|  |
| --- |
| **МИНОБРНАУКИ РОССИИ**  федеральное государственное бюджетное образовательное  учреждение высшего образования  **«Национальный исследовательский университет «МЭИ»** |

|  |  |
| --- | --- |
| **Институт** | ИВТИ |
| **Кафедра** | УИТ |

**Лабораторная работа №1**

**ПОСТРОЕНИЕ И АНАЛИЗ РЕГРЕССИОННЫХ ЗАВИСИМОСТЕЙ**

**Выполнила**

Ишутина Е. И.,

А-01-22

**Проверил**

Назаров Н. А.,

асп. кафедры Управления и   
интеллектуальных технологий

**Оценка  
\_\_\_\_\_\_\_\_\_\_**

Москва, 2024



**Выполнение работы**

1. **Выборка из датасета *SOCR Data - 1035 Records of Heights and Weights of Major League Baseball Players***

Датасет скачан как .csv и из него выделена случайная выборка в терминале. Эта выборка записана в файле sample.csv.

>>> import os

>>> os.chdir("D:\\STUDY\\LVL3\\Интеллектуальный анализ данных\\lr1")

>>> import pandas as p  
>>> data = p.read\_csv ("baseball\_players.csv")  
>>> sample = data.sample(n=115)  
>>> sample.to\_csv("D:\\STUDY\\LVL3\\Интеллектуальный анализ данных\\lr1\\sample.csv")

Проверим:

>>> print(sample.head())

Unnamed: 0 Name Team ... Height(inches) Weight(pounds) Age

0 328 Edwin\_Jackson TB ... 75 190.0 23.47

1 830 Johnny\_Estrada MLW ... 71 209.0 30.67

2 703 Lastings\_Milledge NYM ... 72 187.0 21.90

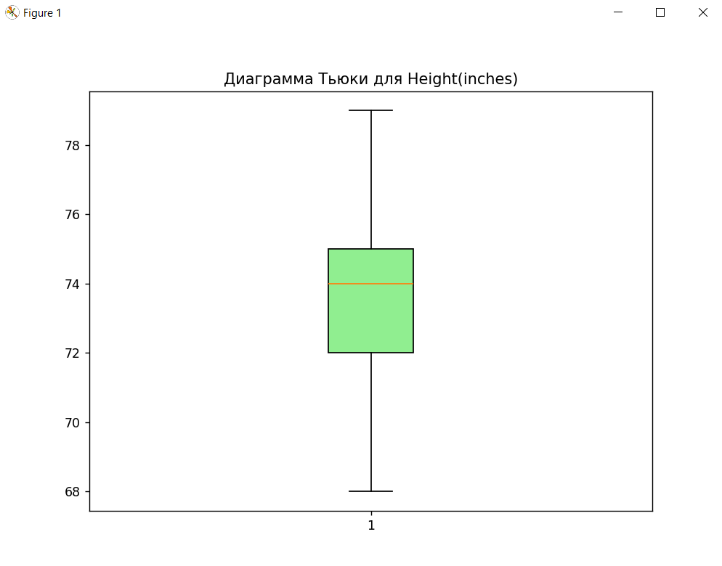
3 727 Humberto\_Quintero HOU ... 73 190.0 27.56

4 916 Shawn\_Hill WAS ... 74 185.0 25.84

[5 rows x 7 columns]  
  
Для облегчения дальнейшей работы выведем имена полей:  
>>> print(sample.columns.to\_list())

['Unnamed: 0', 'Name', 'Team', 'Position', 'Height(inches)', 'Weight(pounds)', 'Age']

* 1. **Разведочный анализ данных**

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

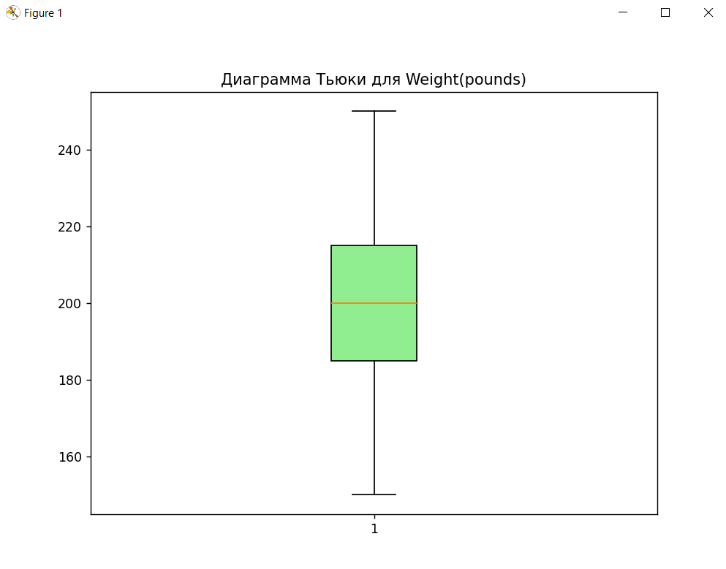
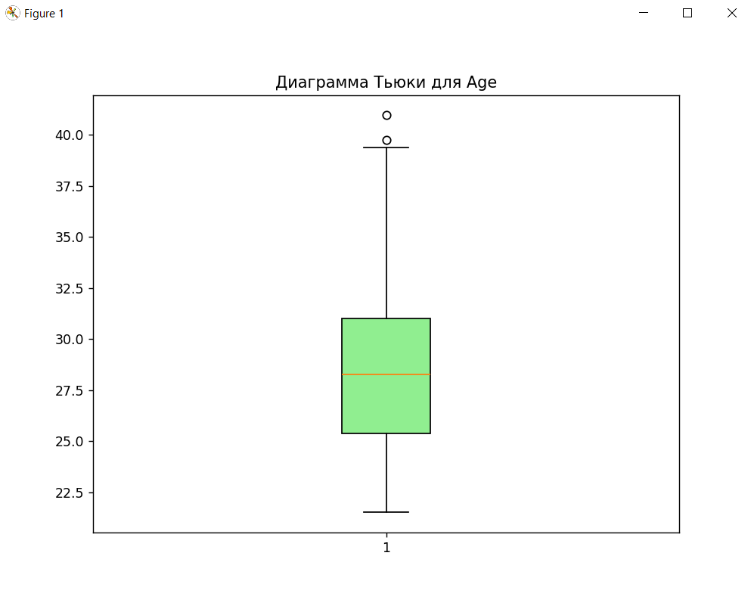
****Диаграмма для роста: медиана находится чуть ближе к верхнему краю ящика, что значит, что большая часть данных сосредоточена в нижней части диапазона. Выбросов нет, усы примерно одинаковой длины, то есть нет аномально высоких или аномально низких людей.

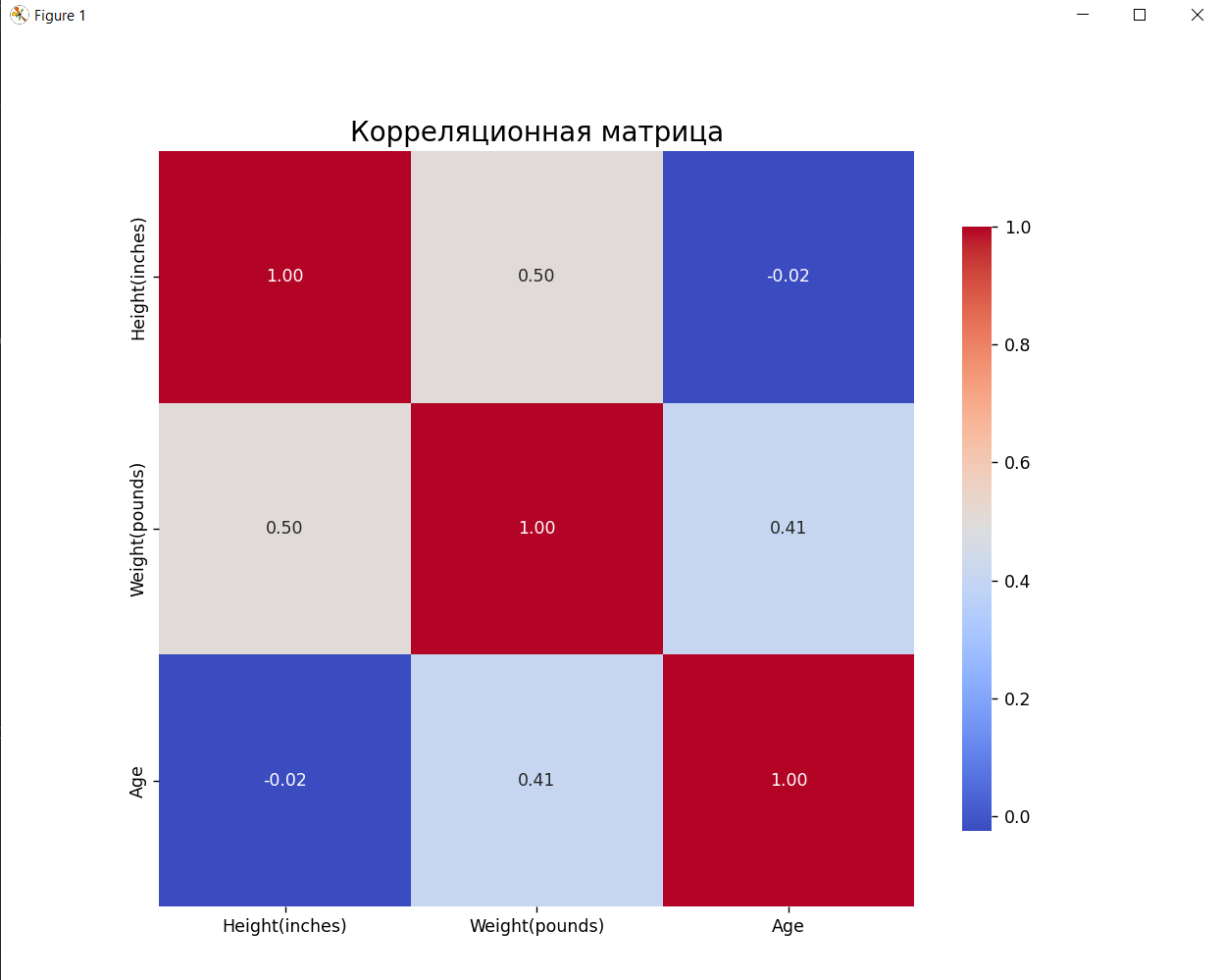
Диаграмма для веса: медиана находится посередние ящика. Выбросов нет, усы одинаковые, ящик в целом выглядит симметрично. Если ящик симметричен относительно медианы, это обычно указывает на нормальное распределение данных. Выбросов нет, усы одинаковой длины.

Диаграмма для возраста: медиана находится посередине ящика. Верхний ус заметно длиннее, к тому же, есть две верхних аномалии. Это значит, что наличие нескольких человек сильно старше медианного значения оказывает влияние на внешний вид ящика.

Код для построения описан в функции tukey() файла task1.py

* + *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.*

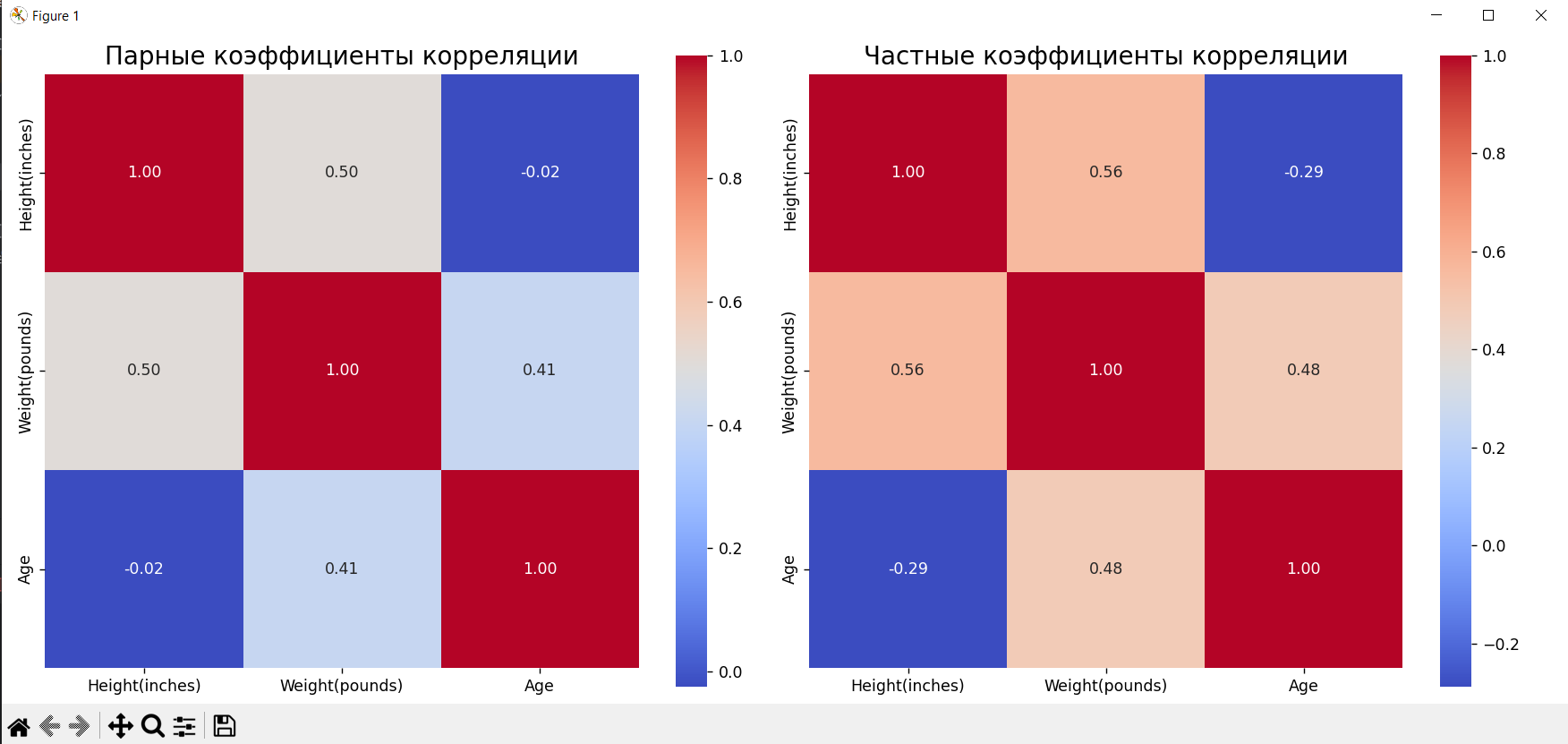
Код для построения описан в функции make\_heatmap() файла task1.py



**Вывод:**

* На главной диагонали, вполне естественно, расположены единицы, так как корреляция между величиной и этой же величиной равна единице.
* Корреляции между возрастом и ростом практически нет (-0.02). Можно обосновать это тем, что корреляция между возрастом и ростом в жизни человека проявляется в детстве и юношестве (примерно до 21 года; зависимость прямая), и в пожилом возрасте (примерно после 60 лет, зависимость обратная, т.к. мышцы становятся слабее и появляется сутулость). Но по диаграмме Тьюки видно, что все спортсмены имеют возраст примерно от 21 до 41 лет (с учетом выбросов), и не попадают в вышеописанные диапазоны возраста.
* Корреляция между весом и возрастом есть (0,41) и по шкале Чеддока она умеренная. Вероятно, это можно связать с замедлением метаболизма с возрастом, или увеличением мышечной массы в связи с продолжительным опытом занятий спортом.
* Корреляция между весом и ростом есть (0.5), также умеренная. Очевидно, у разных людей со спортивным телосложением примерно одинаковый ИМТ, а следовательно, чем больше рост, чем больше вес.
  + *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.*

Код для построения описан в функции pheatmap() файла task1.py



Вывод:

* Если рассматывать частную корреляцию между ростом и возрастом без учёта веса, видно, что взаимосвязь между ними стала более выраженной (слабая по Чеддоку, обратная). Это связано с тем, что между ростом и весом была положительная корреляция, между весом и возрастом была поожительная корреляция, и этот факт повлиял на корреляцию между ростом и возрастом.
* Если рассматривать частную корреляцию между весом и ростом, она также увеличилась. Это может быть связано с тем, что у молодых спортсменов рост, как правило, еще может изменяться, а у взрослых – уже стабилен. Без учета возраста, этот эффект мог сглаживаться.
* Рост был прямо коррелирован с весом, вес – прямо коррелирован с возрастом, поэтому исключая влияние веса, корреляция между весом и возрастом приобрела более выраженный обратный характер.
  + *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.*

Код для построения описан в функции KolmSmir() файла task1.py

H0 – признаки распределены по нормальному закону

H1 – признаки распределены по какому-то другому закону

{'Признак': 'Height(inches)', 'KS-статистика': 0.10546437276670123, 'Критическое значение': 0.12682065392068267, 'Результат': 'Нулевая гипотеза не отвергается'}

{'Признак': 'Weight(pounds)', 'KS-статистика': 0.11991115852136391, 'Критическое значение': 0.12682065392068267, 'Результат': 'Нулевая гипотеза не отвергается'}

{'Признак': 'Age', 'KS-статистика': 0.08548161328256504, 'Критическое значение': 0.12682065392068267, 'Результат': 'Нулевая гипотеза не отвергается'}

Как видно, все признаки распределены нормально. Более отчетливо это заметно для возраста, а менее отчетливо – для веса.

Вывод:

* В целом данные распределены однородно, разве что у возраста несколько игроков значительно старше среднего.
* Наиболее отчетливая корреляция между весом и ростом. Эти переменные больше всего взаимосвязаны, а также наиболее информативны.
* В целом данные распределены близко к нормальному закону, без особо сильных перекосов.
  1. **, 1.5 Регрессионный анализ**

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).

Код для выполнения задания в функции linear() файла task2.py

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.

Для парной регрессии:

MSE: 359.35592058658244

R2: 0.26706732522785337

Скорректированное R2: 0.2321657692863226

Стандартная ошибка: 18.95668537974354

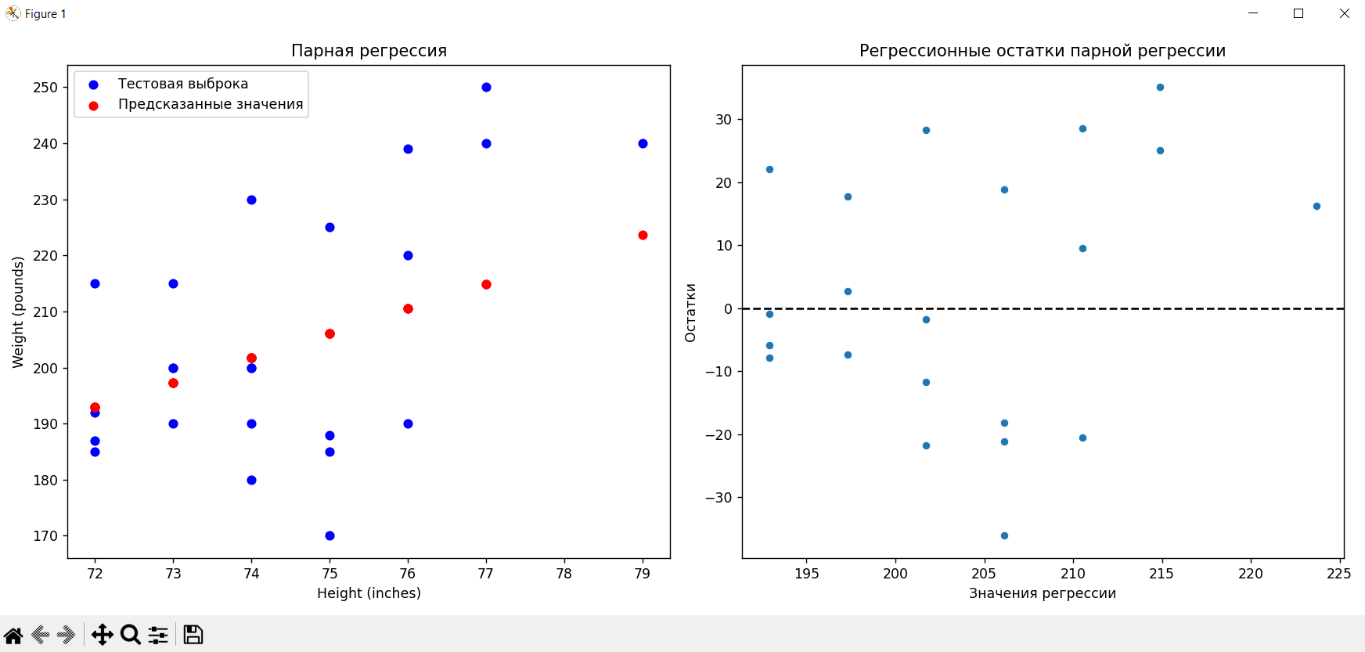
Для множественной регрессии:

MSE: 254.45588378172764

R2: 0.4810186201823897

Скорректированное R2: 0.4291204822006287

Стандартная ошибка: 15.951673385000323



Для множественной регрессии R2 больше (ближе к единице), а стандартная ошибка – меньше, так что она более адекватна.

* 1. **Критерий адекватности Дурбина-Уотсона**

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

Вычиление добавлено в функцию linear() файла task2.py

Критерий Колмогорова-Смирнова по отношению к остаткам:

{'Признак': 'Residuals', 'KS-статистика': 0.11473861383981077, 'Критическое значение': 0.2835796031176217, 'Результат': 'Нулевая гипотеза не отвергается'}

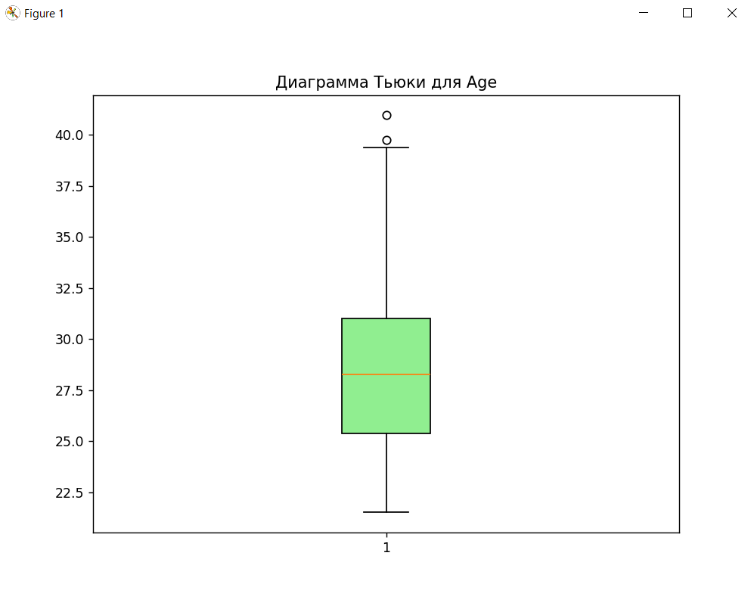
Значение статистики Дурбина-Ватсона: 1.9637923401514257

Как видно, полученное значение статистики Дурбина-Ватсона очень близко к идеальному значению, равному двум. Это показывает, что среди регрессионных остатков практически не наблюдается автокорреляция. В контексте нашего исследования это хорошо, т.к. сильная автокорреляция обычно показывает на то, что модель не адекватна и не учитывает все факторы, влияющие на зависимую переменную, и что оценки получатся смещенными/несостоятельными.

* 1. **Выбросы**

По диаграммам Тьюки из пункта 1.2 видно, что выбросы есть только в величине Age. Видимо, есть два человека, которые старше max + 1.5 (Q3-Q1).

Они не очень сильно влияют на дальнейшие расчеты, т.к. расположены достаточно близко к верхнему усу.

****

* 1. **Предложения по улучшению регрессионной модели**
* Можно удалить выбросы из Age, чтобы они не влияли на предсказания регрессии.
* Можно попробовать выбрать в качестве входа и в качестве выхода другие признаки, чтобы увидеть, как это повлияет на стандартную ошибку и на коэффициент детерминации
* Можно попробовать другую регрессионную модель – полиномиальную или любую нелинейную.

1. **Датасет «*Ирисы Фишера*»**

Перед выполнением, поскольку уже определенные функции подразумевали ввод имени файла, запишем датасет в файл .csv

>>> from sklearn.datasets import load\_iris

>>> import pandas as pd

>>> import numpy as np

>>> iris = load\_iris()

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

>>> iris\_pd.to\_csv("D:\\STUDY\\LVL3\\Интеллектуальный анализ данных\\lr1\\iris.csv")

Имена заголовков и вид данных:

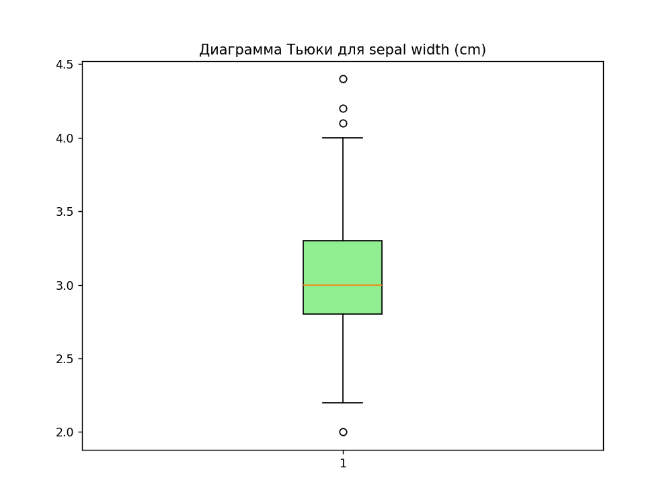
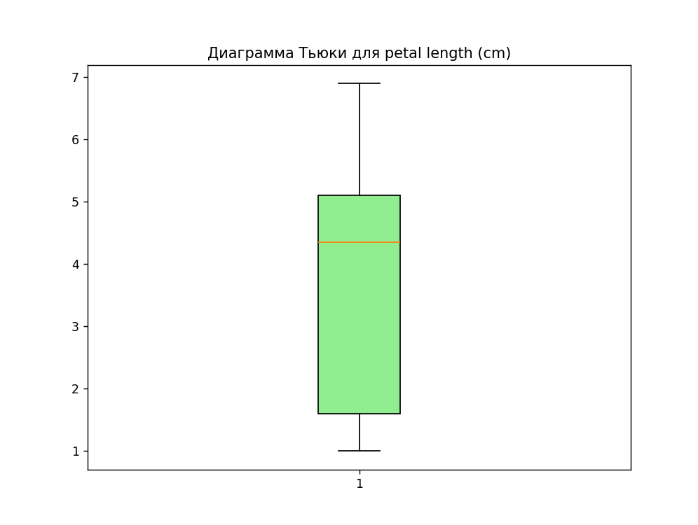
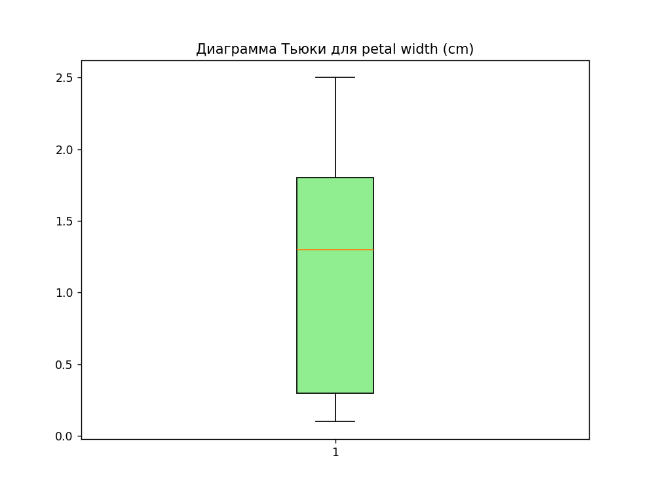
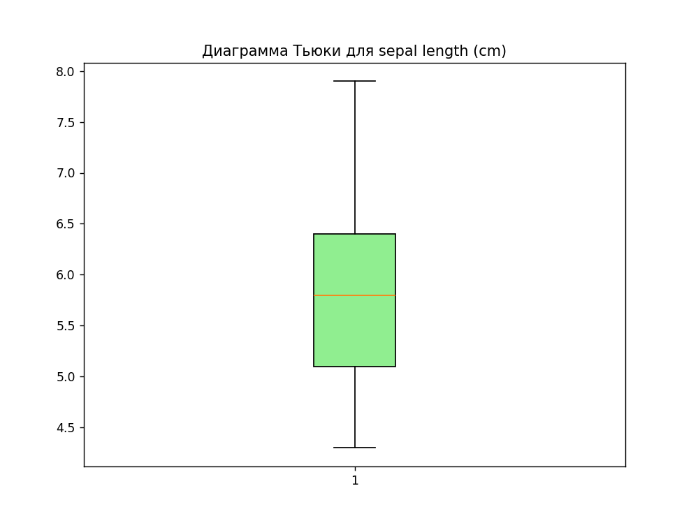
>>> iris\_pd.head

<bound method NDFrame.head of sepal length (cm) sepal width (cm) ... petal width (cm) target

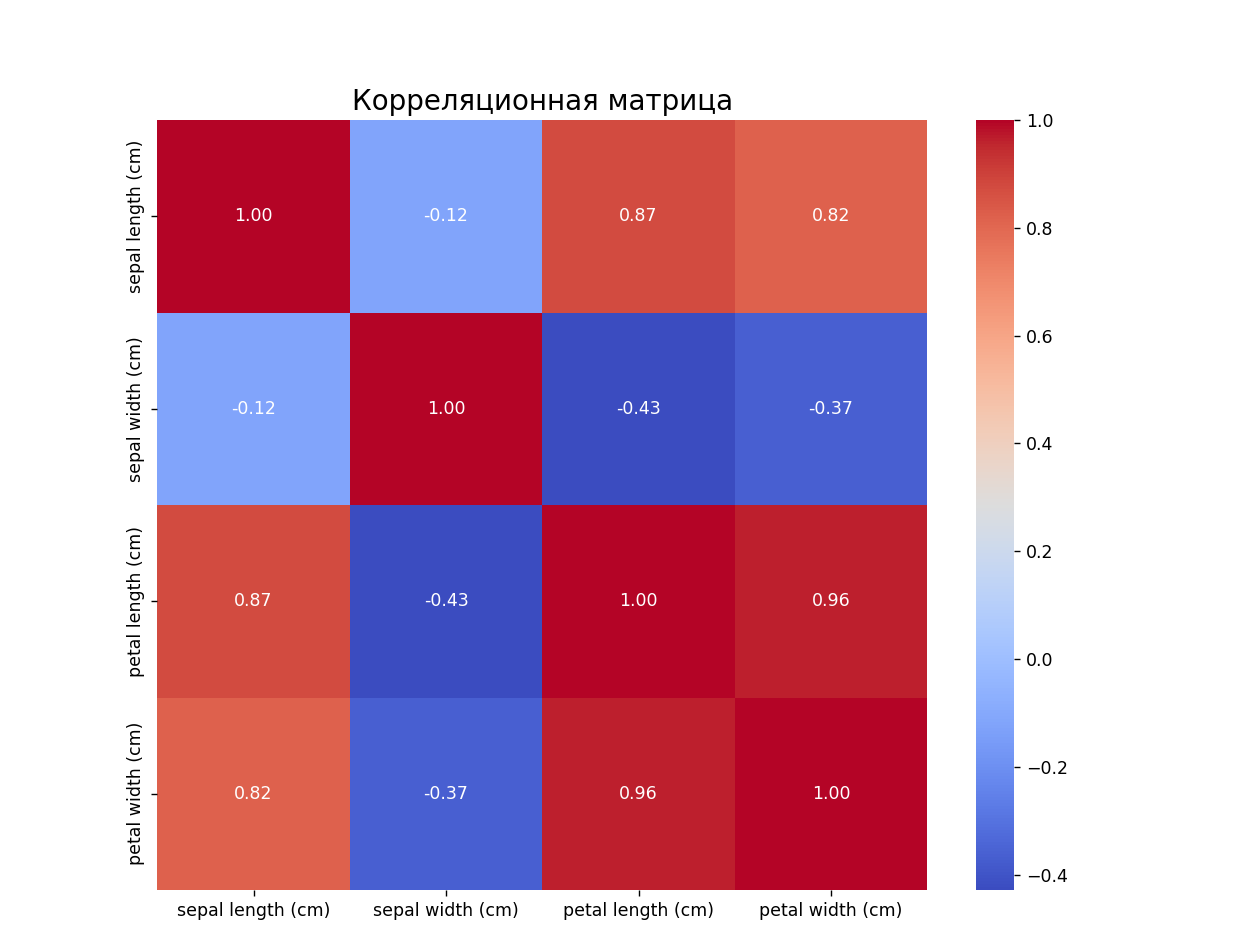
0 5.1 3.5 ... 0.2 0.0

1 4.9 3.0 ... 0.2 0.0

* 1. **Разведочный анализ данных**

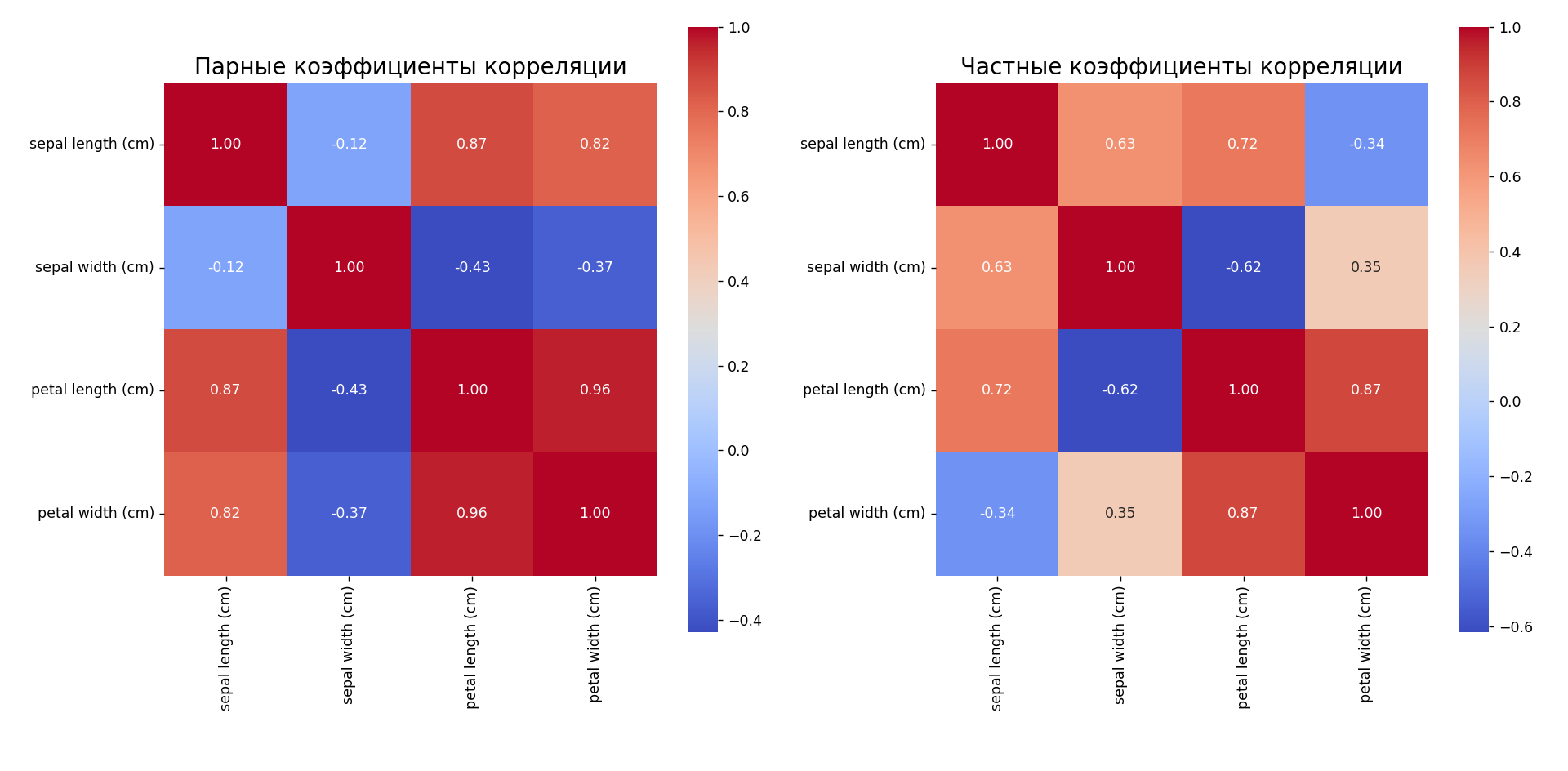
*- построить диаграмму Тьюки, оценить диапазон изменения данных (для этого используем matplotlib). В отчёт включить диаграмму и написать к ней выводы.*

* Для длины чашелистника: медиана находится посередине ящика, то есть распределение приближено к нормальному. Верхний ус заметно длиннее нижнего, но выбросов сверху все-таки нет.
* Для ширины чашелистника: видно, что сам ящик довольно узкий, то есть данные в окрестностях медианы одного порядка и отличаются не слишком сильно. При этом усы по сравнению с ящиком достаточно длинные, и сверху есть три выброса.
* Для длины лепестка и ширины лепестка диаграммы Тьюки выглядят очень похоже. Ящик продолговатый, то есть в центральных 50-ти процентах есть довольно большая дисперсия. Верхний ус сильно длиннее нижнего: есть цветки, которые сильно больше медианного размера. Сама же медиана смещена кверху, это указывает на скошенность распределения влево (negative skew). Это значит, что в данных присутствует больше больших значений, а меньшие значения более редки.
  + *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.*



**Вывод:**

* Длина чашелистника – ширина чашелистника (-0.12). Есть незначительная отрицательная корреляция.
* Длина чашелистника – длина лепестка (0.87). Есть сильная положительная корреляция, что вполне логично: обычно цветки имеют более-менее пропорциональные габариты, и чем больше весь цветок, тем пропорционально больше отдельные части.
* Длина чашелистника – ширина лепестка (0.82). Сильная положительная корреляция, практическое объяснение такое же, как в предыдущем случае.
* Ширина чашелистника – длина лепестка (-0.43). Умеренная отрицательная корреляция, вероятно, может быть связана с тем, что замерялись также и молодые бутоны, которые еще не распустились или распустились частично, и сделовательно, чашелистник гораздо больше и закрывает лепестки.
* Ширина чашелистника – ширина лепестка (-0.37). Умеренная отриательная корреляция. Обьяснение аналогично предыдущему пункту.
* Длина лепестка – ширина лепестка (0,96). Очень выраженная положительная корреляция объясняется по аналогии с первым пунктом вывода.
  + *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.*



**Вывод:**

* Длина чашелистника – ширина чашелистника (0.63). При исключении отрицательного влияния на ширину чашелистика со стороны длины лепестка и ширины лепестка, корреляция выровнялась в положительную сторону
* Длина чашелистника – длина лепестка (0.72). Корреляция не изменилась существенно.
* Длина чашелистника – ширина лепестка (-0.34). При исключении влияния длины лепестка на ширину лепестка, корреляция приобрела умеренный отрицательный характер.
* Ширина чашелистника – длина лепестка (-0.62). При исключении резко положительного влияния между длиной и шириной лепестка, значение приобрело более обратный характер
* Ширина чашелистника – ширина лепестка (0.35). При исключении отрицательного влияния длины лепестка на ширину чашелистика, корреляция приобрела умеренный прямой характер.
* Длина лепестка – ширина лепестка (0,87). Аналогично предыдущим пунктам.
  + *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.*

{'Признак': 'sepal length (cm)', 'KS-статистика': 0.08865361377316228, 'Критическое значение': 0.11104353500617076, 'Результат': 'Нулевая гипотеза не отвергается'}

{'Признак': 'sepal width (cm)', 'KS-статистика': 0.10565879047721255, 'Критическое значение': 0.11104353500617076, 'Результат': 'Нулевая гипотеза не отвергается'}

{'Признак': 'petal length (cm)', 'KS-статистика': 0.19815409613999851, 'Критическое значение': 0.11104353500617076, 'Результат': 'Нулевая гипотеза отвергается'}

{'Признак': 'petal width (cm)', 'KS-статистика': 0.17283424907904044, 'Критическое значение': 0.11104353500617076, 'Результат': 'Нулевая гипотеза отвергается'}

Длина и ширина чашелистика распределены нормально, а длина и ширина лепестка – нет. Датасет «Ирисы Фишера» содержит три различных класса цветков: Setosa, Versicolor и Virginica. Каждый из этих классов имеет свои характерные размеры лепестков. В результате распределения для признаков длины и ширины лепестков могут не быть нормальными из-за того, что каждый класс вносит свою вкладку в общее распределение. Например, у Setosa лепестки гораздо короче и уже по сравнению с другими классами. Когда эти данные объединяются, появляется мультимодальность (несколько пиков), что сильно отклоняется от нормального распределения.

Вывод:

* Судя по диаграммам Тьюки, замеры габаритов лепестков более вариабельны, а замеры чашелистиков имеют гораздо более узкий диапазон изменения значений.
* Судя по тепловой карте, признаки лепестков являются более взаимозависимыми, чем признаки чашелистиков.
* Медианы для всех признаков находятся не строго по центру ящика, что говорит о наличии некоторой скошенности в распределении. Для лепестков скошенность чаще всего выражена, что может указывать на асимметричное распределение данных.
* Тепловые карты показывают, что наиболее информативными переменными для разделения видов ирисов являются длина и ширина лепестков. Они имеют наибольшие коэффициенты корреляции и более четко разделяют классы растений, в отличие от признаков чашелистиков, которые менее информативны для классификации.

**2.2, 2.3, 2.5. Регрессионный анализ**

Значение статистики Дурбина-Ватсона: 1.5162818410676042

Для парной регрессии:

MSE: 0.18860699779525775

R2: 0.11687062059345488

Скорректированное R2: 0.08533028561464973

Стандартная ошибка: 0.4342890716967879

Для множественной регрессии:

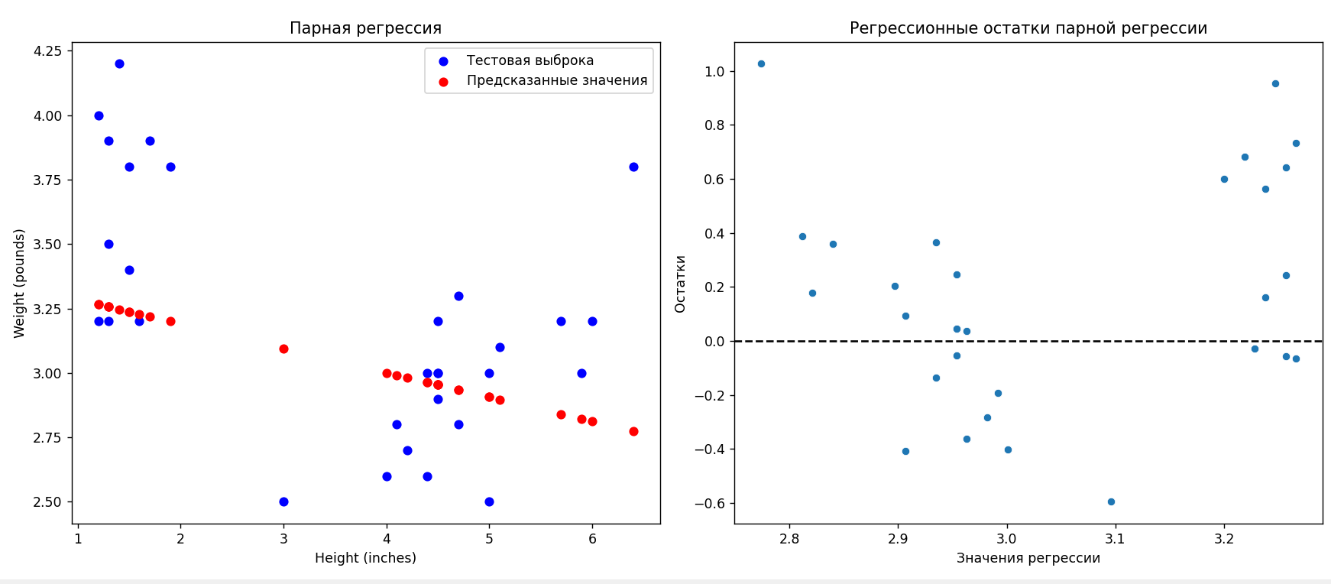
MSE: 0.10446148239795305

R2: 0.5108717852444837

Скорректированное R2: 0.4544339143111549

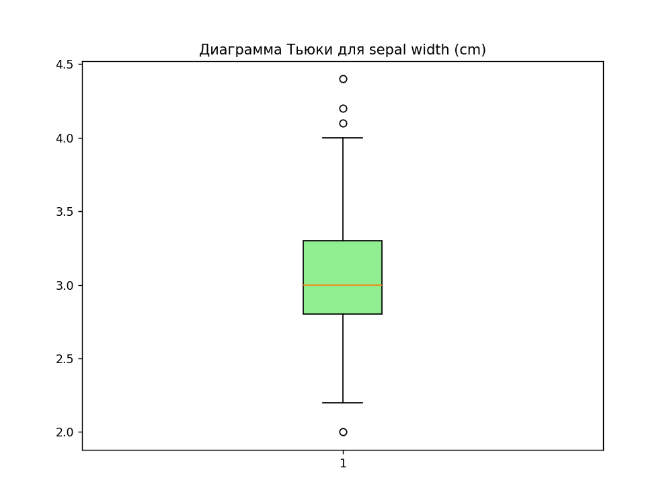
Стандартная ошибка: 0.32320501604701785

Критерий Колмогорова-Смирнова по отношению к остаткам:

{'Признак': 'Residuals', 'KS-статистика': 0.08900277388184441, 'Критическое значение': 0.24830089273567532, 'Результат': 'Нулевая гипотеза не отвергается'}

1. Значение статистики Дурбина-Ватсона 1.51 несколько ниже, чем идеальное значение, равное двум. Есть некоторая положительная автокорреляция остатков (значения остатков имеют тенденцию следовать друг за другом в одном направлении). Это несколько противоречит предположению регрессионного анализа о нелинейности остатков.
2. Множественная регрессия опять показывает более качественный результат. Для нее R2 ближе к единице, а стандартная ошибка ближе к нулю.
3. Регрессионные остатки имеют нормальное распределение, нулевая гипотеза об этом H0 при критерии Колмогорова-Смирнова принимается. Это хорошо, т.к. соответствует предположению регрессионного анализа.

**2.4. Выбросы.**

Выбросы присутствуют только для величины sepal width. Для него значения распределены в относительно узком диапазоне по сравнению с другими признаками. В условиях такого узкого диапазона даже небольшие отклонения от центральных значений могут считаться выбросами на основе метода Тьюки. И, опять же, каждый из трех сортив имеет свои габариты, что влияет на равномерность данных.

**2.6. Предложения по улучшению регрессионной модели.**

* Использовать стандартизованные данные. Габариты отдельных элементов бутона, конечно, имеют одинаковый порядок величин, но, тем не менее, могут негативно влиять на результат регрессии из-за нестандартизованности.
* Устранение выбросов из sepal width.

**3. Общие задания**

**3.1. Повторный расчет регрессии для выборки без выбросов.**

Расчет делается для датасета «Ирисы Фишера». Как было показано ранее, выбросы есть только у величины sepal width, которая в нашей модели является выходной переменной. Задание сделано путем внесения изменений в функцию linear () файла task2.pу.

Значение статистики Дурбина-Ватсона: 2.2890044401997685

Для парной регрессии:

MSE: 0.15077295906365704

R2: 0.1392599863159445

Скорректированное R2: 0.10851927154151397

Стандартная ошибка: 0.3882949382410966

Для множественной регрессии:

MSE: 0.08604342491628998

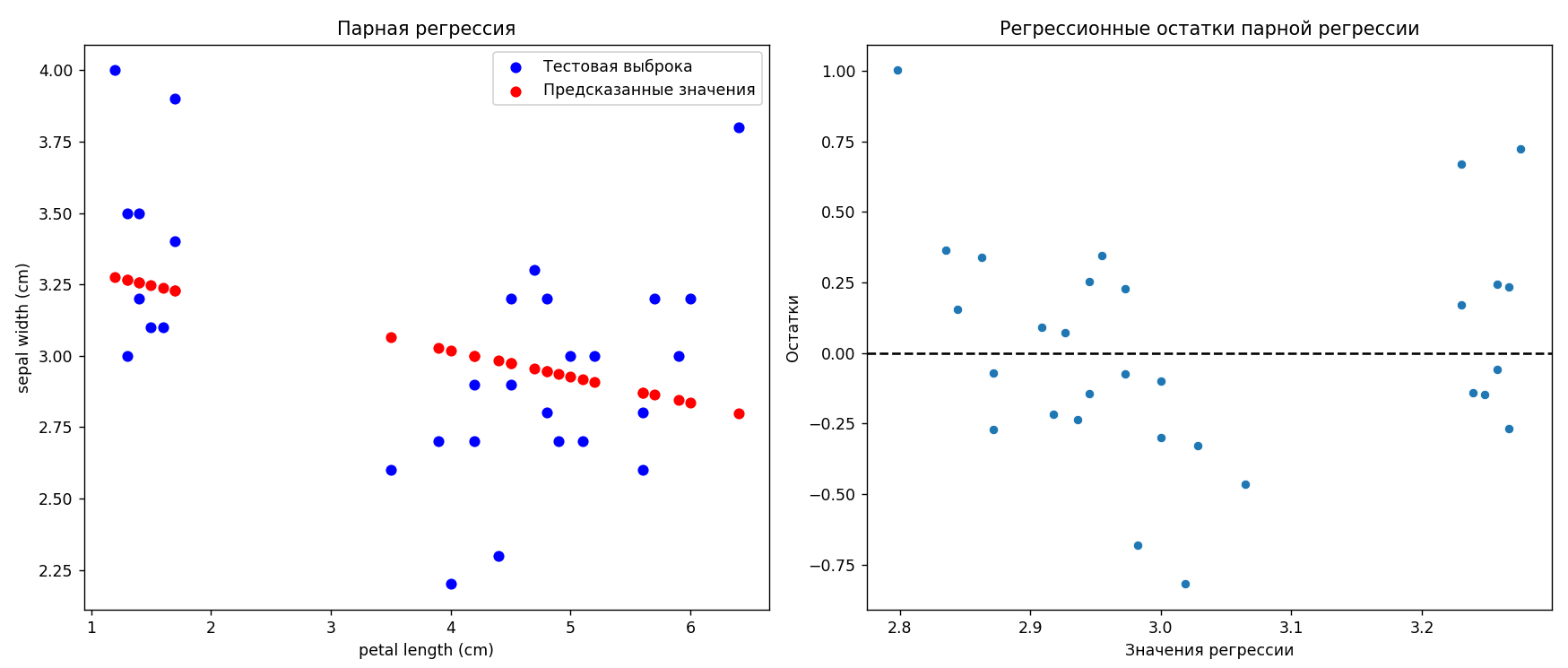
R2: 0.5087911041886395

Скорректированное R2: 0.452113154671944

Стандартная ошибка: 0.293331595496104

Критерий Колмогорова-Смирнова по отношению к остаткам:

{'Признак': 'Residuals', 'KS-статистика': 0.1102589109308757, 'Критическое значение': 0.24830089273567532, 'Результат': 'Нулевая гипотеза не отвергается'}



Видно, что наше редактирование выборки помогло: значение критерия Дурбина-Ватсона теперь больше двух, то есть автокорреляция остатков стала обратной, но тем не менее это число ближе к двум, чем 1.51, так что это хороший результат. Вместе с этим для обеих регрессий увеличились, пусть и незначительно, значения R2, и для обеих регрессий уменьшились стандартные ошибки.

* 1. **Расчет корреляционной матрицы для еще одного датасета.**

Провести корреляционный анализ и сделать выводы о целесообразности построения регрессии для датасетов

Датасет: *Chemical\_process.sta*

Видно, что два параметра вообще не имеют связи между собой. Можно предположить, что это временной ряд или просто стохастический процесс. Регрессионный анализ тут был бы бесполезен.

