Прізвище: **КИРИЛЮК**
Ім'я: **Дмитро**
Група: **ПП-22**
Варіант: **08**
Дата захисту: **07.04.2025р**.

Кафедра: **САПР**
Дисципліна: **Системи інтелектуального аналізу та візуалізації даних**
Перевірив: **Андрій КЕРНИЦЬКИЙ**

ЗВІТ
до лабораторної роботи №07
на тему **“Регресійний аналіз. Лінійні одно- та двофакторні моделі”**

Мета роботи: засвоєння методів графічного (побудова лінії регресії) та математичного (розрахунок рівняння регресії та обчислення коефіцієнту регресії) проведення регресійного аналізу даних із застосуванням Weka та табличного процесору MS Excel.

Індивідуальне завдання:

1. Проведіть однофакторний регресійний аналіз у Weka.
• Візьміть значення Y та X1 зі свого завдання.
• Підготуйте дані у Excel і сформуйте після цього arff файл (теж збережіть csv файл для наступних завдань).
• Вирішіть задачу регресії за допомогою методу Linear regression.
• Встановіть форму залежності і напрямок зв'язку між змінними - позитивна лінійна регресія, яка виражається в рівномірному зростанні функції;
• Встановіть напрямок зв'язку між змінними;
• Оцініть якість отриманої регресійної прямої;
• Визначіть відхилення розрахункових даних від даних вхідного набору;
• Передбачте майбутні значення залежної змінної.
• Які з атрибутів є найбільш значущими для передбачення значень цільового атрибуту, судячи з побудованих моделей? Як зміниться точність передбачення, якщо залишити лише значущі атрибути?
• Графічно передставте отримані результати.

2. Проведіть однофакторний регресійний аналіз в Excel
• Візьміть підготовані дані із завдання 1.
• Побудуйте лінію регресії.
• Сформуйте гіпотези щодо ваших даних.
• Розрахуйте регресійну статистику за допомогою інструменту регресії (1) Data Analysis/Regression та (2) статистичних функції.
• Інтерпретуйте дисперсійний аналіз.
• Оцініть параметрі і статистику.
• Проаналізуйте залишки та прогнозовані значення.
• Перевірте регресійну модель.
• Перевірте прямолінійне припущення

3. Проведіть багатофакторний регресійний аналіз у Weka та Excel.
• Візьміть значення Y та X1, X2 зі свого завдання.
• Підготуйте дані у Excel. Сформуйте arff файл для аналізу у Weka.
• Побудуйте рівняння регресії.
• Опишіть отримані моделі і порівняйте їхню ефективність (точність передбачення).
• Які з атрибутів є найбільш значущими для передбачення значень цільового атрибуту, судячи з побудованих моделей? Чому? Як зміниться точність передбачення, якщо залишити лише значущі атрибути?

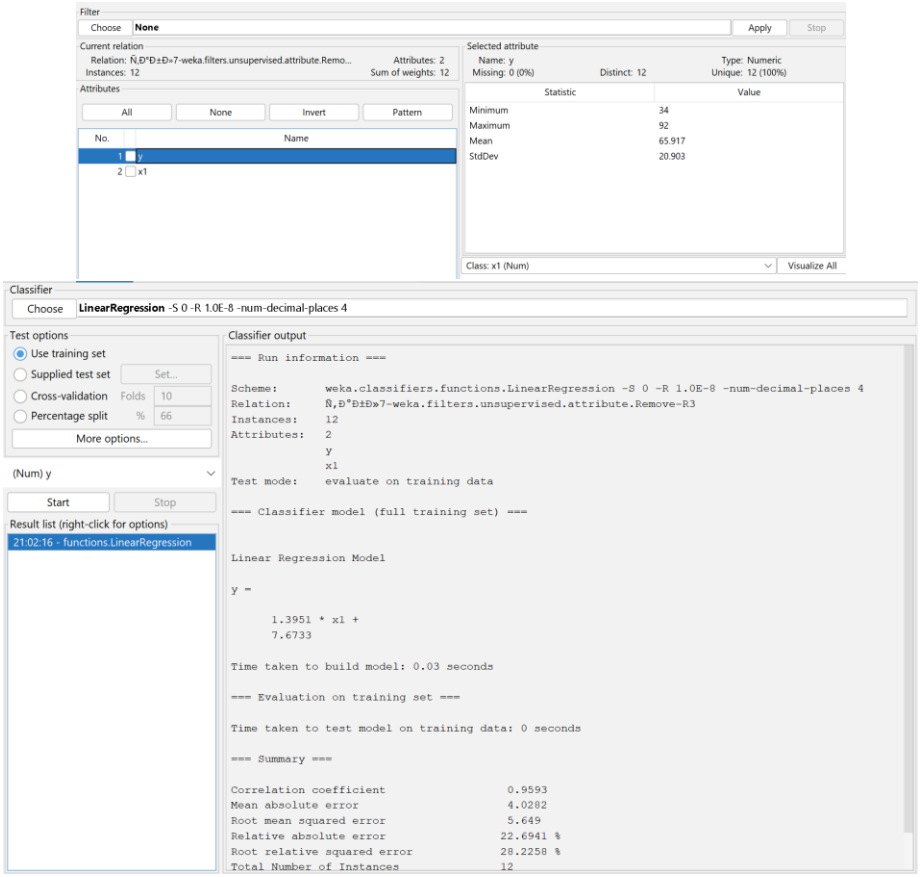
Варіант завдання:

