Кореляційний та регресійний аналіз

1. Основні поняття кореляційного та регресійного аналізу.
2. Парна кореляція та парна лінійна регресія.
3. Множинна лінійна регресія та кореляція.
4. Нелінійна регресія.
5. Оцінка значимості параметрів взаємозв'язку.
6. Непараметричні методи оцінки взаємозв'язку.

1. Основні поняття кореляційного та регресійного аналізу.

Всі явища та процеси, що існують в природі та суспільстві, взаємопов'язані, тому вивчення взаємозв'язків та причинних залежностей є одним з найважливіших завдань статистики. Причинна залежність є головною формою закономірних зв'язків, проте причина сама по собі ще не визначає повною мірою наслідок; останній залежить також від умов, в яких діє причина. Умови і причини являють собою фактори. Ознака, що характеризує наслідок, називається, *результативною*, а та, що характеризує фактор, - *факторною.*

За характером залежності явищ розрізняють функціональні (повні) та кореляційні (неповні) зв'язки.

Функціональний зв'язок передбачає, що певному значенню факторної ознаки завжди відповідає одне або кілька значень результативної ознаки. Функціональні зв'язки характеризуються повною відповідністю між причиною і наслідком. Завдяки цьому функціональну залежність описують точні математичні формули. Функціональні залежності вивчають точні науки, такі як математика, фізика, хімія тощо. У суспільних процесах – це переважно зв'язок між елементами розрахункових формул, напр., залежність фондовіддачі від обсягу виробництва продукції та вартості основних виробничих фондів або прямо пропорційна залежність між продуктивністю праці та виробництвом продукції.

Кореляційний зв'язок проявляється в середньому, для масових спостережень, коли кожному значенню ознаки x відповідає певна множина ознаки y, які варіюють і утворюють ряд розподілу. Прикладом такого зв'язку можна навести залежність між рівнем кваліфікації та продуктивністю праці або залежність між кольором очей та кольором волосся.

Або зв'язок між урожайністю та кількістю внесених добрив. Але для кожного конкретного поля одна й та ж кількість внесених добрив спричинятиме різний приріст врожайності, тому що й інші фактори (погода, стан ґрунту, якість насіння та ін.) впливають на кінцевий результат. Але в середньому такий зв'язок спостерігається – збільшення маси внесених добрив веде до зростання врожайності.

За напрямком зв’язок буває прямим, коли залежна змінна збільшується із збільшенням факторної ознаки; та зворотнім, при якому зростання факторної ознаки спричиняє зменшення результату.

За своєю аналітичною формою зв’язки бувають лінійними та нелінійними. В першому випадку між ознаками проявляються в середньому лінійні співвідношення. Нелінійний взаємозв’язок виражається нелінійною функцією, а змінні пов’язані між собою в основному нелінійно.

За кількістю факторів, що розглядаються розрізняють парний (якщо характеризується зв’язок двох ознак) та множинний, або багатофакторний зв’язок (якщо вичаються більш ніж дві змінні).

За силою розрізняють слабкий та сильний зв’язок. Ця характеристика виражається конкретними величинами.

В найбільш загальному вигляді завдання статистики в галузі вивчення взаємозалежностей складається в кількісній оцінці їх наявності та напрямку, а також в характеристиці сили та форми впливу одних факторів на інші. Для його вирішення застосовують методи кореляційного та регресійного аналізу.

Завдання кореляційного аналізу полягають в вимірі щільності зв’язку між ознаками, визначенні невідомих причинних зв’язків і в оцінці факторів, що мають найбільший вплив на результативну ознаку.

Регресійний аналіз має на меті встановлення форми залежності, визначення функції регресії, використання рівняння для оцінки невідомих значень залежної змінної.

2. Парна кореляція та парна лінійна регресія.

Якщо характеризується зв’язок двох ознак, то його прийнято називати *парним*.

Для кількісної оцінки щільності зв’язку широко застосовують *лінійний коефіцієнт кореляції*. Якщо задані значення змінних *X* та *Y*, то він розраховується за формулою:

$$r_{xy} = \frac{\overline{xy} - \bar{x}\bar{y}}{\sigma_x \sigma_y}.$$

Коефіцієнт кореляції набуває значення від –1 до +1.

Якщо $|r|<0{,}30$, то зв’язок між ознаками слабкий;

$0{,}30 \le |r| \le 0{,}70$ – помірний зв’язок;

$|r|>0{,}70$ – сильний або щільний зв’язок.

