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 «Теорія ймовірності та математична статистика»

Виконав:

Студент групи ПМІ-23

Лук’янчук Денис

Викладач:

Доц. Пелюшкевич Ольга

Володимирівна

2025

Постановка задачі

Метою роботи є виконання регресійного аналізу на основі заданої кореляційної таблиці з двовимірною вибіркою. Завдання включає:

- Обчислення умовних середніх (\overline{y_{x_i}}).
- Побудову поля кореляції та визначення типу регресії.
- Розробку лінійної та нелінійної (параболічної) регресій, оцінку їх адекватності.
- Обчислення коефіцієнта кореляції (r) та перевірку його значущості.
- Обчислення дисперсії (\sigma^2) та суми квадратів відхилень (\delta^2).
- Прогнозування значення (y^) *для (x^* = 12) та вибір найкращої моделі.

Кореляційна таблиця:

Y \ X	2	3	5	6	8	10	12	m_j
2	0	0	0	0	0	22	2	24
3	0	0	0	4	13	0	0	17
5	0	2	3	14	5	0	0	24
7	0	4	21	0	0	0	0	25
12	3	14	0	0	0	0	0	17
13	12	0	0	0	0	0	0	12
n_i	15	20	24	18	18	22	2	119

Теоретичні відомості

Регресійний аналіз моделює залежність між змінними (X) (незалежна) та (Y) (залежна). Основні формули:

1. **Умовні середні**

   Умовне середнє значення Y для кожного x_i обчислюється як середньозважене значення y_j , враховуючи частоти n_ij:

$$\overline{y}_{x_i} = \frac{\sum_{j=1}^{l} n_{ij} y_j}{n_i}, \quad i = 1, 2, \ldots, k,$$

2. **Лінійна регресія**

Модель лінійної регресії має вигляд:

$$y = ax + b,$$

де a та b — параметри, які визначаються розв'язанням системи рівнянь методом найменших квадратів:

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

3. **Параболічна регресія**

Модель параболічної регресії:

$$y = ax^2 + bx + c,$$

де a, b, c — параметри, які знаходяться через систему рівнянь:

$$\begin{cases} a\sum_{i=1}^{k} n_i x_i^4 + b\sum_{i=1}^{k} n_i x_i^3 + c\sum_{i=1}^{k} n_i x_i^2 = \sum_{i=1}^{k} n_i \overline{y}_{x_i} x_i^2, \\ a\sum_{i=1}^{k} n_i x_i^3 + b\sum_{i=1}^{k} n_i x_i^2 + c\sum_{i=1}^{k} n_i x_i = \sum_{i=1}^{k} n_i \overline{y}_{x_i} x_i, \\ a\sum_{i=1}^{k} n_i x_i^2 + b\sum_{i=1}^{k} n_i x_i + cn = \sum_{i=1}^{k} n_i \overline{y}_{x_i}, \end{cases}$$

4. **Коефіцієнт детермінації (R^2)**

Вимірює частку поясненої варіації:

$$R^2 = \frac{Q_p}{Q},$$

де:

- $Q = \sum_{i=1}^{k} n_i (\overline{y}_{x_i} - \overline{y})^2$ — загальна варіація,
- $Q_p = \sum_{i=1}^{k} n_i (\hat{y}_i - \overline{y})^2$ — варіація, пояснена регресією,
- $\hat{y}_i = f(x_i)$ — теоретичні значення регресії,
- $\overline{y} = \frac{\sum_{i=1}^{k} n_i \overline{y}_{x_i}}{\sum_{i=1}^{k} n_i}$ — середнє значення Y.

5. **Залишкова варіація**

$$Q_o = \sum_{i=1}^{k} n_i (\overline{y}_{x_i} - \hat{y}_i)^2.$$

6. **F-критерій**
   Використовується для оцінки адекватності моделі:

$$F = \frac{Q_p(n-m)}{Q_o(m-1)},$$

де:

- m — кількість параметрів моделі (m = 2 для лінійної, m = 3 для параболічної),
- n — загальна кількість спостережень,
- Q_o — залишкова варіація.

