

**Міністерство освіти і науки України**  
**ЛЬВІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ імені ІВАНА ФРАНКА**  
**Факультет прикладної математики та інформатики**  
**Кафедра дискретного аналізу та інтелектуальних систем**

**ІНДИВІДУАЛЬНЕ ЗАВДАННЯ № 3**  
**з курсу “Математична статистика”**

Виконала:  
Студентка групи ПМІ-23  
Богданович Софія

## Постановка задачі

1. За даними кореляційної таблиці обчислити умовні середні  $\bar{y}_{xi}$  ( $i = 1, \dots, k$ ).
2. Побудувати поле кореляції, тобто нанести точки  $M_i(x_i; \bar{y}_{xi})$ ,  $i = 1, \dots, k$ , 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  $\alpha$  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  $Q_0$  обчислити прогнозоване значення  $y^*$  при заданому значенні  $x^*$ .

2.

| $Y \backslash X$ | 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 |    |    |    |    |    |    |

## Короткі теоретичні відомості

**Регресія** — це метод у статистиці й машинному навчанні, який дозволяє передбачати одне значення (змінну) на основі іншого.

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**:

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

За даними кореляційної таблиці обчислюють **умовні середні**  $\bar{y}_{xi}$  ( $i = 1, \dots, k$ )

$$\bar{y}_{xi} = \frac{y_1 n_{i1} + y_2 n_{i2} + \dots + y_l n_{il}}{n_i},$$

Складають **таблицю умовних середніх**  $\bar{y}_x$ :

|             |                 |                 |     |                 |     |                 |
|-------------|-----------------|-----------------|-----|-----------------|-----|-----------------|
| $x$         | $x_1$           | $x_2$           | ... | $x_i$           | ... | $x_k$           |
| $\bar{y}_x$ | $\bar{y}_{x_1}$ | $\bar{y}_{x_2}$ | ... | $\bar{y}_{x_i}$ | ... | $\bar{y}_{x_k}$ |

Для визначення вигляду функції регресії будують точки  $(x_i, \bar{y}_{xi})$  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. Лінійні регресія

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

$$y = ax + b$$

де  $a, b$  – параметри моделі.

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

$$\begin{cases} a \sum_{i=1}^k x_i^2 n_i + b \sum_{i=1}^k x_i n_i = \sum_{i=1}^k x_i n_i \bar{y}_{x_i} \\ a \sum_{i=1}^k x_i n_i + b \sum_{i=1}^k n_i = \sum_{i=1}^k n_i \bar{y}_{x_i} \end{cases}.$$

Перевірка правильності побудови рівняння регресії здійснюється за основним варіаційним рівнянням:

$$Q = Q_p + Q_o$$

Де  $Q = \sum_{i=1}^k (\bar{y}_{x_i} - \bar{y})^2 n_i$  – загальна варіація.

**Варіація регресії:**

$$Q_p = \sum_{i=0}^k (y_i^* - \bar{y})^2 n_i$$

**Варіація залишків:**

$$Q_o = \sum_{i=0}^k (\bar{y}_{x_i} - y_i^*)^2 n_i$$

Адекватність моделі вибіровим даним можна оцінити за **коефіцієнтом детермінації  $R^2$** , що показує частину варіації значень результативної ознаки  $Y$ , що пояснюється рівнянням регресії.

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

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:

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

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  $\alpha$ .

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).

### 2.1. Параболічна кореляція.

Рівняння параболи – параболічної регресії  $Y$  на  $X$  будемо шукати у вигляді

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

де  $a, b, c$  – невідомі параметри.

Невідомі параметри знаходимо з

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

Та підставляємо у рівняння параболи.

### 2.2. Гіперболічна кореляція.

Рівняння гіперболи – гіперболічної регресії  $Y$  на  $X$  будемо шукати у вигляді

$$y = \frac{a}{x} + b$$

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

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

### 2.3. Показникова кореляція.

Рівняння регресії:

$$y = ba^x$$

Розв'язуючи систему,

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

знаходимо  $\lg a$  і  $\lg b$ , а потім параметри  $a$  і  $b$  показникової функції.

