**Лабораторная работа №1**

**ПОСТРОЕНИЕ И АНАЛИЗ РЕГРЕССИОННЫХ ЗАВИСИМОСТЕЙ**

Описание датасетов

2. Набор данных «Iris Data Set» – «Ирисы Фишера». Выборка взята из коллекции UC Irvine Machine Learning Repository и состоит из 150 цветков ирисов, распределенных по трем классам. Каждый класс содержит одинаковое число элементов (50 экземпляров): Ирис щетинистый (Iris setosa), Ирис виргинский (Iris virginica) и Ирис разноцветный (Iris versicolor).

Для каждого элемента выборки приводятся метка класса и значения 4-х признаков (в сантиметрах):

1) Sepal Length (длина чашелистика);

2) Sepal Width (ширина чашелистика);

3) Petal Length (длина лепестка);

4) Petal Width (ширина лепестка).

Описание настроек

Устанавливаем библиотекy:

pip install pingouin

Основные библиотеки в проекте (они будут установлены при установке pingoin)

scikit-learn (sklearn) –для построения моделей и много другого

matplotlib – для построения графиков

pandas – убоство работы с данными

seaborn - для построения графиков

scipy – тут есть всё для статистики и построения моделей.

pingouin – собрана основные статистические методы

Датасет «Ирисы Фишера» есть в библиотеки scikit-learn (так же дополнительно переведём в формат библиотеки pandas)

import pandas as pd

from sklearn.datasets import load\_iris

iris = load\_iris()

iris\_pd=pd.DataFrame(data=np.c\_[iris['data'], iris['target']], columns=iris['feature\_names'] + ['target'])

**ВЫПОЛНЕНИЕ**

**Для каждого датасета провести разведочный анализ данных:**

**построить диаграмму Тьюки, оценить диапазон изменения данных (для этого используем matplotlib). В отчёт включить диаграмму и написать к ней выводы.**

Name: sepal length (cm)

count 150.000000

mean 5.843333

std 0.828066

min 4.300000

25% 5.100000

50% 5.800000

75% 6.400000

max 7.900000

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Name: sepal width (cm)

count 150.000000

mean 3.057333

std 0.435866

min 2.000000

25% 2.800000

50% 3.000000

75% 3.300000

max 4.400000

\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Name: petal length (cm)

count 150.000000

mean 3.758000

std 1.765298

min 1.000000

25% 1.600000

50% 4.350000

75% 5.100000

max 6.900000

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Name: petal width (cm)

count 150.000000

mean 1.199333

std 0.762238

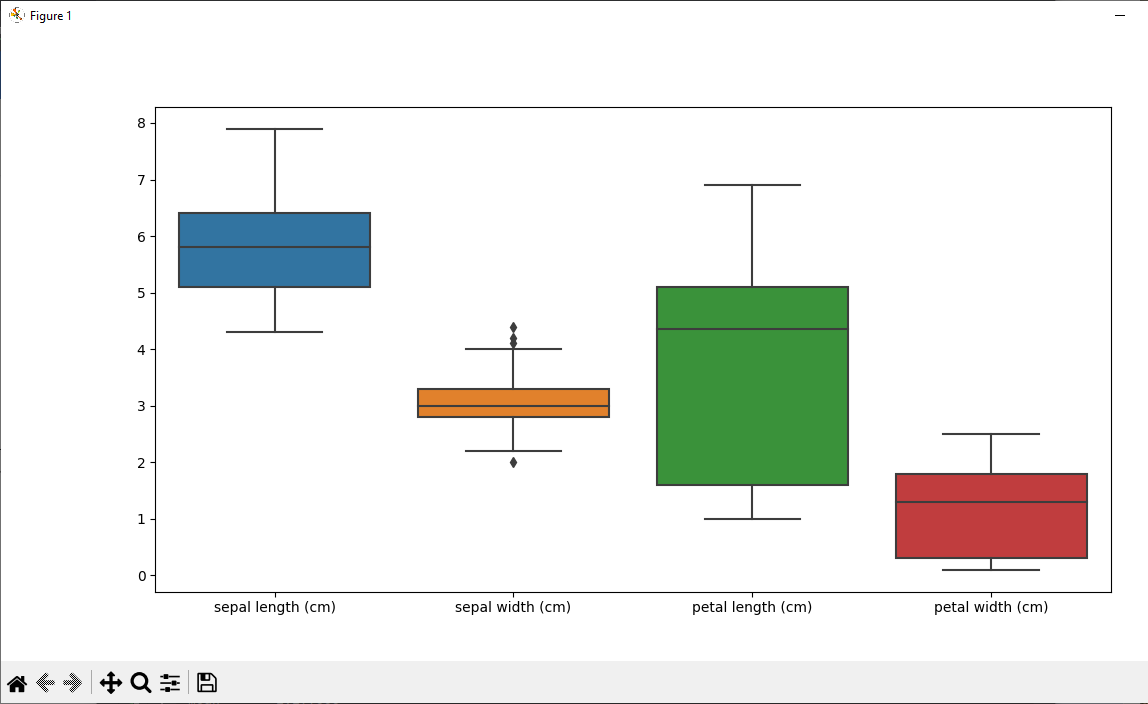
min 0.100000

25% 0.300000

50% 1.300000

75% 1.800000

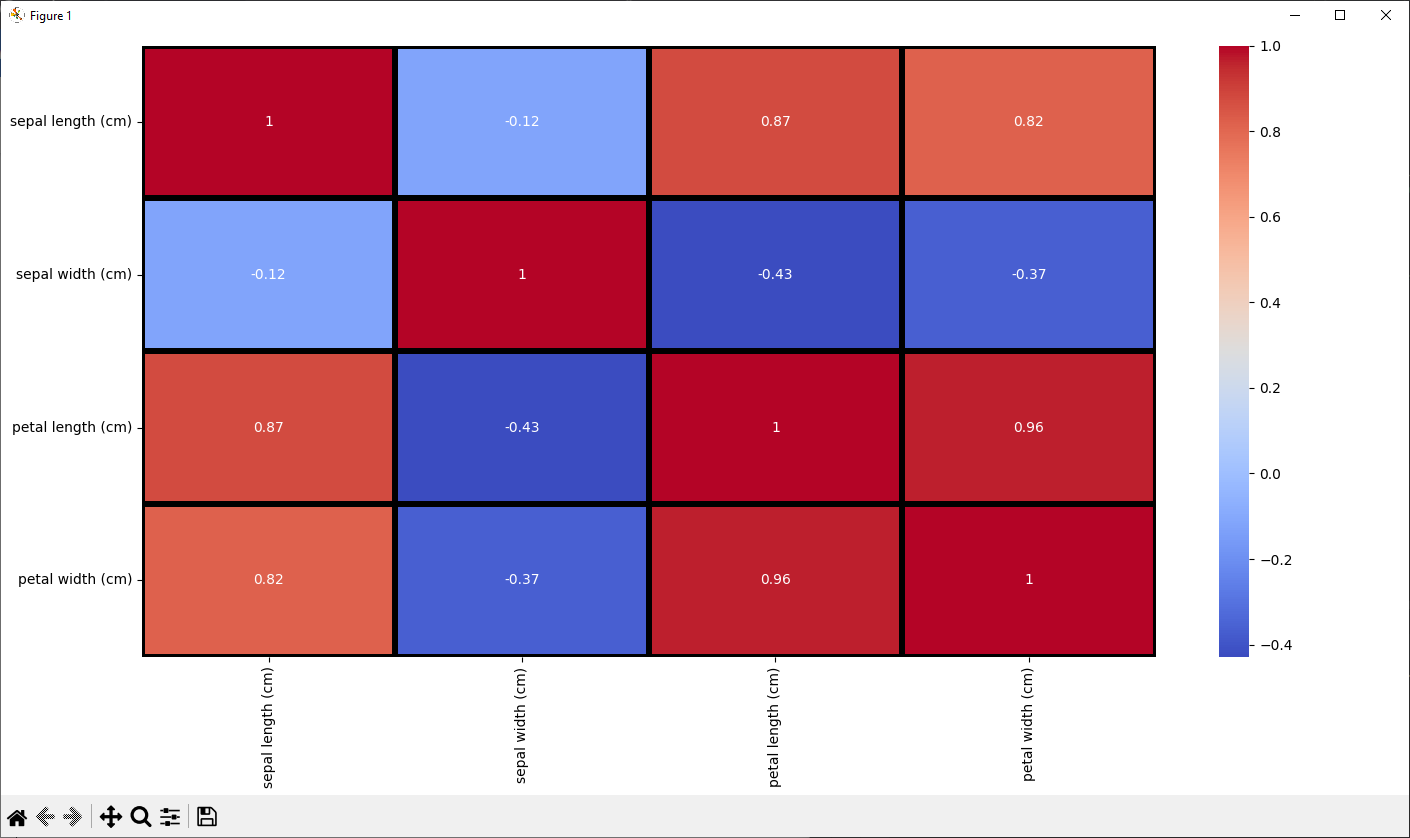
max 2.500000



Анализируя графики можно сделать следующие выводы:

Длина лепестка имеет самый широкий диапазон значений, когда ширина чашелистика самый узкий( в ней также наблюдаются выбросы). Самые большие размеры имеют длины чашелистиков, а самые маленькие- значения ширины лепестка; но их диапазоны Q25-Q75 почти одинаковой ширины и не отличаются так сильно, как в предыдущем случае.

