Statystyka

ZSTA LIO

Łukasz Smaga

Spis treści

Podstawowe informacje

Kontakt

- Prof. UAM dr hab. Łukasz Smaga
  - Zakład Statystyki Matematycznej i Analizy Danych, Wydział Matematyki i Informatyki, Uniwersytet im. Adama Mickiewicza w Poznaniu
  - Pokój: B4-8, ul. Uniwersytetu Poznańskiego 4, Poznań
  - E-mail: ls@amu.edu.pl
  - Tel.: 61 829-5336
  - Strona internetowa: ls.home.amu.edu.pl
  - Dyżury: aktualne dyżury podane są na powyższej stronie internetowej

Zasady zaliczenia

- Egzamin będzie obejmował całość materiału omawianego na wykładach i laboratoriach. Odbędzie się on na ostatnich laboratoriach. Zadania egzaminacyjne będą dotyczyły:
  - (głównie) analizy statystycznej pewnych zagadnień praktycznych z wykorzystaniem programu R i dostępnych danych,
  - podania interpretacji, opisu, itd. wybranych metod statystycznych.
- Ocena końcowa z egzaminu będzie również oceną z laboratoriów.
- Warunkiem koniecznym zaliczenia laboratoriów jest obecność na zajęciach, tj. dopuszczalne są co najwyżej dwie nieusprawiedliwione nieobecności na laboratoriach (nie dotyczy to laboratoriów, na których odbywa się egzamin).
- Egzamin poprawkowy odbędzie się poza zajęciami w podobnej formie.

Plan wykładu

1. Podstawy programu R
2. Statystyka opisowa
3. Model statystyczny
4. Estymacja
5. Weryfikacja hipotez statystycznych
6. Analiza regresji
7. Metody wielowymiarowe

Literatura

1. Biecek P. (2008) Przewodnik po pakiecie R. GIS.
2. Biecek P. (2011) Analiza danych z programem R. Modele liniowe z efektami stałymi, losowymi i mieszanymi. Wydawnictwo Naukowe PWN.
3. Gągolewski M. (2014) Programowanie w języku R. Analiza danych, obliczenia, symulacje. Wydawnictwo Naukowe PWN.
4. Górecki T. (2011) Podstawy statystyki z przykładami w R. BTC.
5. Komsta Ł., Wprowadzenie do środowiska R http://www.r-project.org.

1 Wprowadzenie do programu R

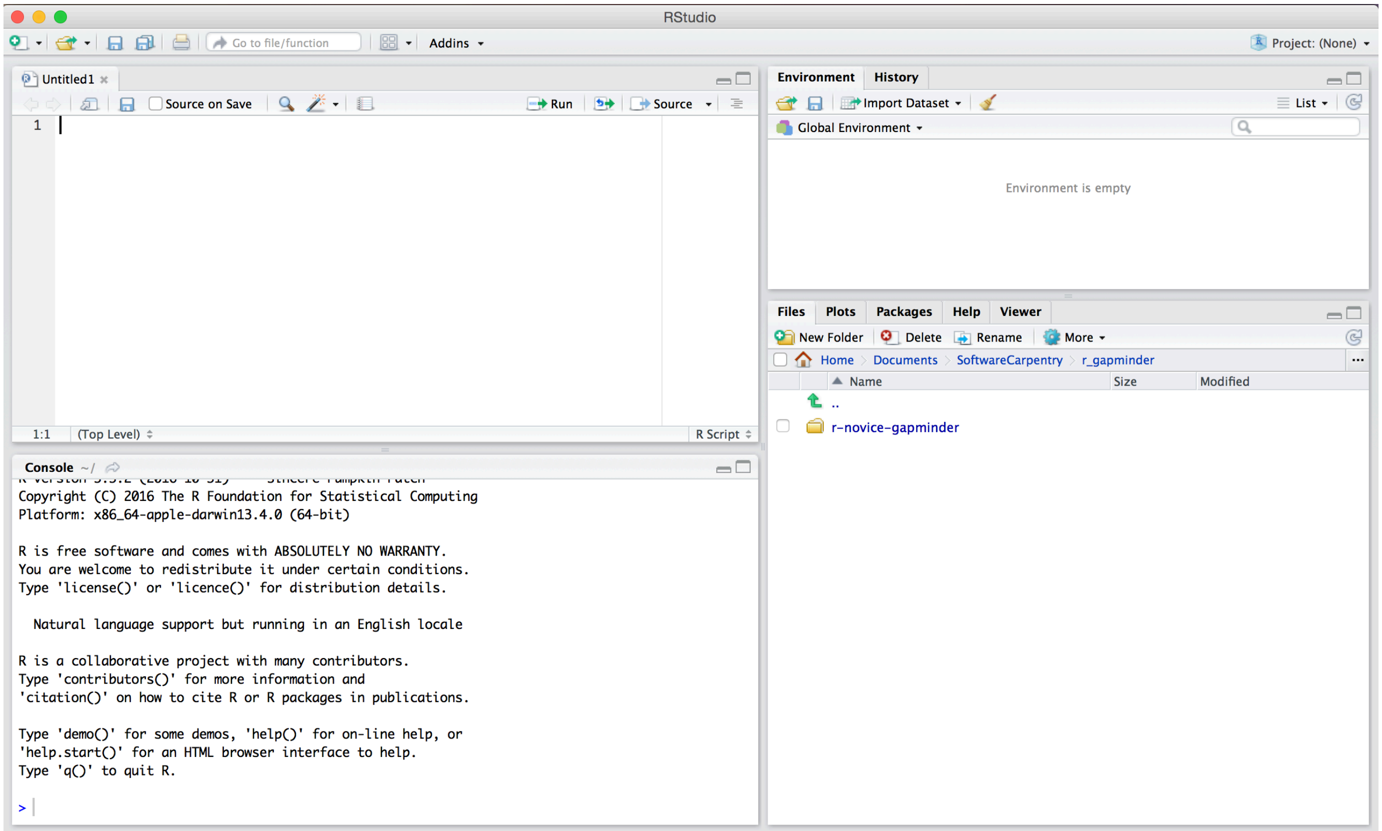
- R jest zaawansowanym pakietem statystycznym jak również językiem programowania istniejącym na platformy Windows, Unix oraz MacOS.
- R jest wolnym (otwartym i darmowym) środowiskiem.
- Język R jest językiem interpretowanym, a nie kompilowanym.
- O sile R stanowi ponad 15000 bibliotek (pakietów), przeznaczonych do najróżniejszych zastosowań.

1.1 RStudio

Okno RStudio składa się z czterech części:

1. edytora kodu źródłowego otwartych plików/skryptów i podglądu własności obiektów (okienko po lewej u góry),
2. konsoli R i terminalu (okienko po lewej u doły),
3. listy zadeklarowanych obiektów i historii poleceń (okienko po prawej u góry),

4. prostego menadżera plików, podglądu rysunków, wykazu dostępnych pakietów R i przeglądarki dokumentacji (okienko po prawej u dołu).



- `CTRL+SHIFT+n` - tworzy nowy plik źródłowy
- `CTRL+ENTER` - przekazuje kod z edytora do konsoli R
- `CTRL+1` i `CTRL+2` - przenoszą karetkę między edytorem a konsolą
- `CTRL+F11` i `CTRL+F12` - przenoszą karetkę między otwartymi skryptami
- `#` - komentarz

1.2 System pomocy

```
?mean
help(mean)
```

1.3 Pakiety

- Pakiet to zestaw narzędzi, takich jak nowe funkcje wraz z dokumentacją oraz nowe zbiory danych, rozszerzających funkcjonalność programu R. Większość z nich znajduje się w repozytorium CRAN (*Comprehensive R Archive Network*).
- `install.packages(nazwa pakietu, dependencies = TRUE)` - instalacja pakietu
- `library(nazwa pakietu)` - ładowanie pakietu
- `detach(package:nazwa pakietu)` - usunięcie pakietu

```
install.packages("car")
library(car)
detach(package:car)
```

1.4 Wektory atomowe

1.4.1 Wektory wartości logicznych

- W R zdefiniowane są dwie stałe logiczne:
  - `TRUE` - prawda,
  - `FALSE` - fałsz.

```
FALSE
```

```
## [1] FALSE
```

- W programie R wielkość liter ma znaczenie (ang. case-sensitive).

```
true
## Error: object 'true' not found
```

- Wektory można tworzyć przez złączanie. Wektor (ciąg) składający się z konkretnych wartości logicznych w określonej kolejności, można utworzyć za pomocą funkcji `c()` (od ang. **c**ombine - złącz).

```
c(TRUE, TRUE, FALSE, FALSE, TRUE)
```

```
## [1]  TRUE  TRUE FALSE FALSE  TRUE
```

```
c(c(TRUE, TRUE, FALSE), c(FALSE, TRUE))
```

```
## [1]  TRUE  TRUE FALSE FALSE  TRUE
```

- Długość wektora zwraca funkcja `length()`.

```
length(c(TRUE, TRUE, FALSE, FALSE, TRUE))
```

```
## [1] 5
```

1.4.2 Wektory liczbowe i zespolone

```
c(1, +2, -3, 2.3, -.4, 5.)
```

```
## [1]  1.0  2.0 -3.0  2.3 -0.4  5.0
```

- Do generowania ciągów arytmetycznych w R służą:
  - operator `:` (różnica równa się 1 lub −1),
  - funkcja `seq()` (od ang. **seq**uence, dowolne różnice).

```
c(-3:2, 4:0)
```

```
##  [1] -3 -2 -1  0  1  2  4  3  2  1  0
```

```
seq(1, 8, by = 2)
```

```
## [1] 1 3 5 7
```

```
seq(1, 8, length.out = 6)
```

```
## [1] 1.0 2.4 3.8 5.2 6.6 8.0
```

1.4.3 Wektory napisów

- Ciągi dowolnych znaków drukowanych, zwane napisami, tworzymy wykorzystując apostrofy lub cudzysłów.

```
c("ZSTA LIO", "informatyka", ",", "statystyka", "!")
```

```
## [1] "ZSTA LIO"    "informatyka" ","           "statystyka"  "!"
```

```
length(c("ZSTA LIO", "informatyka", ",", "statystyka", "!"))
```

```
## [1] 5
```

1.4.4 Nazywanie obiektów

- W R obiekty nazywamy za pomocą jednego z następujących operatorów przypisania (ang. assignment operator):
  - `=`
  - `<-` (w RStudio skrót klawiszowy `ALT+-`)
  - `->`

```
x = 5
5 = x
## Error in 5 = x : invalid (do_set) left-hand side to assignment
y <- 6
6 -> y
x
## [1] 5
y
## [1] 6
```

- Wielu użytkowników programu R nie zaleca stosowania operatora `=`, ponieważ ma on również inne znaczenia, np. używa się go do ustalania wartości funkcji.
- Lepiej nie używać (poza ewentualnie komentarzami) polskich znaków diakrytycznych.
- W R nie trzeba deklarować obiektów (choć można), wystarczy wykorzystać operator przypisania.
- Nazwa obiektu nie jest do niego przypisana na zawsze. Można do niej przypisać nową wartość.

```
(x <- 1)
```

```
## [1] 1
```

```
(x <- 2)
```

```
## [1] 2
```

- Polecenie `ls()` podaje wszystkie aktualnie istniejące obiekty.
- Usunąć jakiś obiekt możemy za pomocą funkcji `rm()`.
- Wszystkie obiekty usuwamy poleceniem `rm(list = ls())`.

```
x <- 1:2
y <- list(1, 2)
ls()
```

```
## [1] "x" "y"
```

```
rm(x)
ls()
```

```
## [1] "y"
```

```r
rm(list = ls())
ls()
```

```
## character(0)
```

1.4.5 Operatory arytmetyczne

- Do działania na wektorach liczbowych (czasem również zespolonych) można używać następujących binarnych operatorów arytmetycznych:
  - `+` (dodawanie),
  - `-` (odejmowanie),
  - `*` (mnożenie),
  - `/` (dzielenie rzeczywiste),
  - `^` (potęgowanie),
  - `%%` (reszta z dzielenie (modulo)),
  - `%/%` (dzielenie całkowite (bez reszty)).
- Operatory arytmetyczne są zwektoryzowane (ang. vectorized), tzn. dla wektorów $\mathbf{x} = (x_1, x_2, \dots, x_n)$ i $\mathbf{y} = (y_1, y_2, \dots, y_n)$ o tej samej długości n w wyniku działania $\mathbf{x}\diamondsuit\mathbf{y}$ uzyskujemy wektor

$$\mathbf{w} = (x_1\diamondsuit y_1, x_2\diamondsuit y_2, \dots, x_n\diamondsuit y_n).$$

  Czyli operacje tego typu wykonywane są element po elemencie (ang. elementwise). Unikamy w tej sposób „jawnej” pętli (pętla jest „ukryta” w kodzie operatora), co może pozwolić na przyśpieszenie obliczeń.

```r
7 %% 3
```

```
## [1] 1
```

```r
1:3 + c(3, 4, 5)
```

```
## [1] 4 6 8
```

- W przypadku, gdy wektory będące argumentami operatorów binarnych są różnej długości, stosowana jest tak zwana reguła zawijania (ang. recycling rule). Powiela ona niejako krótszy wektor tak, aby uzgodnić jego długość dłuższym wektorem. Niech $\mathbf{x} = (x_1, x_2, \dots, x_n)$ i $\mathbf{y} = (y_1, y_2, \dots, y_m)$, gdzie bez straty ogólności $m \geqslant n$. Wtedy wynikiem działania jest m-elementowy wektor postaci (dla odpowiedniego l)

$$\mathbf{x}\diamondsuit\mathbf{y} = (x_1\diamondsuit y_1, \dots, x_n\diamondsuit y_n, x_1\diamondsuit y_{n+1}, x_2\diamondsuit y_{n+2}, \dots, x_l\diamondsuit y_m).$$

```r
x <- c(1, 3, 5, 8, 1, 3, 0, 6)
x * c(1, 3)
```

```
## [1]  1  9  5 24  1  9  0 18
```

```r
x <- c(1, 3, 5, 8, 1, 3, 0)
x * c(1, 3)
```

```
## Warning in x * c(1, 3): długość dłuszego obiektu nie jest wielokrotnością
## długości krótszego obiektu
```

```
## [1]  1  9  5 24  1  9  0
```

1.4.6 Operatory logiczne i relacyjne

- Rozważamy następujące operatory i funkcje logiczne:
  - `!x` (negacja),
  - `x | y` (alternatywa),
  - `x & y` (koniunkcja).
- Do porównywania wektorów służą następujące operatory relacyjne:
  - `x < y` (czy mniejsze?),
  - `x > y` (czy większe?),
  - `x <= y` (czy nie większy?),
  - `x >= y` (czy nie mniejszy?),
  - `x == y` (czy równy?),
  - `x != y` (czy nierówny?).
- Można je stosować na wektorach dowolnych typów. Jednak wynikiem ich działania jest zawsze wektor logiczny.

```
(1:7) == (7:1)
```

```
## [1] FALSE FALSE FALSE  TRUE FALSE FALSE FALSE
```

```
c(TRUE, FALSE) < 1
```

```
## [1] FALSE  TRUE
```

1.5 Indeksowanie wektorów

- Wartości elementów każdego wektora leżą na ściśle określonych pozycjach oznaczonych kolejnymi liczbami naturalnymi (`1:length(x)`).
- Do elementów wektora odwołujemy się poprzez nawiasy kwadratowe `[]`.

```
x <- 1:5
x[2]
```

```
## [1] 2
```

```
x[2:4]
```

```
## [1] 2 3 4
```

```
x[-2]
```

```
## [1] 1 3 4 5
```

```
x[-(2:4)]
```

```
## [1] 1 5
```

```
# x[c(1, -2)]
## Error in x[c(1, -2)] : only 0's may be mixed with negative subscripts
x[c(TRUE, TRUE, FALSE, TRUE, FALSE)]
```

```
## [1] 1 2 4
```

```
x[x < 4]
```

```
## [1] 1 2 3
```

- Nawiasów kwadratowych możemy również użyć do zmiany elementów danego wektora.

```r
x[2] <- 6
x
```

```
## [1] 1 6 3 4 5
```

```r
x[c(2, 4)] <- c(4, 2)
x
```

```
## [1] 1 4 3 2 5
```

```r
x[c(2, 4)] <- 6
x
```

```
## [1] 1 6 3 6 5
```

1.5.1 Listy

- Kolejnym podstawowym typem danych jest lista. Najlepiej postrzegać ją jako ciąg złożony z elementów o dowolnych typach (a więc już niekoniecznie tych samych jak w przypadku wektorów atomowych). W skład listy mogą wchodzić wektory logiczne, liczbowe i napisów, a nawet funkcje, czy też same listy.
- Listy tworzymy zazwyczaj za pomocą funkcji `list()`.

```r
(x <- list(TRUE, 3.5, "ZSTA"))
```

```
## [[1]]
## [1] TRUE
##
## [[2]]
## [1] 3.5
##
## [[3]]
## [1] "ZSTA"
```

```r
(x <- list(logiczna = TRUE, liczba = 3.5, napis = "ZSTA"))
```

```
## $logiczna
## [1] TRUE
##
## $liczba
## [1] 3.5
##
## $napis
## [1] "ZSTA"
```

```r
x[[1]]
```

```
## [1] TRUE
```

```r
x$logiczna
```

```
## [1] TRUE
```

1.6 Ramki danych

- Ramki danych (ang. data frames) to obiekty przechowujące informacje w postaci macierzowej, najczęściej takie, które są np. wynikiem eksperymentów (także numerycznych). Wiersze ramki danych odpowiadają reprezentowanym obiektom, tzw. obserwacjom (ang. observations), bądź przypadkom (ang. cases), np. badanym osobom. Kolumny z kolei podają informacje na temat wartości różnych zmiennych (ang. variables) opisujących ich wybrane własności (mierzalne lub nie).
- W R, ramki danych są reprezentowane przez listy zawierające wektory atomowe o tej samej długości. Każdy element tej szczególnej listy odpowiada kolumnie ramki danych.

```
ramka <- data.frame(
  plec = c("K", "K", "M", "M", "K"),
  wyksztalcenie = c("s", "w", "w", "p", "s"),
  waga = c(60, 55, 80, 75, 62)
)
ramka
```

```
##   plec wyksztalcenie waga
## 1    K             s   60
## 2    K             w   55
## 3    M             w   80
## 4    M             p   75
## 5    K             s   62
```

```
nrow(ramka)
```

```
## [1] 5
```

```
ncol(ramka)
```

```
## [1] 3
```

```
rownames(ramka)
```

```
## [1] "1" "2" "3" "4" "5"
```

```
colnames(ramka)
```

```
## [1] "plec"          "wyksztalcenie" "waga"
```

```
ramka[[3]] # lub ramka$waga lub ramka[, 3]
```

```
## [1] 60 55 80 75 62
```

```
ramka$waga <- c(58, 54, 78, 72, 60)
ramka[ramka$plec == "M", ]
```

```
##   plec wyksztalcenie waga
## 3    M             w   78
## 4    M             p   72
```

```
rbind(ramka[1:2, ], ramka[1:2, ])
```

```
##   plec wyksztalcenie waga
## 1    K             s   58
## 2    K             w   54
## 3    K             s   58
## 4    K             w   54
```

```r
cbind(ramka[1:2, ], wyksztalcenie_2 = as.integer(ramka$wyksztalcenie[1:2]))
```

```
## Warning in data.frame(..., check.names = FALSE): pojawiły się wartości NA na
## skutek przekształcenia
```

```
##   plec wyksztalcenie waga wyksztalcenie_2
## 1    K             s   58              NA
## 2    K             w   54              NA
```

