МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
«ХАРКІВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»
«Кафедра програмної інженерії та інформаційних технологій управління»

Звіт з лабораторних робіт
з дисципліни «Математична статистика»

Виконав:
ст. гр. КН-221в
Шулюпов Є.Р.

Перевірив:
проф. каф. ПІТУ
Козуля Т.В.

Харків 2022

ЛАБОРАТОРНА РОБОТА №4

1. Тема

Регресійний аналіз. Моделі регресійного аналізу. Побудова лінійної регресійної моделі методом найменших квадратів.

2. Мета

Для даного статистичного спостереження надати характеристику даних на основі регресійного аналізу, моделі регресійного аналізу та лінійної регресійної моделі методом найменших квадратів в MS Excel.

3. Постановка завдання

Варіант 7	x	7,0	7,2	7,4	7,6	7,8	8,0	8,2	8,4
	y^*	0,45	0,67	0,84	0,97	0,87	0,99	1,06	1,23

Задана вибірка. Варіант 7

1) Побудувати діаграму розсіювання значень ознак X і Y.
2) Знайти параметри парної лінійної регресійної моделі $y = ax + c$.
3) Перевірити значимість побудованої моделі.
4) Обчислити коефіцієнт детермінації побудованої моделі.
5) Використовуючи інструмент Лінія тренда Excel, випробувати побудова лінійної та інших варіантів регресійних моделей: експоненційної, статечної, логарифмічною, поліноміальних 2–4 ступенів. Результат для кожної моделі представити графічно: діаграма розсіювання, лінія регресії, її рівняння, коефіцієнт детермінації. Вибрати регресію.

Хід виконання роботи

1. Діаграма розсіювання значень ознак X і Y (рис.1).

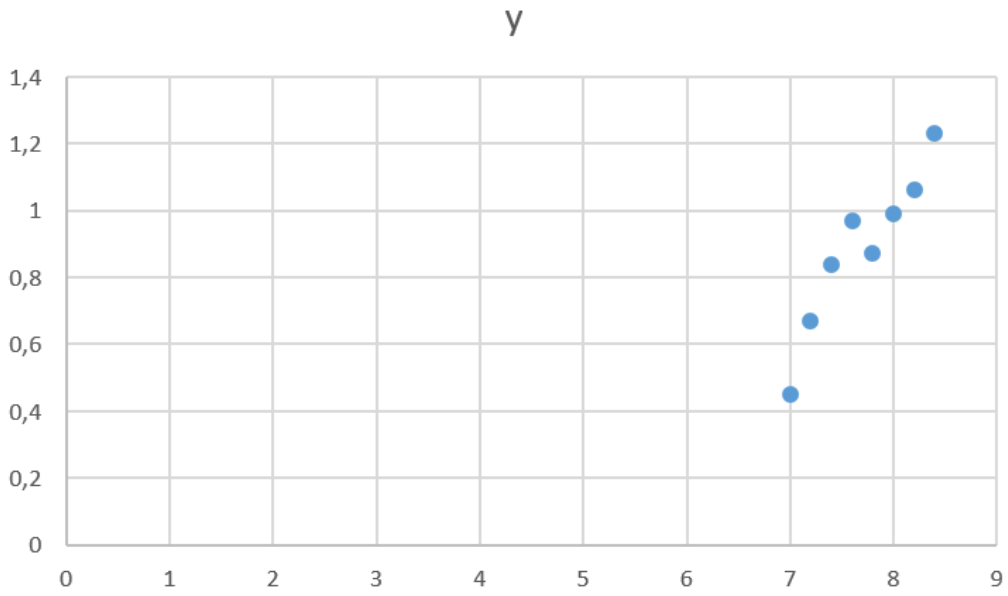

Рисунок 1 «Діаграма розсіювання»

На отриманій діаграмі синім кольором помічено координати точок На діаграмі кожній абсцисі X відповідає значення ординат У, які дані за початковою умовою. Ми можемо стверджувати, що зв’язок між випадковими величинами присутній, бо кореляційний момент наявний, бо міра залежності двох випадкових величин не дорівнює нулю. Обидва значення збільшуються, кореляційний момент є суттєвим – значення вибірки мають незначний розкид, що не надто сильно порушує лінійну структуру.

