



МИНОБРНАУКИ РОССИИ  
Федеральное государственное бюджетное образовательное учреждение  
высшего образования  
«МИРЭА – Российский технологический университет»  
**РТУ МИРЭА**

---

**Институт информационных технологий (ИИТ)**  
**Кафедра прикладной математики (ПМ)**

**КУРСОВАЯ РАБОТА**

по дисциплине «Прикладные задачи математической статистики»

**Тема курсовой работы:** «Анализ статистики данных испытуемых во время лабораторного исследования воздействия стресса при ношении физиологических датчиков и датчиков движения (электродермальная активность)»

Студент группы ИМБО-02-22      Ким Кирилл Сергеевич

  
(подпись)

Руководитель  
курсовой работы

д.т.н., профессор Батенков К.А.

  
(подпись)

Работа представлена к защите      «\_\_» \_\_\_\_\_ 2024 г.

Допущен к защите      «\_\_» \_\_\_\_\_ 2024 г.

Москва 2024 г.



МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение  
высшего образования

«МИРЭА – Российский технологический университет»

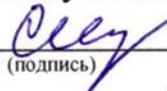
РТУ МИРЭА

Институт информационных технологий (ИИТ)

Кафедра прикладной математики (ПМ)

Утверждаю

Заведующий кафедрой ПМ

  
(подпись) Смоленцева Т.Е.

«20» сентября 2024 г.

## ЗАДАНИЕ

на выполнение курсовой работы

по дисциплине «Прикладные задачи математической статистики»

Студент Ким Кирилл Сергеевич

Группа ИМБО-02-22

**Тема** «Анализ статистики данных испытуемых во время лабораторного исследования воздействия стресса при ношении физиологических датчиков и датчиков движения (электродермальная активность)»

**Исходные данные:** наборы данных репозитория машинного обучения <https://archive.ics.uci.edu/datasets>

**Перечень вопросов, подлежащих разработке, и обязательного графического материала:**

Первичная обработка данных

Интервальные оценки параметров. доверительные интервалы точечных оценок

Идентификация распределений

Генерация распределений, проверка определений известных распределений


Линейная регрессия, оценка адекватности модели, оценка доверительных интервалов параметров

Сглаживание временных рядов

**Срок представления к защите курсовой работы:**

до «16» декабря 2024 г.

**Задание на курсовую работу выдал**


  
Подпись руководителя

Батенков К.А.

(ФИО руководителя)

«20» сентября 2024 г.

**Задание на курсовую работу получил**

  
Подпись обучающегося

Ким К.С.

(ФИО обучающегося)

«20» сентября 2024 г.

# СОДЕРЖАНИЕ

1.	ПЕРВИЧНАЯ ОБРАБОТКА ДАННЫХ .....	6
1.1.	Постановка задачи .....	6
1.2	Ход выполнения работы .....	6
1.3	Работа с целочисленными данными .....	7
1.3.1	Работа с вещественными данными .....	10
1.4	Вывод.....	14
2.	ИНТЕРВАЛЬНЫЕ ОЦЕНКИ ПАРАМЕТРОВ. ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ ТОЧЕЧНЫХ ОЦЕНОК .....	15
2.1	Постановка задачи .....	15
2.2	Ход выполнения работы .....	15
2.2.1	Расчёт точечных оценок математического ожидания и стандартного отклонения.....	16
2.2.2	Расчёт интервальной оценки математического ожидания .....	16
2.2.3	Расчёт интервальной оценки стандартного отклонения .....	18
2.3	Вывод.....	18
3.	ИДЕНТИФИКАЦИЯ РАСПРЕДЕЛЕНИЙ .....	20
3.1	Постановка задачи .....	20
3.2	Ход выполнения работы .....	20
3.2.1	Проверка распределения с помощью критерия Пирсона .....	20
3.2.2	Проверка распределения с помощью метода анаморфоза.....	23
3.2.3	Расчёт параметров распределений .....	25
3.3	Вывод.....	26
4.	ЛИНЕЙНАЯ РЕГРЕССИЯ. ОЦЕНКА АДЕКВАТНОСТИ МОДЕЛИ, ОЦЕНКА ДОВЕРИТЕЛЬНЫХ ИНТЕРВАЛОВ ПАРАМЕТРОВ .....	27
4.1	Постановка задачи .....	27
4.2	Ход выполнения работы .....	28
4.2.1	Оценка коэффициента корреляции Пирсона. Оценка характеристики корреляционной связи по шкале Чеддока.....	28

4.2.2	Проверка статистической значимости коэффициента Корреляции ....	29
4.2.3	Построение модели линейной регрессии .....	30
4.2.4	Оценка адекватности модели .....	31
4.2.5	Оценка значимости коэффициентов модели .....	32
4.2.6	Построение доверительных интервалов коэффициентов модели .....	33
4.2.7	Оценка интервала прогноза линейной модели .....	34
4.2.8	Проведение теста Чоу .....	35
4.2.9	Построение модели линейной регрессии. Оценка значимости коэффициентов и адекватности модели .....	36
4.2.10	Проверка данных на гетероскедастичность при помощи теста Гольдфельда-Квандта и теста Спирмена .....	37
4.3	Вывод .....	42
5.	СГЛАЖИВАНИЕ ВРЕМЕННЫХ РЯДОВ .....	43
5.1	Постановка задачи .....	43
5.2	Ход выполнения работы .....	44
5.2.1	SMA-сглаживание .....	46
5.2.2	WMA-сглаживание .....	47
5.2.3	EMA-сглаживание .....	49
5.2.4	DEMA-сглаживание .....	51
5.3	Вывод .....	53
	Список использованных источников .....	55

# 1. ПЕРВИЧНАЯ ОБРАБОТКА ДАННЫХ

## 1.1. Постановка задачи

1. Для целочисленных данных: построить вариационный ряд и полигон относительных частот. Для вещественных данных: построить таблицу групп и гистограмму.
2. Построить для целочисленных и вещественных данных эмпирическую функцию распределения.
3. Для целочисленных и вещественных данных рассчитать:
  - a) Выборочное среднее.
  - b) Выборочную дисперсию.
  - c) Выборочное стандартное отклонение.
  - d) Выборочную медиану.
  - e) Выборочную вариацию.

## 1.2 Ход выполнения работы

В качестве исходных данных была создана таблица, в которую каждый студент группы ввел свои параметры: рост, месяц рождения и случайное число от нуля до восьми. Фрагмент таблицы представлен на Рисунке 1.1.

	name	num	height	month
0	Timofey	7	1.84	1
1	Kirill	0	1.75	11
2	Bogdan	8	1.80	12
3	Arseniy	2	1.84	9
4	Aisa	1	1.57	1
5	Vanya	0	1.85	12
6	Dima	5	1.83	8
7	Arthur	7	1.80	12

Рисунок 1.1 — Таблица исходных данных

### 1.3 Работа с целочисленными данными

Данные: [7, 0, 8, 2, 1, 0, 5, 7]

Вариационный ряд: [0, 0, 1, 2, 5, 7, 7, 8]

Таблица 1.1 — Статистический ряд

х-варианты	0	1	2	5	7	8
n-абсолютная частота	2	1	1	1	2	1
w-относительная частота	0.25	0.125	0.125	0.125	0.25	0.125

Следующим шагом мы построили полигон относительных частот вариационного ряда (Рисунок 1.2).

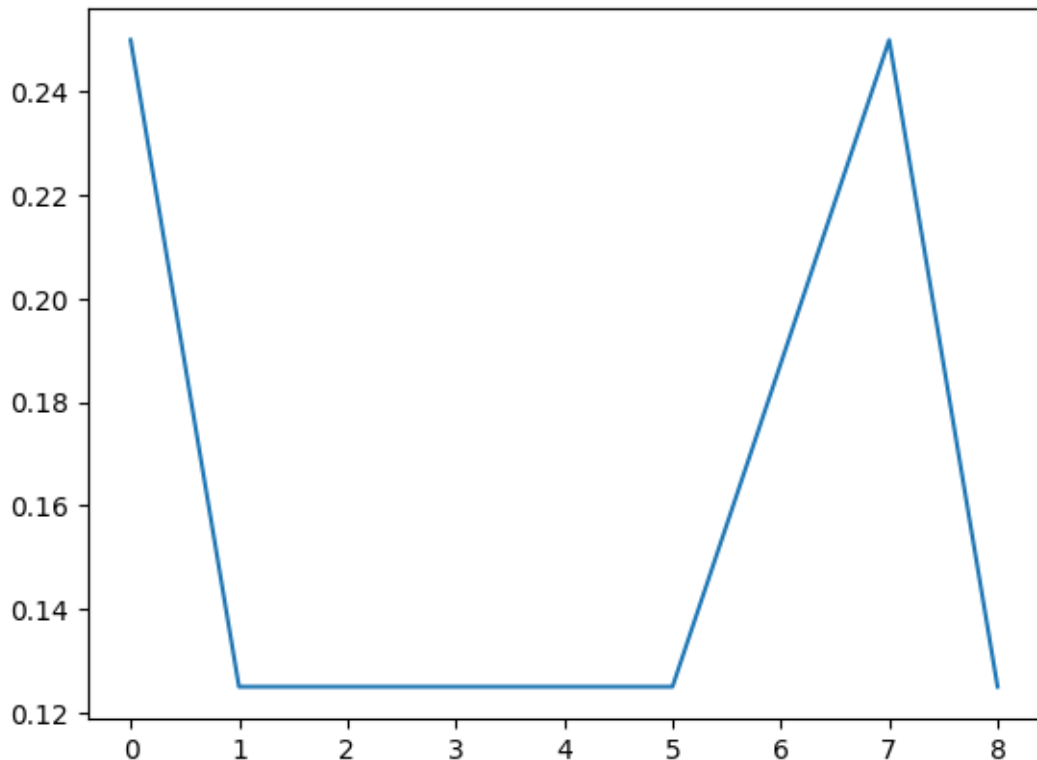


Рисунок 1.2 — Полигон частот

Полигон частот — один из способов графического представления плотности вероятности случайной величины. Представляет собой ломаную, соединяющую точки, соответствующие срединным значениям интервалов группировки и относительным частотам этих интервалов.

Вариационный ряд — ряд значений признака, расположенных в порядке возрастания или убывания. Абсолютная частота — сколько количественно эта варианта встречается в ряду. Относительная частота —

процент вхождения варианты в ряд.

Таблица 1.2 — Эмпирическая функция распределения

x	0	1	2	5	7	8	>8
F	0	0.125	0.25	0.5	0.625	0.875	1

$$F = \begin{cases} 0, & 0 < x \\ 0.125, & 0 \leq x < 1 \\ 0.25, & 1 \leq x < 2 \\ 0.5, & 2 \leq x < 5 \\ 0.625, & 5 \leq x < 7 \\ 0.875, & 7 \leq x < 8 \\ 1, & x > 8 \end{cases}$$



Рисунок 1.3 — Эмпирическая функция распределения

Все статистические характеристики, рассчитанные ниже.

Выборочная средняя рассчитывается по формуле (1.1):

$$\bar{x}_B = \frac{1}{n} \sum_{i=1}^n x_i \quad (1.1)$$

где  $\bar{x}_B$  — выборочная средняя;

$n$  — количество данных в выборке;

$i$  — номер элемента выборки;

$x_i$  —  $i$ -ый элемент выборки.

Выборочная дисперсия рассчитывается по формуле (1.2):

$$D_B = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_B)^2 \quad (1.2)$$

где  $D_B$  — выборочная дисперсия;

$n$  — количество данных в выборке;

$i$  — номер элемента выборки;

$x_i$  —  $i$ -ый элемент выборки;

$\bar{x}_B$  — выборочная средняя.

Выборочное стандартное отклонение (СКО) рассчитывается по формуле (1.3):

$$\sigma = \sqrt{D_B} \quad (1.3)$$

где  $\sigma$  — СКО;

$D_B$  — выборочная дисперсия.

Выборочная вариация рассчитывается по формуле (1.4):

$$\vartheta = \frac{\sigma}{\bar{x}_B} * 100\% \quad (1.4)$$

Результаты вычислений статистических характеристик для показателя «Рост» представлены в Таблице 1.3.

Таблица 1.3 — Результаты вычислений для показателя «Случайного числа»

Характеристика	Значение
Выборочная средняя	3.75
Выборочная дисперсия	9.9375
СКО	3.15
Выборочная медиана	3.5
Выборочная вариация	8.4%



### 1.3.1 Работа с вещественными данными

Данные: [1.84, 1.75, 1.80, 1.84, 1.57, 1.85, 1.83, 1.80]

Вариационный ряд: [1.57, 1.75, 1.8, 1.8, 1.83, 1.84, 1.84, 1.85]

Используя правило Стёрджеса, вычислим количество интервалов.

$$m = 1 + 3.332 * \log_{10}(8) \approx 4$$

Далее вычислим ширину интервала.

$$h = \frac{x_{\max} - x_{\min}}{m} = \frac{1.85 - 1.57}{m} = 0.07$$

Первый интервал рекомендуется начать с такой точки.

$$start = x_{\min} - \frac{h}{2} = 1.57 - \frac{0.07}{2} = 1.535$$

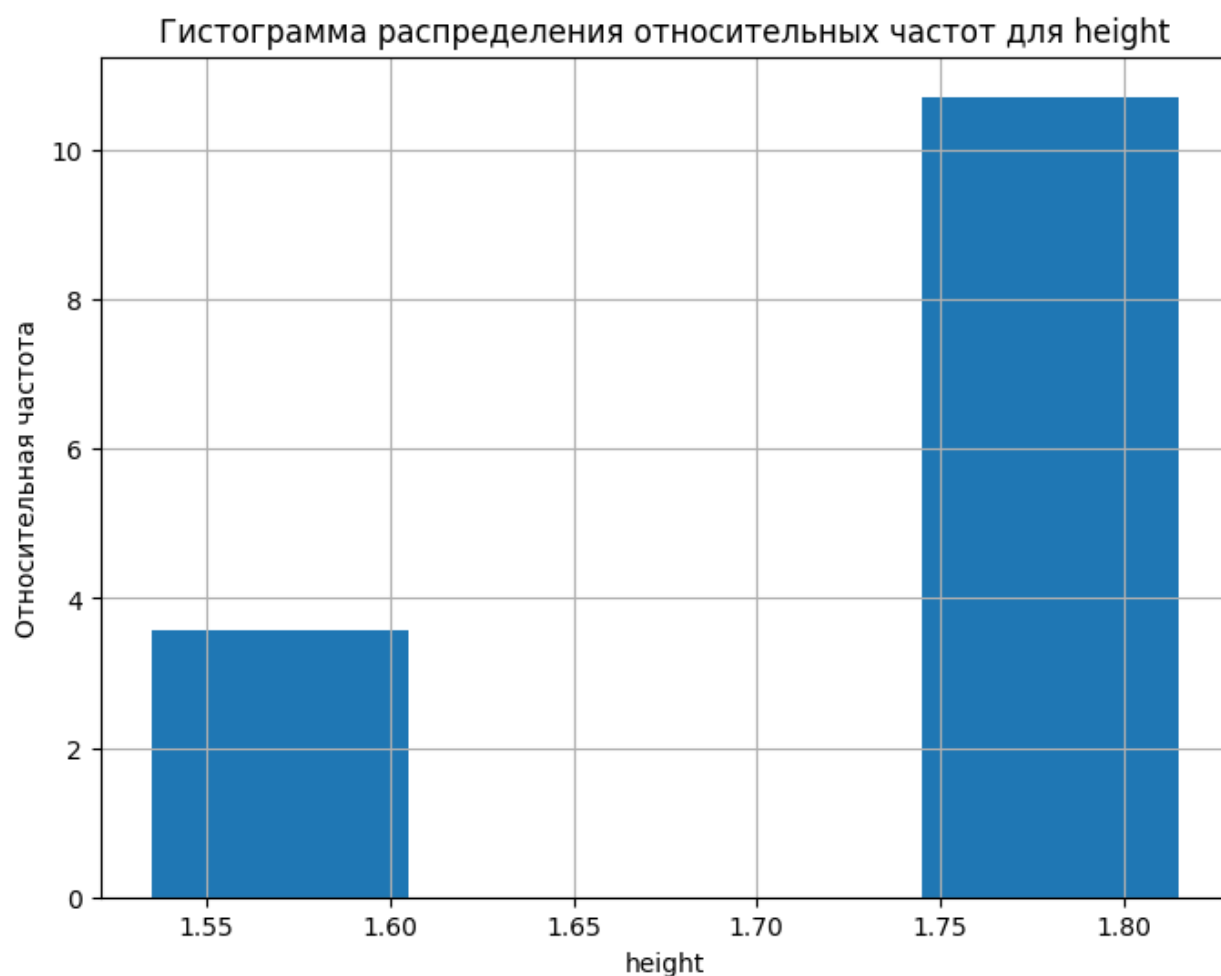
Таблица 1.4 — Интервальный ряд

1.535	1.605	1.675	1.745
1.605	1.675	1.745	1.815

Рассчитаем относительные частоты для каждого интервала.

