримане рівняння регресії використовується для прогнозування сценаріїв розвитку відповідних економічних процесів чи явищ. Прогноз отримаємо внаслідок підстановки в модель певних значень факторів.

9.4. Передумови застосування методу найменших квадратів

При використанні МНК для знаходження оцінок параметрів лінійної багатофакторної моделі потрібно використовувати ряд передумов. Насамперед вони стосуються випадкової змінної e, яка є адитивною складовою, враховуючи помилки вимірювання та специфікації. Ці передумови мають загальний характер і не зв'язані ні з обсягом вибірки, ні з числом включених в аналіз змінних. Перелічимо найбільш суттєві умови, які необхідні для оцінки параметрів моделі МНК:

1. Математичне сподівання залишків дорівнює нулю.

При побудові функції регресії припускається, що результативна змінна Y залежить тільки від пояснювальних змінних $x_j \left(j = \overline{1;m}\right)$, які включені в регресію. Таким чином, при заданих значеннях змінних x_j на змінну Y не впливають жодні систематично діючі фактори та випадковості. Сумарний ефект від дії на залежну змінну

неврахованих факторів і випадковостей враховується збуреною змінною e. При цьому робиться припущення, що для фіксованих значень змінних x_j середнє значення збурення e рівне нулю: $M(e_i)=0$ або для матричної форми

$$M(e)=0. \qquad (16.17)$$

2. Гомоскедастичність (однакова дисперсія) для випадкових величин e_i.

Значення e_i вектора збурення e не залежні між собою і мають постійну дисперсію, тобто:

$$M(e_i^2)=\sigma_e^2. \qquad (16.18)$$

Ця властивість збурюючої змінної e називається гомоскедастичністю. Для кожного об'єкта спостережень у статиці, а при розгляді часових рядів – у різні періоди часу, ці невраховані фактори виявляють одинаковий вплив. Властивість гомоскедастичності може виконуватися лише за умови, що залишки є похибками вимірювання. Якщо залишки нагромаджують загальний вплив змінних, які не враховані в моделі, то зрозуміло, що дисперсія залишків не може бути сталою величиною. У такому випадку маємо справу з явищем гетероскедастичності.

3. Відсутність автокореляції між випадковими величинами e.

Значення випадкової змінної e попарно некорельовані, тобто коваріація збурюючих членів рівна нулеві:

$$M(e_i, e_{i-s})=0, \quad s\neq 0. \qquad (16.19)$$

При дотриманні пункту 6 окреслена умова зводиться до попарної незалежності. Вона є суттєвою у випадку, для якого вихідні дані є часовим рядом. Якщо збурюючі змінні містять тренд або циклічне коливання, то послідовні збурення, які діють у різні моменти часу, корельовані. Такий вид кореляції називають автокореляцією збурень або залишків.

Умови 2 і 3 можна узагальнити, використавши матричну форму запису:

$$M(ee')=\sigma_e^2 E, \qquad (16.20)$$

де E – одинична матриця порядку n. Добуток ee' є симетричною матрицею порядку n. Загальний вигляд математичного сподівання ee' записується так:

$$M(ee') = \begin{bmatrix} M(e_1^2) & M(e_1e_2) & \cdots & M(e_1e_n) \\ M(e_2e_1) & M(e_2^2) & \cdots & M(e_2e_n) \\ \vdots & \vdots & \vdots & \vdots \\ M(e_ne_1) & M(e_ne_2) & \cdots & M(e_n^2) \end{bmatrix}. \qquad (16.21)$$

Елементи, що стоять на головній діагоналі матриці 16.21 є дисперсіями, а поза головною діагоналлю – коваріаціями. Враховуючи умови 2 і 3, вираз (16.21) матиме вид:

$$M(ee') = \begin{bmatrix} \sigma_e^2 & 0 & \cdots & 0 \\ 0 & \sigma_e^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma_e^2 \end{bmatrix}. \qquad (16.22)$$

4. Незалежні змінні моделі утворюють лінійно незалежну систему векторів. Усі пояснювальні змінні, що входять до економетричної моделі мають бути незалежними між собою.

При знаходженні оцінок параметрів у регресії МНК система нормальних рівнянь має розв'язок тільки при існуванні оберненої матриці $(X'X)^{-1}$. Тому припускається, що $X'X$ – невироджена матриця, тобто $rgX = m+1$. Ця умова означає, що число спостережень (обсяг вибірки) повинно перевищувати число параметрів $(n > m)$, в іншому випадку оцінити параметри неможливо.

Таким чином, визначник матриці $X'X$ повинен бути відмінним від нуля: $\det(X'X) \neq 0$, що є необхідною і достатньою умовою існування оберненої матриці $(X'X)^{-1}$. Звідси випливає твердження, що між пояснювальними змінними не повинна існувати лінійна залежність, оскільки для такого випадку $rg = 0$.