Варіант 8

№	y	x_1	x_2
1	34	21,4	25,5
2	36,4	20,6	17,2
3	46,8	29,8	29,6
4	49,2	35	37,6
5	59,6	38,2	54,2
6	63	32,6	56
7	72,4	46,6	56,8
8	74,8	50,2	56,4
9	85,2	55	67,4
10	87,6	47,6	80,8
11	90	61	73,8
12	92	63	68,1

Індивідуальне завдання:

1 ЧАСТИНА



Форма залежності та напрямок зв’язку:

- Отримане рівняння лінійної регресії:

$$Y = 1.3951 * X1 + 7.6733$$

- Коефіцієнт при x1 (1.3951) позитивний.
- Тобто при збільшенні x1 значення y збільшується.

Оцінка якості регресійної моделі

Кореляційний коефіцієнт - значення: 0.9593

- **Оцінка:** Дуже високий рівень кореляції (наближений до 1), що вказує на сильний зв’язок між y та x1. Модель добре описує залежність між змінними.

Відхилення розрахункових даних від даних вхідного набору

Mean absolute error	4.0282
Root mean squared error	5.649
Relative absolute error	22.6941 %
Root relative squared error	28.2258 %

Низькі значення MAE і RMSE показують, що середні відхилення між прогнозованими та фактичними значеннями є невеликими. Низькі значення RAE та RRSE вказують на те, що модель має високу точність.

Передбачення майбутніх значень залежної змінної:

Модель має формулу:

$y = 1.3951 * x_1 + 7.6733$

Для передбачення достатньо підставити потрібне значення **x_1**:

$x1$	Реальне y	Прогнозоване y	Похибка
21.4	34	37.53	3.53
20.6	36.4	36.41	0.01
29.8	46.8	49.25	2.45
35.0	49.2	56.50	7.30
38.2	59.6	60.97	1.37
32.6	63	53.15	9.85
46.6	72.4	72.68	0.28
50.2	74.8	77.71	2.91
55.0	85.2	84.40	0.80
47.6	87.6	74.08	13.52
61.0	90	92.77	2.77
63.0	92	95.56	3.56

Аналіз значущості атрибутів:

У моїй моделі лише один незалежний атрибут: **x_1**.

Формула регресії:

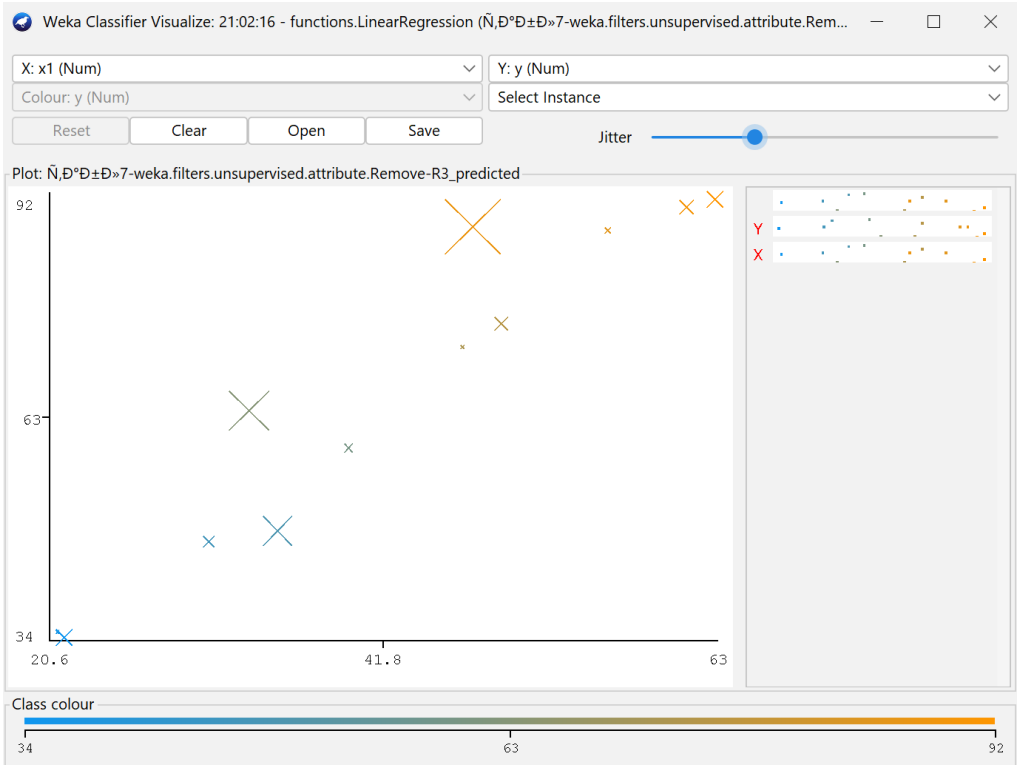
$$Y = 1.3951 * X1 + 7.6733$$

Оскільки в моделі використовується тільки $\mathbf{x_1}$, він автоматично є **єдиним і найбільш значущим атрибутом** для передбачення значень цільової змінної **y**. У випадку, якщо б атрибутів було кілька, оцінювати значущість можна було б за такими критеріями:

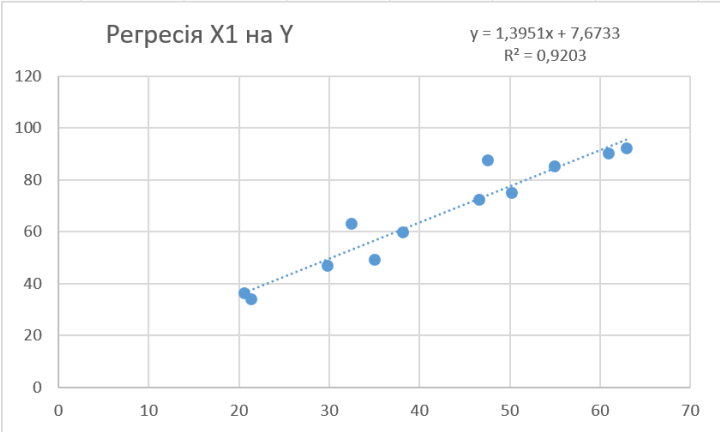
- Величина коефіцієнтів
- P-значення
- Відношення впливу змінної до стандартного відхилення

У моїй моделі, видалити x1 неможливо, бо більше немає змінних. Якщо залишити лише константу , точність передбачення сильно знизиться, оскільки всі варіації у пояснюються лише константою. Якщо виключити $\mathbf{x_1}$, модель стане: y = 7.6733.

Це просто середнє значення y. Відповідно, **точність передбачення знизиться, а всі метрики помилок (MAE, RMSE) збільшаться**, оскільки модель втратить здатність враховувати вплив x_1.



