Міністерство освіти і науки України
Національний технічний університет України «КПІ» імені Ігоря Сікорського
Кафедра обчислювальної техніки ФІОТ

ЗВІТ
з лабораторної роботи №2
з навчальної дисципліни «Аналіз даних для завдань електронної комерції»

Тема:

ДОСЛІДЖЕННЯ АЛГОРИТМІВ ЗГЛАДЖУВАННЯ ЗА НАКОПИЧЕНОЮ ВИБІРКОЮ

Виконав:
Кущ Родіон

Перевірив:
Писарчук
Олексій
Олександрович

Київ 2021

I. Мета:

виявити дослідити та узагальнити особливості застосування методів первинної обробки експериментальних вибірок – виявлення аномальних вимірів та алгоритмів накопиченого згладжування з використанням спеціалізованих пакетів мови програмування Python.

II. Завдання:

Лабораторія провідної IT-компанії реалізує масштабний проект розробки універсальної платформи з обробки Big Data масиву експериментальних даних поточного спостереження для виявлення закономірностей і прогнозування розвитку контрольованого процесу. Платформа передбачає розташування back-end компоненти на власному хмарному сервері з наданням повноважень користувачам заздалегідь адаптованого front-end функціоналу універсальної платформи.

Замовниками ресурсів платформи є: державні та комерційні компанії валютного трейдінгу для прогнозування динаміки зміни курсу валют та ціни інших товарів; метеорологічні служби для прогнозування параметрів метеоумов; департаменти охорони здоров'я для прогнозування зміни показників епідеміологічних ситуацій.

В продовження розвитку задач проекту минулого тижня (лабораторна робота №1) поточний перелік задач (tasks) для реалізації їх у межах лабораторної роботи №2 (Sprint – протягом тижня) для Вас, як Data Science Engineer на проекті включає:

Вам, як Data Science Engineer поставлене наступне завдання.

I. Розробити універсальний скрипт мовою Python що реалізує наступні етапи моделювання та обробки експериментальних даних.

1. Модель експериментальної вибірки з аномаліями відповідно до пунктів.

1.1. Розробити модель дискретних значень виміряних параметрів експериментальної вибірки з характеристиками: трендова модель має квадратичний закон зміни; вибірка має 1000 вимірів; випадкова похибка вимірів розподілена за нормальним законом з нульовим середнім та змінним значенням середньоквадратичної похибки вимірювання; модель виміру – адитивна:

$$B_{вим} = B_{ідеал} + \xi$$

Для виконання даного завдання використати результати лабораторної роботи №1.

1.2. Модель генерації аномальних вимірів випадкової величини:

$$B_{вим} = B_{ідеал} + \xi + (3\sigma + \xi)$$

.

Аномальні виміри складають 10% від загальної кількості вимірів у експериментальної вибірки. Аномальні виміри рівномірно розташовані у межах дискретних значень експериментальної вибірки.

2. Виявлення аномальних вимірів та усунення їх впливу на результати обробки відповідно до підходів, заданих у таблиці Д1 додатку 1.

3. Здійснити згладжування експериментальної вибірки за відсутності аномальних вимірів відповідно до матричної форму метода найменших квадратів (МНК).

4. З використанням методу Монте-Карло дослідити статистичні характеристики (математичне сподівання, середньоквадратичне відхилення, гістограма закону розподілу): закону розподілу випадкової похибки вимірів; вхідної вибірки значень (зашумленої без аномальних вимірів); аномальної вибірки (зашумленої з аномальними вимірами); результатів згладжування МНК.

5. Відобразити результати розрахунків:

5.1. Статистичні характеристики (математичне сподівання, середньоквадратичне відхилення) закону розподілу випадкової похибки вимірів; вхідної вибірки значень (зашумленої без аномальних вимірів); аномальної вибірки (зашумленої з аномальними вимірами); результатів згладжування МНК – у консолі та у формі таблички у звіті з лабораторної роботи.

5.2. Графіки (в одному графічному вікні): квадратичного тренду; зашумленої без аномальних вимірів вибірки; зашумленої з аномальними вимірами вибірки; результатів згладжування МНК.