Таблица 1.5 — Дискретный ряд

Интервал	1	2	3	4
n	1	0	0	3
w	0.25	0	0	0.75



**Рисунок 1.4 — Гистограмма интервального ряда**

Гистограмма распределения относительных частот для рассчитанных интервалов выборки.

Эмпирическая функция распределения:

*Таблица 1.6 — Эмпирическая функция распределения*

x	1.57	1.75	1.8	1.83	1.84	1.85	>1.85
F	0	0.125	0.25	0.625	0.625	0.75	1

$$F = \begin{cases} 0, & 0 < x \\ 0.125, & 1.57 \leq x < 1.75 \\ 0.25, & 1.75 \leq x < 1.8 \\ 0.5, & 1.8 \leq x < 1.83 \\ 0.75, & 1.83 \leq x < 1.84 \\ 0.875, & 1.84 \leq x < 1.85 \\ 1, & x > 1.85 \end{cases}$$



**Рисунок 1.5 — Эмпирическая функция распределения**

Аналогичные вычисления произведены для показателя «Рост» и представны в Таблице 1.7

*Таблица 1.7 — Результаты вычислений для показателя «Рост»*

Характеристика	Значение
Выборочная средняя	1.785
Выборочная дисперсия	0.007525
СКО	0,0867
Выборочная медиана	1.815
Выборочная вариация	4.86%

Выборочная средняя — это среднее значение набора данных, полученное из выборки. Вычисляется как сумма всех значений, деленная на количество значений.

Выборочная дисперсия — это мера разброса значений в выборке относительно выборочной средней. Вычисляется как среднее значение квадратов отклонений каждого значения от выборочной средней. Помогает понять, насколько сильно значения в выборке отличаются друг от друга. Чем больше дисперсия, тем более разбросаны данные.

Стандартное отклонение (СКО) показывает, насколько сильно значения выборки отклоняются от среднего значения. Чем выше стандартное отклонение, тем больше рассеяние значений выборки.

Выборочная медиана — это значение, которое делит упорядоченный набор данных на две равные части. Если количество значений нечетное, медиана — это среднее значение двух центральных чисел. Используется для определения центральной тенденции, особенно в случаях, когда данные имеют выбросы или неравномерное распределение, так как медиана менее чувствительна к крайним значениям по сравнению со средней.

Коэффициент вариации (CV) - это статистический показатель, который характеризует степень рассеяния или вариабельности данных относительно их среднего значения. Показывает, насколько сильно значения данных отклоняются от среднего значения в процентном отношении.

## 1.4 Вывод

В ходе выполнения работы мы познакомились с процессом вычисления стандартных описательных статистик для выборок данных различных типов, были построены различные графики и рассчитаны описательные статистики. Были выполнены следующие задачи:

1. Для целочисленных данных был построен вариационный ряд и полигон относительных частот.
2. Для вещественных данных построена таблица групп и гистограмма.
3. Для целочисленных и вещественных данных построена эмпирическая функция распределения.
4. Для целочисленных и вещественных данных рассчитаны:
  - a. Выборочное среднее.
  - b. Выборочную дисперсию.
  - c. Выборочное стандартное отклонение.
  - d. Выборочную медиану.
  - e. Выборочную вариацию.

## **2. ИНТЕРВАЛЬНЫЕ ОЦЕНКИ ПАРАМЕТРОВ. ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ ТОЧЕЧНЫХ ОЦЕНОК**

### **2.1 Постановка задачи**

1. Скачать папку с исходными данными. В ней находятся 4 ряда данных реализации случайной величины.
2. Для каждого из четырех рядов данных необходимо провести следующие расчёты:
  - подсчитать выборочные статистики для среднего и стандартного отклонения;
  - для выборочного среднего  $\bar{X}_B$  подсчитать границы доверительного интервала по правилу нормального распределения и по правилу  $t$ -распределения Стьюдента;
  - для выборочного среднеквадратического отклонения  $\sigma_B$  подсчитать границы доверительного интервала по оценке  $\chi^2$ -распределения.

### **2.2 Ход выполнения работы**

В качестве исходных данных даны четыре файла, в первом и третьем файлах десять случайных значений, во втором и четвертом — тридцать два значения.

### 2.2.1 Расчёт точечных оценок математического ожидания и стандартного отклонения.

Были рассчитаны выборочные средние и СКО для четырех наборов данных. Выборочная средняя рассчитана по формуле (1.1). СКО рассчитано по формуле (2.1).

$$\sigma_B = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x}_B)^2}{N-1}} \quad (2.1)$$

где  $N$  — количество значений в ряде данных;

$x_i \in \mathbb{R}, i \in \overline{1, N}$  — значения ряда данных.

Результаты вычислений представлены в Таблице 2.1.

Таблица 2.1 — Результаты вычислений

	Первая выборка	Вторая выборка	Третья выборка	Четвертая выборка
<b>Выборочная средняя</b>	39.39	46.44	75.16	43.74
<b>СКО</b>	1.8	7.6	8.8	5.9

### 2.2.2 Расчёт интервальной оценки математического ожидания

Были заданы два выражения, с помощью которых был рассчитан доверительный интервал нормального распределения для заданных таблиц данных по формуле (2.2), используя таблицу критических значений функции Лапласа  $\Phi(x)$ .

$$\hat{X}_B \in \left[ \bar{X}_B - X_\gamma \cdot \frac{\sigma_B}{\sqrt{N}}, \bar{X}_B + X_\gamma \cdot \frac{\sigma_B}{\sqrt{N}} \right], \quad \Phi(X_\gamma) = \frac{\gamma}{2} \quad (2.2)$$

где  $N$  — количество значений в ряде данных;

$\bar{X}_B$  — выборочная средняя;

$X_\gamma$  — критическое значение из таблицы Лапласа.

$\sigma_B$  — выборочное СКО.

Результат представлен на Рисунках 2.1-2.4.

[39.03439118652976; 39.742267714256315]

Рисунок 2.1 — Доверительный интервал для данных №1

[41.71163687944756; 51.16814559910761]

Рисунок 2.2 — Доверительный интервал для данных №2

[73.428393991547; 76.8891184718983]

Рисунок 2.3 — Доверительный интервал для данных №3

[40.10630183547303; 47.38272595613557]

Рисунок 2.4 — Доверительный интервал для данных №4

Далее был рассчитан доверительный интервал распределения Стьюдента для заданных таблиц данных, используя таблицу критических значений  $t_{\gamma, N}$  t-распределения при значении уверенности  $\gamma = 0.95$ . Доверительный интервал рассчитывается по формуле (2.3):

$$\hat{X}_B \in \left[ \overline{X}_B - t_{1-\gamma, N-1} \cdot \frac{\sigma_B}{\sqrt{N}}, \overline{X}_B + t_{1-\gamma, N-1} \cdot \frac{\sigma_B}{\sqrt{N}} \right] \quad (2.3)$$

где  $N$  — количество значений в ряде данных;

$\overline{X}_B$  — выборочная средняя;

$t_{1-\gamma, N-1}$  — критическое значение из таблицы Стьюдента;

$\sigma_B$  — выборочное СКО.

Результат представлен на Рисунках 2.5-2.8.

[39.030779571592376; 39.7458793291937]

Рисунок 2.5 — Доверительный интервал для данных №1

[40.98792447743276; 51.89185800112241]

Рисунок 2.6 — Доверительный интервал для данных №2

[73.41073723399418; 76.90677522945111]

Рисунок 2.7 — Доверительный интервал для данных №3

[39.549432642565186; 47.93959514904342]

Рисунок 2.8 — Доверительный интервал для данных №4



### 2.2.3 Расчёт интервальной оценки стандартного отклонения

Затем был рассчитан доверительный интервал среднего квадрата отклонения для заданных таблиц данных при значении уверенности  $\gamma = 0.95$ . Доверительный интервал рассчитывается по формуле (2.4):

$$\widehat{\sigma_B} \in \left[ \frac{\sigma_B \cdot \sqrt{N-1}}{\sqrt{\chi_{\frac{1+\gamma}{2}, N-1}^2}}, \frac{\sigma_B \cdot \sqrt{N-1}}{\sqrt{\chi_{\frac{1-\gamma}{2}, N-1}^2}} \right] \quad (2.4)$$

где  $N$  — количество значений в ряде данных;

$\chi_{\frac{1-\gamma}{2}, N-1}^2$  — критическое значение из таблицы хи-квадрат;

$\sigma_B$  — выборочное СКО.

Результат представлен на Рисунках 2.9-2.12.

(1.5855113189914307, 2.0977618522036208)

Рисунок 2.9 — Доверительный интервал для данных №1

(5.247218652311081, 13.926847168585525)

Рисунок 2.10 — Доверительный интервал для данных №2

(7.751376999503043, 10.255709168919488)

Рисунок 2.11 — Доверительный интервал для данных №3

(4.0375353632031885, 10.71617968813331)

Рисунок 2.12 — Доверительный интервал для данных №4

## 2.3 Вывод

В ходе выполнения работы были изучены виды распределений и рассчитаны доверительные интервалы для заданных групп данных. Были выполнены следующие задачи:

1. Подсчитаны выборочные статистики для среднего и стандартного отклонения.

2. Для выборочного среднего  $\overline{X}_B$  подсчитаны границы доверительного интервала по правилу нормального распределения и по правилу t-распределения Стьюдента.
3. Для выборочного среднеквадратического отклонения  $\sigma_B$  подсчитаны границы доверительного интервала по оценке  $\chi^2$  - распределения.

### 3. ИДЕНТИФИКАЦИЯ РАСПРЕДЕЛЕНИЙ

#### 3.1 Постановка задачи

Исходные данные для работы представляют собой 4 файла, данные в которых распределены следующим образом:

- в 1 и 4 файлах — по нормальному закону;
- в 3 и 6 файлах — по показательному закону.

Необходимо идентифицировать распределения в каждом файле двумя способами: с помощью критерия согласия Пирсона и методом анаморфоз.

#### 3.2 Ход выполнения работы

Рассчитаем количество интервалов каждого ряда по правилу Стерджесса.

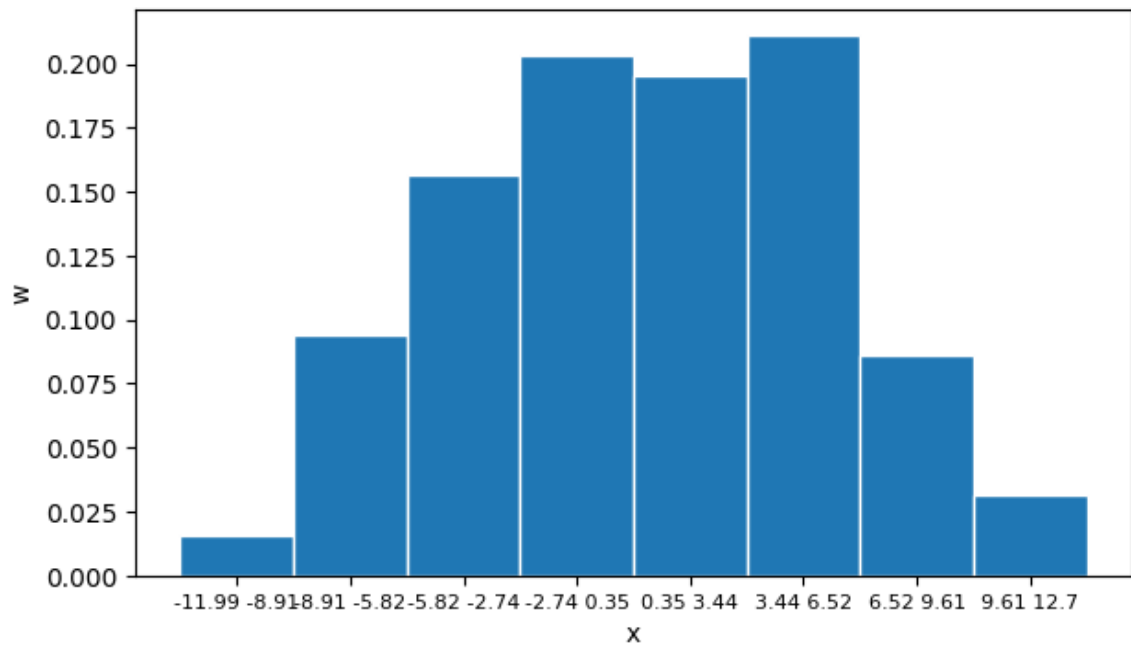
$$m = 1 + \log_2(n) \quad (3.1)$$

Далее вычислим ширину интервала.

$$h = \frac{x_{\max} - x_{\min}}{m} \quad (3.2)$$

##### 3.2.1 Проверка распределения с помощью критерия Пирсона

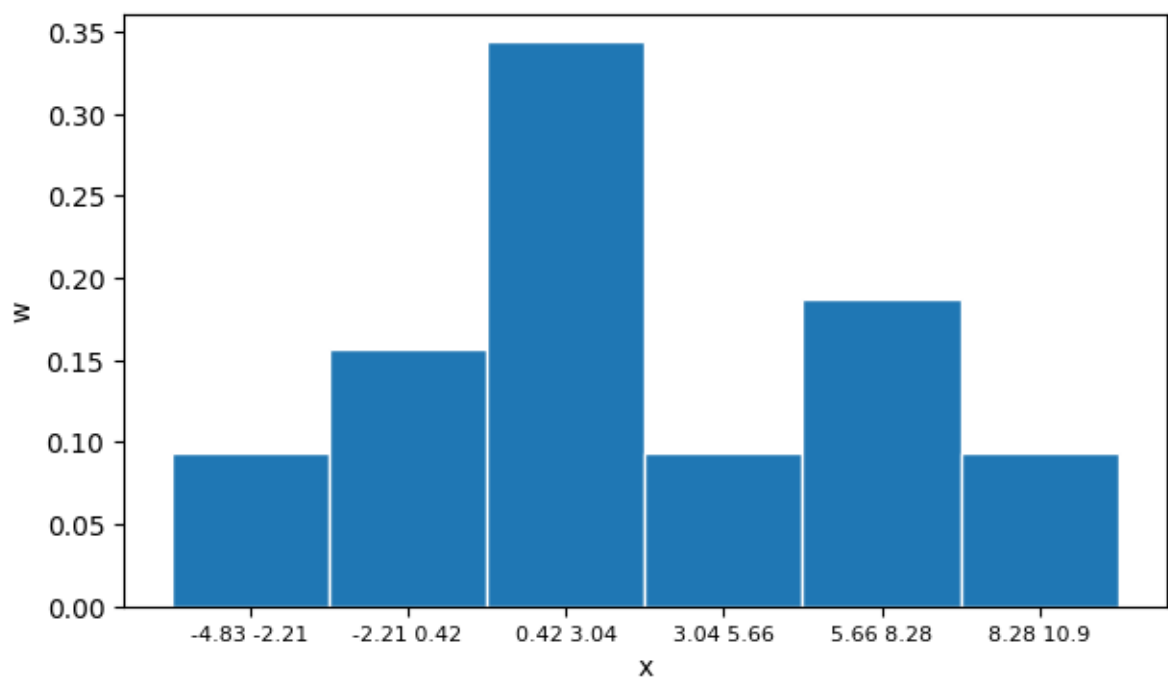
Проведем тест Пирсона для всех четырех выборок. Каждую выборку проверим и на нормальное, и на показательное распределение. Результаты тестирования показаны ниже.



