



Опря А.Т., доктор економічних наук,
Полтавська державна аграрна академія

МАТЕМАТИЧНА ОЦІНКА НАДІЙНОСТІ СТАТИСТИЧНИХ ПАРАМЕТРІВ У КОРЕЛЯЦІЙНОМУ МОДЕЛЮВАННІ ЕКОНОМІЧНИХ ЯВИЩ

Важливою складовою розробки економіко-статистичних моделей є обчислення ймовірних помилок та математична оцінка параметрів досліджуваної сукупності з метою встановлення їх надійності, оскільки вихідна статистична сукупність є вибірковою і не охоплює

всіх, без винятку, аналогічних, однорідних у якісному відношенні одиниць. Тому характеристики, обчислені на базі за такої (вибіркової) інформації, містять певні „помилки”, порівняно з відповідними характеристиками генеральної (всієї) сукупності.

Статистична наука розглядає критерії оцінки кореляційно-регресійних рівнянь із погляду кореляційного і дисперсійного аналізів, а також теорії помилок. Ці оцінки представлені трьома групами критеріїв: а) коефіцієнтів простої, множинної і часткової кореляції; б) критеріями вірогідності коефіцієнта кореляційного рівняння, що ґрунтується на дисперсійному аналізі (критерій Фішера); в) помилками коефіцієнтів рівняння регресії. Всі вони базуються на співвідношенні обсягів варіації різного типу, що визначаються відповідними сумами квадратів відхилень.

У статистичних розрахунках кореляційних моделей найбільш поширеними способами оцінки вибірових характеристик є критерії Ст'юдента (t-критерій) і Фішера (F-критерій).

Треба пам'ятати, що значення коефіцієнта парної кореляції є випадковою величиною, яка залежить від обсягу вибірки. Відомо, що зі зменшенням кількості спостережень надійність коефіцієнта кореляції – як вимірника тісноти зв'язку – знижується. Тому для оцінки його надійності необхідно вивчити вид розподілу коефіцієнта кореляції у вибірковій сукупності. Р. Фішером доведено, що зі збільшенням кількості спостережень розподіл одиниць спостереження наближається до нормального.

При досить значній кількості спостережень коефіцієнт кореляції можна вважати істотним,

Багатомірність ознак і складність економічних моделей потребує поглибленого кількісного аналізу на базі математико-статистичних методів, які дозволяють мати надійні оцінки об'єктивності одержаних результатів. Особливості кореляційно-регресійного методу аналізу економічних явищ потребують розгляду специфіки його використання при виконанні ним аналітичних функцій для забезпечення об'єктивності управлінських рішень.

якщо він перевищує стандартну помилку в три рази і більше, тобто якщо $|r| : \sigma > 3$.

У нечисленних вибірках істотність коефіцієнта кореляції оцінюється за допомогою t-критерію:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}},$$

де n – число парних спостережень.

Порівняння фактичного і табличного t при числі ступенів вільності $\nu = n - 2$ дає змогу оцінити істотність η_{yx} при обраному рівні ймовірності.

Так, якщо вибірка сукупність характеризується параметрами $n=66$; $r_{yx}=0,791$ значення наведеного критерію становитиме:

$$t = \frac{0,791\sqrt{6-2}}{\sqrt{1-0,791^2}} = \frac{0,791 \cdot 2}{\sqrt{0,3743}} = 10,343.$$

Така величина t дає підставу визнати кореляцію істотною, оскільки $t_{\text{табл.}} < t_{\text{факт.}}$.

Ймовірність і довірчий інтервал для часткового коефіцієнта кореляції визначають так само, як і для парних коефіцієнтів кореляції. Різниця лише в числі ступенів вільності. Для часткового коефіцієнта кореляції число ступенів вільності дорівнює: $\nu = N - K$, де K – порядок часткового коефіцієнта кореляції.

Істотність коефіцієнта множинної кореляції (R) перевіряють по t-критерію за співвідношенням: $t_R = R : S_R$, де S_R – середньоквадратична помилка коефіцієнта множинної кореляції, що

визначається за формулою: $S_R = \frac{1-R^2}{\sqrt{n-1}}$.

Якщо розрахункове t_R перевищує табличне значення, то коефіцієнт множинної кореляції визнається істотним. Так, при $n=66$ і $R=0,816$ середньоквадратична помилка становить:

$$S_R = \frac{1-0,816^2}{\sqrt{66-1}} = \frac{0,3341}{0,062} = 0,0414.$$

Розрахунковий показник t-критерію дорівнює: $t_R = 0,816 : 0,0414 = 19,710$ у даному випадку $t_{\text{факт.}}$ значно перевищує $t_{\text{табл.}}$

У практичних розрахунках для оцінки надійності коефіцієнта кореляції, як правило, використовують таблиці вірогідних значень коефіцієнтів кореляції для відповідної кількості спостережень, тобто даного обсягу вибірки і рівня ймовірності.

