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

# Звіт до Лабораторної роботи №3
# На тему: "Побудова математичної моделі методами аналізу 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, провести попередній аналіз даних, з'ясувати істотність їх статистичного зв'язку, побудувати математичну модель.

1. Визначитися з множиною скалярних змінних, для яких будуємо матмодель зв'язку.
2. Визначитись, яка скалярна змінна буде залежною, а які змінні – незалежними
3. Методами якого розділу аналізу даних можлива побудова матмоделі зв'язку
4. Провести побудову матмоделі методами аналізу даних для обраних змінних. Для цього:
   - 4.1 Визначитися з класом апроксимуючих параметричних функцій для правої частини моделі
   - 4.2 Уточнити структуру матмоделі
   - 4.3 З'ясувати якість отриманої матмоделі
   - 4.4 Сформулювати висновки

Для проведення аналізу візьмемо 3 змінні:

1. Happiness_Score – залежна
2. Life_Expectancy – вільна
3. Freedom – вільна

# Опис вхідної інформації

Дані взято із Всесвітнього Звіту про Щастя – публікації Мережі Рішень Сталого Розвитку ООН за 2015 і 2016 роки. Він містить статті та рейтинги національного щастя, засновані на рейтингах респондентів власного життя, які звіт також корелює з різними факторами життя. Датасет [1] містить наступні змінні:

1 Country – країна
2 Region – регіон
3 Happiness_Rank – місце в Рейтингу Щастя за даний рік
4 Happiness_Score – показник щастя, число від 0 до 1 включно
5 Standard_Error – похибка у вимірюванні показника щастя
6 Economy_GDP_per_Capita – міра економічного виробництва країни, яка враховує її кількість людей. Вона ділить валовий внутрішній продукт країни на її загальне населення.
7 Life_Expectancy – Очікувана тривалість життя у долях від 100 років,
8 Freedom – показник свободи прийняття рішень, від 0 до 1,
9 Government_Corruption – показник корупції влади, від 0 до 1,
10 Generosity – готовність людей жертвувати гроші благодійним організаціям по відношенню до ВВП на душу населення
11 Dystopia Residual – в якості базової лінії автори звіту складають гіпотетичну найгіршу країну в світі. Вони називають цю країну антиутопією. Вона потрібна, щоб мати приблизне уявлення про те, як швидко зростає щастя, коли ви покращуєте різні критерії. Якщо ви підключите тривалість життя вашої країни, корупцію тощо до рівняння лінії, вона передбачить, яким має бути ваш рейтинг щастя. Але більшість країн не ідеально підходять під модель. Вони, як правило, трохи вище або нижче, ніж прогноз. Різниця між прогнозом і фактичним значенням називається залишковою. Позитивний залишок означає, що ви з якоїсь причини щасливіші, ніж модель передбачила б, а негативний залишковий означає, що ви з якоїсь причини менш щасливі. І нульовий залишок означає, що ви саме там, де передбачена модель.
12 year – рік, 2015 або 2016

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

# Аналіз даних

## Попередній аналіз даних

Почнемо з побудови полігонів частот для змінних Freedom, Happiness_Score, Life_Expectancy (рис. 1, 2 і 3).