**Рисунок 3.1 — Гистограмма первой выборки**

Полученное значение критерия Пирсона — 3.87, критическое значение — 11.07.

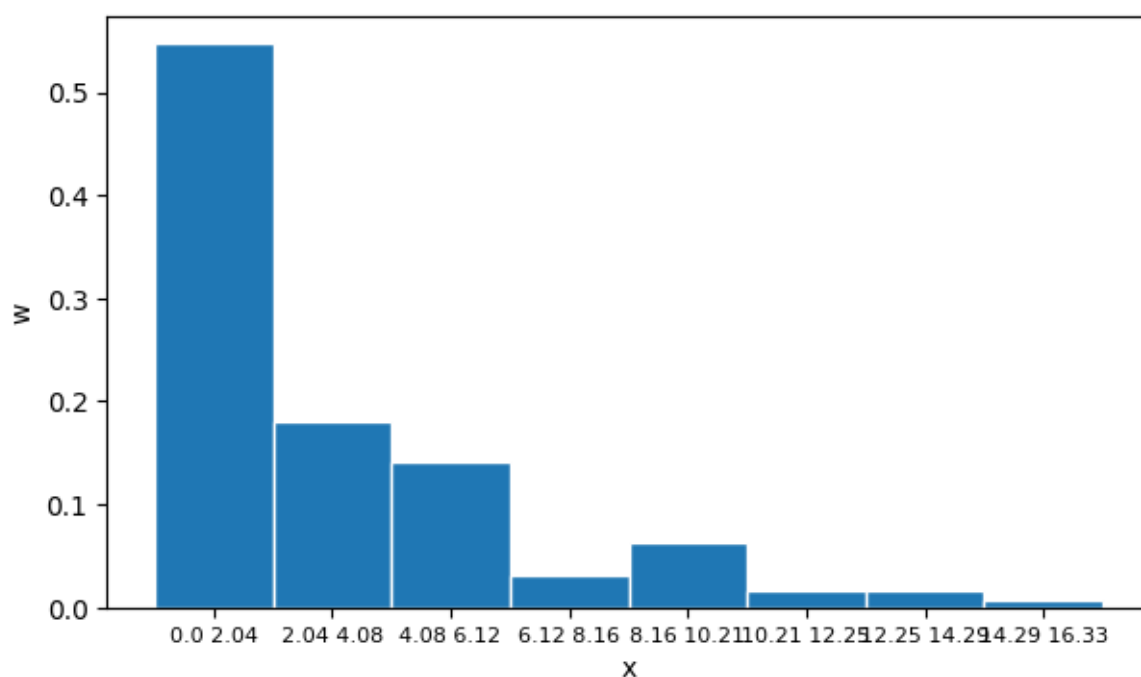
Выборка принадлежит нормальному распределению.



**Рисунок 3.2 — Гистограмма второй выборки**

Полученное значение критерия Пирсона — 5.16, критическое значение — 7.81.

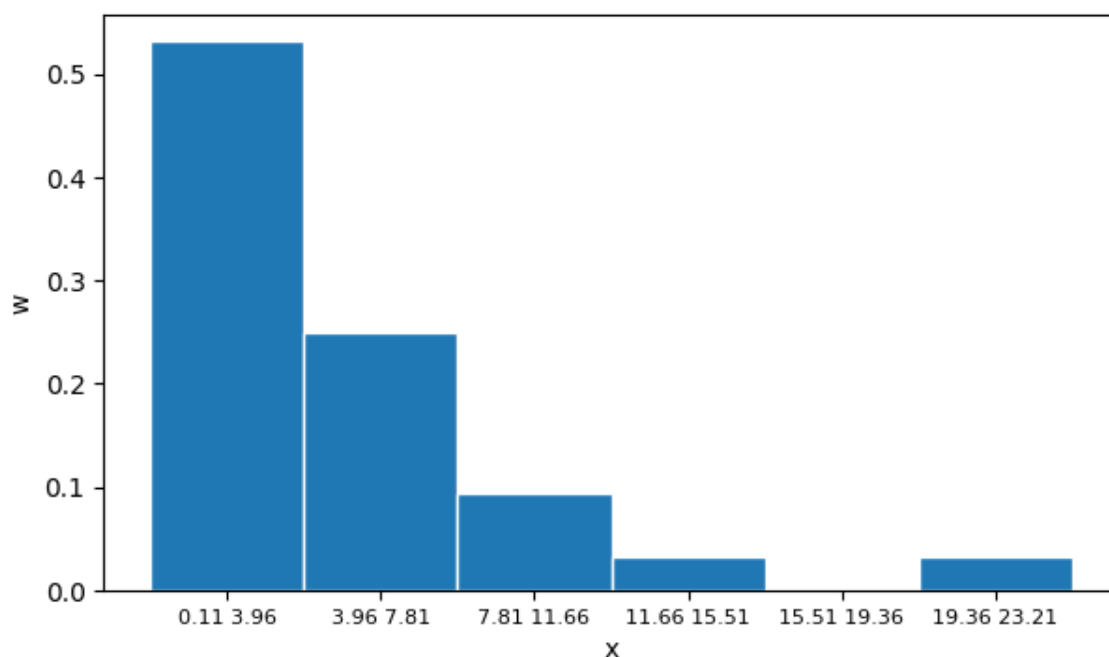
Выборка принадлежит нормальному распределению.



**Рисунок 3.3 — Гистограмма третьей выборки**

Полученное значение критерия Пирсона — 145.9, критическое значение — 11.07.

Выборка НЕ принадлежит нормальному распределению.



**Рисунок 3.4 — Гистограмма четвертой выборки**

Полученное значение критерия Пирсона — 23.13, критическое значение — 7.81.

Выборка НЕ принадлежит нормальному распределению.

### 3.2.2 Проверка распределения с помощью метода анаморфоза

Проверим гипотезу о распределении выборки при помощи анаморфоз. Анаморфоза нормального распределения представлена в формуле (3.3):

$$\ln(p_i) \sim (x_{(i)} - \mu)^2 \quad (3.3)$$

В данном случае при построении анаморфозы нормального распределения  $\mu$  заменяется точечной, полученной ранее.

Анаморфоза равномерного распределения представлена в формуле (3.4):

$$P(x_{(i)}) = \sum_{j=1}^i p_i \sim x_{(i)} \quad (3.4)$$

Анаморфоза экспоненциального распределения представлена в формуле (3.5):

$$\ln(p_i) \sim x_{(i)} \quad (3.5)$$

Соответствующие анаморфозы, построенные для каждой из выборок, представлены на Рисунках 3.5-3.8.

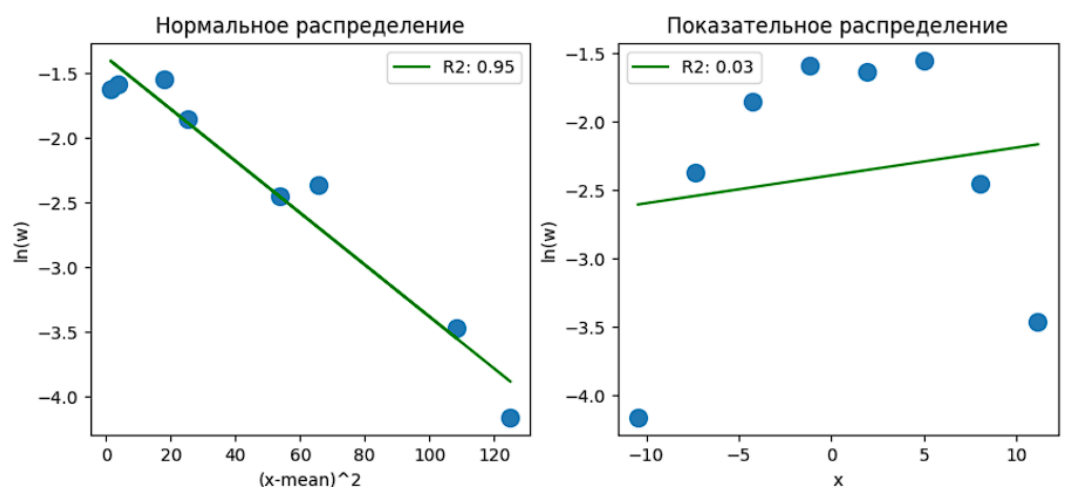


Рисунок 3.5 — Метод анаморфоз для первой группы

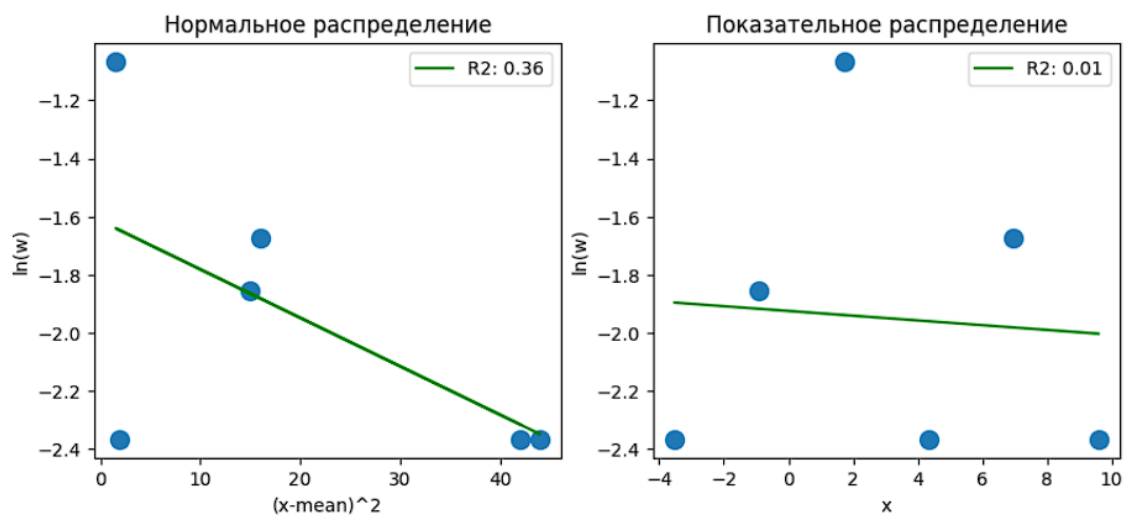


Рисунок 3.6 — Метод анаморфоз для второй группы

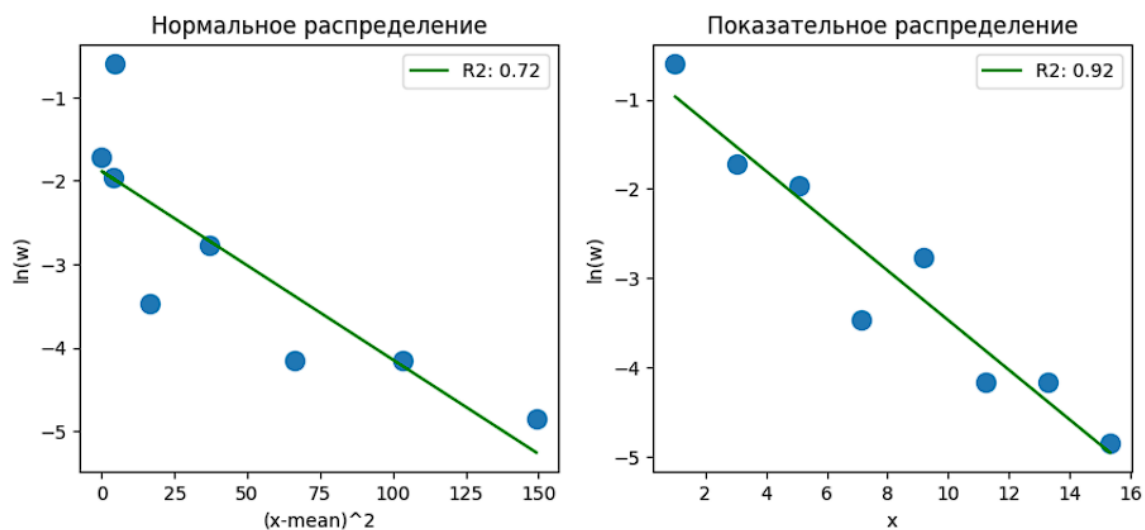


Рисунок 3.7 — Метод анаморфоз для третьей группы

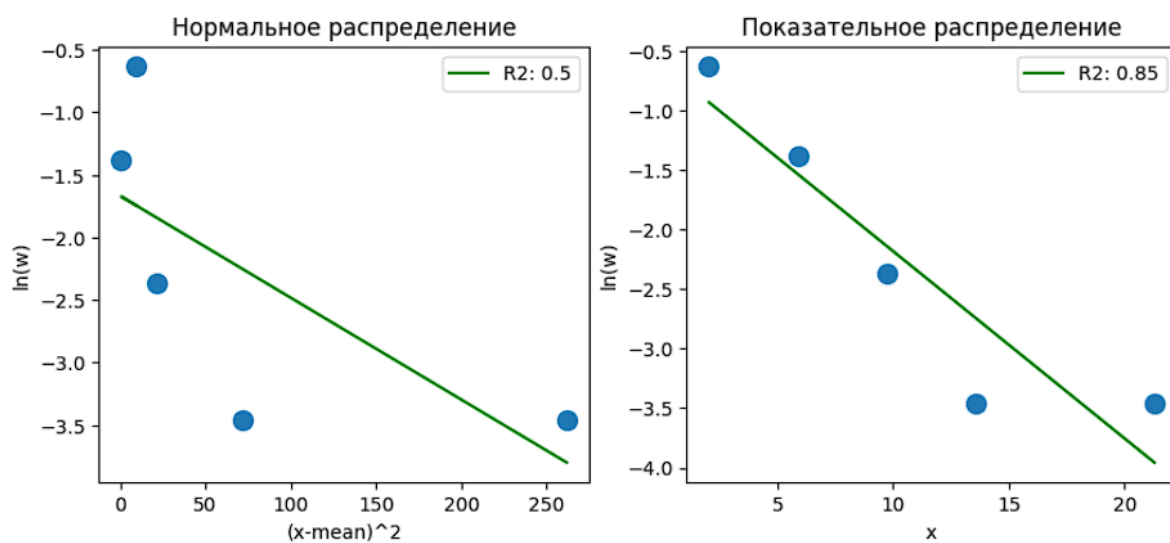


Рисунок 3.8 — Метод анаморфоз для четвертой группы

Чтобы проверить данные гипотезы, построим для выбранных ранее спрямлений линейную регрессию и вычислим коэффициент детерминации между прямой регрессии и частотами выборки. Коэффициент детерминации вычисляется по формуле (3.6):

$$R^2 = r_{p_i y'_i}^2 = \left( \frac{\overline{p_i y'_i} - \bar{p}_i \cdot \bar{y}'_i}{D_{p_i} \cdot D_{y'_i}} \right)^2 \quad (3.6)$$

где  $p_i$  — частоты попадания в интервал с серединой  $x_{(i)}$ ;

$y'_i$  — ордината линии регрессии в точке  $x_{(i)}$ .

Запишем результаты расчетов коэффициентов детерминации  $R^2$  анаморфоз для всех выборок в Таблицу 3.1.