Наявність лінійного зв'язку між пояснювальними змінними називається мультиколінеарністю. Це явище приводить до ненадійності оцінки параметрів моделі, робить їх чутливими до вибраної специфікації моделі та конкретного набору даних. Далі розглянемо методи виявлення мультиколінеарності та способи їх усунення.

5. Пояснювальні змінні не повинні корелювати із збурюючою змінною, тобто має місце:

$$M(x_{ij}, e_i) = 0 \text{ або } M(Xe) = 0. \qquad (16.23)$$

Зазначена умова полягає в тому, що змінна $x_j\left(j=\overline{1;m}\right)$ пояснює змінну Y, але зворотне твердження відсутнє, тобто змінна Y не пояснює змінні x_j. Отже, припускається існування односторонньої залежності змінної Y від x_j і відсутність взаємозв'язку.

6. Збурююча змінна розподілена нормально з параметрами $N\left(0,\sigma_e^2\right)$. Вважається, що вона суттєво не впливає на змінну Y. Ця умова одночасно означає, що залежна змінна Y чи змінні Y і $x_j\left(j=\overline{1;m}\right)$ розподілені нормально. При знаходженні оцінок параметрів регресії дотримання цієї умови не є обов'язковим. Використання статистичних критеріїв при перевірці значимості рівняння регресії та окремих коефіцієнтів регресії, побудова довірчих інтервалів допускає використання окресленої умови.

Отже, для знаходження оцінок параметрів моделі методом регресійного аналізу необхідно щоб виконувались перелічені вище передумови. Крім того, знайдені оцінки повинні володіти такими властивостями: незміщеності, обґрунтованості, ефективності та інваріантності.

9.5. Багатофакторна регресія та її оціночні характеристики

Процес побудови багатофакторної регресійної моделі потребує дотримання певної сукупності умов, як загального, так і особливого характеру відносно рівня адекватності. Такі умови, на жаль, часто ігноруються у деяких наукових розробках, при побудові прикладних моделей функціонування конкретних економічних систем та об'єктів. Особливо це стосується етапу апріорного аналізу.

Опишемо основні умови, які необхідно враховувати при побудові багатофакторних моделей достатнього рівня адекватності.

1) При відборі факторів для моделі, їх кількість повинна бути мінімальною, але достатньою для повної економічної характеристики результативного показника. Відібрані фактори описуються тільки однією характерною ознакою, тим самим включається дублювання показників. Фактори у зв'язку «причина-наслідок» повинні займати один і той же ієрархічний рівень – бути тільки первинними, чи тільки другорядними. Тут важливо, щоб їхнім виміром були не атрибутивні, а тільки кількісні ознаки.

2) При побудові регресійних моделей між вибраними показниками не повинен мати місце функціональний зв'язок.

3) Значний вплив на стійкість та достовірність моделі має репрезентативність і обсяг вибірки. В більшості випадків пропонується, щоби число одиниць досліджуваної сукупності

задовольняло умову шести та більш кратного перевищення його над числом незалежних змінних.

4) Вхідна інформаційна база повинна бути однорідною як в якісному, так і в кількісному відношенні. Якісна однорідність – це однорідність, наприклад, промислових підприємств регіону відносно випуску основного виду продукції, рівня оподаткування та кооперації. Кількісна однорідність полягає в досягненні відсутності у вибірковій сукупності аномальних результатів, наявність яких можна виявити з допомогою, наприклад, коефіцієнта варіації чи τ-критерію.

Якщо коефіцієнт варіації деякої сукупності становить більше 33%, то її можна вважати неоднорідною і з неї необхідно вивести аномальні результати. Спочатку слід виключити той об'єкт, для якого показник X_i чи Y_i мають найбільше відхилення в більшу або меншу сторону від свого середнього значення. Згадана процедура проводиться доти, поки сукупність не буде відповідати зазначеній умові для кожного факторного та результативного показників.

При використанні τ-критерію однорідність досягається тоді, коли після перевірки «крайніх» значень фактична величина окресленого критерію не стане меншою за його критичну.

Крім цього, для досягнення однорідності, можна використати метод групування.

5) Кожному значенню факторного показника X_i повинен відповідати нормальний розподіл результативного Y з однаковою дисперсією. Тобто емпіричний розподіл цих показників повинен бути близьким до нормального закону. Для перевірки цієї гіпотези можна використати критерії Пірсона, Колмогорова та ін. Їх, як правило, використовують для згрупованих даних, а для незгрупованих доцільно користуватися «правилом трьох сигм». Зміст його полягає в тому, що вхідні дані підлягають закону нормального розподілу, якщо в інтервалі $[-3\sigma; 3\sigma]$ знаходиться 99,8 % числа спостережень сукупності.