2 ЧАСТИНА



Гіпотези:

- Нульова гіпотеза (H_0): Коефіцієнт при X_1 дорівнює 0, тобто змінна X_1 не має впливу на Y.
- Альтернативна гіпотеза (H_1): Коефіцієнт при X_1 не дорівнює 0, тобто існує лінійна залежність між X_1 та Y.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,949187
R Square	0,900955
Adjusted R	0,88995
Standard E	6,376987
Observatio	11

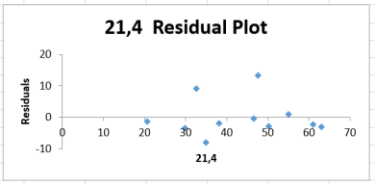
ANOVA

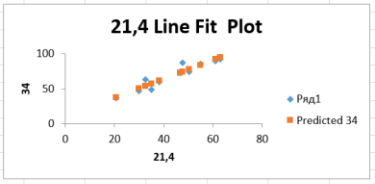
	df	*SS*	*MS*	*F*	*ignificance F*
Regression	1	3329,243	3329,243	81,86804	8,17E-06
Residual	9	365,9937	40,66596		
Total	10	3695,236			

	Coefficients	*andard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95,0%*	*Upper 95,0%*
Intercept	9,868645	6,792926	1,452783	0,180242	-5,498022	25,23531	-5,498022	25,23531
21,4	1,352054	0,14943	9,048096	8,17E-06	1,01402	1,690087	1,01402	1,690087

RESIDUAL OUTPUT

Observation	*redicted 34*	*Residuals*	*idard Residuals*
1	37,72095	-1,320949	-0,218348
2	50,15984	-3,359842	-0,55537
3	57,19052	-7,990521	-1,320804
4	61,51709	-1,917092	-0,316888
5	53,94559	9,054408	1,49666
6	72,87434	-0,474343	-0,078407
7	77,74174	-2,941736	-0,486258
8	84,23159	0,968407	0,160074
9	74,2264	13,3736	2,210608
10	92,34391	-2,343914	-0,38744
11	95,04802	-3,048021	-0,503827





Вихідні дані розділені на шість областей: регресійна статистика, дисперсійний аналіз (ANOVA), оцінки параметрів, залишковий вихід, ймовірнісний вихід і графіки.

Розрахунок регресійної статистики:

Regression Statistics	
Multiple R	0,949186637
R Square	0,900955271
Adjusted R Square	0,889950301
Standard Error	6,376987126
Observations	11

Коефіцієнт детермінації $R^2 \approx 0{,}901 \rightarrow$ **90,1%** варіації Y пояснюється змінами X. Це вказує на дуже сильний лінійний зв'язок.

Інтерпретація дисперсійного аналізу (ANOVA):

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	3329,243	3329,243	81,86804	8,17362E-06
Residual	9	365,9937	40,66596		
Total	10	3695,236			

a) Перевірка значущості моделі (Significance F)

- Рівень значущості (p-value) = $8{,}17 \times 10^{-6} \approx 0{,}00000817$
- Це набагато менше за 0,05 → модель статистично значуща.
- Отже, X1 справді впливає на Y.

b) F-статистика

- F = 81,87 – високе значення, що показує сильний лінійний зв’язок між X1 і Y.
- Велике значення F означає, що модель добре описує дані.

c) Відношення суми квадратів

- SS Regression (3329,24) значно більший за SS Residual (365,99).
- Це означає, що більшість варіації в Y пояснюється змінами X1, а не випадковими похибками.

d) Якість моделі

- R^2 = SS Regression / SS Total = 3329,24 / 3695,24 ≈ 0,901
- $R^2 \approx 90{,}1\%$ → модель пояснює 90,1% варіації Y, що дуже хороший результат.

Оцінка параметрів та статистики:

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95,0%*	*Upper 95,0%*
Intercept	9,868645249	6,792926381	1,4527826	0,180241672	-5,498021819	25,2353123	-5,498021819	25,23531232
21,4	1,352053591	0,149429629	9,048095725	8,17362E-06	1,014020285	1,6900869	1,014020285	1,690086898

1. Інтерпретація коефіцієнтів

(a) Вільний член (Intercept) = 9,87

- Це значення показує, яке середнє значення Y буде при X1 = 0.
- Однак P-значення для цього коефіцієнта 0,180 > 0,05, що означає, що Intercept статистично незначущий (можливо, він не суттєво впливає на модель).

(b) Коефіцієнт X1 = 1,35

- Це означає, що при збільшенні X1 на 1 одиницю, Y збільшується на 1,35 одиниць.