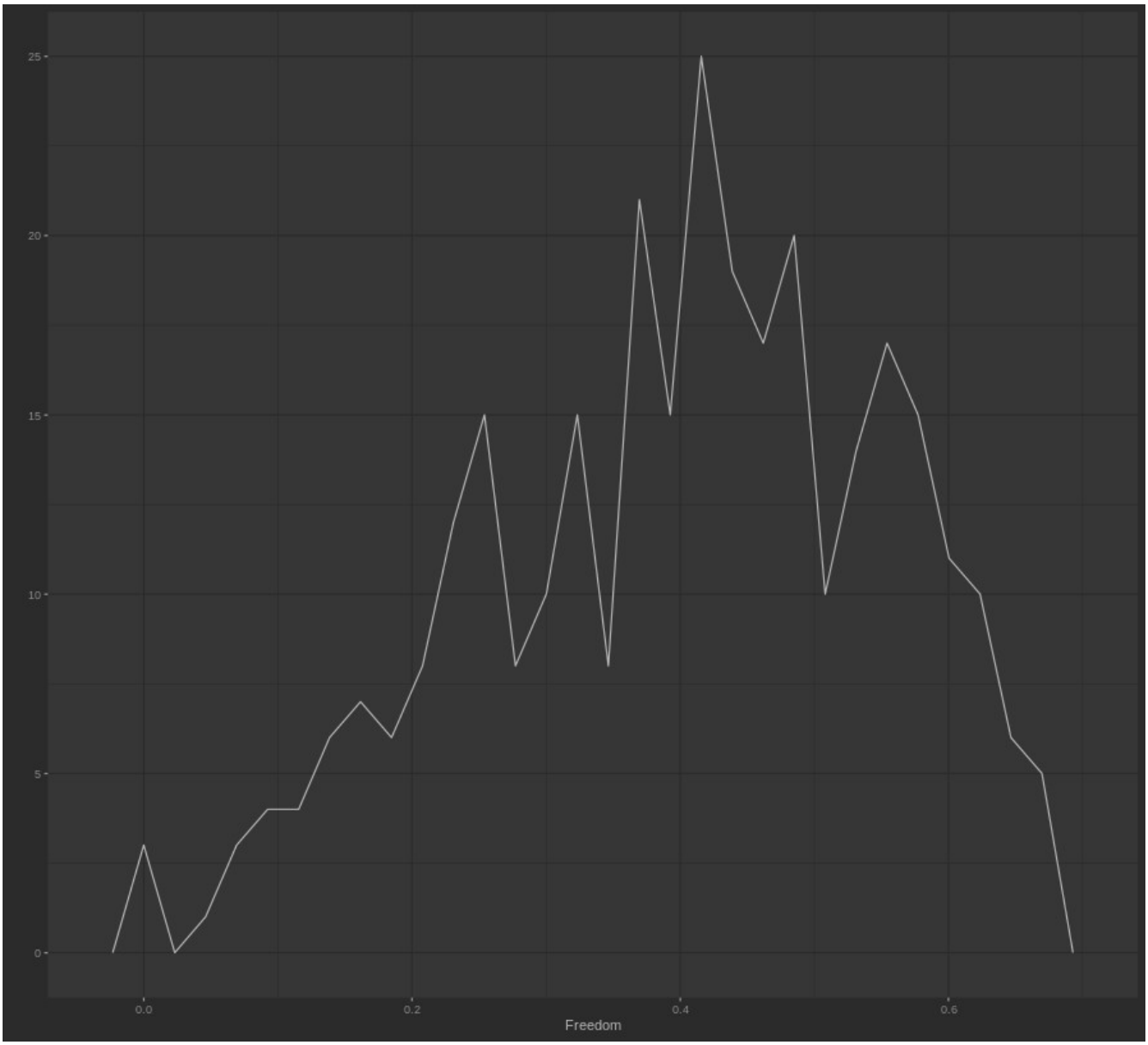



рис. 1 – полігон частот  змінної Freedom

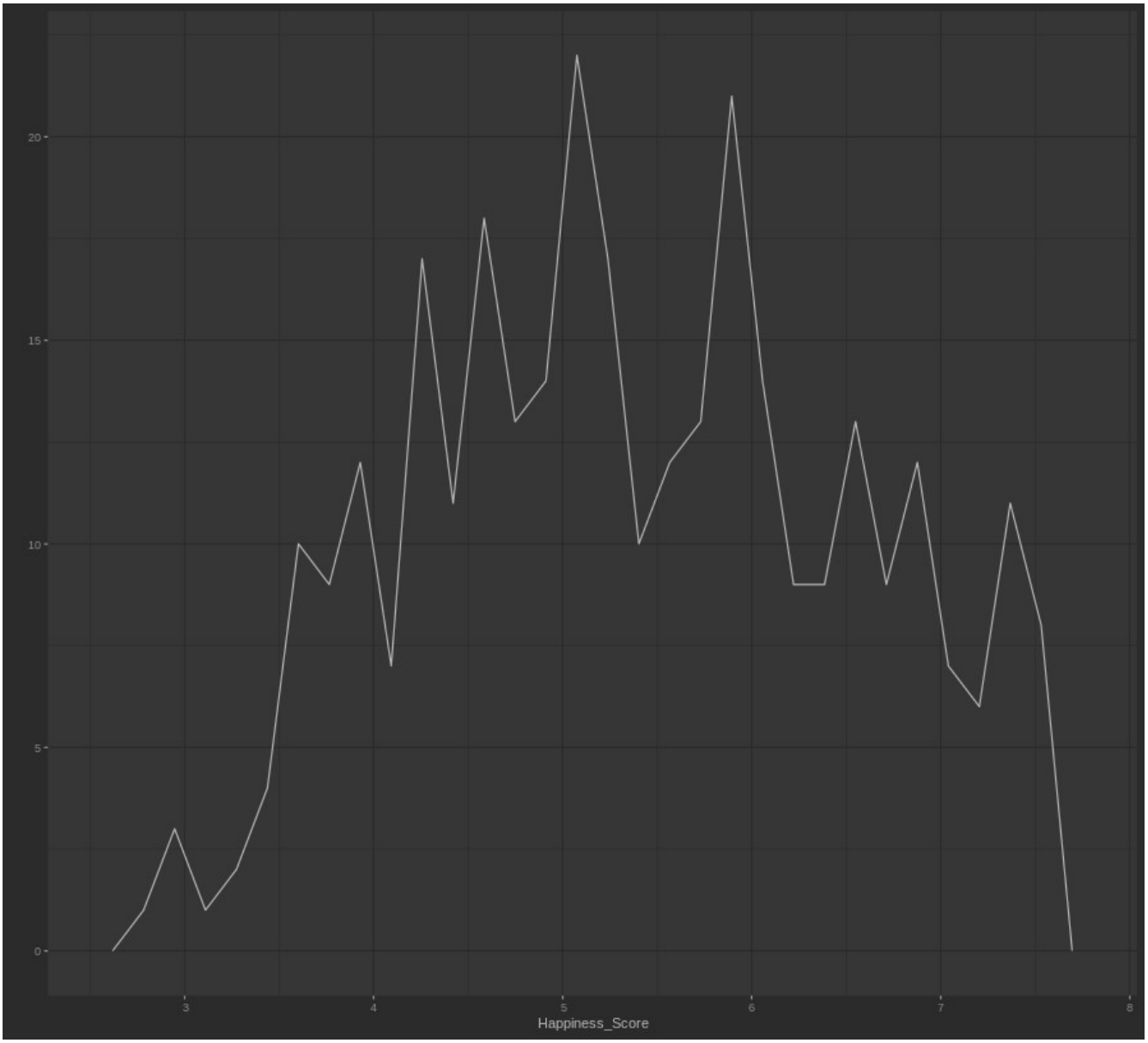



рис. 2 – полігон частот змінної Happiness_Score

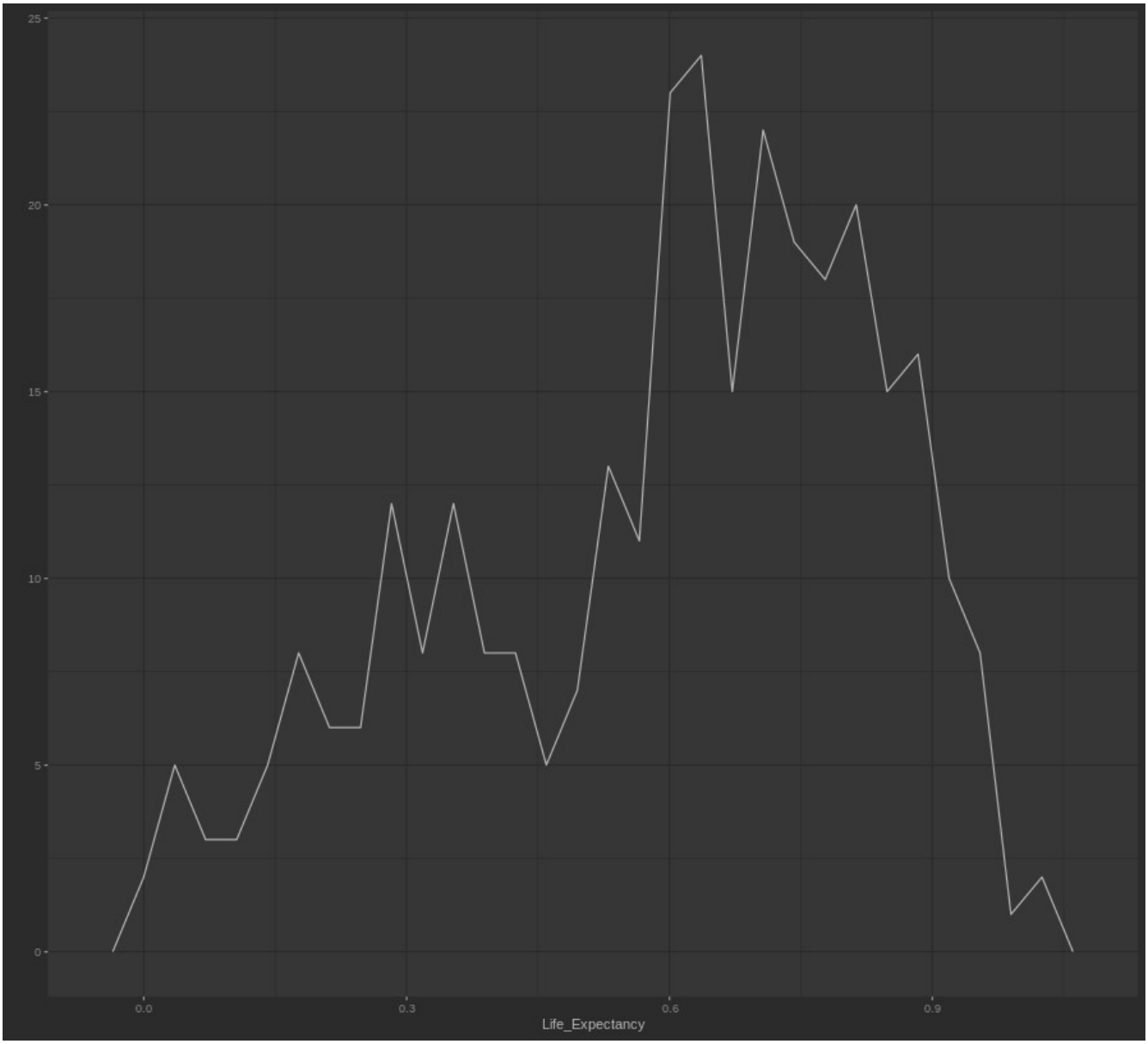



рис. 3 – полігон частот  змінної Life_Expectancy

Знайдемо для кожної змінної найменше значення (min), найбільше (max), квартилі (Qu.), медіану(Median, 2nd Qu.), матсподівання (Mean), децилі (Deciles), геометричне середнє (Geaometric mean), середнє гармонічне (Harmonic mean), моду (Mode), дисперсію (Dispersion), стандартне відхилення (Standard Deviation), коефіцієнт варіації (Coefficient of variation), імовірнісне відхилення (Probabilistic Deviation), розмах вибірки (Sampling span), інтервал концентрації (Concentration

interval), коефіцієнт асиметрії (Skewness) $\beta_1$ і показник ексцесу (Kurtosis) $\beta_2$ для змінних Happiness_Score, Life_Expectancy, Freedom.

Дисперсія, стандартне відхилення, коефіцієнт варіації, імовірнісне відхилення, розмах вибірки та інтервал концентрації є мірами відхилення спостережень випадкової величини від її характеристики положення центра значень. Вони вказують, наскільки суттєво можуть відрізнятися значення випадкової величини від центра зосередження значень.

Якщо $\beta_1<0$ , то розподіл буде скошеним праворуч, якщо $\beta_1>0$ - ліворуч, а для нормальних розподілів $\beta_1=0$.

Якщо $\beta_2>0$ , то розподіл, який досліджується, більш гостроверхий ніж нормальний з відповідними параметрами. Якщо $\beta_2<0$ , то розподіл менш гостроверхий, ніж нормальний з відповідними параметрами. Для нормального розподілу $\beta_2=0$.

```
--------------------------------------------------------------------------------

[1] "Summary of ' Freedom ': "
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.2959  0.4132  0.3999  0.5167  0.6697

[1] "Deciles of ' Freedom ': "
     10%      20%      30%      40%      50%      60%      70%      80%
0.188798 0.257276 0.325052 0.377634 0.413190 0.449224 0.489174 0.543270
     90%
0.590614

[1] "Geometric Mean of ' Freedom ':  0.364284489001505"

[1] "Harmonic Mean of ' Freedom ':  0.274026977578999"

[1] "Mode of ' Freedom ':  0, 0.19847, 0.37938, 0.40672, 0.46074, 0.46582, 0.53466"

[1] "Dispersion of ' Freedom ':  0.0227055212587099"

[1] "Standard Deviation of ' Freedom ':  0.150683513559745"

[1] "Coefficient of Variation of ' Freedom ':  37.6806873307653"

[1] "Probabilistic Deviation of ' Freedom ':  0.11042"

[1] "Sampling Span of ' Freedom ':  0.66973"

[1] "Concentration Interval of ' Freedom ': ( -0.0521546676633627 ,  0.851946413695109 )"

