ІНДИВІДУАЛЬНЕ ЗАВДАННЯ №3 З КУРСУ «МАТЕМАТИЧНА СТАТИСТИКА»

РЕГРЕСІЯ

Нехай вивчається генеральна сукупність, що характеризується системою кількісних ознак (X,Y). Для аналізу залежності між випадковими величинами X і Y зроблена вибірка, причому складова X набула значень $x_1,x_2,...,x_k$, складова Y – $y_1,y_2,...,y_l$, а подія $\{X = x_i,\ Y = y_j\}$ мала частоту появи n_{ij} $(i = 1,..,k\ ;\ j = 1,..,l)$. Результати цих спостережень записують у вигляді кореляційної таблиці:

$Y \mid X$	x_1	x_2	...	x_i	...	x_k	m_j
y_1	n_{11}	n_{21}	...	n_{i1}	...	n_{k1}	m_1
y_2	n_{12}	n_{22}	...	n_{i2}	...	n_{k2}	m_2
...	...	...	...	...	...	...	...
y_j	n_{1j}	n_{2j}	...	n_{ij}	...	n_{kj}	m_j
...	...	...	...	...	...	...	...
y_l	n_{1l}	n_{2l}	...	n_{il}	...	n_{kl}	m_l
n_i	n_1	n_2	...	n_i	...	n_k	n

За даними кореляційної таблиці обчислюють умовні середні $\overline{y_{xi}}$ $(i = 1,...,k)$:

$$\overline{y_{x_1}} = \frac{y_1 n_{11} + y_2 n_{12} + ... + y_l n_{1l}}{n_1}, \quad \overline{y_{x_2}} = \frac{y_1 n_{21} + y_2 n_{22} + ... + y_l n_{2l}}{n_2}, \ ...,$$

$$\overline{y_{x_i}} = \frac{y_1 n_{i1} + y_2 n_{i2} + ... + y_l n_{il}}{n_i}, \ ..., \overline{y_{x_k}} = \frac{y_1 n_{k1} + y_2 n_{k2} + ... + y_l n_{kl}}{n_k}.$$

Складають таблицю умовних середніх $\overline{y_x}$:

x	x_1	x_2	...	x_i	...	x_k
$\overline{y_x}$	$\overline{y_{x_1}}$	$\overline{y_{x_2}}$	...	$\overline{y_{x_i}}$	...	$\overline{y_{x_k}}$

Для визначення вигляду функції регресії будують точки$(x_i, \overline{y_{x_i}})$ та з'єднують їх ламаною, яка називається емпіричною лінією регресії. Якщо емпірична лінія регресії значно наближається до прямої лінії, то висувається гіпотеза про наявність лінійного зв'язку між досліджуваними ознаками.

1. Лінійна регресія

Якщо висунуто гіпотезу про наявність лінійної залежності ознаки Y від X, то рівняння регресії має вид:

$$y = ax + b \qquad (1)$$

де a, b − параметри моделі.

Побудова лінійної регресійної моделі – це знаходження параметрів рівняння (1). Параметри рівняння регресії можна знайти за методом найменших квадратів.

Якщо в рівняння (1) підставити замість x значення $x_1,x_2,\ldots,x_k$, то будуть отримані теоретичні значення $y_i^* = ax_i + b$, які відрізняються від обчислених за кореляційною таблицею умовних середніх $\overline{y_{x_i}}(i = 1,\ldots,k)$. Різниця значень y_i^* і $\overline{y_{x_i}}$ називається помилкою регресійної моделі і позначається e_i. Якщо параметри рівняння підбираються так, щоб сума квадратів помилок була мінімальною, то говорять, що вони отримані за методом найменших квадратів.