1.7 Odczytywanie i zapisywanie danych

- `read.table()`, `load()`, `read.csv()`, `read.csv2()` - wczytanie zbioru danych, odpowiednio z pliku tekstowego, pliku w formacie programu R (z rozszerzeniem `RData`), pliku `csv`, odpowiednio
- `write.table()`, `save()`, `write.csv()`, `write.csv2()` - zapis zbioru danych, odpowiednio do pliku tekstowego, pliku w formacie programu R (z rozszerzeniem `RData`), plików `csv`, odpowiednio
- Przy odczytywaniu i zapisywaniu danych, wygodnie jest najpierw ustalić katalog bieżący na ten, w którym znajdują się lub mają znaleźć się pliki z danymi. Aktualny katalog bieżący sprawdzamy za pomocą funkcji `getwd()`, natomiast zmieniamy go używając funkcji `setwd()`.

```r
getwd()
```

```
## [1] "/home/ls/MEGA/DYDAKTYKA/STA/ZSTA_LIO/ZSTA_LIO_bookdown"
```

```r
# setwd("/home/ls/MEGA/DYDAKTYKA/STA/ZSTA_LIO")
# (odczyt_1 <- read.table("odczyt_1.txt"))
(odczyt_1 <- read.table("http://ls.home.amu.edu.pl/data_sets/odczyt_1.txt"))
```

```
##         V1       V2       V3
## 1 zmienna1 zmienna2 zmienna3
## 2      1.2      1.3      1.4
## 3      2.1      2.2      2.3
## 4      3.1      3.2      3.3
```

```r
(odczyt_1 <- read.table("http://ls.home.amu.edu.pl/data_sets/odczyt_1.txt",
                        header = TRUE))
```

```
##   zmienna1 zmienna2 zmienna3
## 1      1.2      1.3      1.4
## 2      2.1      2.2      2.3
## 3      3.1      3.2      3.3
```

```r
(odczyt_2 <- read.table("http://ls.home.amu.edu.pl/data_sets/odczyt_2.txt",
                        header = TRUE))
```

```
##   zmienna1.zmienna2.zmienna3
## 1                1,2;1,3;1,4
## 2                2,1;2,2;2,3
## 3                3,1;3,2;3,3
```

```r
(odczyt_2 <- read.table("http://ls.home.amu.edu.pl/data_sets/odczyt_2.txt",
                        header = TRUE, sep = ";", dec = ","))
```

```
##   zmienna1 zmienna2 zmienna3
## 1      1.2      1.3      1.4
## 2      2.1      2.2      2.3
```

```
## 3      3.1      3.2      3.3
```

- Pliki z danymi do powyższych przykładów: odczyt_1.txt, odczyt_2.txt
- Można też zaimportować dane klikając na `Import Dataset` w RStudio i w otworzonym okienku ustawić potrzebne parametry.
- Podgląd danych w edytorze kodu źródłowego otrzymujemy za pomocą funkcji `View()`.
- Zapisywanie danych:

```
dane_1 <- data.frame(1:10, 5:14)
write.table(dane_1, "dane_1.txt")
save(dane_1, file = "dane_1.RData")
dane_1 <- read.table("dane_1.txt")
load("dane_1.RData")
```

1.8 Funkcje

- Korzystając z programu R, bardzo szybko odczuwa się potrzebę użycia pewnych fragmentów kodu wielokrotnie, choć być może dla różnych danych.
- Tak jak listy grupują obiekty (być może różnych typów), tak funkcje zbierają określone wyrażenia służące np. do obliczenia pewnych wartości dla zadanych danych.
- Dodatkową zaletą stosowania funkcji jest możliwość dzielenia długiego kodu na łatwiejsze do opanowania części.
- Tworzenie obiektów typu funkcja odbywa się według następującej składni

```
function(lista parametrów) ciało funkcji
```

gdzie `ciało funkcji` jest wyrażeniem do wykonania na obiektach określonych przez `listę parametrów`.

- Wartość obliczonego wyrażenia jest wynikiem działania funkcji. Takim wynikiem może być jeden i tylko jeden obiekt, np. lista.
- Parametrów może być jednak wiele. `lista parametrów` to ciąg oddzielonych przecinkami elementów postaci:
  - `nazwa` parametru (pod taką nazwą będzie dostępny w funkcji obiekt przekazany przy wywołaniu),
  - `nazwa = wyrażenie` (parametr z wartością domyślną),
  - `...` - parametr specjalny, który pozwala przekazać dowolną liczbę argumentów w grupie.

```
szescian <- function(x) x^3 # funkcje zazwyczaj się nazywa
szescian(2)
```

```
## [1] 8
```

```
szescian_2 <- function(x, y) {
  x3 <- x^3
  y3 <- y^3
  return(c(x3, y3))
}
szescian_2(2, 3) # lub szescian_2(x = 2, y = 3)
```

```
## [1]  8 27
```

```
szescian_3 <- function(x = 2, y = 2) {
  x3 <- x^3
```

```
  y3 <- y^3
  return(c(x3, y3))
}
szescian_3()
```

```
## [1] 8 8
```

```
szescian_3(y = 3)
```

```
## [1]  8 27
```

```
str(lapply(list(1, 2, 3), function(x) x^3))
```

```
## List of 3
##  $ : num 1
##  $ : num 8
##  $ : num 27
```

```
szescian_4 <- function(x) {
    if (!is.numeric(x)) {
      stop("non-numeric argument x")
    }
    x^3
}
szescian_4(-3)
## [1] -27
szescian_4("a")
## Error in szescian_4("a") : non-numeric argument x
```

1.9 Instrukcje warunkowe

- Wyrażenie warunkowe `if` ma następującą składnię:

```
if (warunek) wyrazenieTRUE else wyrazenieFALSE
```

- Przykładowo:

```
if (is.numeric("wyrazenie")) {
  print("wyrazenieTRUE")
} else {
  print("wyrazenieFALSE")
}
```

```
## [1] "wyrazenieFALSE"
```

1.10 Pętle

- Pętle umożliwiają wielokrotne wykonywanie tego samego wyrażenia (choć zapewne na różnych obiektach). W programie R mamy do dyspozycji pętle:
  - `while`
  - `repeat`
  - `for`
- Składnia pętli `while` jest następująca:

```
while (warunek) wyrazenie
```

- Zadaniem pętli `while` jest obliczanie `wyrazenia` dopóty, dopóki `warunek` jest spełniony.

```
i <- 1
while (i <= 3) {
  print(i)
  i <- i + 1
}
```

```
## [1] 1
## [1] 2
## [1] 3
```

- Aby pętla nie wykonywała się nieskończoną liczbę razy, zazwyczaj `warunek` będzie konstruowany na danych odczytywanych z pewnego obiektu, który jest modyfikowany za pomocą `wyrazenia`.
- Może się zdarzyć, że `warunek` testowy nigdy nie będzie spełniony i wtedy liczba wykonanych obrotów pętli będzie równa zeru.
- Pętla `repeat` zachowuje się tak jak `while` z `warunkiem` testowym na stałe ustawionym na `TRUE`. Zatem należy zawsze pamiętać o wywołaniu `break`, o ile chcemy doczekać wyniku.

```
i <- 0
repeat {
  i <- i + 1
  print(i)
  if (i == 3) break
}
```

```
## [1] 1
## [1] 2
## [1] 3
```

- Pętla `for` jest chyba najczęściej stosowaną pętlą w programie R. Szczególnie nadaje się ona do „przechodzenia" po elementach wektora atomowego lub listy bądź też wykonywania ciągu wyrażeń zadaną liczbę razy. Jej składnia jest następująca:

```
for (nazwa in wektor) wyrazenie
```

- W pętli `for` każdą kolejną (od pierwszej do ostatniej) wartość `wektora` związujemy z podaną `nazwą` i obliczamy `wyrażenie`. Pętla ta wykonuje się zawsze dokładnie `length(wektor)` razy, o ile nie użyte zostało wyrażenie `break`.

```
for (i in 1:3) print(i)
```

```
## [1] 1
## [1] 2
## [1] 3
```

1.11 Zadania 1

Zadanie 1. Otwórz program RStudio. Następnie utwórz nowy skrypt i zapisz go jako, na przykład, `wprowadzenie_do_R_zadania.R`. W tym skrypcie możesz napisać rozwiązania następujących zadań.

Zadanie 2. Użyj funkcji `rep()`, aby utworzyć wektor logiczny, zaczynając od trzech wartości prawda, następnie czterech wartości fałsz, po których następują dwie wartości prawda i wreszcie pięć wartości fałsz. Przypisz ten wektor logiczny do zmiennej `x`. Na koniec przekonwertuj ten wektor na wektor numeryczny. Jak zmieniły się wartości prawda i fałsz?

```
##  [1]  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE
## [13] FALSE FALSE
```

```
##  [1] 1 1 1 0 0 0 0 1 1 0 0 0 0 0
```

Zadanie 3. Palindromem nazywamy wektor, którego elementy czytane od końca tworzą ten sam wektor co elementy czytane od początku. Utwórz taki wektor 100 liczb przy czym pierwsze 20 liczb to kolejne liczby naturalne, następnie występuje 10 zer, następnie 20 kolejnych liczb parzystych, a pozostałe elementy określone są przez palindromiczność (warunek symetrii).

```
##   [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20  0  0  0  0  0
##  [26]  0  0  0  0  0  2  4  6  8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40
##  [51] 40 38 36 34 32 30 28 26 24 22 20 18 16 14 12 10  8  6  4  2  0  0  0  0  0
##  [76]  0  0  0  0  0 20 19 18 17 16 15 14 13 12 11 10  9  8  7  6  5  4  3  2  1
```

Zadanie 4. Z wektora `letters` wybierz litery na pozycjach 5, 10, 15, 20, 25.

```
## [1] "e" "j" "o" "t" "y"
```

Zadanie 5. Utwórz wektor liczb naturalnych od 1 do 1000, a następnie zamień liczby parzyste na ich odwrotności.

```
## [1] 1 0.5 3 0.25 5 0.1666667 ...
```

Zadanie 6. Uporządkuj elementy wektora $(6, 3, 4, 5, 2, 3)$ od największego do najmniejszego wykorzystując funkcję `order()`.

```
## [1] 6 5 4 3 3 2
```

Zadanie 7. Wyznacz znaki elementów wektora $(-1{,}876; -1{,}123; -0{,}123; 0; 0{,}123; 1{,}123; 1{,}876)$. Następnie zaokrąglij elementy tego wektora do dwóch miejsc po przecinku. Na koniec wyznacz część całkowitą każdego elementu nowego wektora.

```
## [1] -1 -1 -1  0  1  1  1
```

```
## [1] -1.88 -1.12 -0.12  0.00  0.12  1.12  1.88
```

```
## [1] -2 -2 -1  0  0  1  1
```

Zadanie 8. Wyznacz pierwiastek kwadratowy z każdej liczby naturalnej od 1 do 100 milionów. Najpierw wykonaj to polecenie korzystając z odpowiedniej funkcji wbudowanej w R, a następnie wykorzystując potęgowanie. Który sposób działa szybciej? **Wskazówka:** Do badania długości czasu działania programu można wykorzystać funkcję `Sys.time()`.

```
## Time difference of 1.485525 secs
## Time difference of 9.706759 secs
## [1] 1 1.414214 1.732051 2 2.236068 2.44949 ...
```

Zadanie 9. W pakiecie `schoolmath` znajduje się zbiór danych `primlist`, który zawiera liczby pierwsze pomiędzy 1 a 9999999.

- Znajdź największą liczbę pierwszą mniejszą od 1000.
- Ile jest liczb pierwszych większych od 100 a mniejszych od 500?

```
## [1] 997
```

```
## [1] 73
```

Zadanie 10. Wyznacz wszystkie kombinacje wartości wektorów (a, b) i $(1, 2, 3)$ za pomocą funkcji `rep()` i `paste()`.

```
## [1] "a1" "a2" "a3" "b1" "b2" "b3"
```

Zadanie 11. Utwórz wektor 30 napisów następującej postaci: `liczba.litera`, gdzie liczba to kolejne liczby naturalne od 1 do 30 a litera to trzy wielkie litery `X`, `Y`, `Z` występujące cyklicznie.

```
##  [1] "1.X"  "2.Y"  "3.Z"  "4.X"  "5.Y"  "6.Z"  "7.X"  "8.Y"  "9.Z"  "10.X"
## [11] "11.Y" "12.Z" "13.X" "14.Y" "15.Z" "16.X" "17.Y" "18.Z" "19.X" "20.Y"
## [21] "21.Z" "22.X" "23.Y" "24.Z" "25.X" "26.Y" "27.Z" "28.X" "29.Y" "30.Z"
```

Zadanie 12. W pewnych sytuacjach przydatna może się okazać tzw. kategoryzacja zmiennych, czyli inny podział na kategorie niżby wynikał z danych. Wygeneruj 100 obserwacji, które są odpowiedziami na pytania ankiety, każda odpowiedź może przyjąć jedną z wartości: 'a', 'b', 'c', 'd', 'e'. Dokonaj kategoryzacji w taki sposób, aby kategoria 1 obejmowała odpowiedzi 'a' i 'b', 2 odpowiedzi 'c' i 'd' oraz 3 odpowiedź 'e'. **Wskazówka:** Wykorzystaj funkcję `sample()` oraz funkcję `recode()` z pakietu `car`.

```
## Ładowanie wymaganego pakietu: carData
```

```
##   [1] "d" "b" "e" "d" "a" "e" "d" "b" "b" "d" "d" "d" "e" "d" "c" "d" "e" "b"
##  [19] "e" "b" "c" "d" "d" "c" "a" "c" "d" "b" "c" "b" "e" "a" "c" "a" "e" "a"
##  [37] "a" "b" "a" "c" "b" "c" "a" "c" "a" "b" "e" "a" "c" "c" "b" "e" "b" "d"
##  [55] "d" "a" "e" "c" "e" "c" "d" "d" "a" "d" "d" "c" "a" "d" "a" "b" "e" "e"
##  [73] "e" "a" "b" "b" "b" "e" "c" "d" "d" "c" "b" "d" "e" "b" "a" "c" "c" "a"
##  [91] "e" "a" "e" "a" "b" "c" "c" "e" "d" "c"
```

```
##   [1] 2 1 3 2 1 3 2 1 1 2 2 2 3 2 2 2 3 1 3 1 2 2 2 2 1 2 2 1 2 1 3 1 2 1 3 1 1
##  [38] 1 1 2 1 2 1 2 1 1 3 1 2 2 1 3 1 2 2 1 3 2 3 2 2 2 1 2 2 2 1 2 1 1 3 3 3 1
##  [75] 1 1 1 3 2 2 2 2 1 2 3 1 1 2 2 1 3 1 3 1 1 2 2 3 2 2
```

Zadanie 13. Skonstruuj listę o nazwie `moja_lista`, której pierwszym elementem będzie dwuelementowy wektor napisów zawierający Twoje imię i nazwisko, drugim elementem będzie liczba π, trzecim funkcja służąca do obliczania pierwiastka kwadratowego, a ostatni element listy to wektor złożony z liczb $0{,}02; 0{,}04; \ldots; 1$. Następnie usuń elementy numer jeden i trzy z tej listy. Na zakończenie, wyznacz listę zawierającą wartości funkcji gamma Eulera dla elementów listy `moja_lista`.

```
## List of 4
##  $ : chr [1:2] "Łukasz" "Smaga"
##  $ : num 3.14
##  $ :function (x)
##  $ : num [1:50] 0.02 0.04 0.06 0.08 0.1 0.12 0.14 0.16 0.18 0.2 ...
```

```
## List of 2
##  $ : num 3.14
##  $ : num [1:50] 0.02 0.04 0.06 0.08 0.1 0.12 0.14 0.16 0.18 0.2 ...
```

```
## [[1]]
## [1] 2.288038
##
## [[2]]
##  [1] 49.442210 24.460955 16.145727 11.996566  9.513508  7.863252  6.688686
##  [8]  5.811269  5.131821  4.590844  4.150482  3.785504  3.478450  3.216852
## [15]  2.991569  2.795751  2.624163  2.472735  2.338256  2.218160  2.110371
## [22]  2.013193  1.925227  1.845306  1.772454  1.705844  1.644773  1.588641
```

```
## [29]  1.536930  1.489192  1.445038  1.404128  1.366164  1.330884  1.298055
## [36]  1.267473  1.238954  1.212335  1.187471  1.164230  1.142494  1.122158
## [43]  1.103124  1.085308  1.068629  1.053016  1.038403  1.024732  1.011947
## [50]  1.000000
```

Zadanie 14. Utwórz wektor kwadratów 100 pierwszych liczb naturalnych. Następnie zlicz, które cyfry oraz jak często występują na pozycji jedności w kolejnych elementach tego wektora.

```
## [1] 1 4 9 16 25 36 ...
```

```
##
##  0  1  4  5  6  9
## 10 20 20 10 20 20
```

Zadanie 15. Za pomocą funkcji `outer()` wyznacz tabliczkę mnożenia dla liczb mniejszych od 6.

```
##      [,1]        [,2]         [,3]         [,4]         [,5]
## [1,] "1 * 1 = 1" "1 * 2 = 2"  "1 * 3 = 3"  "1 * 4 = 4"  "1 * 5 = 5"
## [2,] "2 * 1 = 2" "2 * 2 = 4"  "2 * 3 = 6"  "2 * 4 = 8"  "2 * 5 = 10"
## [3,] "3 * 1 = 3" "3 * 2 = 6"  "3 * 3 = 9"  "3 * 4 = 12" "3 * 5 = 15"
## [4,] "4 * 1 = 4" "4 * 2 = 8"  "4 * 3 = 12" "4 * 4 = 16" "4 * 5 = 20"
## [5,] "5 * 1 = 5" "5 * 2 = 10" "5 * 3 = 15" "5 * 4 = 20" "5 * 5 = 25"
```

Zadanie 16. Odczytaj zbiór danych dane1.csv a następnie:

1. Z odczytanej ramki danych wyświetl tylko parzyste wiersze.
2. Korzystając z operatorów logicznych wyświetl tylko wiersze odpowiadające pacjentkom starszym niż 50 lat z przerzutami do węzłów chłonnych (`Wezly.chlonne = 1`).

```
##   Wiek Rozmiar.guza Wezly.chlonne Nowotwor Receptory.estrogenowe
## 1   29            1             0        2                   (-)
## 2   29            1             0        2                  (++)
## 3   30            1             1        2                   (-)
## 4   32            1             0        3                  (++)
## 5   32            2             0       NA                   (-)
## 6   33            1             1        3                   (-)
##   Receptory.progesteronowe Niepowodzenia Okres.bez.wznowy VEGF
## 1                     (++)          brak               22  914
## 2                     (++)          brak               53 1118
## 3                      (+)          brak               38  630
## 4                     (++)          brak               26 1793
## 5                     (++)          brak               19  963
## 6                     (++)        wznowa               36 2776
```

```
## ...
```

```
##    Wiek Rozmiar.guza Wezly.chlonne Nowotwor Receptory.estrogenowe
## 2    29            1             0        2                  (++)
## 4    32            1             0        3                  (++)
## 6    33            1             1        3                   (-)
## 8    35            2             1        2                   (+)
## 10   36            1             1        2                   (-)
## 12   37            1             0        3                   (-)
##    Receptory.progesteronowe Niepowodzenia Okres.bez.wznowy VEGF
## 2                      (++)          brak               53 1118
## 4                      (++)          brak               26 1793
```

```
## 6                         (++)        wznowa                36 2776
## 8                         (++)          brak                38 3827
## 10                        (++)          brak                37  834
## 12                         (+)        wznowa                40 3331
```

```
## ...
```

```
##    Wiek Rozmiar.guza Wezly.chlonne Nowotwor Receptory.estrogenowe
## 78   51            1             1        2                  (++)
## 79   51            1             1        2                   (+)
## 81   51            2             1        2                 (+++)
## 84   51            2             1       NA                   (-)
## 88   52            2             1        2                   (+)
## 95   55            1             1        2                  (++)
##    Receptory.progesteronowe Niepowodzenia Okres.bez.wznowy VEGF
## 78                     (++)          brak               33  629
## 79                      (+)          brak               36 2879
## 81                     (++)          brak               52 1098
## 84                      (-)          brak               30 8064
## 88                      (+)        wznowa               48 1927
## 95                     (++)          brak               29  373
```

```
## ...
```

Zadanie 17. Oblicz iloczyn elementów dowolnego wektora `x` za pomocą pętli `while`, `repeat` i `for` (każdej z osobna).

```r
# dla
x <- 1:5
```

```
## [1] 120
```

Zadanie 18. Ile liczb postaci $\binom{n}{r}$ jest większych od miliona dla $1 \leqslant r \leqslant n \leqslant 100$?

```
## [1] 4075
```

Zadanie 19. Napisz funkcję, której argumentem będzie wektor liczbowy a wynikiem wektor zawierający trzy najmniejsze i trzy największe liczby w tym wektorze. W przypadku argumentu krótszego niż trzy liczby, funkcja ma zwracać komunikat o błędzie z komentarzem „za krótki argument”.

```r
# dla
x <- c(2, 6, 1, 5, 7, 3, 4)
```

```
## [1] 1 2 3 5 6 7
```

```r
# dla
x <- c(2, 6)
## Error in command 'extreme_3(x)': za krótki argument
```

2 Czym jest statystyka?

O statystyce, Johnson (Elementary Statistics (wydanie 4)) pisze w następujący sposób:

Statystyka jest uniwersalnym językiem nauki. Statystyka to coś więcej niż „zestaw narzędzi‘‘. Jako potencjalni użytkownicy statystyki musimy opanować „sztukę” prawidłowego korzystania z tych narzędzi. Staranne stosowanie metod statystycznych pozwala na

1. dokładne opisanie wyników badań naukowych,
2. podejmowanie decyzji,
3. dokonywanie oszacowań.

Statystyka obejmuje liczby, podmioty oraz ich wykorzystanie. Słowo „statystyka” ma różne znaczenie dla osób o różnym pochodzeniu i zainteresowaniach. Dla niektórych osób jest rodzaj „hokus-pokus‘‘, w którym osoba znająca zagadnienie przytłacza laika. Dla innych jest to sposób gromadzenia i przedstawiania dużych ilości informacji liczbowych. Dla jeszcze innej grupy jest to sposób na „podejmowanie decyzji w obliczu niepewności”. Z właściwej perspektywy każdy z tych punktów widzenia jest poprawny.

Dziedzinę statystyki można z grubsza podzielić na dwa obszary: statystyka opisowa i wnioskowanie statystyczne. Statystyka opisowa jest tym, co myślą ludzie, gdy słyszą słowo „statystyka”. Obejmuje gromadzenie, prezentację i opis danych liczbowych. Pojęcie wnioskowania statystycznego odnosi się do techniki interpretacji wartości wynikających z technik statystyki opisowej, a następnie wykorzystywania ich do podejmowania decyzji.

Statystyka to coś więcej niż liczby. Używać będziemy następującej definicji:

Statystyka to nauka gromadzenia, klasyfikacji, prezentacji i interpretacji danych liczbowych.

3 Statystyka opisowa

3.1 Podstawowe pojęcia

- populacja - zbiór pewnych elementów
- zmienna lub cecha - funkcja określona na elementach populacji (oznaczamy przez X)
- (rzeczywisty) rozkład populacji - rozkład wartości zmiennej X na elementach populacji
- próba - podzbiór populacji składający się z elementów podlegających badaniu statystycznemu
- dane - zaobserwowane wartości zmiennej X na elementach próby

$$\mathbf{x} = (x_1, \dots, x_n)^\top$$

- (empiryczny) rozkład próby - rozkład zmiennej X na elementach próby

3.2 Metody opisu rozkładu empirycznego

- Niech $\mathbf{x} = (x_1, \dots, x_n)^\top$ będzie próbą, tj. $x_1, \dots, x_n$ są obserwacjami (realizacjami) zmiennej (cechy) X.
- Celem statystyki opisowej jest przedstawienie rozkładu zmiennej X w próbie (rozkład empiryczny) przy użyciu tabeli lub wykresu lub pewnych liczb (statystyk opisowych).
- Często wystarczy podać tylko kilka liczb charakteryzujących ten rozkład.

Rodzaje zmiennych:

1. jakościowa, kategoryczna - Nie ma porządku. Istnieje tylko kilka możliwych wartości zmiennej. W próbie wartości zmiennej powtarzają się. Na przykład: kolor oczu, płeć.
2. ilościowa - istnieje porządek.
   - dyskretna - Istnieje tylko kilka lub przeliczalnie wiele możliwych wartości zmiennej. W próbie wartości zmiennej powtarzają się. Na przykład: liczba zgłoszeń w centrali telefonicznej.
   - ciągła - Istnieje nieskończenie wiele możliwych wartości zmiennej. W próbie wartości zmiennej nie powinny się powtarzać. Na przykład: czas, waga, wysokość, zarobki.

Metody opisu rozkładu empirycznego:

- tabela - szeregi rozdzielcze

- graficzny (oprócz wykresu kołowego, poniższe wykresy można rysować poziomo lub pionowo):
  - wykres słupkowy (jakościowa lub dyskretna zmienna ilościowa)
  - wykres kołowy (jakościowa lub dyskretna zmienna ilościowa)
  - histogram (ciągła zmienna ilościowa)
  - wykres pudełkowy lub ramka-wąsy lub ramkowy (zmienna ilościowa)
- statystyki opisowe:
  - klasyczne (uśrednienie wartości zaobserwowanych w próbie, np. średnia)
  - porządkowe (sortowanie wartości w próbie w porządku rosnącym, np. mediana)

Miary tendencji centralnej rozkładu empirycznego są wartościami liczbowymi, które mają lokalizować w pewnym sensie *środek* zbioru danych. Określenie *średnia* jest często związane z tymi miarami. Każdą z miar tendencji centralnej można nazwać wartością „średnią".

- średnia arytmetyczna (średnia)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

- mediana - wartość zmiennej w próbie uporządkowanej, dla której liczba obserwacji większych niż ta wartość jest równa liczbie obserwacji mniejszych niż ta wartość. Niech

$$x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n)}$$

  będzie próbą uporządkowaną. Wtedy mediana jest wyrażona za pomocą następującego wzoru:

$$Me = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & \text{gdy } n \text{ nieparzyste,} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right), & \text{gdy } n \text{ parzyste.} \end{cases}$$

Mediana jest szczególnym przypadkiem statystyk porządkowych zwanych kwantylami. Kwantyl rzędu $p \in [0,1]$ to liczba, która oddziela $p \cdot 100\%$ najmniejszych danych od $(1-p) \cdot 100\%$ największych obserwacji. Szczególne kwantyle to:

- minimum obserwacji to kwantyl rzędu $p = 0$ (podział $0\% - 100\%$)
- pierwszy (dolny) kwartyl Q_1 jest kwantylem rzędu $p = 1/4$ (podział $25\% - 75\%$)
- drugi kwartyl Q_2 zwany medianą Me jest kwantylem rzędu $p = 1/2$ (podział $50\% - 50\%$)
- trzeci (górny) kwartyl Q_3 jest kwantylem rzędu $p = 3/4$ (podział $75\% - 25\%$)
- maksimum obserwacji to kwantyl rzędu $p = 1$ (podział $100\% - 0\%$)

Po określeniu *środka* zbioru danych szukamy informacji o miarze rozrzutu (rozproszenia, dyspersji). **Miary dyspersji** rozkładu empirycznego opisują stopień rozprzestrzeniania się lub zmienności danych. Ściśle zgrupowane dane będą miały stosunkowo małe wartości tych miar, podczas gdy mniej zgrupowane dane będą miały większe wartości miar rozproszenia.

- odchylenie standardowe

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

  lub

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

  Wyraża średnie zróżnicowanie poszczególnych wartości zmiennych od średniej arytmetycznej.

- wariancja s^2

- współczynnik zmienności
$$V = \frac{s}{\bar{x}}100\%$$
Jest to miara niemianowana wyrażona w procentach, która umożliwia porównanie zmienności zmiennych mierzonych w różnych jednostkach.

Interpretacja odchylenia standardowego

Odchylenie standardowe jest rodzajem miernika, za pomocą którego możemy porównać jeden zestaw danych z innym. Znaczenie tej miary podają twierdzenie Czebyszewa i reguła trzech sigm.

Twierdzenie Czebyszewa. Dla dowolnego rozkładu, proporcja obserwacji mieszczących się w odległości k odchyleń standardowych od średniej wynosi co najmniej

$$1 - \frac{1}{k^2},$$

gdzie k jest dowolną liczbą dodatnią większą niż 1. Twierdzenie to dotyczy każdego rozkładu danych.

Na przykład twierdzenie Czebyszewa mówi, że w obrębie dwóch odchyleń standardowych od średniej ($k = 2$) zawsze znajdziemy co najmniej 75% danych, ponieważ

$$1 - \frac{1}{k^2} = 1 - \frac{1}{2^2} = 0{,}75.$$

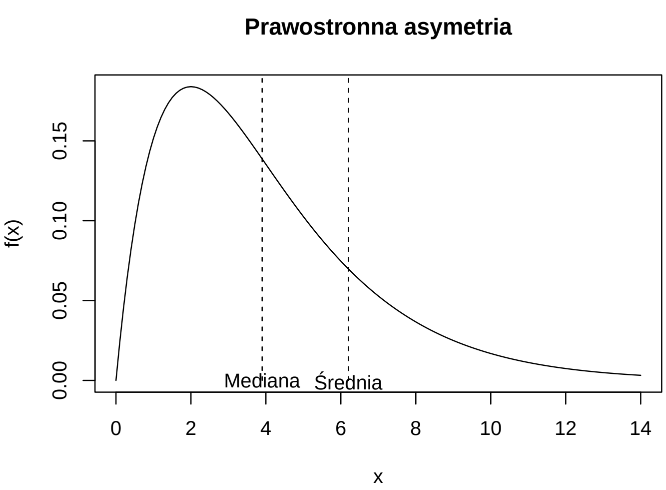
Reguła trzech sigm. Jeśli zmienna ma rozkład normalny, to

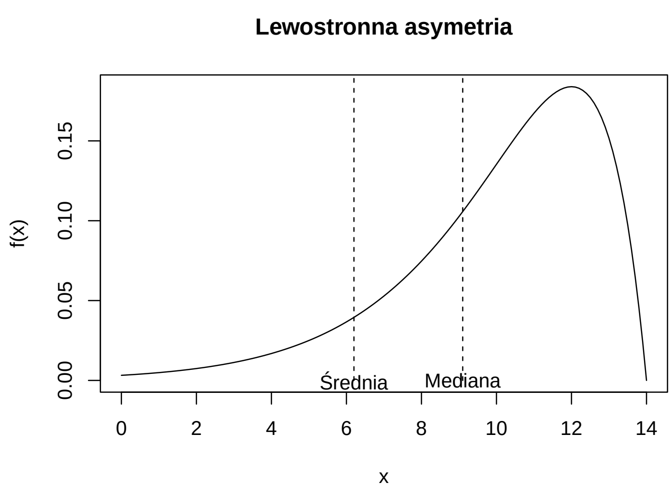
- w obrębie jednego odchylenia standardowego od średniej będzie około 68% danych.
- w obrębie dwóch odchyleń standardowych od średniej będzie około 95% danych.
- w obrębie trzech odchyleń standardowych od średniej będzie około 99,7% danych.

Normalność rozkładu danych można testować stosując regułę trzech sigm.

Miara asymetrii rozkładu

- współczynnik asymetrii (skośności)
$$A = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}$$
- Współczynnik asymetrii
  - równy zeru oznacza symetrię rozkładu zmiennej.
  - przyjmujący wartość dodatnią oznacza prawostronną asymetrię. Prawy ogon jest dłuższy, a masa rozkładu jest skoncentrowana po lewej stronie.
  - przyjmujący wartość ujemną oznacza lewostronną asymetrię. Lewy ogon jest dłuższy, a masa rozkładu jest skoncentrowana po prawej stronie.



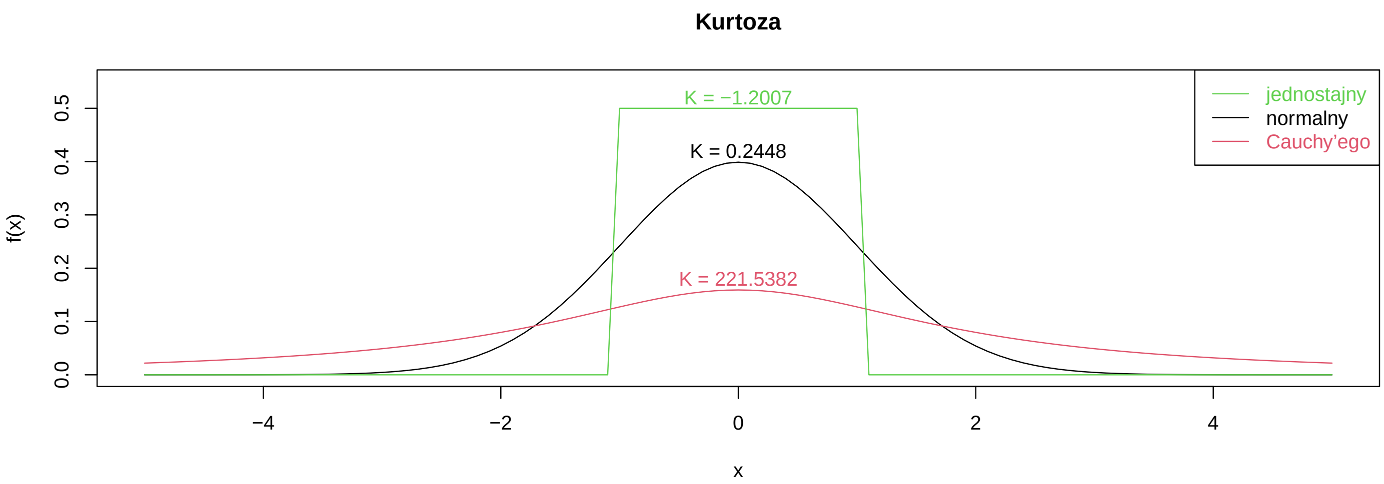


Kurtoza

- kurtoza (współczynnik koncentracji rozkładu)

$$K = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{s^4} - 3$$

- Kurtoza rozkładu normalnego wynosi zero.
- Wbrew stwierdzeniom obecnym w niektórych podręcznikach, kurtoza nie mierzy „spłaszczenia", „wysmukłości" ani „spiczastości" rozkładu (tzn. jak blisko jest rozkład w stosunku do tendencji centralnej).
- Na kurtozę ma wpływ intensywność występowania wartości skrajnych, mierzy więc ona, co się dzieje w „ogonach" rozkładu, natomiast kształt „czubka" rozkładu jest praktycznie bez znaczenia (patrz P.H. Westfall (2014) Kurtosis as Peakedness, 1905-2014. R.I.P., The American Statistician, 68 (3), 191-195).
- Rozkłady prawdopodobieństwa można podzielić ze względu na wartość kurtozy na rozkłady:
  - mezokurtyczne ($K = 0$) - wartość kurtozy wynosi 0, intensywność wartości skrajnych jest podobna do intensywności wartości skrajnych rozkładu normalnego,
  - leptokurtyczne ($K > 0$) - kurtoza jest dodatnia, intensywność wartości skrajnych jest większa niż dla rozkładu normalnego („ogony" rozkładu są „grubsze"),
  - platykurtyczne ($K < 0$) - kurtoza jest ujemna, intensywność wartości skrajnych jest mniejsza niż w przypadku rozkładu normalnego („ogony" rozkładu są „węższe").
- Krótko mówiąc, kurtoza jest miarą występowania wartości odstających.
- Zazwyczaj przyjmuje się wartości kurtozy $K < |2|$ lub $K < |3|$ jako „bezpieczną" wartość dla testów parametrycznych.



3.3 Przykłady 3

Przykład 1. Poniższe dane podają liczbę błędów w grupie 50 osób zdających egzamin testowy. Egzamin składał się z 18 pytań (można popełnić maksymalnie dwa błędy, aby zdać egzamin).

