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

Предварительный, корреляционный и регрессионный анализ неоднородных данных

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

Преподаватель
Малюгин В.И.

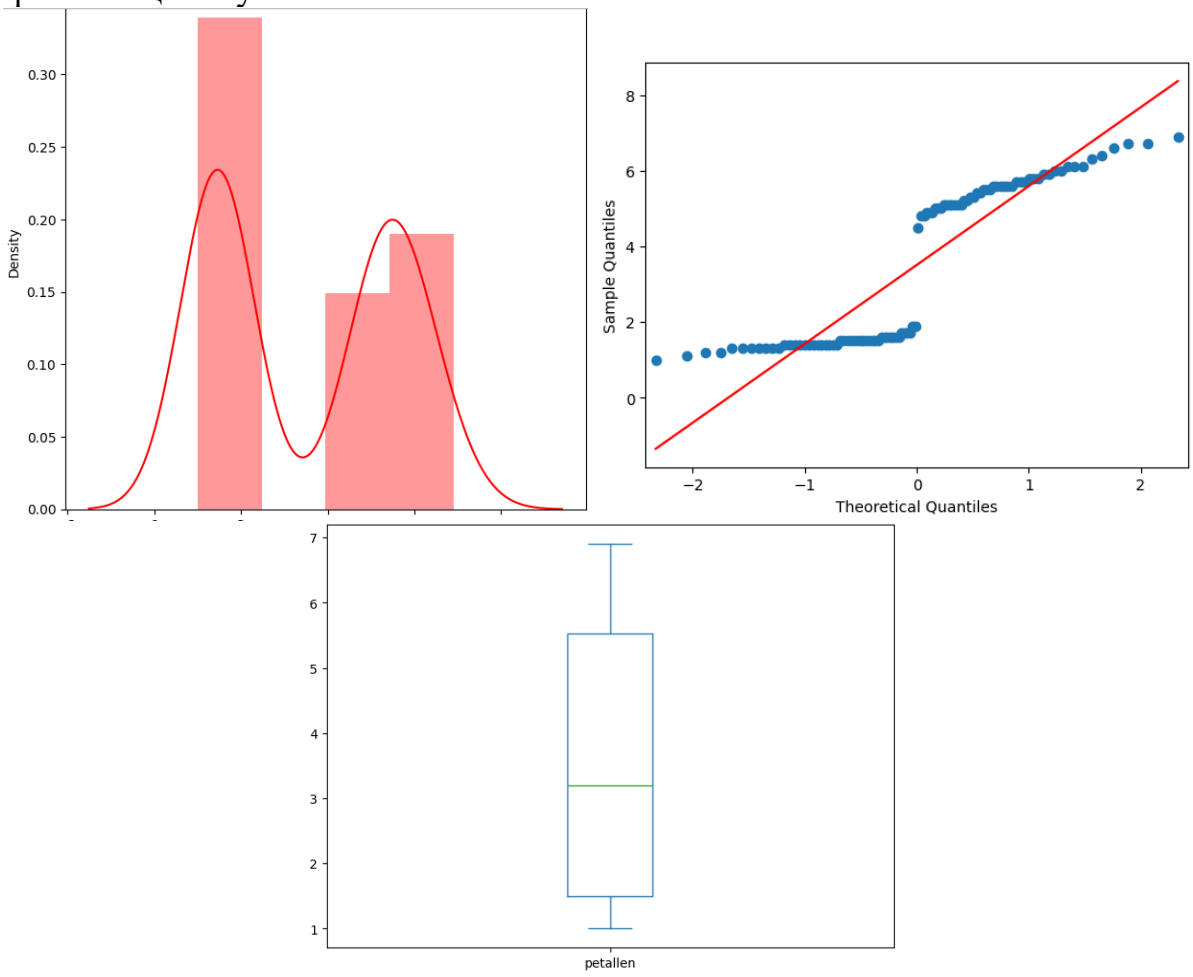
Минск, 2024

1. Условие задания

Оценка влияния неоднородности выборки на вероятностные свойства данных: закон распределения, корреляционные и регрессионные зависимости. Указание: использовать выборки значений переменной (3) для нескольких видов ирисов (13);

2. Оценка влияния неоднородности выборки на вероятностные свойства данных: закон распределения параметра ‘petallen’

В ходе выполнения работы были построены столбчатые диаграммы, графики «ящик с усами» и «квантиль-квантиль»



На графиках очень хорошо прослеживается разбиение данных на две структуры, что уже говорил об отсутствии нормального распределения.