5.3. Гістограми (в одному графічному вікні) похибок: зашумленої без аномальних вимірів вибірки; зашумленої з аномальними вимірами вибірки; результатів згладжування МНК.

Завдання II рівня складності – максимально 8 балів. Варіант 1.

Варіант	Алгоритм виявлення аномальних вимірів	Метод усунення впливу аномальних вимірів
1	За коефіцієнтом старіння інформації	Відновлення вимірів

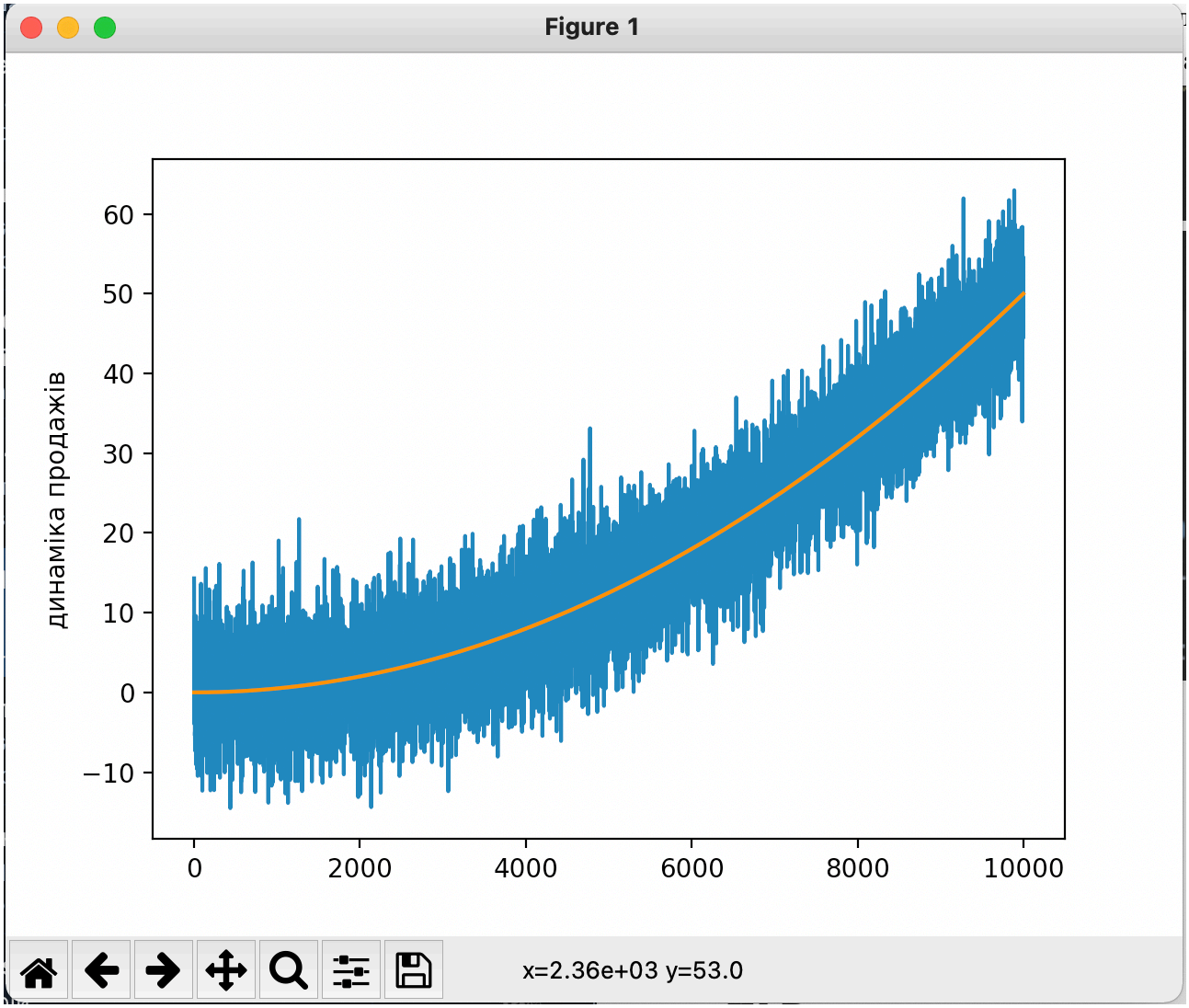
III. Результати виконання лабораторної роботи.

