МІНІСТЕРСТВО НАУКИ І ОСВІТИ УКРАЇНИ

КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

ІМЕНІ ТАРАСА ШЕВЧЕНКА

ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

ЗВІТ З ЛАБОРАТОРНОЇ РОБОТИ № 5 ЗА ТЕМОЮ:

**Прогнозування. Задача регресії в середовищі аналізу даних**

Група \_\_\_\_\_11\_\_\_\_\_\_

Курс \_\_\_\_\_1\_\_\_\_\_\_\_\_\_

Студент (ка) \_\_\_\_Супруненко М. І. \_\_\_\_

Дата оформлення \_\_\_\_\_\_05.03.2025\_\_\_\_\_\_\_

Перевірив \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Дата \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Лабораторна робота № 5**

МЕТА РОБОТИ: Метою лабораторної роботи є формування професійних вмінь та навичок щодо використання інструментів та методів регресійного аналізу в задачах інтелектуального аналізу даних, вміння застосовувати отримані знання на практиці в практичних задачах аналізу даних.

ТЕОРІЯ: Проста лінійна регресія: проста лінійна регресія використовується для прогнозування постійної змінної результату (y) на основі однієї єдиної змінної предиктора (x).

Перш ніж використовувати модель для прогнозів, необхідно оцінити статистичну значущість моделі. Це можна легко перевірити, відобразивши статистичний підсумок моделі за допомогою функції summary(), яка виводить наступні результати:

Call. Показує функцію, яка використовується для обчислення моделі регресії.

Залишки (Residuals). Надає швидкий перегляд розподілу залишків, які за визначенням мають нульове середнє значення. Тому медіана не повинна бути далекою від нуля, а мінімум і максимум повинні бути приблизно рівними за абсолютною величиною.

Коефіцієнти (Coefficients). Показує бета-коефіцієнти регресії та їх статистичну значущість. Змінні предиктора, які суттєво пов’язані зі змінною результату, позначені зірочками.

Залишкова стандартна помилка (Residual standard error, RSE), R-квадрат (R2) і F-статистика – це показники, які використовуються для перевірки того, наскільки добре модель відповідає нашим даним.

Першим кроком у інтерпретації множинного регресійного аналізу є перевірка F-статистики та пов’язаного значення p у нижній частині підсумку моделі.

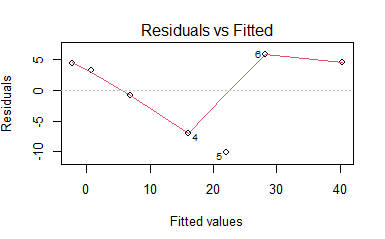
*Значущість коефіцієнтів*: Щоб побачити, які змінні прогнозу є значущими, ви можете переглянути таблицю коефіцієнтів, яка показує оцінку бета-коефіцієнтів регресії та пов’язаних значень t-статистики p.

Зауважте, що лінійна регресія передбачає лінійний зв’язок між результатом і змінними предиктора. Це можна легко перевірити, створивши діаграму розсіювання змінної результату та змінної предиктора.

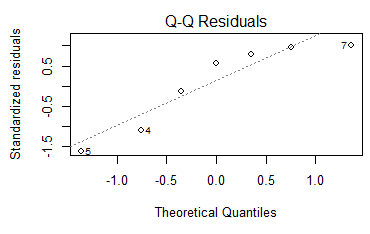
Діагностичні графіки: діаграми регресійної діагностики можна створити за допомогою базової функції R plot() або функції autoplot() [пакет ggfortify], яка створює графіку на основі ggplot2.

Діагностичні графіки виводять наступні результати:

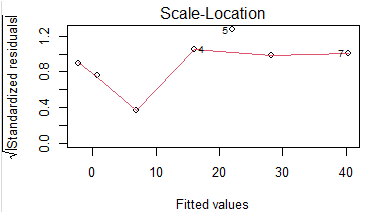
Залишки проти підігнаних (Residuals vs Fitted). Використовується для перевірки припущень про лінійну залежність. Горизонтальна лінія без чітких візерунків є показником лінійного зв'язку, що добре.