Якщо вхідні дані розподіляються за іншими законами (Пуассона, біноміальний та ін.), тоді проведення кореляційно-регресійного аналізу не дасть позитивних результатів. У такому випадку побудована модель буде мати фіктивний характер.

6) Апроксимуюча функція повинна бути найбільш адекватною до процесу дослідження та статистично значима. Тому вибір форми зв'язку є найбільш важливим і відповідальним моментом при побудові моделі, що в свою чергу потребує системного підходу в

процесі дослідження. Статистична значимість побудованої функції оцінюється з допомогою критеріїв Фішера та Стьюдента, коефіцієнтів кореляції та детермінації.

Якщо згадані критерії для різних форм зв'язку відрізняються між собою незначно, то перевагу необхідно віддати простішій функції, як наслідок, вона більш зрозуміла при інтерпретації її характеристичних параметрів.

7) Модель повинна бути позбавлена впливу мультиколінеарності, яка значно погіршує її якість. Для виявлення мультиколінеарності необхідно побудувати матрицю парних коефіцієнтів кореляції і на її основі оцінити тісноту взаємозв'язку між вибраними факторами. Наявність або відсутність колінеарного зв'язку можна оцінити з допомогою системи нерівностей:

$$\begin{cases} r_{YX_i} > r_{X_iX_j}, \\ r_{YX_j} > r_{X_iX_j}, \end{cases} \qquad (16.32)$$

де $r_{YX_i}, r_{YX_j}, r_{X_iX_j}$ – коефіцієнти парної кореляції взаємозв'язку відповідних змінних.

Якщо розраховані значення коефіцієнтів парної кореляції задовольняють цю систему, то можна стверджувати про відсутність колінеарності. В протилежному випадку необхідно позбутися мультиколінеарності з допомогою відомих процедур, які подамо в наступних розділах.

Незважаючи на трудомісткість процесу моделювання метод кореляційно-регресійного аналізу дає можливість одержати досить стійкі та надійні моделі, за умови дотримання перелічених вище умов.

Коефіцієнт множинної кореляції та детермінації.

Основним показником щільності кореляційного зв'язку між результативним показником Y і всіма незалежними змінними $x_j \left(j = \overline{1, m}\right)$, а також ступеня близькості вибраного виду математичної залежності до вибіркових даних є коефіцієнти множинної кореляції та детермінації.

Коефіцієнт множинної детермінації обчислюється за формулою:

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}. \qquad (16.33)$$

Враховуючи рівність

$$\sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \qquad (16.34)$$

маємо:

$$\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2=\sum_{i=1}^{n}(y_i-\overline{y})^2-\sum_{i=1}^{n}(y_i-\hat{y}_i)^2. \qquad (16.35)$$

Тоді вираз для R^2 матиме вигляд:

$$R^2=\frac{\sum_{i=1}^{n}(y_i-\overline{y})^2-\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2}=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\overline{y}_i)^2}. \qquad (16.36)$$

З останньої формули випливає, що якщо $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=0$, то $R^2=1$. Отже, якщо всі вибіркові значення показника розміщені на лінії регресії, то коефіцієнт множинної детермінації дорівнює одиниці. Далі можна зробити висновок: чим ближче вибіркові значення наближаються до лінії регресії, тим ближче R^2 наближаються до одиниці, а отже, тим більше варіація залежної змінної визначається варіацією незалежних факторів. Як бачимо, коефіцієнт множинної детермінації показує частку варіації результативної ознаки, яка знаходиться під впливом досліджуваних факторів, тобто визначає, яка частка варіації ознаки y врахована в моделі та викликана впливом вибраних факторів. Його числове значення змінюється від нуля до одиниці, тобто $R^2\in[0;1]$. Якщо R^2 прямує до нуля, то у вибірці відсутній взаємозв'язок між залежною та незалежними змінними.

Характерною особливістю коефіцієнта детермінації R^2 є те, що він – неспадна функція від кількості факторів, які входять до моделі. Отже, якщо кількість незалежних факторів зростає, то значення R^2 так само зростає.

Тому при співставленні між собою двох регресійних моделей для однакових залежних змінних, але з різною кількістю незалежних факторів, перевагу треба віддати тій моделі, для якої значення R^2 є більшим.

Доповнимо наведену вище методику оцінки якості побудованої моделі ще одним показником – частковим коефіцієнтом детермінації, який розраховується за формулою:

$$\Delta R_j^2 = \frac{1-R^2}{n-j} t_j^2, \qquad (16.43)$$

де j – індекс незалежної змінної, $j = \overline{1,m}$; ΔR_j^2 – частковий коефіцієнт детермінації для j-ої незалежної змінної; $t_j = \frac{a_j - a_j^*}{\sigma_{aj}}$, t_j-статистика для j-го коефіцієнта регресії (a_j^*–довільне задане та обґрунтоване число, наприклад, можна взяти $a_j^* = 0$); σ_{aj} – стандартна помилка оцінки a_j j-го регресійного коефіцієнта.

