Тема 2. Поняття та етапи кореляційного аналізу

План.

1. Етапи кореляційного аналізу.
2. Методи вивчення парної кореляції.
3. Множинний кореляційний аналіз.

2.1 Етапи кореляційного аналізу.

На практиці далеко не всі економічні явища й процеси можуть вивчатися за допомогою методики детермінованого факторного аналізу, тому що в більшості випадків їх не можна звести до функціональних залежностей, коли єдиному значенню факторного показника відповідає єдине значення результативного показника.

Частіше в економічних дослідженнях зустрічаються стохастичні залежності, які відрізняються приблизністю, невизначеністю. Вони проявляються тільки в середньому по значній кількості об'єктів (спостережень). Тут кожній величині факторного показника (аргументу) може відповідати кілька значень результативного показника (функції).

Наприклад, збільшення фондоозброєності праці робітників дає різний приріст продуктивності праці на різних підприємствах навіть за інших рівних умов. Це пояснюється тим, що всі фактори, від яких залежить продуктивність праці, діють у комплексі, взаємозалежно.

Взаємозв'язок між факторами й результативним показником проявляється, якщо для дослідження взяти велика кількість спостережень. Тоді вплив інших (несуттєвих) факторів згладжується, нейтралізується. Це дає можливість установити взаємозв'язок між досліджуваними явищами.

Кореляційний (стохастичний) зв'язок між факторами й результативним показником - це неповна, імовірнісна залежність, що проявляється тільки при великій кількості спостережень.

Економічні явища й процеси господарської діяльності підприємств залежать від великої кількості факторів. Як правило, кожний фактор окремо не визначає досліджуване явище у всій повноті. Тільки комплекс факторів у їхньому взаємозв'язку може дати більш-менш повне подання про характер досліджуваного явища.

Завдання кореляційного аналізу:

- Установлення абсолютного й відносного ступеня залежності впливу факторів на величину результативного показника;
- Розрахунок резервів підвищення рівня досліджуваного показника;
- Планування й прогнозування його величини.

Етапи кореляційного аналізу:

1. Визначення факторів, які впливають на досліджуваний показник і відбір найбільш істотних.
2. Оцінка вихідної інформації.
3. Вивчення характеру зв'язку й моделювання рівняння регресії між факторами й результативним показником, що найбільше точно виражає сутність досліджуваної залежності.
4. Статистична оцінка результатів кореляційного аналізу і їхнє практичне застосування.

Етап 1. *Відбір факторів для кореляційного аналізу є дуже важливим моментом в економічному аналізі. Від того, наскільки правильно він зроблений, залежить точність висновків за підсумками аналізу.*

- Необхідно відбирати самі значимі фактори, які впливають на результативний показник. Критерієм такого відбору є критерій надійності Стьюдента. Якщо надійність фактору менше табличного - фактор у розгляд не приймається.

– Взаємозалежні фактори в кореляційну модель не включаються. Якщо парний коефіцієнт кореляції між двома факторами більше 0,85, то один з факторів необхідно виключити з розгляду.

- Небажано включати фактори, зв'язок яких з результативним показником носить функціональний характер.

Для забезпечення цих умов для вихідних статистичних даних необхідно розраховувати ряд відповідних коефіцієнтів.

Коефіцієнт кореляції (r_{xy}) – визначає тісноту зв'язку між факторними й результативними показниками, r_{xy} є [-1; 1]. Чим ближче його величина до 1, тим більше тісний взаємозв'язок між досліджуваними явищами, і навпаки:

$$r_{xy} = \frac{\mathrm{cov}(x, y)}{\sigma_x \sigma_y}$$

де cov(x,y) – коваріація, показує ступінь узгодження коливань величин x и y, cov(x,y) є [-∞; +∞];

σ_x, σ_y – середньоквадратичне відхилення.

$$\mathrm{cov}(x, y) = \frac{1}{n}\sum (x_i - x_{средн})(y_i - y_{средн});\ \sigma_x = \sqrt{\frac{\sum (x_i - x_{средн})^2}{n}};\ \sigma_y = \sqrt{\frac{\sum (y_i - y_{средн})^2}{n}}$$

де n – обсяг вибірки досліджуваного явища.

