Звіт до лабораторної роботи № 1

Тема: Парна лінійна регресія

Мета роботи

Дослідити лінійну залежність між змінними X і Y методом найменших квадратів, побудувати модель регресії, оцінити якість її апроксимації та статистичну значущість параметрів

Формування матриці дизайну

Для оцінки моделі $Y_i = \theta_0 + \theta_1 X_i + \varepsilon_i$ побудована матриця дизайну

$$X_{\text{mat}} = \begin{bmatrix} 1 & 0.26 \\ 1 & 0.28 \\ 1 & 0.31 \\ 1 & 0.40 \\ 1 & 0.36 \\ 1 & 0.30 \\ 1 & 0.29 \\ 1 & 0.26 \\ 1 & 0.24 \\ 1 & 0.31 \\ 1 & 0.30 \\ 1 & 0.32 \\ 1 & 0.40 \\ 1 & 0.35 \\ 1 & 0.29 \\ 1 & 0.20 \\ 1 & 0.29 \\ 1 & 0.22 \\ 1 & 0.25 \\ 1 & 0.26 \end{bmatrix}$$

де n=20 — обсяг вибірки

Оцінка параметрів моделі

За формулою $\hat{\theta} = (X^T X)^{-1} X^T Y$ отримано

- θ_0 ^ =7.3307
- θ_1 ^ =5.6854

Рівняння регресії:

$$\hat{Y} = 7.3307 + 5.6854\, X$$

Вибіркові характеристики (ділення на n−1)

