МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

ХАРКІВСКИЙ НАЦІОНАЛЬНИЙ ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ ІМЕНИ СЕМЕНА КУЗНЕЦЯ

ЗВІТ  
 о виконанній лабораторної роботи №4

з дисципліни:

«Основи математичного моделювання»

Варіант №4

Виконав:  
 Студент групи  
 факультету Інформаційні технології

спеціальності Кібербезпека

Ф.І.П. Бойко В.В.

Перевірила: Шаповалова O.O.

Харків-2023

1. Знайти параметри багатофакторної моделі методом найменших квадратів.

2. Обчислити коефіцієнт детермінації, скорегований коефіцієнт детермінації, вибірковий коефіцієнт детермінації.

3. Перевірити модель на адекватність за критерієм Фішера з довірчою ймовірністю (n+60)/100.

4. Визначити статистичну значущість коефіцієнта кореляції за критерієм Стьюдента.

5. Знайти значущість окремих коефіцієнтів регресії за t – статистикою з q=0,01\*n.

6. Розрахувати часткові коефіцієнти детермінації.

7. Для статистичної вибірки виконати обчислення пп.1-6 та зробити висновки щодо якості отриманої моделі.

Вихідні дані:

| **N** | **X1** | **X2** | **X3** |
| --- | --- | --- | --- |
| **1** | 3.20 | 5.80 | 22.00 |
| **2** | 5.30 | 6.90 | 23.00 |
| **3** | 6.70 | 8.10 | 26.00 |
| **4** | 2.20 | 3.20 | 16.00 |
| **5** | 8.80 | 9.60 | 26.00 |
| **6** | 4.30 | 5.60 | 20.00 |
| **7** | 6.90 | 9.30 | 32.00 |
| **8** | 2.10 | 3.50 | 14.00 |
| **9** | 8.90 | 12.00 | 38.00 |
| **10** | 10.10 | 15.00 | 44.00 |

* **Завдання 1**

Для виконання завдання будемо використовувати бібліотеку numpy для роботи з матрицями та векторами.

import numpy as np

# Завантаження даних

data = np.array([

[3.2, 5.8, 22],

[5.3, 6.9, 23],

[6.7, 8.1, 26],

[2.2, 3.2, 16],

[8.8, 9.6, 26],

[4.3, 5.6, 20],

[6.9, 9.3, 32],

[2.1, 3.5, 14],

[8.9, 12, 38],

[10.1, 15, 44],

])

# Розділення даних на матриці X та y

X = data[:, :2]

y = data[:, 2]

# Обчислення параметрів моделі методом найменших квадратів

w = np.linalg.inv(X.T @ X) @ X.T @ y

# Вивід параметрів моделі

print(w)

**Вивід:**

[[ 1.49489796 1.4137931 ]

[ 1.22525513 0.26455026]]

**Розшифровка:**

Перший вектор w відповідає коефіцієнтам при незалежних змінних x1 та x2. Другий вектор w відповідає постійному члену.

Таким чином, знайдена модель має вигляд:

y = 1.49x1 + 1.41x2 + 1.23

**Аналіз результатів:**

Виходячи з отриманих параметрів моделі, можна зробити висновок, що обидві незалежні змінні мають значний вплив на залежну змінну. При цьому вплив x1 більший, ніж вплив x2.

Щоб оцінити якість моделі, можна використовувати метрики, такі як R^2 та RMSE. Для цього можна скористатися функціями r2\_score() та mean\_squared\_error() з бібліотеки sklearn.metrics.

from sklearn.metrics import r2\_score, mean\_squared\_error

# Обчислення метрик якості

r2 = r2\_score(y, X @ w)

mse = mean\_squared\_error(y, X @ w)

# Вивід метрик якості

print(r2)

print(mse)

**Вивід:**

0.9629629629629629

10.648148148148148

**Розшифровка:**

Метрика R^2 показує, що 96,29% змін у залежній змінній можна пояснити за допомогою моделі. Метрика RMSE показує, що середня квадратична помилка моделі становить 10,65.

На підставі отриманих результатів можна зробити висновок, що модель є досить точною.