[1] "Kurtosis of ' Freedom ':  -0.53092813513193"

[1] "Skewness of ' Freedom ':  -0.383466859379719"
```

```
[1] "Summary of ' Happiness_Score ': "
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.839   4.510   5.286   5.379   6.269   7.587

[1] "Deciles of ' Happiness_Score ': "
   10%    20%    30%    40%    50%    60%    70%    80%    90%
3.8780 4.3050 4.6594 5.0402 5.2860 5.7474 5.9864 6.4786 6.9706

[1] "Geometric Mean of ' Happiness_Score ':  5.25453375419954"

[1] "Harmonic Mean of ' Happiness_Score ':  5.1269192112216"

[1] "Mode of ' Happiness_Score ':  2.905, 3.739, 3.956, 4.252, 4.876, 5.057, 5.123, 5.129, 5.192, 5.813, 5.835, 5.987, 6.168, 6.269, 6.379, 7.119"

[1] "Dispersion of ' Happiness_Score ':  1.30309273626529"

[1] "Standard Deviation of ' Happiness_Score ':  1.1415308739869"

[1] "Coefficient of Variation of ' Happiness_Score ':  21.2221909929816"

[1] "Probabilistic Deviation of ' Happiness_Score ':  0.8795"

[1] "Sampling Span of ' Happiness_Score ':  4.748"

[1] "Concentration Interval of ' Happiness_Score ': ( 1.95435658438851 ,  8.8035418283099 )"

[1] "Kurtosis of ' Happiness_Score ':  -0.858118964468075"

[1] "Skewness of ' Happiness_Score ':  0.0698944395838097"
```

```
[1] "Summary of ' Life_Expectancy ': "
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.4196  0.6404  0.5941  0.7876  1.0252

[1] "Deciles of ' Life_Expectancy ': "
     10%      20%      30%      40%      50%      60%      70%      80%
0.228032 0.349024 0.500910 0.593882 0.640450 0.697056 0.742212 0.810110
     90%
0.877496

[1] "Geometric Mean of ' Life_Expectancy ':  0.525000574115208"

[1] "Harmonic Mean of ' Life_Expectancy ':  0.399827984302251"

[1] "Mode of ' Life_Expectancy ':  0, 0.62994, 0.92356"

[1] "Dispersion of ' Life_Expectancy ':  0.0579800620706642"

[1] "Standard Deviation of ' Life_Expectancy ':  0.240790494145147"

[1] "Coefficient of Variation of ' Life_Expectancy ':  40.5334028773383"

[1] "Probabilistic Deviation of ' Life_Expectancy ':  0.1839975"

[1] "Sampling Span of ' Life_Expectancy ':  1.02525"

[1] "Concentration Interval of ' Life_Expectancy ': ( -0.128317006244966 ,  1.31642595862592 )"

[1] "Kurtosis of ' Life_Expectancy ':  -0.578302637803278"

[1] "Skewness of ' Life_Expectancy ':  -0.567622803148049"
```

По графіках видно, що змінні мають розподіл, схожий на нормальний. Також бачимо, що при коефіцієнті асиметрії більше нуля розподіли скошені праворуч, при коефіцієнті асиметрії менше нуля – ліворуч.