Оскільки сам коефіцієнт кореляції є свого роду критерієм надійності досліджуваної залежності факторних і результативних ознак, то висновки про його вірогідність може бути поширений і на інші параметри в цілому кореляційно-регресійної моделі економічних явищ.

Вірогідність кореляційного відношення (η) при криволінійній формі зв'язку оцінюється за допомогою F-критерію, який обчислюють за формулою: $F = \frac{\eta^2}{P-1} \cdot \frac{1-\eta^2}{n-P}$,

де P – кількість параметрів рівняння регресії.

Так, якщо досліджується криволінійна кореляційна модель собівартості продукції типу $y = a_0 + a_1 \lg x$ й одержано значення кореляційного відношення $\eta = 0,693$ в сукупності 66 одиниць, то розрахункова величина F-критерію становитиме:

$$F = \frac{0,693^2}{2-1} \cdot \frac{1-0,693^2}{66-2} = 60.$$

При рівні ймовірності 0,95 і ступені вільності $\nu_2 = 66 - 2 = 64$; $\nu_1 = 1$ величина F-критерію становить 4,0. Оскільки $F_p > F_m$ ($60 > 4$), значення кореляційного відношення визнається вірогідним.

В економічних розрахунках здебільшого користуються рівнянням прямої, хоча іноді крива лінія точніше відображає кореляційний зв'язок між досліджуваними ознаками. Проте таки й підхід буде об'єктивним, якщо різниця у коефіцієнтах детермінації ($\eta^2 - r^2$) виявиться неістотною. Для оцінки вірогідності коефіцієнтів детермінації використовують критерій Ст'юдента (при такому застосуванні його називають критерієм криволінійності):

$$t_{\eta^2-r^2} = \frac{\eta^2 - r^2}{\frac{2}{\sqrt{n}} \sqrt{(\eta^2 - r^2) - (\eta^2 - r^2)^2 (2 - \eta^2 - r^2)}}.$$

Чисельник наведеної формули містить оцінку вірогідності коефіцієнтів детермінації (міра криволінійності), знаменник – середню помилку такої різниці. Одержану величину t-критерію порівнюють із табличним його значенням. Так, для 66 одиниць спостереження величина $t_{\eta^2-r^2}$ не повинна перевищувати 2,00.

Множинний кореляційно-регресійний аналіз економічних явищ передбачає розрахунок часткових коефіцієнтів кореляції. Оцінку вірогіднос-

ті їх різниць здійснюють шляхом розрахунку середньої квадратичної помилки.

Для множинної трифакторної моделі формула помилки має вигляд: $\hat{S}_{y1.23} = \frac{1-r_{el.23}^2}{\sqrt{n-m}}$,

де $r_{y1.23}$ – частковий коефіцієнт кореляції; n – чисельність вибірки; m – кількість параметрів регресії (для трифакторної моделі $m=4$).

Розглянемо приклад часткових коефіцієнтів кореляції за множинною регресійною моделлю врожайності картоплі (y) у 101 сільськогосподарському підприємстві. Досліджуваними факторами прийнято: x_1 – внесення мінеральних добрив на 1 га орної землі, ц діючі речовини; x_2 – якісну оцінку ґрунту, балів; x_3 – вартість основних виробничих фондів на 1 га орної землі, грн. Одержані часткові коефіцієнти кореляції становлять: $r_{y1.23} = 0,885$; $r_{y2.13} = 0,778$; $r_{y3.12} = 0,882$.

Висуваємо нульову гіпотезу (H_0): зв'язок між досліджуваними ознаками наявний. Нульова гіпотеза повинна бути прийнятою, якщо фактичний коефіцієнт часткової кореляції виявиться більшим за його табличне значення і, навпаки, відхиленою, якщо фактичні значення виявляться меншими за табличні.

Вихідні дані: $n=101$; $n-m=101-4=97$; $r_{y1.23}=0,885$; $r_{y2.13}=0,778$; $r_{y3.12}=0,882$; $P=0,95$; $\eta_{табл.} = 0,1946$ ($\nu=97$; $P=0,95$).

Порівняння фактичних значень часткових коефіцієнтів кореляції з табличними свідчить про вірогідний зв'язок досліджуваних ознак, адже $r_{факт.} > r_{табл.}$.

