МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
«ХАРКІВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»

Кафедра комп'ютерної математики і аналізу даних

ЗВІТ

про виконання Лабораторної роботи №2 за темою

«Побудова класичної лінійної моделі множинної регресії та оцінка параметрів моделі»
з дисципліни «Аналіз даних і часових рядів»

Група КН-122а

Виконавець Жарський Н.Д.

Викладач Гардер С.Є.

Харків 2025

Лабораторна робота №2

Мета роботи

Перевірити практично процедуру побудови моделі множинної лінійної регресії, оцінити її адекватність і значущість параметрів на двох наборах даних

Теоретичні відомості

1. Модель множинної регресії має вигляд

   $$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$

   де коефіцієнти βі оцінюються методом найменших квадратів:

   $$\hat{\beta} = (X^T X)^{-1} X^T y$$

2. Коефіцієнт детермінації

   $$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

   показує частку дисперсії відгуку, пояснену моделлю

3. F-тест адекватності перевіряє, чи модель загалом статистично значуща:

   $$F = \frac{R^2/p}{(1 - R^2)/(n - p - 1)}$$

   порівнюється з критичним значенням F_{crit}

4. t-тест для множинного коефіцієнта кореляції(R)

   $$t_R = R\sqrt{\frac{n - p - 1}{1 - R^2}}$$

   дає змогу оцінити значущість загальної кореляції

5. t-тести для окремих коефіцієнтів

$$t_i = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}, \quad SE(\hat{\beta}_i) = \sqrt{s^2 \, (X^T X)^{-1}_{ii}}$$

6. Довірчі інтервали для βi:

$$\hat{\beta}_i \pm t_{crit} \, SE(\hat{\beta}_i)$$

Опис даних

- Набір №1 n1 =20 спостережень, p=5 факторів (матриця X1 20×5, вектор y1)
- Набір №2 n2 =29 спостережень, p=4 фактори (матриця X2 29×4, вектор y2)

Методика виконання

1. Створено дизайн-матрицю $X_{design} = [1 \; X]$
2. Оцінено β^ , обчислено прогноз y^ і залишки e=y−y^
3. Знайдено суму квадратів помилок SSE і загальну суму квадратів Syy
4. Обчислено R^2, F-статистику та F_{crit}
5. Проведено t-тест для множинного R і окремих β_i^ ; розраховано стандартні помилки та довірчі інтервали
6. Оцінено дисперсії оцінок: $\mathrm{Var}(\beta_i\text{^}) = s^2 (X^T X)^{-1}_{ii}$

Результати

1. Набір даних №1

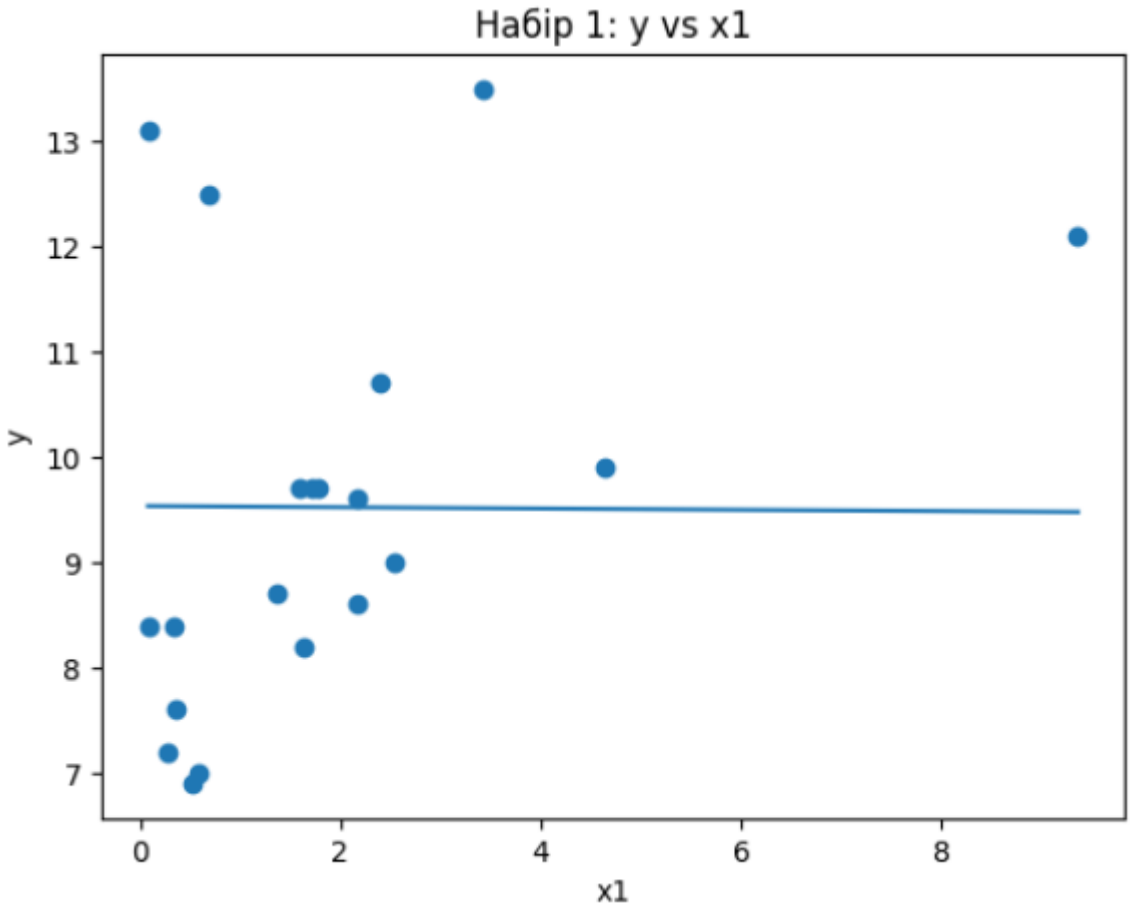
- n=20,p=5
- β^ =[3.5146,−0.0061,15.5425,0.1099,4.4746,−2.9325]
- Стандартна помилка рівняння s=1.5990
- SSE=35.7955,Syy =74.1575,R^2=0.5173
- F=3.0008,Fcrit(1,14;α=0.05) =2.9582 → модель адекватна
- t_R =3.8735,t_{crit} =±2.1448 → R значущий

