Министерство науки и высшего образования Российской Федерации

Федеральное государственное автономное образовательное учреждение высшего образования

«Национальный исследовательский Томский политехнический Университет»



Инженерная школа ядерных технологий

Отделение математики и математической физики

01.03.02 Прикладная математика и информатика

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Лабораторная работа № 4

**Вариант - 6**

по дисциплине:

**Математическая статистика**

|  |  |  |  |
| --- | --- | --- | --- |
| **Выполнил:** |  | | |
| студент группы | 0В21 | Дзебан А.А. |  |
|  |  |  |  |
| **Проверил:** | Шинкеев М.Л. | | |
|  |  |  |  |
|  |  |  |

Томск – 2024

**Цель работы:**  Исследовать влияние на среднее значение результирующего показателя каждого из факторов A и B (провести однофакторный дисперсионный анализ для каждого из факторов), используя линейные регрессионные модели.

Данные исходя из варианта:

|  |  |  |
| --- | --- | --- |
| Вариант | 6 |  |
| Фактор A | Фактор B | Отклик |
| 1 | 1 | 3,33 |
| 1 | 1 | 2,75 |
| 1 | 1 | 3,32 |
| 1 | 1 | 1,44 |
| 1 | 1 | 4,41 |
| 1 | 2 | 5,64 |
| 1 | 2 | 6,96 |
| 1 | 2 | 8,15 |
| 2 | 1 | 6,99 |
| 2 | 1 | 5,6 |
| 2 | 1 | 5,02 |
| 2 | 1 | 5,01 |
| 2 | 2 | 3,24 |
| 2 | 2 | 4,1 |
| 2 | 2 | 5,12 |
| 2 | 2 | 3,16 |
| 2 | 2 | 3,84 |
| 2 | 2 | 3,82 |
| 3 | 1 | 6,51 |
| 3 | 1 | 6,25 |
| 3 | 1 | 6,61 |
| 3 | 1 | 7,17 |
| 3 | 1 | 6,92 |
| 3 | 1 | 4,1 |
| 3 | 2 | 8,21 |
| 3 | 2 | 7,83 |
| 3 | 2 | 5,36 |
| 3 | 2 | 7,86 |

Для остатков каждой из моделей проверить выполняются ли основные предположения дисперсионного анализа (нормальность остатков и однородность дисперсий для различных уровней фактора). В случае, если эффект фактора значим, провести анализ значимости попарных различий средних значений отклика, используя один из методов апостериорного множественного сравнения средних.

При построении моделей полного ранга использовать следующие ограничения на параметры модели:

– эффект k-го уровня

Поскольку реализация идет с использованием библиотеки statsmodel, здесь один из уровней всегда будет приниматься за базовый, то есть обнуляться.

Перенесем данные в pandas DataFrame,

Проведем однофакторный дисперсионный анализ для фактора A и проведем тесты Шапиро-Уилка и для определения нормальности остатков и тест Левена для определения однородности дисперсий остатков.

Однофакторный ANOVA для фактора A:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| A | df | F | PR(>F) | sum\_sq |
| C(A) | 2.0 | 5.848943 | 0.008246 | 29.2565 |
| Residual | 25.0 |  |  | 62.525360 |

Фактор A является значимым, здесь:

Тест Шапиро-Уилка для остатков (фактор A): p-value = 0.9623

Тест Левена для остатков (фактор A): p-value = 0.1204

Здесь остатки являются нормальными и однородными для всех уровней фактора. То есть использование модели линейной регрессии имеет смысл.

Однофакторный ANOVA для фактора B:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| B | df | F | PR(>F) | Sum\_sq |
| C(B) | 1 | 0.752 | 0.393 | 2.583 |
| Residual | 26 |  |  | 89.19 |

Здесь фактор B не является значимым, соответственно и апостериорный анализ проводить нет смысла.