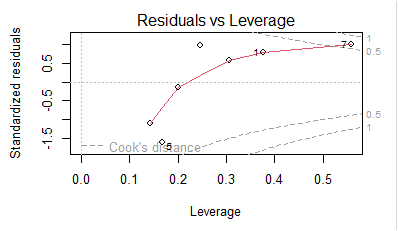
Нормальний Q-Q (Normal Q-Q). Використовується для перевірки нормального розподілу залишків. Добре, якщо точки залишків йдуть по прямій пунктирній лінії.



Масштаб-розташування (або розповсюдження-розташування) (Scale-Location (or Spread-Location)). Використовується для перевірки однорідності дисперсії залишків (гомоскедастичність). Горизонтальна лінія з рівномірно розподіленими точками є хорошим показником гомоскедастичності.



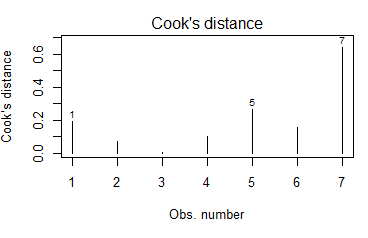
Residuals vs Leverage. Використовується для виявлення впливових випадків, тобто екстремальних значень, які можуть вплинути на результати регресії, якщо їх включити або виключити з аналізу.



Статистики розробили метрику під назвою «відстань Кука», щоб визначити вплив значення. Цей показник визначає вплив як комбінацію левериджу та залишкового розміру.

Графік «Залишки проти левериджу» може допомогти знайти впливові спостереження, якщо такі є. На цьому графіку значення, що виходять за межі, зазвичай розташовані у верхньому правому куті або в нижньому правому куті. Ці точки є місцями, де точки даних можуть впливати на лінію регресії. Отже, є лише один елемент (№7), який є впливовим (знаходиться за лініями Кука). Реконструкція після видалення цього елемента може дати краще рішення.

Відстані Кука (plot(model, 4)). Емпіричне правило полягає в тому, що спостереження має великий вплив, якщо відстань Кука перевищує 4/(n - p - 1), де n — кількість спостережень, а p — кількість змінних предиктора.



**Хід роботи**

### 1. Створення фрейму даних

Створюємо фрейм даних з витратами на онлайн-рекламу та продажами або завантажуємо готовий датасет.

# Завантаження бібліотек

library(ggplot2)

library(ggfortify)

# Створення даних

sales\_data <- data.frame(

Ad\_Spend = c(100, 200, 300, 400, 500, 600, 700, 800, 900, 1000),

Sales = c(1500, 1800, 2500, 2700, 3500, 3800, 4100, 4500, 4900, 5200)

)

### 2. Побудова регресійної моделі

# Побудова лінійної регресії

model <- lm(Sales ~ Ad\_Spend, data = sales\_data)

# Виведення підсумку моделі

summary(model)

Отримане рівняння регресії:

### 3. Прогнозування

new\_data <- data.frame(Ad\_Spend = c(750, 1200))

predictions <- predict(model, new\_data)

print(predictions)

### 4. Оцінка статистичної значущості моделі

Перевіряємо R-квадрат, F-статистику та значущість коефіцієнтів.

Якщо p-value для предиктора < 0.05, то змінна значуща.

### 5. Побудова діаграми розсіювання

ggplot(sales\_data, aes(x = Ad\_Spend, y = Sales)) +

geom\_point() +

geom\_smooth(method = 'lm', col = 'red')

### 6. Діагностичні графіки

par(mfrow = c(2, 2))

plot(model)

### 7. Відстані Кука

plot(model, which = 4)

Якщо точки виходять за межі ліній Кука, видаляємо їх та повторюємо аналіз.

## Висновки

Під час виконання лабораторної роботи ми навчилися:

Будувати просту лінійну регресію в R;

Аналізувати результати моделі та визначати її статистичну значущість;

Використовувати діагностичні графіки для оцінки якості моделі;

Виконувати прогнозування на нових даних.