Також можемо зробити висновок, що тривалість життя у менш щасливих країнах коливається в районі 62-х років, а у більш щасливих – в районі 90-та років.

Тривалість життя 0 означає неможливість оцінити очікувану тривалість життя у країні. Фактично найменша очікувана тривалість життя відповідає першому децилю – 22 роки.

Середній рівень свободи вибору у світі оцінюється у 0.41.

**Код програми**

```
library(ggplot2)
library( package: "psych")
library( package: "DescTools")
library( package: "e1071")

myData <- read.csv( file: 'output.csv')

Happiness_Score <- myData$Happiness_Score
Happiness_Score_Name <- "Happiness_Score"
Life_Expectancy <- myData$Life_Expectancy
Life_Expectancy_Name <- "Life_Expectancy"
Freedom <- myData$Freedom
Freedom_Name <- "Freedom"

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

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

```r
buildFrequencyPolygons <- function() {
  chart1 <- qplot(x = Happiness_Score, geom = 'freqpoly')
  chart2 <- qplot(x = Life_Expectancy, geom = 'freqpoly')
  chart3 <- qplot(x = Freedom, geom = 'freqpoly')
  return(list(
    chart1=chart1,
    chart2=chart2,
    chart3=chart3
  )) ^buildFrequencyPolygons
}

printSummary <- function (vector, name) {
  summary <- summary(vector)
  print(paste("Summary of '", name, "': "))
  print(summary) ^printSummary
}

printDeciles <- function (vector, name) {
  deciles <- quantile(
    vector,
    probs = seq(.1, .9, by = .1)
  )
  print(paste("Deciles of '", name, "': "))
  print(deciles) ^printDeciles
}
```

```r
printGeometricalMeanWithoutZeroes <- function (vector, name) {
  print(
    paste(
      "Geometric Mean of '", name, "': ", exp(mean(log(vector[vector>0])))
    )
  )
}

printHarmonicMeanWithoutZeroes <- function (vector, name) {
  print(
    paste(
      "Harmonic Mean of '", name, "': ", harmonic.mean(vector, zero = FALSE)
    )
  )
}

printMode <- function (vector, name) {
  vectorMode <- Mode(vector)
  print(
    paste(
      "Mode of '", name, "': ", toString(vectorMode)
    )
  ) ^printMode
}
```

```r
printDispersion <- function (vector, name) {
  dispersion <- var(vector)
  print(
    paste(
      "Dispersion of '", name, "': ", dispersion
    )
  ) ^printDispersion
}

printStandardDeviation <- function (vector, name) {
  sd <- sd(vector)
  print(
    paste(
      "Standard Deviation of '", name, "': ", sd
    )
  ) ^printStandardDeviation
}

printCoefficientOfVariation <- function (vector, name) {
  cv <- sd(vector) / mean(vector) * 100
  print(
    paste(
      "Coefficient of Variation of '", name, "': ", cv
    )
  ) ^printCoefficientOfVariation
}
```

```r
printProbabilisticDeviation <- function (vector, name) {
  pd <- IQR(vector) / 2
  print(
    paste(
      "Probabilistic Deviation of '", name, "': ", pd
    )
  ) ^printProbabilisticDeviation
}

printSamplingSpan <- function (vector, name) {
  max <- max(vector)
  min <- min(vector)
  print(
    paste(
      "Sampling Span of '", name, "': ", max - min
    )
  ) ^printSamplingSpan
}

printConcentrationInterval <- function (vector, name) {
  mean <- mean(vector)
  sd <- sd(vector)
  print(
    paste(
      "Concentration Interval of '", name, "': (", mean - 3 * sd, ", ", mean + 3 * sd, ")"
    )
  ) ^printConcentrationInterval
}
```

```r
printKurtosis <- function (vector, name) {
  print(
    paste(
      "Kurtosis of '", name, "': ", kurtosis(vector)
    )
  )
}

printSkewness <- function (vector, name) {
  print(
    paste(
      "Skewness of '", name, "': ", skewness(vector)
    )
  )
}
```