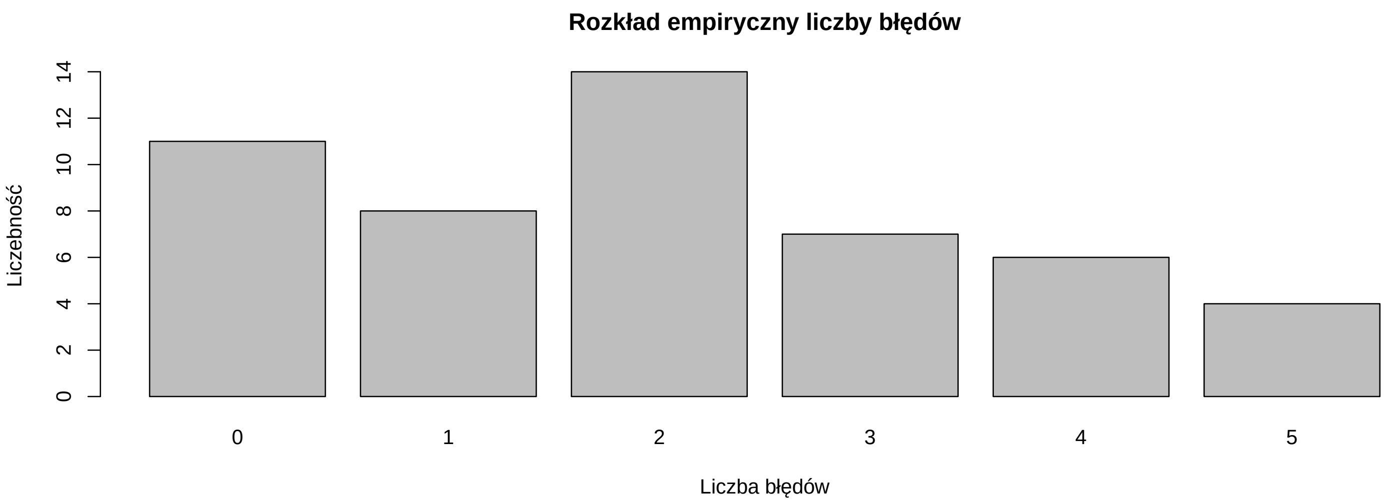
1	1	2	0	1	3	1	4	4	4	0	1	0	0	0	2	3
4	0	1	5	2	3	5	3	2	2	4	0	2	2	0	2	2
3	3	1	3	2	2	0	0	5	4	2	1	5	2	2	0	

Zmienna X to liczba błędów. Jest to dyskretna zmienna ilościowa.

```
liczba_bledow <- c(1, 1, 2, 0, 1, 3, 1, 4, 4, 4, 0, 1, 0, 0, 0, 2, 3,
                   4, 0, 1, 5, 2, 3, 5, 3, 2, 2, 4, 0, 2, 2, 0, 2, 2,
                   3, 3, 1, 3, 2, 2, 0, 0, 5, 4, 2, 1, 5, 2, 2, 0)
# rozkład empiryczny opisany za pomocą szeregu rozdzielczego
data.frame(cbind(liczebnosc = table(liczba_bledow),
                 procent = prop.table(table(liczba_bledow))))
```

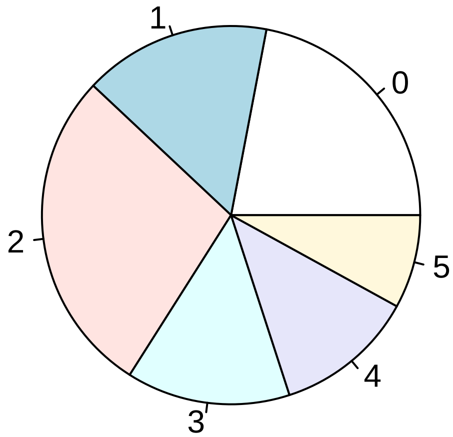
```
##   liczebnosc procent
## 0         11    0.22
## 1          8    0.16
## 2         14    0.28
## 3          7    0.14
## 4          6    0.12
## 5          4    0.08
```

```
# wykres słupkowy
barplot(table(liczba_bledow),
        xlab = "Liczba błędów", ylab = "Liczebność",
        main = "Rozkład empiryczny liczby błędów")
```




```
barplot(prop.table(table(liczba_bledow)),
        xlab = "Liczba błędów", ylab = "Prawdopodobieństwo",
        main = "Rozkład empiryczny liczby błędów")
```

```r
# wykres kołowy
pie(table(liczba_bledow))
```




```r
# średnia
mean(liczba_bledow)
```

```
## [1] 2.02
```

```r
# mediana
median(liczba_bledow)
```

```
## [1] 2
```

```r
# odchylenie standardowe
sd(liczba_bledow)
```

```
## [1] 1.558256
```

```r
# współczynnik zmienności
sd(liczba_bledow) / mean(liczba_bledow) * 100
```

```
## [1] 77.14141
```

Przykład 2. Przeprowadzono 50 niezależnych eksperymentów obejmujących hamowanie pewnego typu samochodu (na suchym asfalcie, z prędkością 40km/h itp.). Notowano długość drogi hamowania w metrach z dokładnością do jednego centymetra. Uzyskane wyniki są zawarte w pliku hamulce.txt. Zmienna X to długość drogi hamowania. Jest to zmienna ilościowa ciągła.

```r
hamulce <- read.table("http://ls.home.amu.edu.pl/data_sets/hamulce.txt", dec = ",")
head(hamulce)
```

```
##      V1
## 1 18.66
## 2 17.81
## 3 18.96
## 4 18.09
## 5 18.73
## 6 18.45
```

```r
data.frame(cbind(liczebnosc = table(cut(hamulce$V1, breaks = seq(17.6, 19, 0.2))),
                 procent = prop.table(table(cut(hamulce$V1, breaks = seq(17.6, 19, 0.2))))))
```

```
##             liczebnosc procent
## (17.6,17.8]          4    0.08
```

```
## (17.8,18]          5    0.10
## (18,18.2]          6    0.12
## (18.2,18.4]        8    0.16
## (18.4,18.6]       11    0.22
## (18.6,18.8]       12    0.24
## (18.8,19]          4    0.08
```

```
(breaks_hist <- hist(hamulce$V1, plot = FALSE)$breaks)
```

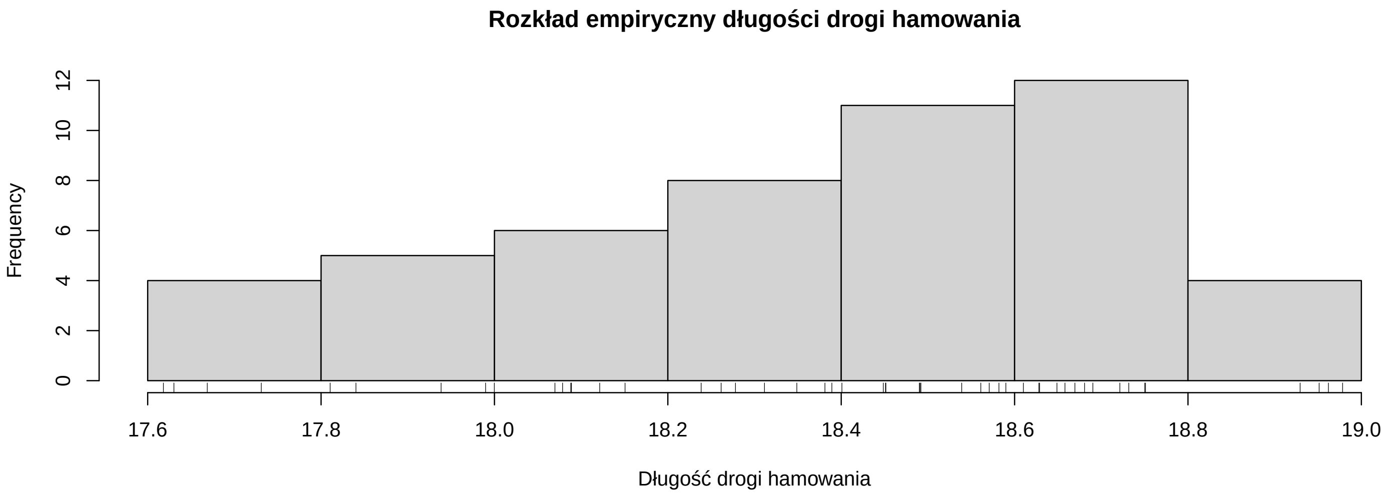
```
## [1] 17.6 17.8 18.0 18.2 18.4 18.6 18.8 19.0
```

```
data.frame(cbind(liczebnosc = table(cut(hamulce$V1, breaks = breaks_hist)),
                 procent = prop.table(table(cut(hamulce$V1, breaks = breaks_hist)))))
```

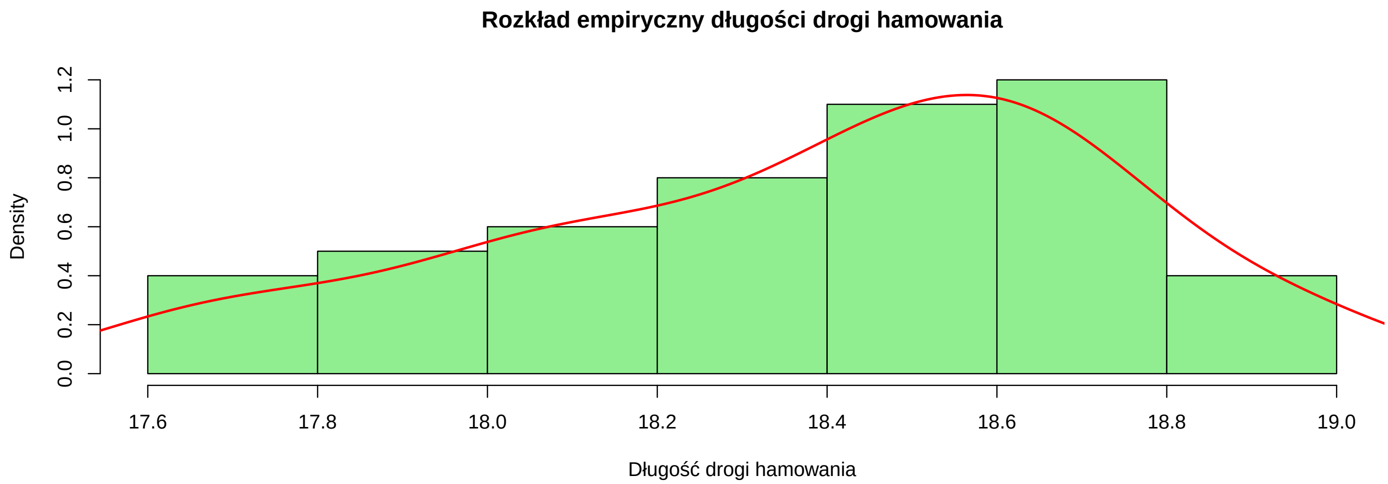
```
##             liczebnosc procent
## (17.6,17.8]          4    0.08
## (17.8,18]            5    0.10
## (18,18.2]            6    0.12
## (18.2,18.4]          8    0.16
## (18.4,18.6]         11    0.22
## (18.6,18.8]         12    0.24
## (18.8,19]            4    0.08
```

Histogram - zestaw sąsiadujących prostokątów, których podstawy, równe rozpiętości przedziałów klasowych, znajdują się na osi odciętych, a wysokości są liczebnościami przedziałów.

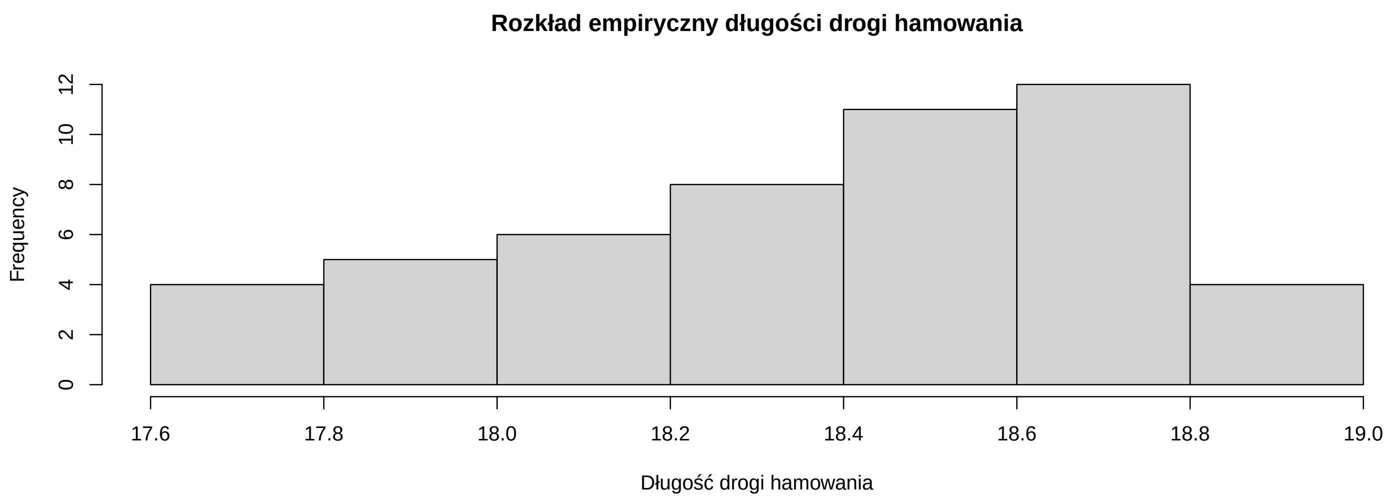
```
# histogram
hist(hamulce$V1,
     xlab = "Długość drogi hamowania",
     main = "Rozkład empiryczny długości drogi hamowania")
rug(jitter(hamulce$V1))
```



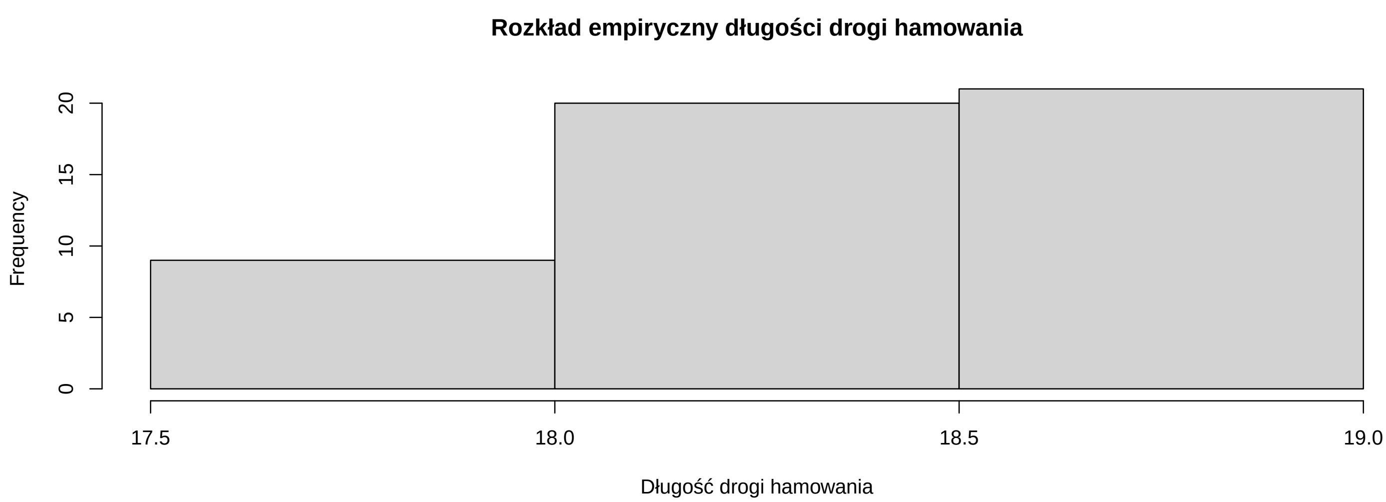

```
# histogram z estymatorem jądrowym gęstości
hist(hamulce$V1,
     xlab = "Długość drogi hamowania",
     main = "Rozkład empiryczny długości drogi hamowania",
     probability = TRUE,
     col = "lightgreen")
lines(density(hamulce$V1), col = "red", lwd = 2)
```




```r
# tylko sugestia liczby grup
hist(hamulce$V1,
     xlab = "Długość drogi hamowania",
     main = "Rozkład empiryczny długości drogi hamowania",
     breaks = 5)
```




```r
hist(hamulce$V1,
     xlab = "Długość drogi hamowania",
     main = "Rozkład empiryczny długości drogi hamowania",
     breaks = 3)
```



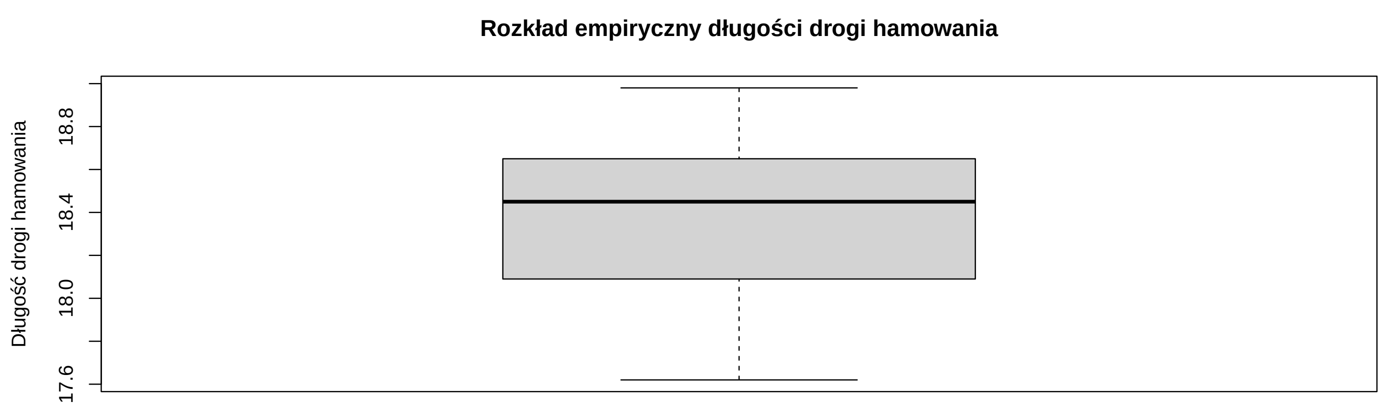
Wykres ramkowy to metoda graficznego przedstawienia danych liczbowych za pomocą ich kwantyli. Tworzymy go poprzez umieszczenie na osi pionowej wartości niektórych parametrów rozkładu (kwantyli).

- Wewnątrz prostokąta znajduje się pogrubiona pozioma linia, która określa wartość mediany.
- Nad osią znajduje się prostokąt (ramka), którego dolny bok jest określony przez pierwszy kwartyl, a górny bok przez trzeci kwartyl. Wysokość pudełka odpowiada wartości rozstępu międzykwartylowego $(Q_3 - Q_1)$.
- Pudełko jest uzupełnione od góry i od dołu segmentami (wąsami). Dolny koniec dolnego segmentu reprezentuje najmniejszą wartość w zestawie danych, zaś górny koniec górnego segmentu jest obserwacją największą. Wartości te muszą spełniać dodatkowy warunek, a mianowicie dolny koniec nie może być mniejszy niż $Q_1 - 1{,}5 \cdot (Q_3 - Q_1)$, a górny większy niż $Q_3 + 1{,}5 \cdot (Q_3 - Q_1)$. Jeśli istnieją obserwacje poza tym zakresem, są one zaznaczane na wykresie indywidualnie jako osobne punkty i są traktowane jako obserwacje odstające.