Частковий коефіцієнт детермінації використовується для обчислення граничного вкладу j-ої незалежної змінної у коефіцієнт детермінації. Він показує величину впливу j-го показника на якість моделі. Тобто наскільки зменшиться коефіцієнт детермінації, якщо j-ий фактор буде виведений з моделі.

Одним з основних показників тісноти кореляційного зв'язку результативного показника y з факторами $x_j\left(j = \overline{1,m}\right)$, а також мірою ступеня відповідності даних $\hat{y}_i\left(i = \overline{1,n}\right)$ є коефіцієнт множинної кореляції. Він визначається як коефіцієнт кореляції між y і $\hat{y}$ та має вигляд:

$$r_{y\hat{y}} = R = \frac{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)\left(\hat{y}_i - \overline{\hat{y}}\right)}{\sqrt{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2 \sum_{i=1}^{n}\left(\hat{y}_i - \overline{\hat{y}}\right)^2}}. \qquad (16.44)$$

Квадрат коефіцієнта множинної кореляції, як і у випадку простої регресії, називається коефіцієнтом детермінації, тобто має місце: $R = \sqrt{R^2}$.

Він характеризує тісноту лінійного зв'язку незалежних факторів $x_j\left(j = \overline{1,m}\right)$ із результативним показником y. Для множинного коефіцієнта кореляції (з врахуванням і без врахування коефіцієнта числа ступенів вільності) характерна така сама зміна числового значення, як і для випадку з коефіцієнтом детермінації.

9.6. Оцінка якості економетричних моделей

Якість лінійних економетричних моделей оцінюється стандартним для економіко-математичних задач способом: за адекватністю та точністю. Адекватність регресійних моделей може бути визначена на основі статистичного аналізу залишкової послідовності. При цьому їх значення отримують внаслідок підстановки у модель фактичних значень всіх включених до моделі факторів.

Для оцінки точності регресійних моделей використовуються статистичні критерії значущості Фішера та Стьюдента.

Стандартні помилки оцінок. Якість вибраної функції регресії можна оцінити на основі стандартних помилок або дисперсій залишків оцінок параметрів моделі.

Розглянемо спочатку алгоритм оцінки якості моделі з допомогою залишків послідовності. Стандартна помилка залишків називається також стандартною помилкою оцінки регресії у зв'язку з інтерпретацією величини u (збурення) як результату помилки специфікації функцій регресії.

Збурююча змінна u є випадковою з визначеним розподілом ймовірності. Математичне сподівання цієї змінної рівне нулю, а дисперсія – σ_u^2 на основі передумов застосування МНК. Таким чином, σ_u^2 – це дисперсія збурення у генеральній сукупності. Проте нам не відомі значення збурення. Про нього можна судити лише на основі залишків e. Знайдена за цими залишками дисперсія σ_e^2 буде оцінкою дисперсії збурюючої змінної. Тоді незміщену оцінку дисперсії збурення знайдено з формули:

$$\sigma_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-(m+1)} = \frac{1}{n-m-1} \cdot e' \cdot e. \qquad (16.57)$$

У знаменнику формули міститься число ступенів вільності $n-(m+1)$, де n – обсяг вибірки, m – число пояснювальних змінних. Такий вираз числа ступенів вільності зумовлений тим, що число залишків повинно задовольняти $m+1$ умовам.

Для безпосереднього проведення розрахунків можна використати такі формули:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} y_i^2 - a_0 \sum_{i=1}^{n} y_i - a_1 \sum_{i=1}^{n} x_{i1} y_i - \ldots - a_m \sum_{i=1}^{n} x_{im} y_i \qquad (16.58)$$

або в матричній формі:

$$\sum_{i=1}^{n} e_i^2 = e'e = Y'Y - A'X'Y.$$

Вираження сум у правій частині (16.58) містяться в робочій таблиці МНК, а оцінки параметрів вже знайдено. Беручи до уваги поняття коефіцієнта детермінації, стає зрозумілим фізичний зміст дисперсії (стандартного відхилення) залишків – це та частка загальної дисперсії σ_y^2, яка не може бути пояснена залежністю змінної від факторів $x_j\left(j=\overline{1,m}\right)$.