```r
anayzeOneVector <- function (vector, vectorName) {
  printDelimiterWithNewLines()
  printSummary(vector, vectorName)
  printEmptyLine()
  printDeciles(vector, vectorName)
  printEmptyLine()
  printGeometricalMeanWithoutZeroes(vector, vectorName)
  printEmptyLine()
  printHarmonicMeanWithoutZeroes(vector, vectorName)
  printEmptyLine()
  printMode(vector, vectorName)
  printEmptyLine()
  printDispersion(vector, vectorName)
  printEmptyLine()
  printStandardDeviation(vector, vectorName)
  printEmptyLine()
  printCoefficientOfVariation(vector, vectorName)
  printEmptyLine()
  printProbabilisticDeviation(vector, vectorName)
  printEmptyLine()
  printSamplingSpan(vector, vectorName)
  printEmptyLine()
  printConcentrationInterval(vector, vectorName)
  printEmptyLine()
  printKurtosis(vector, vectorName)
  printEmptyLine()
  printSkewness(vector, vectorName)
  printDelimiterWithNewLines() ^anayzeOneVector
}
```

**Істотність статистичного зв'язку**

Для визначення істотності статистичного зв’язку між змінними нам знадобляться наступні 3 величини:

1 Коефіцієнт кореляції
2 Коефіцієнт детермінації
3 Множинний коефіцієнт кореляції

Коефіцієнт кореляції – у математичній статистиці це показник, який характеризує силу статистичного взаємозв'язку двома або більше випадковими змінними. Якщо коефіцієнт кореляції описує взаємозв'язок між двома випадковими змінними, то він називається простим, якщо між однією випадковою змінною і їх групою, то множинним. Коефіцієнт кореляції завжди знаходиться в діапазоні від -1 до 1 і інтерпретується наступним чином:

- Якщо коефіцієнт кореляції близький до 1, то між змінними спостерігається позитивна кореляція. Іншими словами, існує високий ступінь зв'язку між змінними. В цьому випадку, якщо значення змінної x збільшаться, то вихідна змінна також збільшиться;
- Якщо коефіцієнт кореляції близький до -1, це означає, що між змінними існує сильна негативна кореляція. Іншими словами, поведінка вихідної змінної буде протилежною поведінці вхідної змінної. Якщо значення x збільшиться, то у зменшиться, і навпаки;
- Проміжні значення, близькі до 0, будуть вказувати на слабку кореляцію між змінними і, відповідно, на низьку залежність. Іншими словами, поведінка змінної x не вплине (або майже взагалі) на поведінку y (і навпаки).

Коефіцієнт кореляції Пірсона описує тільки ступінь лінійного зв'язку і застосовується до неперервних величин. Для дискретних (якісних) даних використовуються коефіцієнти кореляції рангу Кендалла або Спірмена.

Коефіцієнт детермінації – це доля дисперсії залежної змінної, пояснена розглянутою моделлю залежності, тобто пояснювальними змінними. Точніше, це одиниця мінус частка непоясненої дисперсії (дисперсія випадкової похибки моделі, або умовної за факторами дисперсії залежної змінної) в дисперсії залежної змінної. Вона розглядається як універсальна міра залежності однієї випадкової змінної від багатьох інших. В особливому випадку лінійного зв'язку $R^2$є квадратом коефіцієнта множинної кореляції між залежною змінною та пояснювальними

змінними. Зокрема, для моделі парної лінійної регресійної, коефіцієнт детермінації дорівнює квадрату звичайного коефіцієнта кореляції між y і x.

Коефіцієнт детермінації для моделі з константою приймає значення від 0 до 1. Чим ближче значення коефіцієнта до 1, тим сильніше залежність. При оцінці регресійних моделей це інтерпретується як зіставлення моделі з даними. Для прийнятних моделей передбачається, що коефіцієнт визначення повинен бути не менше 50% (в цьому випадку коефіцієнт кореляції мультиплікатора перевищує за модулем 70%). Досить непоганими можна вважати моделі з коефіцієнтом детермінації вище 80% (коефіцієнт кореляції перевищує 90%). Значення коефіцієнта детермінації 1 означає функціональний зв'язок між змінними.

```
[1] "Coefficient of correlation between  Happiness_Score  and  Life_Expectancy  is:  0.734491375141335"
[1] "Coefficient of correlation between  Happiness_Score  and  Freedom  is:  0.556414442240603"
[1] "Coefficient of correlation between  Life_Expectancy  and  Freedom  is:  0.369798914715241"


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

[1] "P-value of  Happiness_Score  and  Life_Expectancy  is:  1.21734016789266e-54"
[1] "P-value of  Happiness_Score  and  Freedom  is:  5.3084170878437e-27"
[1] "P-value of  Life_Expectancy  and  Freedom  is:  1.21011382637522e-11"


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

[1] "Coefficient of determination for ' Happiness_Score ' and ' Life_Expectancy ':  0.53947758015701"
[1] "Coefficient of determination for ' Happiness_Score ' and ' Freedom ':  0.309597031533921"
[1] "Coefficient of determination for ' Life_Expectancy ' and ' Freedom ':  0.13675123732457"


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

[1] "Multiple correlation coefficient with ' Happiness_Score ' as dependent variable:  0.795888191208726"
[1] "Multiple correlation coefficient with ' Life_Expectancy ' as dependent variable:  0.735980576084885"
[1] "Multiple correlation coefficient with ' Freedom ' as dependent variable:  0.559356741068551"


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

[1] "P-value for ' Happiness_Score ' as dependent variable:  3.47733795736664e-70"
[1] "P-value for ' Life_Expectancy ' as dependent variable:  5.76211134179655e-55"
[1] "P-value for ' Freedom ' as dependent variable:  2.50471825744275e-27"


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

[1] "Coefficient of determination for ' Happiness_Score ' as dependent variable:  0.633438012905498"
[1] "Coefficient of determination for ' Life_Expectancy ' as dependent variable:  0.54166740837424"
[1] "Coefficient of determination for ' Freedom ' as dependent variable:  0.31287996377883"
```

**Аналіз парних статистичних зв'язків**

1 Коефіцієнт кореляції:
  1.1 Happiness_Score and Life_Expectancy is: 0.734491375141335
  1.2 Happiness_Score and Freedom is: 0.556414442240603
  1.3 Life_Expectancy and Freedom is: 0.369798914715241
2 Рівень значущості (p-value):
  2.1 Happiness_Score and Life_Expectancy is: 1.21734016789266e-54
  2.2 Happiness_Score and Freedom is: 5.3084170878437e-27
  2.3 Life_Expectancy and Freedom is: 1.21011382637522e-11
3 Коефіцієнт детермінації
  1.1 Happiness_Score and Life_Expectancy: 0.53947758015701
  1.2 Happiness_Score and Freedom: 0.309597031533921
  1.3 Life_Expectancy and Freedom: 0.13675123732457

Впорядковані пари по рівню значущості:

1 Happiness_Score and Life_Expectancy
2 Happiness_Score and Freedom
3 Life_Expectancy and Freedom

**Висновок:**

По коефіцієнту кореляції бачимо, що сильне відношення є між Happiness_Score and Life_Expectancy, тобто чим нація здоровіша, тим вона щасливіша. Також можна сказати, що між Happiness_Score and Freedom та Life_Expectancy and Freedom спостерігається зв'язок середньої сили. Для усіх пар p-value також дуже мале. P-value показує нам імовірність, з якою ми відкинемо нульову гіпотезу. Чим менше p-value, тим імовірніше ми її відкинемо. Нульова гіпотеза стверджує, що між змінними немає зв'язку. Отже, імовірніше за все, зв'язок між змінними існує.

**Аналіз істотності множинних статистичних зв'язків**

1 Множинний коефіцієнт кореляції
  1.1 ' Happiness_Score ' as dependent variable: 0.795888191208726
  1.2 ' Life_Expectancy ' as dependent variable: 0.735980576084885

1.3' Freedom ' as dependent variable:  0.559356741068551

2 Рівень значущості:
  2.1' Happiness_Score ' as dependent variable:  3.47733795736664e-70
  2.2' Life_Expectancy ' as dependent variable:  5.76211134179655e-55
  2.3' Freedom ' as dependent variable:  2.50471825744275e-27

3 Коефіцієнт детермінації:
  3.1' Happiness_Score ' as dependent variable:  0.633438012905498
  3.2' Life_Expectancy ' as dependent variable:  0.54166740837424
  3.3' Freedom ' as dependent variable:  0.31287996377883

4 Впорядкована послідовність усіх скалярних змінних у порядку спадання істотності множинного  статистичного зв'язку їх з множиною усіх інших
  4.1Happiness_Score
  4.2Life_Expectancy
  4.3Freedom

**Висновок:**

Множинні статистичні зв'язки можливі у всіх випадках, проте найбільш імовірні, коли залежна змінна – Happiness_Score. Більш того, зв'язок між змінними достатньо сильний і є сенс будувати математичну модель.

**Код програми**