Wykres pudełkowy jako wskaźnik tendencji centralnej, dyspersji, symetrii, skośności i wielkości ogona:

- dyspersja - odstępy między różnymi częściami pudełka
- symetryczny - pogrubiona linia znajduje się blisko środka pudełka, a długości wąsów są takie same
- prawostronnie asymetryczny - górny wąs jest znacznie dłuższy niż dolny wąs, a linia jest bliższa dolnej części pudełka.
- lewostronnie asymetryczny - dolny wąs jest znacznie dłuższy niż górny wąs, a linia jest bliższa górnej części pudełka
- grube ogony - długość wąsów znacznie przekracza długość pudełka
- cieńkie ogony - długość wąsów jest krótsza niż długość pudełka
- bardzo krótkie ogony (populacja w kształcie litery U, z zanurzeniem w środku zamiast garbu) - wąsy są nieobecne

```r
# wykres ramkowy
boxplot(hamulce$V1,
        ylab = "Długość drogi hamowania",
        main = "Rozkład empiryczny długości drogi hamowania")
```



- statystyki opisowe

```r
# średnia
mean(hamulce$V1)
```

```
## [1] 18.3818
```

```r
# mediana
median(hamulce$V1)
```

```
## [1] 18.45
```

```r
# odchylenie standardowe
sd(hamulce$V1)
```

```
## [1] 0.3603439
```

```r
# współczynnik zmienności
sd(hamulce$V1) / mean(hamulce$V1) * 100
```

```
## [1] 1.96033
```

```r
library(e1071)
# współczynnik asymetrii
skewness(hamulce$V1)
```

```
## [1] -0.452857
```

```r
# kurtoza
kurtosis(hamulce$V1)
```

```
## [1] -0.6738354
```

3.4 Zadania 3

Zadanie 1. Zmienna `wynik` w pliku ankieta.txt opisuje wyniki badania działalności prezydenta pewnego miasta. Wybrano losowo 100 mieszkańców miasta i zadano im następujące pytanie: Jak oceniasz działalność prezydenta miasta? Dostępne były następujące odpowiedzi: zdecydowanie dobrze (`a`), dobrze (`b`), źle (`c`), zdecydowanie źle (`d`), nie mam zdania (`e`). Jakiego typu jest ta zmienna? Jakie są możliwe wartości tej zmiennej?

1. Zaimportuj dane z pliku ankieta.txt do zmiennej `ankieta`.

```
##   plec szkola wynik
## 1    m      p     d
## 2    m      s     e
## 3    m      w     a
## 4    m      s     d
## 5    m      p     c
## 6    m      w     c
```

```
## ...
```

2. Przedstaw rozkład empiryczny zmiennej `wynik` za pomocą szeregu rozdzielczego.

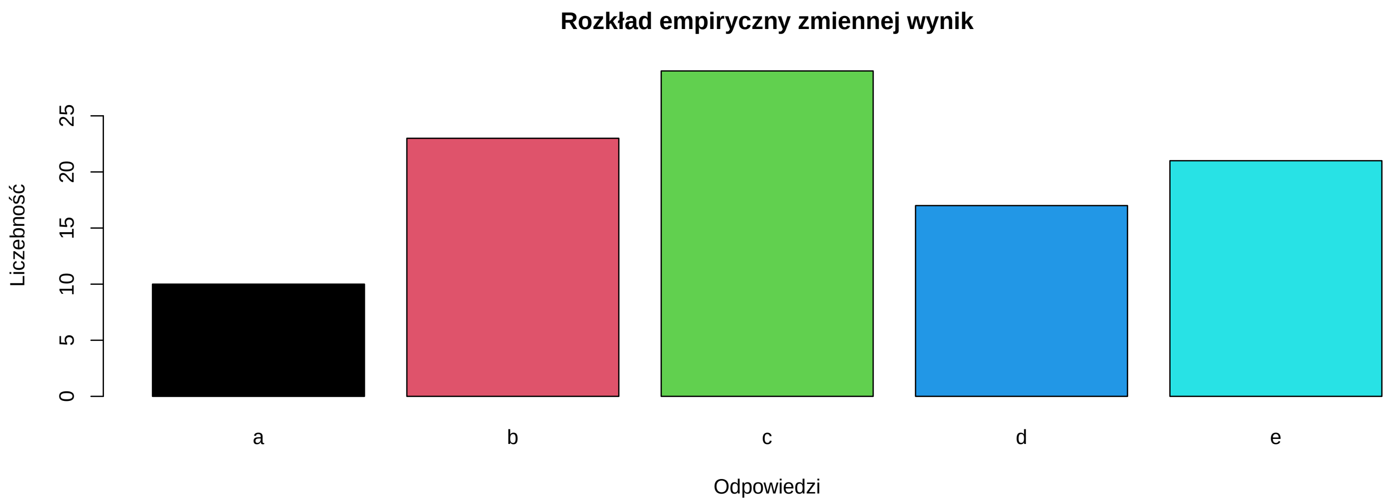
```
##   liczebnosc procent
## a         10    0.10
## b         23    0.23
## c         29    0.29
## d         17    0.17
## e         21    0.21
```

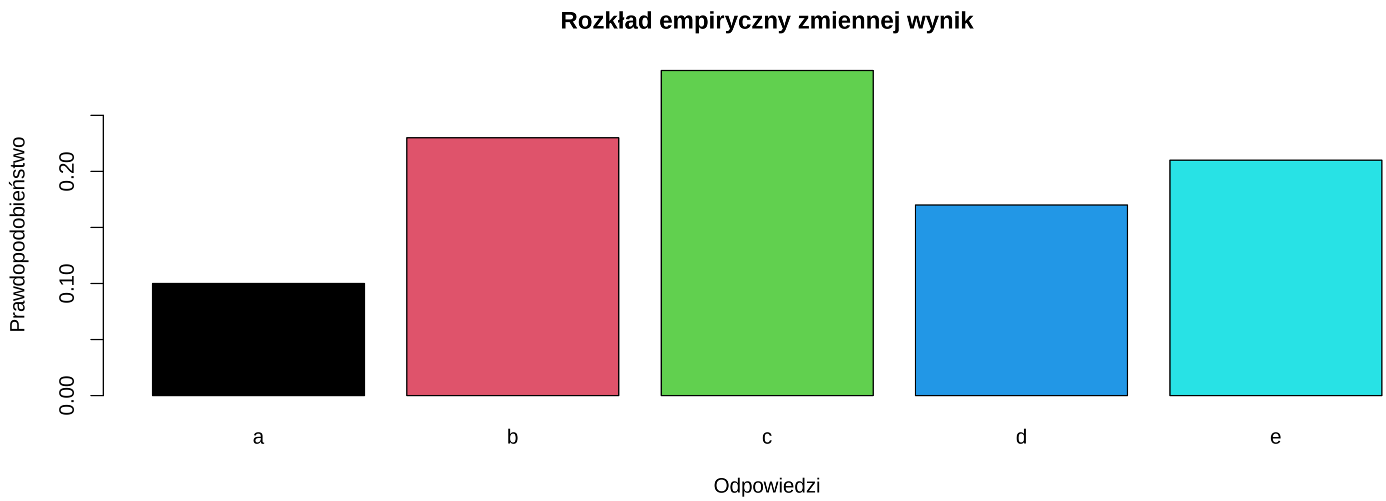
3. Przedstaw rozkład empiryczny zmiennej `wynik` tylko dla osób z wykształceniem podstawowym za pomocą szeregu rozdzielczego.

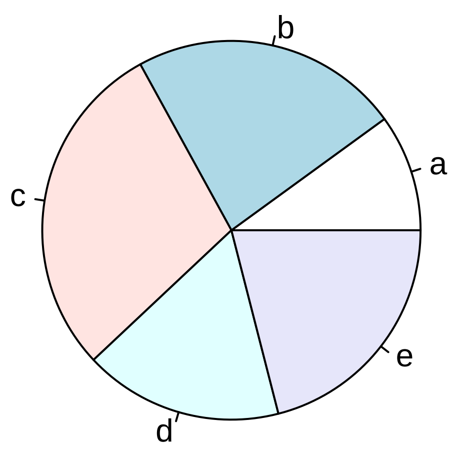
```
##   liczebnosc    procent
## a          2 0.11764706
## b          3 0.17647059
## c          4 0.23529412
```

```
## d          7 0.41176471
## e          1 0.05882353
```

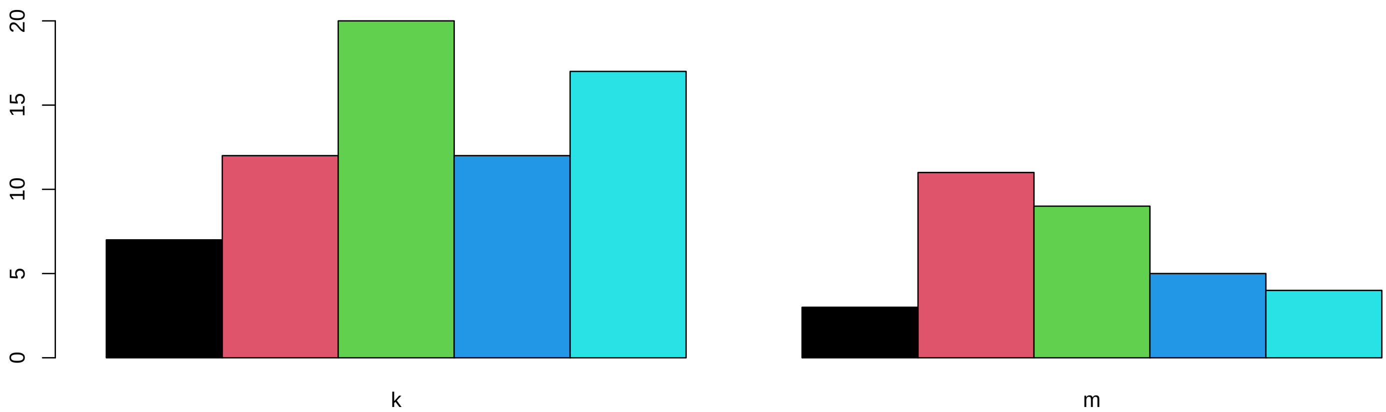
4. Zilustruj wyniki ankiety za pomocą wykresu słupkowego i kołowego.







5. Zilustruj wyniki ankiety za pomocą wykresu słupkowego z podziałem na kobiety i mężczyzn.



6. Zinterpretuj powyższe wyniki (tabelaryczne i graficzne).

Zadanie 2. Przebadano 200 losowo wybranych 5-sekundowych okresów pracy centrali telefonicznej. Rejestrowano liczbę zgłoszeń. Wyniki są zawarte w pliku Centrala.RData. Jakiego typu jest ta zmienna? Jakie są możliwe wartości tej zmiennej?

1. Zaimportuj dane z pliku Centrala.RData.

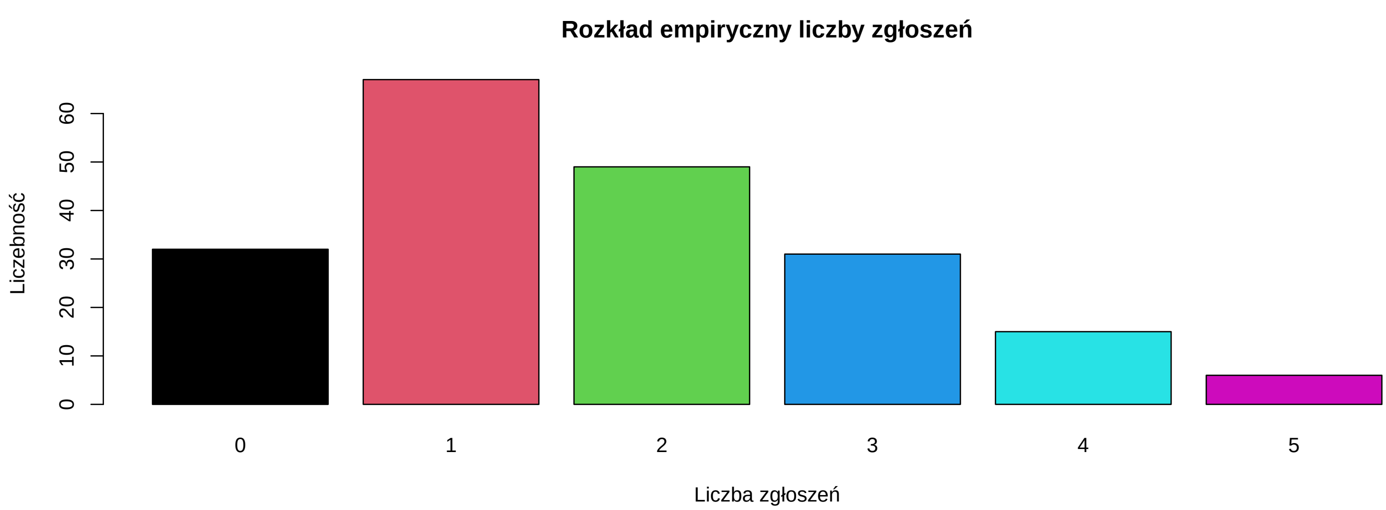
```
##   Liczba
## 1      0
## 2      0
## 3      5
## 4      1
## 5      1
## 6      2
```

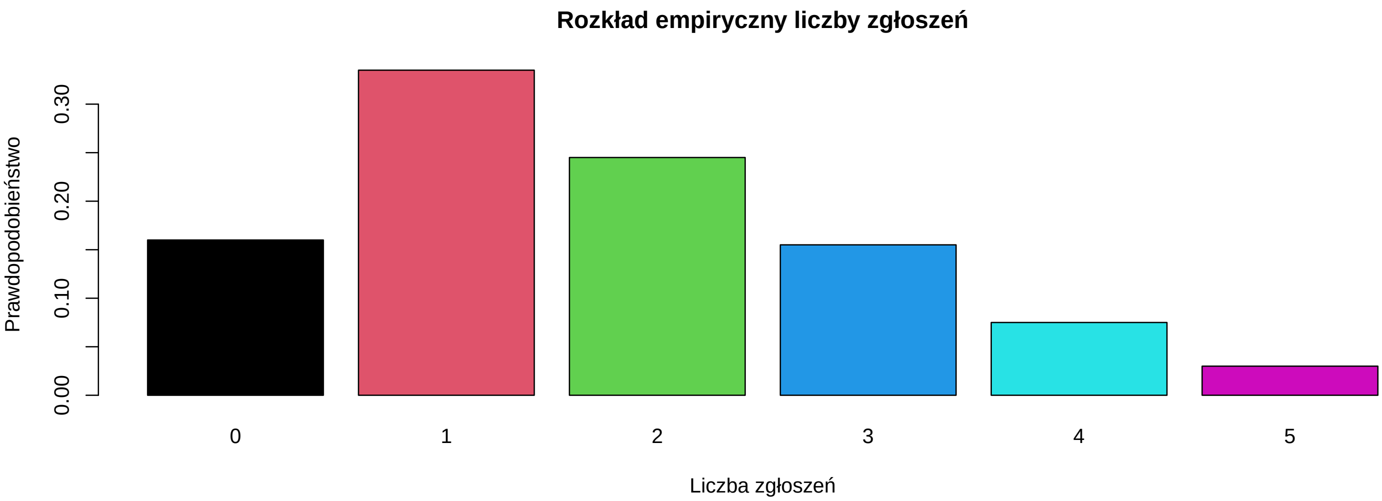
```
## ...
```

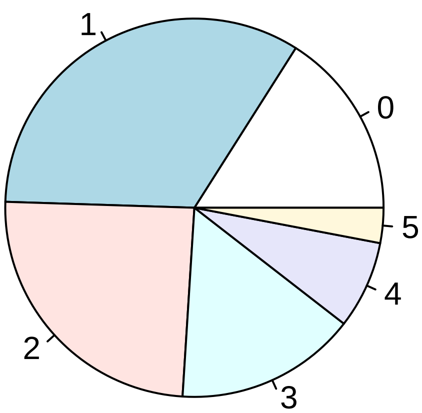
2. Przedstaw rozkład empiryczny liczby zgłoszeń za pomocą szeregu rozdzielczego.

```
##   liczebnosc procent
## 0         32   0.160
## 1         67   0.335
## 2         49   0.245
## 3         31   0.155
## 4         15   0.075
## 5          6   0.030
```

3. Zilustruj liczbę zgłoszeń za pomocą wykresu słupkowego i kołowego.







4. Obliczyć średnią z liczby zgłoszeń, medianę liczby zgłoszeń, odchylenie standardowe liczby zgłoszeń i współczynnik zmienności liczby zgłoszeń.

```
## [1] 1.74
```

```
## [1] 2
```

```
## [1] 1.28086
```

```
## [1] 73.61266
```

5. Zinterpretuj powyższe wyniki (tabelaryczne, graficzne i liczbowe).

Zadanie 3. Zmienna w pliku awarie.txt opisuje wyniki 50 pomiarów czasu bezawaryjnej pracy danego urządzenia (w godzinach). Jakiego typu jest ta zmienna? Jakie są możliwe wartości tej zmiennej?

