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.

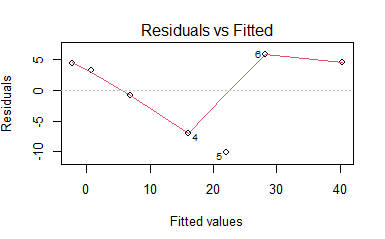
*Значущість коефіцієнтів*: Щоб побачити, які змінні прогнозу є значущими, ви можете переглянути таблицю коефіцієнтів, яка показує оцінку бета-коефіцієнтів регресії та пов’язаних значень t-статистики p.

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.

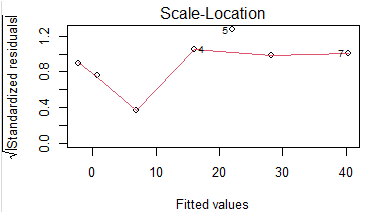
Діагностичні графіки виводять наступні результати:

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.

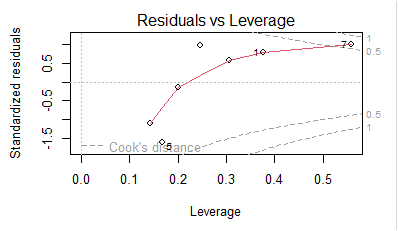


Нормальний Q-Q (Normal Q-Q). Використовується для перевірки нормального розподілу залишків. Добре, якщо точки залишків йдуть по прямій пунктирній лінії.

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. Завантаження бібліотек та даних

|  |
| --- |
| library(ggplot2)  library(dplyr)  # Завантаження CSV-файлу  df <- read.csv("data1.csv", sep=";", stringsAsFactors=FALSE)  # Перетворення числових значень для всіх колонок, крім 'Country.name'  df <- df %>%    mutate(across(.cols = -Country.name,  # Виключаємо 'Country.name'                  .fns = ~ as.numeric(gsub(",", ".", .)),                  .names = "{.col}"))  # Перевірка структури даних  str(df)  summary(df) |
|  |

### Візуалізація зв'язку між змінними

|  |
| --- |
| ggplot(df, aes(x = Log\_GDP\_per\_capita, y = Ladder\_score)) +    geom\_point() +    geom\_smooth(method = "lm", col = "red") +    labs(title = "Залежність Ladder score від Log GDP per capita") |
|  |

### 3. Побудова простої лінійної регресії

|  |
| --- |
| # If needed, after renaming  model <- lm(Ladder\_score ~ Log\_GDP\_per\_capita, data = df)  summary(model) |
|  |

### 4. Перевірка діагностичних графіків

|  |
| --- |
| par(mfrow=c(2,2))  plot(model) |
|  |

### 5. Прогнозування

|  |
| --- |
| new\_data <- data.frame(`Log\_GDP\_per\_capita` = c(10, 20, 30))  predictions <- predict(model, new\_data)  print(predictions) |
|  |
| new\_data <- data.frame(`Log\_GDP\_per\_capita` = c(2.5, 0.1, 3))  predictions <- predict(model, new\_data)  print(predictions) |

**6. Обчислення відстані Кука**

|  |
| --- |
| # Обчислення відстані Кука  cooks\_dist <- cooks.distance(model)  # Візуалізація відстаней Кука  plot(cooks\_dist, type="h", main="Відстань Кука для кожного спостереження", ylab="Cook's Distance")  abline(h = 4/(nrow(df)), col="red", lty=2)  # Граничне значення |
|  |

**7. Ідентифікація впливних точок**

|  |
| --- |
| # Порогове значення  threshold <- 4 / nrow(df)  # Обчислення Cook's distance для первинної моделі  # (припускаємо, що модель була навчена за цими змінними)  model <- lm(Ladder\_score ~ Log\_GDP\_per\_capita, data = df)  cooks\_dist <- cooks.distance(model)  # Визначення впливних точок  influential\_points <- which(cooks\_dist > threshold)  # Видалення впливних точок  df\_cleaned <- df[-influential\_points, ]  # Повторне навчання моделі на очищених даних  model\_cleaned <- lm(Ladder\_score ~ Log\_GDP\_per\_capita, data = df\_cleaned)  summary(model\_cleaned) |
|  |

## Висновки

1. Логарифм ВВП на душу населення (Log\_GDP\_per\_capita) суттєво впливає на рівень щастя (Ladder\_score), що підтверджується значущими коефіцієнтами у моделі.
2. Діагностичні графіки показали, що модель має нормальний розподіл залишків і загалом добре підходить до даних, хоча деякі точки могли впливати на її точність.
3. Впливові точки, знайдені за допомогою відстані Кука, мали значний вплив на результати, тому їх видалення покращило модель.
4. Очищена модель (без впливових точок) є більш стабільною та точнішою, що видно з покращених статистичних показників.
5. Отримана модель може використовуватися для прогнозування рівня щастя на основі економічних показників, але слід враховувати можливу нелінійність та інші фактори.
6. Модель показує, що Log GDP per capita має значний вплив на рівень щастя. Але R2=0.5674R^2 = 0.5674R2=0.5674 говорить про те, що є ще інші фактори, які впливають на щастя (можливо, соціальні умови, здоров’я, освіта тощо).