Тоді, коефіцієнт кореляції має вигляд:

$$r_{xy} = \frac{\sum (x_i - x_{средн})(y_i - y_{средн})}{\sqrt{\sum (x_i - x_{средн})^2 * \sum (y_i - y_{средн})^2}}$$

Коефіцієнт детермінації (d) показує, на скільки відсотків результуючий показник залежить від досліджуваного фактору:

$$d = r_{xy}^2$$

При вивченні тісноти зв'язку треба мати на увазі, що величина коефіцієнтів кореляції є випадковою, залежною від обсягів вибірки. Відомо, що зі зменшенням кількості спостережень надійність коефіцієнтів кореляції падає, і навпаки, при збільшенні кількості спостережень надійність коефіцієнтів кореляції зростає.

Значимість коефіцієнтів кореляції перевіряється за критерієм Стьюдента:

$$t_{расчетн} = \frac{r_{xy}}{\sigma_r}$$

де σ_r – середньоквадратична помилка коефіцієнта кореляції, визначається по формулі:

$$\sigma_r = \frac{1 - r_{xy}^2}{\sqrt{n-1}}$$

Якщо $t_{расчетн} > t_{таблич}(n-1, P)$ зв'язок між результативним і факторним показником є надійною (P – рівень довірчої ймовірності), інакше – фактор х необхідно виключити з розгляду.

Етап 2. Зібрана вихідна інформація повинна бути перевірена на однорідність і достатність обсягу.

Критерієм однорідності інформації служить коефіцієнт варіації (V), що розраховується по кожному факторному й результативному показнику:

$$V = \frac{\sigma_x}{x_{средн}} 100\%$$

Якщо V > 33% - дані неоднорідні й не можуть використовуватися для подальших розрахунків.

На підставі найвищого коефіцієнта варіації розраховується мінімальний необхідний обсяг вибірки:

$$n_{\min} = \frac{V_{\max}^2 * t_{таблич}^2}{m^2}$$

де m – показник точності розрахунків, виражений в %. Прийнятний рівень - 5% - 8%.

Якщо (V < 33%) і ($n >= n_{min}$) то дані можуть бути використані для подальшого аналізу.

Етап 3. Метою даного етапу є побудова рівняння регресії, тобто знаходження функціональної залежності результативного показника від досліджуваних факторів.

Розрізняють:

- Парну кореляцію - зв'язок між двома показниками, один із яких є факторним, а іншої - результативним:

$$Y_{расчетн} = f(x)$$

Процес побудови однофакторної регресійної моделі розглянутий у ПИТАННІ 2.

- Множинну кореляцію - зв'язок між декількома факторними показниками й результативними:

$$Y_{расчетн} = f(x_1, x_2, \ldots x_m)$$

Процес побудови багатофакторної регресійної моделі розглянутий у ПИТАННІ 3.

Етап 4. *Для того щоб переконатися в надійності побудованої регресійної моделі й правомірності її використання для практичної мети, необхідно дати їй статистичну оцінку надійності.*

Для оцінки адекватності отриманого рівняння регресії реальним даним використовується критерій Фишера:

$$F_{расчетн} = \frac{\sum (Y_{i,расчетн} - Y_{среднее,расчетн})^2}{m} * \frac{n-m-1}{\sum (Y_i - Y_{i,расчетн})^2}$$

де m – кількість досліджуваних факторів.

Отримана регресійна модель адекватна реальним даним, якщо $F_{расчетн} > F_{табличн}$ (P, m, n-m).