2. Параметри парної лінійної регресійної моделі y = ax + c (рис.2).

x	y
7	0,45
7,2	0,67
7,4	0,84
7,6	0,97
7,8	0,87
8	0,99
8,2	1,06
8,4	1,23

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественны	0,940065484
R-квадрат	0,883723114
Нормированны	0,864343634
Стандартная ош	0,088658431
Наблюдения	8

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	1	0,358438095	0,358438095	45,60096931	0,00051433
Остаток	6	0,047161905	0,007860317		
Итого	7	0,4056			

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
Y-пересечение	-2,671666667	0,527623202	-5,063588288	0,002302508	-3,962714133	-1,380619201	-3,962714133	-1,380619201
Переменная X 1	0,461904762	0,068401465	6,752848977	0,00051433	0,294532407	0,629277117	0,294532407	0,629277117

ВЫВОД ОСТАТКА

Наблюдение	*Предсказанное Y*	*Остатки*
1	0,561666667	-0,111666667
2	0,654047619	0,015952381
3	0,746428571	0,093571429
4	0,838809524	0,131190476
5	0,931190476	-0,061190476
6	1,023571429	-0,033571429
7	1,115952381	-0,055952381
8	1,208333333	0,021666667

Рисунок 2 «Параметри регресійної моделі»

Регресивна статистика:

R^2 = 0,883 (88,3%), що означає, що розрахункові параметри моделі (рис. 1) на 88,3% поясняють залежність між досліджуваними параметрами. Як відомо, що чим ближче коефіцієнт детермінації до значення 1, тим якісніше модель, то, відповідно, наше значення є достатньо вагомим, аби можна було назвати модель якісною.

У таблиці з коефіцієнтами моделі наведено оцінки β0 (Y-перетин), що має значення -2,67, та β1 (Змінна X1), що має значення 0,46.

Коефіцієнт β0 показує, які значення буде приймати Y, якщо всі змінні в даній моделі будуть рівні 0, тобто на значення аналізованого параметра майже впливають інші фактори, які не описані в моделі. Коефіцієнт β1 показує вагомість впливу змінної X на Y.

Модель має вигляд = 0,4619x - 2,6717, вона є значимою, оскільки значимість p = 0,0005 менше за 0,05, а коефіцієнт детермінації R^2 = 0,883.

Нанесемо лінію регресії на діаграму розсіювання та отримаємо рівняння регресії y = 0,4619x - 2,6717. (рис.3).