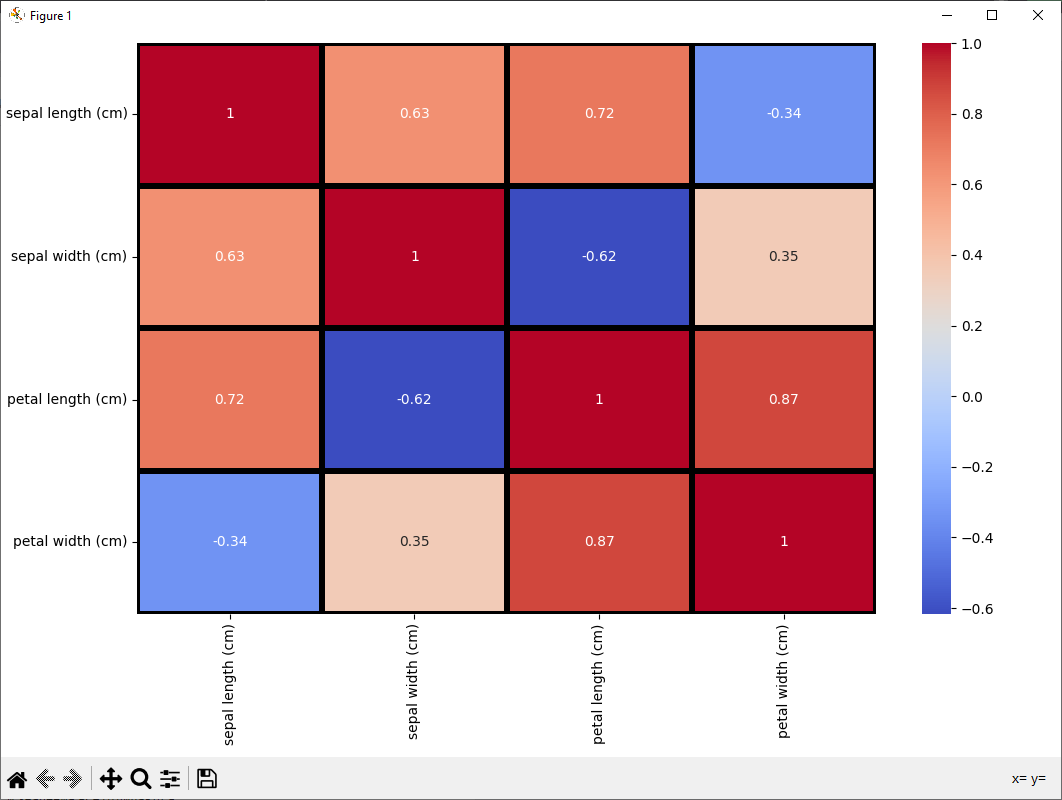
**Проанализировать корреляционные зависимости между исследуемыми переменными. Необходимо построить тепловую карту (для этого используем pandas + seaborn). В отчёт включить тепловую карту и написать к ней выводы.**



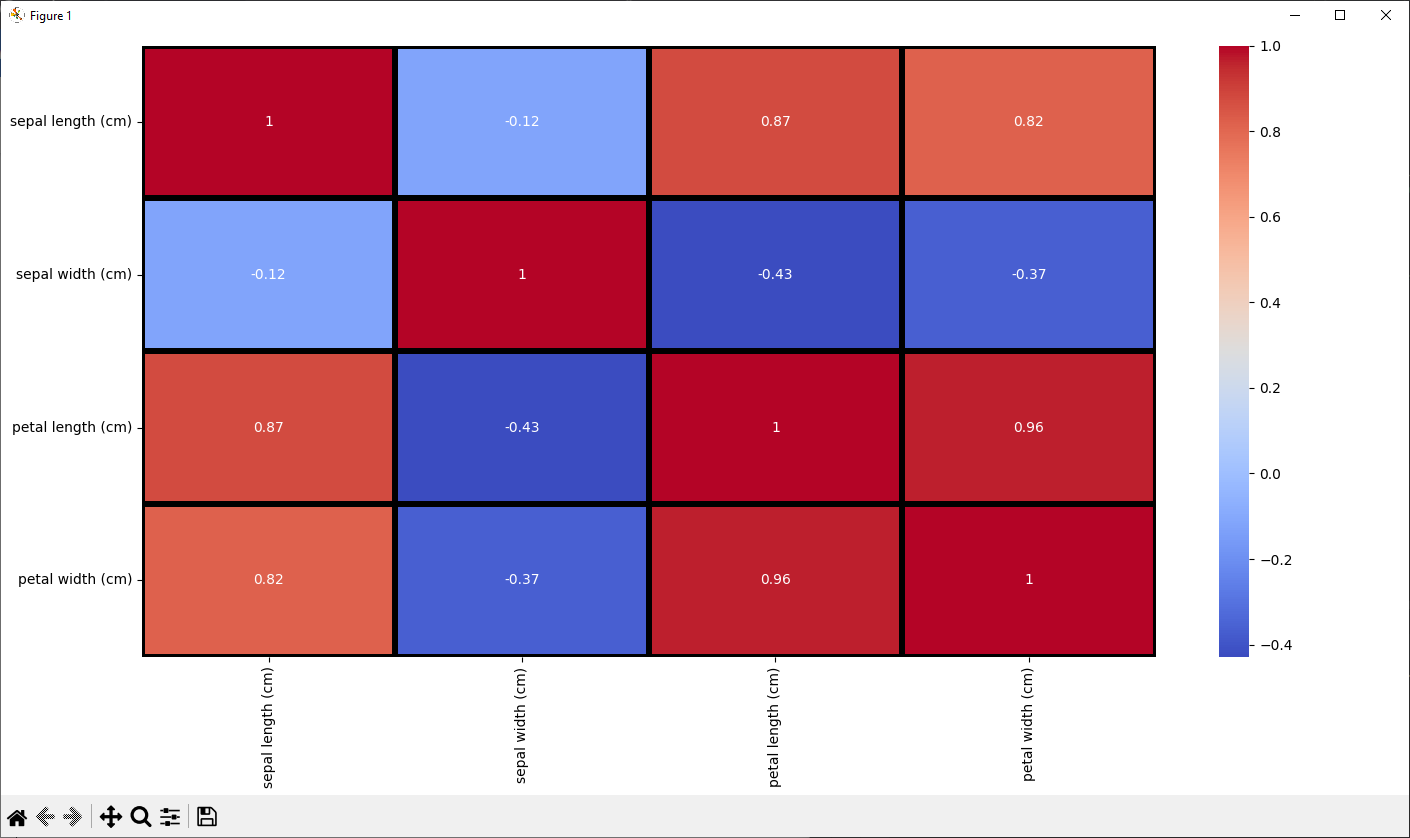
Данная таблица позволяет нам проанализировать зависимость характеристик ирисов дург от друга. Так, например чем больше длина чашелистика, тем больше будут длина и ширина лепестка, однако ширина чашелистика скорее будет меньше. Или например: чем больше ширина чашелистика, тем меньше будут его длина, длина лепестка и ширина лепестка.

**Рассчитать частные коэффициенты корреляции, сравнить их со значениями парных коэффициентов корреляции. Необходимо построить тепловую карту (для этого используем pandas + seaborn + pingouin (у датафрейма пандаса будет доступен метод pcorr)). В отчёт включить тепловую карту и написать к ней выводы.**

Матрица частных коэф-ов корреляции:



Матрица парных коэф-ов:



Как видно, матрицы отличаются друг от друга. Парный коэффициент показывает влияние одной переменной на другую с учетом изменения других переменных (то есть, например: влияние ширины лепестка на его длину при учете ширины и длины чашелистика). Частный коэффициент же показывает влияние характеристик друг на друга при фиксированном значении остальных, то есть устанавливает взаимосвязь между ними без учета остальных характеристик. Так мы можем наблюдать, что при отсутствии влияния других факторов, ширина лепестка и длина лепестка имеют весьма сильную взаимосвязь, а ширина чашелистика и ширина листа, в отличии от парной корреляции, имеют среднюю связь. Или например: без учета остальных данных, длина и ширина чашелистика имеют среднюю, близко к сильной, связь; а длина чашелистика и ширина листа имеют среднюю отрицательную связь.

**Проверить предположение о распределении признаков по нормальному закону критерием Колмогорова-Смирнова. Необходимо рассчитать значения статистик (для этого используем scipy). В отчёт включить таблицу значений и написать к ней выводы.**

sepal length (cm)

KstestResult(statistic=0.999991460094529, pvalue=0.0)

sepal width (cm)

KstestResult(statistic=0.9794298858198347, pvalue=1.9343513094431716e-253)

petal length (cm)

KstestResult(statistic=0.8765328487477231, pvalue=1.4044248603466388e-136)

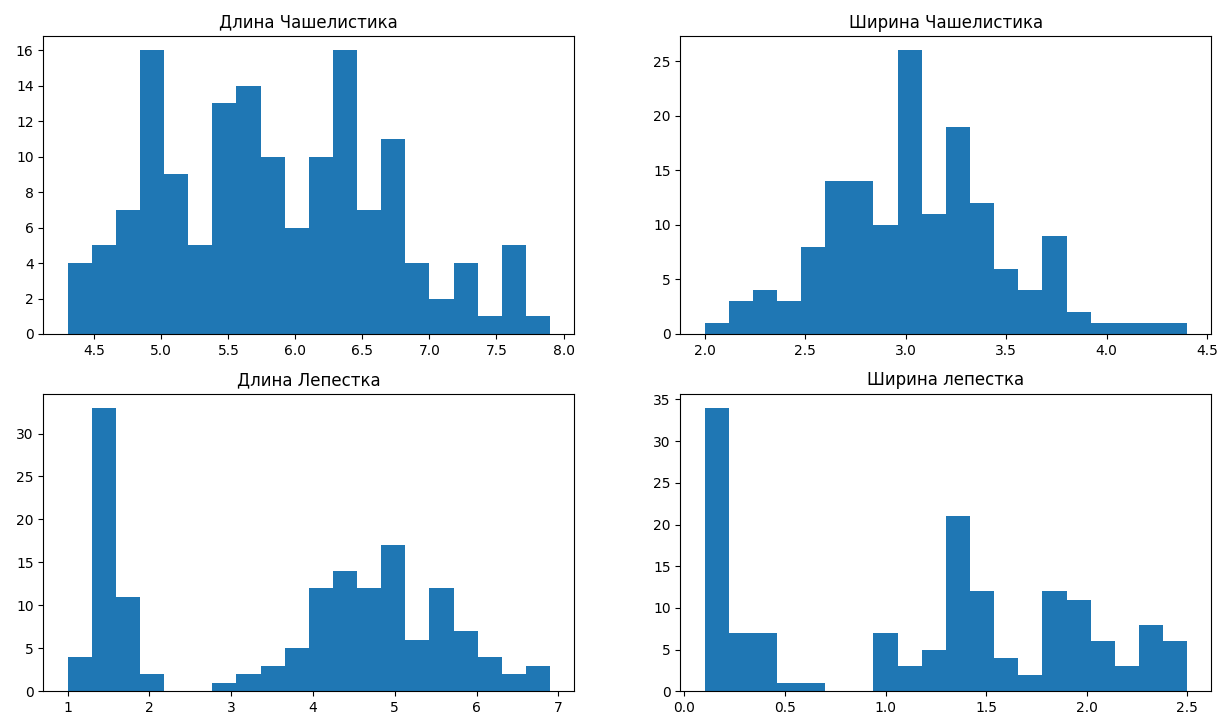
petal width (cm)