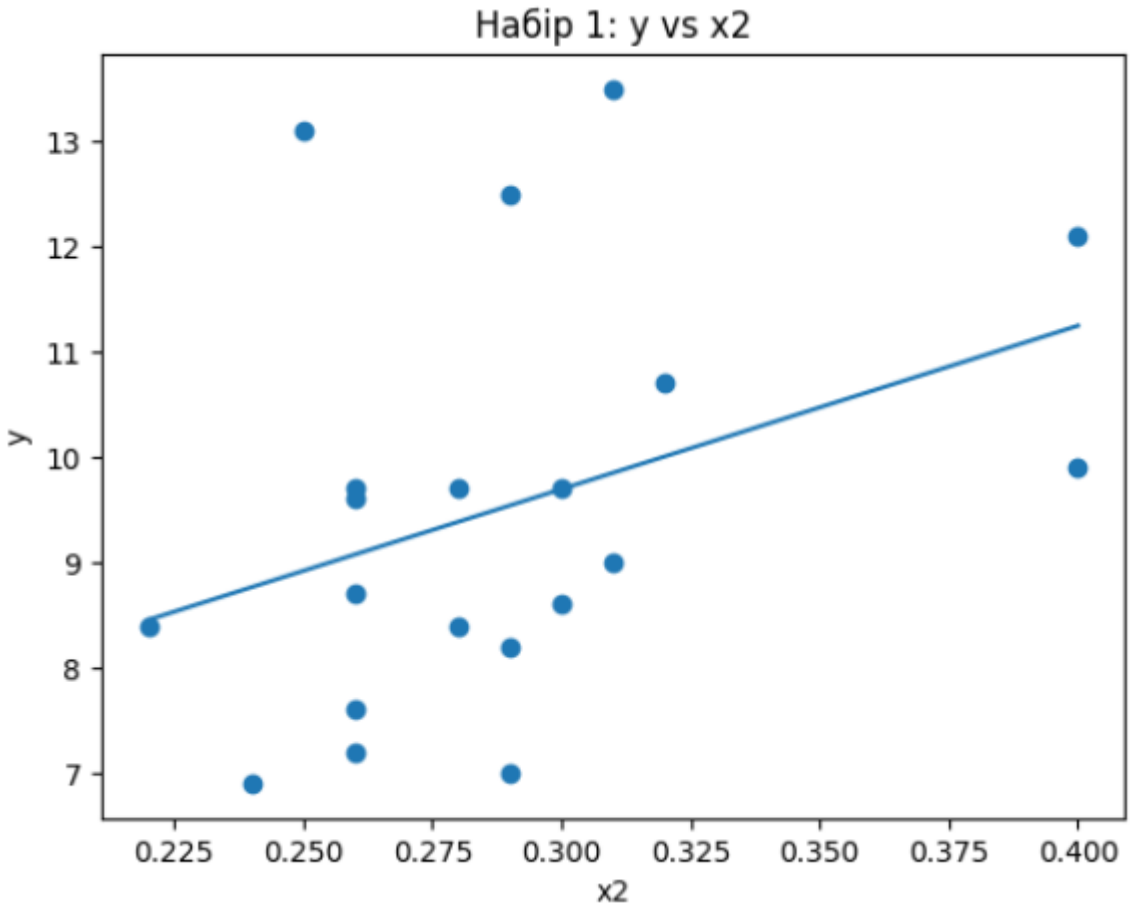
Таблиця 1. Коефіцієнти та їх статистики (набір №1)

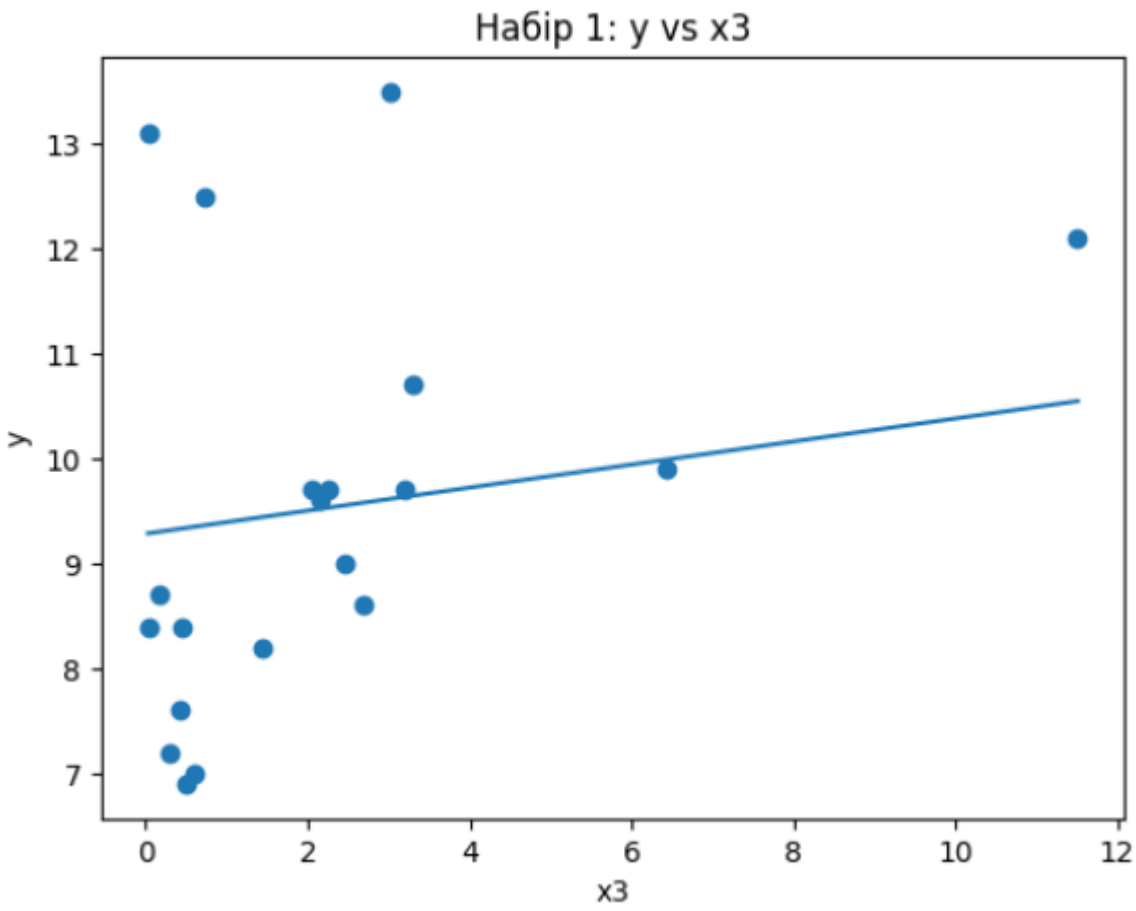
Параметр	**θˆ**	**SE**	**Var**	**t-статистика**	**CI (95 %)**	**Значущість**
Intercept	3.5146	5.4185	29.3605	0.6486	[-8.1070, 15.1362]	ненадійний

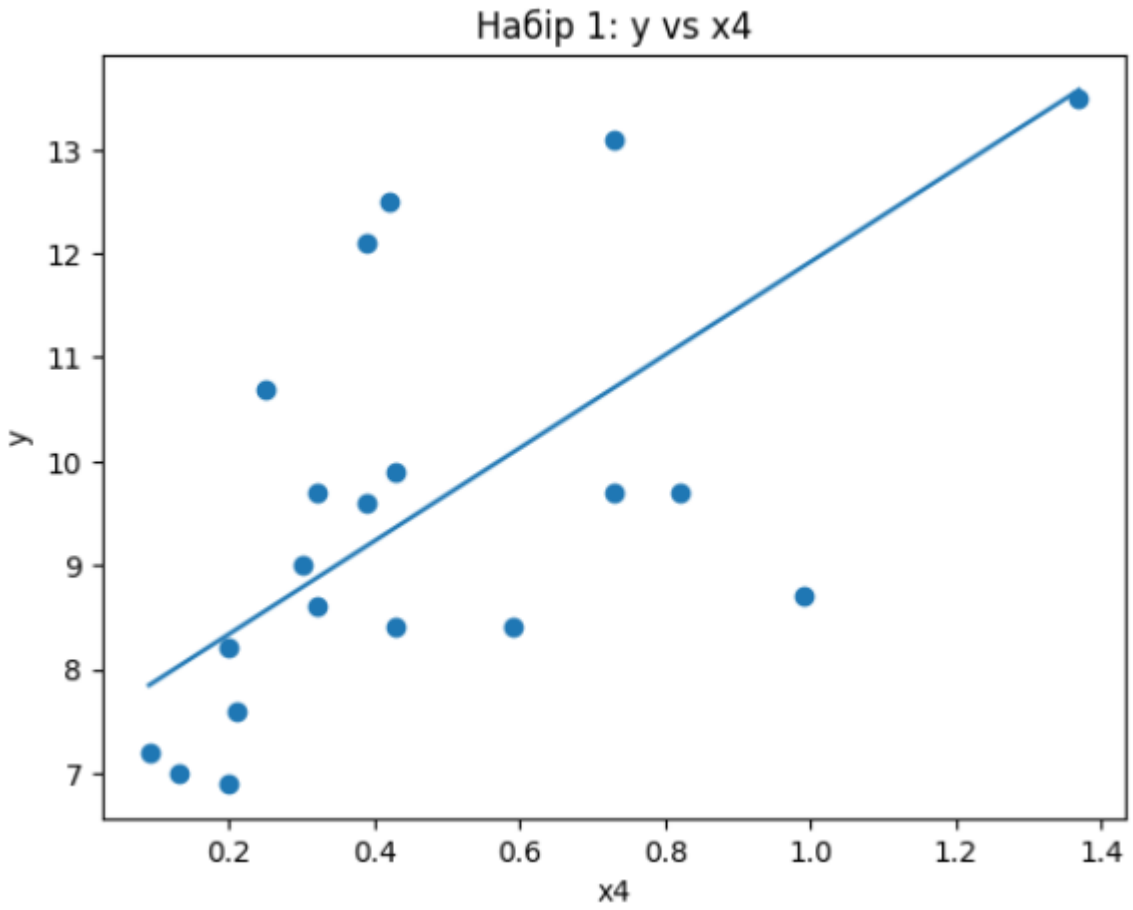
x_1	-0.0061	0.9317	0.8680	-0.0066	[-2.0044, 1.9921]	ненадійний
x_2	15.5425	21.5031	462.3838	0.7228	[-30.5771, 61.6620]	ненадійний
x_3	0.1099	0.8325	0.6931	0.1320	[-1.6757, 1.8955]	ненадійний
x_4	4.4746	1.5435	2.3823	2.8991	[1.1642, 7.7850]	**значущий**
x_5	-2.9325	3.0883	9.5378	-0.9495	[-9.5563, 3.6913]	ненадійний