У випадку лінійної регресії параметри рівняння регресії за методом найменших квадратів знаходяться з системи лінійних алгебраїчних рівнянь:

$$\begin{cases} a\sum_{i=1}^{k} x_i^2 n_i + b\sum_{i=1}^{k} x_i n_i = \sum_{i=1}^{k} x_i n_i \overline{y_{x_i}} \\ a\sum_{i=1}^{k} x_i n_i + b\sum_{i=1}^{k} n_i = \sum_{i=1}^{k} n_i \overline{y_{x_i}} \end{cases} . \qquad (2)$$

Перевірка правильності побудови рівняння регресії здійснюється за основним варіаційним рівнянням:

$Q = Q_p + Q_O$

Де $Q = \sum_{i=0}^{k}(\overline{y_{x_i}} - \bar{y})^2 n_i$ – загальна варіація, тобто сума квадратів відхилень емпіричних значень $\overline{y_{x_i}}$ від середнього $\bar{y} = \frac{\sum_{i=1}^{k} \overline{y_{x_i}} n_i}{n}$

$Q_p = \sum_{i=0}^{k}(y_i^* - \bar{y})^2 n_i$ - варіація регресії, тобто сума квадратів відхилень теоретичних значень Y від середнього, що обумовлена регресією;

$Q_o = \sum_{i=0}^{k}(\overline{y_{x_i}} - y_i^*)^2 n_i$ - варіація залишків, тобто сума квадратів відхилень теоретичних значень Y від емпіричних.

Адекватність моделі вибірковим даним можна оцінити за коефіцієнтом детермінації R^2 , що показує частину варіації значень результативної ознаки Y, що пояснюється рівнянням регресії. Коефіцієнт детермінації розраховується за формулою:

$$R^2 = 1 - \frac{Q_o}{Q} = \frac{Q_p}{Q} .$$

Значення коефіцієнта детермінації знаходяться в інтервалі [0;1]. Чим ближче R^2 до 1, тим краще отримане рівняння регресії пояснює поведінку результативної ознаки.

Для перевірки статистичної значущості рівняння регресії використовується статистика Фішера.

Ми перевіряємо гіпотезу:

- H_0 (нульова гіпотеза): модель регресії не є значущою (коефіцієнти a=0, b=0 тобто немає зв'язку між X і Y).
- H_1 (альтернативна гіпотеза): модель регресії є значущою.

Розраховується *F*-статистика за формулою:

$$F_{емп} = \frac{Q_p(n-m)}{Q_o(m-1)}$$

де n – кількість спостережень, m – кількість параметрів функції регресії (у випадку лінійної моделі $m = 2$). Розраховане значення *F*-статистики порівнюється з критичним значенням $F_{кр}$ розподілу Фішера для степенів свободи $m-1$, $n-m$ та рівня значущості α. Якщо $F_{емп} > F_{кр}$, то нульова гіпотеза відхиляється: модель адекватна.

2. Нелінійна регресія

Якщо графік регресії $y = f(x)$ зображається кривою лінією, то кореляцію називають *нелінійною* (криволінійною).Наприклад, функції регресії *Y* на *X* можуть мати вигляд:

$y = ax^2 + bx + c$ (параболічна кореляція другого порядку)

$y = ax^3 + bx^2 + cx + d$ (поліноміальна кореляція третього порядку)

$y = \frac{a}{x} + b$ (гіперболічна кореляція);

$y = a + b\ lnx$ (логарифмічна кореляція)

$y = bx^a$ (степенева кореляція)

$y = a\sqrt{x} + b$ (коренева кореляція)

$y = ba^x$ (показникова кореляція).

Теорія криволінійної кореляції розв’язує ті самі задачі, що і теорія лінійної кореляції, а саме:

1) за даними кореляційної таблиці встановлюють форму кореляційного зв’язку, тобто визначають вигляд функції $y = f(x)$
2) оцінюють щільність кореляційного зв’язку, тобто дають оцінку ступеню розсіювання значень випадкової величини *Y* навколо побудованої кривої регресії $y = f(x)$.

2.1. **Параболічна кореляція.** У прямокутній системі координат позначимо всі точки, які відповідають парам чисел $(x_i, \overline{y_{x_i}})$, тобто побудуємо *поле кореляції*.

Припустимо, що точки $M_i(x_i, \overline{y_{x_i}})i = 1,\ldots,k$, розташовані приблизно на параболі другого порядку. Рівняння параболи – параболічної регресії *Y* на *X* будемо шукати у вигляді

$$f(x) = ax^2 + bx + c, \qquad (3)$$

де a,b,c – невідомі параметри.