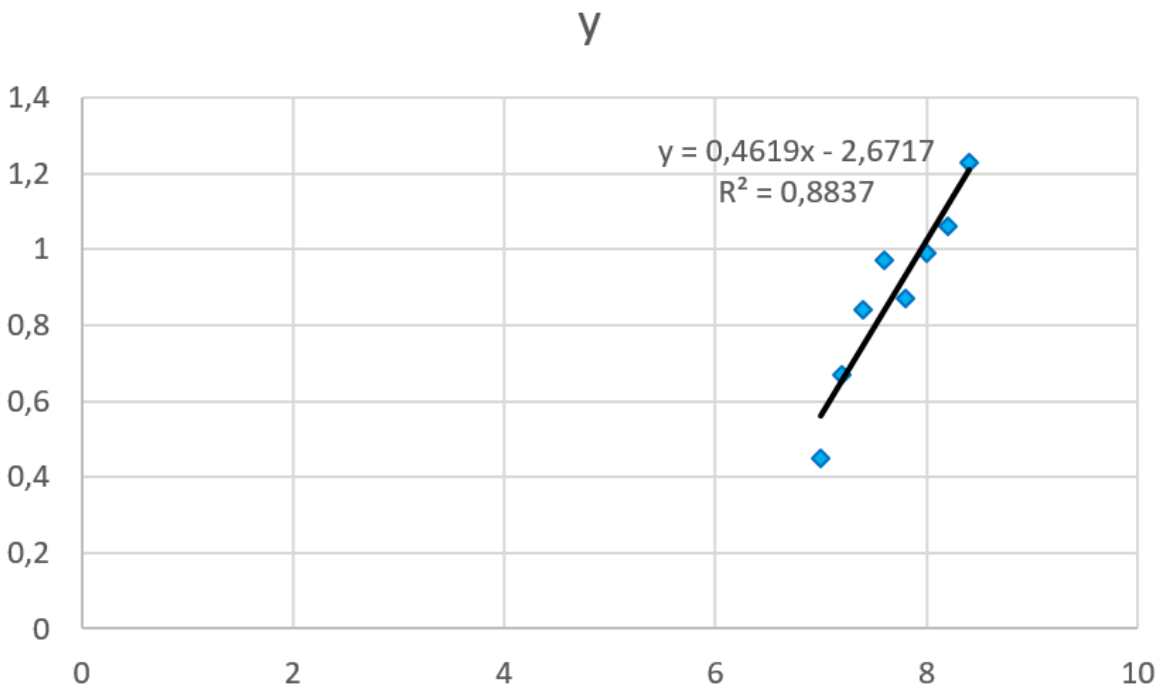

Рисунок 3 «Лінія регресії»

Отримали лінійне рівняння виду y = 0,4619x - 2,6717, а коефіцієнт детермінації R^2 = 0,8837, що співпадає з даними, отриманими попередніми способом. Коефіцієнт детермінації дорівнює 0,8837 ближче до одиниці, ніж до нуля, тобто показує високу адекватність моделі лінійної до реальної залежності між параметрами вимірювань двох факторів.

3. Перевірено значимість побудованої моделі (рис.4).

Дисперсионный анализ					
	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	1	0,358438095	0,358438095	45,60096931	0,00051433
Остаток	6	0,047161905	0,007860317		
Итого	7	0,4056			

Рисунок 4

«Дисперсійний аналіз»

Нам надається таблиця дисперсійного аналізу, де розрахована статистика Фішера, нам наведено значимість по критерію Фішера - 0,0005.

$F_{табл} = 5,99$;

$F_{розрах} = 45,6$;

$F_{табл} < F_{розрах}$, як наслідок, моя регресія адекватна та відповідає вихідним даним.

- Відповідно модель є значимою, адже ймовірність помилки менше за 0,05, тобто *нульова гіпотеза* про відсутність зв'язку між випадковими величинами відкидається. Для того, щоб з'ясувати чи лінійна регресія найбільш адекватно відображає залежність між досліджуваними параметри, побудуємо інші регресійні моделі.

4. Обчислено коефіцієнт детермінації побудованої моделі(рис.5).

Регрессионная статистика	
Множественны	0,940065484
R-квадрат	0,883723114
Нормированны	0,864343634
Стандартная ош	0,088658431
Наблюдения	8

Рисунок 5

«Регресійна статистика»

У таблиці Регресійна статистика наведені, коефіцієнт детермінації R^2 = 0,8837. Стандартна помилка сягає 0,088 або 8,8% - цей рівень приблизного стандартного відхилення статистичної вибіркової сукупності гарантує достовірність моделі.

5. Для покращення лінійної залежності між вибірковими елементами, побудовано експоненційну регресійну модель, використовуючи інструмент Лінія тренду (рис.6).