Таблица 3.1 — Коэффициенты прямой регрессии, коэффициент детерминации и вывод относительно нулевой гипотезы

№ выборки	$R^2$	Вывод
1	0.95	$R^2 > 0.8$ , $H_0$ принимается, выборка распределена нормально
1	0.03	$R^2 < 0.8$ , $H_0$ отвергается, выборка распределена не по нормальному закону
2	0.36	$R^2 < 0.8$ , $H_0$ отвергается, выборка распределена не по нормальному закону
2	0.01	$R^2 < 0.8$ , $H_0$ отвергается, выборка распределена не по нормальному закону
3	0.72	$R^2 < 0.8$ , $H_0$ отвергается, выборка распределена не по нормальному закону
3	0.92	$R^2 > 0.8$ , $H_0$ принимается, выборка распределена нормально
4	0.5	$R^2 < 0.8$ , $H_0$ отвергается, выборка распределена не по нормальному закону
4	0.85	$R^2 > 0.8$ , $H_0$ принимается, выборка распределена нормально

### 3.2.3 Расчёт параметров распределений

Из коэффициентов прямой линейной регрессии найдём оценки параметров генеральной совокупности выборок.

Оценку стандартного отклонения выборок, распределённых нормально, рассчитаем по формуле (3.7):

$$\sigma = \sqrt{\frac{1}{-2k}} \quad (3.7)$$



где  $k$  — угол наклона прямой.

Оценку коэффициента  $\lambda$  экспоненциального распределения найдём по формуле (3.8):

$$\lambda = -k \quad (3.8)$$

Рассчитаем параметр распределения первой выборки:

$$\sigma = \sqrt{\frac{1}{-2 * (-0.02)}} = 5$$

Рассчитаем параметр распределения второй выборки:

$$\sigma = \sqrt{\frac{1}{-2 * (-0.017)}} = 5.42$$

Рассчитаем параметр распределения третьей выборки:

$$\lambda = -k = -(-0.28) = 0.28$$

Рассчитаем параметр распределения четвертой выборки:

$$\lambda = -k = -(-0.157) = 0.157$$

### 3.3 Вывод

В результате выполнения работы был произведен сбор данных, методом Пирсона для каждого полученного ряда данных проверено истинное распределение и ложное. Затем проверено распределение выборок при помощи метода анаморфоз. Для анаморфозов с наибольшим спрямлением построена прямая регрессия, рассчитан коэффициент детерминации, по нему сделан вывод о распределении выборки. При помощи коэффициентов линии регрессии рассчитаны оценки параметров генеральной совокупности выборки.

## **4. ЛИНЕЙНАЯ РЕГРЕССИЯ. ОЦЕНКА АДЕКВАТНОСТИ МОДЕЛИ, ОЦЕНКА ДОВЕРИТЕЛЬНЫХ ИНТЕРВАЛОВ ПАРАМЕТРОВ**

### **4.1 Постановка задачи**

Первый файл данных содержит два ряда. Для первого файла данных:

1. Оценить коэффициент корреляции Пирсона  $r(x, y)$  между двумя переменными в первом и втором столбце.
2. По шкале Чеддока оценить характеристику корреляционной связи между величинами.
3. Проверить статистическую значимость коэффициента корреляции Пирсона с помощью  $t$ -статистики.
4. Построить доверительный интервал для  $r(x, y)$  с надежностью  $\gamma = 0.95$ .
5. Построить линейную регрессию между столбцами, оценить значение коэффициентов линейной зависимости.
6. Оценить адекватность модели с использованием критерия Фишера.
7. Оценить значимость полученных коэффициентов прямой.
8. Построить доверительные интервалы для полученных коэффициентов.
9. Оценить интервал прогноза для линейной модели на  $t = 3$  значения вперед.

Второй файл содержит 4 ряда данных. Для второго файла:

1. С помощью теста Чоу обосновать необходимость деления выборки по одной из качественных факторных переменных.
2. Произвести разбиение и построить две линейных регрессии, оценить коэффициенты моделей.

Третий файл содержит 2 ряда данных. Для третьего файла:

1. Двумя способами (тест Спирмена и тест Гольдфельда-Квандта) определить, присутствует ли в данных гетероскедастичность.
2. Построить линейную регрессию, оценить значения коэффициентов модели.
3. Оценить значимость полученных коэффициентов и адекватность модели.

## 4.2 Ход выполнения работы

### 4.2.1 Оценка коэффициента корреляции Пирсона. Оценка характеристики корреляционной связи по шкале Чеддока

В первом файлом с данными, оценим коэффициент корреляции Пирсона  $r(x, y)$  между двумя переменными в первом и втором столбце.

Коэффициент линейной корреляции Пирсона между рядами данных  $x$  и  $y$  рассчитывается по формуле (4.1):

$$r(x, y) = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \sigma_y}; \overline{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i \quad (4.1)$$

В результате вычислений получен коэффициент корреляции  $r=0.98$ .

По шкале Чеддока оценим характеристику корреляционной связи между величинами.

Согласно шкале (Таблица 4.1), случайные величины, мера тесноты связи которых лежит в интервале  $[0.9, 0.99]$ , имеют высокую силу связи.

Таблица 4.1 — Шкала Чеддока

Количественная мера тесноты связи	Качественная характеристика силы связи
0.1-0.3	Слабая
0.3-0.5	Умеренная
0.5-0.7	Заметная
0.7-0.9	Высокая
0.9-0.99	Весьма высокая

$r = 0.98 \in [0.9, 0.99]$ , следовательно, между  $x$  и  $y$  существует весьма высокая сила связи.

#### 4.2.2 Проверка статистической значимости коэффициента Корреляции

Проверим статистическую значимость коэффициента корреляции Пирсона с помощью  $t$ -статистики.

Рассчитаем  $t$ -статистику для данного ряда данных и коэффициента линейной корреляции Пирсона по формуле (4.2):

$$t_r = |r| \cdot \sqrt{\frac{n-2}{1-r^2}} \quad (4.2)$$

Если полученное значение  $t$ -статистики выходит за границы интервала  $|t_r| < t(n-2)_{1-\frac{\alpha}{2}}$ , то принимается гипотеза  $H_1$ : значение коэффициента линейной корреляции Пирсона значительно отличается от нуля. Если данное значение не выходит за границы интервала, то принимается альтернативная гипотеза  $H_0$ : значение коэффициента линейной корреляции Пирсона незначительно отличается от нуля.

Вычислим значение t-статистики и получим, что  $t_r \approx 54.24$ , а критическое значения  $t(n-2)_{1-\frac{\alpha}{2}} = 1.98$ .

Значение критерия больше критического значения, принимается гипотеза  $H_1$  о том, что коэффициент линейной корреляции Пирсона значительно отличается от нуля и его значение является статистически значимым.

Построим доверительный интервал для  $r(x, y)$  с надежностью  $\gamma = 0.95$ .

Доверительный интервал для коэффициента линейной корреляции Пирсона рассчитывается по формуле (4.3):

$$th\left(\frac{1}{2} \ln \frac{1+r}{1-r} - z_{\alpha} \frac{1}{\sqrt{n-3}}\right) < r < th\left(\frac{1}{2} \ln \frac{1+r}{1-r} + z_{\alpha} \frac{1}{\sqrt{n-3}}\right) \quad (4.3)$$

Получаем доверительный интервал  $r \in (0.97, 0.98)$ .

#### 4.2.3 Построение модели линейной регрессии

Построим линейную регрессию между столбцами, оценить значение коэффициентов линейной зависимости.

Для получения оценки коэффициентов парной линейной регрессии воспользуемся следующими соотношениями (4.4):

$$a = r(x, y) \cdot \frac{\sigma_y}{\sigma_x}; b = \bar{y} - a \cdot \bar{x} \quad (4.4)$$

В результате получаем коэффициенты угла наклона  $a \approx 2.41$ , коэффициент пересечения с осью ординат  $b \approx -2.34$ .

График соответствующей прямой регрессии представлен на Рисунке 4.1:

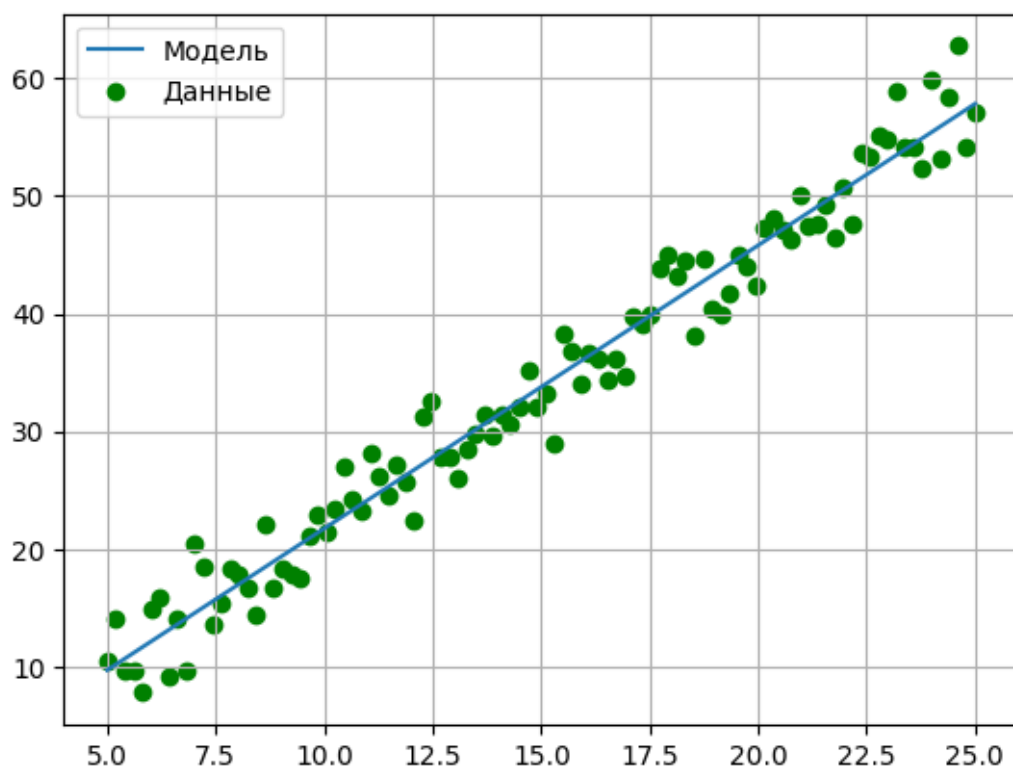


Рисунок 4.1 — График линейной регрессии

#### 4.2.4 Оценка адекватности модели

Оценим адекватность модели с использованием критерия Фишера.

Проверка значимости полученной модели называется проверкой адекватности. Одним из способов проверки значимости линейной модели регрессии является использование критерия Фишера, который заключается в расчёте  $F(n - 2, n - 1)$ -распределенной статистике, определенной по формуле (4.5):

$$F = \frac{S_{LR}^2}{S_{tot}^2}; S_{LR}^2 = \sum_{i=1}^n \frac{(y_i - \hat{y}(x_i))^2}{n - 2}; S_{tot}^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1} \quad (4.5)$$

Если полученное значение F-статистики равно или выше критического значения  $F(n - 2, n - 1)_{1 - \alpha/2}$  при заданном уровне значимости  $\alpha$ , то модель

признаётся неадекватной и принимается гипотеза  $H_1$ , в альтернативном случае принимается гипотеза  $H_0$ , и модель признаётся адекватной.

В результате расчёта получили значение критерия  $F \approx 30.7$ . Критическое значения при заданном уровне значимости равно  $F(98, 99)_{0.975} = 0.72$ . Значение критерия меньше критического значения, вследствие чего принимается нулевая гипотеза  $H_0$  о том, что полученная модель линейной регрессии адекватна.

#### 4.2.5 Оценка значимости коэффициентов модели

Оценим значимость полученных коэффициентов прямой.

В основе проверки лежит гипотеза о равенстве параметров нулю. Для линейной регрессии рассчитываются следующие величины (4.6):

$$m_a = \frac{s_{LR}}{\sigma_x \cdot \sqrt{n}}; m_b = \frac{s_{LR} \cdot \sqrt{\sum_{i=1}^n x_i^2}}{n \cdot \sigma_x}; s_{LR} = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}(x_i))^2}{n - 2}} \quad (4.6)$$

Тогда статистики, рассчитываемые согласно формулам (4.7), имеют t-распределение с  $df = n - 2$  степенями свободы.

$$T_a = \frac{a}{m_a}; T_b = \frac{b}{m_b} \quad (4.7)$$

В связи с данным определением статистик нулевая гипотеза об отсутствии статистической значимости коэффициентов регрессии принимается при условии (4.8):

$$|T_a| \leq t(n - 2)_{1-\frac{\alpha}{2}}; |T_b| \leq t(n - 2)_{1-\frac{\alpha}{2}} \quad (4.8)$$

Если рассчитанный критерий по модулю больше критического значения, то соответствующий коэффициент является статистически значимым.

В результате расчётов  $T_a \approx 54.24$ ,  $T_b \approx -3.28$ . Критическое значение для обоих коэффициентов одинаково и равно  $T_{cv} \approx 1.98$ .

Оба рассчитанных критерия по модулю превышают критическое значение, следовательно, коэффициенты регрессии признаются статистически значимыми.

#### **4.2.6 Построение доверительных интервалов коэффициентов модели**

Построим доверительные интервалы для полученных коэффициентов.

Доверительные интервалы параметров рассчитываются для каждого из параметров рассчитываются согласно формуле (4.9):

$$\begin{cases} \hat{a} \in \left( \alpha - m_a \cdot t(n-2)_{1-\frac{\alpha}{2}}, \alpha + m_a \cdot t(n-2)_{1-\frac{\alpha}{2}} \right), \\ \hat{b} \in \left( b - m_b \cdot t(n-2)_{1-\frac{\alpha}{2}}, b + m_b \cdot t(n-2)_{1-\frac{b}{2}} \right) \end{cases} \quad (4.9)$$

В результате расчётов были получены следующие интервалы для коэффициентов модели линейной регрессии:  $\hat{a} \in (2.32, 2.50)$ ,  $\hat{b} \in (-3.76, -0.93)$ .

Модели, отвечающие граничным значениям полученных интервалов, представлены на Рисунке 4.2:



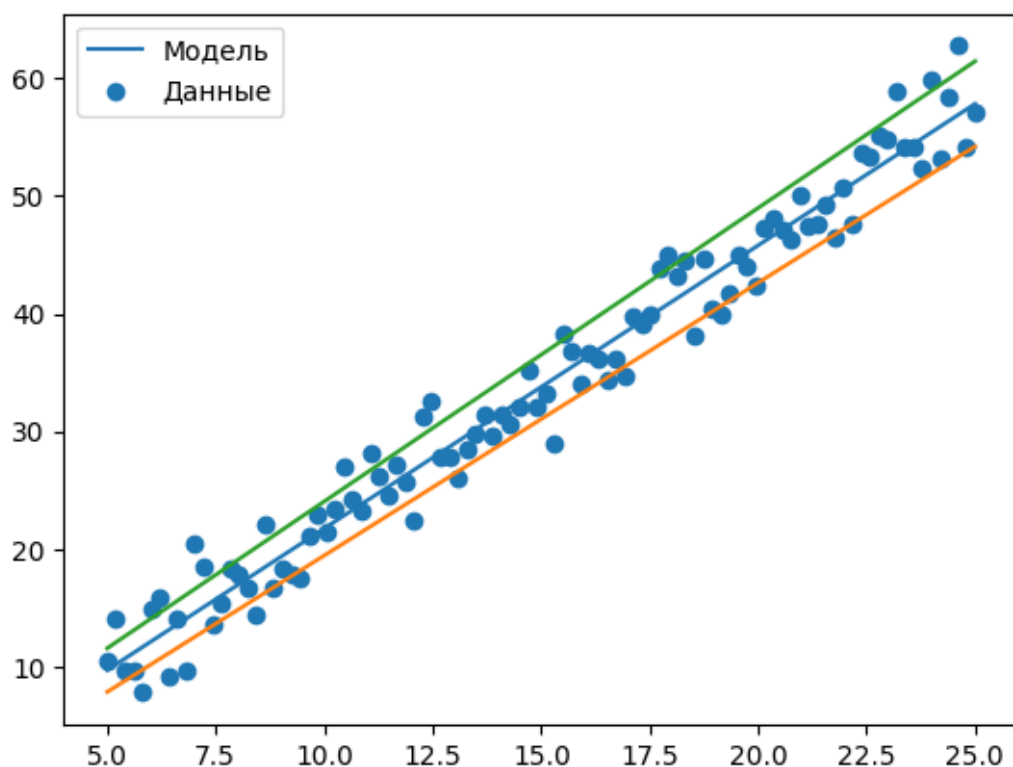


Рисунок 4.2 — Графики моделей линейной регрессии с граничными значениями параметров

#### 4.2.7 Оценка интервала прогноза линейной модели

Оценим интервал прогноза для линейной модели на  $dt = 3$  значения вперед.