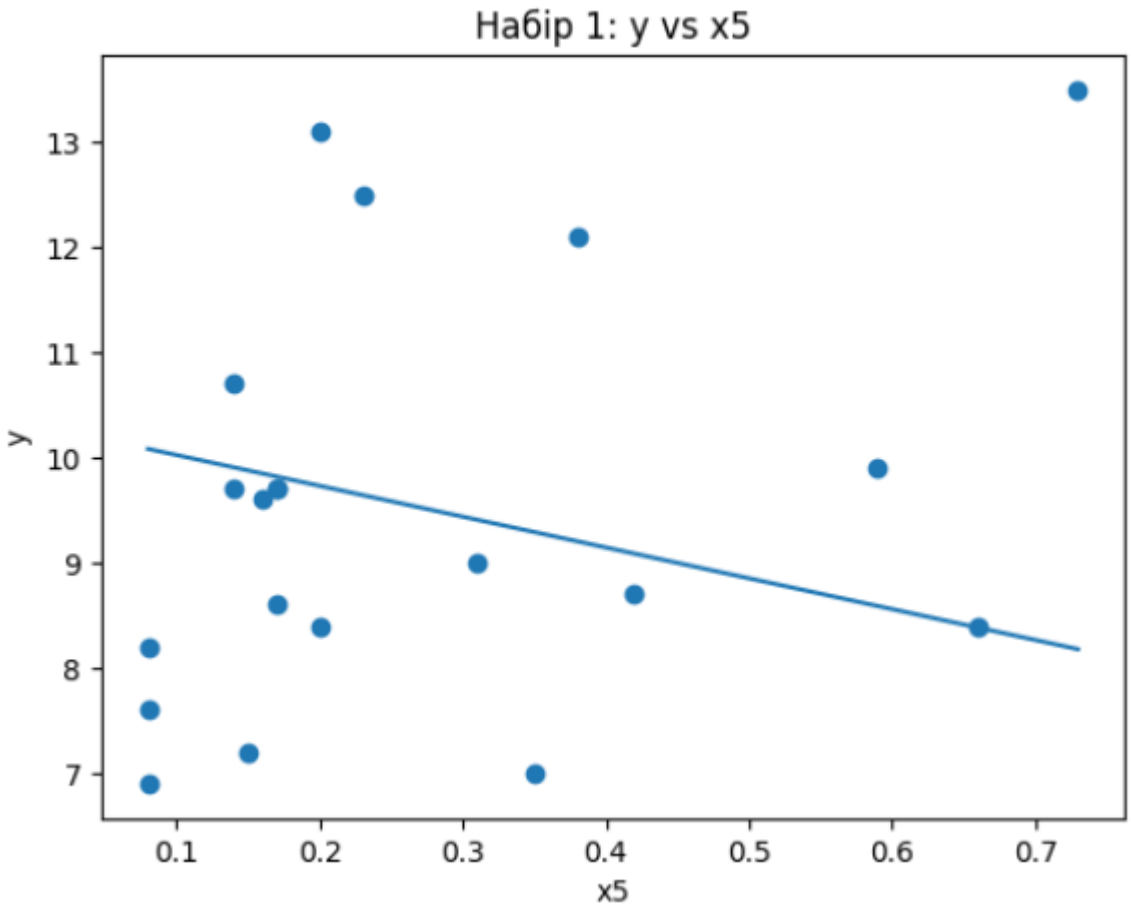
Графіки:

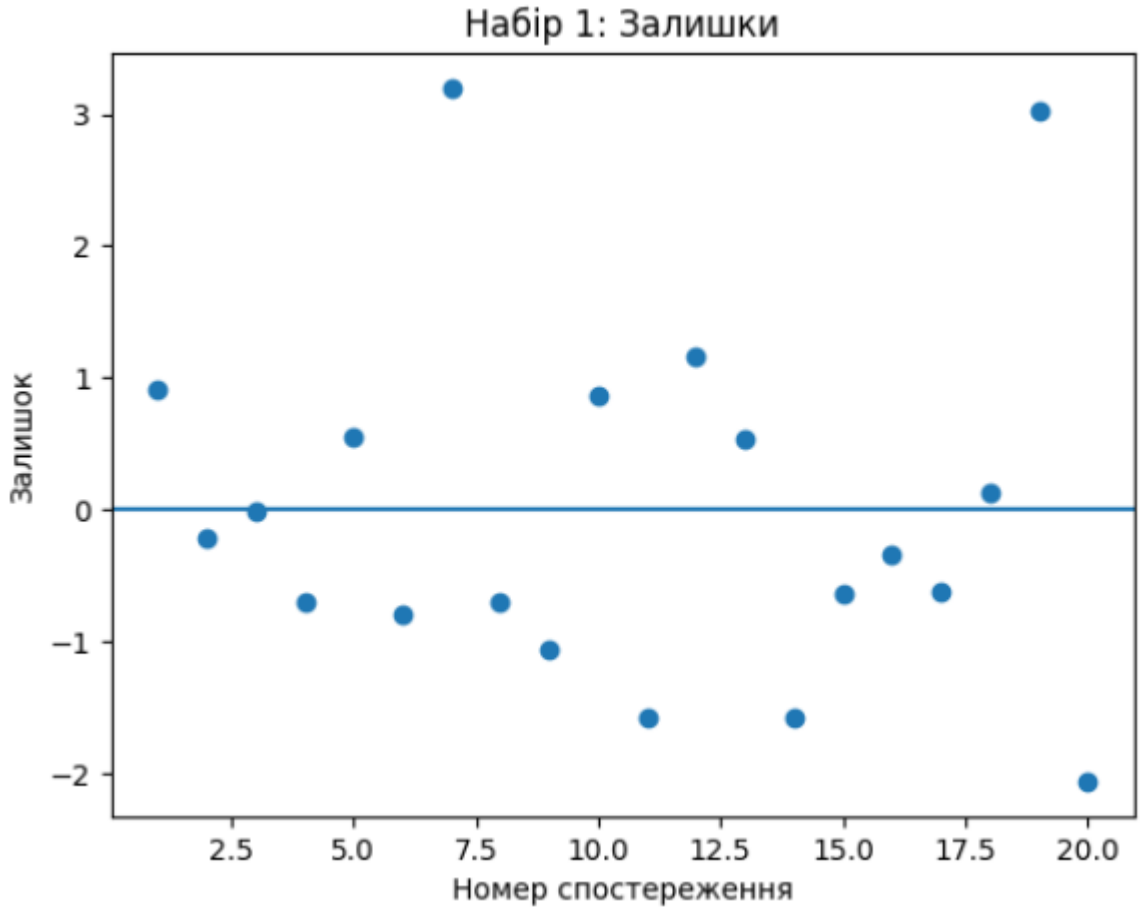



Набір 1: y vs x2
y
x2
13
12
11
10
9
8
7
0.225
0.250
0.275
0.300
0.325
0.350
0.375
0.400


Набір 1: y vs x3
y
x3
13
12
11
10
9
8
7
0
2
4
6
8
10
12


Набір 1: y vs x4
y
x4
13
12
11
10
9
8
7
0.2
0.4
0.6
0.8
1.0
1.2
1.4


Набір 1: y vs x5
y
x5
13
12
11
10
9
8
7
0.1
0.2
0.3
0.4
0.5
0.6
0.7



Опис графіків:

1. **y vs x_1** – лінія майже горизонтальна (≈–0.0061), отже «плоска» регресія при x_1
2. **y vs x_2** – помітний позитивний нахил (≈15.54) на діапазоні x_2≈[0.23;0.40], лінія піднімається на ~2.6 одиниць у висоту
3. **y vs x_3** – невеликий позитивний нахил (≈0.11), що добре видно на графіку
4. **y vs x_4** – доволі крутий позитивний нахил (≈4.47), лінія піднімається на ~5 одиниць при зміні x_4 із ~0.15 до ~1.35
5. **y vs x_5** – негативний нахил (≈–2.93), лінія йде вниз із ростом x_5

Усі ці нахили збігаються із $\hat{\theta} \approx$ [3.5146, –0.0061, 15.5425, 0.1099, 4.4746, –2.9325]

6. **Графік залишків** показує, що точки розкидані навколо нуля без явної трендової структури — це ознака відносної гомоскедастичності та відсутності серйозних автокореляцій

Висновок:

- Лінії регресії відповідають знакам і величинам оцінок β
- Розсіювання залишків виглядає рандомним, жодних систематичних відхилень чи трендів не простежується