Зробимо повторну оцінку вірогідності коефіцієнтів часткової кореляції, скориставшись t-критерієм Ст'юдента:

Фактичне нормоване відхилення становить:

$$t_\phi = \frac{r_{y1.23...n}}{\hat{S}}.$$

За умовами нашого прикладу маємо:

$$\hat{S}_{y1.23} = \frac{1-r_{y1.23}^2}{\sqrt{n-m}} = \frac{1-0,885^2}{\sqrt{101-4}} = 0,022;$$

$$t = \frac{r_{y1.23}}{\hat{S}_{y1.23}} = \frac{0,885}{0,0220} = 40,227;$$

$$\hat{S}_{y2.13} = \frac{1-0,778^2}{\sqrt{97}} = 0,614;$$

$$t = \frac{0,778}{0,614} = 12,671;$$

$$\hat{S}_{y3.12} = \frac{1-0,882^2}{\sqrt{97}} = 0,0226;$$

$$t = \frac{0,882}{0,0226} = 39,022.$$

Висновок щодо вірогідності часткових коефіцієнтів кореляції буде аналогічним попередньому, адже всі фактичні значення t-критерію виявилися вищими за табличні: при рівні ймовірності $P=0,95$ для чисельності вибірки 97 одиниць (101-4) табличне значення t-критерію становить 1,98. Отже, нульова гіпотеза про зв'язок між досліджуваними ознаками приймається.

Вихідною формулою для обчислення стандартної помилки оцінки рівняння регресії є:

$$\hat{S}_{yx} = \sqrt{\frac{\sum (y_i - \bar{y}_x)^2}{n-2}},$$

де $\hat{S}_{yx}$ – стандартна помилка оцінки за рівнянням регресії; y_i – рівні фактичних значень результативної ознаки; $\bar{y}_x$ – рівні теоретичних значень результативної ознаки; n – число одиниць спостереження.

При визначенні різниці між теоретичним (y_x) і фактичним (y_i) рівнями результативної ознаки обидві величини повинні відповідати одному й тому ж значенню аргументу (x_i).

Наведена формула не зовсім зручна в економічних розрахунках. Є простіша робоча формула, яка має майже всі розрахункові величини (крім $\sum y^2$), використовувати при рішенні системи нормальних рівнянь. Вона виведена з основної формули помилки й має вигляд:

$$\hat{S}_{yx}^2 = \frac{\sum y^2 - a_0 \sum y - a_1 \sum xy}{n-2}.$$

У наведених вище формулах з урахуванням втрат на число ступенів вільності замість n береться $n-2$, що викликано використанням двох констант (a_0 і a_1) у рівнянні регресії. Якщо це не враховувати, то стандартна помилка оцінки за рівнянням регресії виявиться менш точною. Пояснюється це тим, що квадрат стандартної помилки за рівнянням регресії (дисперсія) $\hat{S}_{yx}^2$, обчислений за вибірковими даними шляхом ділення суми квадратів відхилень на число одиниць вибірки (n), завжди менше відповідної характеристики у генеральній сукупності i . Такі показники в математичній статистиці називають зміщеними. Щоб одержати незміщені оцінки показників генеральної сукупності, замість n беруть число $n-2$.

У випадках, коли сукупність економічних показників достатня для того, аби вважати, що $n=n-2$, число ступенів вільності не враховують. Якщо розрахунки ведуть на масиві генеральної сукупності, але результати поширюються на ще

більшу сукупність (наприклад, при прогнозуванні), то у формулі помилки оцінки знаменником буде $n-2$; при поширенні даних лише на генеральну сукупність останній береться без поправки на число ступенів вільності.

Для рівняння залежності трудомісткості виробництва молока від рівня механізації очищення приміщень ($\bar{y}_x = 2,98 - 0,84x$) стандартна помилка становить:

$$\hat{S}_{yx}^2 = \frac{419,92 - 2,98 \cdot 156,38 - (-0,84) \cdot 98,97}{64-2} = 0,588.$$

Стандартна помилка за рівнянням регресії виражається у тих же самих одиницях виміру, що й індивідуальні дані результативної ознаки, і трактується так само, як стандартне відхилення від середньої арифметичної. У нашому прикладі абсолютна варіація фактичних показників трудомісткості виробництва 1 центнера молока щодо теоретичних їх значень становить 0,77 людино-дня.