1. Zaimportuj dane z pliku awarie.txt.

```
##    V1
## 1 629
## 2 325
## 3 215
## 4 518
## 5 297
## 6 792
```

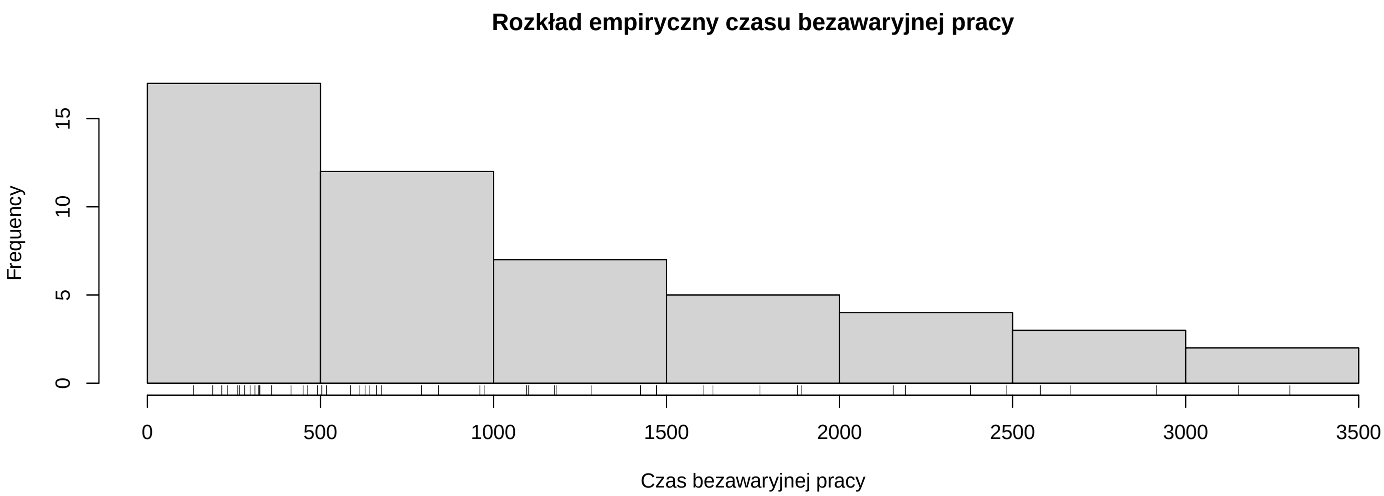
```
## ...
```

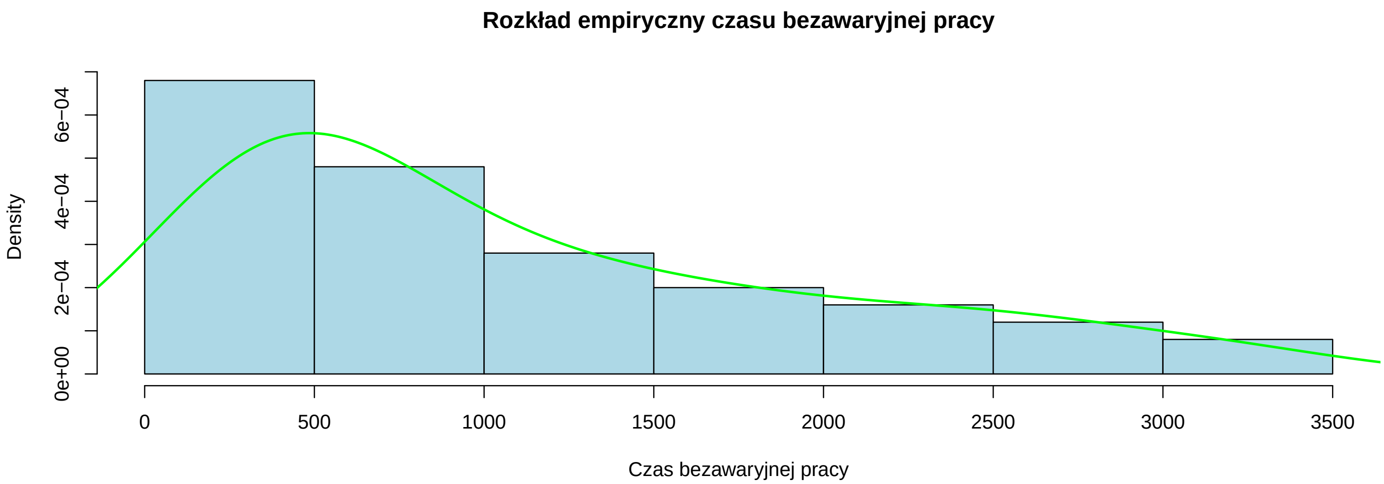
2. Przedstaw rozkład empiryczny czasu bezawaryjnej pracy za pomocą szeregu rozdzielczego.

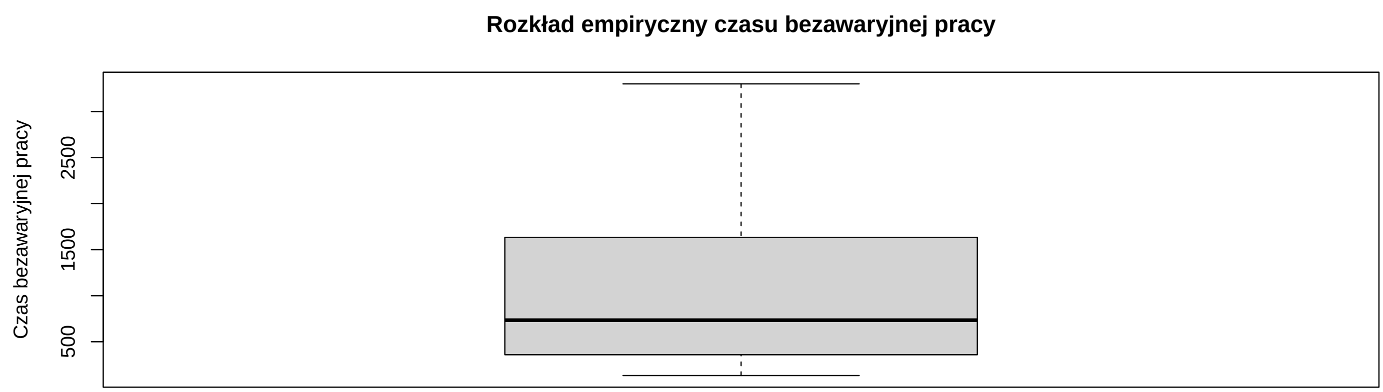
```
##                  liczebnosc procent
```

```
## (0,500]               17    0.34
## (500,1e+03]           12    0.24
## (1e+03,1.5e+03]        7    0.14
## (1.5e+03,2e+03]        5    0.10
## (2e+03,2.5e+03]        4    0.08
## (2.5e+03,3e+03]        3    0.06
## (3e+03,3.5e+03]        2    0.04
```

3. Zilustruj rozkład empiryczny czasu bezawaryjnej pracy za pomocą histogramu i wykresu pudełkowego. Jakie są zalety i wady tych wykresów?







4. Obliczyć średnią, medianę, odchylenie standardowe, współczynnik zmienności, współczynnik asymetrii i kurtozę czasu bezawaryjnej pracy.

```
## [1] 1101.36
## [1] 734
## [1] 883.2735
## [1] 80.19844
## [1] 0.9109508
## [1] -0.354536
```

5. Zinterpretuj powyższe wyniki (tabelaryczne, graficzne i liczbowe).

Zadanie 4. Napisz funkcję `wspolczynnik_zmiennosci()`, która oblicza wartość współczynnika zmienności dla danego wektora obserwacji. Funkcja powinna mieć dwa argumenty:

- `x` - wektor zawierający dane,
- `na.rm` - wartość logiczna (domyślnie `FALSE`), która wskazuje czy braki danych (obiekty `NA`) mają być zignorowane.

Funkcja zwraca wartość współczynnika zmienności wyrażoną w procentach. Ponadto funkcja sprawdza, czy wektor `x` jest wektorem numerycznym. W przeciwnym razie zostanie zwrócony błąd z następującym komunikatem: „argument nie jest liczbą”. Przykładowe wywołania i wyniki funkcji są następujące:

```r
x <- c(1, NA, 3)
wspolczynnik_zmiennosci(x)
## [1] NA
wspolczynnik_zmiennosci(x, na.rm = TRUE)
## [1] 70.71068
wspolczynnik_zmiennosci()
## Error in wspolczynnik_zmiennosci() :
##   argument "x" is missing, with no default
wspolczynnik_zmiennosci(c("x", "y"))
## Error in wspolczynnik_zmiennosci(c("x", "y")) : argument nie jest liczbą
```

4 Model statystyczny

- Model statystyczny (model w ogólności) może być definiowany i interpretowany na wiele sposobów.
- Mówiąc prosto, model statystyczny jest rodzajem teoretycznej reprezentacji danych.
- Model statystyczny jest zwykle określany jako matematyczny związek między jedną lub większą liczbą zmiennych losowych a innymi zmiennymi nielosowymi.
- Model statystyczny reprezentuje, często w znacznie wyidealizowanej formie, proces generowania danych.
- Załóżmy, że dane
$$\mathbf{x} = (x_1, \dots, x_n)^\top$$
są realizacjami wektora losowego
$$\mathbf{X} = (X_1, \dots, X_n)^\top,$$
zwanego **próbą**, o rozkładzie P należącym do pewnej rodziny rozkładów prawdopodobieństwa $\mathcal{P}$.
- Niech badana zmienna X populacji ma rozkład P należący do rodziny rozkładów prawdopodobieństwa $\mathcal{P}$. O obserwacjach $X_1, \dots, X_n$ zakładamy, że są to niezależne zmienne losowe o takim samym rozkładzie

P jak badana zmienna X. W konsekwencji próba

$$\mathbf{X} = (X_1, \dots, X_n)^\top$$

jest nazwana **próbą prostą** z populacji o rozkładzie P.

- Modele można podzielić w następujący sposób:
  1. Modele parametryczne:
     - model dwumianowy,
     - model Poissona,
     - model normalny,
     - model wykładniczy.
  2. Modele nieparametryczne:
     - model zmiennej dyskretnej,
     - model zmiennej ciągłej.
- W modelach parametrycznych możemy indeksować rozkłady $P \in \mathcal{P}$ parametrem θ, więc

$$\mathcal{P} = \{P_\theta : \theta \in \Theta\}.$$

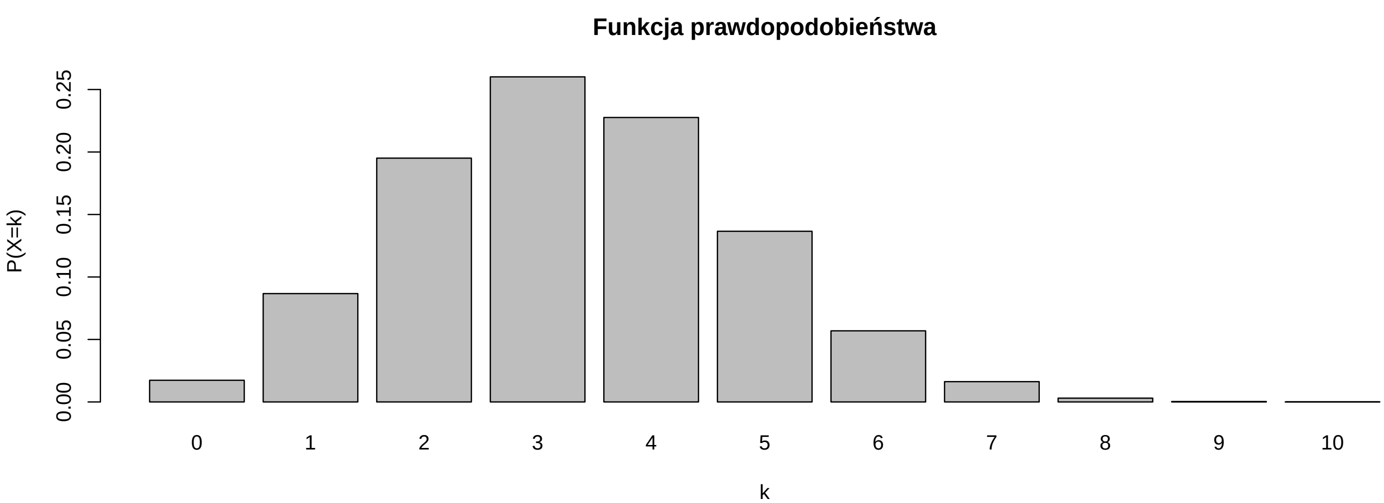
Wybrane rozkłady prawdopodobieństwa:

1. rozkład dwumianowy $b(m, p)$, $m \in \mathbb{N}$, $p \in (0, 1)$

$$\mathbb{P}(X = k) = \binom{m}{k} p^k (1-p)^{m-k},\ k = 0, 1, \dots, m$$

- Funkcja prawdopodobieństwa zmiennej $X \sim b(10, 1/3)$

```
barplot(dbinom(x = 0:10, size = 10, prob = 1 / 3), names.arg = 0:10,
       xlab = "k", ylab = "P(X=k)", main = "Funkcja prawdopodobieństwa")
```

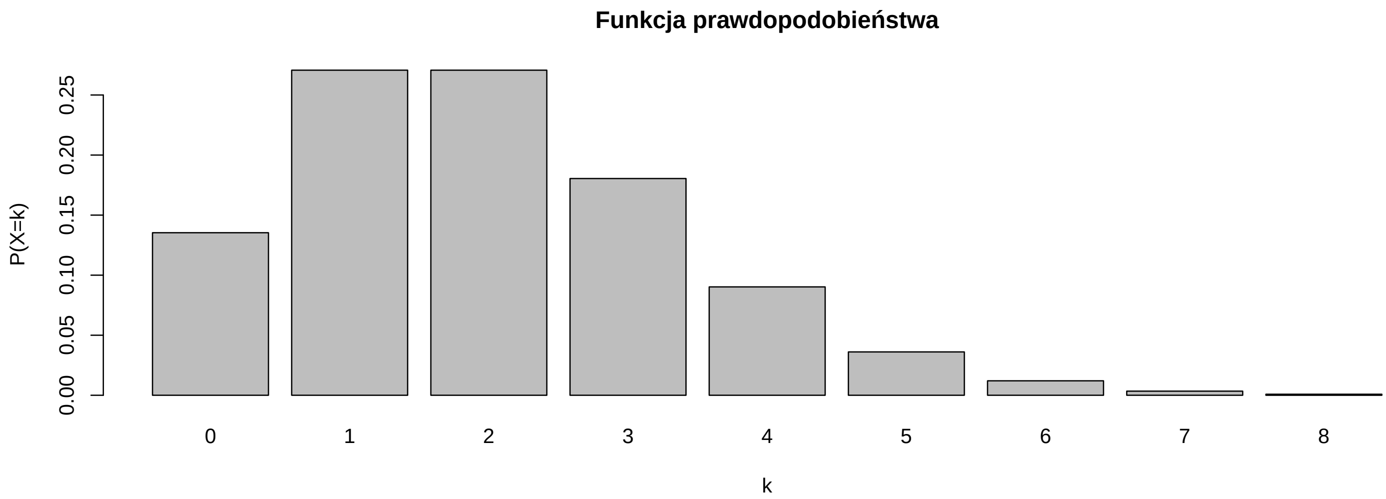


2. rozkład Poissona $\pi(\lambda)$, $\lambda > 0$

$$\mathbb{P}(X = k) = \frac{e^{-\lambda}\lambda^k}{k!},\ k = 0, 1, \dots$$

- Funkcja prawdopodobieństwa zmiennej $X \sim \pi(2)$

```
barplot(dpois(x = 0:8, lambda = 2), names.arg = 0:8,
       xlab = "k", ylab = "P(X=k)", main = "Funkcja prawdopodobieństwa")
```

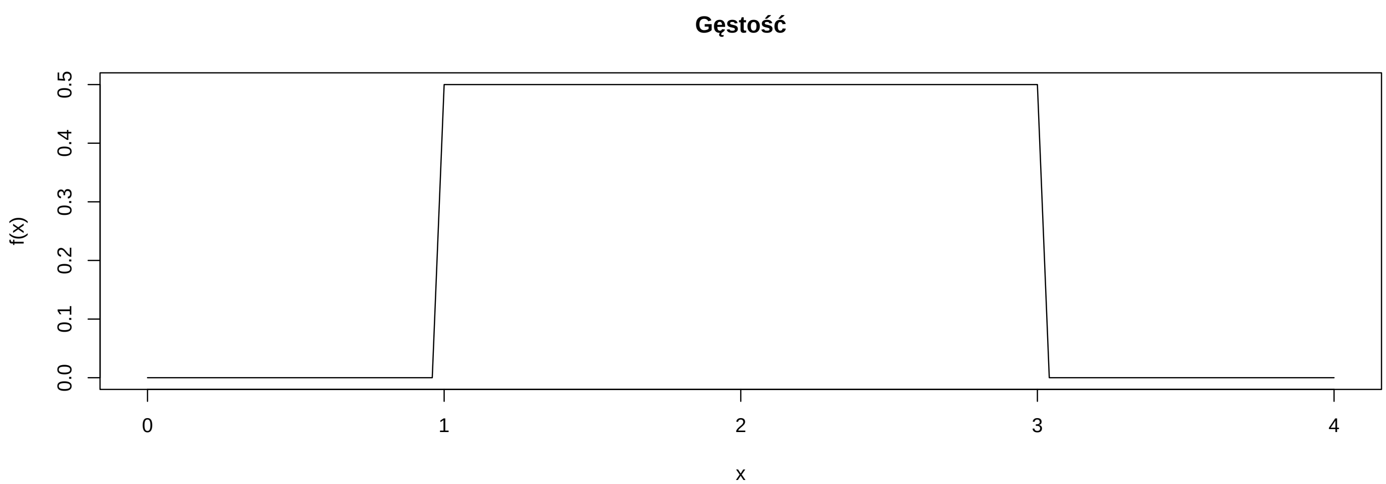


3. rozkład jednostajny $U(a, b)$, $a < b$

$$f_X(x) = \begin{cases} \frac{1}{b-a} & dla \ x \in (a, b) \\ 0 & dla \ x \notin (a, b) \end{cases}$$

- Gęstość zmiennej $X \sim U(1, 3)$

```
curve(dunif(x, min = 1, max = 3), 0, 4, ylab = "f(x)", main = "Gęstość")
```

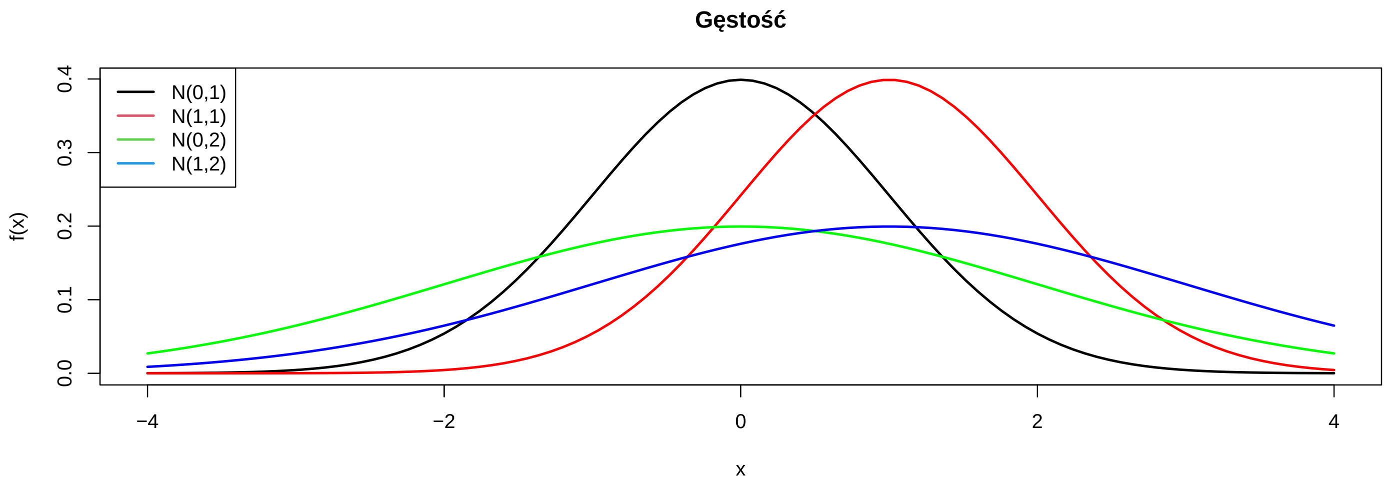


4. rozkład normalny $N(\mu, \sigma)$, $\mu \in \mathbb{R}$, $\sigma > 0$

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- Gęstości rozkładów normalnych

```
curve(dnorm, -4, 4, ylab = "f(x)", main = "Gęstość", lwd = 2)
curve(dnorm(x, mean = 1), col = "red", add = TRUE, lwd = 2)
curve(dnorm(x, sd = 2), col = "green", add = TRUE, lwd = 2)
curve(dnorm(x, mean = 1, sd = 2), col = "blue", add = TRUE, lwd = 2)
legend("topleft", lwd = 2, col = 1:4, legend = c("N(0,1)", "N(1,1)", "N(0,2)", "N(1,2)"))
```



