МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
Факультет прикладной математики и информатики
Кафедра математического моделирования и анализа данных

Корреляционный и регрессионный анализ данных

Отчет по лабораторной работе №2
студентки 3 курса 7 группы
Летецкой М.С.

Преподаватель
Малюгин В.И.

Минск, 2024

1. Постановка задачи

1) Корреляционный и регрессионный анализ однородных данных

Указание:

1) использовать выборку значений переменных (1, 2, 3) без засорений для одного вида ириса (3), в регрессионной модели зависимой является первая переменная;

2) Исследовать эффекты засорений на результаты корреляционного и регрессионного анализа, использовать выборку значений переменных (1, 2, 3) с засорением (переменная 1) для одного вида ириса (3)

2. Выполнение задания без засорений данных

Корреляционный анализ.

В ходе выполнения работы были построены диаграммы рассеяния для каждой пары параметров, заданных в условии (рис.1).

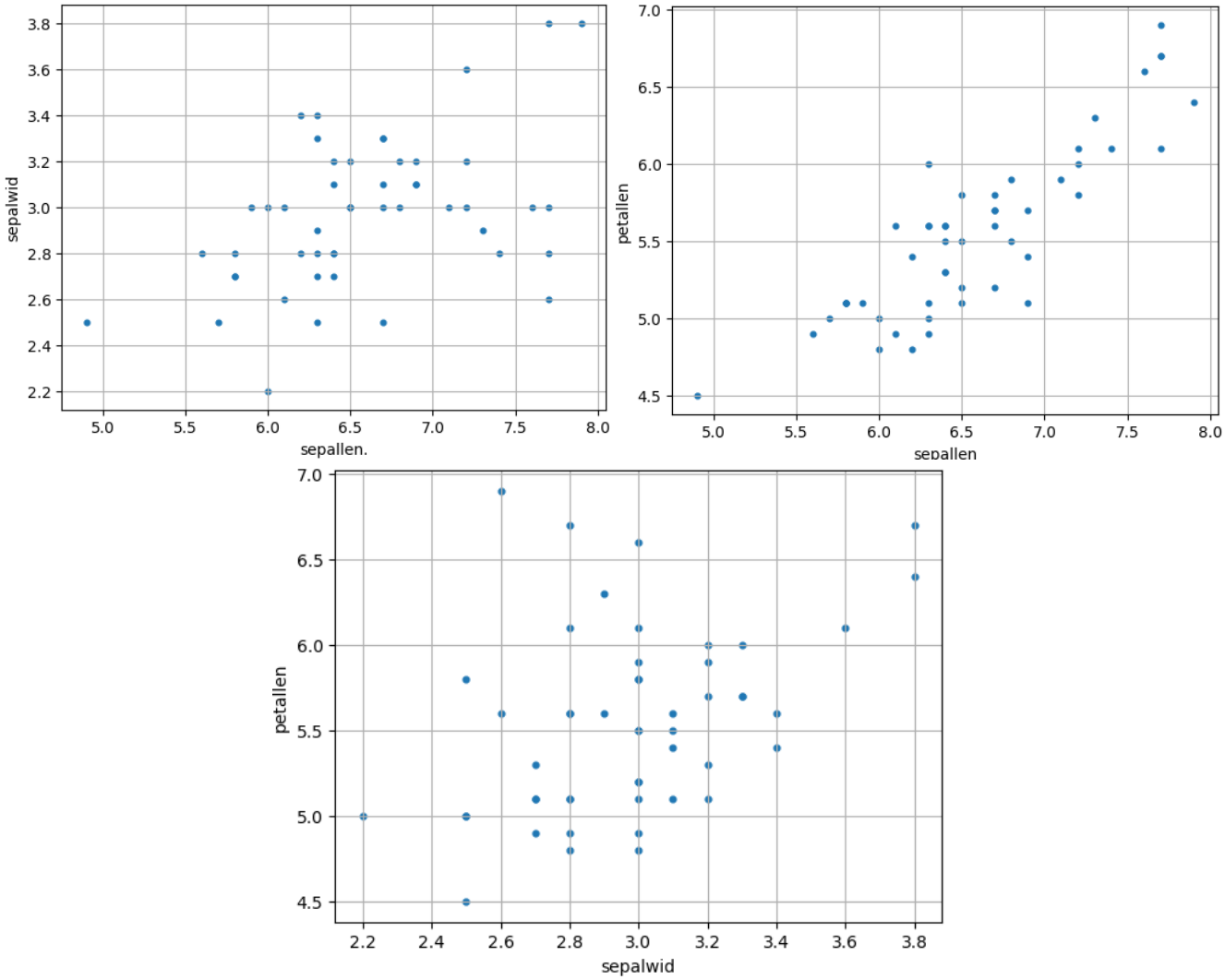


Рис.1

Была построена матрица диаграмм рассеивания (рис.2).