Были проверены некоторые статистические критерии:

- Критерий Шапиро-Уилка

```
shapiro(df_data['petallen'])
```

```
ShapiroResult(statistic=np.float64(0.7805754494200133), pvalue=np.float64(6.5398355204529e-11))
```

pvalue < 0.05, значит можно отвергать нулевую гипотезу (выборка не соответствует нормальному распределению)

- Критерий Д’Агостино

```
normaltest(df_data['petallen'])
```

```
NormaltestResult(statistic=np.float64(1024.7271254795196), pvalue=np.float64(3.0432103607241986e-223))
```

pvalue < 0.05, значит можно отвергать нулевую гипотезу (выборка не соответствует нормальному распределению)

- Критерий Колмогорова-Смирнова

```
stat2, p_value = kstest(df_data['petallen'], stat.norm.cdf)
print(" Kolmogorov-Smirnov Test: statistic= ",stat2," p-value=", p_value)
```

```
 Kolmogorov-Smirnov Test: statistic=  0.8649303297782918  p-value= 3.066167865765473e-87
```

pvalue < 0.05, значит можно отвергать нулевую гипотезу (выборка не соответствует нормальному распределению)

Были получены описательные статистики:

count	100.000000
mean	3.507000
std	2.095221
min	1.000000
25%	1.500000
50%	3.200000
75%	5.525000
max	6.900000
skew	0.09734427380296742
kurtosis	-1.8633622895466364

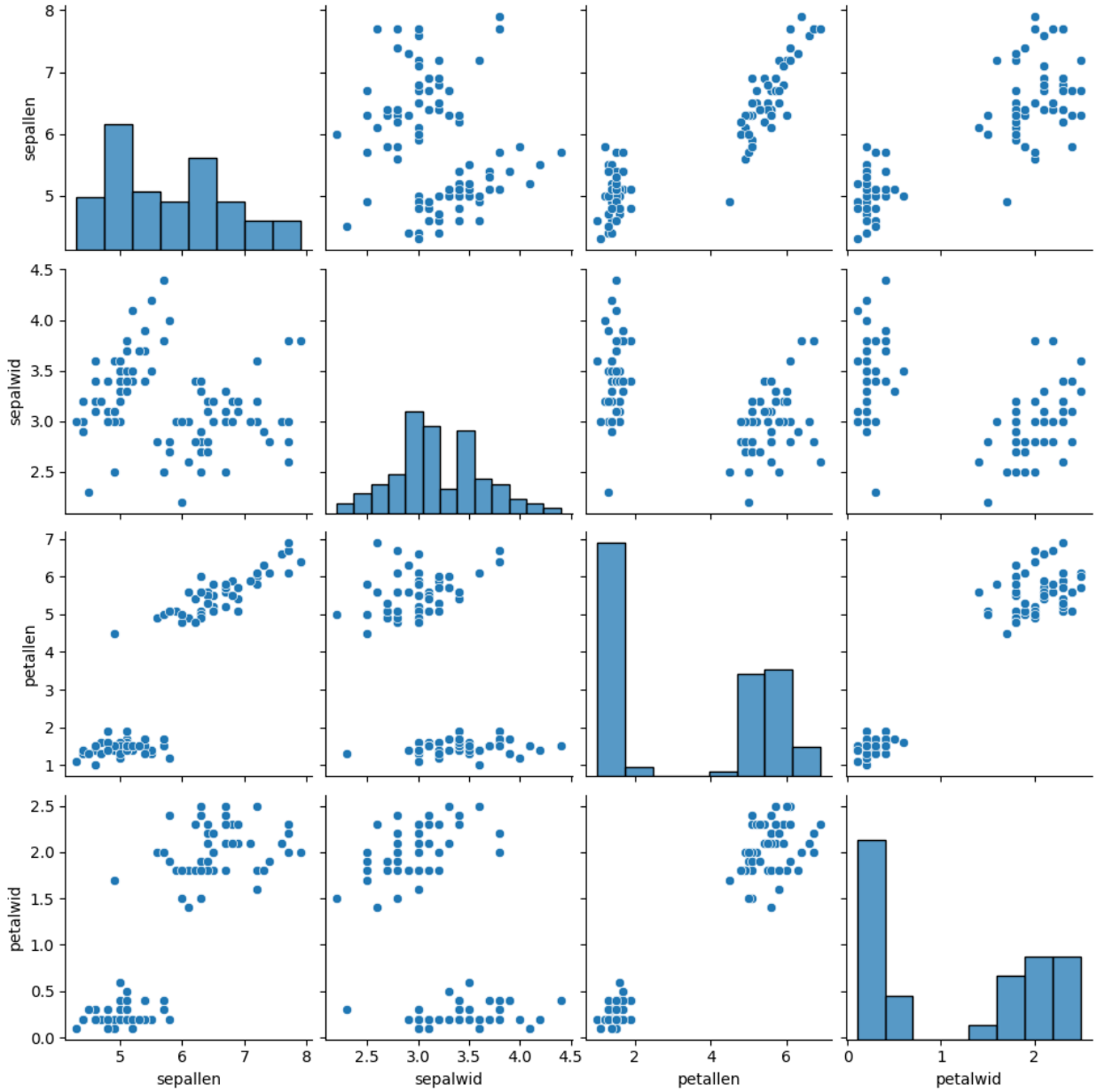
На основе представленных статистических характеристик можно сделать вывод о том, что данные имеют:

- Значительный разброс.
- Правостороннюю асимметрию.
- Значение эксцесса (kurtosis) равное -1.8634 говорит о том, что распределение данных является платикуртическим (плосковершинным).

Основываясь на вышеизложенных фактах, можно утверждать, что данные не соответствуют нормальному распределению.

3. Проверка корреляционных зависимостей

Была построена матрица диаграмм рассеивания, на которой явно прослеживается разбиение данных на две группы:



Были найдены коэффициенты ковариации:

```
ковариация petallen и sepalwid:  -0.4277848484848487
ковариация sepallen и petallen:  1.7921424242424238
ковариация petalwid и petallen:  1.8656040404040413
```

Ковариация (Cov) измеряет, насколько две переменные изменяются совместно. Она может принимать значения как положительные, так и отрицательные, а также ноль:

- Cov(X, Y) > 0: Положительная ковариация. Когда одна переменная (например, X) увеличивается, то и другая переменная (Y) в среднем также имеет тенденцию к увеличению.
- Cov(X, Y) < 0: Отрицательная ковариация. Когда одна переменная (например, X) увеличивается, то другая переменная (Y) в среднем имеет тенденцию к уменьшению.
- Cov(X, Y) ≈ 0: Ковариация близка к нулю. Это означает, что связь между двумя переменными слабая, либо отсутствует (они изменяются независимо).

В данном случае с увеличением значения параметра ‘petallen’ увеличиваются значения параметров ‘sepallen’, ‘petalwid’ и уменьшаются значения параметра ‘sepalwid’.

Были вычислены коэффициенты корреляции:

```
Корреляция sepallen и petallen  0.9048248126076007
Корреляция petalwid и petallen 0.96982426889205
Корреляция sepalwid и petallen -0.4885589157004354
```

Между ‘petallen’ и ‘sepallen’ / ‘petalwid’ очень высокая корреляция (более высоким значениям одного признака соответствуют более высокие значения другого, а более низким значениям одного признака – низкие значения другого).

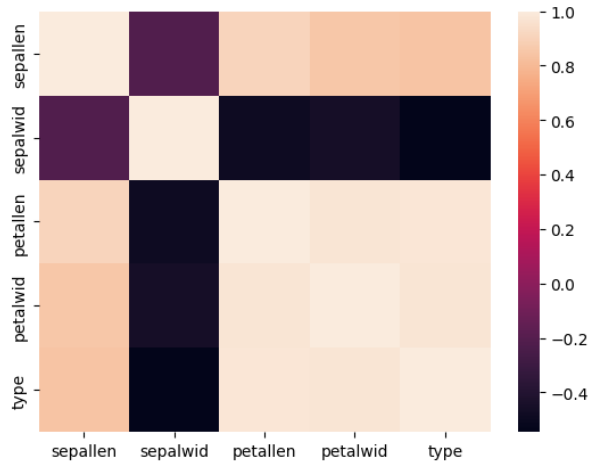
Между ‘petallen’ и ‘sepalwid’ слабая корреляция (более высоким значениям одного признака соответствуют более низкие значения другого, а более низким значениям одного признака – высокие значения другого).