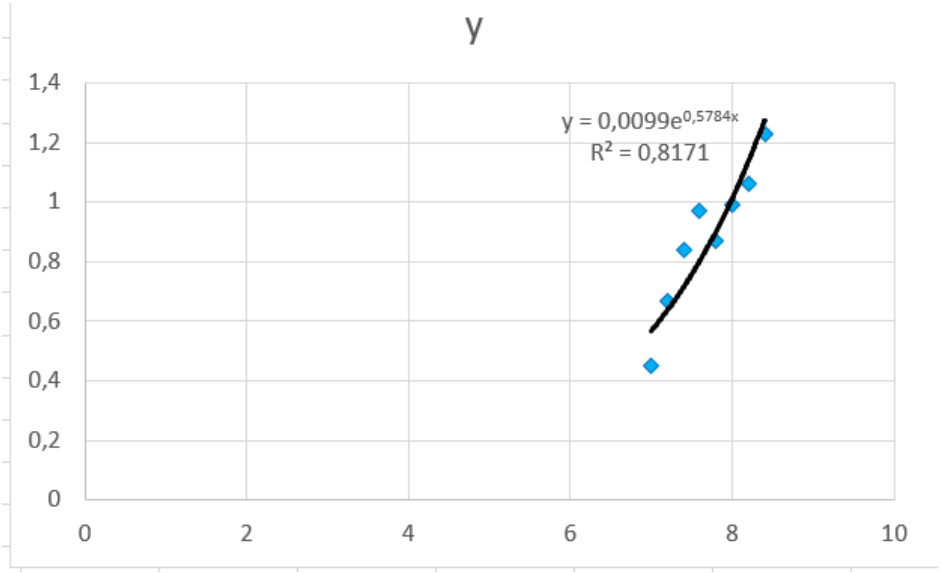

Рисунок 6
«Експоненційна
регресійна модель»

На відміну від лінійної регресійна модель виду $y = 0{,}0099e^{0{,}5784x}$ має лінію тренду більш округлу, а коефіцієнт детермінації R^2 = 0,8171 , що на відміну від моделі лінійної має більш слабку адекватність до реальної залежності між параметрами вимірювань двох факторів.

Для покращення лінійної залежності між вибірковими елементами, побудовано логарифмічну регресійну модель, використовуючи інструмент Лінія тренду (рис.7).

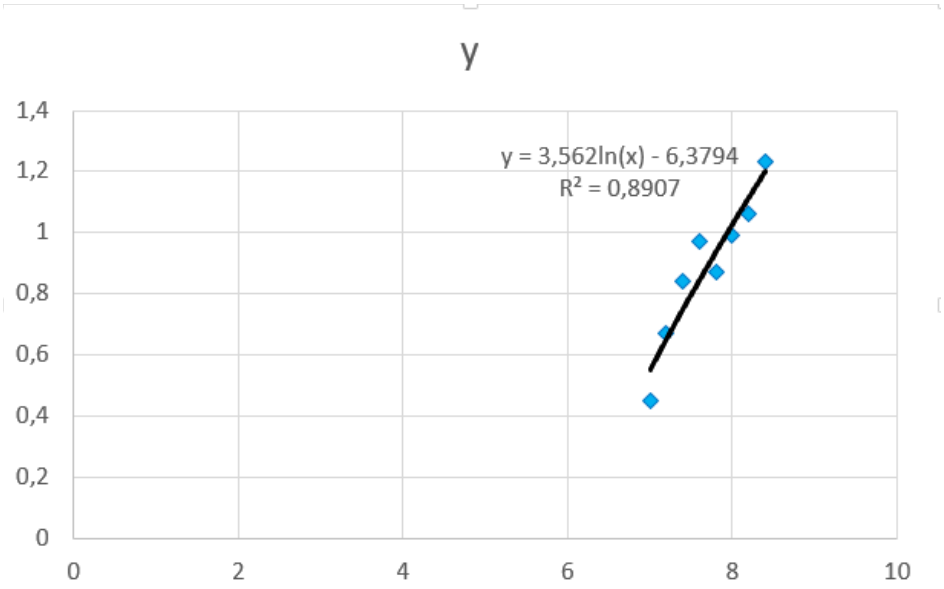

Рисунок 7
«Логарифмічна
регресійна модель»

На відміну від попередніх ліній логарифмічна виду y = 3,562ln(x) - 6,3794 має коефіцієнт детермінації R^2 = 0,8907 , що із порівнянням з попередніми моделями має більшу адекватність моделі лінійної до реальної залежності між параметрами вимірювань двох факторів, але для остаточного висновку треба перевірити ще низку заявлених моделей, щоб переконатися у судженні, або зловити більш уподібнену до точок лінію.