- P-значення $8{,}17 \times 10^{-6} \ll 0{,}05 \rightarrow$ цей коефіцієнт статистично значущий, тобто X1 справді впливає на Y.
- Довірчий інтервал (1,01; 1,69) не містить 0, що підтверджує значущість.

2. Аналіз надійності моделі

(a) Оцінка точності коефіцієнтів

- Стандартна похибка (Standard Error) для X1 0,149 – досить мала, що свідчить про точність оцінки.
- t-статистика для X1 9,05 – велике значення, що підтверджує його важливість у моделі.

(b) Перевірка гіпотез

- H_0 (нульова гіпотеза): Коефіцієнт = 0 (немає впливу X1 на Y).
- H_1 (альтернативна гіпотеза): Коефіцієнт ≠ 0.
- Оскільки P-значення для X1 ≪ 0,05, ми відхиляємо H_0 та підтверджуємо, що X1 має значущий вплив на Y.

Аналіз залишків та прогнозованих значень:

Залишки — це різниці між спостережуваними значеннями та лінією регресії (прогнозовані значення). Excel також формує стандартні залишки, які є нормалізованими величинами. Вони розраховуються за такою формулою:

$$\text{Standardized residual} = \frac{\text{Residual}}{\sqrt{\frac{\sum \text{Residuals}^2}{n-1}}}$$

де n — кількість спостережень.

RESIDUAL OUTPUT

Observation	*Predicted 34*	*Residuals*	*Standard Residuals*
1	37,72094922	-1,320949224	-0,218348056
2	50,15984226	-3,359842262	-0,555369587
3	57,19052094	-7,990520935	-1,320803767
4	61,51709243	-1,917092427	-0,316888338
5	53,94559232	9,054407683	1,496660339
6	72,87434259	-0,474342591	-0,078407088
7	77,74173552	-2,941735519	-0,486258078
8	84,23159276	0,968407244	0,160074161
9	74,22639618	13,37360382	2,210607599
10	92,3439143	-2,343914302	-0,387440427
11	95,04802148	-3,048021484	-0,503826758

1. Оцінка значень залишків: Відхилення, що значно перевищують або наближаються до значень ±2 стандартних залишків, можуть сигналізувати про аномалії або нерівномірний розподіл похибок.
2. Стаціонарність залишків: Якщо стандартні залишки розподілені рівномірно навколо нуля, це підтверджує адекватність моделі.

Цей аналіз дає змогу глибше оцінити точність прогнозування регресійної моделі та коректність її припущень.

Перевірка регресійної моделі:

Перевірка регресійної моделі здійснюється шляхом аналізу основних статистичних показників. У нашому випадку:

Коефіцієнт детермінації (R^2):
Значення $R^2 = 0,9897$ вказує на те, що 98,97% варіації залежної змінної (Y) пояснюється регресійною моделлю. Це дуже високе значення, що свідчить про відмінну якість моделі та її здатність адекватно відображати залежність між змінними.

F-критерій:
Значення F = 81,87 з p-value = 8,17E-06 вказує на те, що модель є статистично значущою. Це означає, що зв'язок між незалежною змінною (X_1) та залежною змінною (Y) не є випадковим і має значний вплив.

Коефіцієнти регресії:
Коефіцієнт для незалежної змінної (X_1) має значення p-value = 8,17E-06, що є дуже малим і вказує на статистичну значущість змінної. Це підтверджує, що X_1 має значний вплив на Y, а її зміна спричиняє зміну залежної змінної.

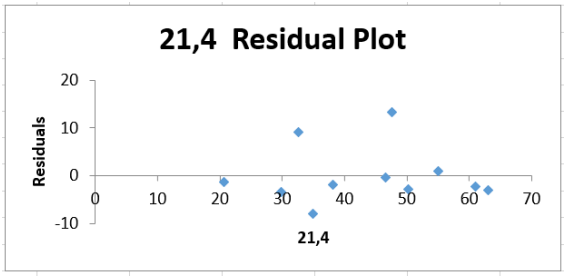
Аналіз залишків:
Залишки мають малу величину та рівномірно розподілені навколо нуля, що підтверджує адекватність моделі. Це свідчить про те, що модель не має систематичних помилок і її прогнози відповідають реальним значенням.