Опишемо основні чинники, від яких залежить стандартна помилка коефіцієнта регресії. Їх можна розділити на різні складові:

1) розсіювання залишків. Чим більша частка варіації значень змінної *Y*, непояснена її залежністю від *x*, тим більша стандартна помилка коефіцієнта регресії. Отже, чим більше фактичні значення змінної *Y* відхиляються від розрахункових значень регресії, тим менш точною є знайдена оцінка параметра;

2) розсіювання значень пояснювальної змінної *x*. Чим сильніше зазначене розсіювання, тим менша стандартна помилка коефіцієнта регресії. Звідси випливає, що при витягнутій хмарці точок на діаграмі розсіювання має надійнішу оцінку функції регресій, ніж при незначному скупченні точок, близько розміщених одна від одної.

3) об'єм вибірки. Чим більший об'єм вибірки, тим менша стандартна помилка коефіцієнта регресії. Тут існує безпосередній зв'язок оцінки параметра моделі з властивістю її асимптотичності незміщеності.

Значущість економетричної моделі. Для перевірки адекватності множинної регресійної моделі, як і у випадку парної регресії, використовується *F*-критерій Фішера. У цьому випадку нульова гіпотеза узагальнюється:

$$H_0: a_1 = a_2 = \ldots = a_m = 0.$$

Тоді альтернативною гіпотезою буде H_1 : хоча б одне значення a_j відмінне від нуля. У випадку невиконання гіпотези H_0 приймається гіпотеза H_1. Отже, не всі параметри незначною мірою відрізняються від нуля. Це свідчить про те, що включені до моделі фактори пояснюють змінну результативного показника.

Для перевірки гіпотези H_0 використовують F-критерій Фішера, з $(m-1)$ та $(n-m-1)$ ступенями вільності:

$$F = F_{m-1,n-m-1} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{m-1} : \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-m-1} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \cdot \frac{(n-m-1)}{(m-1)}, \quad (16.71)$$

де m – кількість незалежних факторів, які включено до моделі разом із фіктивним, n – загальна кількість спостережень.

Можна показати що має місце альтернативне представлення окресленого показника у матричній формі:

$$F_{m-1,n-m-1} = \frac{A'X'Y(n-m-1)}{(Y'Y - A'X'Y)(m-1)}. \quad (16.72)$$

Далі для заданого рівня значущості α і ступенів вільності $k_1=m-1$ і $k_2=n-m-1$ знаходимо табличне значення критерію Фішера – $F_{табл.}(k_1,k_2,\alpha)$. Знайдене розрахункове значення критерію $F_{m-1,n-m-1} = F_{розр.}$. Порівнюємо з табличним: якщо $F_{розр} > F_{табл.}$, тоді гіпотеза H_0 відхиляється і приймається альтернативна, що свідчить про адекватність побудованої моделі, іншими словами, підтверджується наявність істотного зв'язку між залежною та незалежними змінними побудованої економетричної моделі. У протилежному випадку вона приймається і модель вважається неадекватною. ♦

Значущість коефіцієнта детермінації.

Якість моделей множинної регресії можна оцінювати з допомогою коефіцієнтів детермінації.

При реалізації процедури перевірки значущості коефіцієнта детермінації висувається гіпотеза H_0 проти альтернативної H_1, зміст яких полягає в наступному.

H_0: суттєвої різниці між вибірковим коефіцієнтом детермінації та коефіцієнтом детермінації генеральної сукупності $R^2_{ген} = 0$ немає. Ця гіпотеза рівносильна гіпотезі H_0: $a_1 = a_2 = ... = a_m = 0$, тобто жодна із пояснювальних змінних, які включені до моделі, не виявляють суттєвого впливу на залежну змінну.

H_1: вибірковий коефіцієнт детермінації суттєво більший від коефіцієнта детермінації генеральної сукупності $R^2_{ген} = 0$. Прийняття гіпотези H_1 означає, що хоча б одна з m пояснювальних змінних, включені до моделі, виявляє суттєвий вплив на результативний показник.

Для оцінки значущості множинного коефіцієнта детермінації, як і у випадку парної регресії, використаємо статистику F-критерію Фішера.

Покажемо, що між F-критерієм Фішера та множинним коефіцієнтом детермінації існує зв'язок.

Враховуючи (16.33) та (16.71), одержимо:

$$F=\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2(n-m-1)}{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2(m-1)}=\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2-\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}\cdot\frac{n-m-1}{m-1}=$$

$$=\left[\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}:\left(1-\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}\right)\right]\cdot\frac{n-m-1}{m-1}=\frac{R^2}{1-R^2}\cdot\frac{n-m-1}{m-1}. \quad (16.73)$$

Вираз (16.73) показує, що якщо $R^2=0$, то $F=0$. Як бачимо, між F і R^2 існує взаємозв'язок. Оскільки F-критерій є мірою адекватності регресійної моделі, то він є мірою значущості коефіцієнта множинної детермінації. Зауважимо, що при зростанні R^2 значення F-критерію буде зростати.