Отже для покращення лінійної залежності між вибірковими елементами, побудуємо ще поліноміальні регресійні моделі 2-4 ступенів, використовуючи інструмент Лінія тренду (рис.8 - 10).

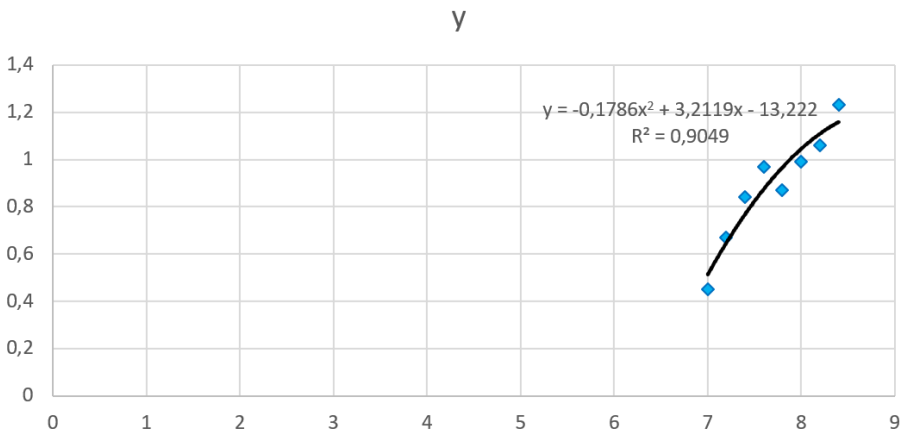

Рисунок 8 «Поліноміальна регресійна модель 2 ст»

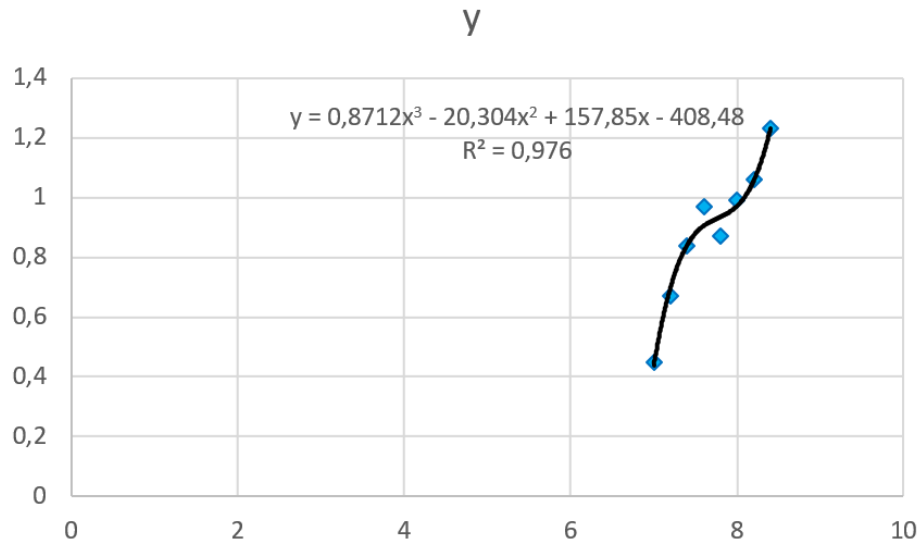

Рисунок 9 «Поліноміальна регресійна модель 3 ст»