Вплив кожного фактору для регресійної моделі виду $Y=b_0+b_1x_1+b_2x_2+\ldots+b_nx_n$ на приріст результативного показника розраховується по формулі:

$$\Delta Y_i = b_i * \Delta x_i$$

У випадку парної нелінійної кореляції необхідно в отримане рівняння підставити спочатку фактичний рівень факторного показника, а потім можливий (прогнозний) і зрівняти отримані результати.

2.2 Методи вивчення парної кореляції.

Метою розгляду даного питання є побудова однофакторної регресійної моделі, що найбільше точно виражає сутність досліджуваної залежності.

Процес побудови однофакторної регресійної моделі містить у собі наступні етапи:

1. Підбор відповідного типу математичної залежності, що найкраще відображає характер досліджуваного зв'язку (лінійна, нелінійна).

Цей етап відіграє важливу роль у кореляційному аналізі, тому що від правильного вибору рівняння регресії залежить результат рішення завдання.

Тип залежності визначається за допомогою побудови діаграм розсіювання, рис. 3.1:

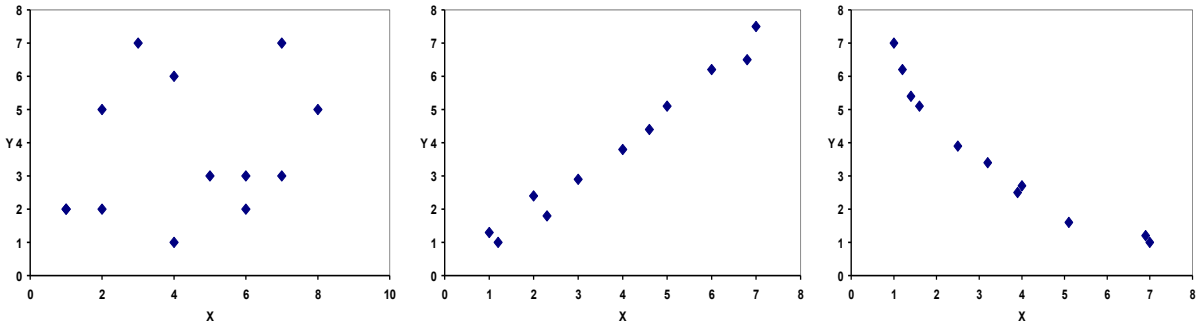

Рис. 3.1 Діаграми розсіювання

По виду скупчення крапок можна висунути гіпотезу про форму залежності між змінними.

Лінійна залежність між двома показниками характеризується рівнянням прямої:

$$Y_{расчетн} = a + bx$$

де x – факторний показник;

$Y_{расчетн}$ – результативний показник;

a і b - параметри рівняння регресії, які потрібно відшукати.

Це рівняння описує такий зв'язок між двома ознаками, при якій зі зміною факторного показника на певну величину спостерігається рівномірне зростання або зменьшення значень результативного показника.

Нелінійна залежність може характеризуватися:

– Параболою другого порядку: $Y_{расчетн} = a + bx + cx^2$;

– Гіперболою: $Y_{расчетн} = a + {}^{b}/_{x}$ і т.п.

2. Розрахунок параметрів рівняння регресії методом найменших квадратів (**МНК**), тобто $\sum (Y_i - Y_{i,расчетн})^2 \longrightarrow \min$

<u>У випадку лінійної залежності маємо</u>:

$$Y_{расчетн} = a + bx,\ S(a,b) = \sum (Y_i - Y_{i,расчетн})^2 \longrightarrow \min$$

маємо: $S(a,b) = \sum (Y_i - a - bx_i)^2 \longrightarrow \min$

Функція буде приймати мінімальне значення, коли часткові похідні будуть дорівнюють нулю:

$$\begin{cases} \frac{\partial S}{\partial a} = 0 \\ \frac{\partial S}{\partial b} = 0 \end{cases} \longrightarrow \begin{cases} \frac{\partial S}{\partial a} = 2\sum (Y_i - a - bx_i) * (-1) = 0 \\ \frac{\partial S}{\partial b} = 2\sum (Y_i - a - bx_i) * (-x_i) = 0 \end{cases}$$

$$\begin{cases} \sum (Y_i - a - bx_i) = 0 \\ \sum (Y_i - a - bx_i) * (-x_i) = 0 \end{cases} \longrightarrow \begin{cases} \sum a + b\sum x_i = \sum Y_i \\ a\sum x_i + b\sum x_i^2 = \sum x_i Y_i \end{cases}$$

$$\begin{cases} na + b\sum x_i = \sum y_i \\ a\sum x_i + b\sum x_i^2 = \sum x_i y_i \end{cases}$$ звідси знаходимо a і b.

$$a = \frac{\sum y_i \sum x_i^2 - \sum x_i \sum x_i y_i}{n\sum x_i^2 - \sum x_i \sum x_i} \quad b = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \sum x_i \sum x_i}$$

<u>У випадку параболи другого порядку</u> $Y_{расчетн} = a + bx + cx^2$ маємо:

$$Y_{расчетн} = a + bx + cx^2,\ S(a,b,c) = \sum (Y_i - Y_{i,расчетн})^2 \longrightarrow \min$$

тобто $S(a,b,c) = \sum (Y_i - a - bx_i - cx_i^2)^2 \longrightarrow \min$

Функція буде приймати мінімальне значення, коли часткові похідні будуть дорівнюють нулю:

$$\begin{cases} \frac{\partial S}{\partial a} = 0 \\ \frac{\partial S}{\partial b} = 0 \\ \frac{\partial S}{\partial c} = 0 \end{cases} \longrightarrow \begin{cases} \frac{\partial S}{\partial a} = 2\sum (Y_i - a - bx_i - cx_i^2) * (-1) = 0 \\ \frac{\partial S}{\partial b} = 2\sum (Y_i - a - bx_i - cx_i^2) * (-x_i) = 0 \\ \frac{\partial S}{\partial c} = 2\sum (Y_i - a - bx_i - cx_i^2) * (-x_i^2) = 0 \end{cases}$$

$$\begin{cases} \sum (Y_i - a - bx_i - cx_i^2) = 0 \\ \sum (Y_i - a - bx_i - cx_i^2) * (-x_i) = 0 \\ \sum (Y_i - a - bx_i - cx_i^2) * (-x_i^2) = 0 \end{cases} \longrightarrow \begin{cases} na + b\sum x_i + c\sum x_i^2 = \sum Y_i \\ a\sum x_i + b\sum x_i^2 + c\sum x_i^3 = \sum x_i Y_i \\ a\sum x_i^2 + b\sum x_i^3 + c\sum x_i^4 = \sum x_i^2 Y_i \end{cases}$$

звідси знаходимо a, b і c.

<u>У випадку гіперболи $Y_{расчетн} = a + {}^{b}/_{x}$ маємо</u>:

$$Y_{расчетн} = a + \frac{b}{x},\ S(a,b) = \sum (Y_i - Y_{i,расчетн})^2 \longrightarrow \min$$

тобто $S(a,b) = \sum (Y_i - a - \frac{b}{x_i})^2 \longrightarrow \min$

Функція буде приймати мінімальне значення, коли часткові похідні будуть дорівнюють нулю:

$$\begin{cases} \frac{\partial S}{\partial a} = 0 \\ \frac{\partial S}{\partial b} = 0 \end{cases} \longrightarrow \begin{cases} \frac{\partial S}{\partial a} = 2\sum (Y_i - a - \frac{b}{x_i}) * (-1) = 0 \\ \frac{\partial S}{\partial b} = 2\sum (Y_i - a - \frac{b}{x_i}) * (-\frac{1}{x_i}) = 0 \end{cases}$$

$$\begin{cases} \sum (Y_i - a - \frac{b}{x_i}) = 0 \\ \sum (Y_i - a - \frac{b}{x_i}) * (-\frac{1}{x_i}) = 0 \end{cases} \longrightarrow \begin{cases} na + b\sum \frac{1}{x_i} = \sum Y_i \\ a\sum \frac{1}{x_i} + b\sum \frac{1}{x_i^2} = \sum \frac{1}{x_i} Y_i \end{cases}$$

звідси знаходимо a і b.