Нам необходимо определиться со значением  $x$  для которого будет получено прогнозное значение  $y$  по модели  $y(x)$ :  $x \approx 25.6$ . Получившееся значение будет лежать в интервале с значимостью  $\alpha$ , который будет рассчитан далее.

Интервальные оценки для прогноза линейной моделью регрессии рассчитываются следующим образом (4.10):

$$\begin{cases} \hat{y} \in (\hat{y}(x) - E, \hat{y}(x) + E) \\ E = t(n-2)_{1-\frac{\alpha}{2}} \cdot S_{LR} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{n \cdot \sigma_x^2}} \end{cases} \quad (4.10)$$

В результате получена следующая интервальная оценка прогноза:  
 $\hat{y} \in (53.66, 64.92)$ .

#### 4.2.8 Проведение теста Чоу

Во втором файле с данными, с помощью теста Чоу обоснуем необходимость деления выборки по одной из качественных факторных переменных.

Для всех моделей в тесте считается  $RSS$  классическим образом, для первой модели – сумма остатков по данным, которые доступны только ей, для второй модели аналогично, для общей модели – сумма квадратов остатков по всем данным. Далее рассчитывается  $F$ -статистика по формуле (4.11):

$$F_{chow} = \frac{(RSS - RSS_1 - RSS_2) / k}{(RSS_1 + RSS_2) / (n - 2k)} \quad (4.11)$$

где  $RSS_1$  и  $RSS_2$  — сумма квадратов остатков первой и второй модели соответственно;

$RSS$  — сумма квадратов остатков стандартной модели на общей выборке;

$k$  — количество параметров модели линейной регрессии.

Разделим выборку по первому категориальному признаку.

#### 4.2.9 Построение модели линейной регрессии. Оценка значимости коэффициентов и адекватности модели

Потом построим для каждой из полученных подвыборок модель линейной регрессии. График моделей линейной регрессии представлен на Рисунке 4.3.

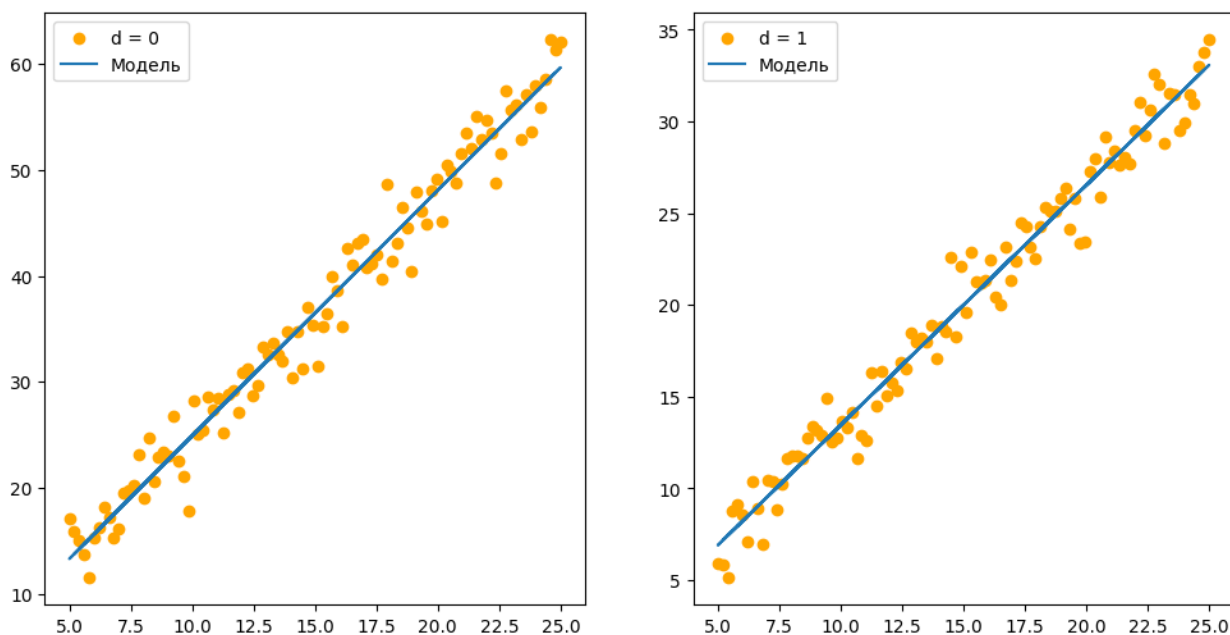


Рисунок 4.3 — График моделей линейной регрессии

Коэффициенты полученных моделей представлены в Таблице 4.2:

Таблица 4.2 — Коэффициенты построенных моделей

cat1	$w_1$	$b$
0	2.32	1.70
1	1.31	0.36

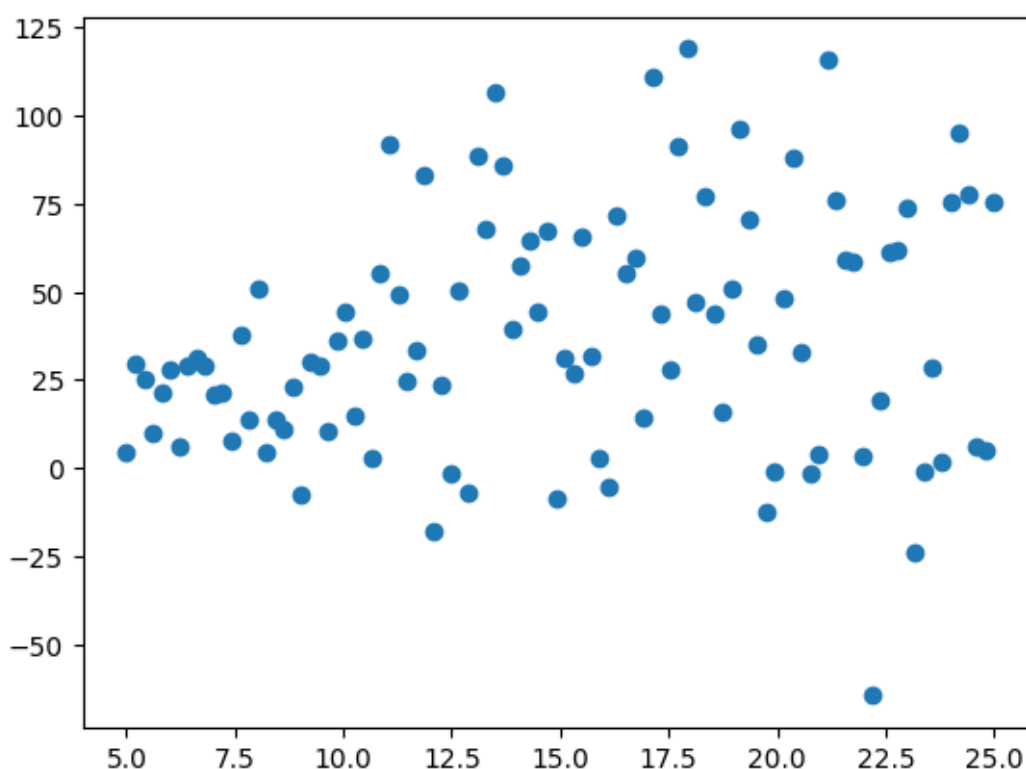
Рассчитаем критическое значение по распределению Фишера  $F(k, n - 2k)_{0.95}$  с заданным уровнем значимости .

$F_{cv} \approx F(2, 196)_{0.95} \approx 3.042$ .  $F_{chow} \approx 2240.60$ . Полученное значение критерия больше критического значения, следовательно, принимается альтернативная гипотеза  $H_1$  о разнородности выборок и необходимости строить две разные модели  $\hat{y}_1(x)$  и  $\hat{y}_2(x)$ , разбивая выборки.

#### 4.2.10 Проверка данных на гетероскедастичность при помощи теста Гольдфельда-Квандта и теста Спирмена

В третьем файле с данными, определяем двумя способами (тест Спирмена и тест Гольдфельда-Квандта), присутствует ли в данных гетероскедастичность.

Построим график зависимости показательной переменной от факторной (Рисунок 4.4).



**Рисунок 4.4 — Зависимость показательной переменной от факторной**

Отсортируем выборку по факторной переменной. Полученную выборку разделим на две подвыборки с 0-го по  $m_1$  элемент и с  $n - m_2$  по  $n$  элемент соответственно ( $m_1 = m_2 = 3 * n / 8$ ). Построим модели линейной регрессии для данных выборок. Графики полученных моделей представлены на Рисунках 4.5-4.6.

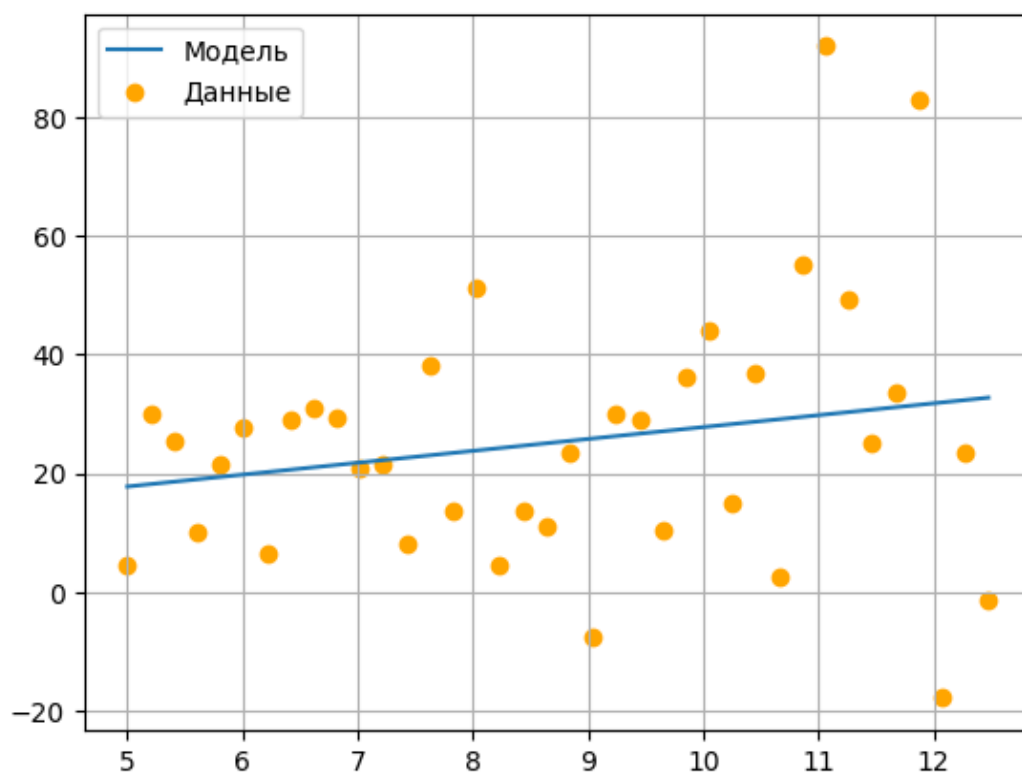


Рисунок 4.5 — Модель линейной регрессии для первой подвыборки

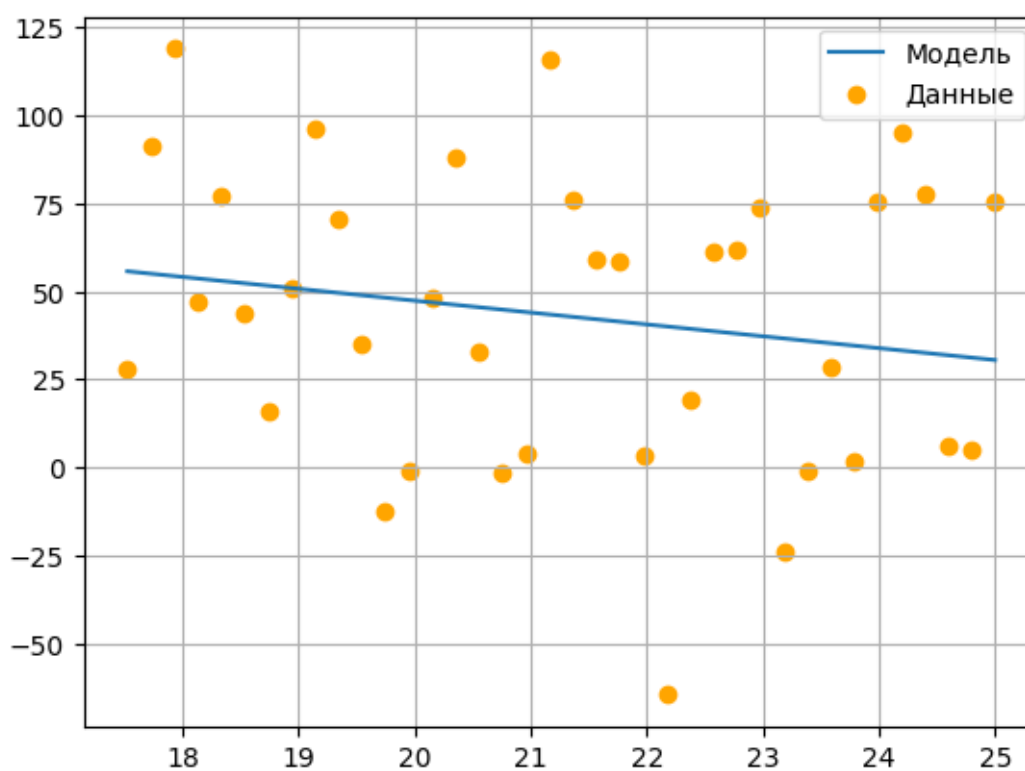


Рисунок 4.6 — Модель линейной регрессии для второй подвыборки

Для полученных двух оценок регрессионной модели находят суммы квадратов остатков и рассчитывают F-статистику, равную отношению большей суммы квадратов остатков к меньшей (4.12):

$$F = \frac{\sum_{i=1}^{m_1} (\hat{y}_1(x_i) - y_i)^2 / (m_1 - k)}{\sum_{i=n-m_2+1}^n (\hat{y}_2(x_i) - y_i)^2 / (m_2 - k)} \quad (4.12)$$

где  $k$  — число факторных (объясняющих) переменных в линейной зависимости,

$\hat{y}_1(x_i)$  — модель на первых  $m_1$  записях отсортированных данных по объясняющей переменной;

$\hat{y}_2(x_i)$  — модель на последних  $m_2$  записях отсортированных данных по объясняющей переменной.

Данный критерий имеет распределение Фишера  $F(m_1 - k, m_2 - k)$ . Если подсчитанная статистика по значению больше критического значения распределения Фишера с заданными степенями свободы и уровнем значимости  $F(m_1 - k, m_2 - k)_{1-\alpha/2}$ , то нулевая гипотеза отвергается и гетероскедастичность имеет место для заданной линейной зависимости.

Для нашей выборки получен критерий  $F \approx 4.51$ . Критическое значение равно  $F_{cv} = F(37, 37)_{0.95} \approx 1.74$ .  $F > F_{cv}$ , из чего следует, что мы принимает гипотезу  $H_1$  о том, что в данных присутствует гетероскедастичность.

Тест ранговой корреляции Спирмена — непараметрический статистический тест, позволяющий проверить гетероскедастичность случайных ошибок регрессионной модели. Особенность теста заключается в том, что не конкретизируется форма возможной зависимости дисперсии случайных ошибок модели от той или иной переменной.