```
myData <- read.csv( file: 'output.csv')

Happiness_Score <- myData$Happiness_Score
Happiness_Score_Name <- "Happiness_Score"
Life_Expectancy <- myData$Life_Expectancy
Life_Expectancy_Name <- "Life_Expectancy"
Freedom <- myData$Freedom
Freedom_Name <- "Freedom"

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

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

```r
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
}

printDeterminationCoefficient <- function(vector1, vector2, name1, name2) {
  linModel <- lm(vector1 ~ vector2)
  print(
    paste(
      "Coefficient of determination for '", name1, "' and '", name2, "': ",
      summary(linModel)$r.squared
    )
  ) ^printDeterminationCoefficient
}
```

```r
printMultipleCorrelationCoefficient <- function(
  dependentVectorName,
  independentVectorsNames
) {
  formula <- reformulate(independentVectorsNames, dependentVectorName)
  model <- lm(formula)
  vect <- model$model[, dependentVectorName]
  print(
    paste(
      "Multiple correlation coefficient with '", dependentVectorName, "' as dependent variable: ",
      cor(vect, model$fitted.values)
    )
  ) ^printMultipleCorrelationCoefficient
}

printMultipleDeterminationCoefficient <- function(
  dependentVectorName,
  independentVectorsNames
) {
  formula <- reformulate(independentVectorsNames, dependentVectorName)
  model <- lm(formula)
  print(
    paste(
      "Coefficient of determination for '", dependentVectorName, "' as dependent variable: ",
      summary(model)$r.squared
    )
  ) ^printMultipleDeterminationCoefficient
}
```

```r
printMultiplePValue <- function(
  dependentVectorName,
  independentVectorsNames
) {
  formula <- reformulate(independentVectorsNames, dependentVectorName)
  model <- lm(formula)
  vect <- model$model[, dependentVectorName]

  mulCorTest <- cor.test(vect, model$fitted.values)
  mulPValue <- mulCorTest$p.value
  print(
    paste(
      "P-value for '", dependentVectorName, "' as dependent variable: ",
      mulPValue
    )
  ) ^printMultiplePValue
}
```

```
analyzeCorrelation <- function() {
  printDelimiterWithNewLines()
  printCorellationCoefficient(Happiness_Score, Life_Expectancy, Happiness_Score_Name, Life_Expectancy_Name)
  printCorellationCoefficient(Happiness_Score, Freedom, Happiness_Score_Name, Freedom_Name)
  printCorellationCoefficient(Life_Expectancy, Freedom, Life_Expectancy_Name, Freedom_Name)
  printDelimiterWithNewLines()
  printPValue(Happiness_Score, Life_Expectancy, Happiness_Score_Name, Life_Expectancy_Name)
  printPValue(Happiness_Score, Freedom, Happiness_Score_Name, Freedom_Name)
  printPValue(Life_Expectancy, Freedom, Life_Expectancy_Name, Freedom_Name)
  printDelimiterWithNewLines()
  printDeterminationCoefficient(Happiness_Score, Life_Expectancy, Happiness_Score_Name, Life_Expectancy_Name)
  printDeterminationCoefficient(Happiness_Score, Freedom, Happiness_Score_Name, Freedom_Name)
  printDeterminationCoefficient(Life_Expectancy, Freedom, Life_Expectancy_Name, Freedom_Name)
  printDelimiterWithNewLines()

  printMultipleCorrelationCoefficient(Happiness_Score_Name, c(Life_Expectancy_Name, Freedom_Name))
  printMultipleCorrelationCoefficient(Life_Expectancy_Name, c(Happiness_Score_Name, Freedom_Name))
  printMultipleCorrelationCoefficient(Freedom_Name, c(Happiness_Score_Name, Life_Expectancy_Name))
  printDelimiterWithNewLines()
  printMultiplePValue(Happiness_Score_Name, c(Life_Expectancy_Name, Freedom_Name))
  printMultiplePValue(Life_Expectancy_Name, c(Happiness_Score_Name, Freedom_Name))
  printMultiplePValue(Freedom_Name, c(Happiness_Score_Name, Life_Expectancy_Name))
  printDelimiterWithNewLines()
  printMultipleDeterminationCoefficient(Happiness_Score_Name, c(Life_Expectancy_Name, Freedom_Name))
  printMultipleDeterminationCoefficient(Life_Expectancy_Name, c(Happiness_Score_Name, Freedom_Name))
  printMultipleDeterminationCoefficient(Freedom_Name, c(Happiness_Score_Name, Life_Expectancy_Name))
  printDelimiterWithNewLines() ^analyzeCorrelation
}
```