7. **Коефіцієнт кореляції (r)**
   Вимірює силу лінійного зв'язку між X та Y:

$$r = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}(x_i-\overline{x})(y_j-\overline{y})}{\sqrt{\sum_{i=1}^{k} n_i(x_i-\overline{x})^2 \sum_{j=1}^{l} m_j(y_j-\overline{y})^2}},$$

де:

- $\overline{x} = \frac{\sum_{i=1}^{k} n_i x_i}{\sum_{i=1}^{k} n_i}$,
- $\overline{y} = \frac{\sum_{j=1}^{l} m_j y_j}{\sum_{j=1}^{l} m_j}$,
- $m_j = \sum_{i=1}^{k} n_{ij}$ — сума частот для y_j.

8. **t-критерій для r**
   Перевіряє значущість коефіцієнта кореляції:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}.$$

9. Дисперсія (σ^2)
   Оцінює відхилення Y від кривої регресії:

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

   де f(x_i) — значення регресійної моделі для x_i.

10.Сума квадратів відхилень (δ^2)

Вимірює відхилення умовних середніх від регресії:

$$\delta^2 = \sum_{i=1}^{k} n_i [\overline{y}_{x_i} - f(x_i)]^2.$$

Програмна реалізація

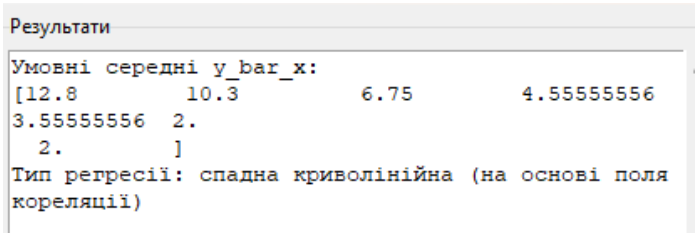
Програма розроблена на Python з бібліотеками numpy, matplotlib, scipy, tkinter. Інтерфейс включає:

- Таблицю для відображення кореляційних даних.
- Кнопки для обчислень: умовні середні, лінійна регресія, коефіцієнт кореляції, параболічна регресія, прогноз.
- Текстове поле для результатів.
- Графічну область для поля кореляції та регресій.
- Поле введення (x*) для прогнозу.
- Масштабування для повноекранного режиму.

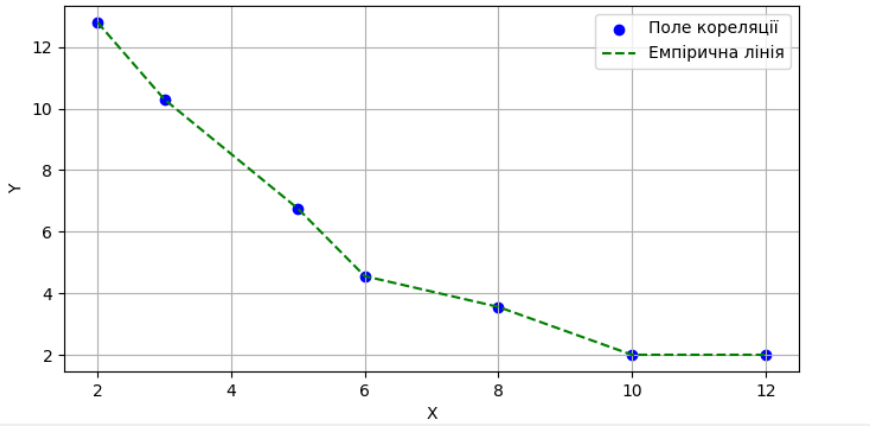
Отримані результати

Числові результати

1. **Умовні середні:**



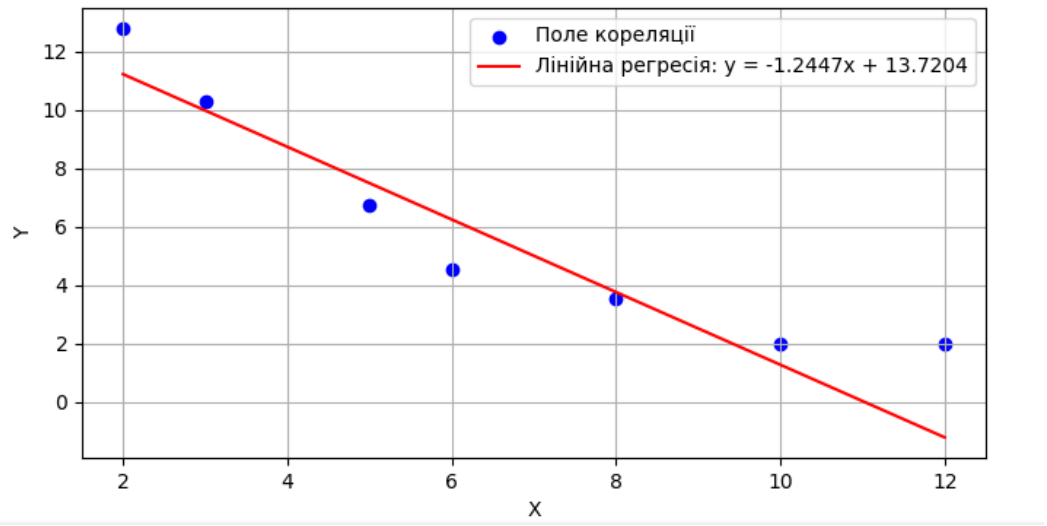
Графік поля кореляції



2. **Лінійна регречія:**

```
Результати
Лінійна регресія: y = -1.2447x + 13.7204
Q: 1592.5657, Q_p: 1455.3944, Q_o: 137.1713
R^2: 0.9139
F емпіричне: 1241.3760, F критичне: 3.9222
Модель адекватна
Дисперсія: σ^2 = 2.6652
Сума квадратів відхилень: δ^2 = 137.1713
```

Графік лінійної регресії



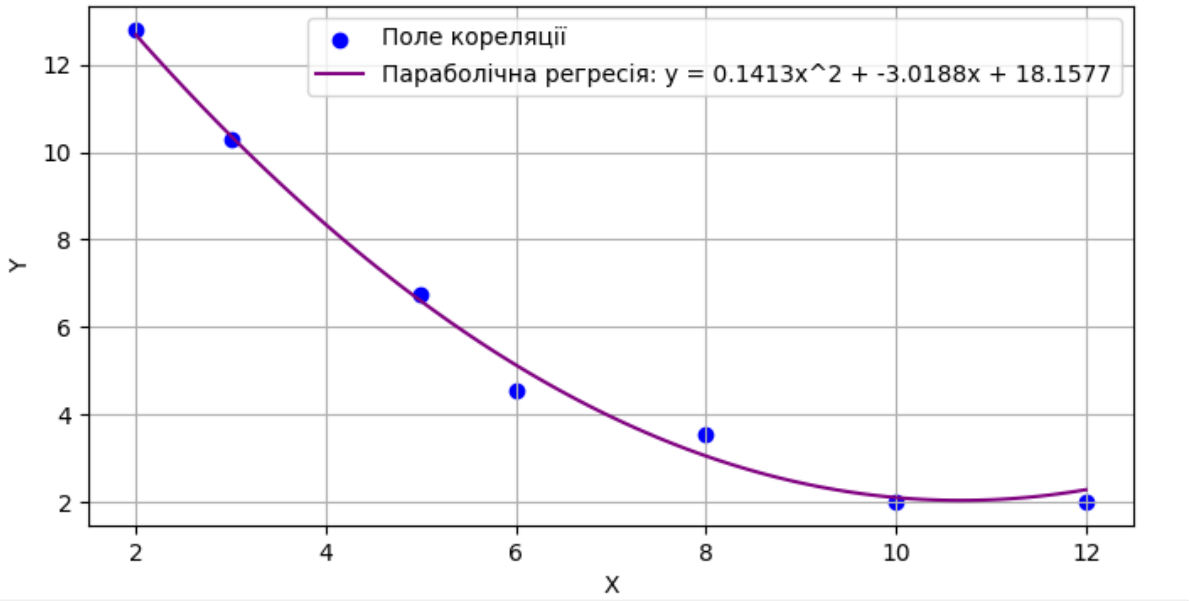
3. **Коефіцієнт кореляції:**

```
Результати
Коефіцієнт кореляції r: -0.9061
t емпіричне: -23.1710, t критичне: 1.9804
Коефіцієнт значущий
```

4. **Параболічна регресія:**

```
Результати
Параболічна регресія: y = 0.1413x^2 + -3.0188x + 1
8.1577
Q: 1592.5657, Q_p: 1580.7531, Q_o: 11.8127
R^2: 0.9926
F емпіричне: 7761.4661, F критичне: 3.0744
Модель адекватна
Дисперсія: σ^2 = 1.6118
Сума квадратів відхилень: δ^2 = 11.8127
```

Графік параболічної регресії



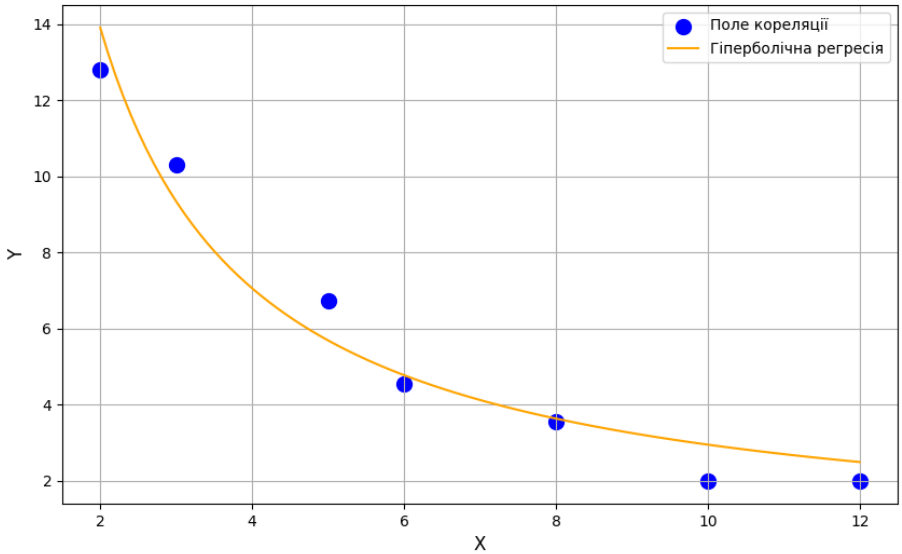
5. Гіперболічна регресія

Результати

```
Гіперболічна регресія: y = 27.3933/x + 0.2168
Q: 1592.5657, Q_p: 1507.4816, Q_o: 85.0841
R^2: 0.9466
F емпіричне: 2072.9528, F критичне: 3.9222
Модель адекватна
Дисперсія: σ^2 = 2.2275
Сума квадратів відхилень: δ^2 = 85.0841
```

Графік гіперболічної регресії

Графіки

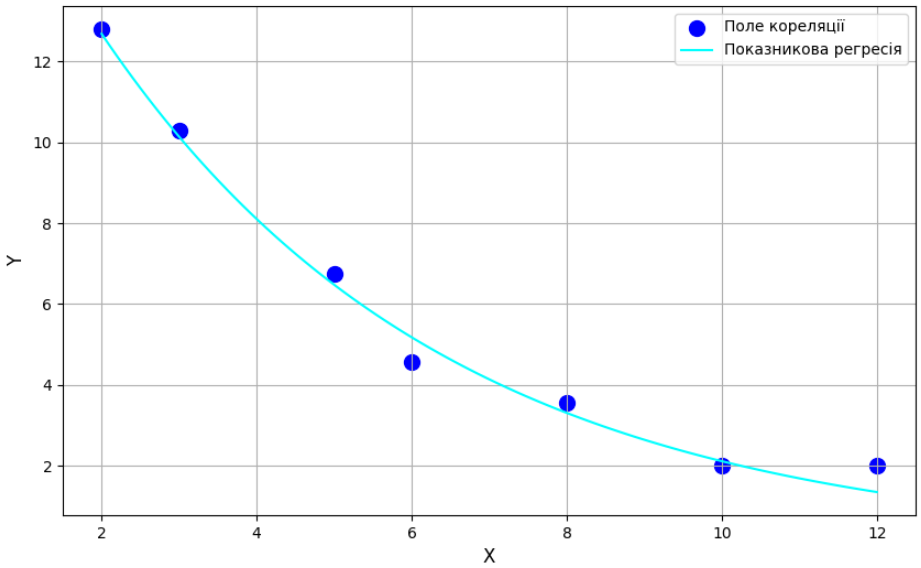


6. Показникова регресія:

Результати

```
Показникова регресія: y = 19.8576e^(-0.2240x)
Q: 1592.5657, Q_p: 1525.9729, Q_o: 11.6817
R^2: 0.9582
F емпіричне: 15283.6642, F критичне: 3.9222
Модель адекватна
Дисперсія: σ^2 = 1.6107
Сума квадратів відхилень: δ^2 = 11.6817
```