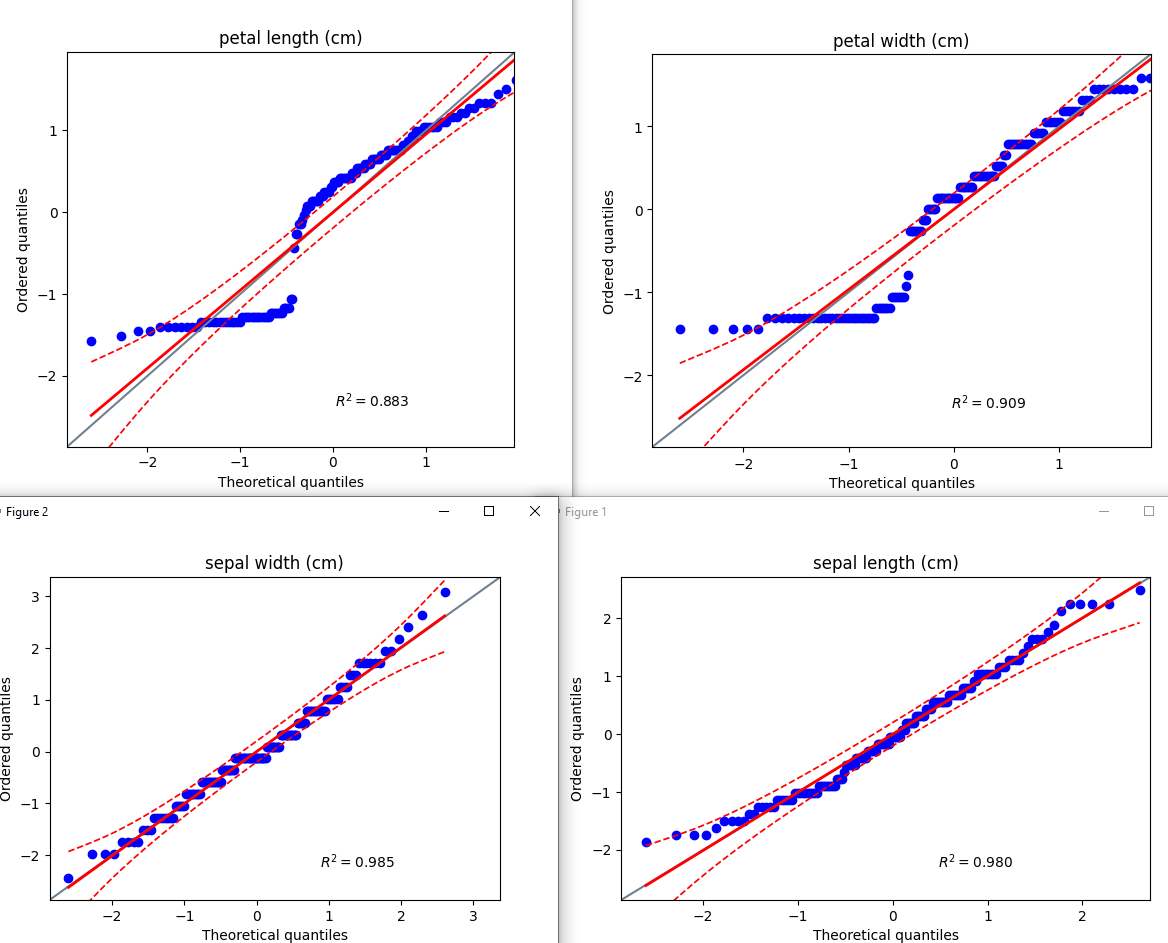
KstestResult(statistic=0.5459263761057697, pvalue=1.876499271371568e-42)

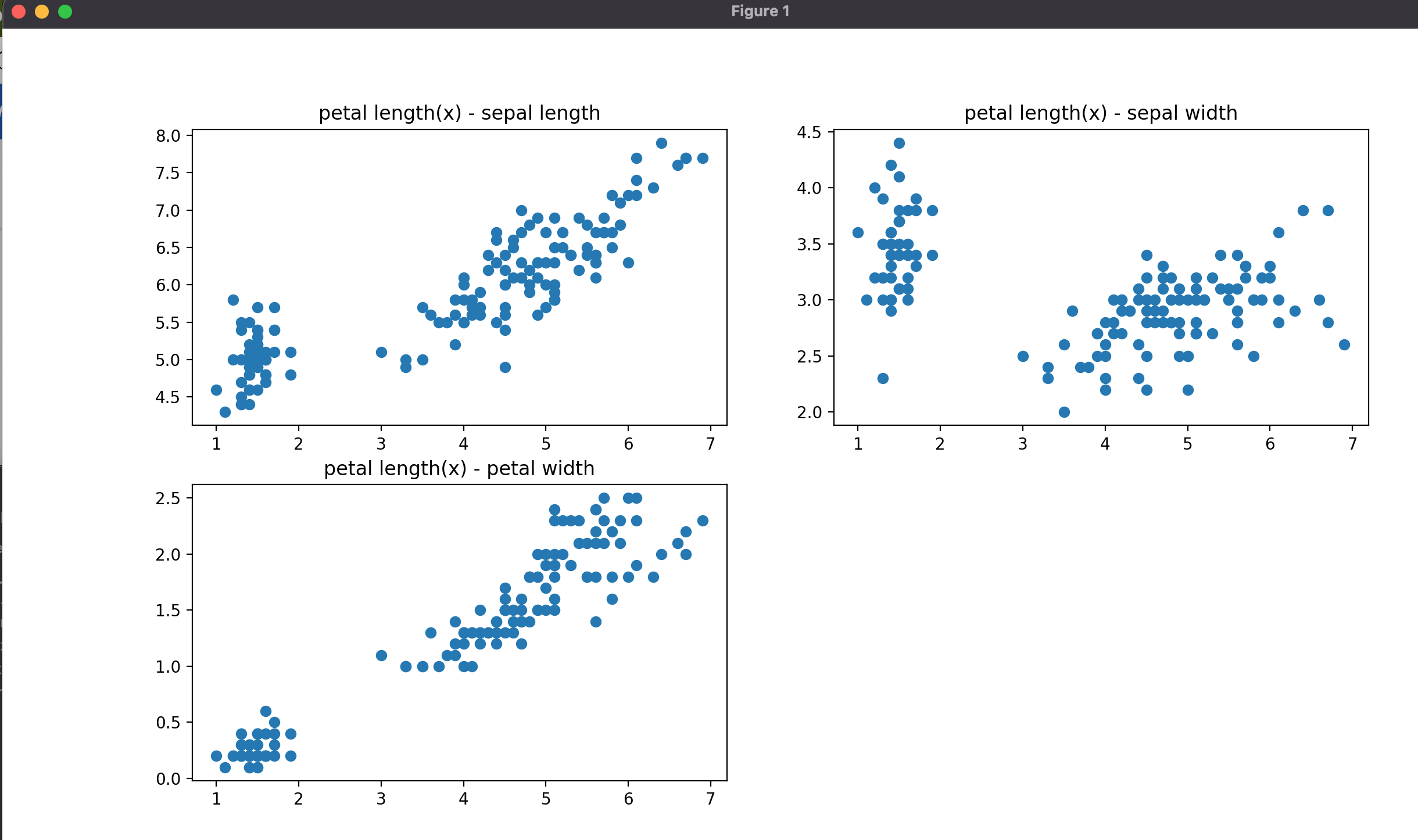
Т.к все pvalue < 0.05 У нас достаточно оснований считать что признаки распределены ненормально.

**Сделать выводы о степени однородности данных, силе зависимости между переменными, виде функции распределения, наиболее информативных переменных (Вставляем в конец абзаца).**

Анализируя графики, можно сделать следующие выводы: больше всех с остальными коррелирует характеристика длины лепестка. Характеристики лепестков распределены не нормально и имеют по два, явно выраженных, пика. Характеристики чашелистика распределены более однородно, распределения имеют сходство с нормальным, но таковыми не являются.







**Определить входные и выходные переменные. Построить парную регрессию и множественную регрессию (Используем sklearn). Сравнить результаты парной и множественной регрессии. В отчёт включить график с предсказанными значения (на одном графике), график предсказанных значений и остатков (на одном графике), и значения множественного коэффициента корреляции, коэффициента детерминации, скорректированного коэффициента детерминации и стандартной ошибки. Выбрать наиболее адекватную модель (в отчёте написать обоснование).**

Выходную переменную, она же зависимая переменная, стоит выбрать длину лепестка, т.к она сильнее всех коррелирует с остальными переменными, распределение этой характеристики имеет неплохую линейную взаимосвязь с тремя другими переменными, поэтому стоит начать исследование именно этого вида модели, выбрав ширину лепестка и чашелистика и длину чашелистика за независимые переменные.

**Проведем анализ парной регрессии между переменными:**

Анализируя полученные результаты можем сделать следующие выводы: распределение остатков от всех тех зависимостей сохраняют гомоскедастичность, однако из-за выбросов, наблюдаемых в ширине чашелистика, мы также наблюдаем выбросы в остатках, большую стандартную ошибку и низкий коэф детерминации. Самый высокий кф детерминации показала зависимость между длиной лепестка и шириной лепестка, значит для случая парной регрессии лучше всего изменчивость длины лепестка объясняет ширина лепестка. Если бы мы не применяли множественную регрессию, то мы бы выбрали именно эту, моследнюю модель.