Для модели линейной регрессии  $\hat{y}(x) = a \cdot x + b$ , обученной на всей выборке, необходимо рассчитать остатки  $e_i = y_i - \hat{y}(x_i)$ ,  $i = 1, 2, \dots, n$  ( $n$  — количество элементов в выборке).

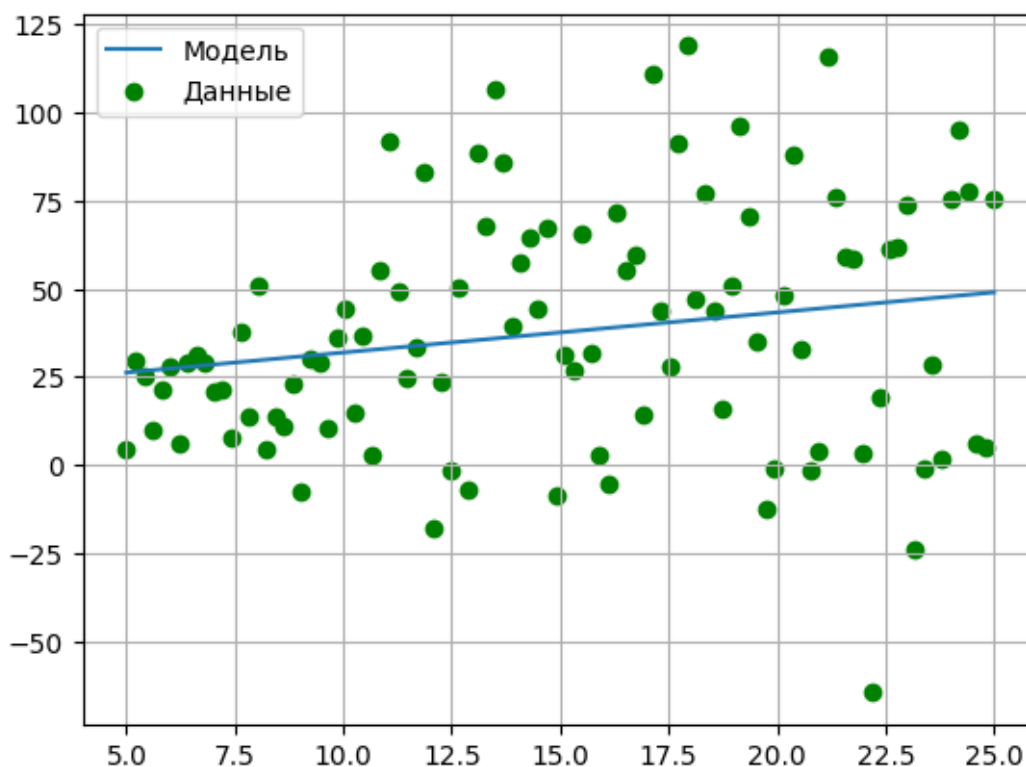
Для ранжированных данных ошибок  $e_i$  и факторной переменной  $x_i$  производится подсчёт коэффициента ранговой корреляции Спирмена (4.13):

$$r_s(e, x) = 1 - \frac{6 \cdot \sum_{i=1}^n (\text{rank}(e_i) - \text{rank}(x_i))^2}{n \cdot (n^2 - 1)} \quad (4.13)$$

Если значение статистики  $r_s(e, x)\sqrt{(n-1)}$  по модулю превышает критическое значение стандартного нормального распределения  $N(0, 1)$  с заданным уровнем значимости  $\alpha$ , т. е.  $\left| r_s(e, x)\sqrt{(n-1)} \right| > N(0, 1)_{1-\frac{\alpha}{2}}$ , то гетероскедастичность признается значимой и гипотеза  $H_1$  принимается. В случае, если значение данной статистики находится в пределах критического значения стандартного нормального распределения, то принимается нулевая гипотеза  $H_0$  об отсутствии гетероскедастичности.

В результате расчётов получаем  $R = r_s(e, x)\sqrt{(n-1)} \approx 0.05$ . Критическое значение равно  $N_{cv} = N(0, 1)_{0.975} \approx 1.98$ .  $|R| < N_{cv}$ , в силу чего принимается нулевая гипотеза  $H_0$  о том, что в данных отсутствует гетероскедастичность.

Построим линейную регрессию, оценим значения коэффициентов модели.



**Рисунок 4.7 — График модели линейной регрессии**

График полученной модели линейной регрессии представлен на Рисунке 4.7. Угловой коэффициент  $a \approx 1.14$ , смещение  $b \approx 20.60$ .

Оценим значимость полученных коэффициентов и адекватность модели.

Проверим значимость коэффициентов. В результате расчётов  $T_a \approx 1.95$ ,  $T_b \approx 2.20$ . Критическое значение для обоих коэффициентов одинаково и равно  $T_{cv} \approx 1.98$ .

Оба рассчитанных критерия меньше критического значения, следовательно, коэффициенты регрессии признаются статистически незначимыми.

Проверим адекватность модели. В результате расчёта получили значение критерия  $F \approx 1.03$ . Критическое значения при заданном уровне значимости равно  $F(98, 99)_{0.975} = 0.72$ . Значение критерия меньше критического значения, вследствие чего принимается нулевая гипотеза  $H_0$  о том, что полученная модель линейной регрессии адекватна.



### **4.3 Вывод**

В результате выполнения работы была осуществлена оценка связи между значениями первого набора данных, проведен тест Чоу для второй выборки, на основании результатов которого было осуществлено деление на подвыборки относительно категориального признака, а также осуществлена проверка третьей выборки на гетероскедастичность при помощи тестов Спирмена и Голдфельда–Квандта. Для каждой выборки построена модель линейной регрессии, проведена оценка статистической значимости параметров и выполнена проверка модели на адекватность.

## 5. СГЛАЖИВАНИЕ ВРЕМЕННЫХ РЯДОВ

### 5.1 Постановка задачи

В папке два файла, которые содержат разные временные ряды. В первом файле находится ряд с синусоидальным трендом. Во втором - с линейным.

Необходимо выделить тренд используя 4 метода:

- простое скользящее среднее (SMA);
- взвешенное скользящее среднее (WMA) особого типа;
- экспоненциальное сглаживание (EMA);
- двойное экспоненциальное сглаживание (DEMA).

Каждый метод требует подбора некоторых параметров:

1. SMA и WMA - размер окна,
2. EMA - параметр сглаживания  $A$ ,
3. DEMA - параметр сглаживания вокруг тренда  $A$  и параметр сглаживания самого тренда  $B$ .

Для весов в WMA использовать экспоненциальную весовую функцию.

Необходимо подобрать оптимальные значения соответствующих параметров, используя Q-статистику Льюнг-Бокса при  $m = 5$ . Оптимальными параметрами будем считать те, что минимизируют приведенную статистику.

В качестве размеров окна  $w = 2 * m + 1$  перебрать значения  $m = 3, 5, 7, 9$ ; в качестве параметров сглаживания:  $\alpha, \gamma = 0.1, 0.2, \dots, 0.9$ . Обратите внимание, что метод DEMA двухпараметрический, что требует выбрать оптимальную комбинацию сразу двух параметров  $\alpha, \gamma$ .

После подбора оптимальных параметров провести тест Дарбина-Уотсона ( $m = 1, \alpha = 0.95$ ) на данных после исключения выделенного тренда для каждого метода и каждого ряда.

Изобразить графики исходных данных, графики трендов при оптимальных параметрах у каждого метода для каждого ряда, расчетные формулы, а также результаты тестов Дарбина-Уотсона.

## 5.2 Ход выполнения работы

Импортированы данные. На Рисунке 5.1 представлена визуализация данных из первого файла.

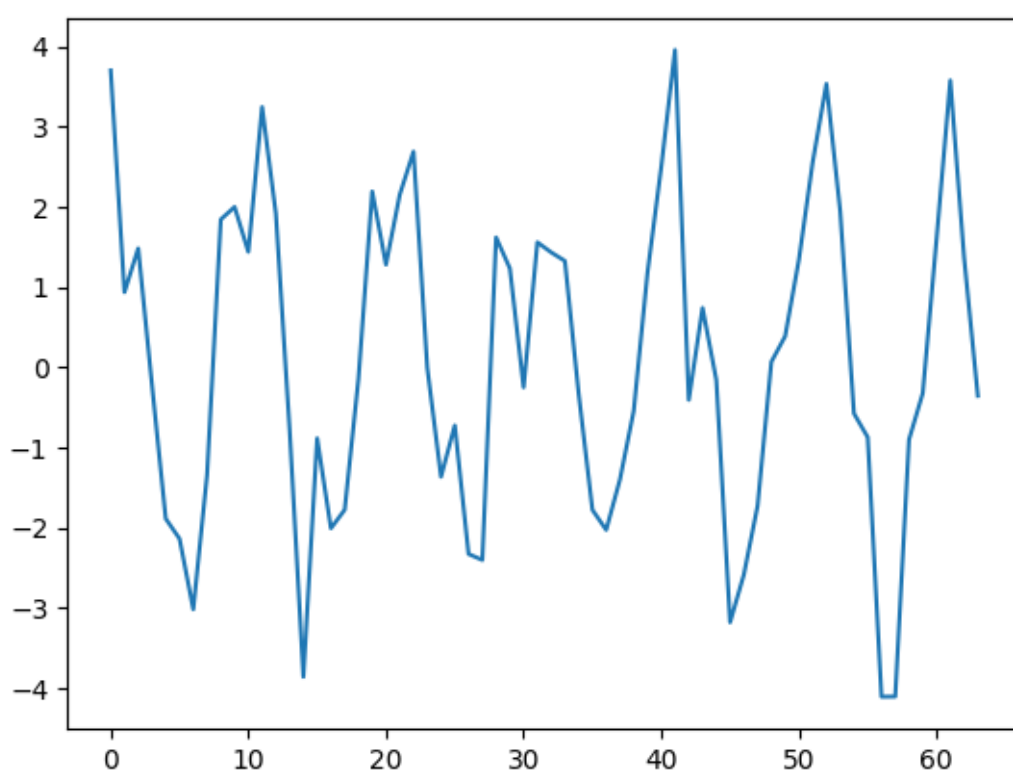


Рисунок 5.1 — Визуализация данных с синусоидальным трендом

На Рисунке 5.2 представлена визуализация данных из второго файла.

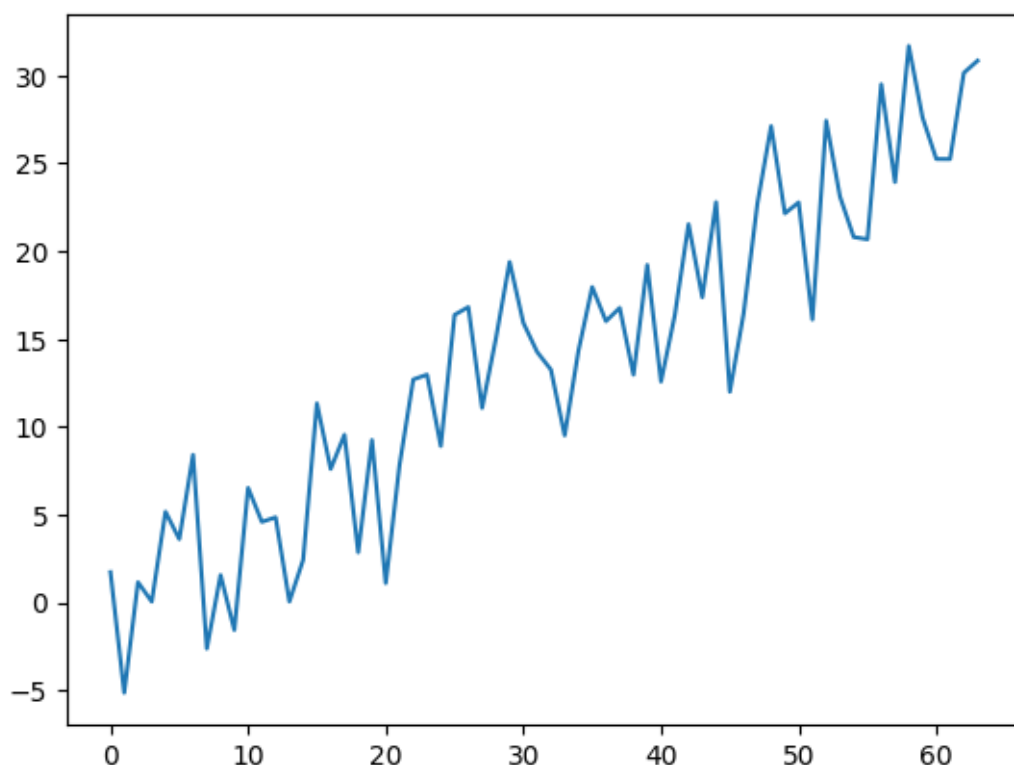


Рисунок 5.2 — Визуализация данных с линейным трендом

Простое скользящее среднее:

Метод простого скользящего среднего, с размером окна  $w = 2 \cdot m + 1$ , где  $m$  – количество членов ряда в сумме по одной стороне от центрального значения, является частным случаем метода взвешенного скользящего среднего с равными весовыми коэффициентами (5.1):

$$\tilde{y}_t = \sum_{i=-m}^m \omega_i y_{t+i}, \quad \omega_i = \frac{1}{2 \cdot m + 1} \quad (5.1)$$

где  $y_t$  – исходные значения временного ряда в дискретных отсчётах  $t$ ;

$\omega_i$  – весовые коэффициенты окна сглаживания;

$\tilde{y}_t$  – сглаженный ряд данных  $y_t$ .

Найдем оптимальный параметр сглаживания с помощью Q-статистики Льюнг-Бокса (5.2):

$$Q = n(n+2) \sum_{k=1}^m \frac{r^2(k)}{n-k} \quad (5.2)$$

Где  $r$  – выборочная оценка автокорреляционной функции (5.3):

$$r(k) = \frac{(n-k) \sum_{t=1}^{n-k} x_t x_{t+k} - \sum_{t=1}^{n-k} x_t \sum_{t=1}^{n-k} x_{t+k}}{\sqrt{(n-k) \sum_{t=1}^{n-k} x_t^2 - (\sum_{t=1}^{n-k} x_t)^2} \sqrt{(n-k) \sum_{t=1}^{n-k} x_{t+k}^2 - (\sum_{t=1}^{n-k} x_{t+k})^2}} \quad (5.3)$$

Оптимальное значение параметра  $m$ , найденное с использованием Q-статистики Льюнг-Бокса равно 3.

Проведем тест Дарбина-Уотсона, он является одним из самых распространенных методов для выявления линейной автокорреляции остатков первого порядка.

Считается следующая статистика (5.4):

$$d = \frac{\sum_{i=2}^n (\varepsilon_i - \varepsilon_{i-1})^2}{\sum_{i=1}^n \varepsilon_i^2} \quad (5.4)$$

Определяются критические значения  $d_L$  и  $d_U$  по специальным таблицам.

### 5.2.1 SMA-сглаживание

График сглаживания для первого временного ряда при разных значениях  $m$  представлен на Рисунке 5.3.