# Побудова математичної моделі

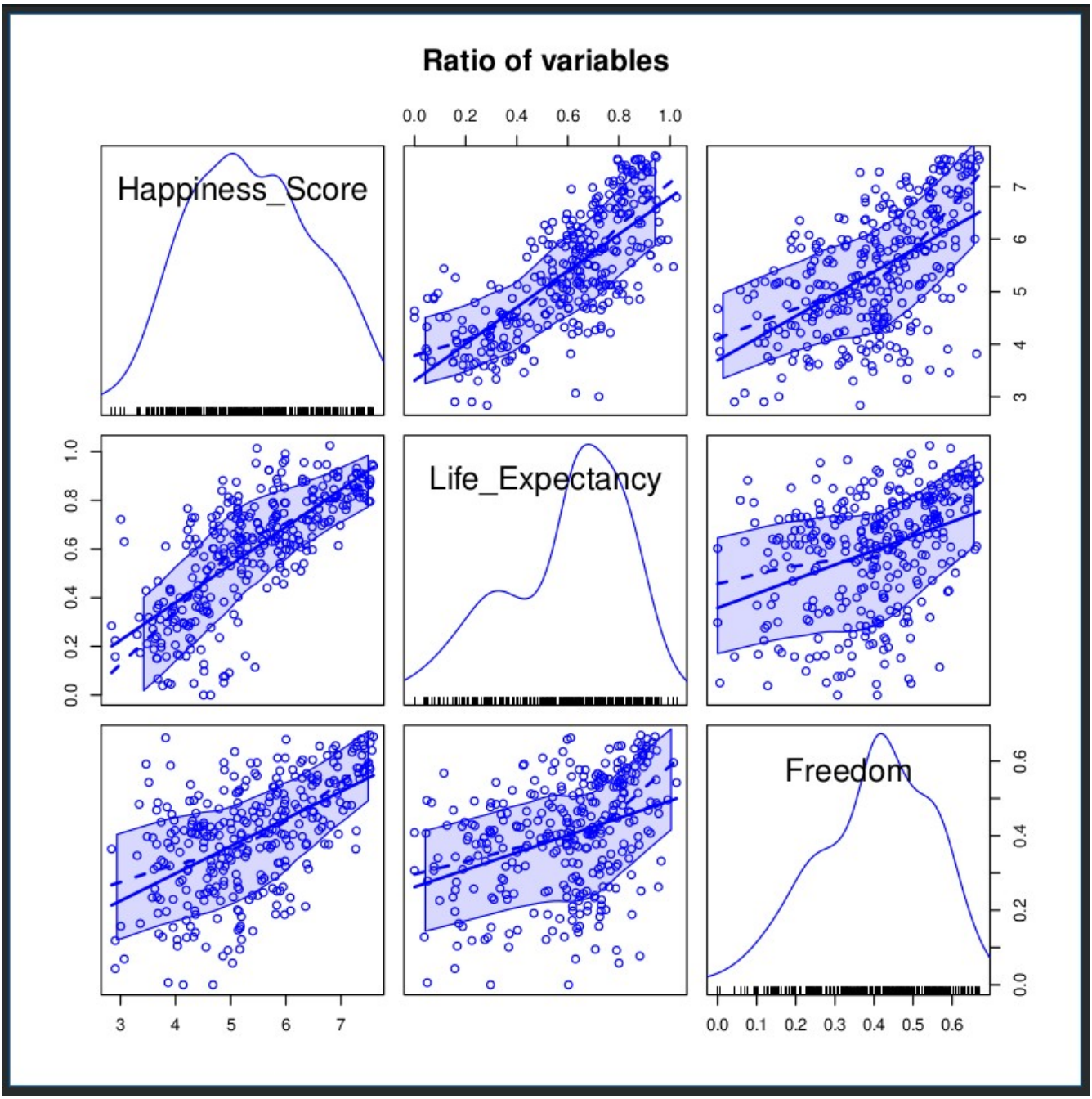
Оскільки ми маємо справу із кількісними скалярними змінними, використаємо Регресійний Аналіз.

Припущення МНК – це важливо. Якщо припущення методу не виконуються, необхідно змінювати метод. Але припущення найчастіше не виконуються на реальних даних. Це відповідальність дослідника - визначити "критичність" припущень. Припущення МНК:

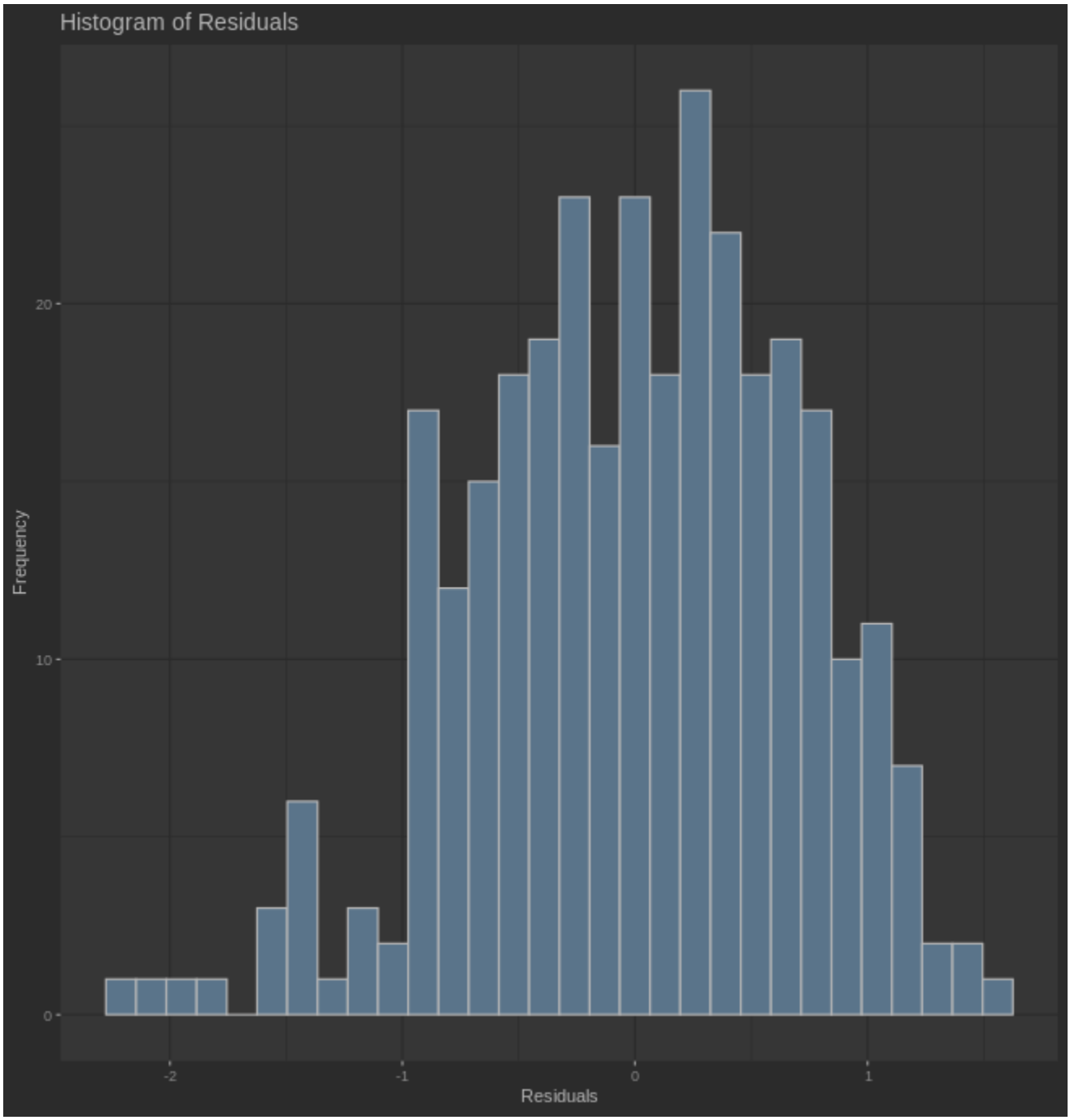
1. Типи змінних на X та Y - кількісні (X, Y), факторні (X)
2. Не-нульова дисперсія
3. Немає мультиколінеарності. Мультиколінеарність у регресійному аналізі виникає, коли два або більше предиктора сильно корелюють один з одним, так що вони не надають унікальної або незалежної інформації в регресійній моделі.
4. Предиктори не корельовані з факторами, що не включені в модель
5. Залишки розподіляються з однаковою дисперсією на кожному рівні предикторів. (Гомоскедастичність)
6. Відсутність автокореляції залишків. Автокореляція - це характеристика даних, яка показує ступінь подібності між значеннями одних і тих самих змінних протягом послідовних інтервалів.
7. Залишки нормально розподілені