### 2.4. Коренева кореляція.

Рівняння регресії має вигляд:

$$y = a\sqrt{x} + b$$

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

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

## 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

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

Щоб перевірити значущість коефіцієнта кореляції, треба перевірити гіпотезу:

- $H_0$  (нульова гіпотеза):  $\rho=0$  (у генеральній сукупності немає лінійної залежності).
- $H_1$  (альтернативна гіпотеза):  $\rho \neq 0$  (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 Стюдента 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).

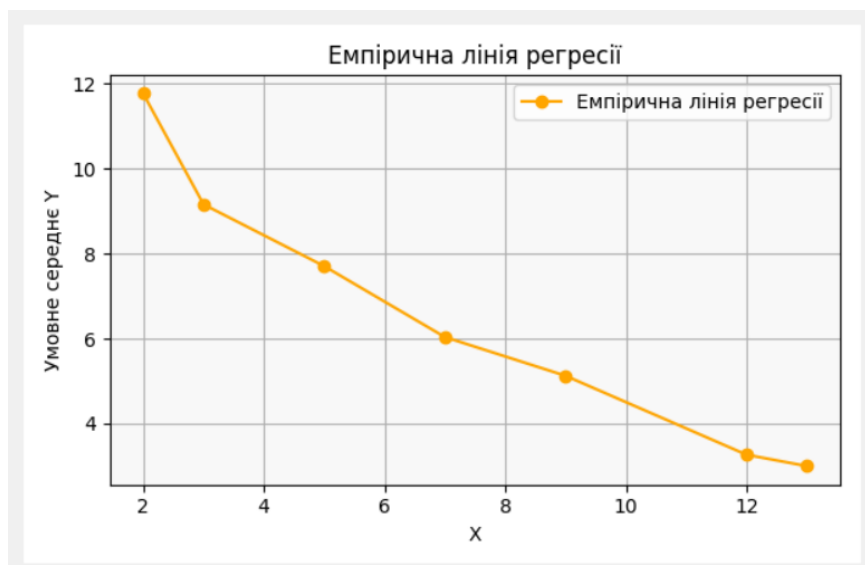
## Програмна реалізація та отримані результати

1. За даними кореляційної таблиці обчислити умовні середні  $\bar{y}_{xi}$  ( $i = 1, \dots, k$ ).

Таблиця умовних середніх:

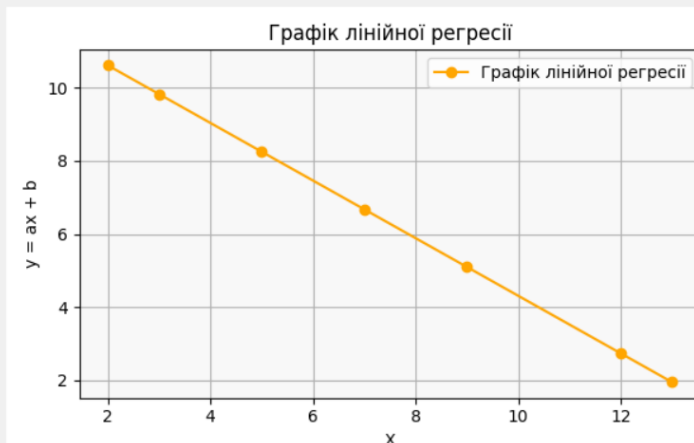
| x <sub>i</sub> | умовне середнє |
|----------------|----------------|
| 2.0            | 11.769231      |
| 3.0            | 9.160000       |
| 5.0            | 7.705882       |
| 7.0            | 6.037037       |
| 9.0            | 5.125000       |
| 12.0           | 3.266667       |
| 13.0           | 3.000000       |

2. Побудувати поле кореляції, тобто нанести точки  $M_i(x_i; \bar{y}_{xi})$ ,  $i = 1, \dots, k$ , на координатну площину, та емпіричну лінію регресії.



### 3. Побудувати лінійне рівняння регресії та намалювати графік.

a = -0.787213  
b = 12.183652



### 4. Обчислити коефіцієнт детермінації та перевірити адекватність побудованої лінійної моделі.

Варіаційне рівняння для перевірки правильності побудови моделі:

$$Q = Q_p + Q_o$$

Підставляємо наші значення варіацій:

$$Q = 1082.966661$$

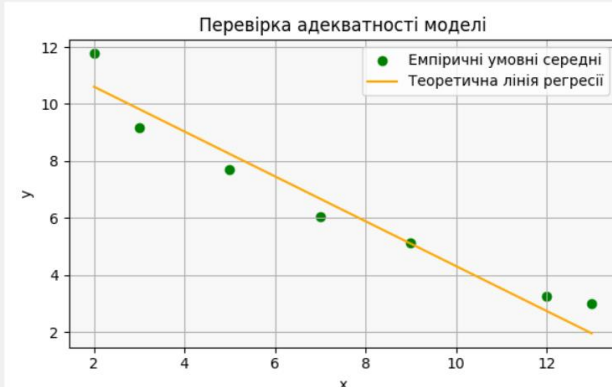
$$Q_p = 1012.475831$$

$$Q_o = 70.49083$$

Переконуємося, що наша модель побудована правильно:  $1082.966661 \approx 1082.966661$

Розрахуємо коефіцієнт детермінації:  $R^2 \approx 0.93491$

Оскільки значення коефіцієнта детермінації наближене до 1, робимо висновок, що лінійне рівняння регресії добре пояснює поведінку результативної ознаки.



### 5. Обчислити вибірковий лінійний коефіцієнт кореляції та за рівня значущості $\alpha$ перевірити його значущість



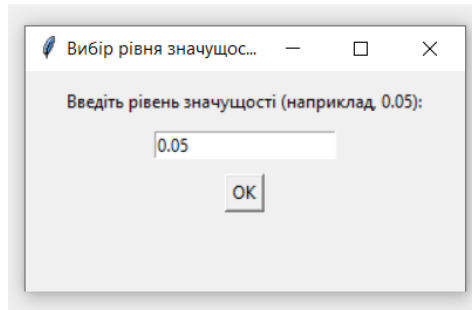
#### Рахуємо вибіркового лінійний коефіцієнт кореляції

$r = -0.966907$

Абсолютне значення коефіцієнта вказує на силу зв'язку, а знак на його напрям.

У нашому випадку між змінними є дуже сильний зворотний (негативний) лінійний зв'язок: коли одна змінна збільшується, інша, ймовірно, зменшується.

Обираємо рівень значущості:



#### Перевіримо статистичну значущість коефіцієнта

$H_0$  (нульова гіпотеза):  $\rho=0$  (у генеральній сукупності немає лінійної залежності).

$H_1$  (альтернативна гіпотеза):  $\rho \neq 0$  (лінійна залежність є)

Обраний рівень значущості: 0.05

Емпіричне значення статистики  $t$ : -44.197273

Критичне значення статистики  $t$ : 1.977561

Оскільки  $|t_{\text{емп}}| > t_{\text{крит}}$ , нульова гіпотеза відхиляється.

Коефіцієнт кореляції є статистично значущим, тобто зв'язок між змінними існує на рівні генеральної сукупності.

6. 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. Перевірити адекватність побудованої нелінійної моделі за  $F$ -критерієм.

Для кожного вигляду функції нелінійної регресії обирається рівень значущості, як показано на рисунку.

#### а. Параболічна регресія

##### Припускаємо параболічний вигляд функції нелінійної регресії

Тобто, що рівняння кривої регресії має вигляд  $y = ax^2 + bx + c$

$a = 0.050635$

$b = -1.477948$

$c = 13.917209$

Рівняння кривої регресії:  $y = 0.05x^2 + -1.48x + 13.92$

##### Перевірка адекватності побудованої моделі за $F$ -критерієм

$H_0$  (нульова гіпотеза): модель регресії не є значущою

$H_1$  (альтернативна гіпотеза): модель регресії є значущою.

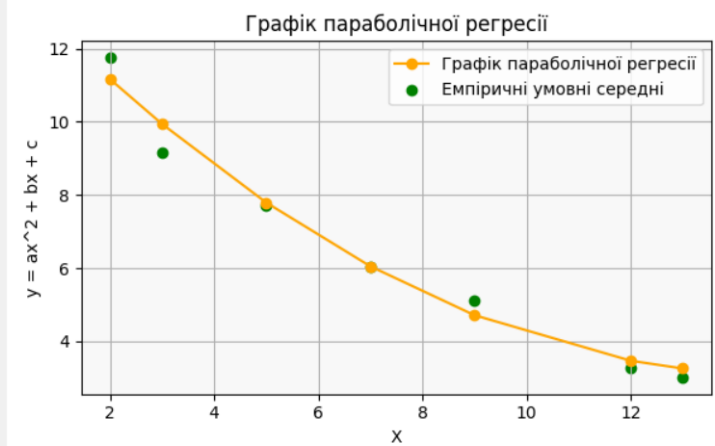
$Q = 1082.966661$

$Q_p = 1053.221188$

$Q_o = 29.745473$

Емпіричне значення статистики  $F$ : 2390.025231

Критичне значення статистики  $F$ : 3.063204



## в. Гіперболічна регресія

Припускаємо гіперболічний вигляд функції нелінійної регресії

Тобто, що рівняння кривої регресії має вигляд  $y = a/x + b$

$a = 18.087349$

$b = 3.039613$

Рівняння кривої регресії:  $y = 18.09/x + 3.04$

Перевірка адекватності побудованої моделі за F-критерієм

$H_0$  (нульова гіпотеза): модель регресії не є значущою

$H_1$  (альтернативна гіпотеза): модель регресії є значущою.

$Q = 1082.966661$

$Q_p = 1023.964846$

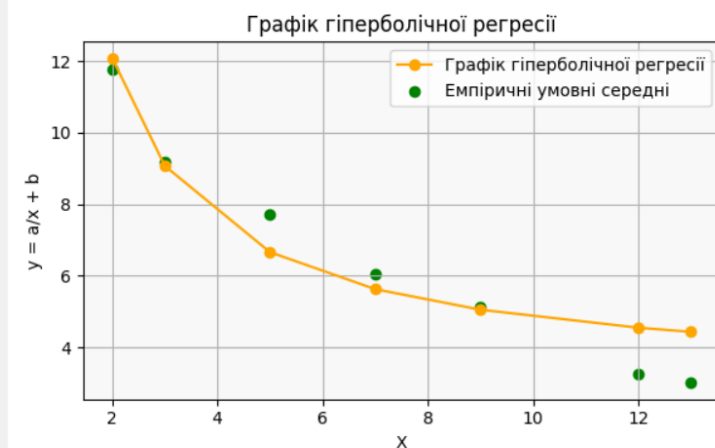
$Q_0 = 59.001815$

Емпіричне значення статистики F: 2360.253122

Критичне значення статистики F: 3.910747

Оскільки  $F_{\text{емпіричне}} > F_{\text{критичне}}$ , відхиляємо нульову гіпотезу.

Модель є адекватною при рівні значущості 0.05.



## с. Показникова регресія

Припускаємо показниковий вигляд функції нелінійної регресії

Тобто, що рівняння кривої регресії має вигляд  $y = ba^x$

$a = 0.888078$

$b = 14.06959$

Рівняння кривої регресії:  $y = 14.07 * 0.89^x$

Перевірка адекватності побудованої моделі за F-критерієм

$H_0$  (нульова гіпотеза): модель регресії не є значущою

$H_1$  (альтернативна гіпотеза): модель регресії є значущою.

$Q = 1082.966661$

$Q_p = 1027.952347$

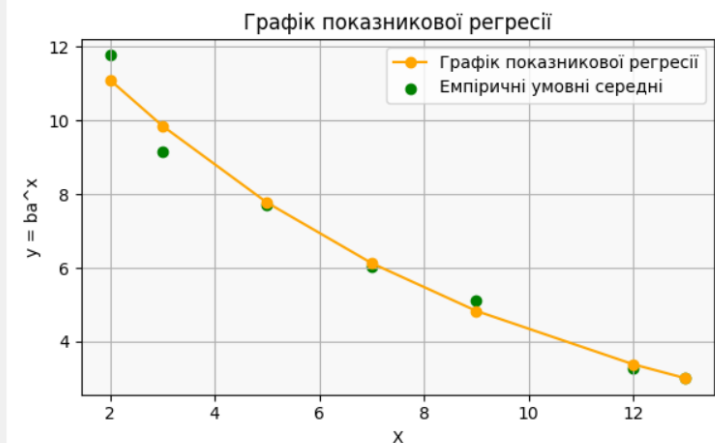
$Q_0 = 26.374502$

Емпіричне значення статистики F: 5300.631517

Критичне значення статистики F: 3.910747

Оскільки  $F_{\text{емпіричне}} > F_{\text{критичне}}$ , відхиляємо нульову гіпотезу.

Модель є адекватною при рівні значущості 0.05.



#### d. Коренева регресія

Оскільки  $F_{\text{емпіричне}} > F_{\text{критичне}}$ , відхиляємо нульову гіпотезу.  
 Модель є адекватною при рівні значущості 0.05.