Для апостериорного анализа будем использовать метод Тьюки,

1. **Вычисление разницы средних**:
   * Для каждой пары групп вычисляется разница между средними значениями.
2. **Проверка значимости**:
   * Если разница средних превышает критическое значение, различия считаются статистически значимыми.
3. **Скорректированные p-значения**:
   * Метод Тьюки также вычисляет скорректированные p-значения для каждой пары групп.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Уровень 1 | Уровень 2 | Разн средних | P-value | Нижн граница интерв | Верхн граница  интерв | Отклоняем H0 |
| 1 | 2 | 0.09 | 0.9921 | -1.7785 | 1.9585 | Нет |
| 1 | 3 | 2.18 | 0.0198 | 0.3135 | 4.0505 | Да |
| 2 | 3 | 2.09 | 0.0177 | 0.3304 | 3.8536 | Да |

Таким образом можно сделать вывод о том, что различия между средними значениями отклика для уровней 1 и 2 не значимы, в отличие от 1,3 и 2,3.

**Вывод:**

Были изучены возможности дисперсионного однофакторного анализа (ANOVA), а так же изучен метод апостериорного анализа средних значений отклика при различных уровнях групп.

**Приложение:**

# %%

import pandas as pd

import statsmodels.api as sm

from statsmodels.formula.api import ols

from statsmodels.stats.multicomp import pairwise\_tukeyhsd

from scipy.stats import shapiro, levene

import matplotlib.pyplot as plt

# %%

data = pd.DataFrame({

    'A': [1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3],

    'B': [1, 1, 1, 1, 1, 2, 2, 2, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2],

    'Отклик': [3.33, 2.75, 3.32, 1.44, 4.41, 5.64, 6.96, 8.15, 6.99, 5.60, 5.02, 5.01, 3.24, 4.10, 5.12, 3.16, 3.84, 3.82, 6.51, 6.25, 6.61, 7.17, 6.92, 4.10, 8.21, 7.83, 5.36, 7.86]

})

# %%

data

# %%

# Однофакторный дисперсионный анализ для фактора A

model\_a = ols('Отклик ~ C(A)', data=data).fit()

anova\_table\_a = sm.stats.anova\_lm(model\_a, typ=2)

print("Однофакторный ANOVA для фактора A:")

print(anova\_table\_a)

# Проверка остатков на нормальность (фактор A)

residuals\_a = model\_a.resid

shapiro\_test\_a = shapiro(residuals\_a)

print(f"Тест Шапиро-Уилка для остатков (фактор A): p-value = {shapiro\_test\_a[1]:.4f}")

# Проверка однородности дисперсий (фактор A)

levene\_test\_a = levene(\*[data[data['A'] == level]['Отклик'] for level in data['A'].unique()])

print(f"Тест Левена для остатков (фактор A): p-value = {levene\_test\_a.pvalue:.4f}")

# %%

# Однофакторный дисперсионный анализ для фактора B

model\_b = ols('Отклик ~ C(B)', data=data).fit()

anova\_table\_b = sm.stats.anova\_lm(model\_b, typ=2)

print("Однофакторный ANOVA для фактора B:")

print(anova\_table\_b)

# Проверка остатков на нормальность (фактор B)

residuals\_b = model\_b.resid

shapiro\_test\_b = shapiro(residuals\_b)

print(f"Тест Шапиро-Уилка для остатков (фактор B): p-value = {shapiro\_test\_b[1]:.4f}")

# Проверка однородности дисперсий (фактор B)

levene\_test\_b = levene(\*[data[data['B'] == level]['Отклик'] for level in data['B'].unique()])

print(f"Тест Левена для остатков (фактор B): p-value = {levene\_test\_b.pvalue:.4f}")

# %%

# Апостериорный анализ

if anova\_table\_a['PR(>F)'][0] < 0.05:  # Если фактор A значим

    tukey\_a = pairwise\_tukeyhsd(endog=data['Отклик'], groups=data['A'], alpha=0.05)

    print("Попарные сравнения для фактора A:")

    print(tukey\_a)

if anova\_table\_b['PR(>F)'][0] < 0.05:  # Если фактор B значим

    tukey\_b = pairwise\_tukeyhsd(endog=data['Отклик'], groups=data['B'], alpha=0.05)

    print("Попарные сравнения для фактора B:")

    print(tukey\_b)