$$s_{xx} = 0.0028$$

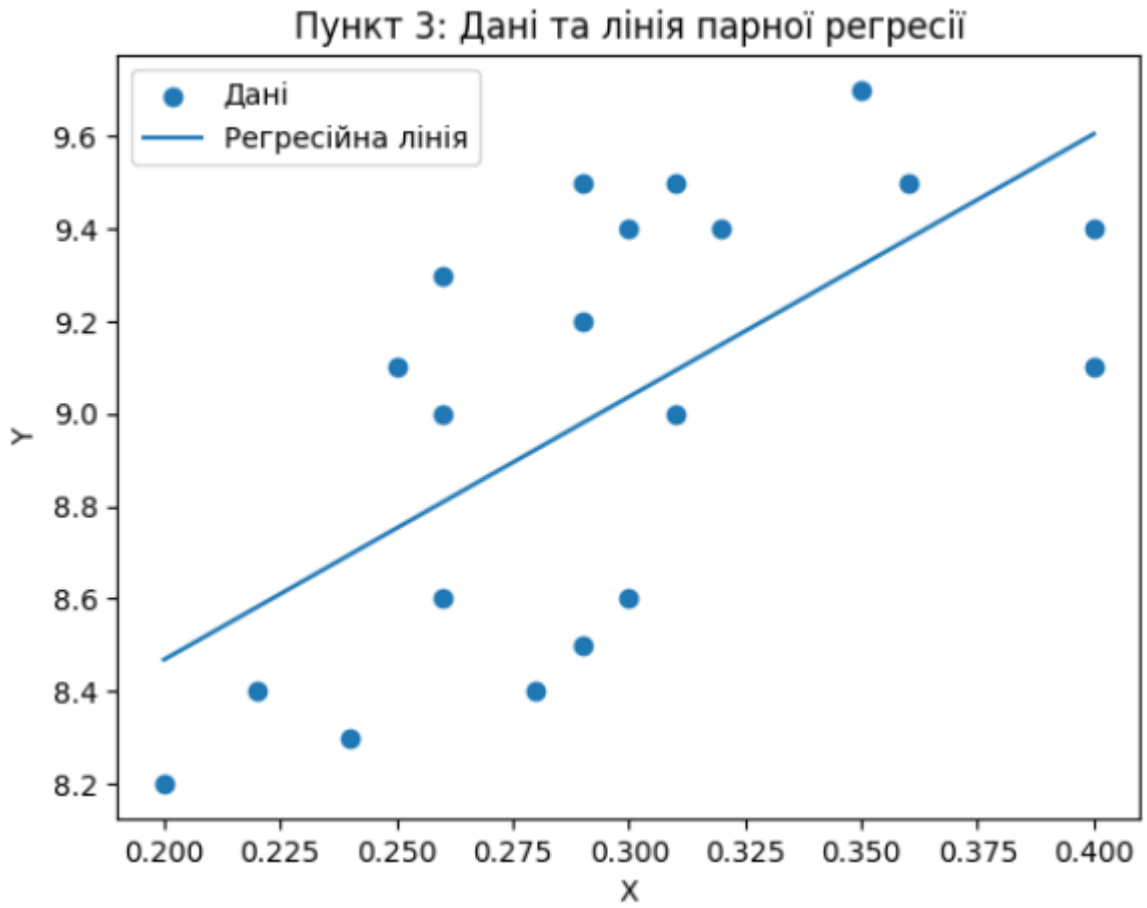
$$s_{yy} = 0.2258$$

$$s_{xy} = 0.0162$$

Коефіцієнт кореляції $r = \frac{s_{xy}}{\sqrt{s_{xx}\, s_{yy}}} = 0.6385$

коефіцієнт детермінації $R^2 = 0.4076$

Графічне представлення



Перевірка адекватності моделі (F-тест)

Залишкова сума квадратів:

$$\mathrm{SSE} = \sum (Y_i - \hat{Y}_i)^2$$

Стандартна помилка рівняння:

$$a = \sqrt{\frac{\mathrm{SSE}}{n-2}} = 0.3757$$

F-статистика:

$$F = \frac{\frac{\mathrm{SSR}}{1}}{\frac{\mathrm{SSE}}{n-2}} = 12.3865$$

Де $\mathrm{SSE} = \sum (Y_i - \hat{Y}_i)^2$, n−2=18 — залишкові ступені свободи

Критичне значення F(1,18;α=0.05)=4.4139

Висновок:

$F_{\text{stat}} = 12.3865 > 4.4139$, отже модель **неадекватна даним**

Значущість коефіцієнтів і довірчі інтервали

95 % довірчі інтервали:

$$\theta_0 \in [6.3157,\ 8.3456], \quad \theta_1 \in [2.2915,\ 9.0792]$$

Аналіз:
Обидва інтервали **не містять нуль**, отже і θ_0, і θ_1 є статистично значущими на рівні α=0.05