Така оцінка використовується з метою побудови довірчих границь для оцінок за рівняння регресії. Для цього обирається досить висока ймовірність P_1 як гарантія того, що оцінювана одиниця y' не буде нижчою оцінки $\bar{y}_x$ більше, ніж на $t_{P_1} \hat{S}_{yx}$; обирається також досить висока ймовірність P_2 як гарантія того, що оцінювана одиниця y не буде вищою оцінки більше, ніж на $t_{P_2} \hat{S}_{yx}$. Тоді довірчі границі виражаться: верхня – $\bar{y}_x + t_{P_2} \hat{S}_{yx}$; нижня – $\bar{y}_x - t_{P_1} \hat{S}_{yx}$.

Величини t_{P_1} і t_{P_2} є „коефіцієнтами довіри”, що залежать від ймовірності P , з якою гарантуються розміри граничної помилки.

Якщо розподіл ймовірностей помилок $y - \bar{y}_x$ відомий, тоді ймовірність того, що випадкова одиниця (x, y) потрапить між цими границями, тобто в довірчу область

$$a_0 + a_1 x - t_{P_1} \hat{S}_{yx} < y < a_0 + a_1 x + t_{P_2} \hat{S}_{yx},$$

$$\text{дорівнює: } P = P_1 + P_2.$$

Якщо розподіл симетричний (наприклад, нормальний) і обрано $P_1 = P_2$, то $t_{P_1} = t_{P_2}$. У цьому випадку довірчі границі виражаються симетричними лініями навколо прямої регресії.

Розглянемо випадок визначення довірчих інтервалів (за відсутності додаткової інформації) теоретичного значення трудомісткості виробництва 1 центнера молока для $i_{\text{го}}$ підприємства, де показник рівня механізації очищення приміщень становить 0,86 (86%). Теоретичний рівень трудомісткості виробництва 1 центнера молока для цього підприємства дорівнює:

$$\bar{y}_x = 2,9837 - 0,8442 \times 0,86 \text{ людино-дня.}$$

За таблицями t-розподілу знаходимо, що при рівні ймовірності 0,95 і ступенях вільності варіації 64–2 $t=2,0$. Отже, гранична помилка становить: $\Delta_{yx} = t \cdot \hat{S}_{yx} = 2,0 \times 0,77 = 1,54$, а інтервали довіри, – відповідно, $2,26 - 1,54 \leq \bar{y}_x \leq 2,26 + 1,54$, або $0,72 \leq \bar{y}_x \leq 3,80$.

Одержані інтервали довіри відображають неточність передбачень теоретичного рівня результативної ознаки. Така неточність пояснюється тим, що досліджувана залежність трудомісткості продукції і рівня механізації виробничих процесів є не функціональною, а кореляційною, тобто неповною.

Подібно до рівня простої регресії, в рівнянні множинної регресії обчислене значення результативної ознаки ($\bar{y}_{x_1 \dots x_n}$) за будь-якої комбінації факторів ($x_1, \dots, x_n$) є лише кращою оцінкою, найбільш імовірною величиною очікуваного результату. Фактичні значення завжди певною мірою відрізняються від обчислених. Якщо використати геометричне тлумачення, то точки, що відображають фактичні величини в $K+1$ -мірному просторі, розкидані навколо гіперплощини $\bar{y}_{x_1 \dots x_k} = a_0 + a_1 x_1 + \dots + a_k x_k$. Звичайно, за таких умов необхідно користуватися помилкою оцінки за рівнянням регресії. Мірою помилки оцінки тут буде ϵ – стандартне відхилення фактичних величин результативної ознаки від їх відповідних теоретичних значень. Саму стандартну помилку оцінки визначають за формулою:

$$\hat{S}_{yx_1 \dots x_k}^2 = (S_y^2 - a_1 C_{yx_1} - a_2 C_{yx_2} - \dots - a_k C_{yx_k}) \frac{n}{n - (k + 1)},$$

де $\hat{S}_{yx_1 \dots x_k}^2$ – стандартна помилка оцінки за рівнянням множинної регресії;

C_{yx_k} – коваріація (середній добуток відхилень двох змінних від відповідних середніх; так, коваріація між y і x_1 визначається як

$$C_{yx_1} = \left[\sum (y - \bar{y})(x_1 - \bar{x}_1) \right] : n, \text{ між}$$

$$y \text{ і } x_2 - C_{yx_2} = \left[\sum (y - \bar{y})(x_2 - \bar{x}_2) \right] : (n - k) \text{ і т.д.);}$$

$a_1, a_2, \dots, a_k$ – параметри рівняння; n – кількість одиниць спостереження; k – кількість факторів, включених у кореляційну модель.

Як і при простій кореляції, стандартна помилка за рівнянням множинної регресії використовується для побудови інтервалів довіри $\bar{y}_{x_1 \dots x_k}$. Вона завжди виражається в одиницях виміру результативної ознаки.