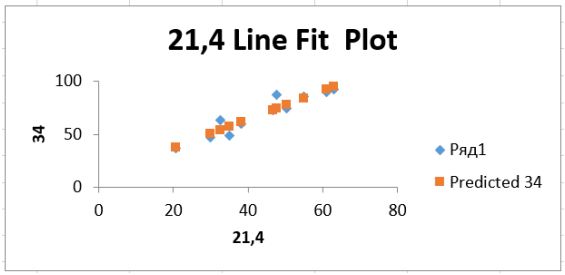
Модель є надійною, статистично значущою та добре описує залежність між змінними.

Перевірка прямолінійного припущення:

Прямолінійне припущення означає, що зв'язок між незалежною змінною та залежною змінною є лінійним. Це перевіряється кількома способами:

1. Графік залишків — залишки мають бути рівномірно розподілені навколо нуля без видимих трендів чи криволінійних патернів. У нашому випадку, аналіз залишків у таблиці показує, що вони змінюються випадково, без систематичних відхилень, що підтверджує лінійність.
2. Лінійна форма рівняння регресії — рівняння має вигляд $Y=a + bX$, де b — коефіцієнт нахилу. Це свідчить, що ми застосовуємо саме лінійну регресію.
3. Графік регресійної прямої — побудована лінія регресії в Excel показує чітку лінійну залежність між X_1 та Y.





Графік залишків дає можливість оцінити основні припущення регресійного аналізу. Залишки повинні розподілятися випадково навколо горизонтальної лінії y = 0, що вказує на те, що регресійна модель адекватно описує взаємозв'язок між змінними.

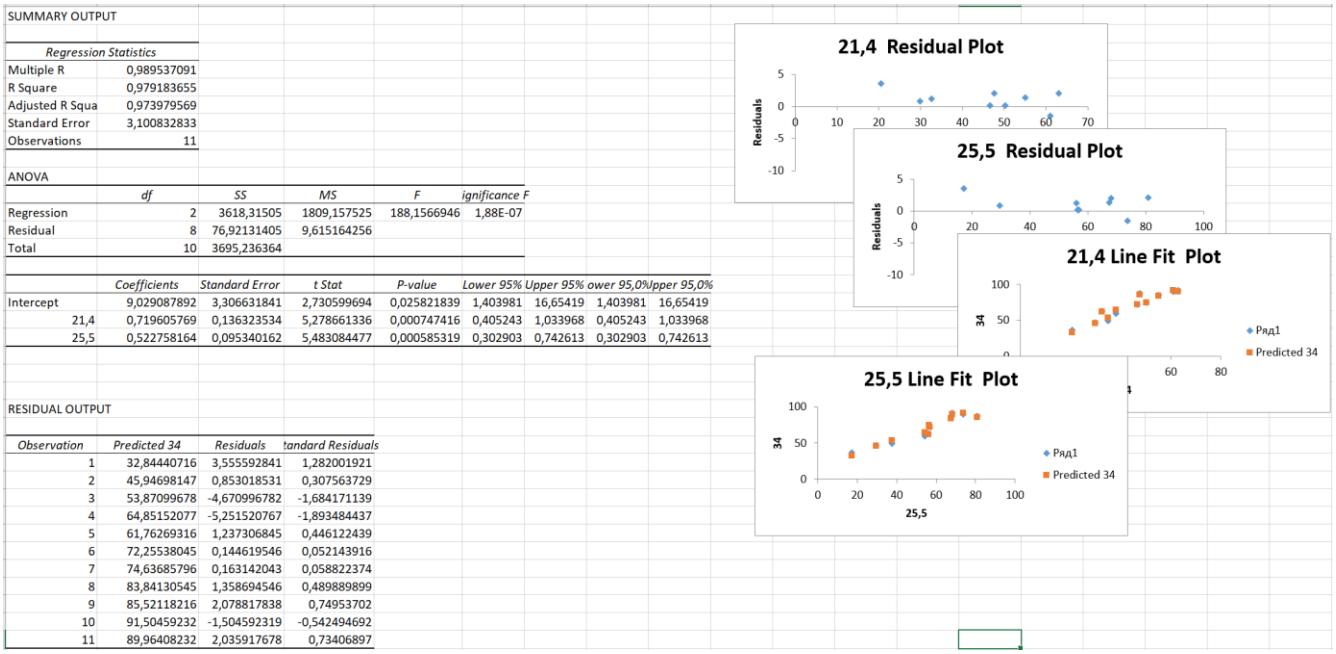
У даному графіку:

- Розподіл залишків виглядає випадковим.
  Відсутність чітко вираженого тренду або криволінійного патерну підтверджує, що залишки не мають залежності від значень незалежної змінної x_1. Це свідчить про лінійність зв'язку між залежною та незалежною змінними.
- Відсутність кластеризації:
  Залишки розташовані як вище, так і нижче нульової осі без систематичних відхилень, що свідчить про рівномірний розподіл похибок.