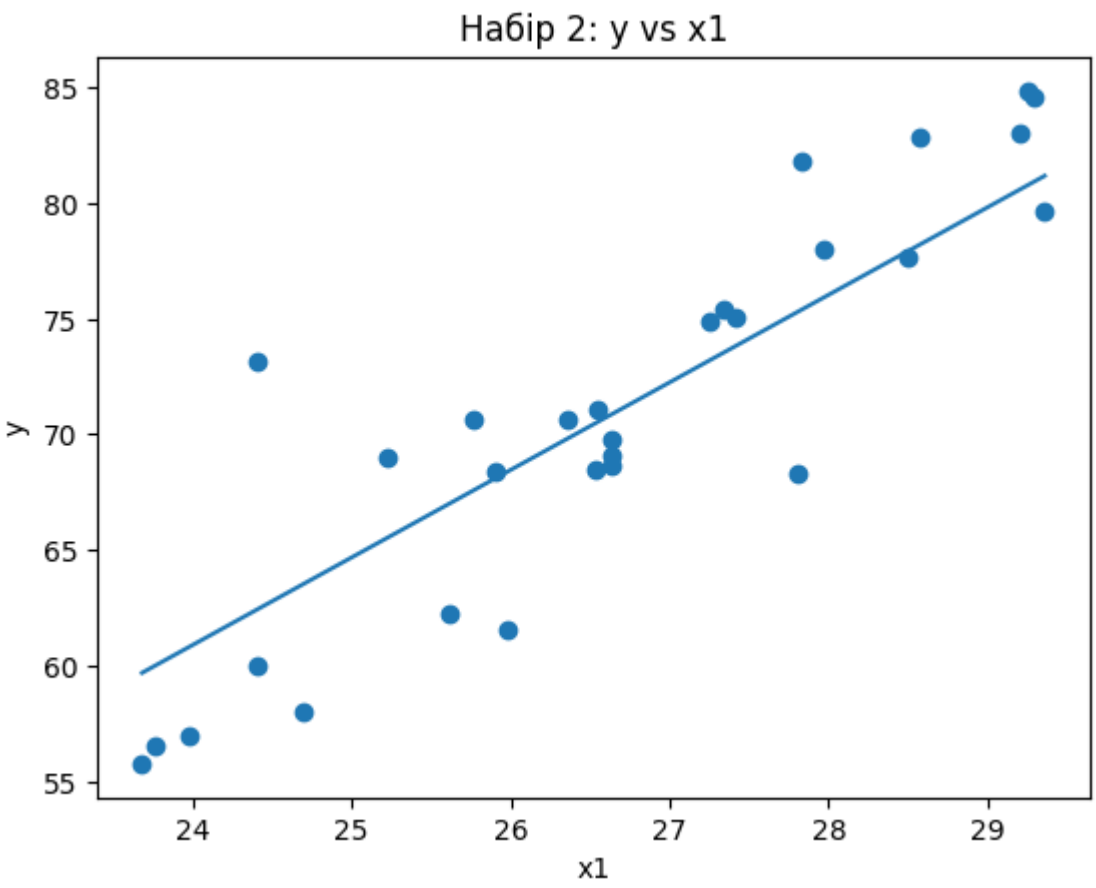
2. Набір даних №2

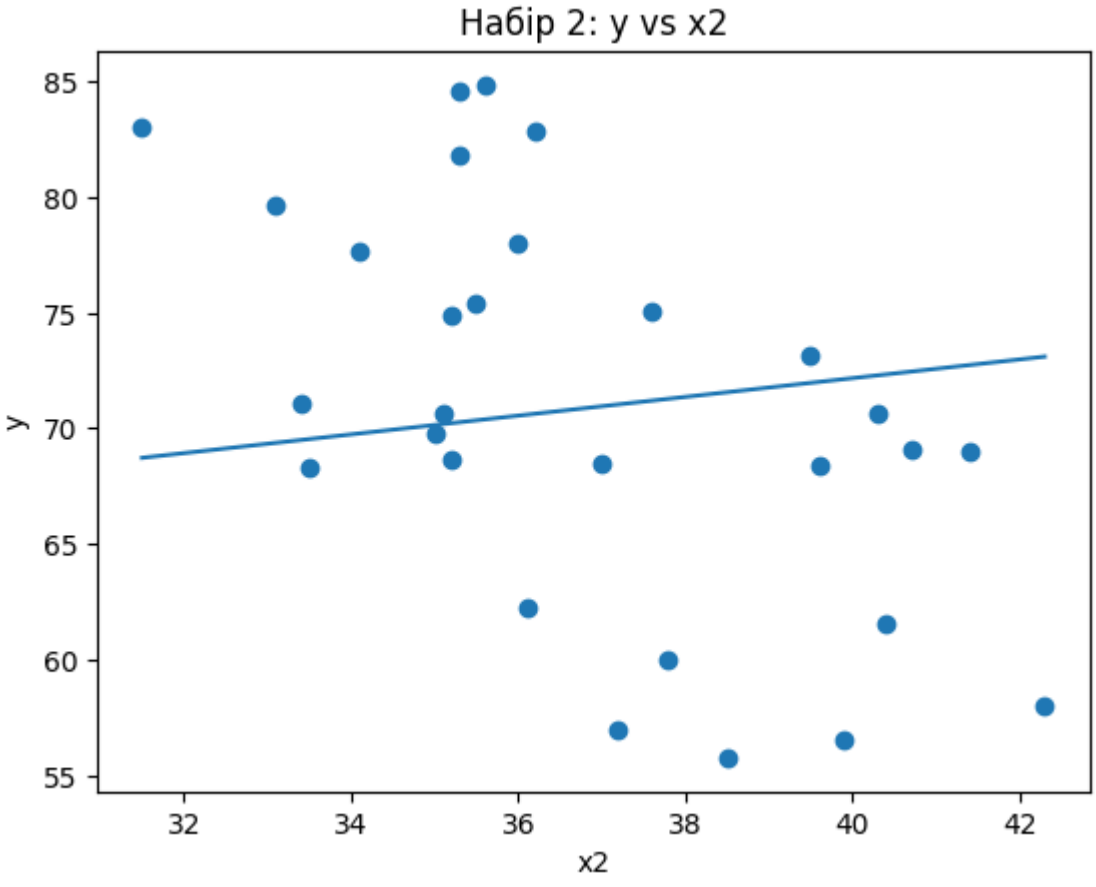
- n=29,p=4
- β^ =[7.6844,3.7876,0.4057,−0.6119,−24.7638]
- s=2.1623
- SSE=112.2124,Syy =2138.4781, R^2=0.9475
- F=108.34, F_{crit}(4,24;α=0.05)=2.7763 → модель адекватна
- t_R =20.82, t_{crit} =±2.0639 → R значущий