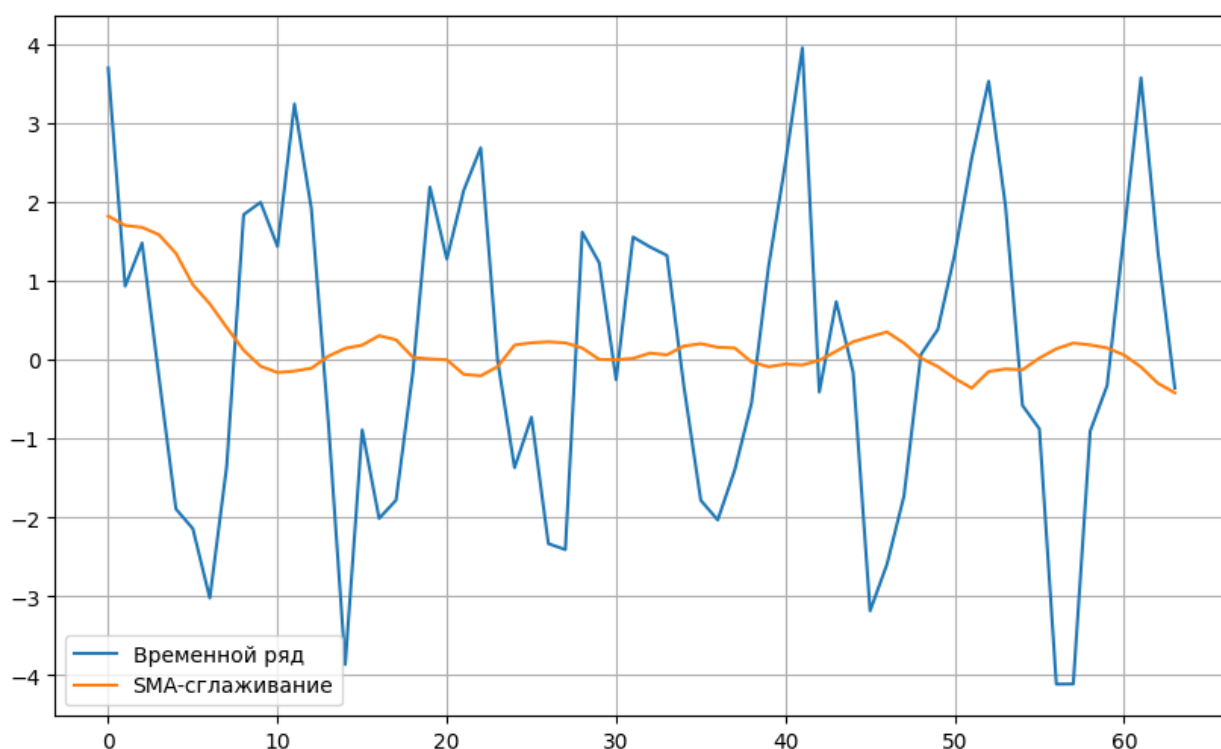


Рисунок 5.3 — SMA для первого ряда

Оптимальное значение  $m$  равно 3.

В результате теста Дарбина-Уотсона при  $m = 1$ ,  $\alpha = 0.95$  мы отклоняем гипотезу  $H_0$  о том, что автокорреляция остатков отсутствует (автокорреляция присутствует),  $d = 0.72 < d_U = 1.767 < 4 - d_U = 2.233$ .

График сглаживания для второго временного ряда при различных значениях  $m$  представлен на Рисунке 5.4.

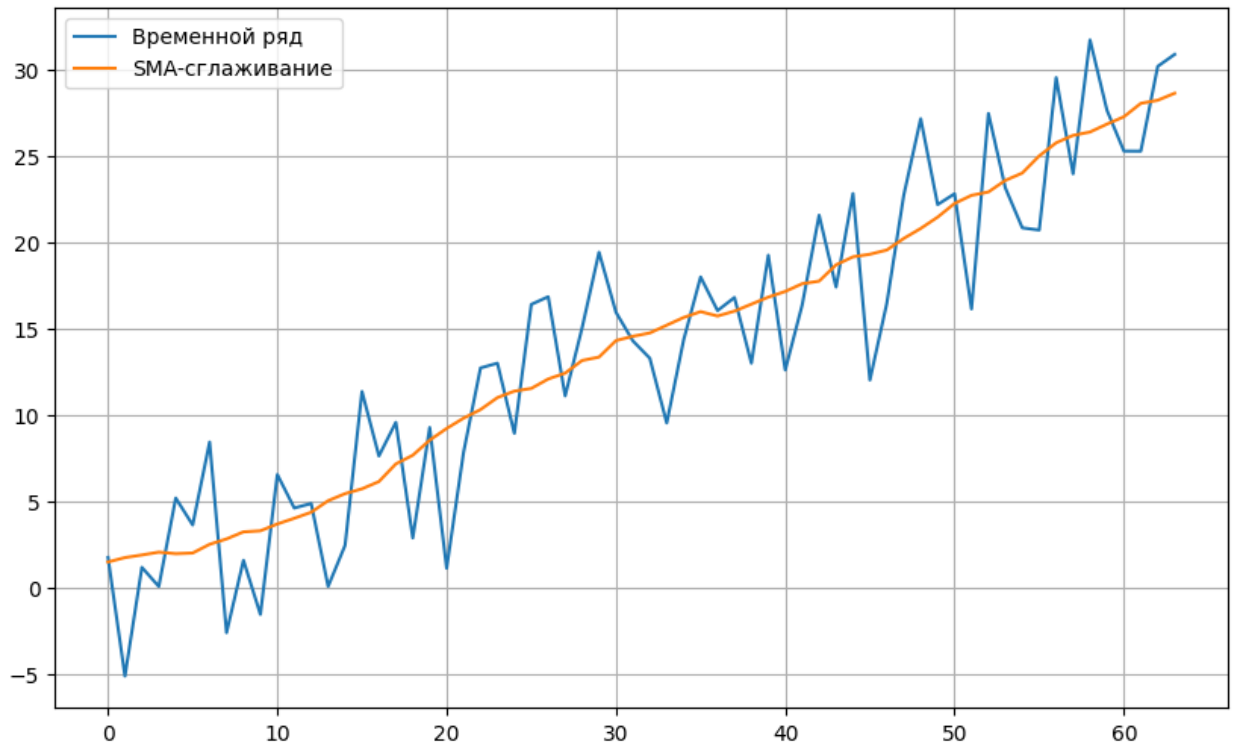


Рисунок 5.4 — SMA для второго ряда

Оптимальное значение  $m$  равно 4.

В результате теста Дарбина-Уотсона при  $m = 1$ ,  $\alpha = 0.95$  принимаем гипотезу о том, что автокорреляция присутствует, так как  $d = 2.117 < 4 - d_U = 2.37 < 4 - d_L = 3.43$ .

### 5.2.2 WMA-сглаживание

Взвешенное скользящее среднее:

Метод взвешенного скользящего среднего работает идентично методу простого скользящего среднего, за исключением необходимости определять саму весовую функцию метода сглаживания. Весовая функция – метод

определения значений весов исходя из определенного правила отображения номера элемента окна в его значение.

В данной работе используется экспоненциальная весовая функция:

$$\omega_i = \frac{e^{-\varepsilon \cdot |i|}}{\sum_{j=-m}^m e^{-\varepsilon \cdot |j|}}; i = -m, (-m+1), \dots, m; \varepsilon = 0.3.$$

График сглаживания для первого временного ряда при различных значениях  $m$  представлен на Рисунке 5.5.

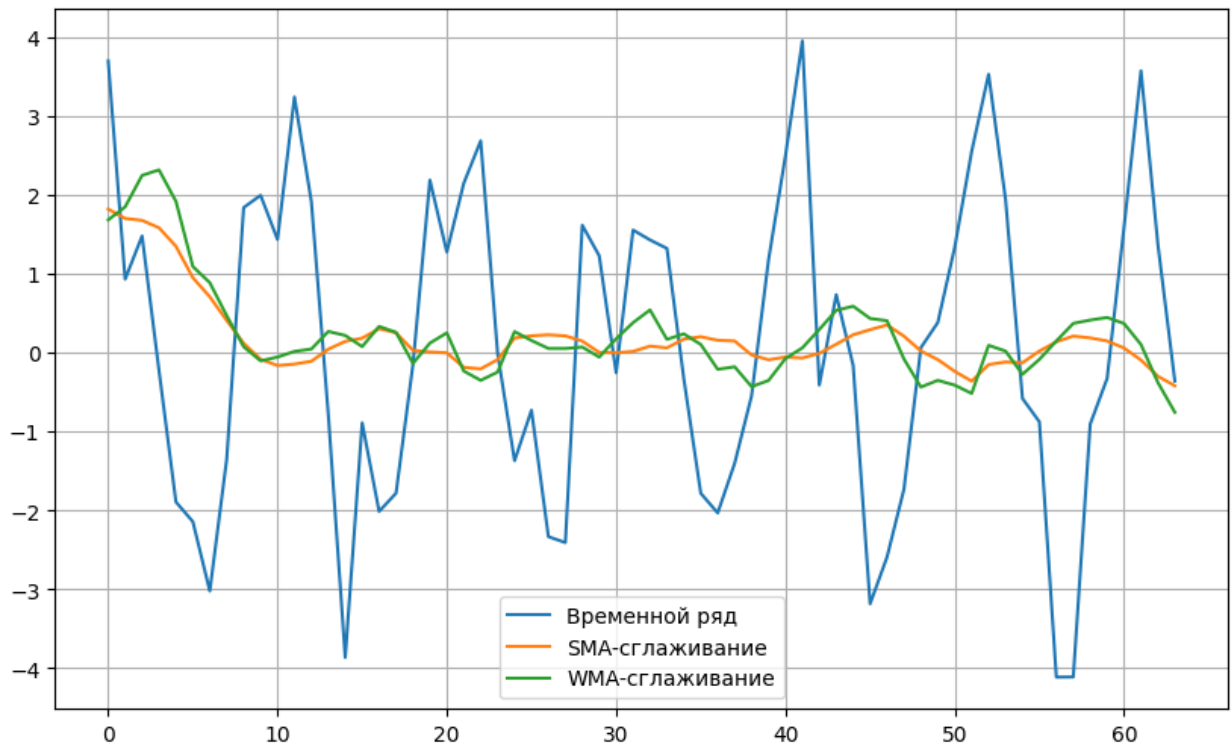
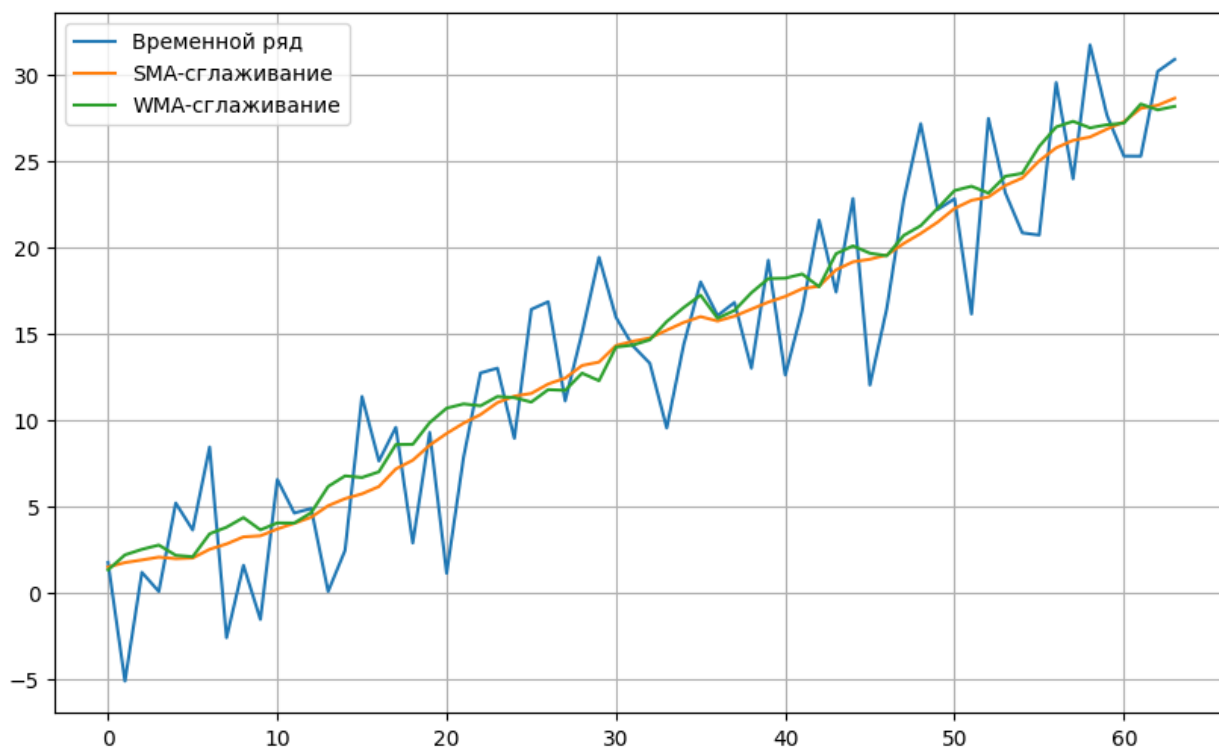


Рисунок 5.5 — WMA для первого ряда

Оптимальное значение  $m$  равно 3.

В результате теста Дарбина-Уотсона при  $m = 2$ ,  $\alpha = 0.95$  мы отклоняем гипотезу об отсутствии автокорреляции (выявлена положительная автокорреляция), так как  $d = 0.71 < d_U = 0.715 < 4 - d_U = 3.285$ .

График сглаживания для второго временного ряда при различных значениях  $m$  представлен на Рисунке 5.6.



**Рисунок 5.6 — WMA для второго ряда**

Оптимальное значение  $m$  равно 9.

В результате теста Дарбина-Уотсона при  $m = 2$ ,  $\alpha = 0.95$  мы принимаем альтернативную гипотезу  $H_1$  о существовании отрицательной автокорреляции, так как  $d = 1.911 > d_L = 1.675$ .

### 5.2.3 ЕМА-сглаживание

Экспоненциальное сглаживание:

При таком сглаживании усреднение ведется не по окну фиксированного размера, а по всему ряду от начала до текущего момента, при этом веса, с которыми учитываются давние измерения, убывают экспоненциально.

Экспоненциальное скользящее среднее вычисляется по следующей рекуррентной формуле (5.5):

$$\tilde{y}_t = \alpha y_t + (1 - \alpha) \tilde{y}_{t-1} \quad (5.5)$$



где  $\alpha$  – коэффициент сглаживания (сила сглаживания) принимает значения в диапазоне от 0 до 1 в действительной области.

Коэффициент  $\alpha$  влияет на степень восприятия истории. Чем ниже значение коэффициента, тем сильнее происходит именно сглаживание.

Покажем работу алгоритма сглаживания на обоих рядах при различных значениях  $\alpha$ .

График сглаживания для первого временного ряда при различных значениях  $\alpha$  представлен на Рисунке 5.7.

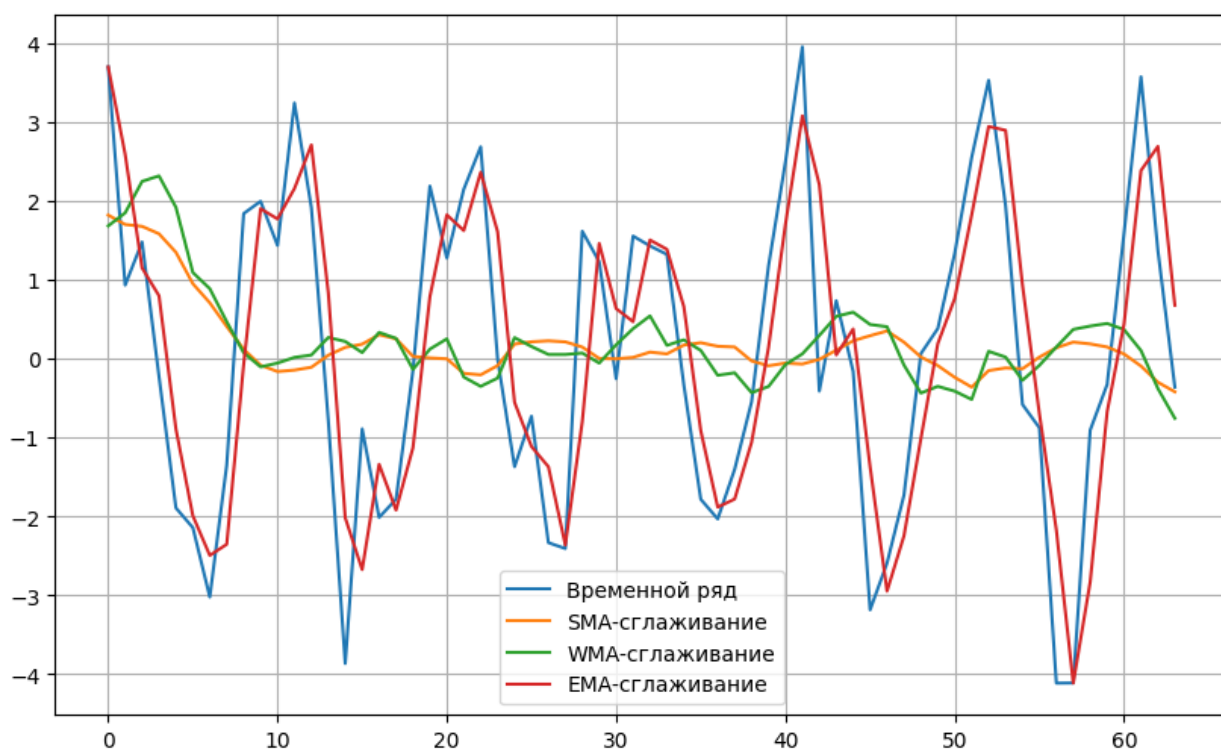


Рисунок 5.7 — ЕМА для первого ряда

Оптимальное значение параметра  $\alpha$ , найденное с использованием Q-статистики Льюнг-Бокса равно 0.9.