Коли $|r|=1$ – зв’язок функціональний.

Якщо $|r| \approx 0$, то лінійний зв’язок між *X* та *Y* відсутній. Але можливе нелінійна взаємодія, а це потребує додаткової перевірки.

Для характеристики впливу змін *X* на варіацію *Y* використовують методи

регресійного аналізу. В випадку парної лінійної залежності рівняння регресії має наступний вигляд:

$$Y_{теор} = a_0 + a_1 X, \quad i = \overline{1, n},$$

де n – число спостережень;

a_0, a_1 – невідомі параметри рівняння;

$Y_{i\ теор}$ – розраховане вирівняне значення результативної ознаки після підстановки в рівняння X_i.

Параметри a_0 та a_1 оцінюються за допомогою метода найменших квадратів. Його суть складається в тому, що найкращі оцінки a_0 та a_1 отримують, коли:

$$\sum_{i=1}^{n} (Y_i - Y_{теор})^2 \to \min.$$

Для обчислення параметрів рівняння регресії a_0 та a_1 складають і розв'язують систему рівнянь:

$$\begin{cases} n a_0 + a_1 \sum X = \sum Y \\ a_0 \sum X + a_1 \sum X^2 = \sum YX \end{cases}$$

Можна використати й інші формули, що отримують з методу найменших квадратів, наприклад:

$$a_1 = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2}, \quad a_0 = \bar{Y} - a_1 \bar{X}$$

Параметр a_1 – це коефіцієнт регресії, що характеризує вплив зміни ознаки X на результативну ознаку Y. Він показує, на скільки одиниць в середньому зміниться Y при зміні X на одну одиницю. Якщо $a_0>0$, то спостерігається позитивний зв'язок. Якщо $a_1<0$, то збільшення X на одиницю спричинить зменшення Y в середньому на величину a_1.

Параметр a_0 – це постійна величина в рівнянні регресії. Іноді вона характеризує початкове значення Y.

Значення функції $Y_{теор} = a_0 + a_1 X$ називається розрахунковим значенням і на графіку зображує теоретичну лінію регресії.

3. Множинна лінійна регресія та множинна кореляція.

Парна кореляція та регресія можуть розглядатися як окремий випадок множинної (багатофакторної) регресії (кореляції), коли характеризується зв'язок множини незалежних змінних з результативною ознакою.

На практиці теоретичні положення про наявність взаємозв'язків підтверджуються парними коефіцієнтами кореляції між залежною та

незалежними змінними. Для розрахунку множинного рівняння регресії обирають найбільш значимі з всієї множини незалежних змінних відповідно до коефіцієнтів кореляції.

Множинна регресія вивчає зв'язок між трьома або більше ознаками та має наступний вигляд:

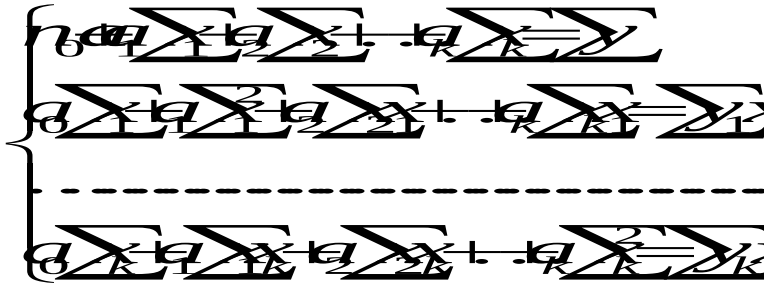
де $Y_{теор}$ – розрахункове значення регресії, яке є оцінкою очікуваного значення Y при фіксованих значеннях ознак $X_1, ..., X_k$;

$X_1, ..., X_k$ – найбільш значимі незалежні змінні;

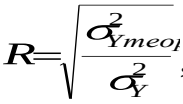
a_0 – параметр, що показує усереднений вплив на результативний показник факторів, що не включені до моделі (або не виділені для дослідження);

$a_1, ..., a_k$ – коефіцієнти регресії, кожний з яких показує на скільки одиниць зміниться Y зі зміною відповідної ознаки X на одиницю за умови, що останні ознаки не зміняться.

Параметри рівняння множинної регресії, як правило, знаходять методом найменших квадратів:



При оцінці лінійного множинного зв'язку розраховують *коефіцієнт множинної кореляції*. Він відображає щільність зв'язку між залежною змінною та варіаціями всіх незалежних змінних, що включені до аналізу:



де  - факторна дисперсія;

 - загальна дисперсія.

Коли оцінюється тіснота зв'язку між результативною Y та двома факторними ознаками X_1, X_2, то множинний коефіцієнт кореляції можна визначити за формулою:

де r – парні коефіцієнти кореляції між ознаками.

Множинний коефіцієнт кореляції змінюється в межах від 0 до 1 і є позитивною величиною: 0<R<1:

R≤0,3 – зв'язок практично відсутній (або не всі важливі фактори взаємозв'язку враховані, або вибрано невірну форму рівняння регресії. Необхідно переглянути змінні, що ввійшли в модель, та можливо її вид);

0,3<R≤0,5 – слабкий зв'язок;

0,5<R≤0,7 – помірний зв'язок;

R>0,7 – сильний зв'язок.

4. Нелінійна регресія.

Надання зв'язку через лінійну функцію в тому випадку, коли існує нелінійне співвідношення, викличе помилки та спрощені або навіть неправильні висновки на основі аналітичного рівняння.

Питання про не лінійність форми рівняння необхідно вирішувати на стадії теоретичного аналізу. Аналіз повинен спиратися на суті взаємодіючих явищ та процесів і підкріплятися різними статистичними критеріями.

Існують різні форми нелінійних рівнянь регресії, але в загальному вигляді можна виділити два їх класи.

1. Регресії нелінійні відносно включених в дослідження змінних, але лінійне за параметрами.

Це, наприклад, поліноми. В випадку парної регресії рівняння має наступний вигляд:

Множинна регресія $Y=f(X_1,X_2)$ маємо наступне рівняння:

Можливе застосування гіперболи, інших функцій. За допомогою стандартних програм для ЕВМ може бути створено будь-яке нелінійне поєднання змінних, що є лінійним відносно коефіцієнтів рівняння. Остання оцінюються за допомогою метода найменших квадратів.

2. Регресії з нелінійними параметрами.

Найбільш роз поширеною є ступенева функція:

$Y=a_0X^{a_1}$ - парна регресія;

 - множинна регресія.

Використання цих функцій обмежується складністю оцінювання параметрів

рівняння. Це потребує спеціальних прийомів, програм для ЕВМ.

5. Оцінка значимості параметрів взаємозв'язку.

Коли отримані оцінки кореляції та регресії, необхідно перевірити їх на відповідність істинним параметрам взаємозв'язку.

Для оцінки значимості коефіцієнта парної кореляції розраховують стандартну помилку коефіцієнта кореляції:

$$\sigma_{r_{XY}} = \sqrt{\frac{1 - r_{XY}^2}{n-2}}.$$

Необхідно, щоб $\sigma_{r_{XY}} < r_{XY}$. Значимість r_{XY} перевіряється його співставленням з $\sigma_{r_{XY}}$, при цьому отримують:

$$t_{розр} = r_{XY} \sqrt{\frac{n-2}{1 - r_{XY}^2}},$$

де $t_{розр}$ – так зване розрахункове значення *t*-критерію.

Якщо $t_{розр}$ більше теоретичного (табличного) значення критерію Стьюдента ($t_{табл}$) для заданого рівня ймовірності та (n-2) ступенів свободи, то можна стверджувати, що r_{XY} значимий.

Аналогічним чином на основі відповідних формул розраховують стандартні помилки параметрів рівняння регресії, а після й *t*- критерій для кожного параметру. Необхідно виконання умови $t_{розр} > t_{табл}$. В іншому випадку довіряти отриманій оцінці параметру немає підстав.

Висновок про правильність вибору виду взаємозв'язку та характеристику значимості всього рівняння регресії отримують за допомогою *F*-критерію (критерій Фішера- Снедекора), що розраховується за наступною формулою:

$$F_{розр} = \frac{R^2 (n-m)}{(1 - R^2)(m-1)}$$

де n – число спостережень;

m – число параметрів рівняння регресії.

$F_{розр}$ має бути більше $F_{табл}$ при $v_2=m-1$ та $v_1=n-m$ ступенях свободи. В іншому випадку необхідно переглянути форму рівняння, перелік змінних тощо.

6. Непараметричні методи оцінки взаємозв'язку.

Важливою задачею статистики є розробка методики статистичної оцінки соціальних явищ, яка ускладнюється тим, що багато соціальних явищ не мають кількісної оцінки.

Для визначення щільності зв'язку двох якісних ознак, кожний з яких складається лише з двох груп, застосовують *коефіцієнти асоціації та контингенції.*

Для їх обчислення будують таблицю, яка показує зв'язок між двома явищами, кожне х яких повинне бути альтернативним, тобто складатися з двох якісно відмінних один від іншого значень ознаки (наприклад, хороший, поганий).

Таблиця для обчислення коефіцієнтів асоціації та контингенції

a	b	a+b
c	d	c+d
a+c	b+d	a+b+c+d

Коефіцієнт асоціації: $K_a = \frac{ad+bc}{ad+bc}$

Коефіцієнт контингенції: $K_k = \sqrt{\frac{abdc}{(ab)(cd)(ac)(bd)}}$

Коефіцієнт контингенції завжди менший коефіцієнта асоціації. Зв'язок вважається підтвердженим, якщо $К_а > 0,5$, або $К_к > 0,3$.

Приклад. Досліджувалась характеристика випадкових споживачів наркотиків в залежності від їх сімейного стану в одному з регіонів Росії (тис. чол.).

Групи споживачів наркотиків	Сімейний стан		Всього
	жонатий (заміжня)	нежонатий (незаміжня)	
Споживав	10,0	14,5	24,5
Не споживав	2,5	4,5	7,0
Разом	12,5	19,0	31,5

Таким чином, споживання наркотиків не залежить від сімейного стану випадкових споживачів.

Коли кожна з якісних ознак складається більше ніж з двох груп, то для визначення щільності зв'язку можливе застосування коефіцієнтів взаємного сполучення Пірсона та Чупрова:

$$K_П = \sqrt{\frac{\varphi^2}{1+\varphi^2}}$$

$$K_Ч = \sqrt{\frac{\varphi^2}{\sqrt{(K_1-1)\cdot(K_2-1)}}}$$

де $\varphi^2 = \sum \frac{n_{xy}^2}{n_x \cdot n_y} - 1$ - показник взаємного сполучення, визначається як сума відношень квадратів частот кожної клітки таблиці n_{xy} до добутку підсумкових частот відповідного стовпчика n_y та строки n_x мінус одиниця;

$$1+\varphi^2 = \sum \frac{\sum \frac{n_{xy}^2}{n_x}}{n_y} = \sum \frac{\sum \frac{n_{xy}^2}{n_y}}{n_x};$$

K₁ – число значень (груп) першої ознаки;

K₂ – число значень (груп) другої ознаки.

Чим більш наближені величини показників $K_П$ та $K_Ч$ до 1, тим тісніше зв'язок між показниками.

Приклад. Досліджувалась залежність між оцінкою рівня життя респондентів

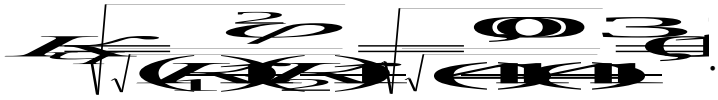
Москви та типом підприємства, на якому вони працюють

Тип підприємства	Оцінка рівня життя				Всього
	цілком задовільний	скоріше задовільний	скоріше не задовільний	зовсім не задовільний	
державне	31	35	35	35	136
акціонерне	17	13	14	9	53
орендне	4	2	1	1	8
приватне	8	5	4	3	20
Разом	60	55	54	48	217

$$K_П = \sqrt{\frac{\varphi^2}{1+\varphi^2}}$$

$$1+\varphi^2 = \sum \frac{\sum \frac{n_{xy}^2}{n_y}}{n_x} = \frac{\frac{31^2}{60}+\frac{35^2}{55}+\frac{35^2}{54}+\frac{35^2}{48}}{136} + \frac{\frac{17^2}{60}+\frac{13^2}{55}+\frac{14^2}{54}+\frac{9^2}{48}}{53} + \frac{\frac{4^2}{60}+\frac{2^2}{55}+\frac{1^2}{54}+\frac{1^2}{48}}{8} + \frac{\frac{8^2}{60}+\frac{5^2}{55}+\frac{4^2}{54}+\frac{3^2}{48}}{20} = \text{[illegible]}$$

$\varphi^2 = \text{[illegible]};$



Таким чином, оцінка рівня життя респондентів не залежить від типа підприємства, на якому вони працюють.