Для демонстрації результатів роботи програми була сгенерована вибірка з **10000** елементів.

Була сгенерована випадкова величина за квадратичним законом та накладеною на неї вибірки, сгенерованої за нормальним законом, в якості нормальної похибки.

```
normalS = np.random.normal(dm, dsig, iter)
```

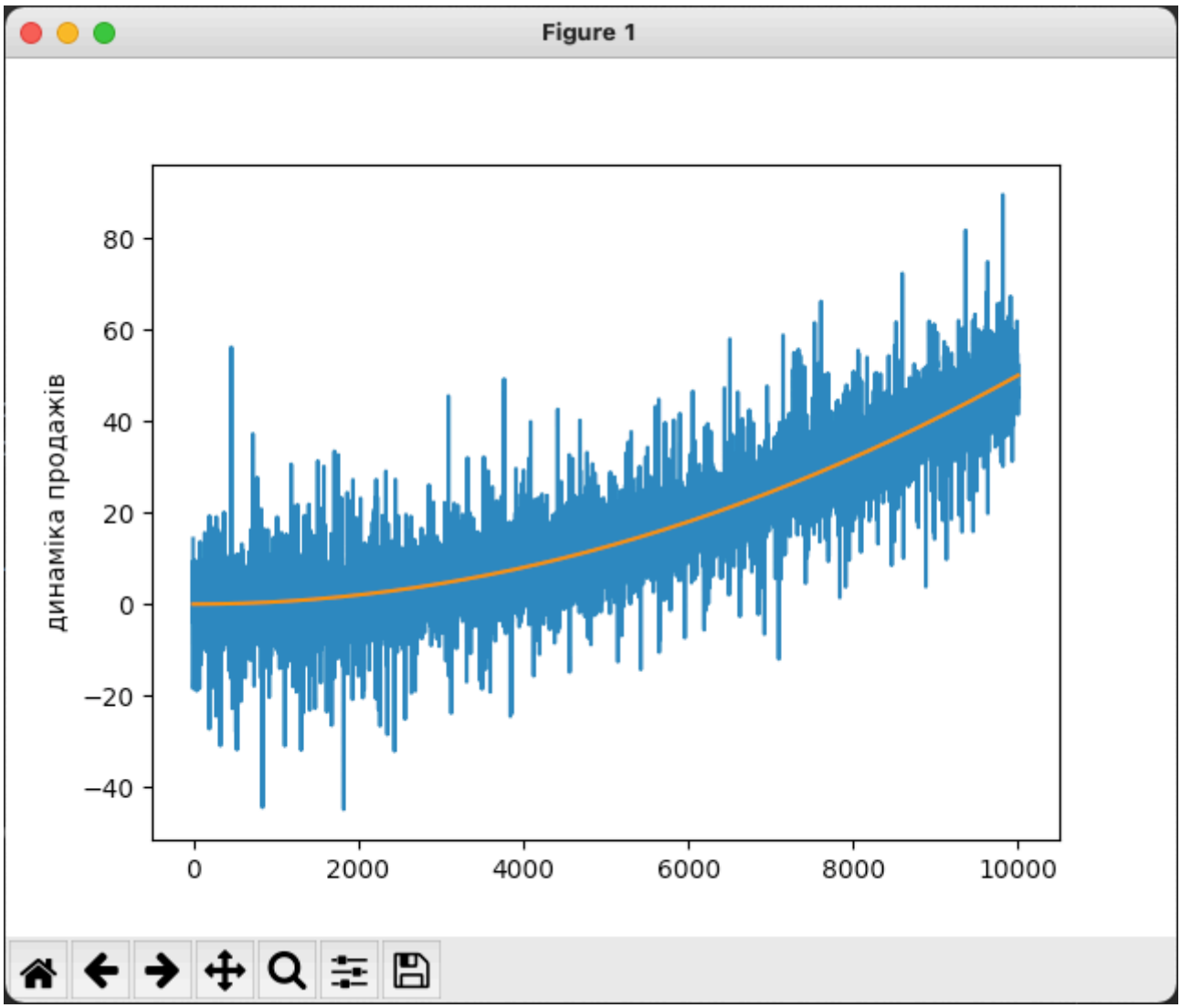
```
quadraticS = np.zeros(n)
quadraticSV = np.zeros(n)

SV0 = np.zeros(n)
quadraticSAV = np.zeros(n)

for i in range(n):
    quadraticS[i] = (0.0000005 * i * i)
    quadraticSV[i] = quadraticS[i] + normalS[i]
    SV0 = abs(quadraticSV[i] - quadraticS[i])
    quadraticSAV[i] = quadraticSV[i]
```