В результате теста Дарбина-Уотсона при  $m = 2$ ,  $\alpha = 0.95$  принимается гипотеза  $H_0$  о том, что отсутствует автокорреляция остатков, так как  $d_U = 1.672 < d = 2.99 < 4 - d_U = 2.328$ .

График сглаживания для второго временного ряда при различных значениях  $\alpha$  представлен на Рисунке 5.8.

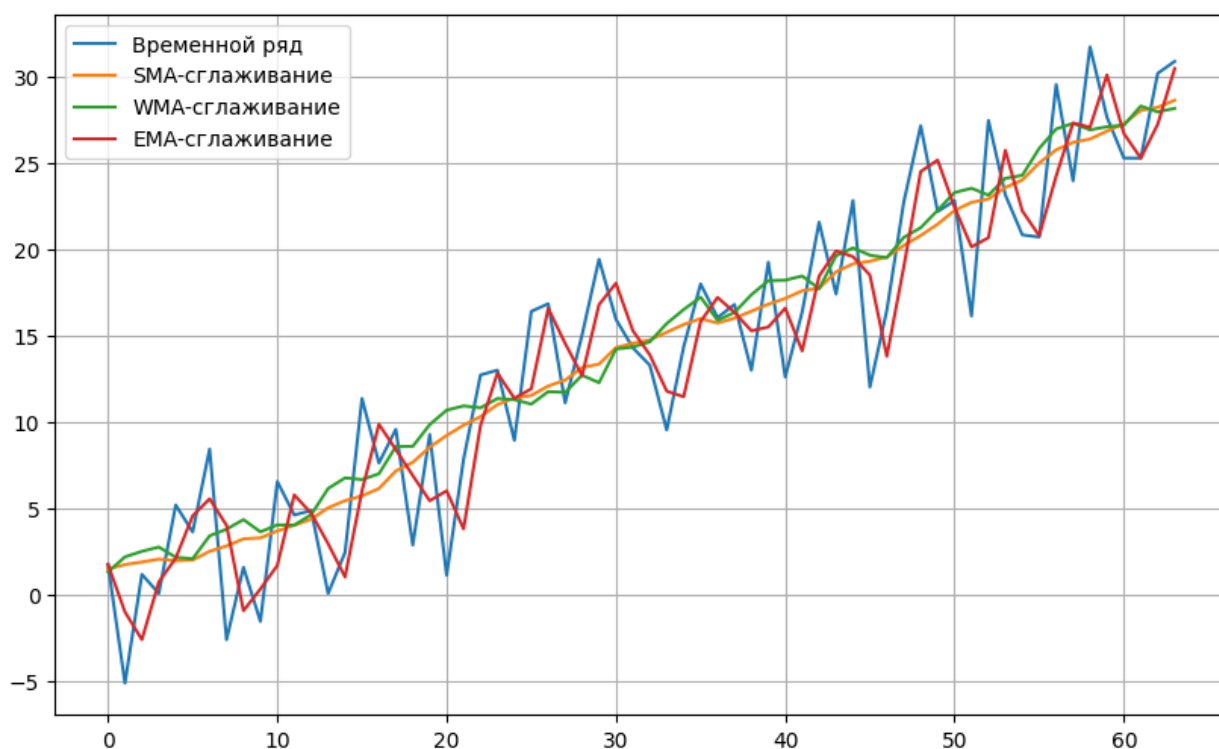


Рисунок 5.8 — ЕМА для второго ряда

Оптимальное значение параметра  $\alpha$ , найденное с использованием Q-статистики Льюнг-Бокса равно 0.9.

В результате теста Дарбина-Уотсона при  $m = 2$ ,  $\alpha = 0.95$  мы принимаем альтернативную гипотезу  $H_1$  о том, что существует положительная автокорреляция остатков, так как  $d = 2.99 > d_L = 1.54$ .

#### 5.2.4 ДЕМА-сглаживание

Двойное экспоненциальное сглаживание:

Двойное экспоненциальное сглаживание осуществляется по следующим формулам с коэффициентами  $\alpha$ ,  $\gamma$ , варьирующихся в пределах от 0 до 1 в действительной оси.

$$\begin{aligned}\tilde{y}_t &= \alpha y_t + (1 - \alpha)(\tilde{y}_{t-1} + b_{t-1}), \\ b_t &= \gamma(\tilde{y}_t + \tilde{y}_{t-1}) + (1 - \gamma)b_{t-1}, \\ \tilde{y}_1 &= y_1, b_1 = y_2 - y_1\end{aligned}$$

Продemonстрируем работу алгоритма сглаживания на обоих рядах при различных значениях  $\alpha$  и  $\gamma$ .

График сглаживания для первого временного ряда при различных значениях  $\alpha$  и  $\gamma$  представлен на Рисунке 5.9.

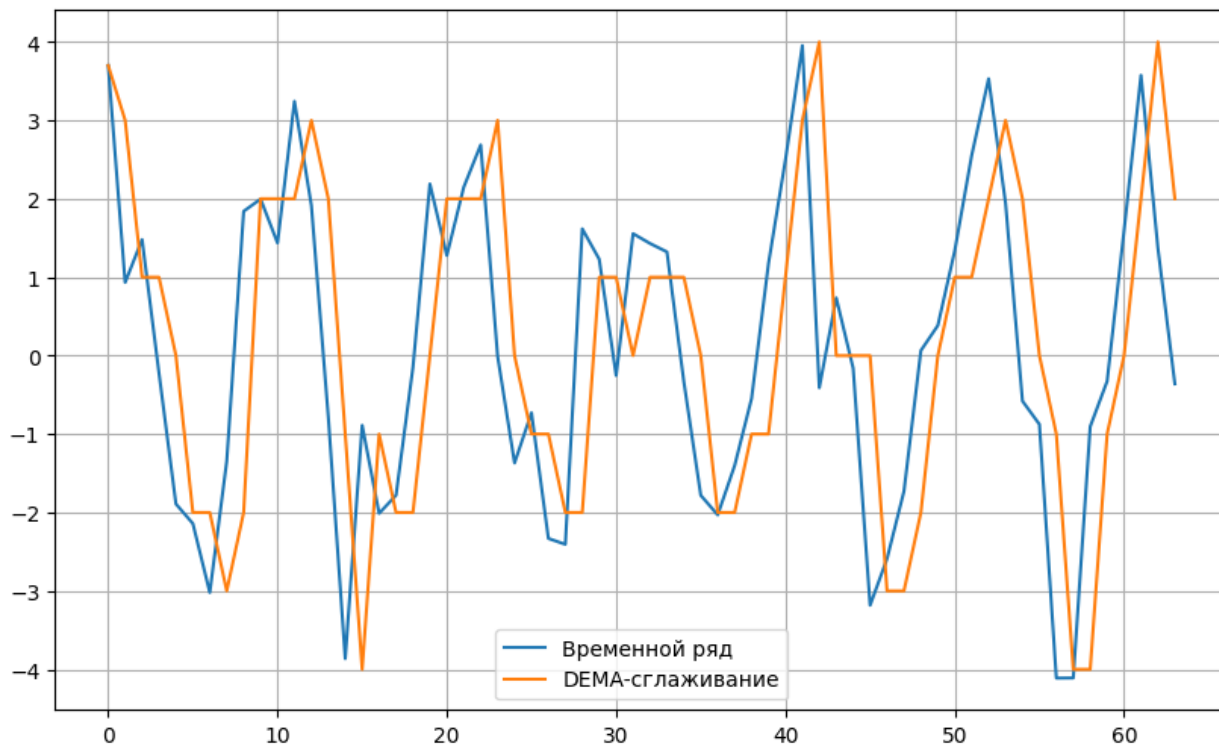
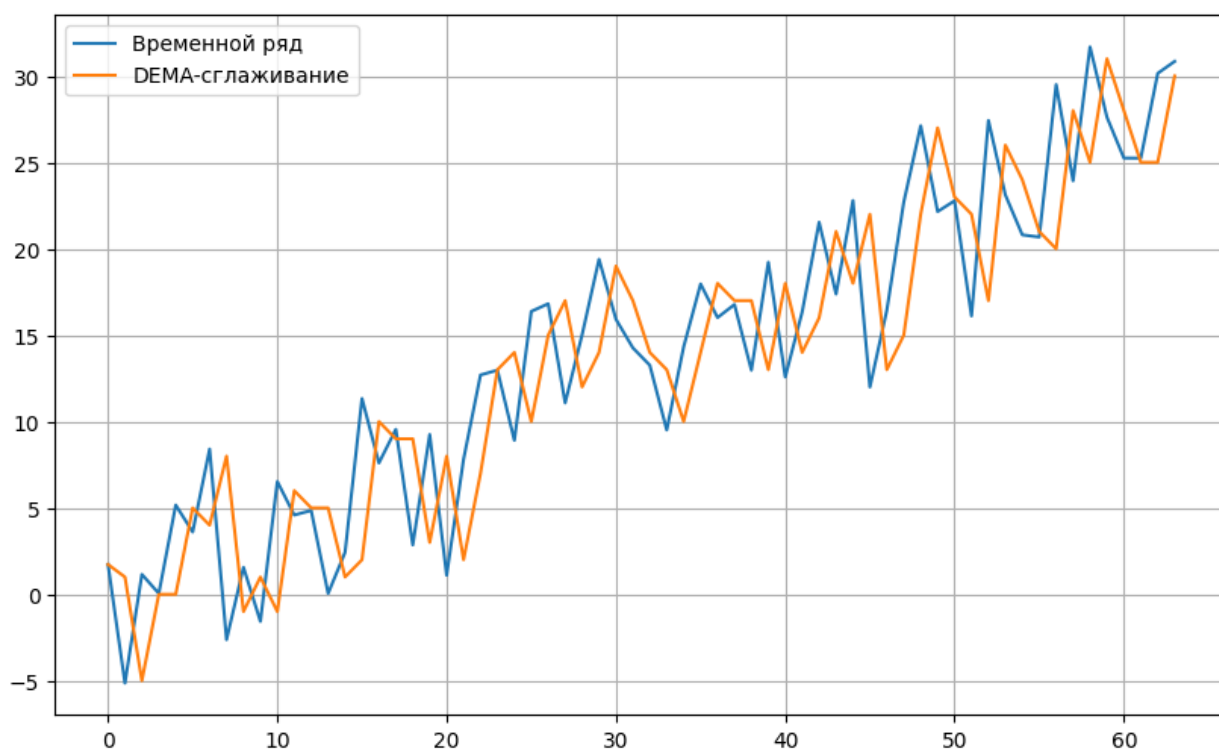


Рисунок 5.9 — DEMA для первого ряда

Оптимальные значения параметров  $\alpha = 0.9$ ,  $\gamma = 0.2$ , найдены методом перебора и выявлением минимального значения MSE.

В результате теста Дарбина-Уотсона при  $m = 2$ ,  $\alpha = 0.95$  мы отвергаем гипотезу  $H_0$  о том, что отсутствует автокорреляция остатков (автокорреляция остатков присутствует), так как  $d_U = 1.672 < 4 - d_U = 2.328 < d = 1.714$ .

График сглаживания для второго временного ряда при различных значениях  $\alpha$  и  $\gamma$  представлен на Рисунке 5.10.



**Рисунок 5.10 — DEMA для второго ряда**

Оптимальные значения параметров  $\alpha = 0.9$ ,  $\gamma = 0.1$ , найдены методом перебора и выявлением минимального значения MSE.

В результате теста Дарбина-Уотсона при  $m = 2$ ,  $\alpha = 0.95$  мы отвергаем гипотезу  $H_0$  о том, что отсутствует автокорреляция остатков (автокорреляция остатков присутствует), так как  $d_U = 1.735 < 4 - d_U = 2.265 < d = 2.77$ .

### 5.3 Вывод

В ходе выполнения работы были изучены основные методы сглаживания временных рядов и их визуализация на графиках. Были выполнены следующие задачи:

1. Выделен тренд, используя 4 метода:
  - простое скользящее среднее (SMA);
  - взвешенное скользящее среднее (WMA) особого типа;
  - экспоненциальное сглаживание (EMA);
  - двойное экспоненциальное сглаживание (DEMA).

2. Подобраны оптимальные значения соответствующих параметров, используя Q-статистику Льюнг-Бокса.
3. После подбора оптимальных параметров проведен тест Дарбина-Уотсона на данных после исключения выделенного тренда для каждого метода и каждого ряда.
4. Изобразить графики исходных данных, графики трендов при оптимальных параметрах у каждого метода для каждого ряда, расчетные формулы, а также результаты тестов Дарбина-Уотсона.

## СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. С. Е. Демин, Е. Л. Демина; М-во образования и науки РФ; ФГАОУ ВО «УрФУ им. первого Президента России Б.Н.Ельцина», Нижнетагил. технол. ин-т (фил.). – Нижний Тагил: НТИ (филиал) УрФУ, 2016. – 284 с.
2. О.И. Хайруллина, О.В. Баянова; Министерство сельского хозяйства Российской Федерации, федеральное государственное бюджетное образовательное учреждение высшего образования «Пермский аграрно-технологический университет имени академика Д.Н. Прянишникова». – Пермь: ИПЦ «Прокрость», 2019 – 176 с.
3. К. О. Кизбикенов. – Барнаул: АлтГПУ, 2017 – 115 с.
4. Карасёва Л. А. Статистика // Всемирная история экономической мысли: В 6 томах / Гл. ред. В. Н. Черковец. — М.: Мысль, 1987. — Т. I. От зарождения экономической мысли до первых теоретических систем политической жизни. — С. 484—494. — 606 с. — 20 000 экз. — ISBN 5-244-00038-1.
5. Миклашевский И. Н. Статистика теоретическая // Энциклопедический словарь Брокгауза и Ефрона: в 86 т. (82 т. и 4 доп.). — СПб., 1890—1907.
6. Норман Дрейпер, Гарри Смит. Прикладной регрессионный анализ. Множественная регрессия = Applied Regression Analysis. — 3-е изд. — М.: «Диалектика», 2007. — С. 912. — ISBN 0-471-17082-8.
7. Орлов А. И. Прикладная статистика. Учебник. — М.: Экзамен, 2006. — 671 с.
8. Дарелл Хафф. Как лгать при помощи статистики = How to Lie with Statistics. — М.: Альпина Паблишер, 2015. — 163 с. — ISBN 978-5-9614-5212-9.

9. Глинский В. В., Ионин В. Г. Статистический анализ. — М.: Инфра-М, 2002. — 241 с. — (Высшее образование). — 5000 экз. — ISBN 5-16-001293-1.

10. Кендалл М., Стьюарт А. Многомерный статистический анализ и временные ряды. — М.: Наука, 1976. — 736 с.

11. White C. Unkind cuts at statisticians (англ.) // The American Statistician. — 1964. — Vol. 18, no. 5. — P. 15—17.