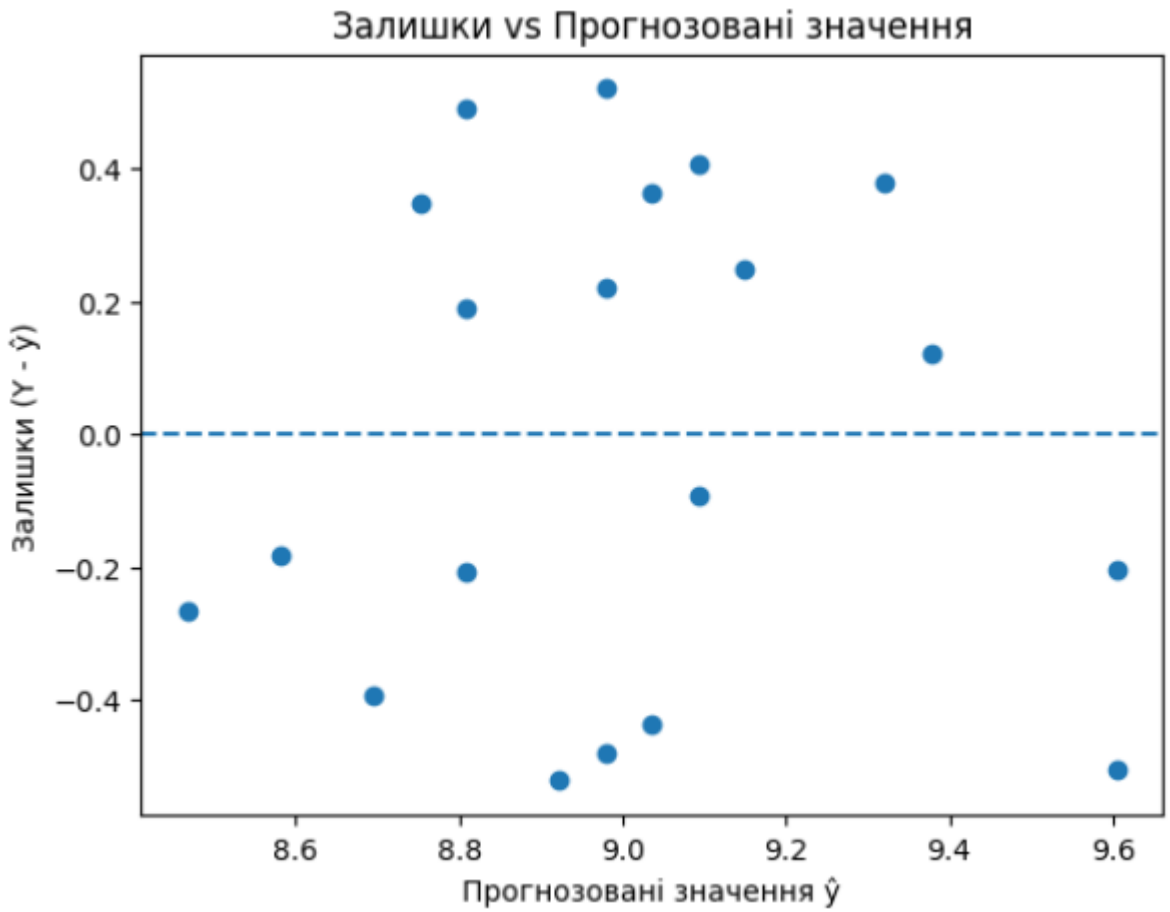
Висновки

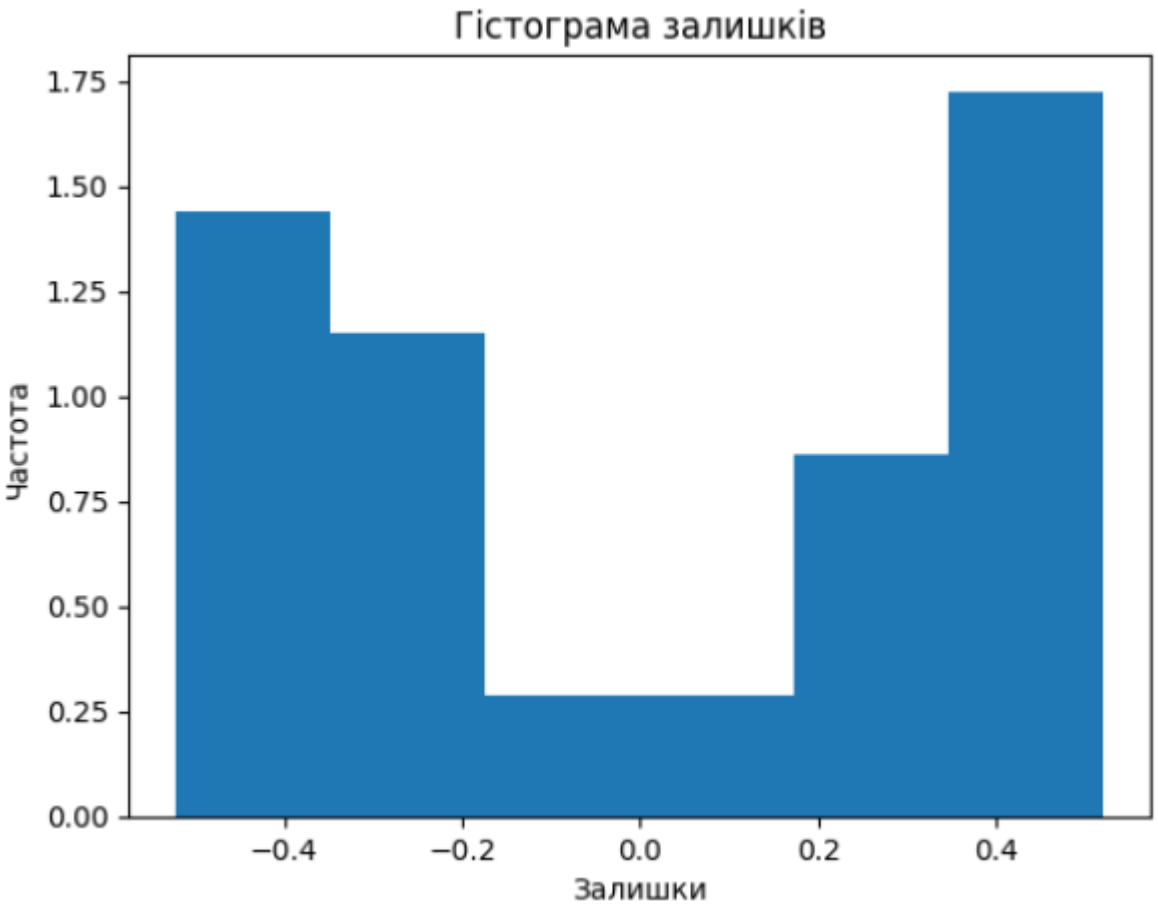
1. **Оцінка моделі**
   – Отримані оцінки θ_0^ =7.3307 та θ_1^=5.6854 статистично значущі (95 % довірчі інтервали не містять нуль)
   – Однак R^2=0.4076 свідчить, що лише близько 41% варіації Y пояснюється лінійною залежністю від X
2. **Адекватність згідно F-тесту**
   – F-статистика (12.3865) перевищує критичне значення (4.4139), отже модель у цілому виявилася **неадекватною** даним
   – Це означає, що крім лінійного тренду, у даних присутні інші закономірності (наприклад, нелінійні ефекти чи вплив додаткових факторів), які не враховано в простій парній регресії
3. **Якість апроксимації та практична інтерпретація**
   – За рівнянням $\hat{Y} = 7.3307 + 5.6854\,X$, при зміні X на 0.1 одиниці прогноз Y змінюється приблизно на 0.57