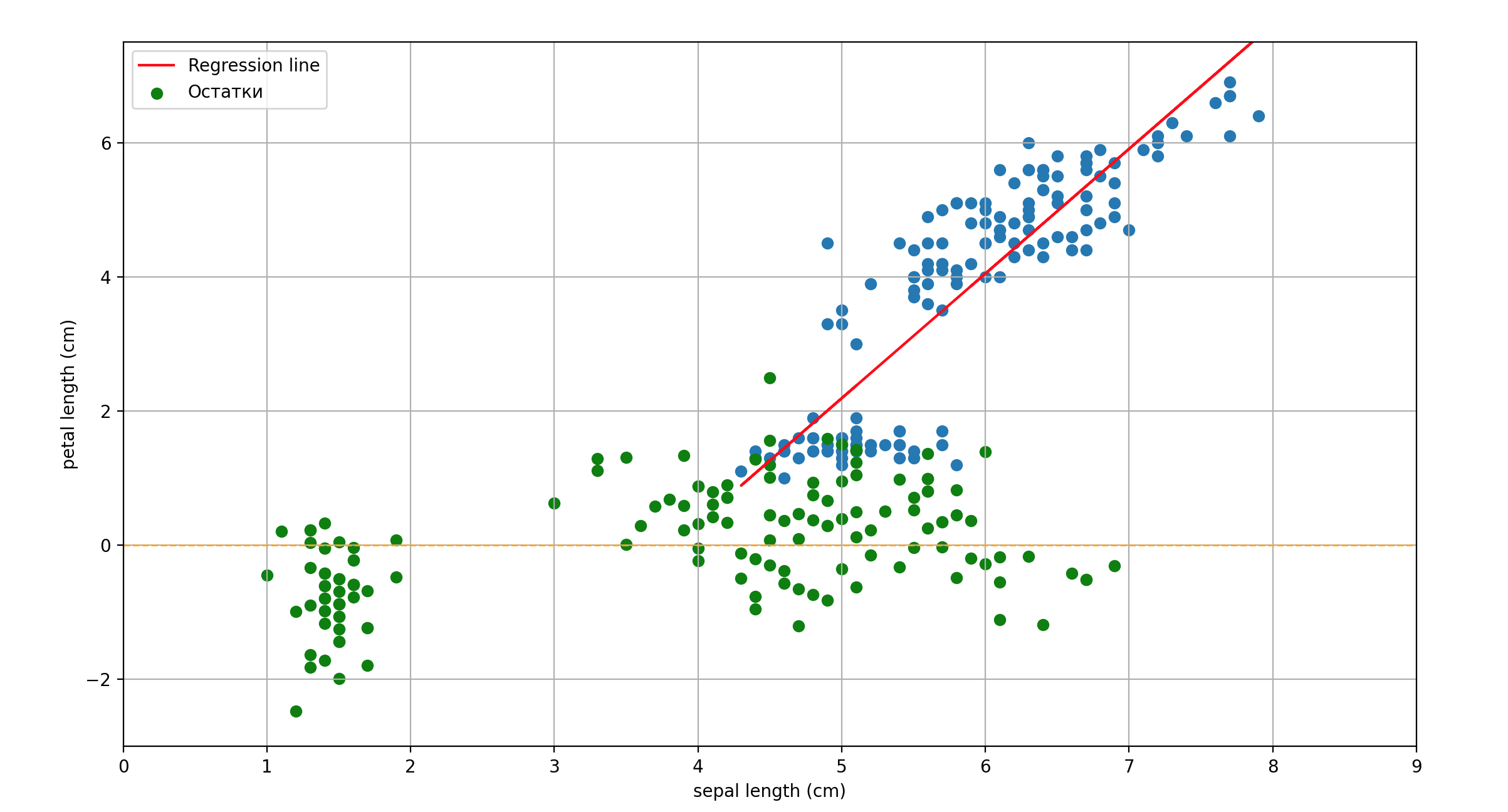
Анализ парной регрессии между длиной лепестка и длиной чашелистика

b coef: [[1.85843298]]

intercept: [-7.10144337]

Coef determination: 0.759954645772515

MAE: 0.7067088106229915



\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

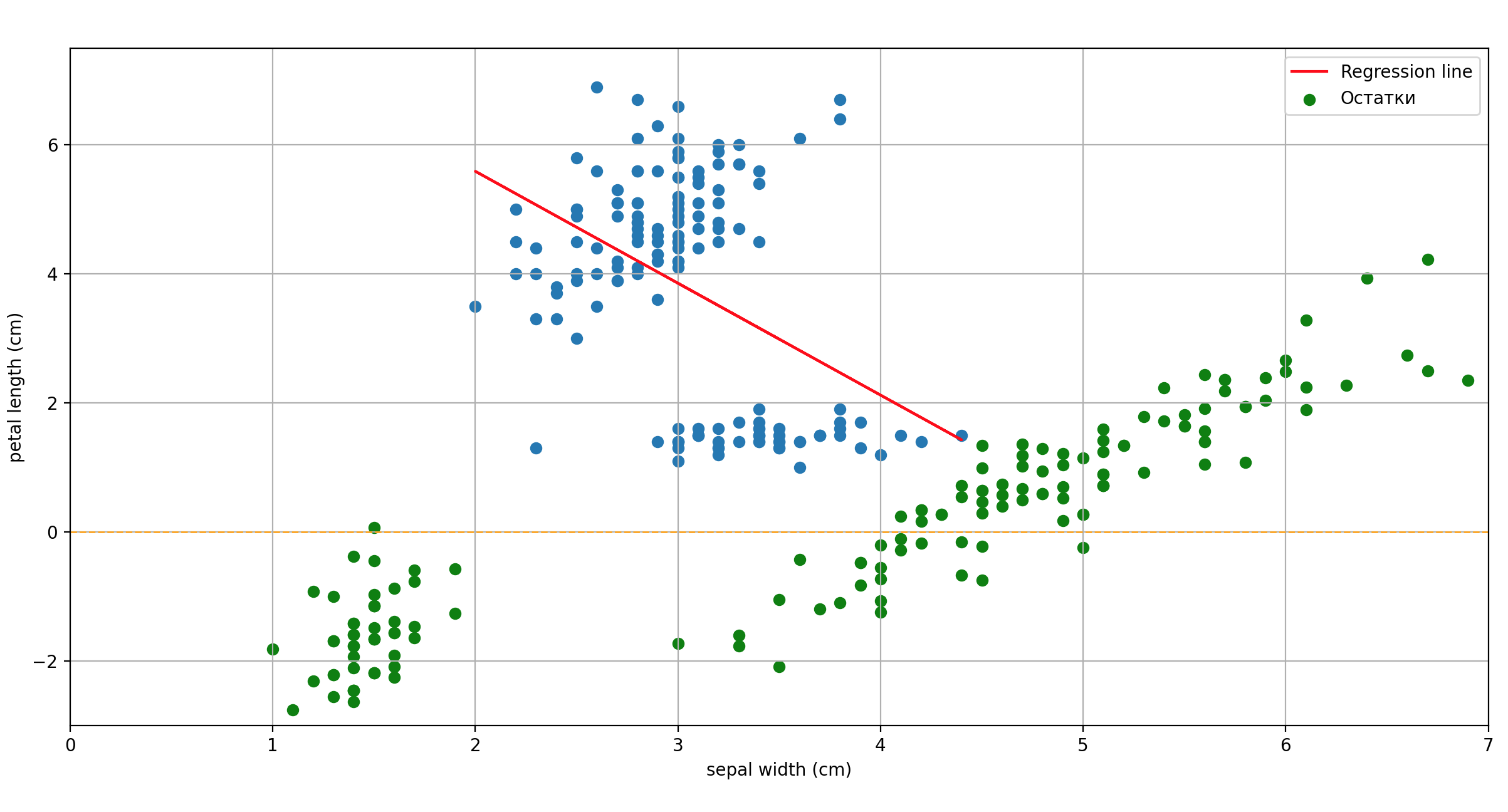
Анализ парной регрессии между длиной лепестка и шириной чашелистика

b coef: [[-1.73522152]]

intercept: [9.06315061]

Coef determination: 0.1835609229987637

MAE: 1.358156701302861



\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

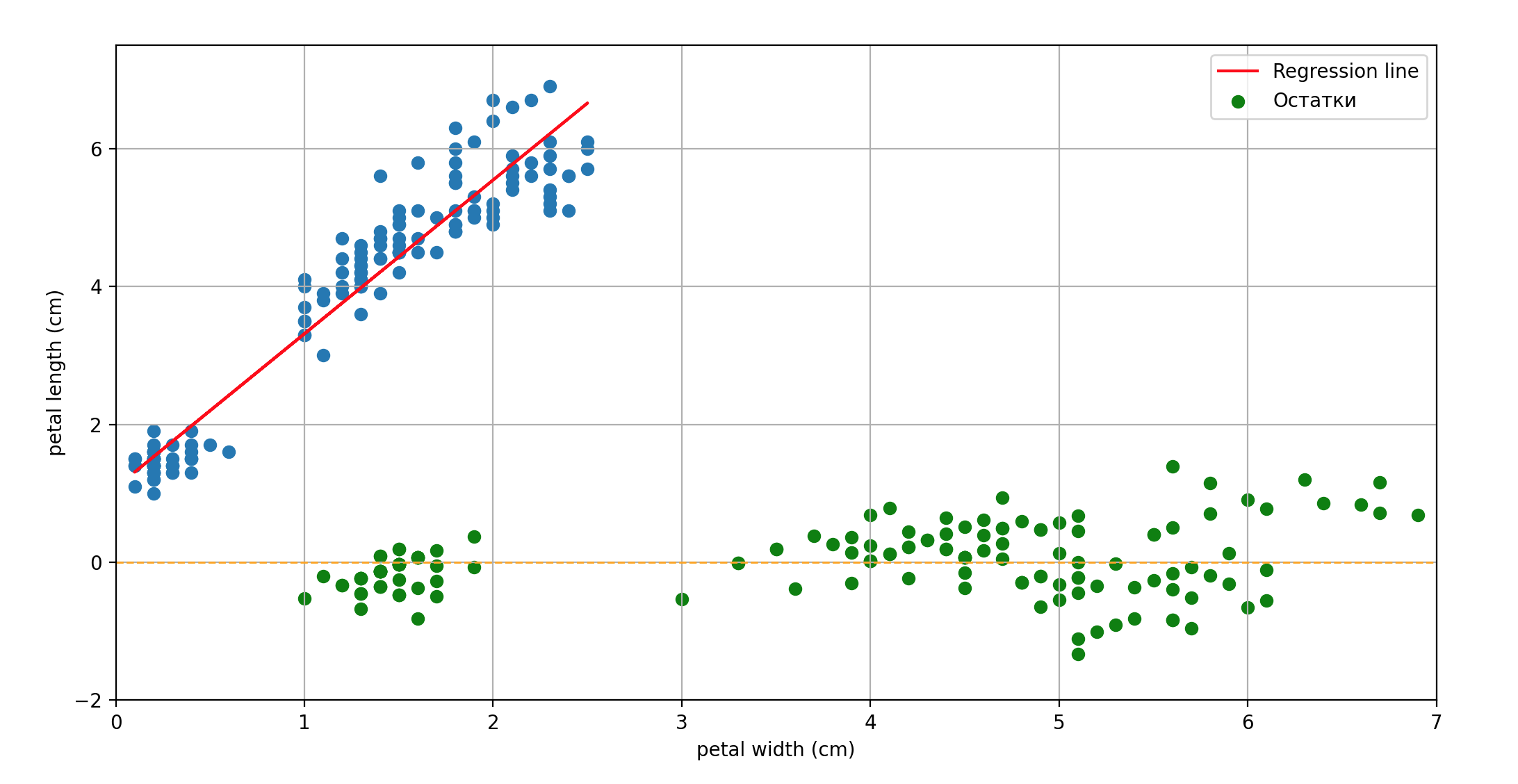
Анализ парной регрессии между длиной лепестка и шириной лепестка

b coef: [[2.2299405]]

intercept: [1.08355803]

Coef determination: 0.9271098389904927

MAE: 0.36579694412763775



\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**Проведем анализ для случая множественной регрессии:**

Для всех 3:

Coeffs

sepal length (cm) 0.729138

sepal width (cm) -0.646012

petal width (cm) 1.446793

intercept: -0.2627111975741818

determin coeff: 0.9680117693912218

mulipile corr coef 0.9838758912541875

Скорректированный кф детерминации 0.9673544769814524

MAE: 0.24008457241575176

Выкинули ширину чашелистика

Coeffs

sepal length (cm) 0.542256

petal width (cm) 1.748103

intercept: -1.5071383768459268

determin coeff: 0.9485236349446816

mulipile corr coef 0.9739217807117169

Скорректированный кф детерминации 0.947823276236446

MAE: 0.31369294102123063

Выкинули длину чашелистика

Coeffs

sepal width (cm) -0.355035

petal width (cm) 2.155611

intercept: 2.258163512047224

determin coeff: 0.933764185485493

mulipile corr coef 0.9663147445245224

Скорректированный кф детерминации 0.9324031755982086

MAE: 0.344127313072561

Выкинули ширину лепестка

Coeffs

sepal length (cm) 1.775593

sepal width (cm) -1.338623

intercept: -2.524761511833407

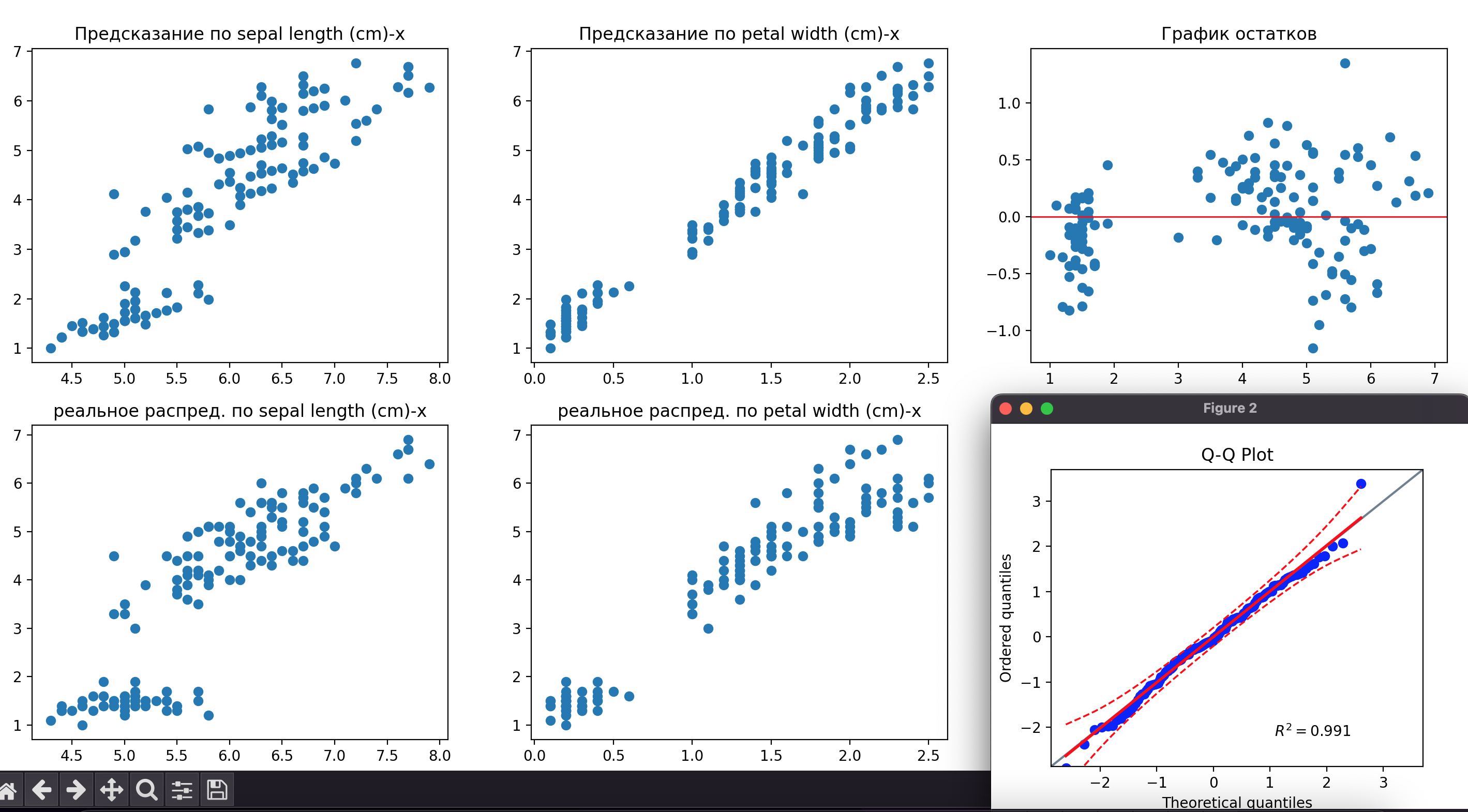
determin coeff: 0.8676860089345194

mulipile corr coef 0.9314966499856666

Скорректированный кф детерминации 0.8649672282961876

MAE: 0.5162246726456129

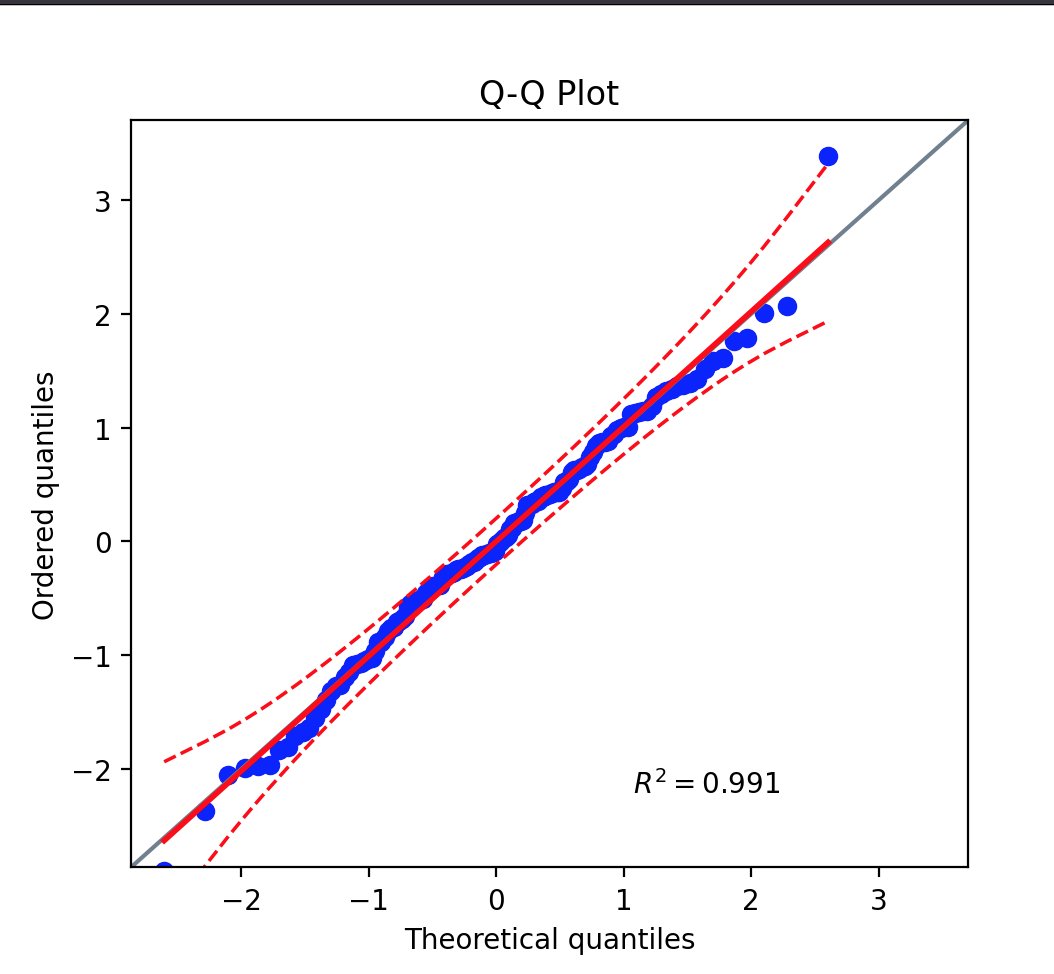
Возьмем модель, в которой не учитывается ширина чашелистика, т.к в этой характеристике наблюдаются выбросы, что может отрицательно сказаться на предсказываемой величине.



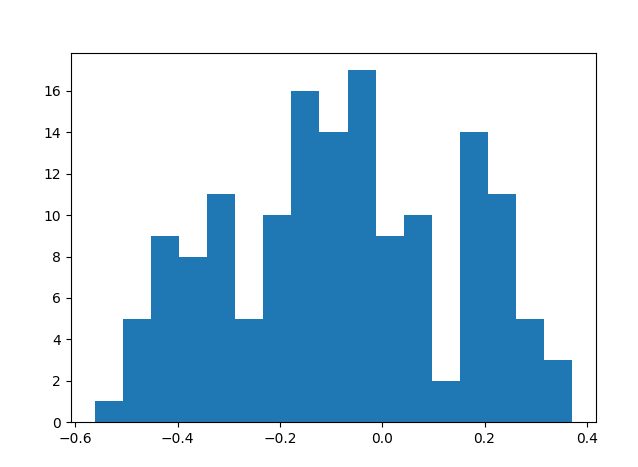
Как видно выше: распределение остатков похоже на нормальное и сохраняет гомоскедастичность.

Как итог: было принято решение использовать модель с двумя независимыми переменными: длиной чашелистика и шириной листа. Мы убрали ширину чашелистика, потому что ранее в этой характеристике наблюдались выбросы, что могло повлиять на наши предсказания. В нашей модуле коэффициент детерминации достаточно высок, а также низкая стандартная ошибка, что позволяет нам сделать вывод, что это модель является достаточно оптимальной, для предсказания длины лепестка.

**Проверить гипотезу о нормальном распределении остатков. Рассчитать статистику Дурбина-Уотсона. В отчёт включить рассчитанные значения и написать, как определили исходя из полученного значения статистики.**



KstestResult(statistic=0.23142153806530008, pvalue=1.5149175653583154e-07)

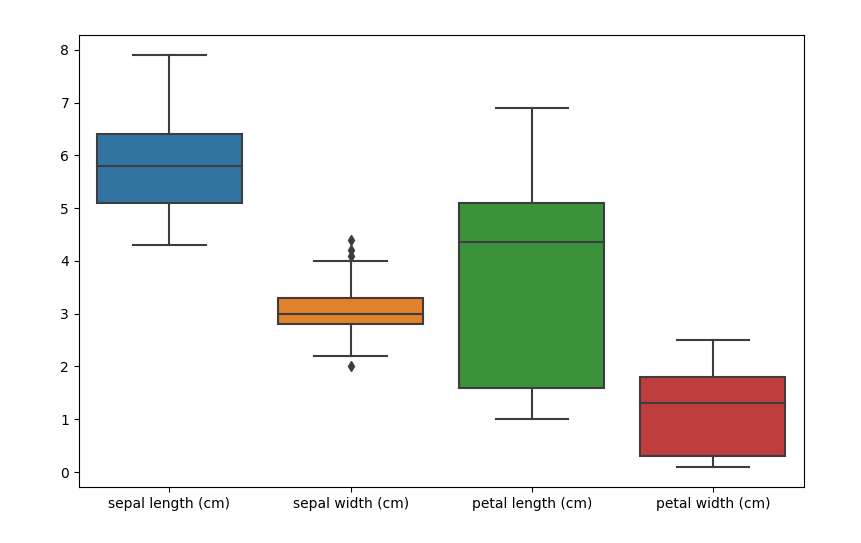


Поскольку pvalue теста Колмогорова-Смирнова меньше 0.05, то у нас достаточно оснований полагать, что распределение остатков не является нормальным, что также можно заметить и по гистограме.

Дурбин Вотсон: 1.3391185441384632

Статистика относительна близка к 2, есть некоторые основания считать, последовательная корреляция отсутствует.

**Проанализировать наличие выбросов (смотрим по «диаграмме Тьюки»). Сделать вывод о наличии (отсутствии) выбросов (с учетом результатов, полученных в предыдущих пунктах). В отчёт продублировать диаграмму и написать выводы.**



Выбросы наблюдаются в ширине чашелистика, обучая нашу модель, мы учли этот факт и выбросили его из независимых переменных, также было видно, что при удалении этой переменной, сильно изменились коэффициенты b1 и b2, а при удалении других, эти коэффициенты сильно не менялись, что также можно считать за обоснованное решение.

**Провести предсказание зависимой переменной по вычисленному уравнению множественной регрессии (Predict Dependent Variable), оценить точность. Для этого делим выборку на тестовую и обучающую. После чего рассчитываем СКО, коэффициент детерминации и строим график ошибок. В отчёт включить график, значение СКО, коэффициента детерминации и написать выводы о качестве модели.**

Количество элементов в тренировочной модели- 120 элементов. Тестовая выборка -30 элементов

TRAIN

Coeffs

sepal length (cm) 0.494126

petal width (cm) 1.779993

intercept: -1.2823982281800896

coef determination: 0.9548129226675086

Скорректированный кф детерминации 0.9540404939951583

coef correlation: 0.9771452925064463

MSE : 0.14276100827253616

TEST

Coeffs

sepal length (cm) 0.494126

petal width (cm) 1.779993

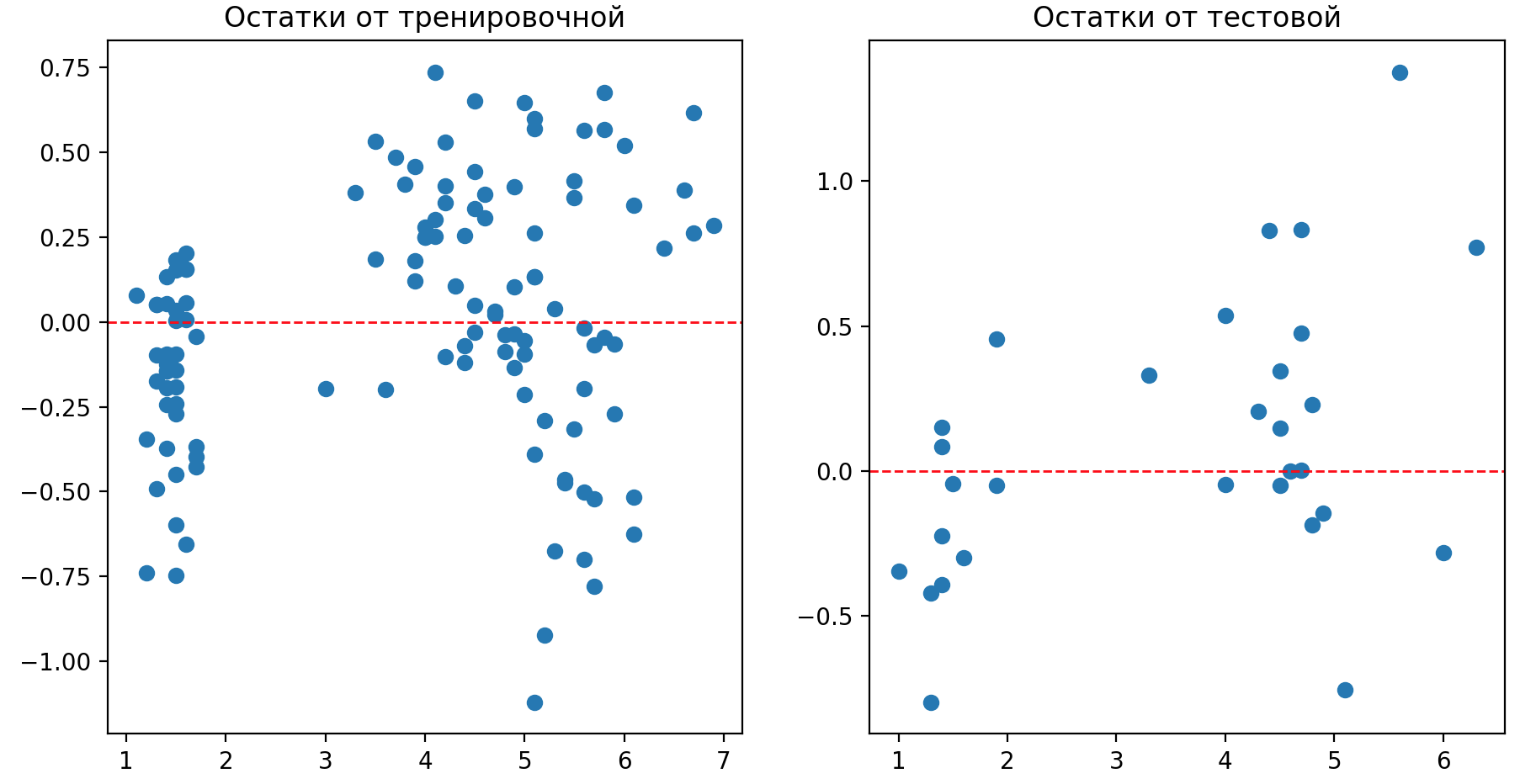
intercept: -1.2823982281800896

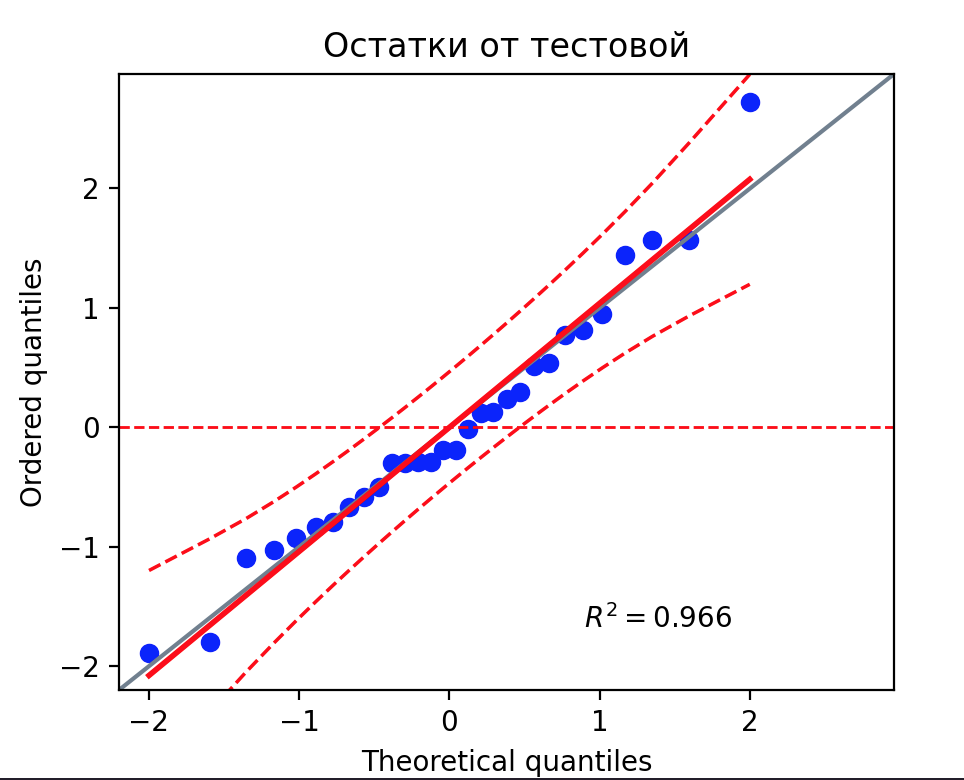
coef determination: 0.9169579087500086

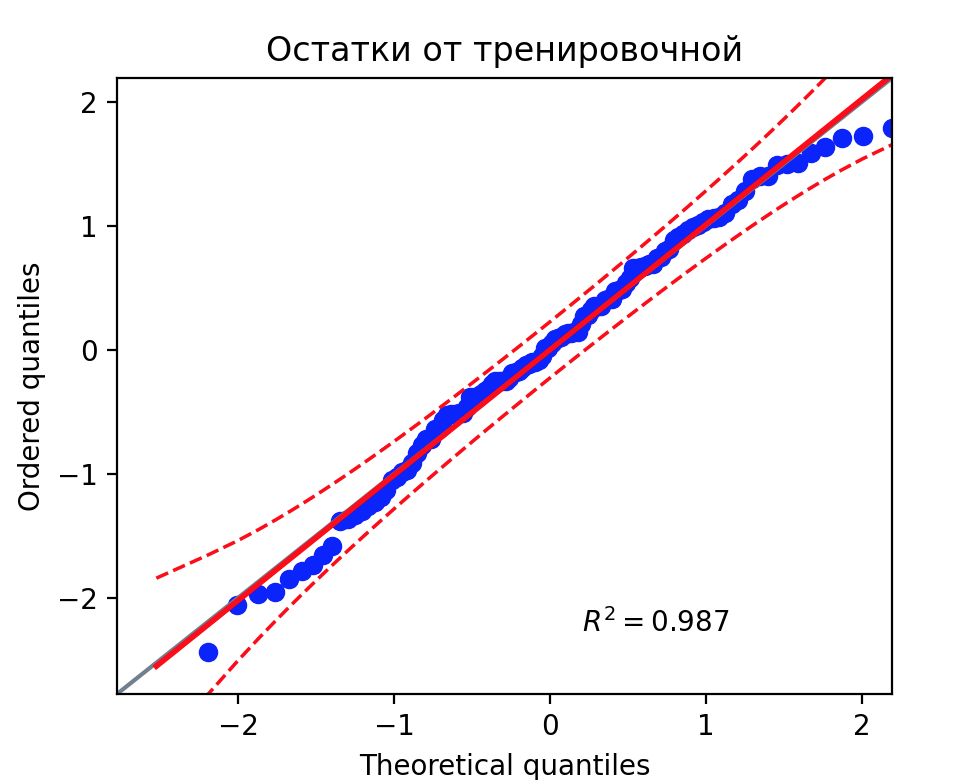
Скорректированный кф детерминации 0.9108066427314907

coef correlation: 0.9575791918948576

MSE : 0.23029970897369842







Анализируя полученные данные, можно сделать следующие выводы: наша модель достаточно хорошо обрабатывает тестовые выборки, это можно сказать, учитывая хороший коэффициент детерминации и не очень высокую среднеквадратическую ошибку средних. График остатков сохраняет гомоскедастичность, распределение остатков похоже на нормальное, однако в распределении остатков тренировочной выборки наблюдаются выбросы, что может сказываться на предсказаниях нашей модели и это следует учитывать.

**Сформулировать предложения по улучшению регрессионной модели.**

Первое что нам захочется сделать — это увеличить объемы выборок, например тренировать нашу модель на 300 элементах, а тестировать на 100, тогда мы будем точнее предсказывать значения. Также мы можем проанализировать выбросы и возможно удалить такие данные, если они окажутся аномальными. Также наша модель, возможно, будет лучше предсказывать результат, если мы учтем и другие характеристики ирисов: например, количество солнечных дней или частота поливания. Однако следует опасаться мультикорелляции, ведь при увеличении количества независимых переменных мы рискуем ухудшить предсказательную способность нашей модели.

**Провести корреляционный анализ и сделать выводы о целесообразности построения регрессии для датасета:**

Таблица имеет следующий вид:

**- Chemical\_process.sta;**

**YIELD HCL NH3 H20 CATALYST**

**0 59,5668403 12,0718253 12,4206519 12,226249 9,34286019**

**1 61,2219727 9,93095187 10,8588672 10,124134 8,18025758**

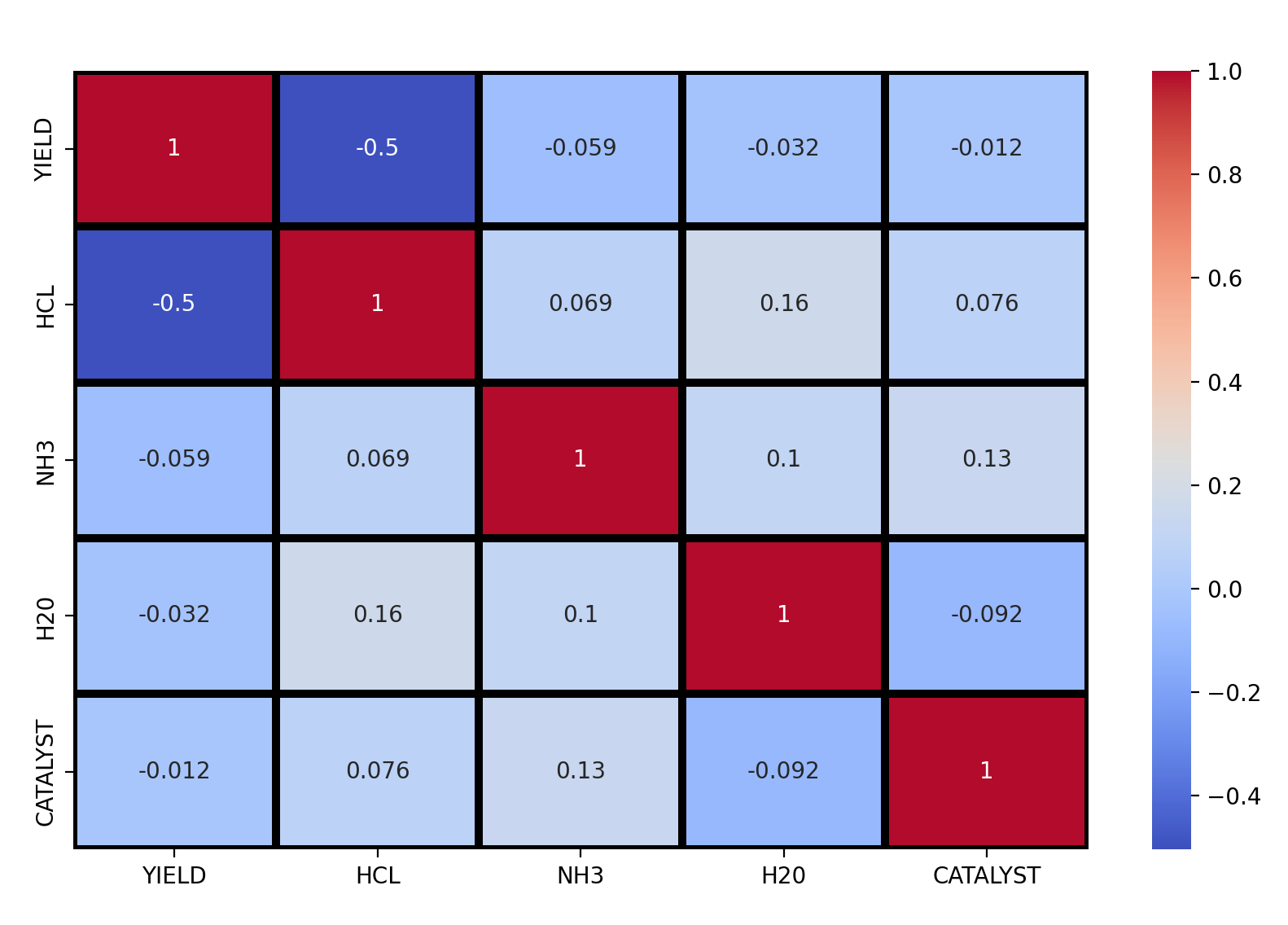
**2 61,0310989 9,77530747 7,83677175 7,77680288 8,28966643**