Із всіх парабол такого виду шукана найближче розташована (згідно з методом найменших квадратів) до точок M_1 , M_2 ,...,M_k , причому точка M_i вибирається n_i разів, $i = 1,...,k$ (скільки разів зустрічаються у розподілі значення x_i).

Невідомі коефіцієнти *a,b,c* визначимо таким чином, щоб сума відповідних відхилень була мінімальною. Застосуємо метод найменших квадратів. Для цього складемо функцію:

$$F(a,b,c) = \sum_{i=1}^{k} n_i (f(x_i) - \overline{y}_{x_i})^2 = \sum_{i=1}^{k} (ax_i^2 + bx_i + c - \overline{y}_{x_i})^2 n_i$$

.

Це функція трьох незалежних змінних *a,b,c*. Необхідна умова екстремуму функції (рівність нулю частинних похідних за змінними *a,b* і *c*) дає три рівняння. Наведемо кінцевий вигляд системи рівнянь відносно параметрів *a,b,c*:

$$\begin{cases} (\sum_{i=1}^{k} n_i x_i^4)a + (\sum_{i=1}^{k} n_i x_i^3)b + (\sum_{i=1}^{k} n_i x_i^2)c = \sum_{i=1}^{k} n_i \overline{y}_{x_i} x_i^2; \\ (\sum_{i=1}^{k} n_i x_i^3)a + (\sum_{i=1}^{k} n_i x_i^2)b + (\sum_{i=1}^{k} n_i x_i)c = \sum_{i=1}^{k} n_i \overline{y}_{x_i} x_i; \\ (\sum_{i=1}^{k} n_i x_i^2)a + (\sum_{i=1}^{k} n_i x_i)b + nc = \sum_{i=1}^{k} n_i \overline{y}_{x_i}. \end{cases} \quad (4)$$

Розв’язуючи її методом Гаусса, знайдемо параметри *a,b,c*, які підставимо в (3).

2.2. Гіперболічна кореляція. Припустимо, що аналіз залежності між змінними *X* і *Y* , вираженої кореляційною таблицею, приводить до вибору форми кореляційної залежності *Y* на *X* у вигляді рівняння гіперболи

$$y = \frac{a}{x} + b \quad (5)$$

Регресії такого типу називаються *гіперболічними*.

За методом найменших квадратів невідомі параметри *a* і *b* шукаємо з системи рівнянь:

$$\begin{cases} a \sum_{i=1}^{k} \frac{1}{x_i} n_i + bn = \sum_{i=1}^{k} \overline{y}_{x_i} n_i; \\ a \sum_{i=1}^{k} \frac{1}{x_i^2} n_i + b \sum_{i=1}^{k} \frac{1}{x_i} n_i = \sum_{i=1}^{k} \frac{1}{x_i} \overline{y}_{x_i} n_i. \end{cases} \quad (6)$$

2.3. Показникова кореляція.

Розглянемо випадок, коли аналіз зв’язку між змінними *X* та *Y*, заданими кореляційною таблицею, приводить до вибору форми кореляційної залежності *Y* на *X* у вигляді показникової функції

$$\overline{y} = ba^x, \quad (7)$$

Логарифмуючи обидві частини рівності (7), одержимо $\lg y = x \lg a + \lg b$. Отже, якщо між X та Y існує кореляційна залежність Y на X з параметрами a і b, то між $\lg Y$ і X – лінійна кореляційна залежність з параметрами $\lg a$ і $\lg b$. Тому система рівнянь для визначення $\lg a$ і $\lg b$ буде мати вигляд

$$\begin{cases} \lg a \sum\limits_{i=1}^{k} n_i x_i + n \lg b = \sum\limits_{i=1}^{k} n_i \lg \overline{y}_{x_i}; \\ \lg a \sum\limits_{i=1}^{k} n_i x_i^2 + \lg b \sum\limits_{i=1}^{k} n_i x_i = \sum\limits_{i=1}^{k} n_i x_i \lg \overline{y}_{x_i}. \end{cases} \quad (8)$$

Розв'язуючи її, знаходимо $\lg a$ і $\lg b$, а потім параметри a і b показникової функції (7).

2.4. **Коренева кореляція.** Припустимо, що аналіз залежності між змінними X і Y, вираженої кореляційною таблицею, приводить до вибору форми кореляційної залежності Y на X у вигляді рівняння

$$y = a\sqrt{x} + b \quad (9)$$

У цьому випадку невідомі параметри a і b будемо шукати з системи рівнянь

$$\begin{cases} a \sum\limits_{i=1}^{k} n_i \sqrt{x_i} + bn = \sum\limits_{i=1}^{k} \overline{y}_{x_i} n_i; \\ a \sum\limits_{i=1}^{k} n_i x_i + b \sum\limits_{i=1}^{k} n_i \sqrt{x_i} = \sum\limits_{i=1}^{k} n_i \overline{y}_{x_i} \sqrt{x_i}. \end{cases} \quad (10)$$

3.Вибірковий лінійний коефіцієнт кореляції

Вибірковий коефіцієнт кореляції r показує силу та напрямок лінійної залежності між змінними X і Y у вибірці. Він обчислюється за формулою

$$r_{12} = \frac{c_{12}}{s_1 s_2} = \frac{\sum\limits_{i=1}^{k}\sum\limits_{j=1}^{l} n_{ij}(x_i - \overline{x})(y_j - \overline{y})}{\sqrt{\sum\limits_{i=1}^{k} n_i (x_i - \overline{x})^2}\sqrt{\sum\limits_{j=1}^{l} m_j (y_j - \overline{y})^2}} \quad (11)$$

Щоб перевірити значущість коефіцієнта кореляції, треба перевірити гіпотезу:

- H_0 (нульова гіпотеза): $\rho=0$ (у генеральній сукупності немає лінійної залежності).
- H_1 (альтернативна гіпотеза): $\rho\neq0$ (лінійна залежність є)

Для перевірки треба обчислити t-статистику :

$$t_{\text{емп}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

і порівняти її із критичним значенням з таблиці Стьюдента для обраного рівня значущості та ступенів свободи d.f.=n−2. Якщо $|t_{\text{емп}}| > t_{\text{кр}}$, то нульова гіпотеза відхиляється (коефіцієнт кореляції значущий)

ЗАВДАННЯ ДО ЛАБОРАТОРНОЇ РОБОТИ

1. За даними кореляційної таблиці обчислити умовні середні $\bar{y}_{xi}$ ($i = 1,\ldots,k$).
2. Побудувати поле кореляції, тобто нанести точки $M_i(x_i;\bar{y}_{xi})$, $i = 1,\ldots,k$, на координатну площину, та емпіричну лінію регресії.
3. Побудувати лінійне рівняння регресії та намалювати графік.
4. Обчислити коефіцієнт детермінації та перевірити адекватність побудованої лінійної моделі.
5. Обчислити вибірковий лінійний коефіцієнт кореляції.
6. За рівня значущості α перевірити значущість коефіцієнта кореляції.
7. Зробити припущення про вигляд функції нелінійної регресії (парабола, гіпербола і т.д.). В залежності від вигляду функції регресії скласти відповідну систему рівнянь. Розв’язати її і знайти невідомі параметри вибраної функції нелінійної регресії.
8. Записати рівняння кривої регресії Y на X : $y = f(x)$ та побудувати її графік
9. Перевірити адекватність побудованої нелінійної моделі за F-критерієм
10. За моделлю з найменшою залишковою варіацією Q_o обчислити прогнозоване значення y* при заданому значенні x*.

Структура звіту:
1) Постановка задачі;
2) Короткі теоретичні відомості;
3) Програмна реалізація (без тексту програми);
4) Отримані результати (графічні та числові) та їх аналіз;
5) Висновки

Максимальна кількість балів – 10.

1.

$Y \mid X$	3	6	7	10	13	15	17
1	22						
1,5	2	31					
2		1	25	4			
2,5			2	18	3		
3,5				1	30	8	
4						12	2

2.

$Y \mid X$	2	3	5	7	9	12	13
3						13	4
5				1	21	2	
6				24	3		
7		7	13	2			
10	3	18	4				
12	23						

3.

$Y \mid X$	0	1	2	3	4	5	6
2	30	3	5				
3	2	20					
5		5	10	2			
10			7	12	10		
17					20	15	
30						5	5

4.

$Y \mid X$	0	0,5	1	1,5	2	2,5	3
5	3	18	2	3			
25		2	1	10	5		
40					7		
55						10	
70						1	10
100							35

5.

$Y \mid X$	3	4	7	10	11	14	17
1	18						
2	2	18	3				
2,5		4	25	2			
3				30	2	5	
4					16	4	4
4,5						22	3

6.

$Y \mid X$	2	3	5	8	10	11	13
3						19	2
4				3	31	2	
6			1	16	3		
8		2	21	4			
10	3	31	5				
12	30	2					

7.

$Y \mid X$	0	4	6	7	8	9	10
5	25		2				
20	10	60					
40		2	22	2			
62				1	2		
78						28	
95							21

8.

$Y \mid X$	0	1	2	3	4	5	6
1	29	5	10	15			
10		1	2	50	8		
20			1	1	10	9	
30					1	20	
40						5	
64						4	20

9.

$Y \mid X$	3	5	6	9	12	14	19
1,5	21						
2,5	4	31	3				
3		5	28	3	4		
3,5				25	4	3	
4					17	3	5
4,5						29	2

10.

$Y \mid X$	2	3	5	7	9	12	13
3						21	1
4			2	3	20		
5		2	31	12	4		
6		15	3				
10	3	7					
12	25						

11.

$Y \mid X$	0	1	2	3	4	5	6
7	50	1					
11	2	15	3				
20		20	17	4			
35			15	13	7		
50				7	42	20	
75					1	16	2

12.

$Y \mid X$	0	0,5	1	1,5	2	2,5	3
1	2	15	25		10	2	
5		3	30	45	10	5	
10			2	1	20	15	
15			1	1	3	25	
25				1	5	18	3
27					1		18

13.

$Y\|X$	4	5	7	9	12	15	17
1	12						
1,5	3	19					
2,5		3	31	1			
3			2	18	7		
3,5				1	20	4	
4						17	2

14.

$Y\|X$	2	3	5	6	8	10	12
2						22	2
3				4	13		
5		2	3	14	5		
7		4	21				
12	3	14					
13	12						

15.

$Y\|X$	4	5	7	9	10	11
4					15	1
15			7	11	15	
20	18	3	2			
25	2	20		1		
30	3	5	9	1	1	
35	11	10	4	1	3	

16.

$Y\|X$	0	1	2	3	4	5	6
1	10	20	30	50	18	40	22
6		2	5	45	15	10	30
12			2	1	2	18	40
20					3	15	30
28						1	10
30							1

17.

$Y\|X$	3	5	7	9	13	15	17
1	23						
1,5	2	19					
2		3	32	2			
3			8	23	5		
3,5				2	17	4	
4						20	3

18.

$Y\|X$	1	2	4	6	9	11	12
3						7	31
4				2	21	4	
5			4	12	6		
7		3	22	5			
10	4	20					
12	23						

19.

$Y\|X$	0	1	2	3	4	5	6
1	45	4	5				
10	1	4	8	10			
20			7	20			
25				1	44		
30					3	28	
44						15	11

20.

$Y\|X$	0	0,5	1	1,5	2	2,5	3
1	50	15	30	20			
10	1	2	12	60	23		
20			1	2	20	20	
30				1	2	22	
40					1	25	
60						1	57

21.

$Y\|X$	4	6	8	11	13	15	17
1,5	13	2					
2	7	21	1				
3			20	7			
3,5				18	2		
4					25	3	4
4,5						16	1

22.

$Y\|X$	0	1	2	3	4	5	6
2	18	3	2				
3	2	20					
5	3	5	10	2			
10			7	12	5		
17					20	3	
26						45	5

23.

$Y\|X$	2	3	5	7	9	12	13
3						21	1
4			2	3	20		
5		2	31	12	4		
6		15	3				
10	3	7					
12	25						

24.

$Y\|X$	0	4	6	7	8	9	10
7	19	3	2				
13	2	14					
40		3	22	2			
80					5		
120				1	22	28	
200						2	21