Значення статистики, яке обраховане за (16.73), порівнюється з критичним значенням цієї статистики, знайденим за таблицями при заданому рівні значущості α та відповідних числах ступеня вільності. Якщо $F_{розр}>F_{табл}$, то знайдений коефіцієнт детермінації значно відрізняється від нуля. Цей висновок гарантується ймовірністю (1-α).

Значущість коефіцієнта кореляції. Відомо, що коефіцієнт множинної кореляції є вибірковою характеристикою, тому при перевірці якості побудованої моделі доцільно провести оцінку його значущості. Ця оцінка ґрунтується на *t*-статистиці Стьюдента:

$$t_{розр} = \frac{R\sqrt{n\text{-}m\text{-}1}}{\sqrt{1\text{-}R^2}}, \qquad (16.74)$$

де R – коефіцієнт множинної кореляції; R^2 – коефіцієнт детермінації; $(n–m–1)$ – число ступенів вільності. Обчислене значення *t*-статистики за формулою (16.74) порівнюють із критичним значенням $t_{k,\alpha}$, знайденим за таблицею *t*-розподілу для рівня значущості α і числа ступенів вільності $k=n–m–1$.

Прийняття чи відхилення гіпотези про значущість коефіцієнта множинної кореляції проводиться за тими ж правилами, що і для випадку парної регресії.

Якщо $|t_{розр}| > t_{k,б}$, то гіпотеза H_0 відхиляється і приймається гіпотеза H_1. Тому можна зробити висновок про значущість коефіцієнта множинної кореляції між залежною та незалежними змінними. У протилежному випадку приймається нульова гіпотеза.

Значущість коефіцієнта множинної кореляції можна також оцінювати на основі проведення процедури перевірки значущості коефіцієнта детермінації, оскільки між ними є зв'язок: $R = \sqrt{R^2}$.

Значущість оцінок параметрів моделі. Для розгляду значущості знайдених оцінок параметрів багатофакторної моделі побудуємо такі гіпотези:

H_0: $a_j = b_j$, що вказує на відсутність суттєвої різниці між оцінкою параметра регресії, отриманої за результатами вибірки, і дійсним значенням b_j (параметра регресії генеральної сукупності);

H_1: $a_j \neq b_j$, що вказує на наявність суттєвої різниці між оцінкою параметра регресії та відповідним параметром генеральної сукупності.

Альтернативна гіпотеза може бути сформульованою таким чином:

H_1: $a_j > b_j$ або $a_j < b_j$, тобто оцінка параметра суттєво більша або суттєво менша від параметра генеральної сукупності.

Для прийняття відповідних гіпотез використовується t-критерій Стьюдента:

$$t_{розр} = t_j = \frac{|a_j - b_j|}{\sigma_{a_j}}, \text{ при } k = n - m - 1, \qquad (16.75)$$

де a_j – оцінка параметра b_j, отримана за методом найменших квадратів; σ_{a_j} – середньоквадратичне відхилення оцінки j-го параметра; k – число ступенів вільності.

Обчислене значення t_j порівнюється із критичним значенням $t_{k,\alpha}$ знайденим за таблицями при заданому рівні значущості α і числом степенів вільності k. Якщо $t_j > t_{k,\alpha}$, то a_j значно відрізняється від b_j, тобто не можна припускати, що вибірка взята з генеральної сукупності з параметром регресії b_j.

На практиці буває дуже складно вказати завчасно числове значення параметра регресії b_j генеральної сукупності, тому часом доводиться висувати інше припущення:

H_0: $a_j = 0$, тобто пояснювальна змінна x_j не виявляє суттєвого впливу на залежну змінну y;

H_1: $a_j \neq 0$, тобто змінна x_j виявляє суттєвий вплив на y. У даному випадку для перевірки нульової гіпотези використовується t-статистика:

$$t_{розр} = t_j = \frac{|a_j|}{\sigma_{a_j}}, \qquad (16.76)$$

яка має k ступенів вільності.

У матричній формі (16.76) матиме вигляд:

$$t_j = \frac{|a_j|}{\sqrt{\sigma_e^2 C_{jj}}},\ j = \overline{1,m}\ , \qquad (16.77)$$

де C_{jj} – діагональний елемент матриці $(XX)^{-1}$; σ_e^2 – дисперсія залишків.

Величина $\sigma_{a_j} = \sqrt{\sigma_e^2 C_{jj}}$ називається стандартною оцінкою j-го параметра моделі.