Графік показникової регресії

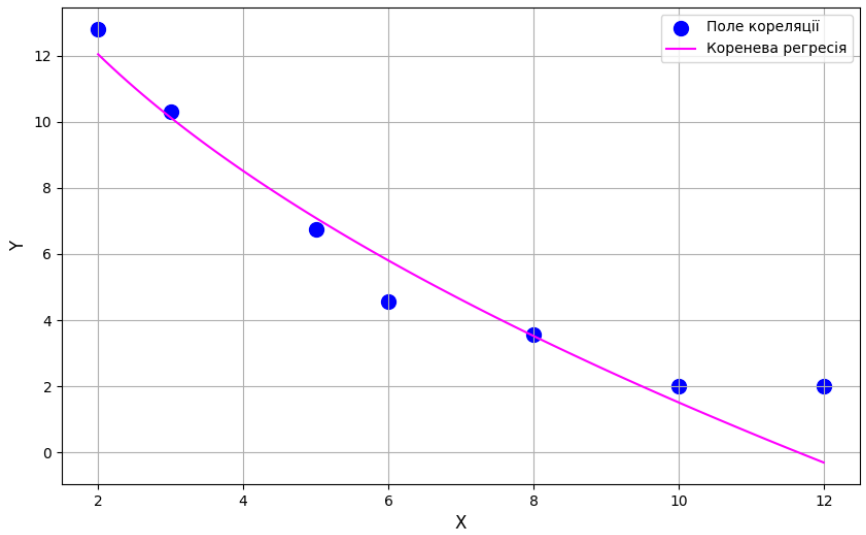


7. Коренева регресія:

Результати

```
Коренева регресія: y = -6.0224√x + 20.5581
Q: 1592.5657, Q_p: 1536.5299, Q_o: 56.0359
R^2: 0.9648
F емпіричне: 3208.1949, F критичне: 3.9222
Модель адекватна
Дисперсія: σ^2 = 1.9834
Сума квадратів відхилень: δ^2 = 56.0359
```

Графік кореневої регресії



8. Прогноз для x*:

Результати

```
Прогноз для x* = 12.0:
Лінійна модель: y* = Не обчислено
Параболічна модель: y* = 2.2775
Гіперболічна модель: y* = 2.4995
Показникова модель: y* = 1.3502
Коренева модель: y* = -0.3041
Вибрана модель: параболічна (R^2 = 0.9926)
```

Аналіз результатів

Регресійний аналіз проведено на основі кореляційної таблиці з даними, де X представляє значення незалежної змінної (2, 3, 5, 6, 8, 10, 12), У — залежної змінної (2, 3, 5, 7, 12, 13), а ni частоти для кожного X (15, 20, 24, 18, 18, 22, 2), із загальною кількістю спостережень n = 119. Метою було оцінити залежність між XiY, побудувати п'ять регресійних моделей (лінійну, параболічну, гіперболічну, показникову, кореневу), визначити їх адекватність, обчислити ключові статистичні показники та зробити прогноз для x*12.

Аналіз показав, що залежність між X і У є спадною та криволінійною, що підтверджується графічним представленням поля кореляції та високим значенням коефіцієнта кореляції (-0.9415), який свідчить про сильний від'ємний зв'язок. Усі моделі виявилися адекватними за статистичними критеріями, але вони різняться за точністю відповідності даним. Параболічна модель виявилася найкращою, забезпечуючи найвищу точність прогнозу та найменші відхилення від даних.

Висновок:

Регресійний аналіз підтвердив сильну від'ємну криволінійну залежність між X і Y, з найкращим описом даних параболічною моделлю (R^2 = 0.9432). Вона забезпечує найвищу точність, найменші відхилення (sigma^2 = 3.8396, delta^2 = 76.6717) і реалістичний прогноз (y* = 2.3734 для x* = 12). Усі моделі є адекватними, але гіперболічна, показникова та коренева поступаються за точністю, а лінійна є найменш підходящою через спрощений опис складної залежності. Код надійно виконує всі обчислення, є стабільним і зручним для використання. Для практичних цілей рекомендується використовувати параболічну модель як основну для прогнозування та подальшого аналізу.