Київський національний університет імені Тараса Шевченка
Факультет Комп'ютерних наук і кібернетики
Кафедра Системного аналізу і теорії прийняття рішень

Звіт до Лабораторної роботи №2
На тему: "Кореляційний аналіз даних"

Студента 3 курсу
Групи САТР-3
Арзамасцева Владислава Олександровича

Київ – 2022

Зміст

Постановка задачі

Визначитися з множиною скалярних змінних (Не менше трьох змінних), взятих з довільного датасету і з'ясувати істотність їх статистичного зв'язку.

1. Визначитися з множиною скалярних змінних, для яких маєте намір з'ясувати істотність їх статистичного зв'зку;
2. На основі результатів попередньої обробки обраного набору даних визначитися, які характеристики статистичного зв'язку потрібно використати при подальшому їх кореляційному аналізі
3. Провести аналіз істотності парних статистичних зв'язків для усіх пар скалярних змінних
4. Провести аналіз істотності множинних статистичних зв'язків між кожною обраною в якості залежної скалярною змінною і множиною усіх інших скалярних змінних (Які виступають у ролі незалежних змінних).

Для проведення аналізу візьмемо 3 змінні:

1. age
2. chol
3. trestbps

Опис вхідної інформації

Дані, які використовуються для проведення цього аналізу, взяті з набору даних, зібраного чотирма лікарнями в Клівленді, Угорщині, Швейцарії та Лонг-Біч [1]. Дані називають набором даних про хвороби серця UCI. Цей набір даних складається з 303 осіб з 14 атрибутами, де 138 осіб представлені без серцево-судинних захворювань і 165 осіб, представлених ССЗ. Спочатку було 76 атрибутів, але опубліковані експерименти стосуються використання підмножини лише 14 атрибутів. Змінні після заміни числових позначень категорій на буквені:

1. age: Вік у роках
2. sex: Стать: (M – чол., F – жін.)
3. cp: Тип болю у грудях: A – типова стенокардія, B – нетипова стенокардія, C – неангінальний біль, D – асимптоматична
4. trestbps: Кров'яний тиск у спокої (в мм рт.ст. при надходженні в лікарню)
5. chol: Сироватковий холестерин в мг/дл
6. fbs: (рівень цукру в крові натще > 120 мг/дл) (T = так; F = ні)
7. restecg: Електрокардіографічні результати у стані спокою — A: нормальні, B: наявність аномалії хвиль ST-T (інверсії T-хвиль та/або висота/спад ST > 0,05 мВ), C: відображення ймовірної або певної гіпертрофії лівого шлуночка за критеріями Естеса
8. thalach: Найбільша частота серцебиття
9. exang: Біль, що з'являється під час вправ (T = так; F = ні)
10. oldpeak: Спад ST, спричинений вправами, порівняно зі станом спокою
11. slope: Нахил ST-сегмента — A: підйомний, B: плоский, C: нисхідний
12. ca: Кількість основних судин (0-3), забарвлених флюороскопією
13. thal: A = у нормі; B = фіксоване відхилення; C = оборотний дефект
14. target: діагностика серцевих захворювань (стан ангіографічного захворювання) — L – звуження діаметра < 50%, M – звуження діаметра > 50%

Змінні 1, 4, 5, 8, 10, 12, 13 – кількісні скалярні, змінні 2, 3, 6, 7, 9, 11, 14 – скалярні якісні номінальні. Всі змінні – категоризовані (дискретні).

Аналіз даних

Для визначення істотності статистичного зв’язку між змінними нам знадобляться наступні 4 величини:

1. Коефіцієнт кореляції
2. Коефіцієнт детермінації
3. Множинний коефіцієнт кореляції

```
[1] "Coefficient of correlation between  age  and  trestbps  is:  0.279350906561288"
[1] "Coefficient of correlation between  age  and  chol  is:  0.213677956559562"
[1] "Coefficient of correlation between  chol  and  trestbps  is:  0.123174206532391"


================================================================================

[1] "P-value of  age  and  trestbps  is:  7.76226907480995e-07"
[1] "P-value of  age  and  chol  is:  0.0001786286434144 9"
[1] "P-value of  chol  and  trestbps  is:  0.0320820536108711"


================================================================================

[1] "Coefficient of determination for 'age' and 'trestbps':  0.078036928996614"
[1] "Coefficient of determination for 'age' and 'chol':  0.0456582691194699"
[1] "Coefficient of determination for 'trestbps' and 'chol':  0.015171885154884"


================================================================================

[1] "Multiple correlation coefficient with 'age' as dependent variable:  0.33267017733609"
[1] "Multiple correlation coefficient with 'trestbps' as dependent variable:  0.28680976401047"
[1] "Multiple correlation coefficient with 'chol' as dependent variable:  0.223672721594632"


================================================================================

[1] "P-value for 'age' as dependent variable:  2.90704996915259e-09"
[1] "P-value for 'trestbps' as dependent variable:  3.79561198580997e-07"
[1] "P-value for 'chol' as dependent variable:  8.59317211772679e-05"
```

```
[1] "Coefficient of determination for 'age' and 'trestbps':  0.078036928996614"
[1] "Coefficient of determination for 'age' and 'chol':  0.0456582691194699"
[1] "Coefficient of determination for 'trestbps' and 'chol':  0.015171885154884"


========================================================================

[1] "Multiple correlation coefficient with 'age' as dependent variable:  0.33267017733609"
[1] "Multiple correlation coefficient with 'trestbps' as dependent variable:  0.28680976401047"
[1] "Multiple correlation coefficient with 'chol' as dependent variable:  0.223672721594632"


========================================================================

[1] "P-value for 'age' as dependent variable:  2.90704996915259e-09"
[1] "P-value for 'trestbps' as dependent variable:  3.79561198580997e-07"
[1] "P-value for 'chol' as dependent variable:  8.59317211772679e-05"


========================================================================

[1] "Coefficient of determination for 'age' as dependent variable:  0.110669446888826"
[1] "Coefficient of determination for 'age' as dependent variable:  0.110669446888825"
[1] "Coefficient of determination for 'trestbps' as dependent variable:  0.0822598407317421"


========================================================================
```

Аналіз істотності парних статистичних зв'язків

1. Коефіцієнт кореляції:
    1. age and trestbps is: 0.279350906561288
    2. age and chol is: 0.213677956559562
    3. chol and trestbps is: 0.123174206532391
2. Рівень значущості (p-value):
    1. age and trestbps is: 7.76226907480995e-07
    2. age and chol is: 0.00017862864341449
    3. chol and trestbps is: 0.0320820536108711
3. Коефіцієнт детермінації
    1. 'age' and 'trestbps': 0.078036928996614
    2. 'age' and 'chol': 0.0456582691194699
    3. 'trestbps' and 'chol': 0.0456582691194699
4. Впорядковані пари по рівню значущості:

1. chol and trestbps
2. age and chol
3. age and trestbps

Висновок:

Бачимо, що для всіх пар змінних коефіцієнт кореляції не перевищує 0.28, тобто відношення між змінними у парах слабке. Проте рівень значущості для пар (age and trestbps), (age and chol) і (chol and trestbps) близький до нуля. Це означає, що між змінними в цих парах є зв'язок. Також бачимо, що коефіцієнт детермінації близький до нуля, тож залежність між змінними – слабка.

Аналіз істотності множинних статистичних зв'язків

1. Множинний коефіцієнт кореляції
   1. with 'age' as dependent variable: 0.33267017733609
   2. with 'trestbps' as dependent variable: 0.28680976401047
   3. with 'chol' as dependent variable: 0.223672721594632
2. Рівень значущості:
   1. for 'age' as dependent variable: 2.90704996915259e-09
   2. for 'trestbps' as dependent variable: 3.79561198580997e-07
   3. for 'chol' as dependent variable: 8.59317211772679e-05
3. Коефіцієнт детермінації:
   1. for 'age' as dependent variable: 0.110669446888826
   2. for 'age' as dependent variable: 0.110669446888825
   3. for 'trestbps' as dependent variable: 0.0822598407317421
4. Впорядкована послідовність усіх скалярних змінних у порядку спадання істотності множинного статистичного зв'язку їх з множиною усіх інших
   1. chol
   2. trestbps
   3. age

Висновок:

Множинні статистичні зв'язки можливі у всіх випадках, проте найбільш імовірні, коли залежна змінна – age. Також бачимо, що значення коефіцієнта детермінації збільшилось, проте зв'язок усе ще слабкий.

Список використаних джерел

1. Датасет: https://www.kaggle.com/datasets/amdirfan/predicte-heart-disease?resource=download
2. Опис змінних: https://towardsdatascience.com/heart-disease-classification-8359c26c7d83
3. Аналіз Даних, 2001, Слабоспицький
4. https://ru.wikipedia.org/wiki/%D0%9A%D0%BE%D1%8D%D1%84%D1%84%D0%B8%D1%86%D0%B8%D0%B5%D0%BD%D1%82_%D0%B4%D0%B5%D1%82%D0%B5%D1%80%D0%BC%D0%B8%D0%BD%D0%B0%D1%86%D0%B8%D0%B8
5. https://www.scribbr.com/statistics/p-value/
6. https://dataschool.com/fundamentals-of-analysis/correlation-and-p-value/
7. https://dzone.com/articles/what-is-p-value-in-layman-terms