Знайдене за (16.77) значення t_j порівнюють із значенням $t_{k,\alpha}$. Якщо $t_j > t_{k,\alpha}$, то відповідна оцінка параметра економетричної моделі є достовірною. Довірчі інтервали для параметрів b_j побудуємо на основі формули:

$$b_j = a_j \pm t_j \sqrt{\sigma_e^2 C_{jj}}\ . \qquad (16.78)$$

Довірчі інтервали параметрів моделі. Наявність точкових або асимптотних розподілів оцінок параметрів регресії та вибіркового коефіцієнта кореляції дають можливість провести оцінку значущості згаданих статистичних характеристик і побудувати інтервальні оцінки. Точкові оцінки визначаються одним числом, інтервальні – двома числами: кінцями інтервалу або його межами.

Надійність оцінки визначається ймовірністю, з якою робиться висновок, що побудований за результатами вибірки довірчий інтервал містить невідомий інтервал генеральної сукупності. Ймовірність інтервальної оцінки параметра називають довірчою і позначають через P, причому $P \in (0{,}95; 0{,}99)$. Тоді можна сподіватися, що в множині спостережень параметр генеральної сукупності буде правильно оцінений (довірчий інтервал покриє дійсне значення цього параметра) приблизно в $P \cdot 100\%$ випадках і лише в $(1-P)\cdot 100\%$ випадках буде помилковим. Ризик помилки визначається рівнем

значущості α, причому $\alpha = 1 - P$ і називається довірчим рівнем, який відповідає цьому інтервалу. В більшості випадків приймається $P = 0{,}95$, тим самим $\alpha = 0{,}05$ (ризик помилки складає 5 %).

Нехай параметр генеральної сукупності позначається через δ, а його оцінка – через α. Враховуючи наведене означення довірчого інтервалу, маємо:

$$P\left(\alpha - \nu\sigma_\alpha \le \delta \le \alpha + \nu\sigma_\alpha\right) = 1 - \alpha, \qquad (16.79)$$

де ν – довірчий множник, що означає частку стандартного відхилення, яка повинна бути врахована, щоби з наперед заданою ймовірністю P довірчий інтервал $\alpha \pm \nu\sigma_\alpha$ покривав параметр генеральної сукупності. Зрозуміло, що значення ν залежить від довірчої ймовірності P або від рівня значущості α, а також від обсягу вибірки. Якщо в дослідженні використовується t-статистика Стьюдента, тоді $\nu = t$ із відповідними ступенями вільності.

Довірчий інтервал можна подати у виді $\delta \in \left[\alpha - \nu\sigma_\alpha; \alpha + \nu\sigma_\alpha\right]$, де $\nu\sigma_\alpha$ називається точністю оцінки. Чим менша зазначена величина, тим менша ширина довірчого інтервалу. Це в свою чергу свідчить про високу якість вибраної оцінки.

Далі розглянемо процедуру побудови довірчих інтервалів для параметрів лінійної регресії. Для цього замінимо параметр α оцінкою параметра регресії $a_j\left(j = \overline{0,m}\right)$. Тоді покладемо $\nu = t_{k,\alpha}$, причому $k = n - m - 1$. Замість σ_α підставимо σ_{a_j}. Внаслідок таких підстановок одержимо довірчі границі, в середині яких при заданому рівні значущості α або при довірчій ймовірності $P = 1 - \alpha$ міститься невідомий параметр b_j генеральної сукупності, тобто маємо інтервал виду:

$$\left[a_j - t_j\sqrt{\sigma_e^2 c_{jj}}; a_j + t_j\sqrt{\sigma_e^2 c_{jj}};\right]. \qquad (16.80)$$

Довірчий інтервал коефіцієнта кореляції. При побудові довірчого інтервалу для коефіцієнта кореляції генеральної сукупності ρ необхідно використати перетворення Фішера, завдяки якому розподіл параметра r може бути наближено приведений до нормального:

$$Z = 0,5 \cdot ln\frac{1+r}{1-r} = 1,1513 \cdot lg\frac{1+r}{1-r}. \qquad (16.81)$$

Підставивши значення r у (16.81), отримаємо значення Z. Значення σ_Z знаходимо з формули:

$$\sigma_Z = \frac{1}{\sqrt{n-3}}, \qquad (16.82)$$

де n – об'єм вибірки.

Довірчий множник у цьому випадку є квантилем стандартного нормального розподілу $\lambda\alpha$. Тоді довірчі границі для величини Z при рівні значущості α будуть $Z \pm \sigma_Z$, а довірчий інтервал

$$\left[Z - \lambda_\alpha \sigma_Z; Z + \lambda_\alpha \sigma_Z\right].$$

Для рівні значущості α=0,05 квантиль стандартного нормального розподілу $\lambda_{0,05}$=1,96.