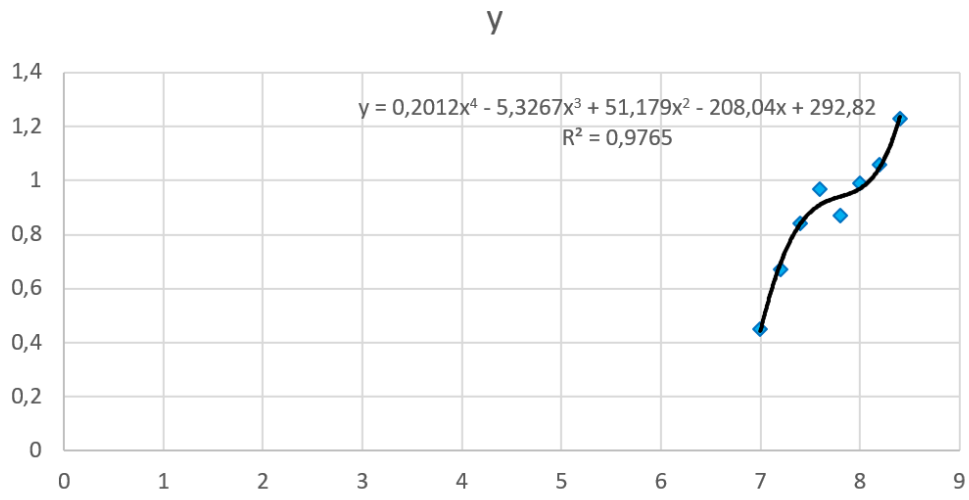

Рисунок 10 «Поліноміальна регресійна модель 4 ст»

З переліку поліноміальних моделей

- $y = -0{,}1786x^2 + 3{,}2119x - 13{,}222$
- $y = 0{,}8712x^3 - 20{,}304x^2 + 157{,}85x - 408{,}48$
- $y = 0{,}2012x^4 - 5{,}3267x^3 + 51{,}179x^2 - 208{,}04x + 292{,}82$

другого, третього та четвертого ступеня відповідно отримано наступні висновки:

2 ступінь - зовні схожа на логарифмічну, але більш випукла. Має коефіцієнт детермінації $R^2 = 0{,}9048$;

3 ступінь - зовні не має ознак, які зустрічались раніше у моделях, але ми починаємо бачити передумови для точок максимума та мінімума, які відхиляють лінію тренду до точок із найбільшим розкидом, що вже підвищує коефіцієнт детермінації до $R^2 = 0{,}976$;

4 ступінь – тендеція на кривизну у необхідні нам точки візуально ще більш зростає, що і за результатами аналізу та розрахунку дає нам найвищий коефіцієнт детермінації з усіх попередніх моделей, а саме $R^2 = 0{,}9765$;

Для покращення лінійної залежності між вибірковими елементами, побудовано степеневу регресійну модель, використовуючи інструмент Лінія тренду (рис.11).

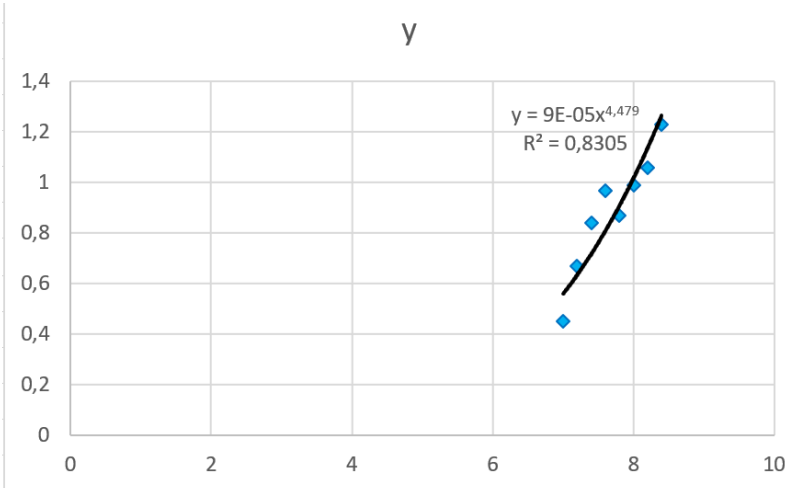

Рисунок 11 «Степенева регресійна модель»

На відміну від попередніх степенева задається рівнянням $y = 9E\text{-}05x^{4,479}$. Схожа на лінійну, але більш випукла. Має коефіцієнт детермінації $R^2 = 0,8305$, що показує велику адекватність степеневої моделі.