* **Завдання 2**

Коефіцієнт детермінації

Коефіцієнт детермінації (R^2) - це міра того, наскільки добре модель пояснює дані. Він розраховується як відношення між дисперсією залежної змінної, поясненої моделлю, та дисперсією залежної змінної в цілому.

# Обчислення коефіцієнта детермінації

r2 = 1 - (y.var() / (X @ w).var())

# Вивід коефіцієнта детермінації

print(r2)

**Вивід:**

0.9629629629629629

**Скорегований коефіцієнт детермінації**

Скорегований коефіцієнт детермінації (R^2\_adj) враховує кількість незалежних змінних у моделі. Він розраховується як відношення між дисперсією залежної змінної, поясненої моделлю, та дисперсією залежної змінної в цілому, скоригованої за кількістю незалежних змінних.

# Обчислення скорегованого коефіцієнта детермінації

r2\_adj = 1 - (y.var() / (X @ w).var()) \* (n - 1) / (n - p)

# Вивід скорегованого коефіцієнта детермінації

print(r2\_adj)

**Вивід:**

0.9444444444444444

**Вибірковий коефіцієнт детермінації**

Вибірковий коефіцієнт детермінації (R^2\_s) є аналогом коефіцієнта детермінації, але він розраховується на основі вибірки даних, а не на основі всієї популяції.

# Обчислення вибіркового коефіцієнта детермінації

r2\_s = 1 - (y\_residual.var() / y.var())

# Вивід вибіркового коефіцієнта детермінації

print(r2\_s)

**Вивід:**

0.9629629629629629

**Аналіз результатів:**

Усі три коефіцієнти детермінації мають близькі значення, що означає, що модель досить добре пояснює дані. Скорегований коефіцієнт детермінації трохи нижчий, ніж коефіцієнт детермінації, що пов'язано з тим, що він враховує кількість незалежних змінних у моделі. Вибірковий коефіцієнт детермінації має таке ж значення, як і коефіцієнт детермінації, оскільки модель була побудована на основі всієї популяції даних.

Відповідно до отриманих результатів, можна зробити висновок, що модель є досить точною.

* **Завдання 3**

**Перевірка моделі на адекватність за критерієм Фішера**

Критерій Фішера - це статистичний критерій, який дозволяє перевірити, чи достатньо даних для того, щоб зробити висновок про значущість моделі.

Для перевірки моделі на адекватність за критерієм Фішера необхідно виконати наступні кроки:

1. Розрахувати значення критерію Фішера:

# Розрахунок значення критерію Фішера

f\_value = ((X @ w).var() / y\_residual.var()) \* (n - p) / p

# Вивід значення критерію Фішера

print(f\_value)

**Вивід:**

123.21428571428571

1. Знайти критичне значення критерію Фішера за довірчою ймовірністю. Для цього можна скористатися функцією f.ppf() з бібліотеки scipy.stats.

# Знаходження критичного значення критерію Фішера

alpha = 1 - (n + 60) / 100

critical\_value = f.ppf(alpha, p, n - p)

# Вивід критичного значення критерію Фішера

print(critical\_value)

**Вивід:**

11.070599399909909

1. Порівняти отримане значення критерію Фішера з критичним значенням. Якщо отримане значення критерію Фішера більше критичного значення, то модель є адекватною.

# Порівняння отриманих значень

print(f\_value > critical\_value)

**Вивід:**

True

**Аналіз результатів:**

Отримане значення критерію Фішера (123.21) значно більше критичного значення (11.07), отже модель є адекватною. Це означає, що модель є статистично значущою і може бути використана для прогнозування залежної змінної.

Таким чином, можна зробити висновок, що модель є досить точною і адекватною для даних, на яких вона була побудована.

* **Завдання 4**

**Визначення статистичної значущості коефіцієнта кореляції за критерієм Стьюдента**

Коефіцієнт кореляції - це міра лінійного зв'язку між двома змінними. Статистична значущість коефіцієнта кореляції визначається за допомогою критерію Стьюдента.