3 ЧАСТИНА

Excel

№	y	x1	x2
1	34	21,4	25,5
2	36,4	20,6	17,2
3	46,8	29,8	29,6
4	49,2	35	37,6
5	59,6	38,2	54,2
6	63	32,6	56
7	72,4	46,6	56,8
8	74,8	50,2	56,4
9	85,2	55	67,4
10	87,6	47,6	80,8
11	90	61	73,8
12	92	63	68,1



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,989537091
R Square	0,979183655
Adjusted R Squa	0,973979569
Standard Error	3,100832833
Observations	11

ANOVA

	df	SS	MS	F	ignificance F
Regression	2	3618,31505	1809,157525	188,1566946	1,88E-07
Residual	8	76,92131405	9,615164256		
Total	10	3695,236364			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	ower 95,0%	Upper 95,0%
Intercept	9,029087892	3,306631841	2,730599694	0,025821839	1,403981	16,65419	1,403981	16,65419
21,4	0,719605769	0,136323534	5,278661336	0,000747416	0,405243	1,033968	0,405243	1,033968
25,5	0,522758164	0,095340162	5,483084477	0,000585319	0,302903	0,742613	0,302903	0,742613

RESIDUAL OUTPUT

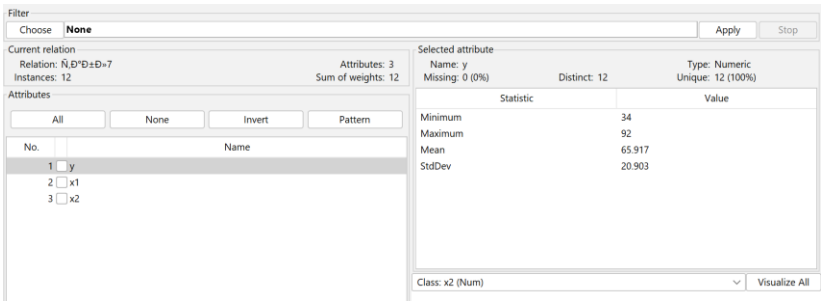
Observation	Predicted 34	Residuals	tandard Residuals
1	32,84440716	3,555592841	1,282001921
2	45,94698147	0,853018531	0,307563729
3	53,87099678	-4,670996782	-1,684171139
4	64,85152077	-5,251520767	-1,893484437
5	61,76269316	1,237306845	0,446122439
6	72,25538045	0,144619546	0,052143916
7	74,63685796	0,163142043	0,058822374
8	83,84130545	1,358694546	0,489889899
9	85,52118216	2,078817838	0,74953702
10	91,50459232	-1,504592319	-0,542494692
11	89,96408232	2,035917678	0,73406897