Далі була сгенерована множина аномальнх вимірів та накладена на попередню вибірку, розмір якої дорівнює 10% від початкової вибірки.

```
for i in range(nAV):
    SAV[i] = mt.ceil(np.random.randint(1, iter))
```

```
#=============================================================
SSAV = np.random.normal(dm, (3*dsig), nAV)
for i in range(nAV):
    k = int(SAV[i])
    quadraticSAV[k] = quadraticS[k] + SSAV[i]
```



За допомогою коефіцієнта сатріння інформації виявляємо аномальні виміри.

$(S_{cm}^{опт}+4)(S_{cm}^{опт}-1)^5 - A_k^2(S_{cm}^{опт}+4)^4 = 0$, де

$A_k^2 = \frac{t_{он}\alpha_n^2}{\sigma_k^2}$ - інтенсивність зміни досліджуваного процесу

$\alpha_n^2 = \vartheta_n - \vartheta_{n-1}$ - друга кінцева різниця за виміряними значеннями

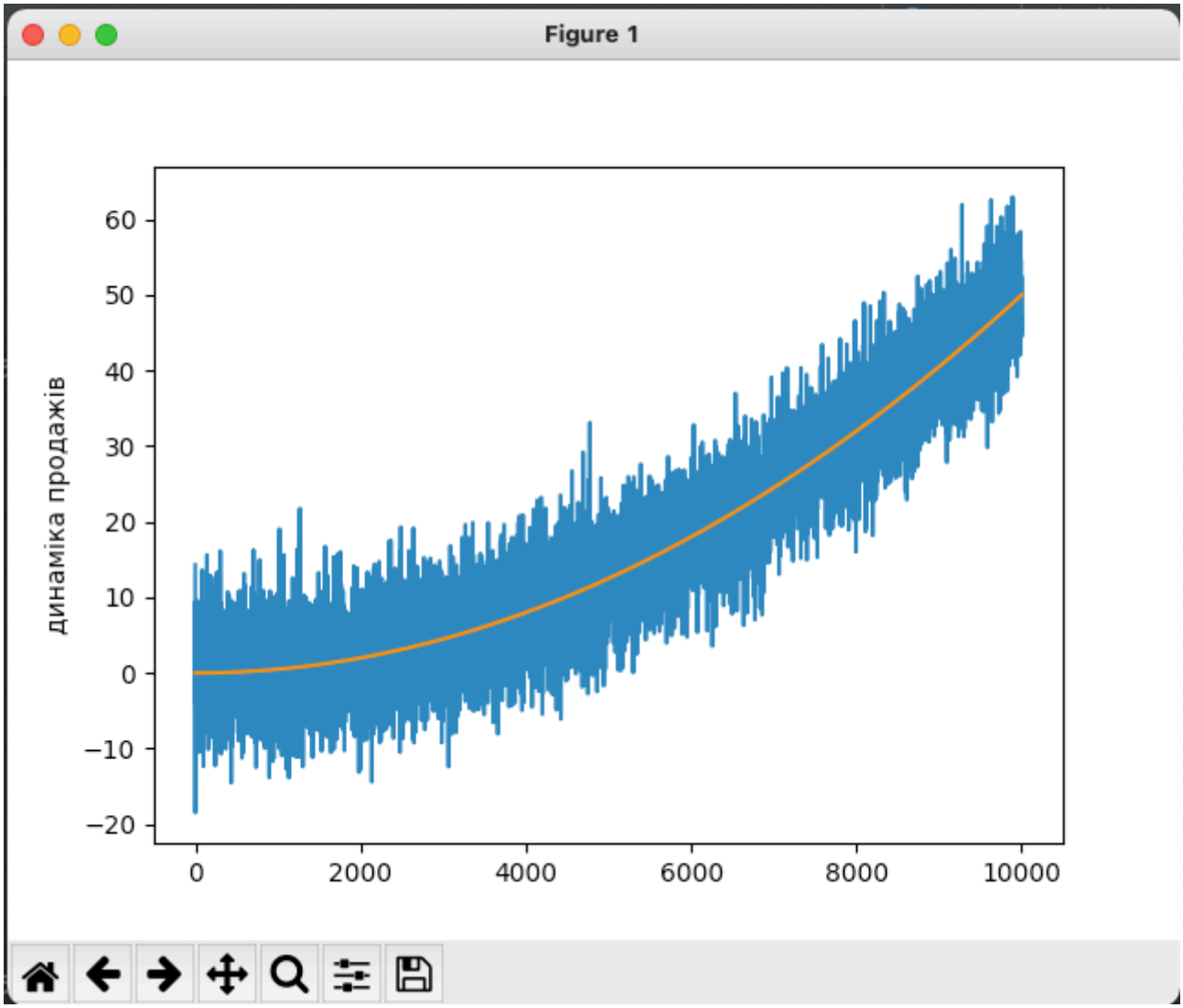
σ_k^2 - дисперсія вимірювання

$t_{он}$ - дискретність оновлення інформації

Розраховуємо інтенсивність зміни досліджуваного потоку та підставляємо замість кореня рівняння межі розв'язку a = 0 та b = 4. Якщо знак результату не змінвся - це означає,

що розв’язок рівняння виходить за встановлені межі, а одже даний вимір аномальний. Після виявлення аномального виміру ми заміняємо його середнім значенням між попереднім та наступним виміром.

```
for i in range(1, n-1):
    ai = (quadraticSAV[i+1] - quadraticSAV[i]) - (quadraticSAV[i] - quadraticSAV[i-1])
    Ak = i * ai/dQuadraticSAV
    x1 = (a + 4) * mt.pow((a - 1), 5) - Ak * mt.pow((a + 4), 4)
    x2 = (b + 4) * mt.pow((b - 1), 5) - Ak * mt.pow((b + 4), 4)
    if np.sign(x1) == np.sign(x2):
        quadraticSAV[i] = (quadraticSAV[i-1] + quadraticSAV[i+1])/2
```



Далі виконаємо згладжування отриманих вибірок за допомогою метода найменших квадратів.