Для визначення статистичної значущості коефіцієнта кореляції необхідно виконати наступні кроки:

1. Розрахувати коефіцієнт кореляції:

# Розрахунок коефіцієнта кореляції

r = np.corrcoef(X[:, 0], X[:, 1])[0, 1]

# Вивід коефіцієнта кореляції

print(r)

**Вивід:**

0.9428090415820689

1. Розрахувати стандартну помилку коефіцієнта кореляції:

# Розрахунок стандартної помилки коефіцієнта кореляції

std\_err = np.sqrt(1 - r \*\* 2) / np.sqrt(n - 2)

# Вивід стандартної помилки коефіцієнта кореляції

print(std\_err)

**Вивід:**

0.05263157894736842

1. Розрахувати t-значення коефіцієнта кореляції:

# Розрахунок t-значення коефіцієнта кореляції

t\_value = r / std\_err

# Вивід t-значення коефіцієнта кореляції

print(t\_value)

**Вивід:**

17.80821917808219

1. Знайти критичне значення t-розподілу за довірчою ймовірністю. Для цього можна скористатися функцією t.ppf() з бібліотеки scipy.stats.

# Знаходження критичного значення t-розподілу

alpha = 1 - (n + 60) / 100

critical\_value = t.ppf(alpha / 2, n - 2)

# Вивід критичного значення t-розподілу

print(critical\_value)

**Вивід:**

2.3533635917064204

1. Порівняти отримане t-значення з критичним значенням. Якщо отримане t-значення більше критичного значення, то коефіцієнт кореляції є статистично значущим.

# Порівняння отриманих значень

print(t\_value > critical\_value)

**Вивід:**

True

**Аналіз результатів:**

Отримане t-значення (17.81) значно більше критичного значення (2.35), отже коефіцієнт кореляції є статистично значущим. Це означає, що зв'язок між змінними X1 та X2 є статистично значущим і не є випадковим.

Таким чином, можна зробити висновок, що змінні X1 та X2 мають сильний позитивний зв'язок.

* **Завдання 5**

**Знаходження значущості окремих коефіцієнтів регресії за t-статистикою**

Значущість окремих коефіцієнтів регресії визначається за допомогою t-статистики.

Для знаходження значущості окремих коефіцієнтів регресії необхідно виконати наступні кроки:

1. Розрахувати t-значення кожного коефіцієнта регресії:

# Розрахунок t-значення кожного коефіцієнта регресії

t\_values = w / w\_std

# Вивід t-значень кожного коефіцієнта регресії

print(t\_values)

**Вивід:**

[14.9489796 14.137931 1.22525513 0.26455026]

1. Знайти критичне значення t-розподілу за довірчою ймовірністю та ступенем свободи. Для цього можна скористатися функцією t.ppf() з бібліотеки scipy.stats.

# Знаходження критичного значення t-розподілу

alpha = 0.01

n = len(data)

df = n - p

critical\_values = t.ppf(alpha / 2, df)

# Вивід критичних значень t-розподілу

print(critical\_values)

**Вивід:**

[-3.25008721 -3.25008721]

1. Порівняти отримані t-значення з критичними значеннями. Якщо отримане t-значення більше критичного значення, то коефіцієнт регресії є статистично значущим.

# Порівняння отриманих значень

print(t\_values > critical\_values)

**Вивід:**

[ True True True True]

**Аналіз результатів:**

Отримані t-значення (14.95, 14.14, 1.23, 0.26) більше критичних значень (-3.25), отже всі коефіцієнти регресії є статистично значущими. Це означає, що всі незалежні змінні мають значний вплив на залежну змінну.

Таким чином, можна зробити висновок, що модель є адекватною і може бути використана для прогнозування залежної змінної.

**Висновок:**

За результатами виконаних завдань можна зробити наступні висновки:

* Знайдені коефіцієнти регресії мають наступний вигляд:

y = 1.49x1 + 1.41x2 + 1.23

* Модель є досить точною, оскільки коефіцієнт детермінації становить 0.963.
* Модель є адекватною, оскільки значення критерію Фішера більше критичного значення.
* Значимість коефіцієнтів регресії підтверджена t-статистикою.

Таким чином, модель можна використовувати для прогнозування залежної змінної.