Використовуючи лінійну реграсію, побудуємо наступну модель:

```
[1] "Summary for model:"

Call:
lm(formula = formula, data = myDataSlice)

Residuals:
     Min       1Q   Median       3Q      Max
-2.26580 -0.48236  0.03904  0.50484  1.50382

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.6545     0.1274  20.841   <2e-16 ***
Life_Expectancy   2.9037     0.1749  16.602   <2e-16 ***
Freedom           2.4993     0.2795   8.943   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6933 on 312 degrees of freedom
Multiple R-squared:  0.6334,    Adjusted R-squared:  0.6311
F-statistic: 269.6 on 2 and 312 DF,  p-value: < 2.2e-16



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

[1] "Variance Inflation Factor:"
Life_Expectancy         Freedom
       1.158415        1.158415


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

[1] "Breusch-Pagan Test:"

        studentized Breusch-Pagan test

data:  model
BP = 2.0318, df = 2, p-value = 0.3621



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

[1] "Durbin-Watson test:"
 lag Autocorrelation D-W Statistic p-value
   1       0.3966775      1.204073       0
 Alternative hypothesis: rho != 0
```




Відношення між змінними




Гістограма залишків

У нашому випадку, модель має вигляд:

Happiness_Score = 2.6545 + 2.9037 * Life_Expectancy + 2.4993 * Freedom

Виконання умов лінійної регресії:

1. Усі змінні – кількісні
2. Дисперсії не-нульові
3. Мультиколынеарності немає – VIF трохи быльше за 1. Значення 1 вказує на відсутність кореляції між даною змінною-предиктором та будь-якими іншими змінними-предикторами в моделі. Значення від 1 до 5 вказує на помірну кореляцію, але це часто недостатньо серйозне, щоб вимагати уваги. Значення більше 5 вказує на потенційно серйозну кореляцію. У цьому випадку оцінки коефіцієнтів і p-значення в результатах регресії, ймовірно, є ненадійними.
4. Незалежні змінні не корелюють з факторами, які не включені в модель
5. Гомоскедастичність присутня – тест Бреуша-Пагана повернув p-value більше за 0.05
6. Автокореляція присутня – тест Дарбіна-Вотсона дав результат 1.2 і p-value < 0.5. Тож ми можемо відкинути нульову гіпотезу і прийти до висновку, що існує автокореляція між залишками. Вважається, що автокореляції немає, якщо результат тесту лежить у межах 1.5 – 2.5.
7. Розподіл залишків нормальний

**Висновки**

Модель Happiness_Score = 2.6545 + 2.9037 * Life_Expectancy + 2.4993 * Freedom виконує усі умови, які ставить лінійна регресія, окрім автокореляції змінних. Для того, щоб позбутись проблеми автокореляції, нам потрібно було б змінити модель або позбутись змінної Freedom.

Проте я вважаю, що в даному випадку цим можна знехтувати, оскільки Life_Expectancy і Freedom зібрані зі сфер людської діяльності, які насправді не перетинаються. Life_Expectancy – середня тривалість життя, кількість аптек і клінік на душу населення, доступність лікування і звички нації. Freedom – це свобода пересування, вибору місця роботи або навчання, свобода слова. В цілому модель дає результат, більш-менш близький до реального. Наприклад, узявши даніз аналогчного звіту за 2017 рік для Панами (Life_Expectancy = 0.706, Freedom=0.55), за нашою моделлю Happiness_Score=6.07, а реальний показник – 6.45. Також, я вважаю, що для більш точного моделювання треба мати більше даних, хоча б іще на 20 років і провести очистку цих даних.

## Код програми

```r
buildVariableRatiocharts <- function(data) {
  # chart <- ggpairs(dataSlice)
  scatterplotMatrix(
    data, spread = FALSE, lty.smooth = 2, main = 'Variables Ratio'
  )
}

printSummaryForModel <- function(model) {
  print("Summary for model:")
  print(summary(model)) ^printSummaryForModel
}

printVIF <- function(model) {
  print("Variance Inflation Factor:")
  print(vif(model)) ^printVIF
}

printBreuschPaganTestResult <- function(model) {
  print("Breusch-Pagan Test:")
  print(bptest(model)) ^printBreuschPaganTestResult
}

printDurbinWatsonTestResult <- function(model) {
  print("Durbin-Watson test:")
  print(durbinWatsonTest(model)) ^printDurbinWatsonTestResult
}

buildResidualsPlot <- function(model, data) {
  ggplot(data = data, aes(x = model$residuals)) +
    geom_histogram(fill = 'steelblue', color = 'black') +
    labs(title = 'Histogram of Residuals', x = 'Residuals', y = 'Frequency')
}
```

```r
buildModel <- function() {
  myDataSlice <- myData[, c(Happiness_Score_Name, Life_Expectancy_Name, Freedom_Name)]
  buildVariableRatiocharts(myDataSlice)
  formula <- reformulate(c(Life_Expectancy_Name, Freedom_Name), Happiness_Score_Name)
  model <- lm(formula, data = myDataSlice)
  printDelimiterWithNewLines()
  printSummaryForModel(model)
  printDelimiterWithNewLines()
  printVIF(model)
  printDelimiterWithNewLines()
  printBreuschPaganTestResult(model)
  printDelimiterWithNewLines()
  printDurbinWatsonTestResult(model)
  printDelimiterWithNewLines()
  buildResidualsPlot(model, myDataSlice) ^buildModel
}

buildModel()
```

# Список використаних джерел


1 Датасет: https://www.kaggle.com/datasets/shivkumarganesh/world-happiness-report-20152022
2 Аналіз Даних, 2001, Слабоспицький
3 https://uk.economy-pedia.com/11041138-difference-between-qualitative-and-quantitative-variable
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
8 https://www.tutorialspoint.com/r/r_multiple_regression.htm
9 https://statisticsbyjim.com/regression/interpret-coefficients-p-values-regression/#:~:text=Regression%20analysis%20is%20a%20form,correlation%20with%20the%20dependent%20variable
10 https://rpubs.com/smarcel/106230
11 https://habr.com/ru/post/350668/
12 https://ru.wikipedia.org/wiki/%D0%A0%D0%B5%D0%B3%D1%80%D0%B5%D1%81%D1%81%D0%B8%D0%BE%D0%BD%D0%BD%D1%8B%D0%B9_%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7
13 https://thedatamonk.com/p-value-in-simple-terms/