Додатки

```r
dataHospital <- read.csv( file: 'data_hospital.csv')

printDelimiterWithNewLines <- function() {
  cat("\n\n==========================================================================\n\n")
}

printEmptyLine <- function() {
  cat("\n")
}

printCorellationCoefficient <- function(vector1, vector2, name1, name2) {
  correlation <- cor(vector1, vector2)
  print(
    paste("Coefficient of correlation between ", name1, " and ", name2, " is: ", correlation)
  ) ^printCorellationCoefficient
}

printPValue <- function(vector1, vector2, name1, name2) {
  res <- cor.test(vector1, vector2)
  pValue <- res$p.value
  print(
    paste("P-value of ", name1, " and ", name2, " is: ", pValue)
  ) ^printPValue
}

analyze <- function() {
  printCorellationCoefficient(dataHospital$age, dataHospital$trestbps, name1: "age", name2: "trestbps")
  printCorellationCoefficient(dataHospital$age, dataHospital$chol, name1: "age", name2: "chol")
  printCorellationCoefficient(dataHospital$chol, dataHospital$trestbps, name1: "chol", name2: "trestbps")
  printDelimiterWithNewLines()
  printPValue(dataHospital$age, dataHospital$trestbps, name1: "age", name2: "trestbps")
  printPValue(dataHospital$age, dataHospital$chol, name1: "age", name2: "chol")
  printPValue(dataHospital$chol, dataHospital$trestbps, name1: "chol", name2: "trestbps")
```

```r
printPValue(dataHospital$age, dataHospital$trestbps, name1: "age", name2: "trestbps")
printPValue(dataHospital$age, dataHospital$chol, name1: "age", name2: "chol")
printPValue(dataHospital$chol, dataHospital$trestbps, name1: "chol", name2: "trestbps")
printDelimiterWithNewLines()
age <- dataHospital$age
trestbps <- dataHospital$trestbps
chol <- dataHospital$chol
age.model <- lm(age ~ trestbps + chol)
print(
  paste(
    "Multiple correlation coefficient with 'age' as dependent variable: ",
    cor(age.model$model$age, age.model$fitted.values)
  )
)
trestbps.model <- lm(trestbps ~ age + chol)
print(
  paste(
    "Multiple correlation coefficient with 'trestbps' as dependent variable: ",
    cor(trestbps.model$model$trestbps, trestbps.model$fitted.values)
  )
)
chol.model <- lm(chol ~ trestbps + age)
print(
  paste(
    "Multiple correlation coefficient with 'chol' as dependent variable: ",
    cor(chol.model$model$chol, chol.model$fitted.values)
  )
)
printDelimiterWithNewLines()
ageTrestbpsLM <- lm(age ~ trestbps)
print(
  paste(
    "Coefficient of determination for 'age' and 'trestbps': ",
```

```
  paste(
    "Coefficient of determination for 'age' and 'trestbps': ",
    summary(ageTrestbpsLM)$r.squared
  )
)
ageCholLM <- lm(age ~ chol)
print(
  paste(
    "Coefficient of determination for 'age' and 'chol': ",
    summary(ageCholLM)$r.squared
  )
)
trestbpsCholLM <- lm(trestbps ~ chol)
print(
  paste(
    "Coefficient of determination for 'age' and 'chol': ",
    summary(trestbpsCholLM)$r.squared
  )
)
printDelimiterWithNewLines()
ageMulCorTest <- cor.test(age.model$model$age, age.model$fitted.values)
ageMulPValue <- ageMulCorTest$p.value
print(
  paste(
    "P-value for 'age' as dependent variable: ",
    ageMulPValue
  )
)
trestBpsMulCorTest <- cor.test(trestbps.model$model$trestbps, trestbps.model$fitted.values)
trestbpsMulPValue <- trestBpsMulCorTest$p.value
print(
  paste(
    "P-value for 'trestbps' as dependent variable: "
```

```
  ageMulPValue <- ageMulCorTest$p.value
  print(
    paste(
      "P-value for 'age' as dependent variable: ",
      ageMulPValue
    )
  )
  trestBpsMulCorTest <- cor.test(trestbps.model$model$trestbps, trestbps.model$fitted.values)
  trestbpsMulPValue <- trestBpsMulCorTest$p.value
  print(
    paste(
      "P-value for 'trestbps' as dependent variable: ",
      trestbpsMulPValue
    )
  )
  cholMulCorTest <- cor.test(chol.model$model$chol, chol.model$fitted.values)
  cholMulPValue <- cholMulCorTest$p.value
  print(
    paste(
      "P-value for 'chol' as dependent variable: ",
      cholMulPValue
    )
  )
  printDelimiterWithNewLines() ^analyze
}

analyze()
```