Была построена регрессионная модель:

```
                  Results: Ordinary least squares
==================================================================
Model:              OLS              Adj. R-squared:     0.896
Dependent Variable: sepallen         AIC:                50.1518
Date:               2025-01-21 11:48 BIC:                60.5725
No. Observations:   100              Log-Likelihood:     -21.076
Df Model:           3                F-statistic:        285.2
Df Residuals:       96               Prob (F-statistic): 1.14e-47
R-squared:          0.899            Scale:              0.092965
------------------------------------------------------------------
               Coef.   Std.Err.     t     P>|t|    [0.025   0.975]
------------------------------------------------------------------
Intercept      1.7427    0.3117   5.5915  0.0000   1.1240   2.3613
sepalwid       0.6900    0.0843   8.1894  0.0000   0.5228   0.8573
petallen       0.6890    0.0613  11.2459  0.0000   0.5673   0.8106
petalwid      -0.5023    0.1373  -3.6597  0.0004  -0.7748  -0.2299
------------------------------------------------------------------
Omnibus:              0.212        Durbin-Watson:           2.455
Prob(Omnibus):        0.899        Jarque-Bera (JB):        0.054
Skew:                 0.054        Prob(JB):                0.973
Kurtosis:             3.034        Condition No.:           56
==================================================================
```

В ней все параметры значительно влияют на отклик.

Была построена heatmap для коэффициентов корреляции:



Можно заметить, что присутствует мультиколлениарность. Была проведена проверка её значимости:

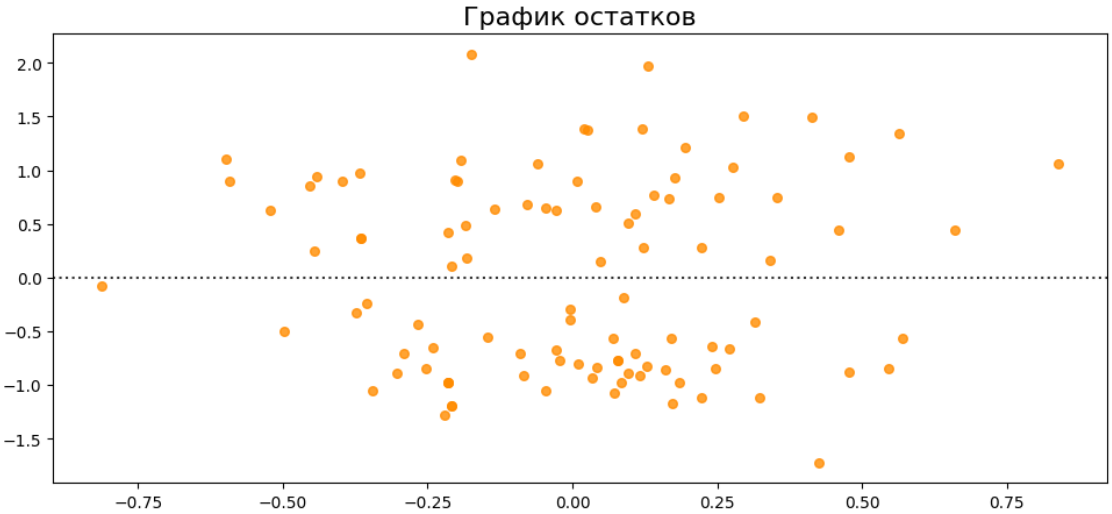
```
X = add_constant(df_data.drop('petallen', axis=1))

VIFs = pd.DataFrame()
VIFs['Variable'] = X.columns
VIFs['VIF'] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
print(VIFs)
```

```
   Variable         VIF
0     const  191.461032
1  sepallen    4.812751
2  sepalwid    2.182200
3  petalwid   23.376173
4      type   29.350568
```

Высокие значения VIF указывают на потенциальные проблемы с мультиколлинеарностью. Как правило, значение VIF выше 5 требует внимания, а выше 10 — серьезного рассмотрения изменений в модели.