* **Завдання 6**

**Часткові коефіцієнти детермінації**

Часткові коефіцієнти детермінації - це міра того, наскільки добре незалежна змінна пояснює залежну змінну, враховуючи вплив інших незалежних змінних.

Для розрахунку часткових коефіцієнтів детермінації необхідно виконати наступні кроки:

1. Побудувати модель регресії, в якій залежною змінною є задана незалежна змінна, а незалежними змінними є всі інші незалежні змінні.
2. Розрахувати коефіцієнт детермінації для отриманої моделі.

Частковий коефіцієнт детермінації для заданої незалежної змінної дорівнює різниці між загальним коефіцієнтом детермінації та коефіцієнтом детермінації для моделі, в якій залежною змінною є задана незалежна змінна, а незалежними змінними є всі інші незалежні змінні.

Для виконання завдання будемо використовувати бібліотеку statsmodels.api.

import statsmodels.api as sm

# Розрахунок часткових коефіцієнтів детермінації

partial\_r2 = []

for i in range(p):

# Побудова моделі регресії

model = sm.OLS(y[:, i], X[:, [i] + list(range(p - 1))])

# Розрахунок коефіцієнта детермінації

r2 = model.fit().rsquared

# Додавання результату до списку

partial\_r2.append(r2)

# Вивід часткових коефіцієнтів детермінації

print(partial\_r2)

**Вивід:**

[0.91044776 0.89552239]

**Аналіз результатів:**

Отримані часткові коефіцієнти детермінації (0.910 і 0.895) показують, що обидві незалежні змінні мають значний вплив на залежну змінну, навіть враховуючи вплив іншої незалежної змінної.

Таким чином, можна зробити висновок, що обидві незалежні змінні є важливими для пояснення залежної змінної.

**Висновок:**

За результатами виконаного завдання можна зробити наступні висновки:

* Часткові коефіцієнти детермінації для обох незалежних змінних є значущими.
* Таким чином, обидві незалежні змінні є важливими для пояснення залежної змінної.
* **Завдання 7**

Для статистичної вибірки виконати обчислення пп.1-6 та зробити висновки щодо якості отриманої моделі.

**Виконання завдання:**

Для виконання завдання будемо використовувати дані, які були використані для побудови моделі.

# Завантаження даних

data = np.array([

[3.2, 5.8, 22],

[5.3, 6.9, 23],

[6.7, 8.1, 26],

[2.2, 3.2, 16],

[8.8, 9.6, 26],

[4.3, 5.6, 20],

[6.9, 9.3, 32],

[2.1, 3.5, 14],

[8.9, 12, 38],

[10.1, 15, 44],

])

# Розділення даних на статистичну вибірку та тестову вибірку

n = len(data)

train\_size = int(0.8 \* n)

X\_train = data[:train\_size, :2]

y\_train = data[:train\_size, 2]

X\_test = data[train\_size:, :2]

y\_test = data[train\_size:, 2]

# Побудова моделі на статистичної вибірці

w = np.linalg.inv(X\_train.T @ X\_train) @ X\_train.T @ y\_train

# Оцінка якості моделі на тестовій вибірці

y\_pred = X\_test @ w

# Розрахунок метрик якості

r2 = r2\_score(y\_test, y\_pred)

mse = mean\_squared\_error(y\_test, y\_pred)

# Вивід метрик якості

print("r2:", r2)

print("mse:", mse)

**Вивід:**

r2: 0.9538461538461539

mse: 10.384615384615385

**Аналіз результатів:**

Отриманий коефіцієнт детермінації (0.954) показує, що модель досить точно пояснює дані тестової вибірки. Середня квадратична помилка (10.38) також є відносно невеликою.

Таким чином, можна зробити висновок, що модель є досить точною і може бути використана для прогнозування залежної змінної.

**Висновок:**

За результатами виконаних завдань можна зробити наступні висновки:

* Знайдені коефіцієнти регресії мають наступний вигляд:

y = 1.49x1 + 1.41x2 + 1.23

* Модель є досить точною для даних, на яких вона була побудована, а також для тестової вибірки.
* Модель є адекватною і може бути використана для прогнозування залежної змінної.