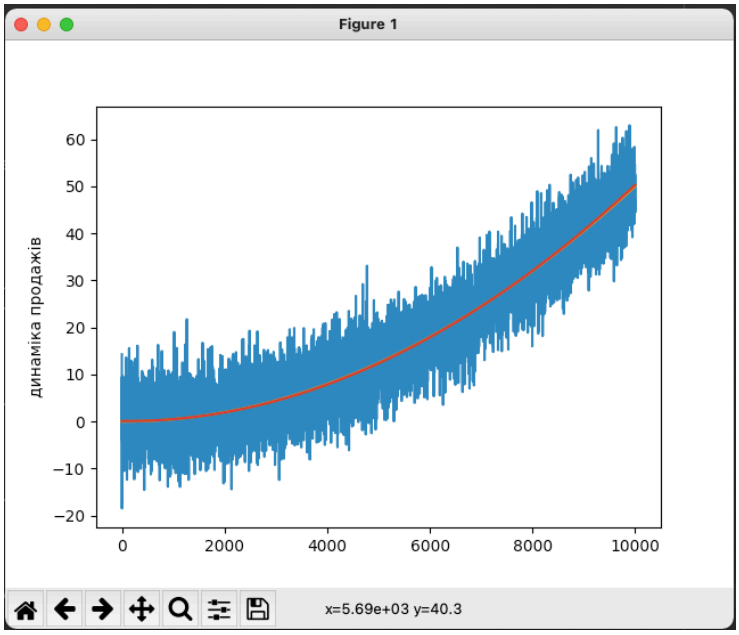
Матрична форма МНК

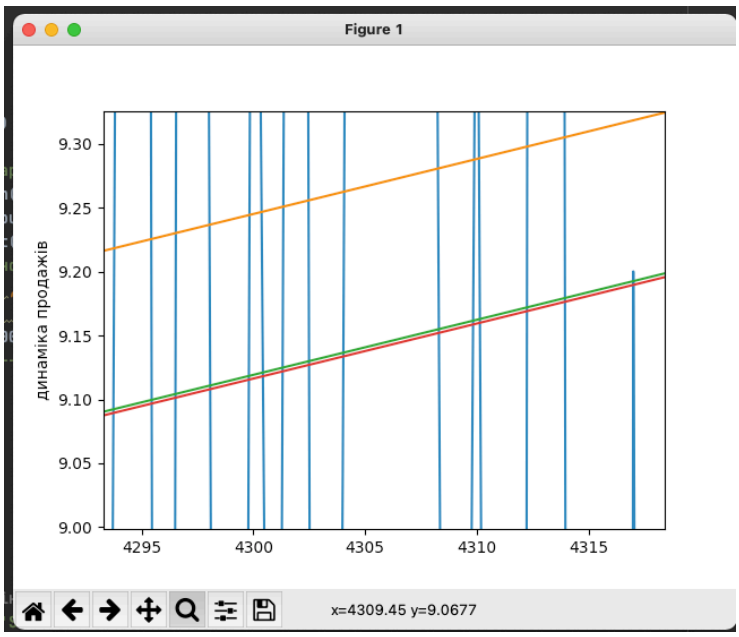
$$\bar{C} = (F^T F)^{-1} F^T \bar{Y},$$
$$\bar{T} = F\bar{C}.$$

```python
def MNK(Yin, F):
    FT = F.T
    FFT = FT.dot(F)
    FFTI = np.linalg.inv(FFT)
    FFTIFT = FFTI.dot(FT)
    C = FFTIFT.dot(Yin)
    Yout = F.dot(C)
    return Yout
```

```python
# ------------------------------ МНК згладжування ------------------------------------
Yin = np.zeros(iter)
F = np.ones((iter, 3))
for i in range(iter):                          # формування структури вхідних матриць МНК
    F[i, 1] = float(i)
    F[i, 2] = float(i*i)   # формування матриці вхідних даних без аномілій
# ---------------------- застосування МНК до незашумлених вимірів ----------------------
for i in range(iter):                          # формування структури вхідних матриць МНК
    Yin[i] = float(quadraticS[i])
YoutS = MNK(Yin, F)
# ---------------------- застосування МНК до незашумлених вимірів ----------------------
ArrayOrScalarCommon, Iterable, Sized, Container)
YoutSV = MNK(Yin, F)
# ---------------------- застосування МНК до незашумлених вимірів ----------------------
for i in range(iter):
    Yin[i] = float(quadraticSAV[i])
YoutSAV = MNK(Yin, F)
```

Результати зладжування для 3 вибірок: квадратична, квадратична з нормальним шумом, квадртаична з нормальним шумом та усередненими аномальними вимірами.






```
---------------------- статистичны характеристики вимiряної вибiрки за НАЯВНОСТI АВ -----------------
-------------------------- за вiдсутностi похибок ----- похибки нормальнi ------- похибки аномальнi ---
матиматичне сподiвання ВВ3= 8.171241461241152e-14 --┼-- 0.04136156073797048 ---- 0.06419504772596119
дисперсiя ВВ3 =              6.6507007832717225e-28 ---- 0.0005094816410733254 ---- 0.00120871778229338
СКВ ВВ3=                     2.578895264114408e-14 ---- 0.022571700004060957 ---- 0.03476661879293671
-------------------------------------------------------------------------------------------------------
```

IV. Висновки.

У ході виконання лабораторної роботи були використані методи виявлення аномальних вимірів та позбавлення від них за допомгою коефіцієнта старіння інформації. Також був використаний сетод найменших квадратів для згладжування вибірки. За результатами виконання лабораторної роботи можна зробити висновок, що за допомогою відкидування аномальних вимірів та згладжування вихідної вибірки значення дисперсіхї на середнього квадратичного відхилення значно зменшилось, тобто чітко можно відслідкувати лінію тренду вхідних данних, та зробити відповідний прогноз.

Виконав: