

# ІНДИВІДУАЛЬНЕ ЗАВДАННЯ №3 З КУРСУ «МАТЕМАТИЧНА СТАТИСТИКА»

## НЕЛІНІЙНА 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  $x_1, x_2, \dots, x_k$ , складова  $Y$  –  $y_1, y_2, \dots, y_l$ , а подія  $\{X = x_i, Y = y_j\}$  мала частоту появи  $n_{ij}$  ( $i = 1, \dots, k$ ;  $j = 1, \dots, l$ ). Результати цих спостережень записують у вигляді кореляційної таблиці:

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

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

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

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

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

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

Аналогічно можна скласти таблицю умовних середніх  $\overline{x_{yj}}$ :

$y$	$y_1$	$y_2$	...	$y_j$	...	$y_l$
$\overline{x_y}$	$\overline{x_{y_1}}$	$\overline{x_{y_2}}$	...	$\overline{x_{y_j}}$	...	$\overline{x_{y_l}}$

Для визначення вигляду функції регресії будують точки  $(x; \overline{y_x})$  (або  $(y; \overline{x_y})$ ) і за їх розміщенням роблять висновок про приблизний вигляд функції регресії.

Якщо графік регресії  $\overline{y_x} = f(x)$  або  $\overline{x_y} = \phi(y)$  зображається кривою лінією, то кореляцію називають *нелінійною* (криволінійною).

Наприклад, функції регресії  $Y$  на  $X$  можуть мати вигляд:

$$\overline{y_x} = ax^2 + bx + c \text{ (параболічна кореляція другого порядку);}$$

$\bar{y}_x = ax^3 + bx^2 + cx + d$  (параболічна кореляція третього порядку);

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

$\bar{y}_x = ba^x$  (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  $\bar{y}_x = f(x)$  або  $\bar{x}_y = \phi(y)$ ;

2) оцінюють щільність кореляційного зв'язку, тобто дають оцінку ступеню розсіювання значень випадкової величини  $Y$  навколо побудованої кривої регресії  $\bar{y}_x$  (або значень випадкової величини  $X$  навколо  $\bar{x}_y$ ).

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

Припустимо, що точки  $M_i(\bar{x}_i; y_{xi}), i = 1, \dots, k$ , розташовані приблизно на параболі другого порядку. Рівняння параболі – параболічної регресії  $Y$  на  $X$  будемо шукати у вигляді

$$f(x) = ax^2 + bx + c, \quad (1)$$

де  $a, b, c$  – невідомі параметри.

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

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

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

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  $a, b, c$ :

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

Розв'язуючи її методом Гаусса, знайдемо параметри  $a, b, c$ , які підставимо в (1).

У випадку параболічної регресії  $X$  на  $Y$  необхідно знайти функцію  $\phi(y) = a_1y^2 + b_1y + c_1$ . У результаті одержуємо систему рівнянь відносно параметрів  $a_1, b_1, c_1$ , в якій порівняно з системою (2)  $x$  і  $y$  міняються місцями.

**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

$$\bar{y}_x = \frac{a}{x} + b, \quad (3)$$

а у випадку регресії  $X$  на  $Y$  – гіперболи

$$\bar{x}_y = \frac{c}{y} + d. \quad (4)$$

Регресії такого типу називаються *гіперболічними*.

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

$$\begin{cases} a \sum_{i=1}^k \frac{1}{x_i} n_i + bn = \sum_{i=1}^k \bar{y}_{x_i} n_i; \\ a \sum_{i=1}^k \frac{1}{x_i^2} n_i + b \sum_{i=1}^k \frac{1}{x_i} n_i = \sum_{i=1}^k \frac{1}{x_i} \bar{y}_{x_i} n_i. \end{cases} \quad (5)$$

У випадку гіперболічної регресії  $X$  на  $Y$  система рівнянь для визначення параметрів  $c, d$  рівняння (4) знаходиться аналогічно.

### 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

$$\bar{y}_x = ba^x, \quad (6)$$

а при розгляді регресії  $X$  на  $Y$  – показникової функції

$$\bar{x}_y = dc^y. \quad (7)$$

Логарифмуючи обидві частини рівності (6), 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

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

Розв'язуючи її, знаходимо  $\lg a$  і  $\lg b$ , а потім параметри  $a$  і  $b$  показникової функції (6). Аналогічно можна одержати систему рівнянь для визначення логарифмів параметрів  $c$  і  $d$  рівняння (7).

**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

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

а у випадку регресії  $X$  на  $Y$  – рівняння

$$\bar{x}_y = c\sqrt{y} + d. \quad (10)$$

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

$$\begin{cases} a \sum_{i=1}^k n_i \sqrt{x_i} + bn = \sum_{i=1}^k \bar{y}_{x_i} n_i; \\ a \sum_{i=1}^k n_i x_i + b \sum_{i=1}^k n_i \sqrt{x_i} = \sum_{i=1}^k n_i \bar{y}_{x_i} \sqrt{x_i}. \end{cases} \quad (11)$$

Для відшукування параметрів  $c$  і  $d$  рівняння (10) складаємо аналогічну до (11) систему рівнянь, де змінні  $x$  і  $y$  міняються місцями.

**5. Оцінка щільності кореляційного зв'язку.** За побудованою кривою регресії  $\bar{y}_x = f(x)$  (або  $\bar{x}_y = \phi(y)$ ) можна оцінити відхилення значень випадкової величини  $Y$  від кривої регресії  $\bar{y}_x$  (або значень випадкової величини  $X$  від кривої регресії  $\bar{x}_y$ ). Зокрема, обчислюють дисперсію величини  $Y$  відносно кривої регресії  $Y$  на  $X$ :

$$\sigma^2(y, \bar{y}_x) = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^l (y_j - f(x_i))^2 n_{ij} = \frac{\Delta}{n},$$

$$\begin{aligned} \Delta = & n_{11}[y_1 - f(x_1)]^2 + n_{21}[y_1 - f(x_2)]^2 + \dots + n_{k1}[y_1 - f(x_k)]^2 + \\ & + n_{12}[y_2 - f(x_1)]^2 + n_{22}[y_2 - f(x_2)]^2 + \dots + n_{k2}[y_2 - f(x_k)]^2 + \dots + \\ & + n_{1l}[y_l - f(x_1)]^2 + n_{2l}[y_l - f(x_2)]^2 + \dots + n_{kl}[y_l - f(x_k)]^2. \end{aligned} \quad (12)$$

За міру розсіювання значень випадкової величини  $Y$  від кривої регресії  $y_x$  можна також взяти, наприклад, суму квадратів відхилень  $\delta^2$  умовних середніх

$$\overline{y_{x_i}} = \frac{1}{n_i} \sum_{j=1} y_j n_{ij}$$

обчислених за даними кореляційної таблиці, від значень  $f(x_i)$  функції регресії:

$$\delta^2 = \sum_{i=1}^k \delta_i^2 n_i = \sum_{i=1}^k |\overline{y_{x_i}} - f(x_i)|^2 n_i \quad (13)$$

### ЗАВДАННЯ ДО ЛАБОРАТОРНОЇ РОБОТИ

1. За даними кореляційної таблиці обчислити умовні середні  $\overline{y_{xi}}$  ( $i = 1, \dots, k$ ).
2. Побудувати поле кореляції, тобто нанести точки  $M_i(x_i; \overline{y_{xi}})$ ,  $i = 1, \dots, k$ , на координатну площину. На основі цього зробити припущення про вигляд функції регресії (парабола, гіпербола і т.д.)
3. В залежності від вигляду функції регресії ((1), (3), (6) чи (9)) скласти відповідну систему рівнянь ((2), (5), (8) чи (11)). Розв'язати її і знайти невідомі параметри вибраної функції регресії.
4. Записати рівняння кривої регресії  $Y$  на  $X$ :  $\overline{y_x} = f(x)$  (з конкретною знайденою в пункті 3 функцією регресії  $f(x)$ ) та побудувати її графік.
5. Обчислити дисперсію (12) величини  $Y$  відносно кривої регресії  $Y$  на  $X$ .
6. Визначити суму квадратів відхилень  $\delta^2$  умовних середніх від значень функції регресії за формулою (13).

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\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\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\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\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\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\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\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\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\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\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\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\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