5. rozkład wykładniczy $Ex(\lambda)$, $\lambda > 0$

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & dla\ x > 0 \\ 0 & dla\ x \leqslant 0 \end{cases}$$

- Gęstości rozkładów wykładniczych

```
curve(dexp, 0, 5, ylim = c(0, 2), ylab = "f(x)", main = "Gęstość", lwd = 2)
curve(dexp(x, rate = 2), col = "red", add = TRUE, lwd = 2)
curve(dexp(x, rate = 0.5), col = "green", add = TRUE, lwd = 2)
legend("topright", lwd = 2, col = 1:3, legend = c("Ex(1)", "Ex(2)", "Ex(0.5)"))
```

6. rozkład Rayleigha $R(\lambda)$, $\lambda > 0$

$$f_\lambda(x) = \frac{2}{\lambda} x \exp\left(-\frac{x^2}{\lambda}\right) I_{(0,\infty)}(x)$$

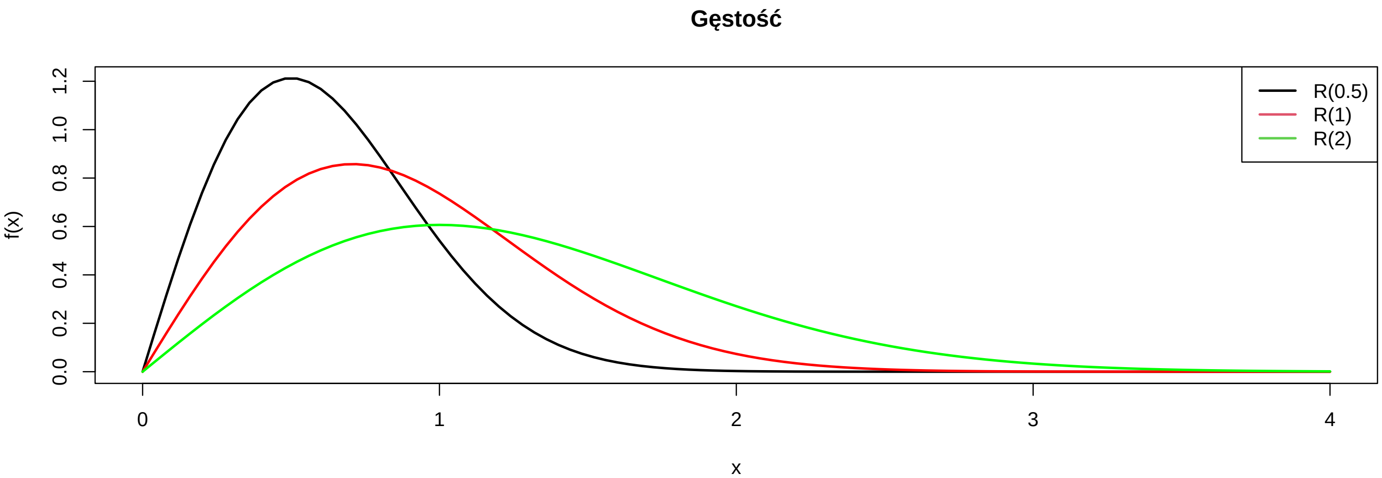
Uwaga. Rozkład Rayleigha jest zaimplementowany w pakiecie `VGAM` z następującą funkcją gęstości

$$f_\sigma(x) = \frac{x}{\sigma^2} \exp\left(-\frac{x^2}{2\sigma^2}\right) I_{(0,\infty)}(x),$$

więc w naszej notacji $\sigma = \sqrt{\frac{\lambda}{2}}$.

- Gęstości rozkładów Rayleigha

```
lambda <- 0.5
curve(VGAM::drayleigh(x, sqrt(lambda / 2)),
      xlim = c(0, 4), ylab = "f(x)", main = "Gęstość", lwd = 2)
lambda <- 1
curve(VGAM::drayleigh(x, sqrt(lambda / 2)),
      col = "red", add = TRUE, lwd = 2)
lambda <- 2
curve(VGAM::drayleigh(x, sqrt(lambda / 2)),
      col = "green", add = TRUE, lwd = 2)
legend("topright", lwd = 2, col = 1:3, legend = c("R(0.5)", "R(1)", "R(2)"))
```



Rozkłady prawdopodobieństwa w programie R

Rozkład	Dystrybuanta	Gęstość/Funkcja prawd.	Kwantyl	Generator
dwumianowy	pbinom	dbinom	qbinom	rbinom
Poissona	ppois	dpois	qpois	rpois
ujemny dwumianowy	pnbinom	dnbinom	qnbinom	rnbinom
geometryczny	pgeom	dgeom	qgeom	rgeom
hipergeometryczny	phyper	dhyper	qhyper	rhyper
jednostajny	punif	dunif	qunif	runif
beta	pbeta	dbeta	qbeta	rbeta
wykładniczy	pexp	dexp	qexp	rexp
gamma	pgamma	dgamma	qgamma	rgamma
normalny	pnorm	dnorm	qnorm	rnorm
logarytmiczno-normalny	plnorm	dlnorm	qlnorm	rlnorm
Weibulla	pweibull	dweibull	qweibull	rweibull
chi-kwadrat	pchisq	dchisq	qchisq	rchisq
t-Studenta	pt	dt	qt	rt
Cauchy'ego	pcauchy	dcauchy	qcauchy	rcauchy
F-Snedecora	pf	df	qf	rf
Rayleigha	VGAM::prayleigh	VGAM::drayleigh	VGAM::qrayleigh	VGAM::rrayleigh

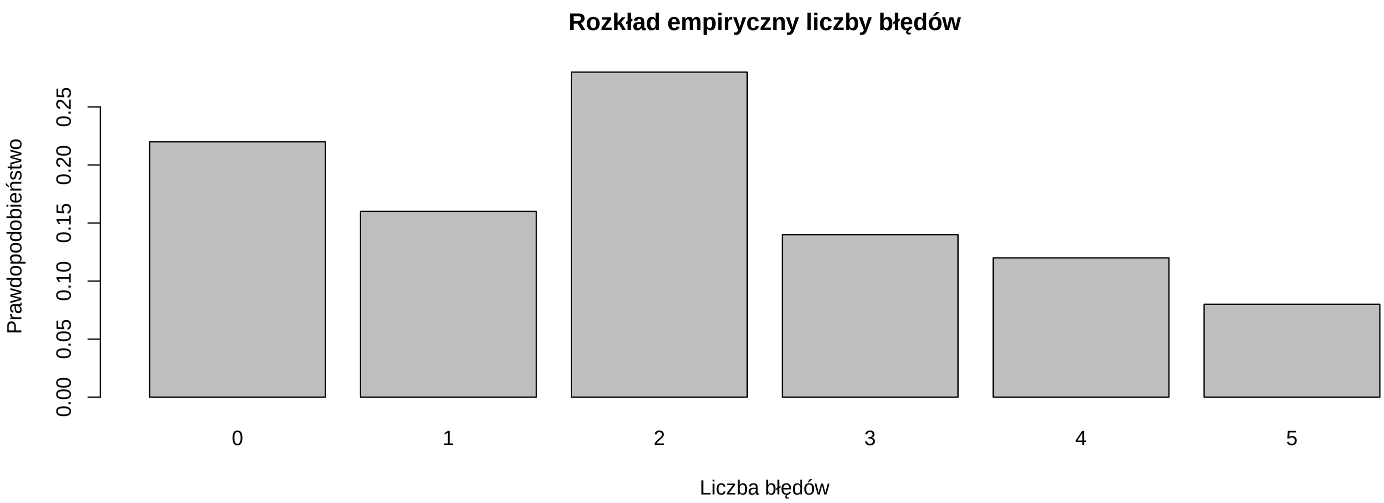
Przykład 1. Poniższe dane podają liczbę błędów w grupie 50 osób zdających egzamin testowy. Egzamin

składał się z 18 pytań (można popełnić maksymalnie dwa błędy, aby zdać egzamin).

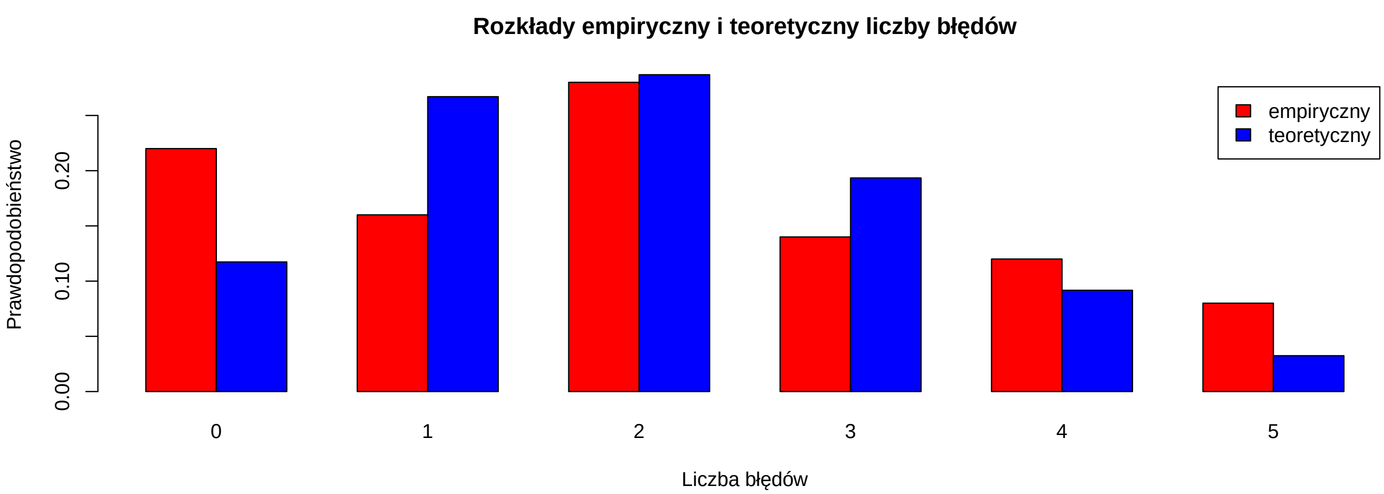
1	1	2	0	1	3	1	4	4	4	0	1	0	0	0	2	3
4	0	1	5	2	3	5	3	2	2	4	0	2	2	0	2	2
3	3	1	3	2	2	0	0	5	4	2	1	5	2	2	0	

Zmienna X to liczba błędów. Jest to dyskretna zmienna ilościowa.

```
liczba_bledow <- c(1, 1, 2, 0, 1, 3, 1, 4, 4, 4, 0, 1, 0, 0, 0, 2, 3,
                   4, 0, 1, 5, 2, 3, 5, 3, 2, 2, 4, 0, 2, 2, 0, 2, 2,
                   3, 3, 1, 3, 2, 2, 0, 0, 5, 4, 2, 1, 5, 2, 2, 0)
# wykres słupkowy
barplot(prop.table(table(liczba_bledow)),
        xlab = "Liczba błędów", ylab = "Prawdopodobieństwo",
        main = "Rozkład empiryczny liczby błędów")
```



- model: rozkład dwumianowy z $m = 18$
- $\mathcal{P} = \{b(18, p) : p \in (0, 1)\}$
- $\Theta = (0, 1)$ oraz $\theta = p$

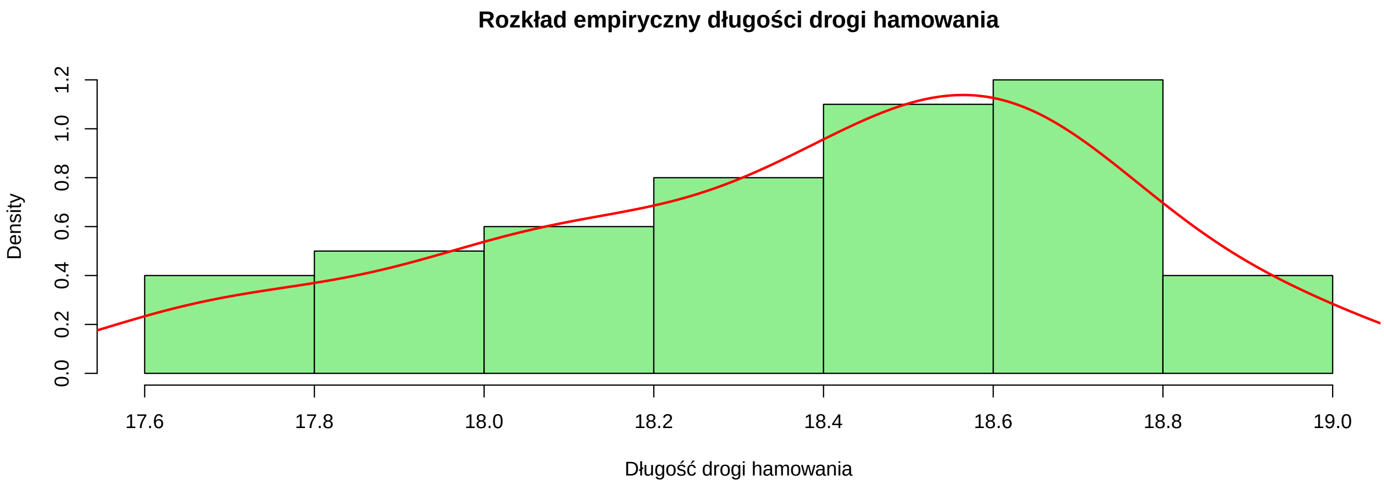


Przykład 2. Przeprowadzono 50 niezależnych eksperymentów obejmujących hamowanie pewnego typu samochodu (na suchym asfalcie, z prędkością 40km/h itp.). Notowano długość drogi hamowania w metrach z dokładnością do jednego centymetra. Uzyskane wyniki są zawarte w pliku hamulce.txt. Zmienna X to długość drogi hamowania. Jest to zmienna ilościowa ciągła.

```r
hamulce <- read.table("http://ls.home.amu.edu.pl/data_sets/hamulce.txt", dec = ",")
head(hamulce)
```

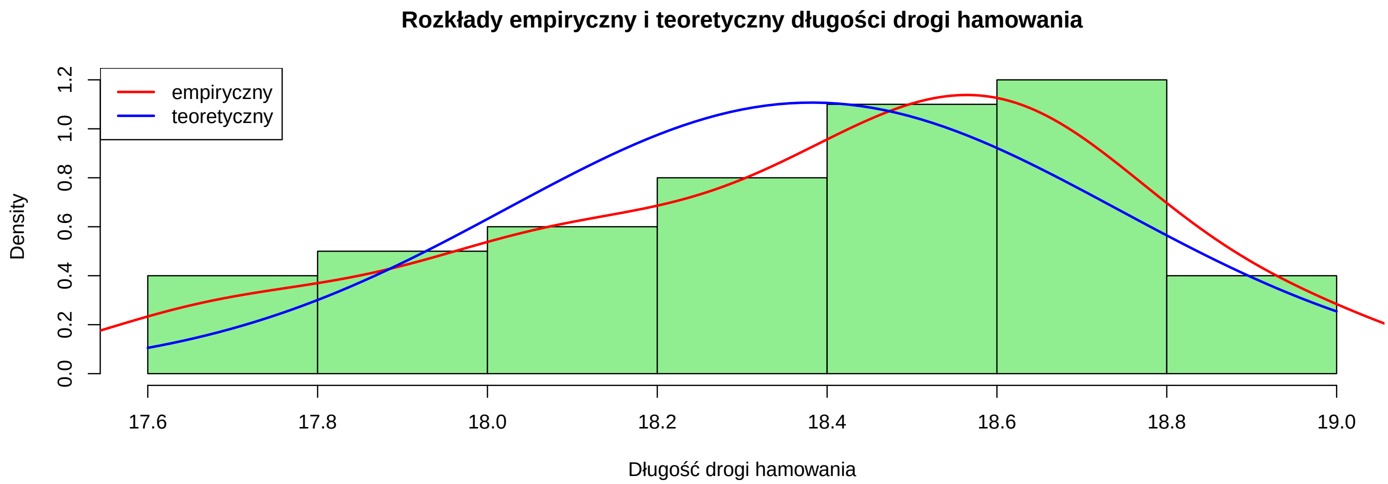
```
##      V1
## 1 18.66
## 2 17.81
## 3 18.96
## 4 18.09
## 5 18.73
## 6 18.45
```

```r
# histogram z estymatorem jądrowym gęstości
hist(hamulce$V1,
     xlab = "Długość drogi hamowania",
     main = "Rozkład empiryczny długości drogi hamowania",
     probability = TRUE,
     col = "lightgreen")
lines(density(hamulce$V1), col = "red", lwd = 2)
```



- model: rozkład normalny
- $\mathcal{P} = \{N(\mu, \sigma) : \mu \in \mathbb{R}, \sigma > 0\}$
- $\Theta = \mathbb{R} \times (0, \infty)$ oraz $\theta = (\mu, \sigma)$

```
##      V1
## 1 18.66
## 2 17.81
## 3 18.96
## 4 18.09
## 5 18.73
## 6 18.45
```



4.1 Estymacja punktowa

- W modelach parametrycznych występują pewne parametry, które są nieznane. Musimy je oszacować na podstawie próby. W tym celu wykorzystujemy estymatory (punktowe) i przedziały ufności.
- Niech $\mathbf{X} = (X_1, \ldots, X_n)^\top$ będzie próba prostą z populacji o rozkładzie P_θ, gdzie $\theta \in \Theta$ jest parametrem.
- Niech $g(\theta)$ będzie funkcją parametryczną, np. $g(\theta) = \theta$, $g(\theta) = \sqrt{\theta}$.

Definicja. Statystyką nazywamy funkcję mierzalną $T(\mathbf{X})$ próby $\mathbf{X}$.

- Statystyka jest zmienną lub wektorem losowym.

Definicja. Każda statystyka $T(\mathbf{X})$ o wartościach w zbiorze $g(\Theta)$ jest nazywana **estymatorem** funkcji parametrycznej $g(\theta)$ (oznaczamy również $\widehat{g}(\mathbf{X})$).

4.1.1 Metoda największej wiarogodności

- Niech rozkłady $P_\theta \in \mathcal{P}$ będą opisane za pomocą funkcji prawdopodobieństwa (gęstości) p_θ (f_θ) i niech $\theta \in \Theta \subseteq \mathbb{R}^d$.