ВИСНОВКИ

У ході лабораторної роботи мною за методичними вказівками було отримано відповідно до завдання різні моделі регресійного аналізу методом найменших квадратів за допомогою програми статистичного аналізу MS Excel.

З вірогідністю 95% я відкидаю нульову гіпотезу через порівняння критичного та розрахованого критерію Фішера, отже ми приймаємо лінійну регресійну модель.

Знайдшов основні значення дисперсійного аналізу за такими формулами, як:

1) $\sigma_e^2 = 1/n \sum_1^n (y - Y_T)^2$, залишкова дисперсія;

2) $r = \frac{\sum_1^n xy - n\bar{x}\bar{y}}{n\sqrt{\sigma_x^2 \sigma_y^2}}$, лінійний коефіцієнт кореляції, використовується для лінійного зв’язку;

3) $r_b = \frac{\left(\sum n_{uv} uv - n\bar{u}\bar{v}\right)}{\left(n\tilde{\sigma}_u \tilde{\sigma}_v\right)}$, вибірковий коефіцієнт кореляції;

4) $\overline{y_x} = ax^2 + bx + c$, параболічна кореляція другого порядку;

5) $\overline{y_x} = ax^3 + bx^2 + cx + d$, параболічна кореляція третього порядку.

Побудовано такі моделі та встановлено їх адекватність щодо залежності наданих параметрів:

1 лінійна регресійна модель

$y = 0,4619x - 2,6717$, коефіцієнт детермінації дорівнює 0,8837, що досить близько до одиниці, тобто має сильну адекватність до реальної моделі. (див. рис 3).

2 експоненційна регресійна модель

$y = 0{,}0099e^{0{,}5784x}$, коефіцієнт детермінації дорівнює 0,8171, що досить близько до одиниці, тобто має сильну адекватність до реальної моделі, меньш адекватна ніж лінійна регресійна модель. (див. рис 6)

3 логарифмічна регресійна модель

$y = 3{,}562\ln(x) - 6{,}3794$, коефіцієнт детермінації дорівнює 0,8907, що досить близько до одиниці, тобто має сильну адекватність до реальної моделі. (див. рис 7)

4 поліноміальна регресійна модель 2 ступеня

$y = -0{,}1786x^2 + 3{,}2119x - 13{,}222$, коефіцієнт детермінації дорівнює 0,9048, що досить близько до одиниці, тобто має сильну адекватність до реальної моделі, також має найсильнішу адекватність. (див. рис 8)

5 поліноміальна регресійна модель 3 ступеня

$y = 0{,}8712x^3 - 20{,}304x^2 + 157{,}85x - 408{,}48$, коефіцієнт детермінації дорівнює 0,976, що досить близько до одиниці, тобто має сильну адекватність до реальної моделі, також має найсильнішу адекватність. (див. рис 9)

6 поліноміальна регресійна модель 4 ступеня

$y = 0{,}2012x^4 - 5{,}3267x^3 + 51{,}179x^2 - 208{,}04x + 292{,}82$, коефіцієнт детермінації дорівнює 0,9765, що досить близько до одиниці, тобто має сильну адекватність до реальної моделі, також має найсильнішу адекватність *з усієї* низки регресійних моделей. (див. рис 10)

7 степенева регресійна модель

$y = 9E\text{-}05x^{4,479}$, коефіцієнт детермінації дорівнює 0,8305, що досить близько до одиниці, тобто має сильну адекватність до реальної моделі. (див. рис 11)

Таким чином, дана вибірка відображає сильну кореляційну залежність, поліноміальна регресійна модель 4 ступеня має найвищий коефіцієнт детермінації з усіх розгянутих моделей($R^2 = 0,9765$), тобто найбільш адекватний по відношенню до реальних значень, тому саме цю модель потрібно обрати як найбільш достовірну.