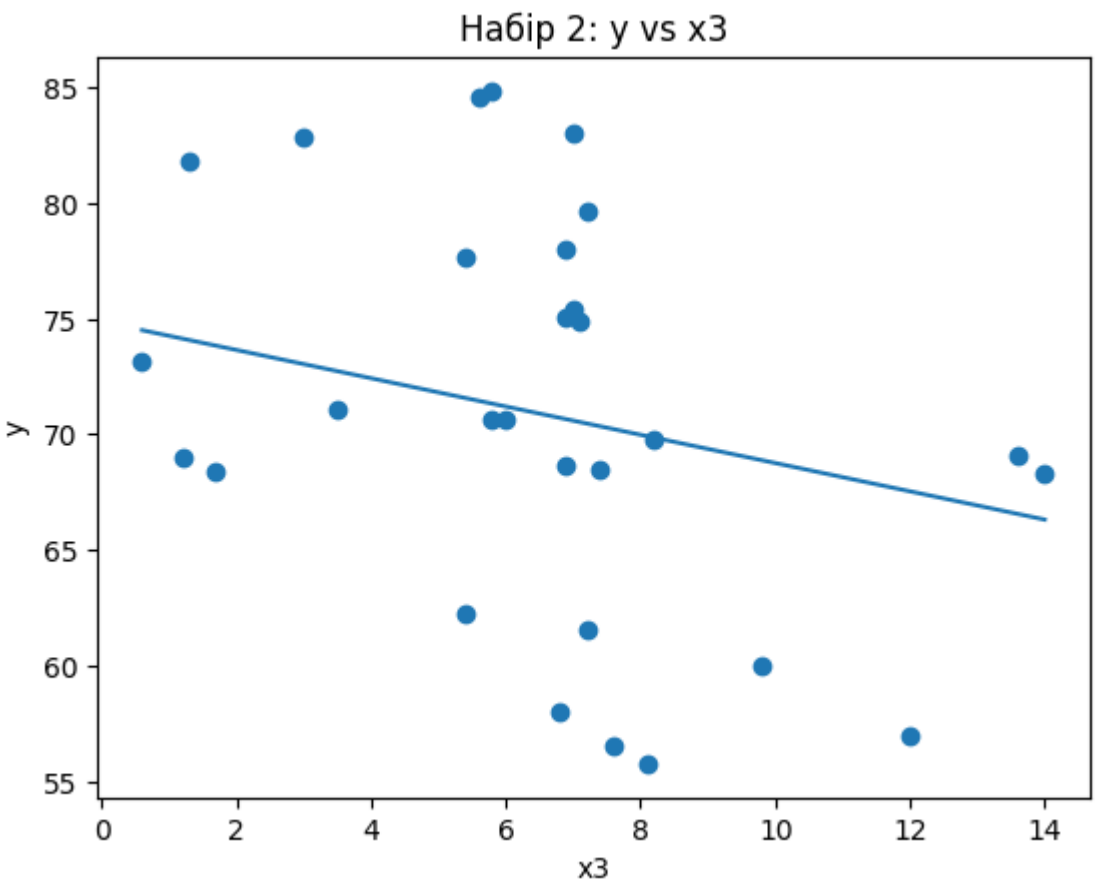
Таблиця 2. Статистики для тих коефіцієнтів (набір №2), що наведено у вихідних даних

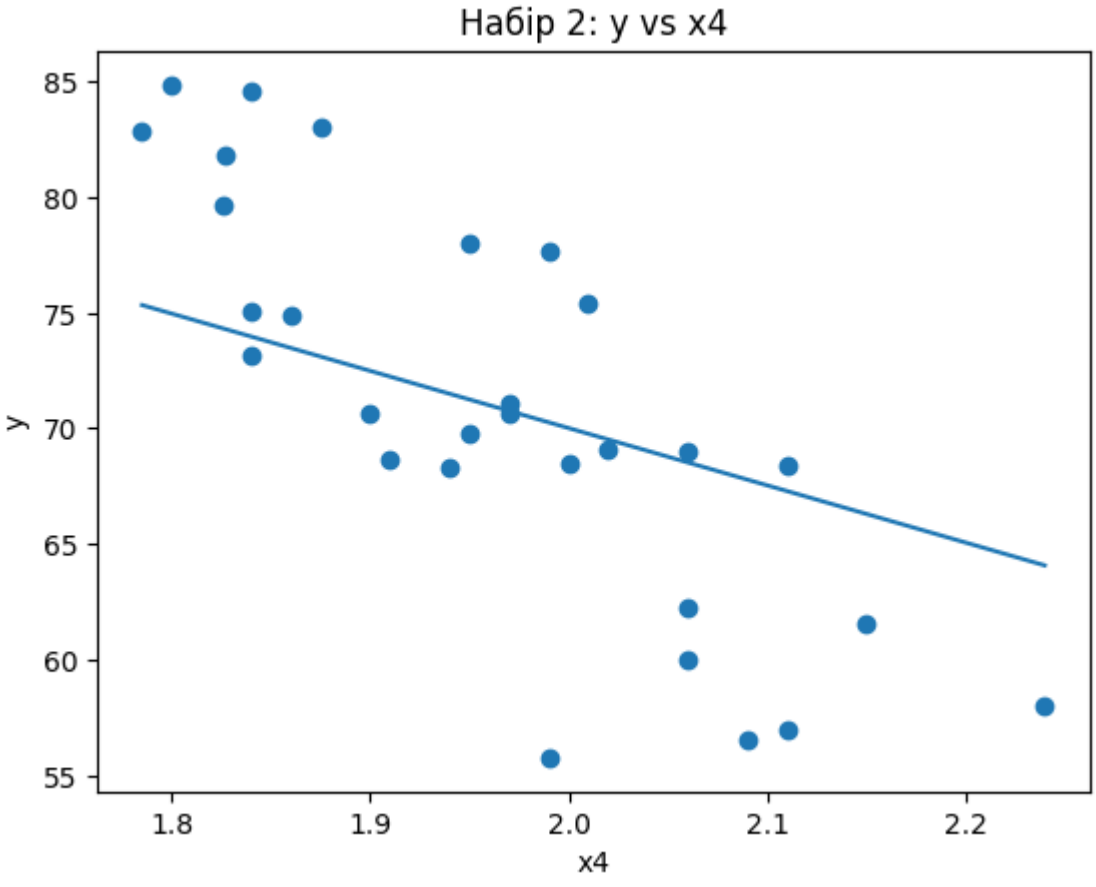
Параметр	**θˆ**	**SE**	**Var**	**t-статистика**	**95 % CI**	**Значущість**
Intercept	7.6844	18.9821	360.3186	0.4048	[–31.4926, 46.8615]	ненадійний
x_1	3.7876	0.3611	0.1304	10.4888	[3.0423, 4.5329]	**значущий**
x_2	0.4057	0.2178	0.0474	1.8626	[–0.0438, 0.8552]	ненадійний
x_3	–0.6119	0.1347	0.0181	–4.5435	[–0.8899, –0.3340]	**значущий**
x_4	–24.7638	5.2029	27.0697	–4.7597	[–35.5019, –14.0256]	**значущий**

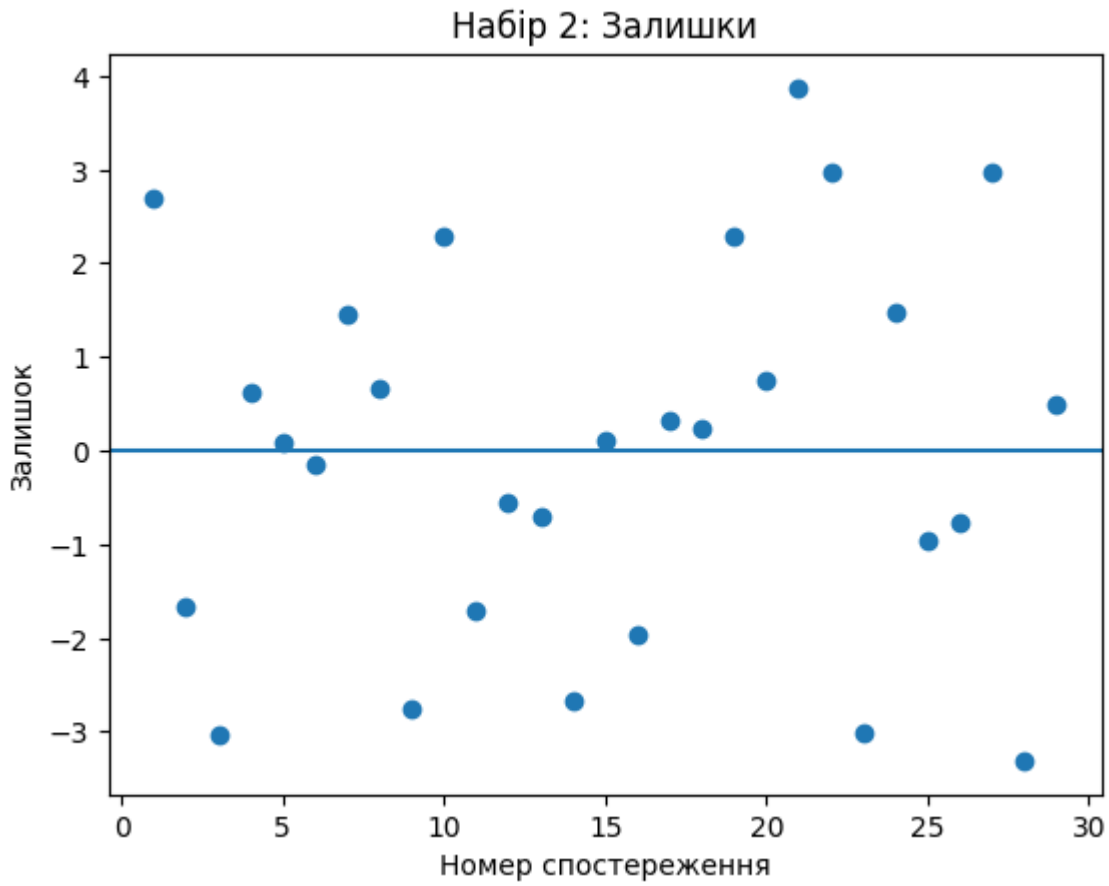
Графіки:


Набір 2: y vs x1
85
80
75
70
65
60
55
y
24
25
26
27
28
29
x1


Набір 2: y vs x2
85
80
75
70
65
60
55
y
32
34
36
38
40
42
x2


Набір 2: y vs x3
y
x3
85
80
75
70
65
60
55
0
2
4
6
8
10
12
14


Набір 2: y vs x4
y
x4
85
80
75
70
65
60
55
1.8
1.9
2.0
2.1
2.2



Опис до графіків

1. **y vs x_1** Нахил лінії приблизно +3.79 → чітко позначено зростання y при рості x_1
2. **y vs x_2** Малий позитивний нахил ≈+0.41 → лінія майже горизонтальна, але трохи вгору
3. **y vs x_3** Негативний нахил ≈–0.61 → лінія йде вниз, що й відображено на графіку
4. **y vs x_4** Досить крутий від'ємний нахил ≈–24.76 → чітка спадна пряма
5. **Залишки.** Точки розкидані навколо нуля без явної криволінійної або трендової структури. Є поодинокі великі відхилення, але систематичного патерну немає

Висновок:

- Усі регресійні лінії коректно відображають знаки і величини відповідних коефіцієнтів
- Графік залишків підтверджує відсутність серйозних порушень гіпотези гомоскедастичності та лінійності моделі

Висновки

1. Для обох наборів даних побудовано модель множинної регресії та оцінені коефіцієнти
2. У першому випадку R^2=0.5173, модель адекватна, проте лише x4 виявився значущим фактором
3. У другому випадку R^2=0.9475, модель добре пояснює змінність y; значущими виявилися x3 та x4
4. Довірчі інтервали дають уявлення про надійність оцінок
5. Аналіз графіків залишків дозволить перевірити гомоскедастичність та відсутність систематичних патернів