6. Прогнозування розвитку економічних процесів

Важливою метою економетричного моделювання є розробка прогнозування функціонування об'єкта дослідження або процесу на перспективу. Переважно термін прогнозування використовується в тих ситуаціях, коли необхідно передбачити стан економічної системи чи процесу в майбутньому. Тобто перед нами стоїть завдання побудови прогнозних сценаріїв їх функціонування та розвитку. Побудовані сценарії сприятимуть передбаченню ймовірнісних шляхів розвитку та поведінки деякої економічної системи і її складових на деяку перспективу. Така задача є досить складною та важливою при прийнятті стратегічних рішень на різних ієрархічних рівнях економічними процесами. Як уже відзначалося раніше, дані інформаційні бази можуть не мати часової структури, але і в цих випадках також може виникнути задача оцінки значень залежної змінної для деякої сукупності незалежних пояснювальних змінних, яких немає у точкових спостереженнях. Як побудову оцінки залежної змінної необхідно розуміти прогнозування в економетриці.

Процедура прогнозування має багато різних аспектів, серед яких можна виділити точкове та інтервальне прогнозування. У першому випадку – це конкретне число, а в другому – інтервал, в якому дійсне значення змінної знаходиться із заданим рівнем довіри. Крім цього, для часових рядів при знаходженні прогнозу суттєвим є наявність або відсутність кореляції в часі між помилками.

Відправною точкою в економетричному прогнозуванні є побудова економетричних моделей. Якщо побудована модель оцінена на адекватність за F-критерієм Фішера і в результаті є прийнятною, то її можна використовувати для прогнозу залежної змінної.

При використанні побудованої моделі для прогнозування робиться припущення про збереження на період прогнозування існуючих раніше взаємозв'язків між змінними.

Припустимо, що побудована економетрична модель такого виду:

$$Y = AX + e, \qquad (16.83)$$

де Y – вектор значень залежної змінної; X – матриця незалежних змінних розміром $n\times(m+1)$; A – вектор оцінки параметрів моделі; e– вектор оцінки залишків.

Використаємо окреслену модель у подальшому для знаходження прогнозних значень вектора Y_n, якщо сподіване значення незалежних змінних буде становити X_n. Як було зазначено вище, цей прогноз може бути точковим або інтервальним. Враховуючи (16.83), знайдемо незміщену оцінку прогнозу:

$$M\left[Y_n\left(X_n\right)\right] = X_n A + e. \qquad (16.84)$$

Можна довести, що дисперсія прогнозу у матричній формі буде:

$$\sigma_n^2 = D\left[Y_n\left(X_n\right)\right] = M\left\{\hat{Y}_n - M\left[\hat{Y}_n\left(X_n\right)\right]\right\}^2 = \sigma_e^2 X_n \left(X'X\right)^{-1} X_n'. \qquad (16.85)$$

Звідси, середньоквадратична (стандартна) помилка прогнозу буде:

$$\sigma_n = \sigma_e \sqrt{X_n \left(X'X\right)^{-1} X_n'} \ . \qquad (16.86)$$

Запишемо формулу для знаходження *t*-критерію розподілу Стьюдента:

$$t_\alpha = \frac{\hat{Y}_n - M\left(Y_n\left(X_n\right)\right)}{\sigma_e \sqrt{X_n \left(X'X\right) X_n'}} \qquad (16.87)$$

при $(n–m–1)$ ступенях вільності та рівні значущості α.

Довірчий інтервал для прогнозних значень має вигляд:

$$\hat{Y}_n - t_\alpha \sigma_e \sqrt{X_n (X'X)^{-1} X_n'} \le M(Y_n(X_n)) \le \hat{Y}_n + t_\alpha \sigma_e \sqrt{X_n (X'X)^{-1} X_n'}. \qquad (16.88)$$

Вираз $\hat{Y}_n = X_n A$ можна розглядати як точкову оцінку математичного сподівання прогнозного значення Y_n, а також як індивідуальне значення Y_n для вектора змінних X_n, що лежить за межами базового періоду.

Для знаходження інтервального прогнозу індивідуального значення $\hat{Y}_n$ насамперед необхідно знайти відповідну стандартну помилку $\sigma^2_{e(i)}$:

$$\sigma^2_{e(i)} = \sigma^2_e + \sigma^2_n = \sigma^2_e + \sigma^2_e X_n' \left(X'X\right)^{-1} X_n - \sigma^2_e \left(1 + X_n' \left(X'X\right)^{-1} X_n\right). \qquad (16.89)$$

Тоді інтервальний прогноз індивідуального значення визначиться за формулою:

$$\hat{Y}_n - t_\alpha \sigma_e \sqrt{1 + X_n (X'X)^{-1} X_n'} \le Y_n \le \hat{Y}_n + t_\alpha \sigma_e \sqrt{1 + X_n (X'X)^{-1} X_n'}. \qquad (16.90)$$