Часть 3

Задания:

3.1. Проверить гипотезу о независимости переменных по критерию Хи-

квадрат (2 балла)

3.2. Вычислить оценку ковариации, коэффициента корреляции (2 балла).

Проверить гипотезу о незначимости коэффициента корреляции (2 балла).

3.3. Оценить параметры линейной регрессии (1 балл), вычислить

коэффициент детерминации (1 балл), проверить значимость модели по

критерию Фишера (2 балла).

Код:

```
import pandas as pd
import numpy as np
from sklearn.linear_model import LinearRegression
import matplotlib.pyplot as plt
import warnings

def is_ind(len_sepal, width_sepal):
    len_sepal = pd.qcut(len_sepal, 5)
    width_sepal = pd.qcut(width_sepal, 5)

    df = pd.concat([len_sepal, width_sepal], axis=1)
    table = pd.crosstab(df['lenght_of_petal'], df['width_of_sepal'])

    table['vit'] = table.sum(axis=1)
    vtj = pd.Series(name='vtj')
    table = table.append(vtj, ignore_index=False)
    table.iloc[5] = table.sum(axis=0)
    print(table)
    s = 0
    for i in range(5):
        for j in range(5):
            s += np.power(table.iloc[j][i], 2) / (table.iloc[5][i] *
table.iloc[j][5])
    hi_prac = (s - 1) * 50
    hi_teor = 26.3


    if hi_teor > hi_prac:
        print('Переменные независимы, т.к', hi_prac ,'<', hi_teor)
    else:
        print('Переменные зависимы, т.к', hi_prac ,'>=', hi_teor)

def main():
    warnings.filterwarnings('ignore')
    df = pd.read_csv("iris1.data", delimiter=',')
    df = df[df['class'].isin(['Iris-setosa'])]

    width_petal = df['lenght_of_petal']
    width_sepal = df['width_of_sepal']
    print('3.1  Проверить гипотезу о независимости переменных по критерию Хи-
квадрат')
    is_ind(width_petal, width_sepal)
    print('3.2 Вычислить оценку ковариации коэффициента корреляции. Проверить
гипотезу о незначимости коэффициента корреляции')
    m_len = width_petal.mean()
    m_width = width_sepal.mean()
    alpha = 0.05

    xy = np.sum(np.multiply(width_petal, width_sepal))  # сумма произведений
    xy_m = xy / 50  # среднее от произведения переменных

    s_len = np.sum(np.power(np.subtract(width_petal, m_len), 2)) #  сумма
дисперсий лепестка
    s_len_m = s_len / 50

    s_width = np.sum(np.power(np.subtract(width_sepal, m_width), 2))  # сумма
дисперсий чашелистника
    s_width_m = s_width / 50

    std_dev_len = np.sqrt(s_len_m)
    std_dev_width = np.sqrt(s_width_m)
```

```
    c = xy_m - m_len * m_width
    r = c / (std_dev_len * std_dev_width)

    print("Коэффициент корреляции:", r)
    print("Коэффициент ковариации:", c)

    T = (r / np.sqrt(1 - r ** 2)) * (np.sqrt(50 - 2))
    tss = s_width

    # T табличное
    T_table = 1.96

    if T > T_table:
        print("Коэффициент корреляции значим, гипотеза принимется\n\n")
    else:
        print("Коэффициент корреляции значим, гипотеза отвергается\n\n")

    print('3.3')
    x = width_sepal.values.reshape(-1, 1)
    y = width_petal.values.reshape(-1, 1)
    reg = LinearRegression()
    reg.fit(x, y)
    plt.scatter(width_sepal, width_petal)
    plt.plot(width_sepal, reg.predict(x), color='red', linewidth=2)
    plt.show()
    print("Уравнение линейной регрессии: Y = {:.5} +
{:.5}X".format(reg.intercept_[0], reg.coef_[0][0]),
          "\nВсе нужные параметры можно легко увидеть")

    sq_len = np.sum(np.power(width_petal, 2))
    beta1_lid = (xy_m - m_len * m_width) / ((sq_len / 50) - (m_len ** 2))
    beta0_lid = m_width - beta1_lid * m_len

    rss = 0
    ess = 0
    for i in range(50):
        y_lid = beta0_lid + beta1_lid * width_petal[i]

        rss += (width_sepal[i] - y_lid) ** 2
        ess += (y_lid - m_width) ** 2

    print("tss", tss)
    print("rss", rss)
    print("ess", ess)
    determ_ko = 1 - rss / tss
    print("Коэффициент детерминации по формуле 1 - rss/tss", determ_ko)

    determ_ko2 = ess / tss
    print("Коэффициент детерминации по формуле ess/tss", determ_ko2)
    determ_ko3 = r * r
    print("Коэффициент детерминации по формуле r*r", determ_ko3)

    f = (determ_ko3 / (1 - determ_ko3)) * ((50 - 1 - 1) / 1)
    f_table = 4.03
    if f table > f:
        print("регрессия считается незначимой, т.к", f_table, '>', f)
    else:
        print("регрессия считается значимой, т.к", f_table, '<=', f)
```

Конец кода

Вывод:

3.1

Проверить гипотезу о независимости переменных по критерию Хи-квадрат

Переменные независимы, т.к. 13.112899005756152 < 26.3

3.2

Вычислить оценку ковариации коэффициента корреляции.
Проверить гипотезу о незначимости коэффициента корреляции

Коэффициент корреляции: 0.17669462869681588

Коэффициент ковариации: 0.011448000000000569

Коэффициент корреляции значим, гипотеза отвергается.

3.3

Уравнение линейной регрессии: Y = 1.189 + 0.080463X

Все нужные параметры можно легко увидеть

tss 7.1138

rss 6.8917001084598715

ess 0.22209989154014956

Коэффициент детерминации по формуле 1 - rss/tss 0.031220991810302356

Коэффициент детерминации по формуле ess/tss 0.031220991810305257

Коэффициент детерминации по формуле r*r 0.031220991810305628

регрессия считается незначимой, т.к 4.03 > 1.546903467381109

График