Был проведён анализ остатков. График остатков:



На графике не прослеживается явной закономерности расположения точек, проверен ряд критериев:

- Критерий Шапиро-Уилка

```
shapiro(result_linear_ols.resid)
```

```
ShapiroResult(statistic=np.float64(0.9946273547530884), pvalue=np.float64(0.9643237428432809))
```

- Критерий Д’Агостино

```
normaltest(result_linear_ols.resid)
```

```
NormaltestResult(statistic=np.float64(0.21201391699876276), pvalue=np.float64(0.8994183894518422))
```

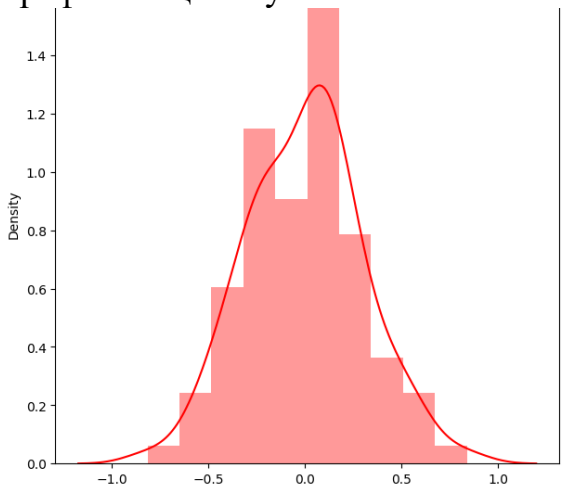
- Критерий Колмогорова-Смирнова

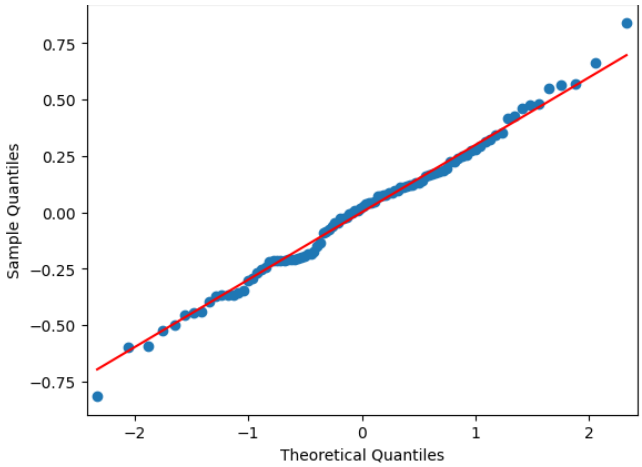
```
stat2, p_value = kstest(result_linear_ols.resid, stat.norm.cdf)
print(" Kolmogorov-Smirnov Test: statistic= ",stat2," p-value=", p_value)
```

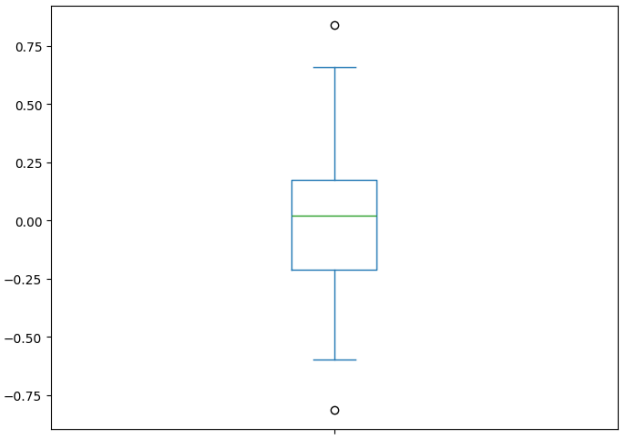
```
 Kolmogorov-Smirnov Test: statistic=  0.2749863364078761  p-value= 3.5623816410310056e-07
```

Первые два критерия говорят о том, что нет оснований отклонять нулевую гипотезу о нормальном распределении, третий гипотезу отклоняет.

Для более полного исследования были построены столбчатая диаграмма, график «ящик с усами» и «квантиль-квантиль»:







Можно сказать, что распределение остатков регрессионной модели близко к нормальному, однако данные имеют аномальные значения.

ВЫВОДЫ.

В ходе выполнения работы была отклонена гипотеза о наличии нормального распределения данных из неоднородной выпорки, выявлены сильные корреляционные зависимости между параметрами и их мультиколлениарность, с которой в будущих исследованиях придется бороться, построена регрессионная модель, остатки которой имеют распределение, близкое к нормальному, и аномальные значения.