МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
Факультет прикладной математики и информатики

Кафедра математического моделирования и анализа данных

Корреляционный и регрессионный анализ данных Фишера

Отчет по лабораторной работе №2
студентки 3 курса 7 группы
Бовта Тимофея

Преподаватель
Сафиуллин Т. Т.

Минск, 2023

Оглавление

Корреляционный и регрессионный анализ данных Фишера

Цели

В ходе корреляционного анализа данных с помощью графических и статистических результатов исследовать зависимости между случайными величинами. В ходе регрессионного анализа данных с помощью графических и статистических результатов исследовать влияние одних случайных величин на другие.

Постановка задачи

Для данных Фишера:

1) Провести корреляционный анализ однородных данных;
2) Провести регрессионный анализ однородных данных;

Указания:

- использовать выборку значений переменных 1, 2, 3 без засорений для одного вида ириса 1, в регрессионной модели зависимой является первая переменная;
- Исследовать эффекты засорений на результаты корреляционного и регрессионного анализа, использовать выборку значений переменных 1, 2, 3 с засорением переменной 3 для одного вида ириса 1.

Результаты проведенного исследования

Корреляционный анализ однородных данных

Проведем исследование значения случайной величины SEPALLEN для ирисов 1-го вида (SETOSA). В первом отчете мы проверили эту выборку на нормальность. Эта информация понадобится нам в данном исследовании.

Сперва исследуем связь значения случайной величины SEPALLEN с остальными случайными величинами SEPALWID, PETALLEN.

	sepal length (cm)	sepal width (cm)	petal length (cm)
0	5.1	3.5	1.4
1	4.9	3.0	1.4
2	4.7	3.2	1.3
3	4.6	3.1	1.5
4	5.0	3.6	1.4
5	5.4	3.9	1.7
6	4.6	3.4	1.4
7	5.0	3.4	1.5
8	4.4	2.9	1.4
9	4.9	3.1	1.5

Таблица 1.1 – Набор данных, для которого проводится исследование

Для того чтобы увидеть общую картину построим матрицу диаграмм рассеяния (Рисунок 1.1).

На рисунке 1.1 хорошо просматривается корреляция между случайными величинами SEPALLEN и SEPALWID. Отчетливая связь между SEPALLEN и PETALLEN не прослеживается.

Для более точной характеристики проведем по-отдельности исследование корреляции между исследуемыми случайными величинами.

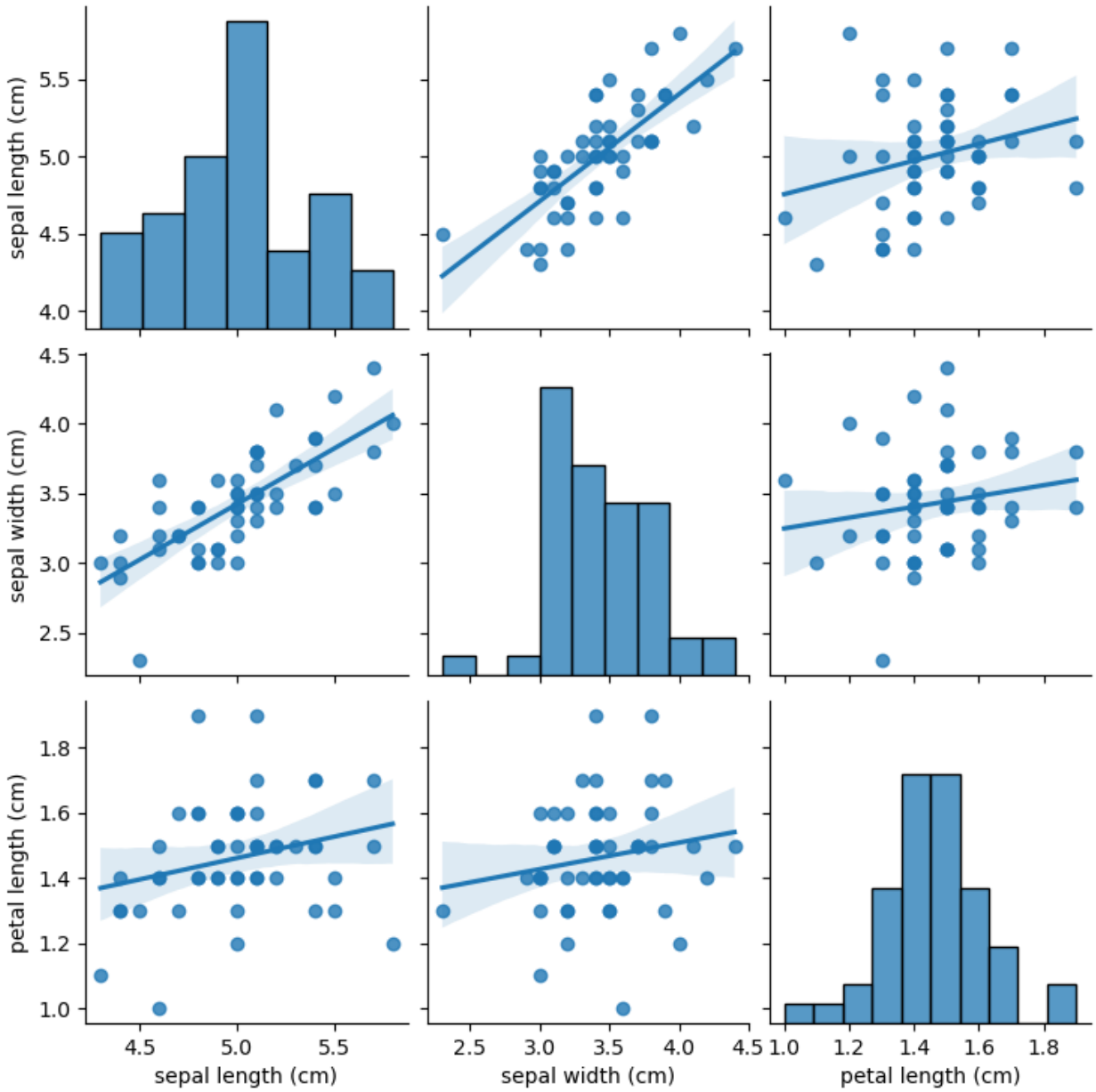

Рисунок 1.1 - Матрица диаграмм рассеяния

Корреляция между SEPALLEN и SEPALWID

Для исследования корреляции между случайными величинами SEPALLEN и SEPALWID построим диаграмму рассеяния (Рисунок 1.2).

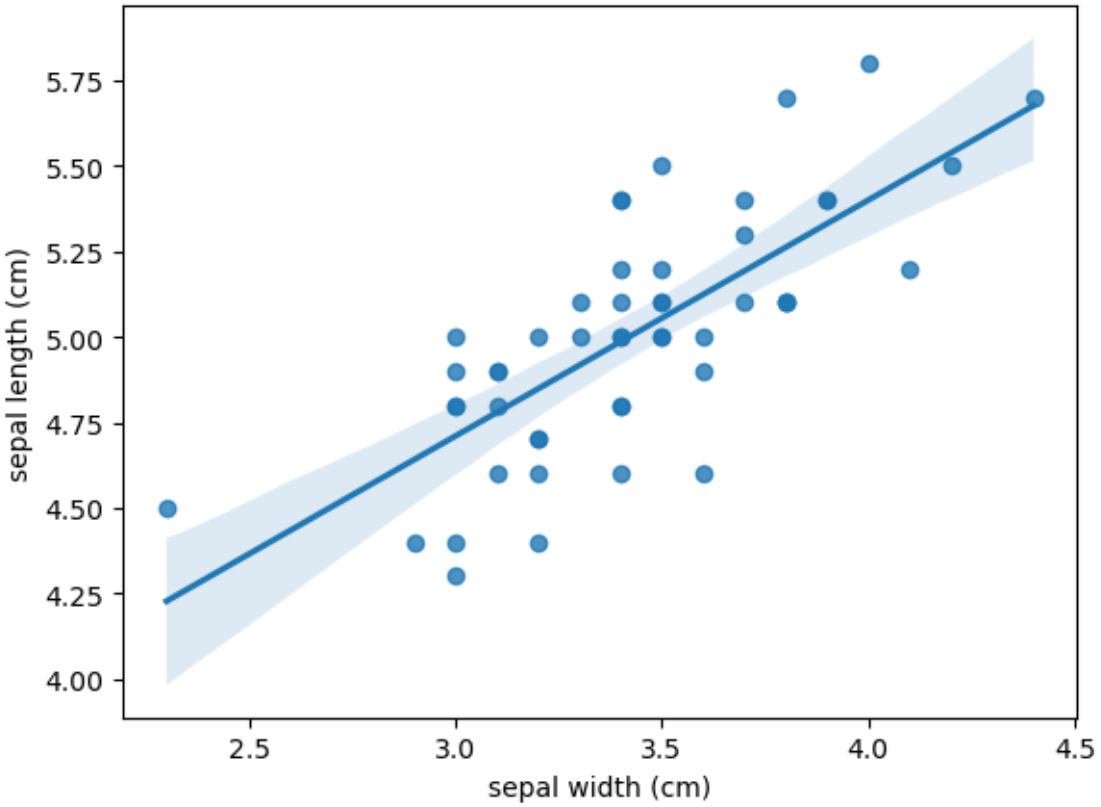

Рисунок 1.2 - Диаграмма рассеяния случайной величины SEPALLEN относительно SEPALWID

На графике (Рисунок 1.2) видно, что с увеличением переменной SEPALWID растет и SEPALLEN. Попробуем применить коэффициент корреляции Пирсона. Для того, чтобы применить его, нам необходимо выполнение следующих условий:

1. Исследуемые переменные X и Y должны быть распределены нормально.
2. Исследуемые переменные X и Y должны быть измерены в интервальной шкале или шкале отношений.
3. Количество значений в исследуемых переменных X и Y должно быть одинаковым.

Условия 2 и 3 очевидно выполняются. Проверим выполнение условия 1. В прошлом отчете мы приняли гипотезу о нормальности распределения случайной величины SEPALLEN. Теперь проверим, имеет ли SEPALWID нормальное распределение. Выдвинем гипотезу о том, что случайная величина SEPALWID имеет нормальное распределение. Построим гистограмму распределения этой случайной величины (Рисунок 1.3).

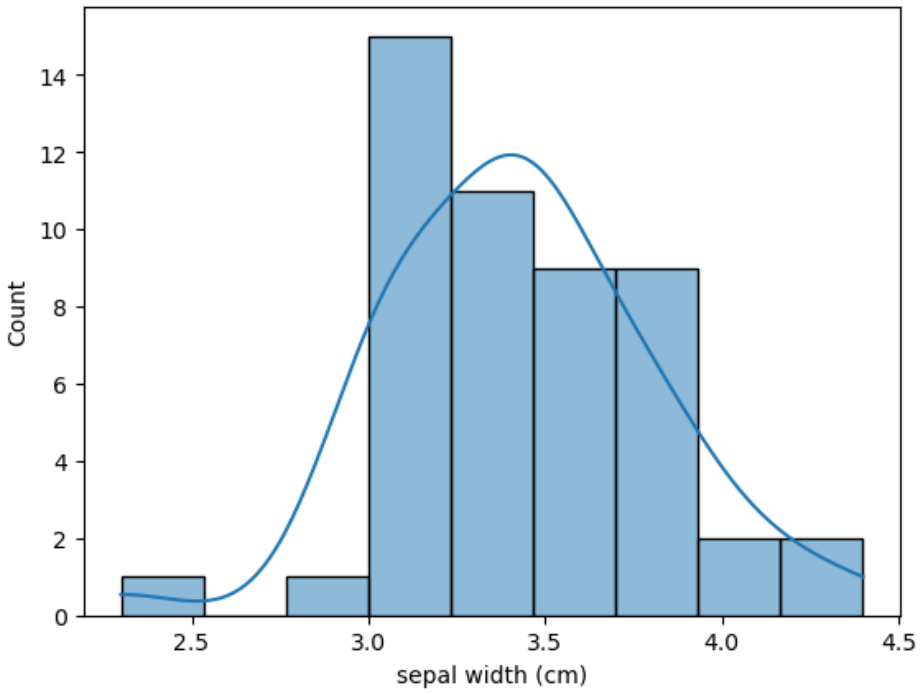

Рисунок 1.3 - Гистограмма распределения переменной SEPALWID

На гистограмме отсутствует ярко выраженная асимметрия. Проверим P-значение критериев Колмогорова-Смирнова, Шапиро-Уилка и Хи-квадрата Пирсона.

```
Shapiro-Wilk test: 0.2715126574039459
Kolmogorov-Smirnov test: 0.5486851446031328
Chi-Square test: 0.025349250037615446
```

Критерии Колмогорова-Смирнова, Шапиро-Уилка имеют P-значение больше 0.05. Соответственно мы не отклоняем гипотезу о нормальности распределения SEPALWID.

Таким образом, мы можем воспользоваться коэффициентом корреляции Пирсона. Также выдвинем гипотезу о равенстве истинного коэффициента корреляции нулю. Для проверки этой гипотезы воспользуемся t-критерием Стьюдента.

```
pearson correlation coefficient: 0.7425466856651596
T-test P-value: 5.666981945473851e-39
```

Мы получили коэффициент корреляции по Пирсону равный 0.74, при этом P-значение t-критерия очень мало, следовательно, мы имеем право отклонить гипотезу о том, что значение истинного коэффициента корреляции равно нулю. Таким образом, мы можем оценить истинный коэффициент по Пирсону, то есть

$$Corr(SEPALLEN, SEPALWID) = 0.74.$$

Корреляция между SEPALLEN и PETALLEN

Для исследования корреляции между случайными величинами SEPALLEN и PETALLEN построим диаграмму рассеяния (Рисунок 1.4).

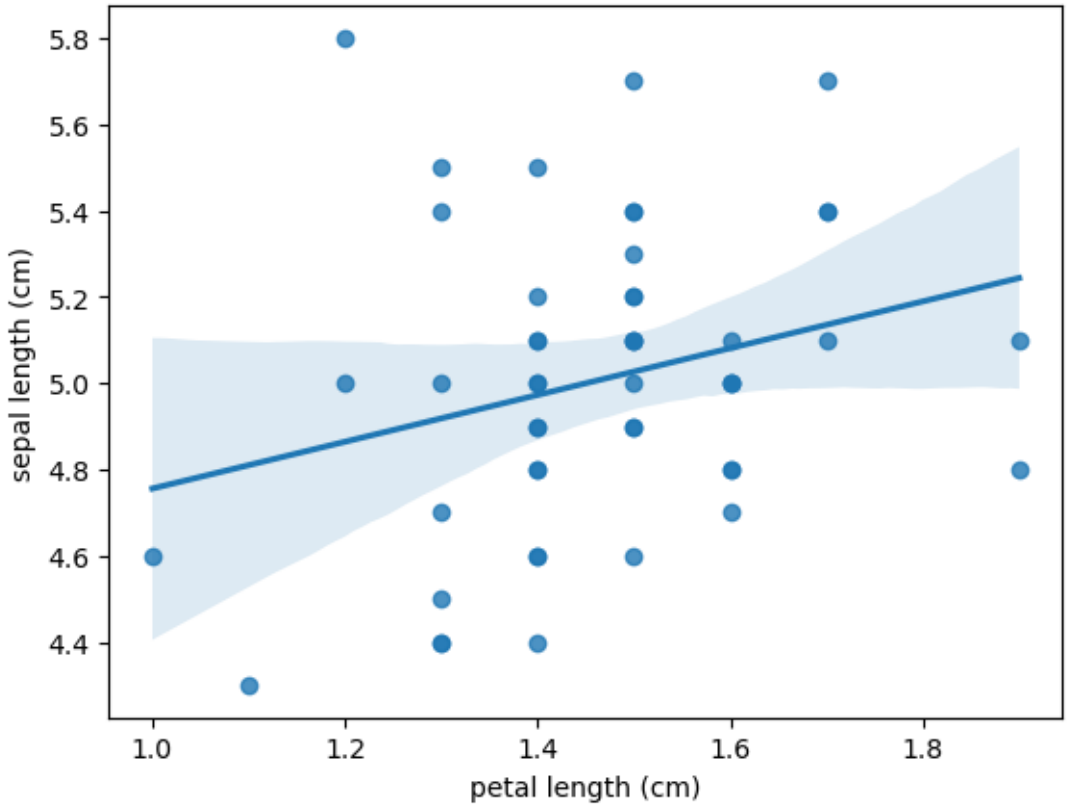

Рисунок 1.4 - Диаграмма рассеяния случайной величины SEPALLEN относительно PETALLEN

На этом графике (Рисунок 1.4) зависимость SEPALLEN от PETALLEN не просматривается.

Аналогично предыдущему исследованию воспользуемся коэффициентом корреляции Пирсона. Для этого нам необходимо, чтобы распределение PETALLEN было нормальным. Выдвинем гипотезу о том, что случайная величина PETALLEN имеет нормальное распределение. Построим гистограмму этого распределения (Рисунок 1.5).

На гистограмме отсутствует ярко выраженная асимметрия. Воспользуемся критериями Колмогорова-Смирнова, Шапиро-Уилка и Хи-квадрата Пирсона для проверки выдвинутой гипотезы.

```
Shapiro-Wilk test: 0.05481128394603729
Kolmogorov-Smirnov test: 0.5486851446031328
Chi-Square test: 5.7631046419157774e-05
```

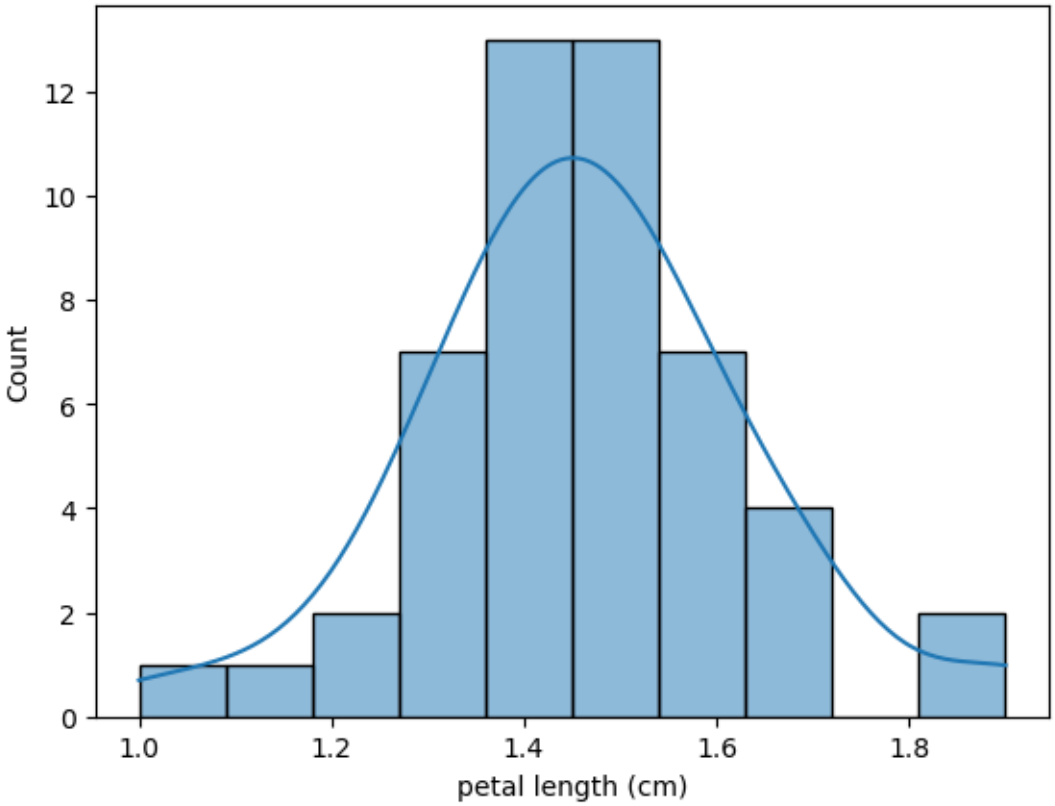

Рисунок 1.5 - Гистограмма распределения переменной PETALLEN

Критерии Колмогорова-Смирнова, Шапиро-Уилка имеют P-значение больше 0.05. Соответственно мы принимаем гипотезу о том, что случайная величина PETALLEN распределена нормально. Следовательно, мы можем воспользоваться коэффициентом корреляции Пирсона и применить t-критерий для проверки гипотезы о равенстве истинного коэффициента корреляции нулю.

```
pearson correlation coefficient: 0.2671757588687571
T-test P-value: 1.3140296801836079e-81
```

Коэффициент корреляции Пирсона между SEPALLEN и PETALLEN равен 0.26, а P-значение t-критерия сильно меньше 0.05. Что позволяет нам отклонить гипотезу о равенстве истинного коэффициента корреляции нулю и, следовательно, считать его значение по Пирсону. Отсюда

$$Corr(SEPALLEN, PETALLEN) = 0.26$$

Регрессионный анализ однородных данных

Проведем регрессионный анализ влияния факторов SEPALWID и PETALLEN на поведение SEPALLEN для ирисов 1-го вида (SETOSA), чтобы построить регрессионную модель для предсказания поведения зависимой переменной SEPALLEN. Для проведения регрессионного анализа нам необходимо выполнение следующих условий:

1. Линейная зависимость переменных;
2. Нормальное распределение остатков;
3. Гетероскедастичность;
4. Проверка на мультиколлинеарность;
5. Нормальное распределение переменных;

Линейная зависимость и нормальное распределение переменных следует непосредственно из корреляционного анализа. Остатки мы исследуем после построения регрессионной модели.

Для повышения качества модели сперва исключим мультиколлинеарные факторы. Чтобы найти мультиколлинеарные факторы (независимые переменные с сильной корреляционной связью ($\geq 0{,}7$)) построим корреляционную матрицу (Таблица 2.1).

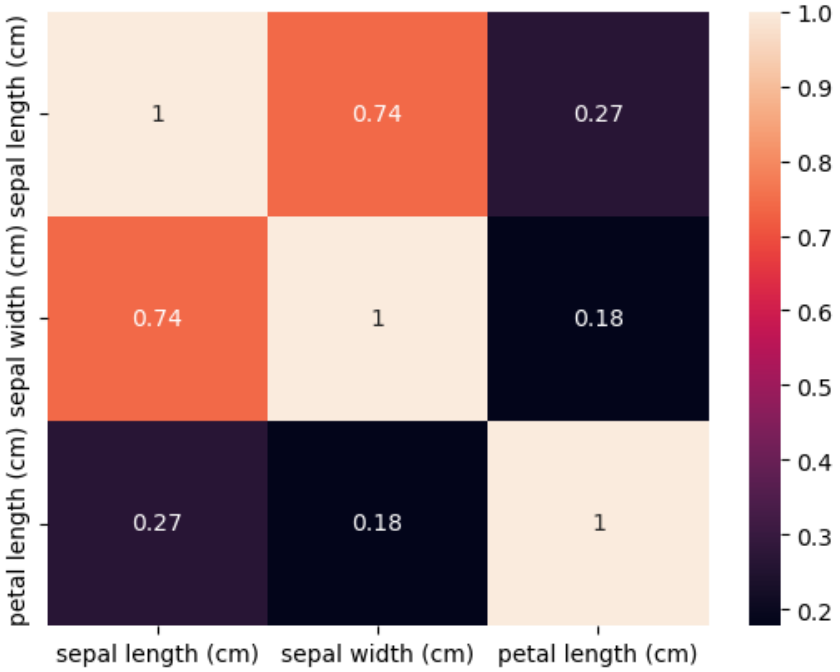

Таблица 2.1 - Корреляционная матрица

В таблице 2.1 заметно, что сильная корреляционная связь присутствует лишь между зависимой и независимой переменной, отсюда следует, что наша модель не содержит в себе мультиколлениарных факторов.

Проведем анализ регрессионного уравнения. Регрессионное уравнение в нашем случае имеет вид

$$sepallen = w_0 + w_1 * sepalwid + w_2 * petallen$$

Для анализа этого уравнения построим таблицу с уровнями значимости (Таблица 2.2). Значения коэффициентов были рассчитаны с помощью метода наименьших квадратов.

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.3037	0.385	5.979	0.000	1.529	3.079
sepal width (cm)	0.6674	0.090	7.387	0.000	0.486	0.849
petal length (cm)	0.2834	0.197	1.437	0.157	-0.113	0.680

Таблица 2.2 - Сводка регрессии по 2-ем переменным

Проанализируем таблицу 2.2. С помощью этой таблицы мы проверяем две гипотезы о равенстве коэффициентов w_1, w_2 нулю с помощью t-критерия. В колонке “P>|t|” мы получили P-значения для соответствующих гипотез. Следовательно, мы получаем следующее: гипотеза о равенстве коэффициента при SEPALWID отклоняется; гипотеза о равенстве коэффициента при PETALLEN не отклоняется. Значит мы можем исключить третье слагаемое из нашего уравнения регрессии и это никак не скажется на качестве нашей модели. Теперь наше уравнение регрессии имеет вид

$$sepallen = w_0 + w_1 * sepalwid$$

Снова построим таблицу с уровнями значимости (Таблица 2.3).

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.6390	0.310	8.513	0.000	2.016	3.262
sepal width (cm)	0.6905	0.090	7.681	0.000	0.510	0.871

Таблица 2.3 - Итоговая сводка регрессии

Из таблицы 2.3 мы видим, что у нас уже нет факторов, которые не оказывали бы существенное влияние на поведение зависимой переменной. Следовательно, мы можем построить регрессионное уравнение для

предсказания зависимой переменной, используя коэффициенты, найденные по методу наименьших квадратов:

$$sepallen = 2.639 + 0.6905 * sepalwid$$

Также при составлении таблицы с уровнями значимости для коэффициентов регрессии была составлена таблица 2.4, из которой мы выясняем, что значение коэффициента детерминации в данном случае R=0.55, то есть примерно 55% изменчивости зависимой переменной объясняется построенной моделью. Также была вычислена F-статистика и соответствующее P-значение для построенной модели. Это позволяет нам принять гипотезу о том, что наша модель в принципе позволяет нам объяснить поведение нашей зависимой переменной.

Dep. Variable:	sepal length (cm)	**R-squared:**	0.551
Model:	OLS	**Adj. R-squared:**	0.542
Method:	Least Squares	**F-statistic:**	58.99
Date:	Tue, 10 Oct 2023	**Prob (F-statistic):**	6.71e-10
Time:	22:51:19	**Log-Likelihood:**	1.7341
No. Observations:	50	**AIC:**	0.5319
Df Residuals:	48	**BIC:**	4.356
Df Model:	1		
Covariance Type:	nonrobust		

Таблица 2.4 - Итоговая сводка регрессии

Для предварительного анализа качества модели и применимости регрессионного анализа проведем анализ остатков (разностей фактических значений отклика и значений, предсказанных по уравнению регрессии).

Построим модель линейной регрессии и обучим её на нашей выборке, где SEPALLEN зависимая переменная, а SEPALWID – независимая переменная. После этого попробуем предсказать значения с помощью построенной модели.

Вычислим остатки, отняв от реальных значений SEPALLEN значения, предсказанные нашей моделью. Проведем проверку распределения остатков на нормальность, выдвинув гипотезу о том, что остатки имеют нормальное распределение. Для этого построим гистограмму распределения остатков (Рисунок 2.1).

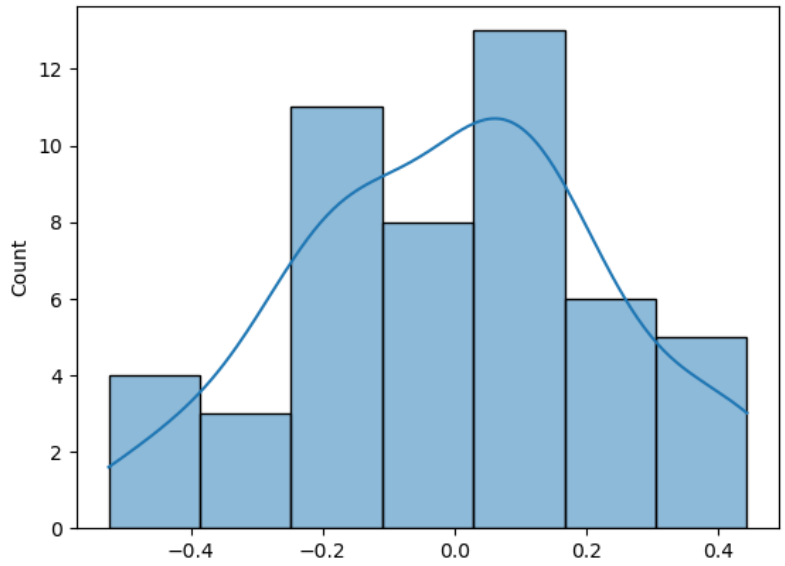

Рисунок 2.1 - Гистограмма распределения остатков

На гистограмме отсутствует ярко выраженная асимметрия, поэтому гипотеза о нормальности не отклоняется.

Для более полного изучения построим нормально-вероятностный график остатков (Рисунок 2.2).

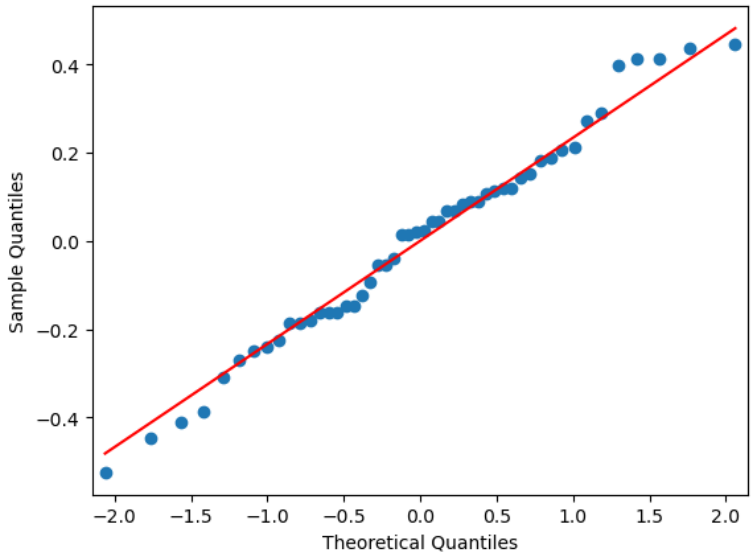

Рисунок 2.2 - Нормально-вероятностный график остатков

График (Рисунок 2.2) показывает, что систематических отклонений фактических данных от теоретической нормальной прямой не наблюдается. Это дает нам возможность принять гипотезу о нормальности распределения остатков. Следовательно, мы можем применять регрессионную модель к нашим данными.

Проверим отсутствие зависимости остатков от предсказанных по уравнению регрессии значений отклика. Для этого построим диаграмму рассеяния предсказанных значений SEPALLEN от остатков (Рисунок 2.3).

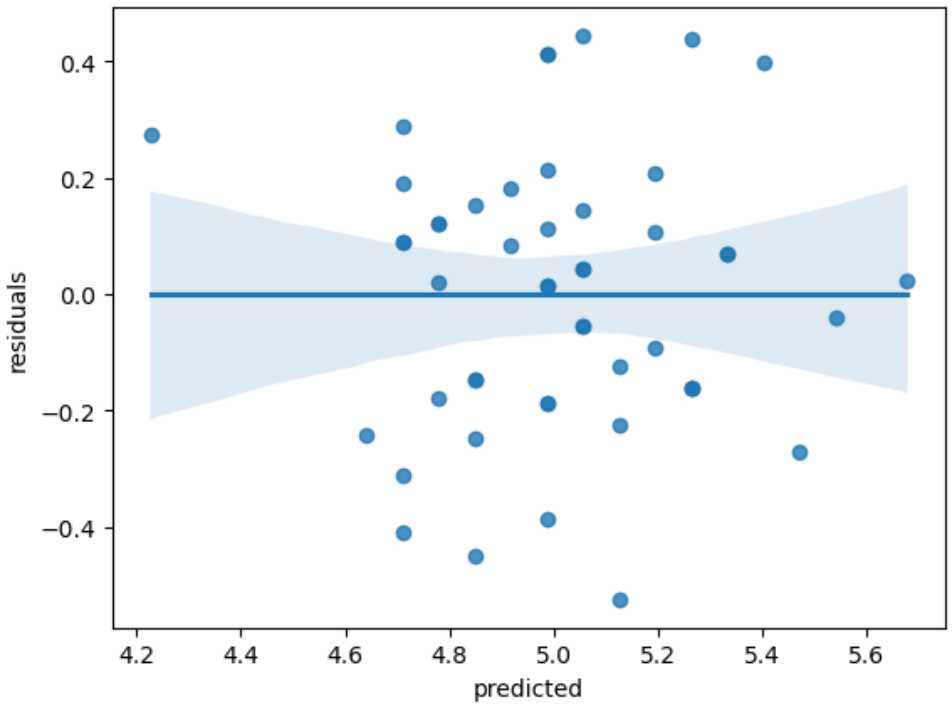

Рисунок 2.3 - Диаграмма рассеяния предсказанных значений и остатков

На этом графике (Рисунок 2.3) мы можем посмотреть на соотношение остатков и предсказанных значений. Нетрудно заметить, что точки не имеют системности в своем расположении. Соответственно, мы можем сказать, что остатки не зависят от предсказанных значений.

Оба условия выполнены. Следовательно, анализ остатков показал, что модель достаточно качественная. В заключение оценим приемлемость модели в целом с помощью дисперсионного анализа. Для этого проведем однофакторный дисперсионный анализ ANOVA (Таблица 2.5). В однофакторном дисперсионном анализе мы проверяем гипотезу о равенстве средних сумм квадратов двух групп. Первая группа будет сформирована из предсказываемых нашей моделью значений, а вторая – на основе примитивного прогноза, что все значения SEPALLEN будут равны среднему.

	sum_sq	df	F	PR(>F)
sepalwid	3.356885	1.0	58.99373	6.709843e-10
Residual	2.731315	48.0	NaN	NaN

Таблица 2.4 - ANOVA

Поскольку P-уровень значимости меньше 0.05, то мы можем утверждать, что наша модель приемлема и будет работать лучше, чем наивный прогноз по средним значениям.

Коэффициент детерминации $R^2 = 0{,}55$, значит 55% факторов мы учли в своей модели.

Следовательно, **мы можем использовать данную модель линейной регрессии для предсказывания значения случайной величины SEPALLEN**.

Корреляционный анализ однородных данных с засорениями

Исследуем влияние засорений в значениях случайной величины PETALLEN на результаты корреляционного анализа. Для получения новой выборки мы изменим 5 первых значений в исходной выборке на случайные значения.

	sepal length (cm)	sepal width (cm)	petal length (cm)
0	5.1	3.5	5.431271
1	4.9	3.0	2.524563
2	4.7	3.2	7.988240
3	4.6	3.1	8.362760
4	5.0	3.6	1.525346
5	5.4	3.9	1.700000
6	4.6	3.4	1.400000
7	5.0	3.4	1.500000
8	4.4	2.9	1.400000
9	4.9	3.1	1.500000
10	5.4	3.7	1.500000

Таблица 3.1 – Набор данных, для которого проводится исследование

Для того чтобы увидеть общую картину построим матрицу диаграмм рассеяния (Рисунок 3.1).

На рисунке 3.1 можно заметить, что засорения в переменной PETALLEN никоим образом не сказались на корреляции между SEPALLEN и SEPALWID. Исследуем поведение корреляции между SEPALLEN и PETALLEN.

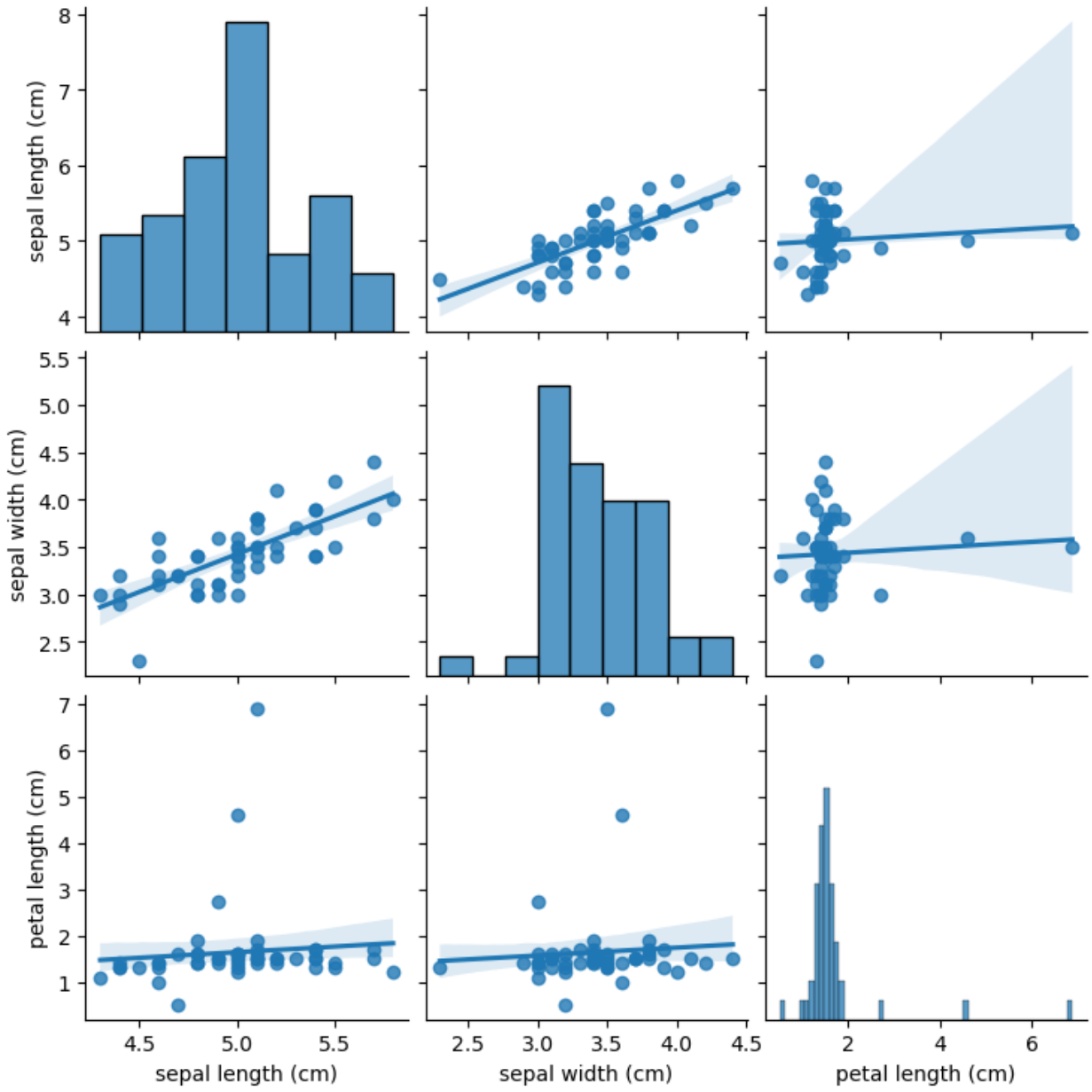

Рисунок 3.1 - Матрица диаграмм рассеяния

Корреляция между SEPALLEN и PETALLEN

Для исследования корреляции между случайными величинами SEPALLEN и PETALLEN построим диаграмму рассеяния (Рисунок 3.4).

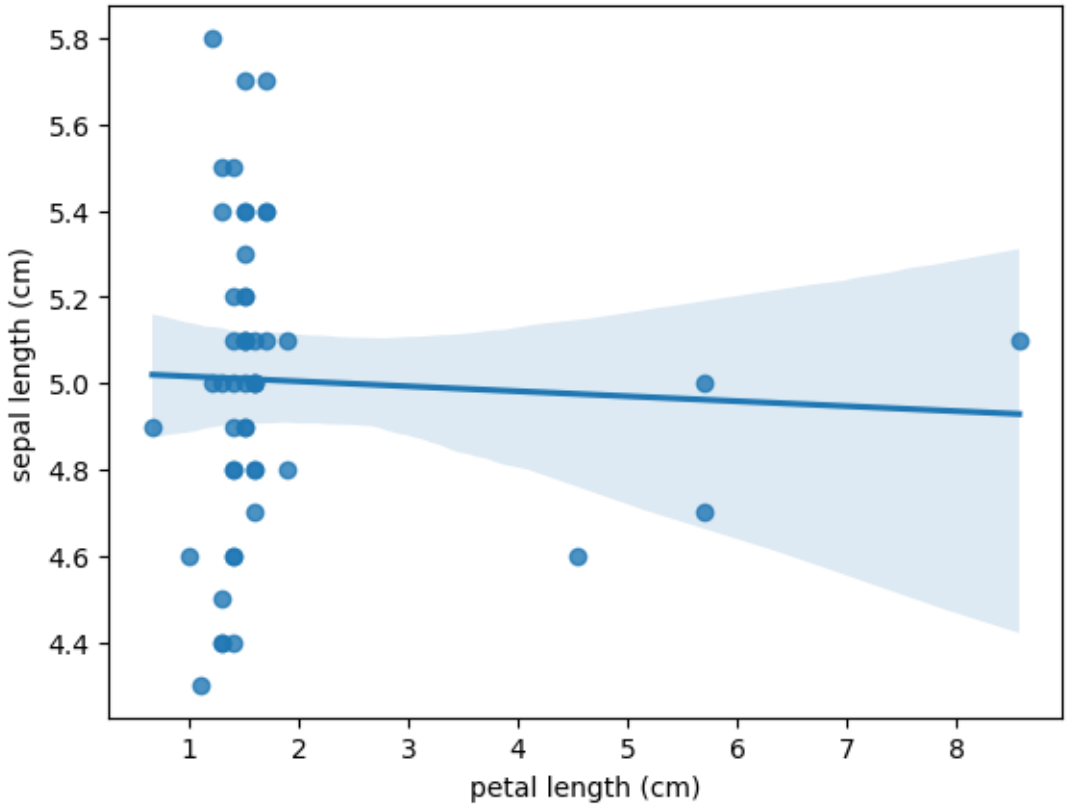

Рисунок 3.4 - Диаграмма рассеяния случайной величины SEPALLEN относительно PETALLEN

На этом графике (Рисунок 3.4) зависимость SEPALLEN от PETALLEN не просматривается.

Аналогично предыдущему исследованию воспользуемся коэффициентом корреляции Пирсона. Для этого нам необходимо, чтобы распределение PETALLEN было нормальным. Выдвинем гипотезу о том, что случайная величина PETALLEN имеет нормальное распределение. Построим гистограмму этого распределения (Рисунок 3.5).

На гистограмме присутствует ярко выраженная асимметрия. Воспользуемся критериями Колмогорова-Смирнова, Шапиро-Уилка и Хи-квадрата Пирсона для проверки выдвинутой гипотезы.

```
Shapiro-Wilk test: 1.3935754676966394e-12
Kolmogorov-Smirnov test: 1.3867885687360081e-05
Chi-Square test: 2.1722521309710315e-06
```

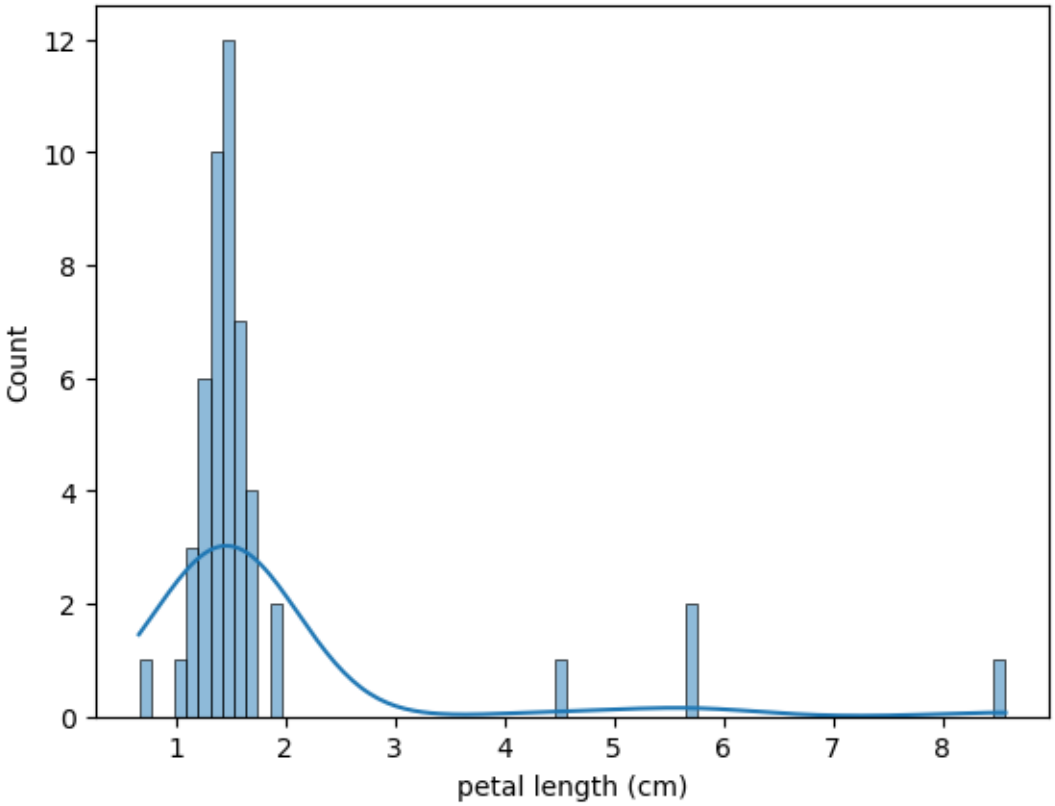

Рисунок 3.5 - Гистограмма распределения переменной PETALLEN

Все три критерия позволяют нам отклонить гипотезу о том, что PETALLEN с засорениями имеет нормальное распределение. Следовательно, мы воспользуемся коэффициентом корреляции Спирмена. Также выдвинем гипотезу о равенстве истинного коэффициента корреляции нулю и проверим ее с помощью критерия Спирмена.

```
spearman correlation coefficient: 0.22457039672596352
spearman P-value: 0.11690694096066725
```

Коэффициент корреляции Спирмена между SEPALLEN и PETALLEN равен 0.22, а P-значение критерия Спирмена сильно больше 0.05. Что позволяет нам принять гипотезу о равенстве истинного коэффициента корреляции нулю и. Отсюда следует, что линейная связь между переменными SEPALLEN и PETALLEN с засорениями отсутствуем. Таким образом, мы можем не использовать переменную PETALLEN при построении регрессионной модели.

Регрессионный анализ однородных данных с засорениями

В корреляционном анализе данных с засорениями мы пришли к выводу о том, что мы можем не использовать в нашей модели переменную PETALLEN, так как она не имеет линейной связи с предсказываемой величиной SEPALLEN. Отсюда следует, что можно исключить из исследования переменную PETALLEN. А из-за этого результаты регрессионного анализа не изменятся, следовательно, на поведение нашей регрессионной модели засорения в переменной PETALLEN никак не скажутся.

Таким образом, мы можем использовать регрессионную модель

$$sepallen = 2.639 + 0.6905 * sepalwid.$$