**3 61,8486942 9,44677572 12,42 10,5518509 11,2684256**

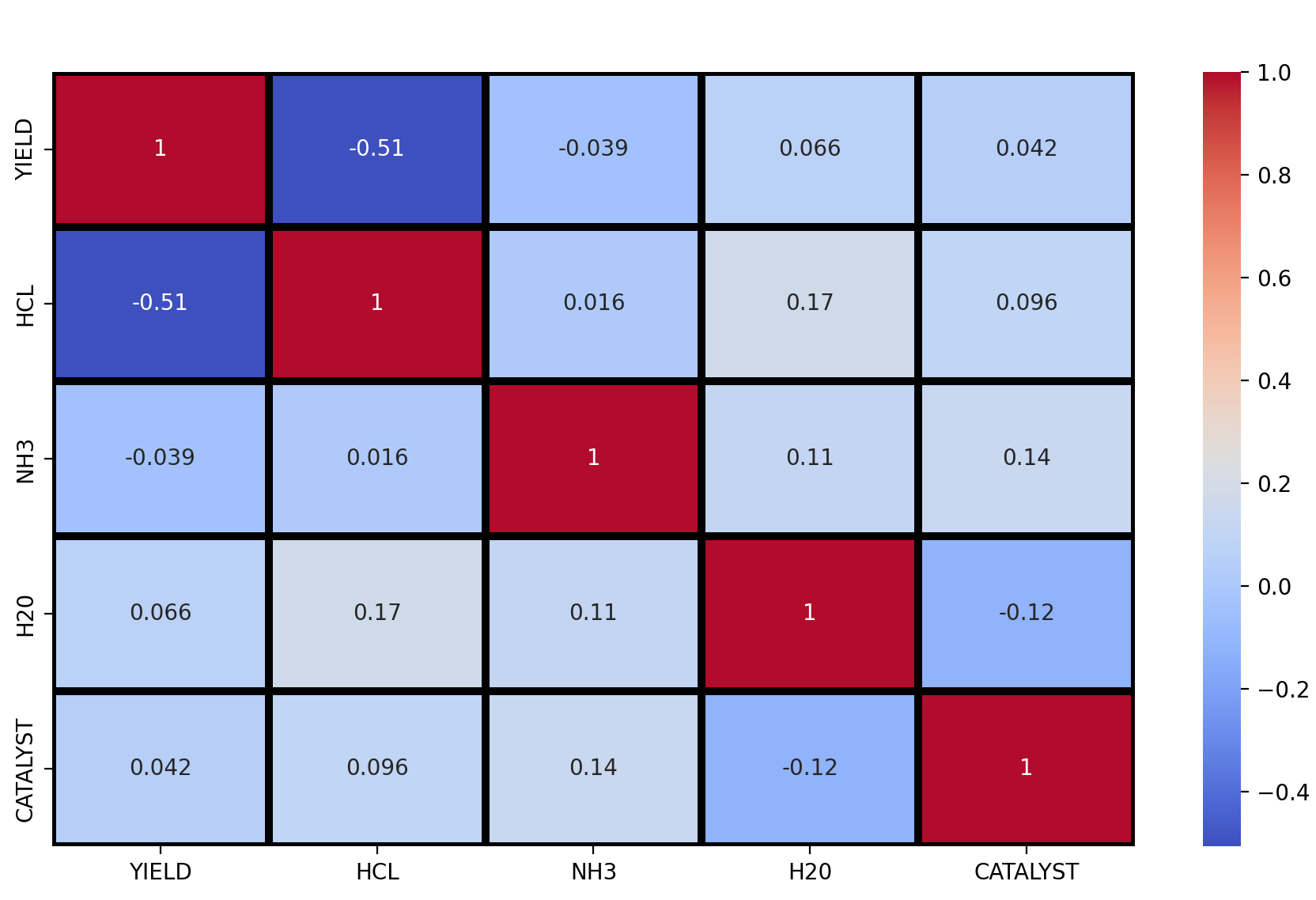
**4 59,4272822 11,9083987 12,39 8,2083346 10,3972747**

Для того, чтобы применять регрессионный анализ, у нас должны соблюдаться следующие факторы: переменный должны иметь линейную зависимость, остатки должны иметь нормальное распределение и обладать гомоскедастичностью. Ну и конечно, между переменными должна быть корреляция.

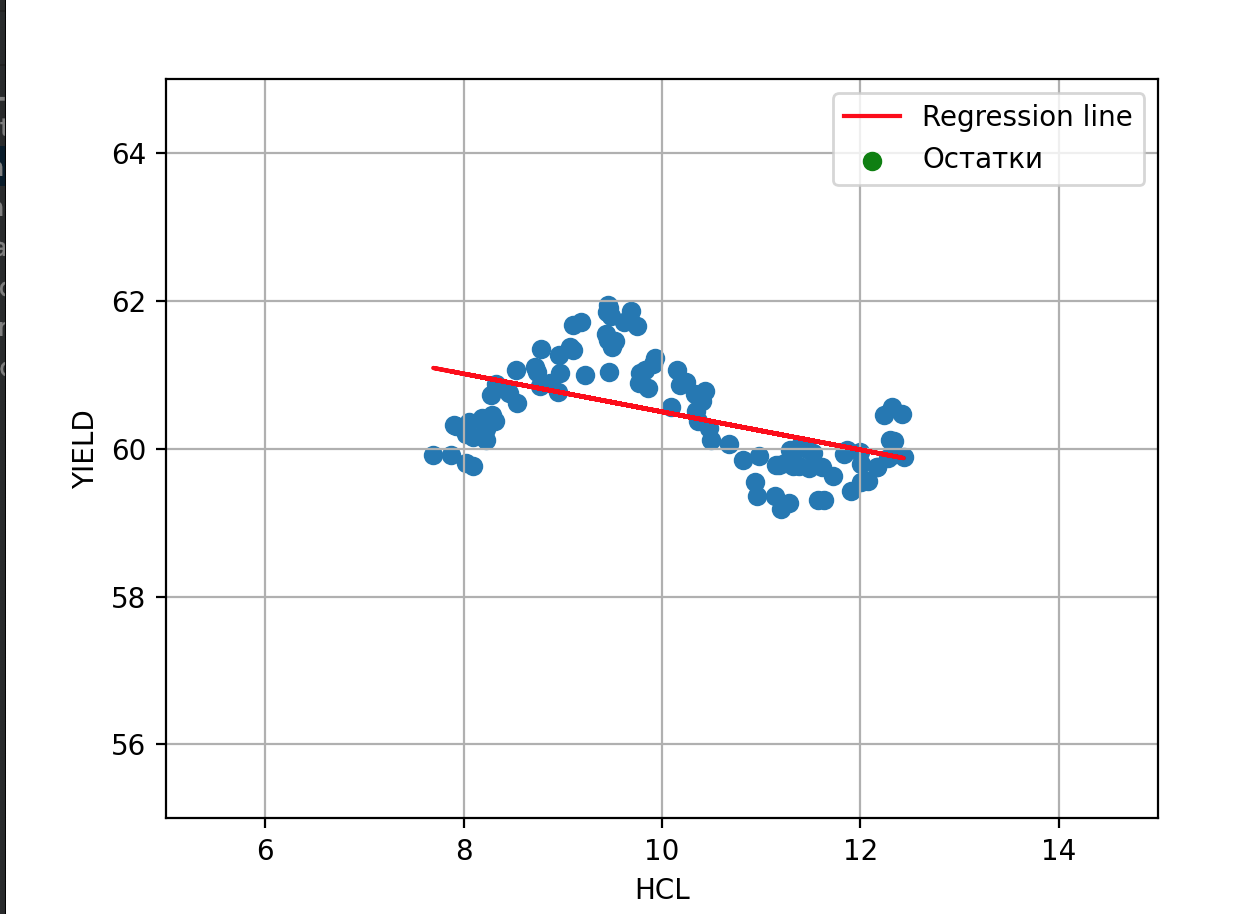
**Парная карта:**

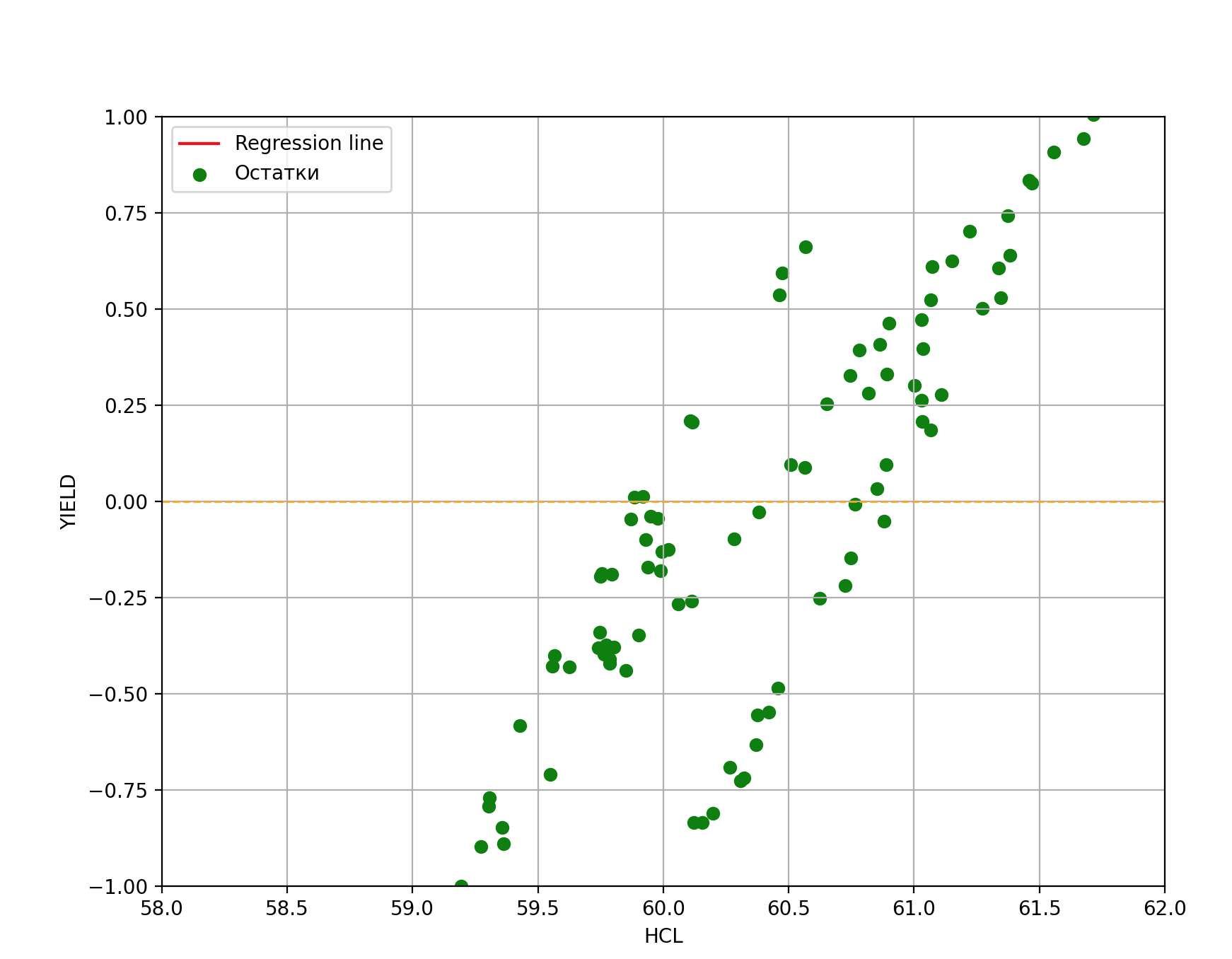


Множественная карта



Как следует из коэффициентов корреляции, сильная корреляция между переменными отсутствует. Между HCL и YIELD присутствует средняя взаимосвязь. Только исходя из этой информации, мы уже можем заключить, что построение регрессионной модели для этих данных - нецелесообразно. Чтобы лучше в этом убедиться, давайте построим зависимость, за выходную переменную возьмем YIELD.





Как видно из данных графиков, зависимость HCL от YIELD имеет сильную нелинейность, остатки также распределены ненормально и свойство гомоскедастичности не сохраняется. Отсюда можно сделать вывод, что модель линейной регрессии нецелесообразно применять к этим данным.