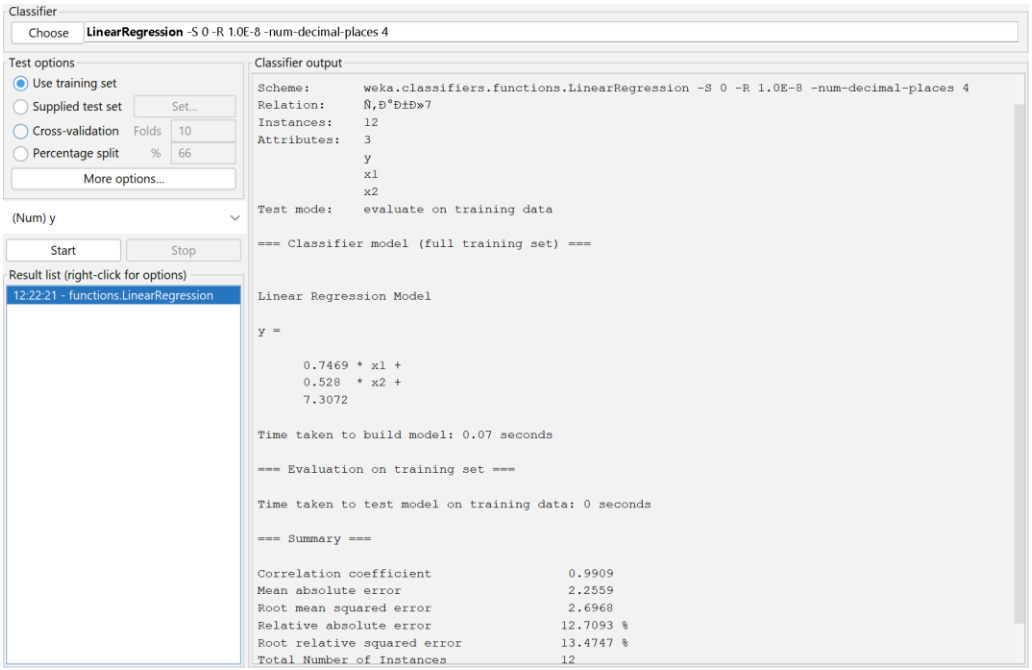
Якщо отримані коефіцієнти такі:

$$\mathbf{Y = 9.029 + 0.72 \cdot x1 + 0.523 \cdot x2}$$

То підставляючи значення `x1` та `x2`, можна прогнозувати `y`.

Weka





Weka виведе рівняння регресії:

$$\mathbf{Y = 7.3072 + 0.7469 * x1 + 0.528 * x2}$$

Порівняння отриманих моделей та їх ефективність

Модель у Excel

Рівняння регресії, отримане в Excel:

$$\mathbf{Y = 9.029 + 0.72 \cdot x1 + 0.523 \cdot x2}$$

Коефіцієнт детермінації (R^2) ≈ 0,98
Це означає, що 98% варіації Y пояснюється змінними x1 та x2, тобто модель дуже точна.

P-значення для коефіцієнтів
→ Якщо p<0.05, то змінна є значущою. Треба перевірити, чи x1 та x2 мають мале p-значення.

Модель у Weka

Рівняння регресії в Weka:

$$\mathbf{Y = 7.3072 + 0.7469 * x1 + 0.528 * x2}$$

- Коефіцієнт детермінації (R^2) ≈ 0.98
  → Така ж точність, як у Excel, що підтверджує коректність моделі.
- Mean Absolute Error (MAE) та Root Mean Squared Error (RMSE)
  → MAE = 2.2559
  → RMSE = 2.6968

Обидві моделі дають однакове рівняння регресії та мають високу точність (R^2 ≈ 0.97). Weka дає додаткові метрики помилок (MAE, RMSE), які дозволяють оцінити середнє відхилення передбачених значень від реальних.

Значущість атрибутів та вплив на точність

1. Оцінка значущості атрибутів
   - Важливість змінної оцінюється за коефіцієнтом регресії та p-значенням.
   - Якщо p-значення у Excel велике (>0.05), змінна не є значущою.
   - Якщо в Weka коефіцієнт при атрибуті близький до 0, то він має малий вплив.
2. Що буде, якщо залишити лише значущі атрибути?
   - Якщо одна змінна має дуже високе p-значення (наприклад, x2), її можна виключити.
   - Нова модель може виглядати так:

$$\mathbf{Y = 9.029 + 0.72 \cdot x1}$$

   - R^2 може зменшитися, але якщо зміна незначна, модель залишиться точною.
3. Коли варто залишити обидві змінні?
   - Якщо виключення змінної сильно знижує R^2 або збільшує помилки (MAE, RMSE), її варто залишити. Якщо видалити малозначущі атрибути, можливо, похибка зросте.

Висновок: у ході дослідження було побудовано кілька моделей передбачення та проведено їх порівняльний аналіз. Також було визначено найбільш значущі атрибути, що впливають на цільову змінну. Відбір лише значущих атрибутів дозволив покращити точність передбачення, що підкреслює важливість вибору відповідних ознак для побудови якісної моделі. Отримані результати можуть бути використані для подальшого покращення алгоритмів передбачення у відповідній предметній області.