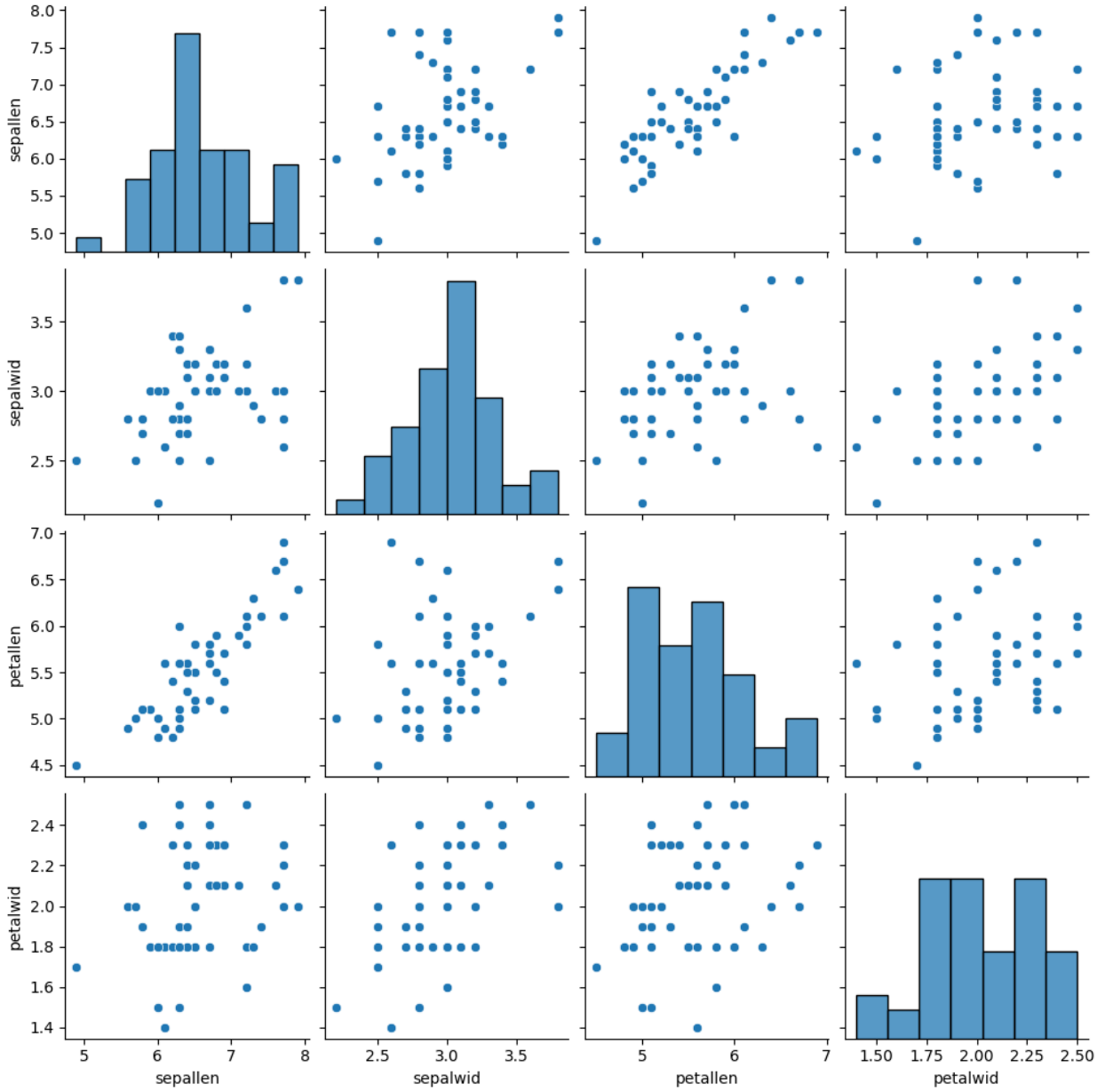


Рис.2

По результатам построения диаграмм рассеивания можно заметить, что между параметрами ‘petallen’ и ‘sepallen’ прослеживается четкая связь.

Была найдена ковариация между параметром ‘sepallen’ и ‘petallen’, ‘sepalwid’ (по условию ‘sepallen’ – зависимая переменная):

```
ковариация sepallen и sepalwid:  0.09376326530612239
ковариация sepallen и petallen:  0.3032897959183673
```

Ковариация - мера совместной вариативности (линейной зависимости) двух случайных величин. Если обе величины демонстрируют однонаправленное изменение, то ковариация положительная, а если разнонаправленное — отрицательная. Если ковариация близка к нулю, то величины независимы. По полученным результатам можно сделать вывод о том, что переменная ‘sepallen’ в той или иной мере зависима от ‘sepalwid’ и ‘petallen’.

Была найдена корреляция между данными параметрами двумя способами:

- Вручную

```
def covariance(xs, ys):
    dx = xs - xs.mean()
    dy = ys - ys.mean()
    return (dx * dy).sum() / (dx.count() - 1)

def variance(xs):
    x_hat = xs.mean()
    n = xs.count()
    n = n - 1 if n in range( 1, 30 ) else n
    return sum((xs - x_hat) ** 2) / n

def standard_deviation(xs):
    return np.sqrt(variance(xs))

def correlation(xs, ys):
    return covariance(xs, ys) / (standard_deviation(xs) *  standard_deviation(ys))

print ("Корреляция sepallen и sepalwid",correlation (df_data['sepallen'], df_data['sepalwid']))
print ("Корреляция sepallen и petallen",correlation (df_data['sepallen'], df_data['petallen']))
```

```
Корреляция sepallen и sepalwid 0.46655899632052344
Корреляция sepallen и petallen 0.8818619723832409
```

- При помощи стандартной функции библиотеки

```
print ("Корреляция sepallen и sepalwid",df_data['sepallen'].corr( df_data['sepalwid']))
print ("Корреляция sepallen и petallen",df_data['sepallen'].corr(df_data['petallen']))
```

```
Корреляция sepallen и sepalwid 0.45722781639411275
Корреляция sepallen и petallen 0.8642247329355761
```

По полученным результатам видно, что корреляция между ‘sepallen’ и ‘sepalwid’ умеренная, между ‘sepallen’ и ‘petallen’ – высокая.

Регрессионный анализ.

Были построены диаграммы рассеивания для ‘sepallen’ и ‘sepalwid’, ‘petallen’ (рис.3).

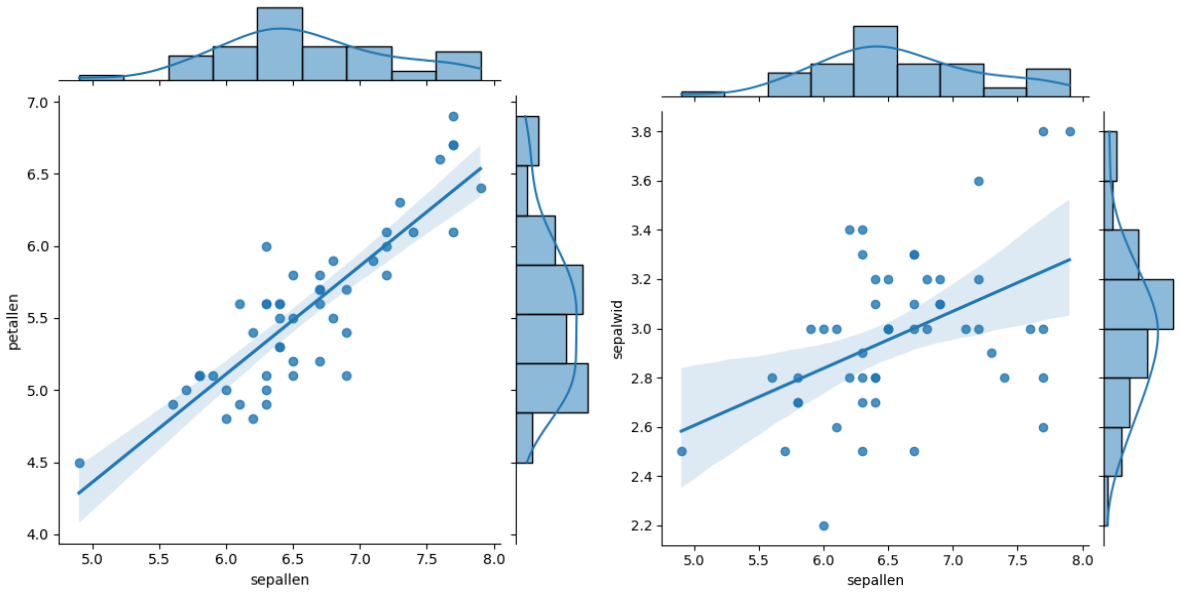


Рис.3

Была построена регрессионная модель:

```
                Results: Ordinary least squares
================================================================
Model:              OLS              Adj. R-squared:     0.750
Dependent Variable: sepallen         AIC:                31.1539
Date:               2024-11-26 22:00 BIC:                38.8020
No. Observations:   50               Log-Likelihood:     -11.577
Df Model:           3                F-statistic:        49.98
Df Residuals:       46               Prob (F-statistic): 1.62e-14
R-squared:          0.765            Scale:              0.10112
----------------------------------------------------------------
               Coef.   Std.Err.     t     P>|t|    [0.025  0.975]
----------------------------------------------------------------
Intercept      0.6999    0.5336   1.3116  0.1962  -0.3742  1.7740
sepalwid       0.3303    0.1743   1.8949  0.0644  -0.0206  0.6812
petallen       0.9455    0.0907  10.4223  0.0000   0.7629  1.1281
petalwid      -0.1698    0.1981  -0.8570  0.3959  -0.5685  0.2289
----------------------------------------------------------------
Omnibus:              0.056        Durbin-Watson:           1.922
Prob(Omnibus):        0.973        Jarque-Bera (JB):        0.039
Skew:                 0.032        Prob(JB):                0.981
Kurtosis:             2.879        Condition No.:           81
================================================================
```

Есть два фактора, у которых расчетный уровень значимости P>|t| превышает 0.05. Удаляем тот фактор, у которого расчетный уровень значимости больше (petalwid):

```
                Results: Ordinary least squares
================================================================
Model:              OLS              Adj. R-squared:     0.751
Dependent Variable: sepallen         AIC:                29.9460
Date:               2024-11-26 22:00 BIC:                35.6821
No. Observations:   50               Log-Likelihood:     -11.973
Df Model:           2                F-statistic:        75.02
Df Residuals:       47               Prob (F-statistic): 2.36e-15
R-squared:          0.761            Scale:              0.10055
----------------------------------------------------------------
            Coef.    Std.Err.      t      P>|t|     [0.025   0.975]
----------------------------------------------------------------
Intercept   0.6248     0.5249    1.1904   0.2399   -0.4311   1.6807
petallen    0.9348     0.0896   10.4330   0.0000    0.7546   1.1151
sepalwid    0.2600     0.1533    1.6953   0.0966   -0.0485   0.5684
----------------------------------------------------------------
Omnibus:              0.061        Durbin-Watson:           1.834
Prob(Omnibus):        0.970        Jarque-Bera (JB):        0.197
Skew:                 -0.073       Prob(JB):                0.906
Kurtosis:             2.729        Condition No.:           76
================================================================
```

У ‘sepalwid’ расчетный уровень значимости P>|t| превышает 0.05, убираем его из модели:

```
                Results: Ordinary least squares
================================================================
Model:              OLS              Adj. R-squared:     0.742
Dependent Variable: sepallen         AIC:                30.9137
Date:               2024-11-26 22:00 BIC:                34.7377
No. Observations:   50               Log-Likelihood:     -13.457
Df Model:           1                F-statistic:        141.6
Df Residuals:       48               Prob (F-statistic): 6.30e-16
R-squared:          0.747            Scale:              0.10448
----------------------------------------------------------------
             Coef.    Std.Err.      t      P>|t|    [0.025   0.975]
----------------------------------------------------------------
Intercept    1.0597     0.4668    2.2702   0.0277   0.1212   1.9982
petallen     0.9957     0.0837   11.9011   0.0000   0.8275   1.1640
----------------------------------------------------------------
Omnibus:              0.060        Durbin-Watson:           1.764
Prob(Omnibus):        0.970        Jarque-Bera (JB):        0.256
Skew:                 0.015        Prob(JB):                0.880
Kurtosis:             2.651        Condition No.:           59
================================================================
```

Была построена модель, у которой все параметры значительно влияют на отклик.

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

Был построен график остатков (рис.4).

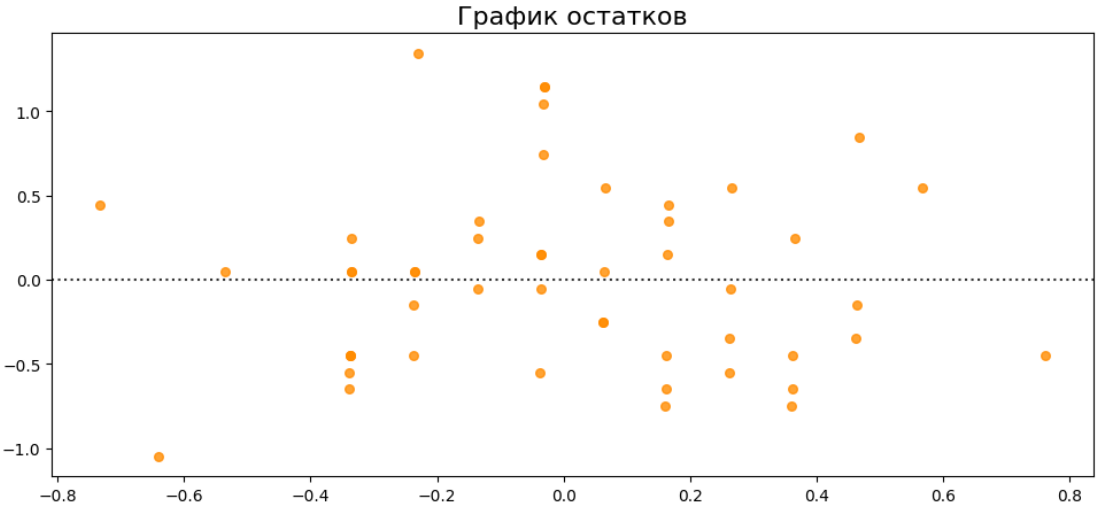


Рис.4

Точки не имеют системности, значит остатки не зависят от предсказанных значений.

Была проведена проверка некоторых критериев на нормальность распределения:

- Критерий Шапиро-Уилка:

```
ShapiroResult(statistic=np.float64(0.9831871348860054), pvalue=np.float64(0.6917523626039768))
```

pvalue = 0.69 > 0.05, значит не можем отвергать нулевую гипотезу (нет док-в, что выборка не соответствует нормальному распределению).

- Критерий Д’Агостино:

```
NormaltestResult(statistic=np.float64(0.06036675211158883), pvalue=np.float64(0.9702675933896836))
```

pvalue = 0.97> 0.05, значит не можем отвергать нулевую гипотезу (нет док-в, что выборка не соответствует нормальному распределению)

Для остатков была построена столбчатая диаграмма, график «квантиль-квантиль» и «ящик с усами» (рис.5).

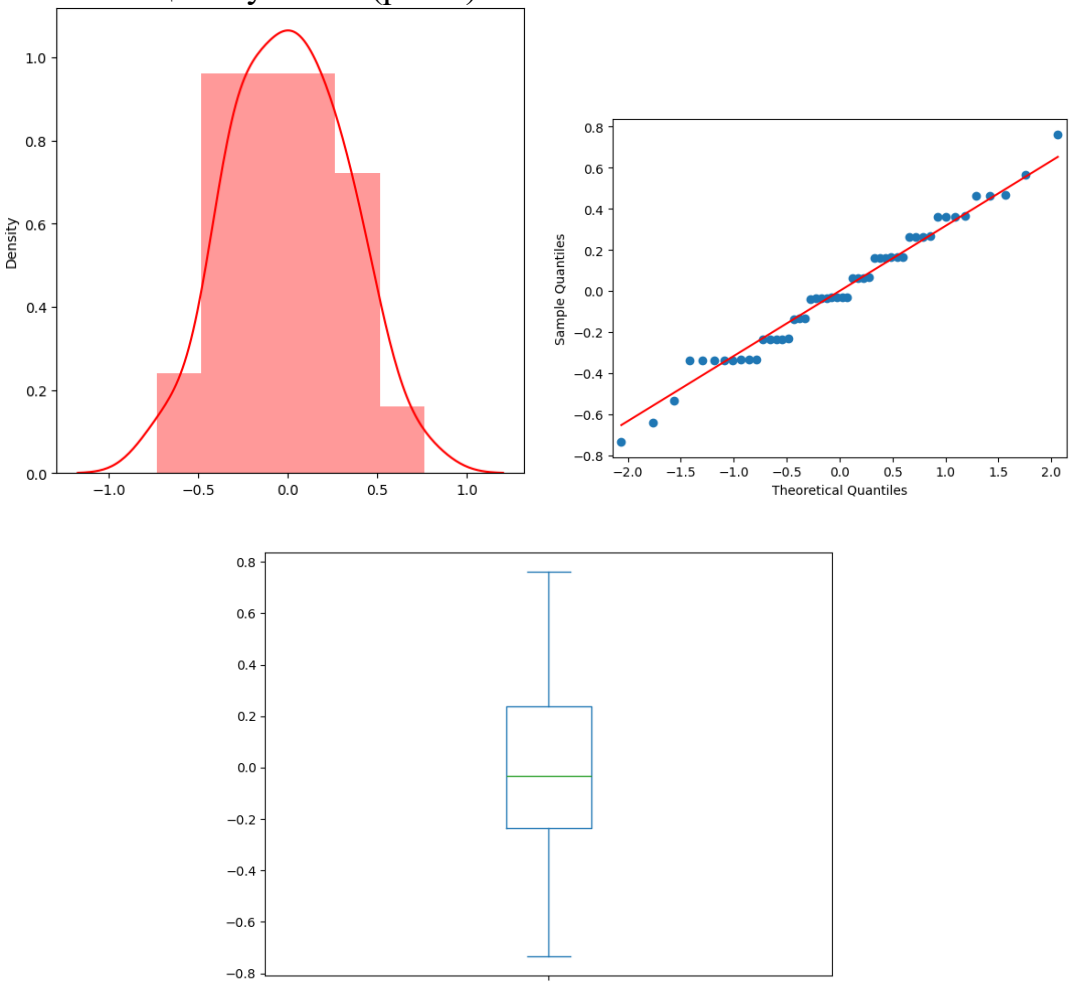


Рис.5

Вывод: остатки имеют нормальное распределение.

```
f_oneway(df_data['sepallen'], df_data['petallen'])
```

```
F_onewayResult(statistic=np.float64(75.69824052323709), pvalue=np.float64(7.947332254531585e-14))
```

Уровень значимости p-value < 0.05, можно утверждать, что потроенная модель приемлема.

3. Выполнение задания для данных с засорением

Была построена матрица диаграмм рассеивания для каждого параметра (рис.6).

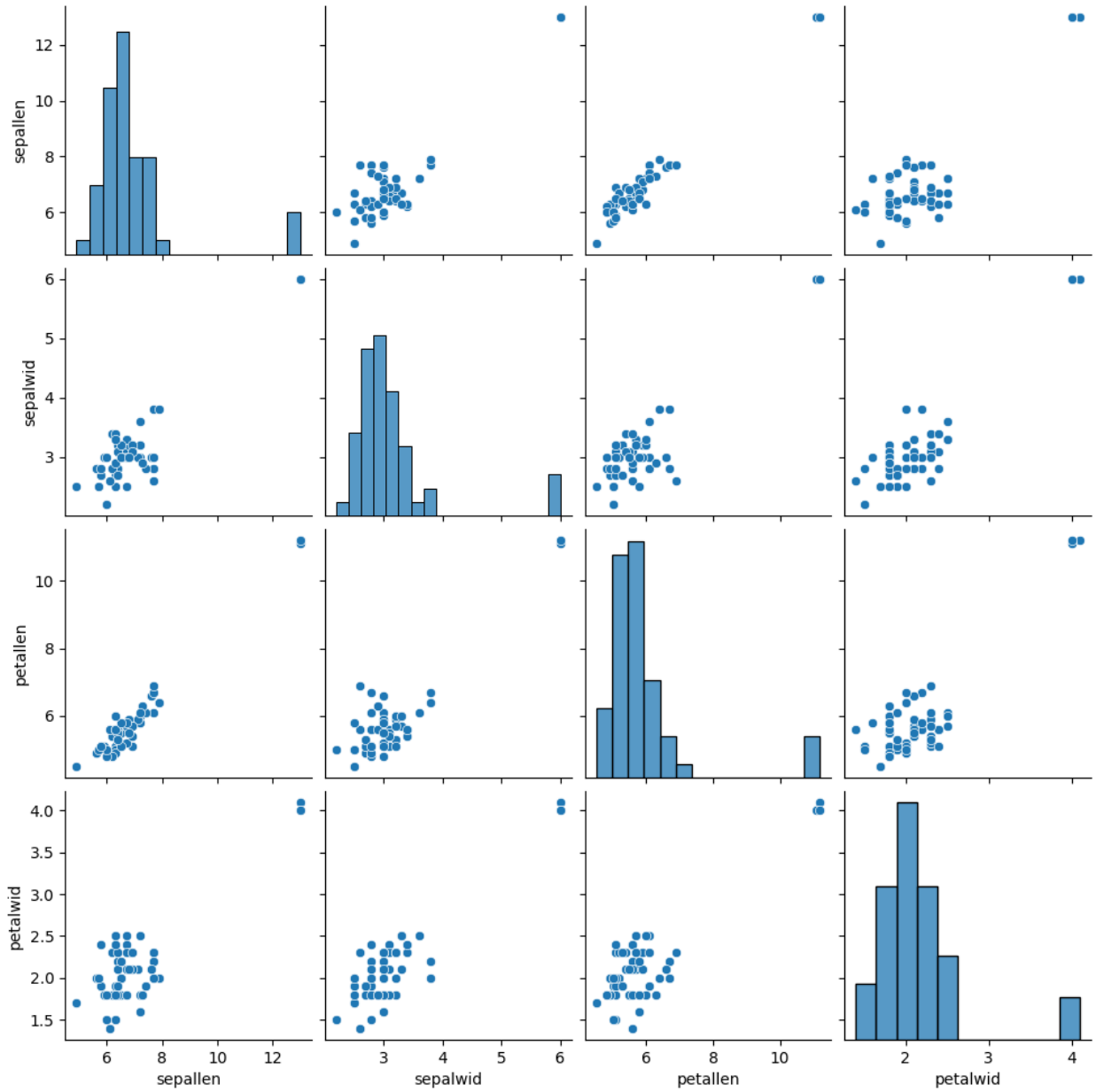


Рис.6

По рис.6 видно, что данные имеют аномальные значения.
Коэффициенты корреляции:

```
Корреляция sepallen и sepalwid без засорения  0.9165979465328571
Корреляция sepallen и petallen без засорения 0.9804659927178373
Корреляция sepallen и sepalwid с засорением  0.9165979465328571
Корреляция sepallen и petallen с засорением 0.9804659927178373
```

Коэффициенты ковариации:

```
ковариация sepallen и sepalwid без засорения:  1.1643476621417799
ковариация sepallen и petallen без засорения:  2.272752639517346
ковариация sepallen и sepalwid с засорением:  1.1643476621417799
ковариация sepallen и petallen с засорением:  2.272752639517346
```

Можно заметить, что полученные значения для выборки с засорением не значительно отличаются от данных выборки без засорения.

Была построена регрессионная модель:

```
                Results: Ordinary least squares
================================================================
Model:              OLS              Adj. R-squared:     0.963
Dependent Variable: sepallen         AIC:                30.6517
Date:               2025-01-21 10:55 BIC:                38.4567
No. Observations:   52               Log-Likelihood:     -11.326
Df Model:           3                F-statistic:        444.4
Df Residuals:       48               Prob (F-statistic): 5.31e-35
R-squared:          0.965            Scale:              0.098055
----------------------------------------------------------------
               Coef.   Std.Err.     t     P>|t|    [0.025  0.975]
----------------------------------------------------------------
Intercept      0.3045    0.1900   1.6027  0.1156  -0.0775  0.6865
sepalwid       0.3672    0.1640   2.2394  0.0298   0.0375  0.6968
petallen       0.9868    0.0763  12.9325  0.0000   0.8333  1.1402
petalwid      -0.1414    0.1941  -0.7285  0.4699  -0.5318  0.2489
----------------------------------------------------------------
Omnibus:              0.203        Durbin-Watson:           2.008
Prob(Omnibus):        0.903        Jarque-Bera (JB):        0.027
Skew:                 0.056        Prob(JB):                0.986
Kurtosis:             2.998        Condition No.:           38
================================================================
```

Удаляем тот фактор(среди тех, у которых расчетный уровень значимости P>|t| превышает 0.05), у которого расчетный уровень значимости больше (petalwid):

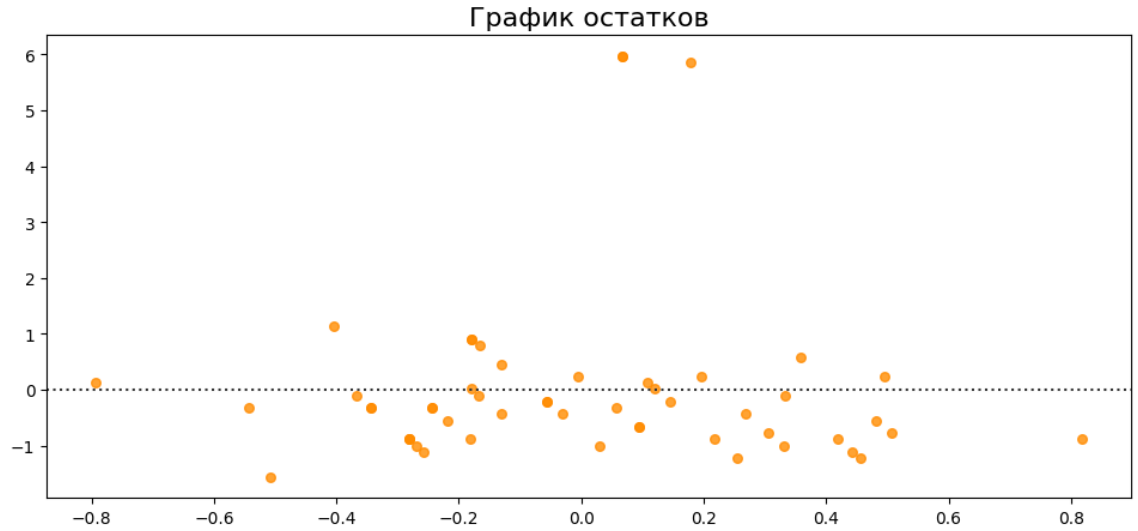
```
                Results: Ordinary least squares
================================================================
Model:              OLS              Adj. R-squared:     0.963
Dependent Variable: sepallen         AIC:                29.2235
Date:               2025-01-21 10:55 BIC:                35.0772
No. Observations:   52               Log-Likelihood:     -11.612
Df Model:           2                F-statistic:        672.8
Df Residuals:       49               Prob (F-statistic): 2.35e-36
R-squared:          0.965            Scale:              0.097116
----------------------------------------------------------------
            Coef.    Std.Err.      t      P>|t|     [0.025   0.975]
----------------------------------------------------------------
Intercept   0.2894     0.1879    1.5396   0.1301   -0.0883   0.6670
sepalwid    0.2998     0.1347    2.2250   0.0307    0.0290   0.5706
petallen    0.9738     0.0738   13.1878   0.0000    0.8254   1.1222
----------------------------------------------------------------
Omnibus:              0.070        Durbin-Watson:           1.952
Prob(Omnibus):        0.966        Jarque-Bera (JB):        0.033
Skew:                 -0.035       Prob(JB):                0.984
Kurtosis:             2.898        Condition No.:           31
================================================================
```

Перестроенная модель:

```
                Results: Ordinary least squares
================================================================
Model:              OLS              Adj. R-squared:     0.961
Dependent Variable: sepallen         AIC:                32.2284
Date:               2025-01-21 10:55 BIC:                36.1309
No. Observations:   52               Log-Likelihood:     -14.114
Df Model:           1                F-statistic:        1242.
Df Residuals:       50               Prob (F-statistic): 5.59e-37
R-squared:          0.961            Scale:              0.10479
----------------------------------------------------------------
             Coef.    Std.Err.      t      P>|t|     [0.025   0.975]
----------------------------------------------------------------
Intercept    0.3532     0.1929    1.8304   0.0731   -0.0344   0.7407
petallen     1.1232     0.0319   35.2483   0.0000    1.0592   1.1872
----------------------------------------------------------------
Omnibus:               0.177        Durbin-Watson:           1.823
Prob(Omnibus):         0.915        Jarque-Bera (JB):        0.246
Skew:                  0.129        Prob(JB):                0.884
Kurtosis:              2.783        Condition No.:           27
================================================================
```

Данные с засорением не имеют нормального распределения.

Был проведен анализ остатков.



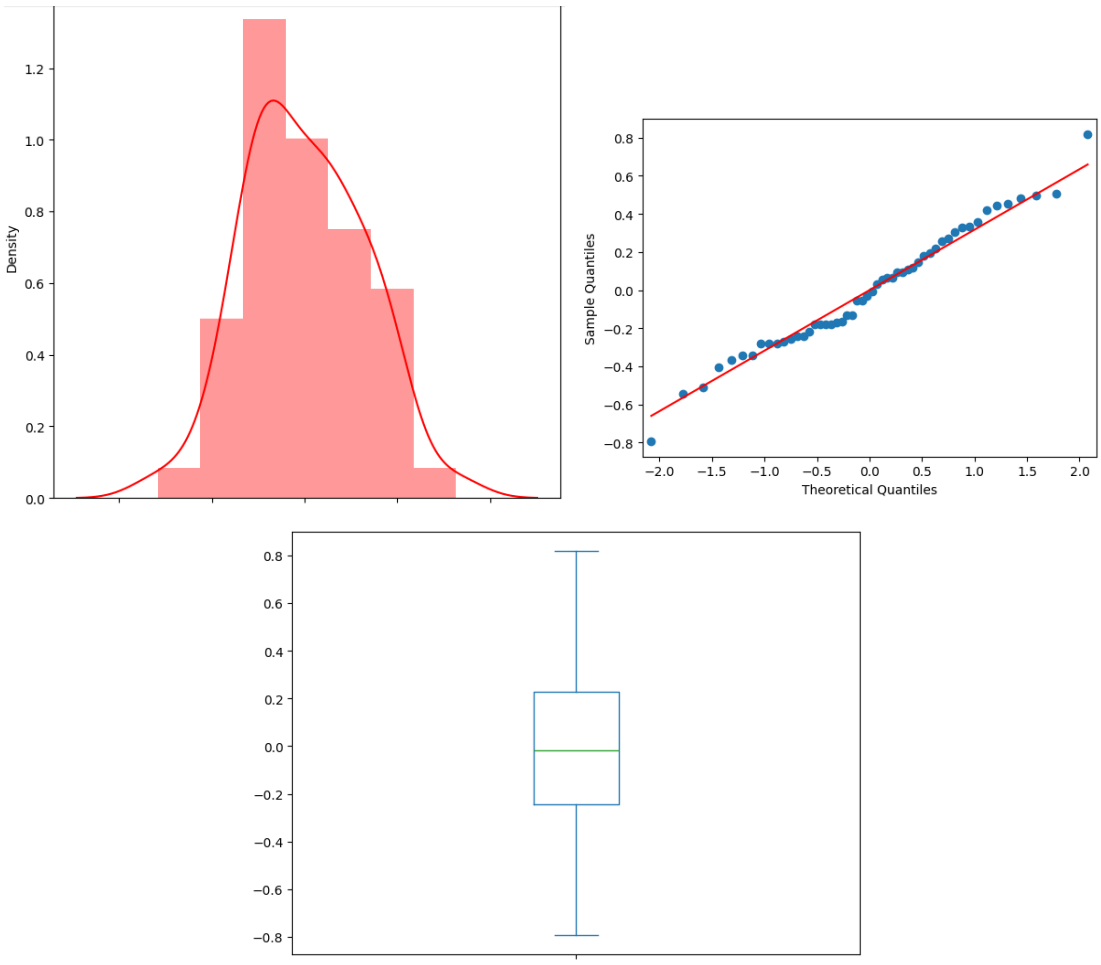
По графику остатков видно, что есть «выбросы», но закономерность не прослеживается.

Результаты проверки критериев:

- Критерий Шапиро-Уилка: pvalue = 0.770587169726827
- Критерий Д’Агостино: pvalue = 0.9150987003968815

Оба критерия не дают оснований отвергать гипотезу о нормальном распределении остатков.

Для остатков была построена столбчатая диаграмма, график «квантиль-квантиль» и «ящик с усами»:



Вывод: остатки построенной модели имеют нормальное распределение.

Построение регрессионной модели для “чистых” данных — это более простая задача, которая позволяет получить точные и интерпретируемые результаты. При работе с “засоренными” данными важно учитывать влияние шума и выбросов, тщательно проводить предобработку данных.