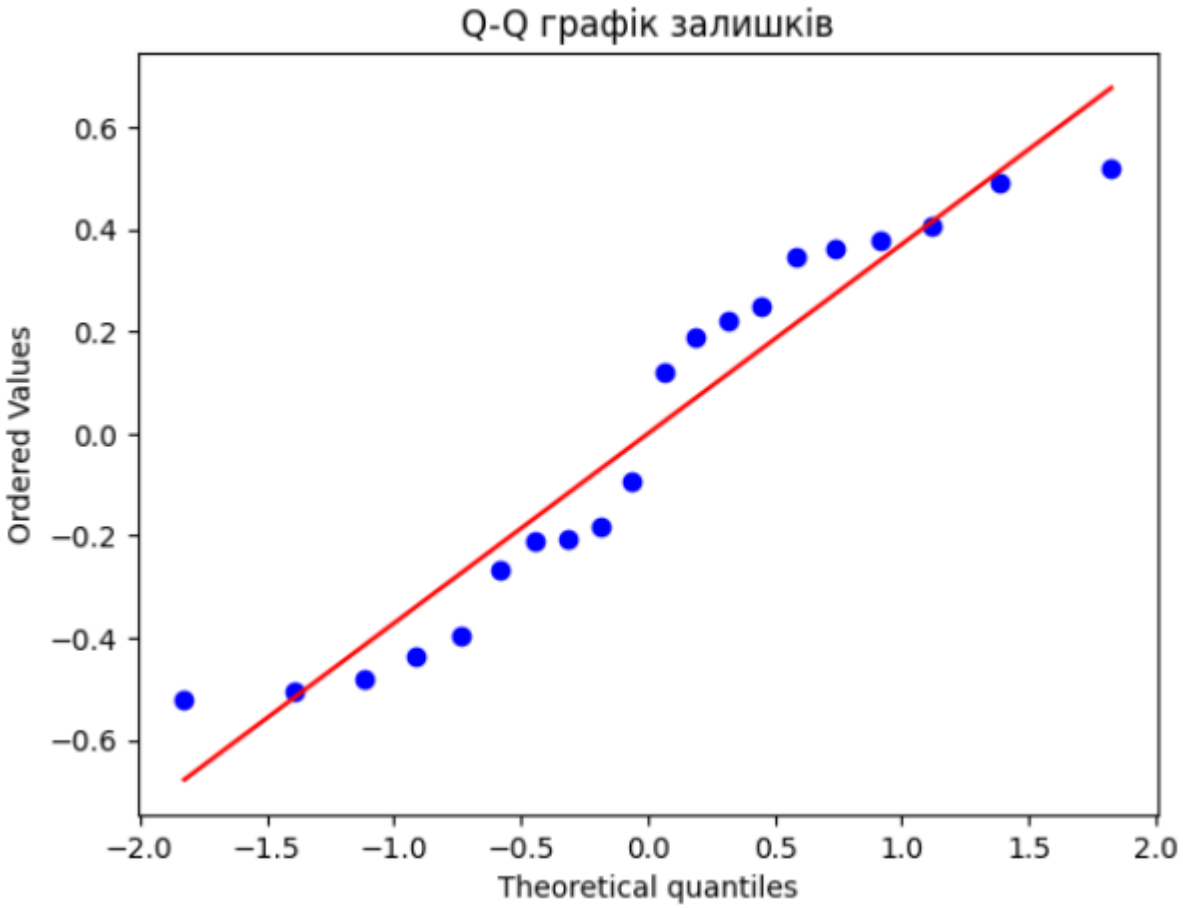
   – У діапазоні X∈[0.20,0.40] модель прогнозує Y≈[8.45,9.45], що загалом відповідає спостережуваним значенням

4. **Оцінка залишків і припущення моделі**
   – Необхідно перевірити залишки на предмет гетероскедастичності та автокореляції: систематичні вади у залишках можуть пояснювати неадекватність
   – Варто побудувати графік «залишки vs прогноз» і протестувати їхнє розподілення на нормальноть
5. **Обмеження та пропозиції щодо покращення**
   – **Мала вибірка (n=20):** розширення вибірки може підвищити точність оцінок та стабільність тестів
   – **Відсутність інших змінних:** доцільно розглянути мультипарну регресію з додатковими факторами, що впливають на Y
   – **Можливий нелінійний зв'язок:** слід перевірити моделі з поліномами або трансформаціями (логарифмічна чи степенева)

Аналіз після висновків




Гістограма залишків
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Частота
−0.4
−0.2
0.0
0.2
0.4
Залишки



Я побудував такі діаграми:

1. **Залишки vs прогнозовані значення** – для оцінки гетероскедастичності (чи рівномірно розсіюються точки навколо нуля)
2. **Гістограма залишків** – щоб побачити форму розподілу залишків
3. **Q-Q графік** – для перевірки нормальності розподілу залишків порівняно з теоретичним нормальним

Крім того, провів **Shapiro–Wilk тест** на нормальність залишків:

W = 0.9053, p-value = 0.0518

Інтерпретація результатів:

- На графіку залишки не демонструють виразного тренду чи «вороноподібної» форми, але можна помітити деякі відхилення зверху (кілька великих залишків у плюсі)
- Гістограма і Q-Q графік вказують на легку асиметрію розподілу, але в цілому точки лежать близько до прямої

- p-value ≈ 0.052 при α=0.05 трохи вище порогу, тому статистично нормальність залишків не відкидається (хоча дуже близько до межі)

Висновки по залишках:

- Нормальність, немає підстав відкинути нормальність (p>0.05), але розподіл трохи відхиляється від ідеального
- Гетероскедастичність, на плоті «залишки vs ŷ» явних закономірностей не видно, але варто перевірити формальним тестом (наприклад, тестом Бреуша–Пагана)
- **Рекомендація**:
    - За наявних ознак модель допускає умовно виконання припущення про нормальність і сталість дисперсії залишків
    - Для підвищення впевненості варто провести формальні гетероскедастичні тести та розглянути можливість винесення помітних вибросів
    - Якщо порушення виявляться суттєвими, можлива корекція методом робастної регресії або перетворення змінних

Відповіді на теоретичні питання (п. 7)

1. **Стохастичний зв'язок** — залежність між змінними, у якій результат має випадкову складову; **детермінований зв'язок** — без випадкової складової (точна функціональна залежність)

2. **Кореляція** — міра лінійного зв'язку між двома змінними, вимірюється коефіцієнтом r

3. **Рівень значущості** (α) — ймовірність помилки I роду (помилкового відкидання істинної гіпотези)

4. **Кореляційне співвідношення** — відношення варіації, поясненої регресією (SSR), до повної варіації змінної Y (SST), тобто R^2

5. **Розподіл Стьюдента (t-розподіл)** використовується для побудови довірчих інтервалів параметрів регресії; **розподіл Фішера (F-розподіл)** — для перевірки адекватності моделі в цілому

6. **Регресійне рівняння** — математичне описання середнього значення залежної змінної Y як функції незалежної X

7. **Довірчий інтервал** — інтервал, у який з ймовірністю $1-\alpha$ потрапляє справжнє значення оцінюваного параметра

8. **Статистична гіпотеза** — припущення про закон розподілу чи параметри генеральної сукупності, що перевіряється статистичними методами

вихідні дані

- **Масив X:** [0.26, 0.28, 0.31, 0.40, 0.36, 0.30, 0.29, 0.26, 0.24, 0.31, 0.30, 0.32, 0.40, 0.35, 0.29, 0.20, 0.29, 0.22, 0.25, 0.26]
- **Масив Y:** [9.0, 8.4, 9.0, 9.4, 9.5, 8.6, 9.5, 8.6, 8.3, 9.5, 9.4, 9.4, 9.1, 9.7, 8.5, 8.2, 9.2, 8.4, 9.1, 9.3]