2.3 Множинний кореляційний аналіз.

Процес побудови багатофакторної регресійної моделі містить у собі наступні етапи:

1. Підбір відповідного типу математичної залежності, що найкраще відображає характер досліджуваного зв'язку (лінійна, нелінійна).

При моделюванні зв'язку між факторними й результативними показниками у випадку множинної регресії при підборі рівняння, що найкраще описує досліджувані залежності, також виконують побудову діаграм розсіювання.

Розрізняють:

– Лінійну залежність. *Якщо зв'язок всіх факторних показників з результативним носить прямолінійний характер, то для записів цих залежностей можна використати лінійну функцію виду:*

$$Y_{расчетн} = b_0 + b_1 x_1 + b_2 x_2 + ... + b_n x_n$$

– Статечну залежність. *Якщо зв'язок між результативним і факторним показниками носить криволінійний характер, то може бути використана статечна функція виду:*

$$Y_{расчетн} = b_0 x_1^{b_1} x_2^{b_2} ... x_n^{b_b}$$

Логарифмуючи дане вираження, приходимо до лінійної функції:

$$Lg(Y_{расчетн}) = Lg(b_0) + b_1 Lg(x_1) + b_2 Lg(x_2) + ... + b_n Lg(x_n)$$

Для зручності обчислень і в першому й у другому випадку вводять фіктивну змінну $x_0=1$. З обліком цього рівняння лінійної множинної регресії представляється у вигляді:

$$Y_{расчетн} = b_0 x_0 + b_1 x_1 + b_2 x_2 + ... + b_n x_n \text{, або}$$
$$Lg(Y_{расчетн}) = Lg(b_0) x_1 + b_1 Lg(x_1) + b_2 Lg(x_2) + ... + b_n Lg(x_n)$$

У випадках, коли важко обґрунтувати форму залежності, рішення завдання можна провести по різних моделях і порівняти отримані результати. Адекватність моделей визначається за критерієм Фишера.

2. Розрахунок параметрів рівняння регресії методом найменших квадратів (МНК).

Усі компоненти рівняння регресії представляються у вигляді відповідних матриць (для лінійної регресії):

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_m \end{bmatrix},\ X = \begin{bmatrix} x_{10} & x_{11} & x_{12} & \dots & x_{1n} \\ x_{20} & x_{21} & x_{22} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ x_{m0} & x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1n} \\ 1 & x_{21} & x_{22} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix},\ B = \begin{bmatrix} b_0 \\ b_1 \\ \dots \\ b_n \end{bmatrix}$$

Параметри рівняння регресії розраховуються за формулою:

$$B = \left[[X]^T [X] \right]^{-1} [X]^T Y$$

Параметри рівняння регресії виду $Y_{расчетн} = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_n x_n$ можна знайти використовуючи функцію Microsoft Excel *ЛИНЕЙН*. Для цього необхідно виконати кроки:

- Відзначити блок осередків, де повинні перебувати звітні дані. Ширина блоку - число оцінюваних параметрів (n+1), висота - 5 рядків.

– Викликати функцію *ЛИНЕЙН*:

1 аргумент - стовпець Y;

2 аргумент - матриця [X];

3, 4 аргументи - істина.

- нажати <F2>, <CTRL>+<SHIFT>+<ENTER>

З'явиться таблиця:

b_2	b_1	b_0
...	...	...
...	...	...