$a = -3.884805$   
 $b = 16.571298$

Рівняння кривої регресії:  $y = 16.57 * -3.88^x$

**Перевірка адекватності побудованої моделі за F-критерієм**

$H_0$  (нульова гіпотеза): модель регресії не є значущою

$H_1$  (альтернативна гіпотеза): модель регресії є значущою.

$Q = 1082.966661$

$Q_p = 1054.410207$

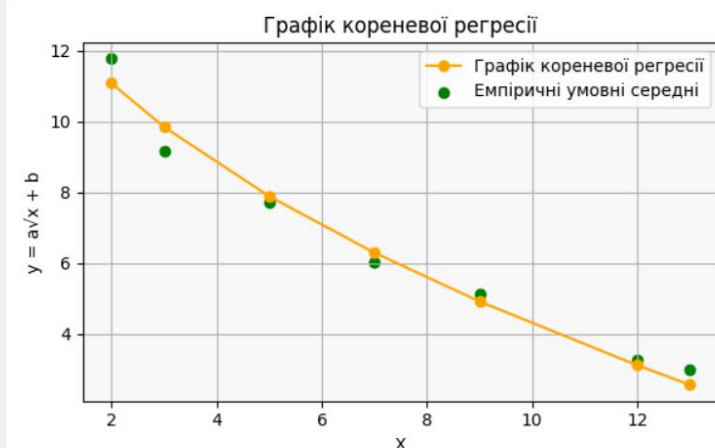
$Q_0 = 28.556453$

Емпіричне значення статистики F: 5021.624568

Критичне значення статистики F: 3.910747

Оскільки  $F_{\text{емпіричне}} > F_{\text{критичне}}$ , відхиляємо нульову гіпотезу.

Модель є адекватною при рівні значущості 0.05.



#### Аналіз результатів:

Зі значення коефіцієнта детермінації для лінійної моделі можна зробити висновок, що вона добре описує залежність між статистичними змінними.

Крім того, перевірка адекватності нелінійних моделей за допомогою критерію Фішера показала, що всі ці моделі є адекватними при стандартних рівнях значущості 0.01, 0.05 і 0.1.

Однак, залишкова варіація  $Q_0$ , яка відображає похибку моделі, виявилася найменшою саме для показникової моделі.

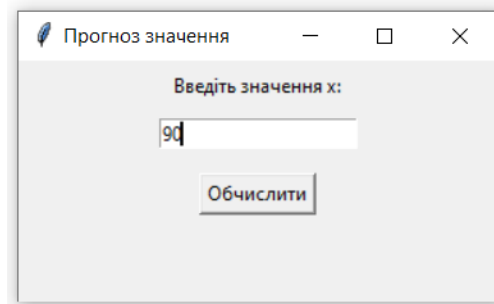
На основі цього можна зробити висновок, що показникова модель найкраще описує залежність у досліджуваному випадку.

7. За моделлю з найменшою залишковою варіацією  $Q_0$  обчислити прогнозоване значення  $y^*$  при заданому значенні  $x^*$ .

Обчислення прогнозованого значення  $y^*$  при заданому значенні  $x^*$

Моделлю з найменшим значенням залишкової варіації  $Q_0$  є показникова модель.

Вибіркові значення  $x^*$ : 2, 3, 5, 7, 9, 12, 13



Прогнозоване значення для  $x = 90.0$ :  
 $y = 0.000$

## Висновок

У ході лабораторної роботи було виконано кореляційно-регресійний аналіз на основі кореляційної таблиці. Розраховано умовні середні, побудовано поле кореляції та емпіричну лінію регресії. Знайдено рівняння лінійної регресії, обчислено коефіцієнт детермінації, перевірено адекватність моделі та значущість коефіцієнта кореляції.

Крім того, було досліджено чотири варіанти нелінійної регресії (параболічну, гіперболічну, показникову, логарифмічну). Для кожної з них складено відповідну систему рівнянь, знайдено параметри, побудовано графіки та перевірено адекватність моделей за F-критерієм. На основі моделі з найменшою залишковою варіацією зроблено прогноз значення  $y^*$  при заданому  $x^*$ . Проведене дослідження дозволило порівняти точність різних моделей та зробити висновок про найбільш адекватну для подальших прогнозів.