У таблиці наведено параметри множинної кореляційної моделі прибутку від м'ясного скотар-

ства (66 підприємств Лісостепу). У дослідженні аналізується дія таких факторів: x_1 – обсяг виробництва яловичини одним підприємством, ц; x_2 – повна собівартість 1 центнера яловичини, грн. x_3 – середня ціна реалізації 1 центнера яловичини, грн.; x_4 – щільність поголів'я на 100 га сільськогосподарських угідь, гол.; x_5 – виробництво яловичини на 1 людино-день, кг. За результативну ознаку прийнято середньорічний розмір прибутку з розрахунку на одне підприємство ($\bar{y}_{x_1 \dots x_5}$). У рівнянні зв'язку одержано такі параметри при x : $a_1=45,02$; $a_2=233,92$; $a_3=764,48$; $a_4=271,89$; $a_5=264,82$. У цій же таблиці наведено також значення даних параметрів при виключенні одного з досліджуваних факторів.

Для знайденого рівняння регресії стандартна помилка оцінки становить:

$$\hat{S}_{yx_1 \dots x_5} = (1085307136 - 45 \times 14846543 - 234 \times 209260 - 784 \times 247870 - 271 \times 114909 - 265 \times 41509) \times$$

$$\times \frac{66}{65 - (5 + 1)} = 144953 \text{ тис.};$$

$$\hat{S}_{yx_1 \dots x_5} = 12040.$$

Розглянемо, як можна на прикладі підприємств одного адміністративного району розрахувати найбільш імовірний рівень результативної ознаки, використавши одержане рівняння множинної регресії. Вихідні дані по підприємствах у середньому по досліджуваних факторах становлять: x_1 – 1202 ц; x_2 – 101,68 грн.; x_3 – 125,52 грн.; x_4 – 29,2 гол.; x_5 – 9,7 кг.

Підставивши відповідні значення величин $x_1, \dots, x_5$, одержуємо $\bar{y}_{x_1 \dots x_5} = 37,296$.

Порівняння фактичного рівня з розрахунковим дає змогу встановити вплив не включених у рівняння регресії факторів.

Фактичний рівень щодо теоретичного дорівнює: $(34,690 : 37,296) \times 100 \geq 93\%$, тобто не досягає його на 7%. Різниця $37,296 - 34,690 = 2,606$ менша від стандартної помилки оцінки за рівнянням ($\hat{S}_{yx_1 \dots x_5} = 12,040$), тому її можна вважати результатом дії випадкових причин. Зауважимо, що випадковий характер розбіжності теоретичного та фактичного рівнів цим не доводиться, а встановлюється, що така розбіжність не є істотною.

У таблиці наведено значення стандартних помилок оцінки при різних можливих комбінаціях факторів за умов виключення одного з них. Як бачимо, найбільший розмір помилки одержано при виключенні фактора обсягу виробництва продукції (x_1).

МЕТОДИ І МЕТОДИКИ ДОСЛІДЖЕНЬ

Числові характеристики для обчислення стандартної помилки за рівнянням множинної регресії

Дисперсія загальна	$\hat{S}_y^2$	1085307,136	Примітка
Коваріація	C_{yx_1}	14846,543	
$C_{yx} = \frac{\sum (y - \bar{y})(x - \bar{x})}{n}$	C_{yx_2}	209,260	
	C_{yx_3}	247,870	
	C_{yx_4}	114,909	
	C_{yx_5}	41,509	
Коефіцієнти регресії за різних комбінацій факторів	a_1	1-4-248,80 2-46,60 5-45,48 3-50,70 6-45,02	
	a_2	1-430,16 4-248,80 2-5-235,13	
	a_3	1-1376,84 4-727,52 2-872-19 5-809,59 3-6-764,48	Для уникнення громіздкості таблиці комбінації факторів пронумеровано:
	a_4	1-753,38 4-2-350,53 5-286,99	
		3-23,09 6-271,89	
			1-2345,
			2-1345,
		1-1543,67 4-335,41	3-1245,
	a_5	2-284,76 5-	4-1235,
		3-872,63 6-264,82	5-1234,
			6-12345
Стандартна помилка оцінки за рівнянням за різних комбінацій факторів	$\hat{S}_{2345}$	23,317	
	$\hat{S}_{1345}$	12,826	
	$\hat{S}_{1245}$	16,052	
	$\hat{S}_{1235}$	14,713	
	$\hat{S}_{1234}$	12,044	
	$\hat{S}_{12345}$	12,040	