Definicja. Funkcję L daną wzorem

$$L(\theta; \mathbf{x}) = p_\theta(\mathbf{x})$$

nazywamy **funkcją wiarogodności**.

- Funkcja wiarogodności jest funkcją parametru θ podczas, gdy obserwacje $\mathbf{x}$ są ustalone (oznaczamy to pisząc $\mathbf{x}$ jako argument funkcji L po średniku).

Definicja. Estymatorem największej wiarogodności (ENW) parametru θ jest statystyka $\widehat{\theta}(\mathbf{X})$ spełniająca następujący warunek

$$\forall \mathbf{x} \in \mathcal{X} : L\left(\widehat{\theta}(\mathbf{x}); \mathbf{x}\right) = \sup_{\theta \in \Theta} L(\theta; \mathbf{x}).$$

- Dla danego parametru θ, ENW może nie istnieć lub być wyznaczony niejednoznacznie.
- ENW funkcji parametrycznej $g(\theta)$ jest statystyka

  $$g(\widehat{\theta}(\mathbf{X})),$$

  gdzie $\widehat{\theta}(\mathbf{X}) = ENW(\theta)$.
- Zazwyczaj przy wyznaczaniu ENW łatwiej jest wykorzystać funkcję $\ln L$ niż funkcję L.

- W przypadku próby prostej mamy

$$L(\theta;\mathbf{x}) = p_\theta(\mathbf{x}) = \prod_{i=1}^{n} p_\theta(x_i),$$

gdzie $p_\theta(x)$ jest funkcją prawdopodobieństwa (gęstości) zmiennej X, a $\mathbf{x} = (x_1, \ldots, x_n)^\top$ jest wektorem zawierającym dane.

4.1.2 Estymator nieobciążony

Definicja. Estymator $\widehat{g}(\mathbf{X})$ nazywany jest estymatorem nieobciążonym (EN) funkcji parametrycznej $g(\theta)$, gdy

$$\forall \theta \in \Theta : E_\theta(\widehat{g}(\mathbf{X})) = g(\theta).$$

- Dla danej funkcji parametrycznej $g(\theta)$, zbiór estymatorów nieobciążonych może być pusty.

Twierdzenie. Załóżmy, że zmienna X populacji ma rozkład o wartości oczekiwanej μ. Niech $\mathbf{X} = (X_1, \ldots, X_n)^\top$ będzie próbą prostą z tej populacji. Wtedy statystyka

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

jest nieobciążonym estymatorem wartości oczekiwanej μ.

Twierdzenie. Załóżmy, że zmienna X populacji ma rozkład o wartości oczekiwanej μ oraz skończonej i niezerowej wariancji σ^2. Niech $\mathbf{X} = (X_1, \ldots, X_n)^\top$, $n > 1$ będzie próbą prostą z tej populacji. Wtedy statystyka

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

jest nieobciążonym estymatorem wariancji σ^2.

4.1.3 Estymator nieobciążony o minimalnej wariancji

Definicja. Niech A będzie niepustym zbiorem estymatorów nieobciążonych funkcji parametrycznej $g(\theta)$ o skończonej wariancji (dla każdego $\theta \in \Theta$). Statystyka $\widehat{g}_* \in A$ jest nazywana **estymatorem nieobciążonym o minimalnej wariancji** (ENMW) funkcji parametrycznej $g(\theta)$, gdy

$$\forall \widehat{g} \in A \ \forall \theta \in \Theta : Var_\theta(\widehat{g}_*) \leqslant Var_\theta(\widehat{g}).$$

Twierdzenie. Jeżeli istnieje estymator nieobciążony o minimalnej wariancji dla danej funkcji parametrycznej $g(\theta)$, to jest on wyznaczony jednoznacznie (z dokładnością do zbioru miary zero).

4.1.4 Przykłady estymatorów w wybranych rozkładach

Niech

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

$$\widetilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

- rozkład dwumianowy $b(m, p)$, $m \in \mathbb{N}$, $p \in (0, 1)$

$$ENW(p) = EN(p) = ENMW(p) = \frac{1}{m}\bar{X}$$

- rozkład Poissona $\pi(\lambda)$, $\lambda > 0$

$$ENW(\lambda) = EN(\lambda) = ENMW(\lambda) = \bar{X}$$

- rozkład jednostajny $U(0, \theta)$, $\theta > 0$

$$\begin{aligned} ENW(\theta) &= \max\{X_1, \dots, X_n\} \\ EN(\theta) = ENMW(\theta) &= \frac{n+1}{n}\max\{X_1, \dots, X_n\} \end{aligned}$$

- rozkład normalny $N(\mu, \sigma)$, $\mu \in \mathbb{R}$, $\sigma > 0$

$$\begin{aligned} ENW(\mu) = EN(\mu) = ENMW(\mu) &= \bar{X} \\ ENW(\sigma^2) &= \widetilde{S}^2 \\ EN(\sigma^2) = ENMW(\sigma^2) &= S^2 \\ ENW(\sigma) &= \widetilde{S} \\ EN(\sigma) = ENMW(\sigma) &= \sqrt{\frac{n-1}{2}}\frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}S \end{aligned}$$

- rozkład wykładniczy $Ex(\lambda)$, $\lambda > 0$

$$\begin{aligned} ENW(\lambda) &= \frac{1}{\bar{X}} \\ EN(\lambda) = ENMW(\lambda) &= \frac{n-1}{n\bar{X}} \end{aligned}$$

4.1.5 Wykres kwantyl-kwantyl

- Wykres kwantyl-kwantyl (Q-Q plot), jest wykresem zaobserwowanych statystyk porządkowych z losowej próby (kwantyle empiryczne) do odpowiadającym im (oszacowanym) wartościom średniej lub mediany w oparciu o założony rozkład lub w stosunku do empirycznych kwantyli innego zestawu danych.
- Wykresy kwantyl-kwantyl służą do oceny, czy dane pochodzą z określonego rozkładu lub czy dwa zestawy danych mają ten sam rozkład. Jeśli rozkłady mają ten sam kształt (ale niekoniecznie te same parametry położenia lub skali), wówczas wykres układa się mniej więcej na linii prostej. Jeśli rozkłady są dokładnie takie same, wówczas wykres układa się mniej więcej na linii prostej $y = x$.
- Najpierw wybiera się zbiór kwantyli pewnych rzędów. Punkt (x, y) na wykresie odpowiada jednemu z kwantyli drugiego rozkładu (współrzędna y) wykreślonemu względem kwantyla tego samego rzędu pierwszego rozkładu (współrzędna x).
- „qqline” dodaje linię do „teoretycznego” wykresu kwantyl-kwantyl, która przechodzi przez kwantyle rzędów `probs = c(0.25, 0.75)`, czyli domyślnie pierwszy i trzeci kwartyl.

Przykład 1 (cd.). Poniższe dane podają liczbę błędów w grupie 50 osób zdających egzamin testowy. Egzamin składał się z 18 pytań (można popełnić maksymalnie dwa błędy, aby zdać egzamin).

1	1	2	0	1	3	1	4	4	4	0	1	0	0	0	2	3
4	0	1	5	2	3	5	3	2	2	4	0	2	2	0	2	2
3	3	1	3	2	2	0	0	5	4	2	1	5	2	2	0	

Zmienna X to liczba błędów. Jest to dyskretna zmienna ilościowa.

- model: rozkład dwumianowy z $m = 18$
- $\mathcal{P} = \{b(18, p) : p \in (0, 1)\}$
- $\Theta = (0, 1)$ oraz $\theta = p$

```
liczba_bledow <- c(1, 1, 2, 0, 1, 3, 1, 4, 4, 4, 0, 1, 0, 0, 0, 2, 3,
                   4, 0, 1, 5, 2, 3, 5, 3, 2, 2, 4, 0, 2, 2, 0, 2, 2,
                   3, 3, 1, 3, 2, 2, 0, 0, 5, 4, 2, 1, 5, 2, 2, 0)
m <- 18
# estymator
(p_est <- mean(liczba_bledow) / m)
```

```
## [1] 0.1122222
```

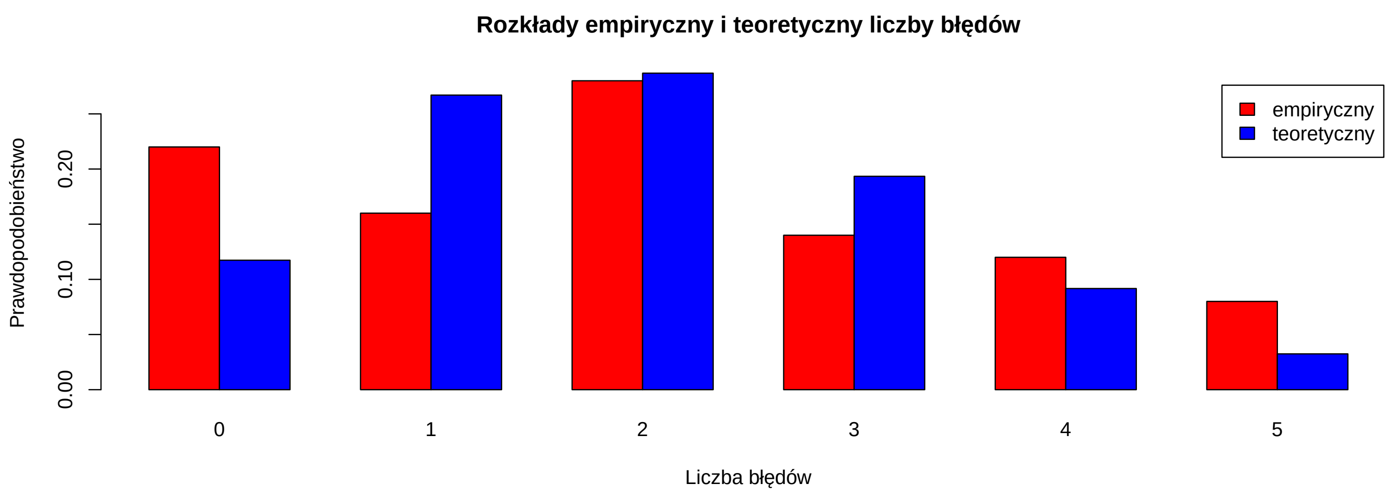
```
probs <- dbinom(sort(unique(liczba_bledow)), size = m, prob = p_est)
sum(probs)
```

```
## [1] 0.9887985
```

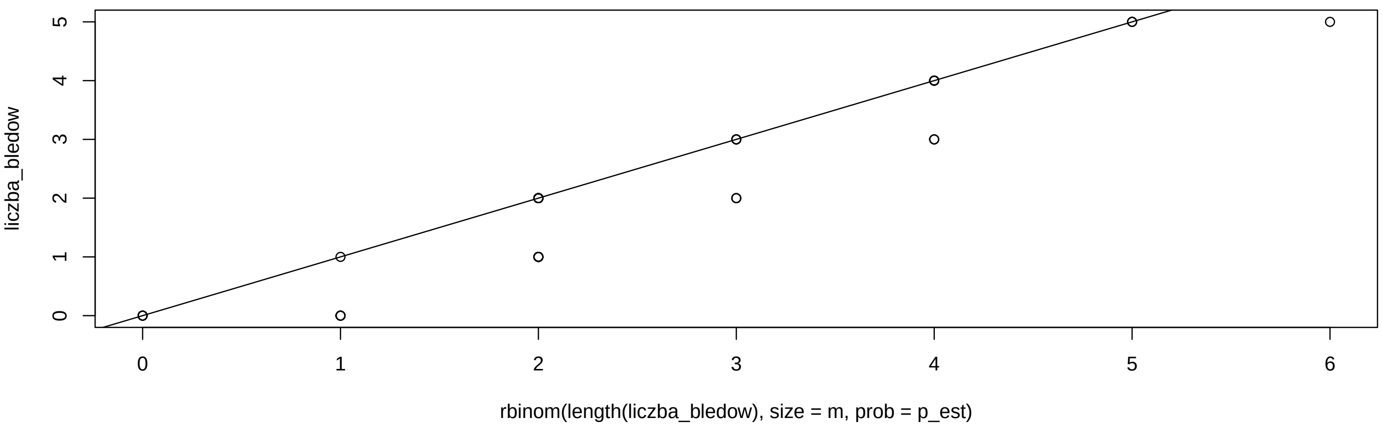
```
counts <- matrix(c(prop.table(table(liczba_bledow)), probs), nrow = 2, byrow = TRUE)
rownames(counts) <- c("empiryczny", "teoretyczny")
colnames(counts) <- sort(unique(liczba_bledow))
counts
```

```
##                     0         1         2         3          4          5
## empiryczny  0.2200000 0.1600000 0.2800000 0.1400000 0.12000000 0.08000000
## teoretyczny 0.1173483 0.2670078 0.2868914 0.1934153 0.09168466 0.03245109
```

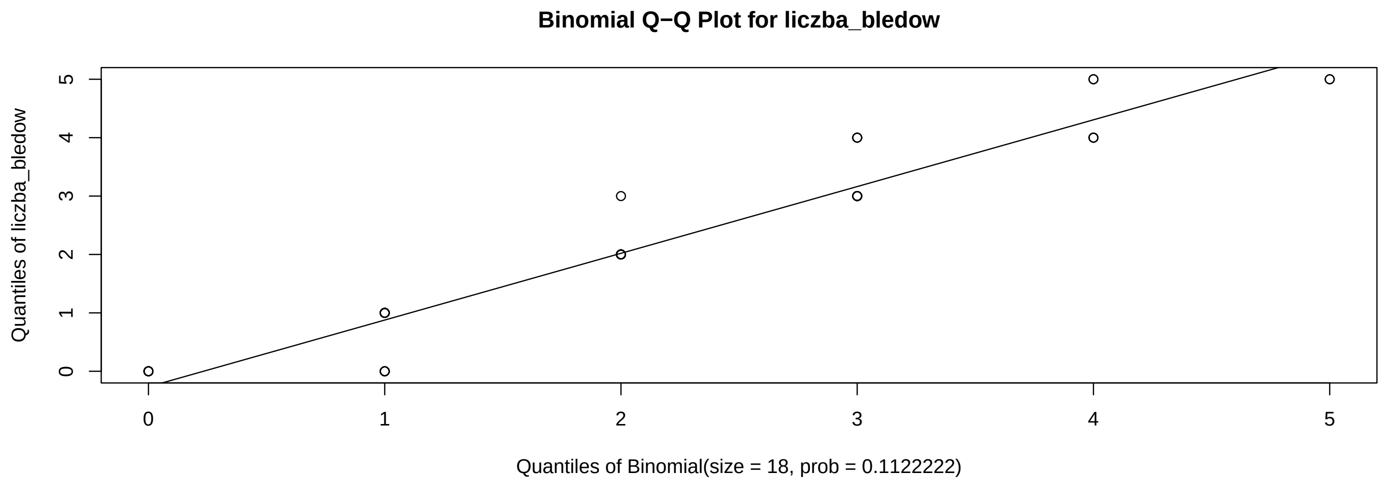
```
barplot(counts,
        xlab = "Liczba błędów", ylab = "Prawdopodobieństwo",
        main = "Rozkłady empiryczny i teoretyczny liczby błędów",
        col = c("red", "blue"), legend = rownames(counts), beside = TRUE)
```




```
# wykres kwantyl-kwantyl
qqplot(rbinom(length(liczba_bledow), size = m, prob = p_est), liczba_bledow)
qqline(liczba_bledow, distribution = function(probs) { qbinom(probs, size = m, prob = p_est) })
```




```
# lub
library(EnvStats)
qqPlot(liczba_bledow,
       distribution = "binom",
       param.list = list(size = m, prob = p_est),
       add.line = TRUE)
```



Przykład 2 (cd.). Przeprowadzono 50 niezależnych eksperymentów obejmujących hamowanie pewnego typu samochodu (na suchym asfalcie, z prędkością 40km/h itp.). Notowano długość drogi hamowania w metrach z dokładnością do jednego centymetra. Uzyskane wyniki są zawarte w pliku hamulce.txt. Zmienna X to długość drogi hamowania. Jest to zmienna ilościowa ciągła.

- model: rozkład normalny
- $\mathcal{P} = \{N(\mu, \sigma) : \mu \in \mathbb{R}, \sigma > 0\}$
- $\Theta = \mathbb{R} \times (0, \infty)$ oraz $\theta = (\mu, \sigma)$

```
hamulce <- read.table("http://ls.home.amu.edu.pl/data_sets/hamulce.txt", dec = ",")
# estymatory
(mu_est <- mean(hamulce$V1))
```

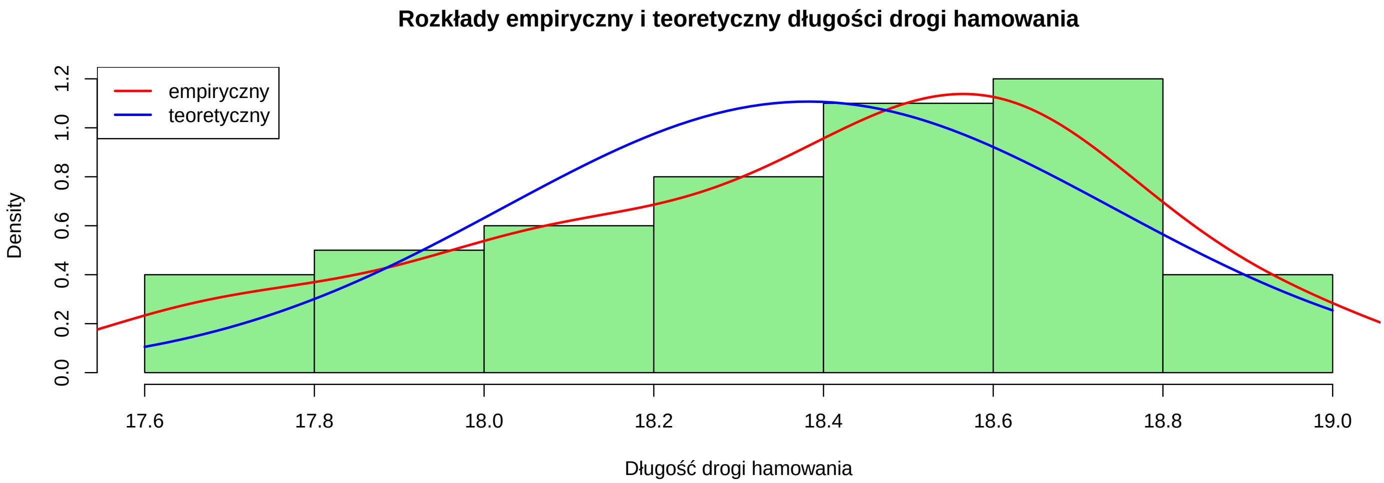
```
## [1] 18.3818
```

```
(sigma_est <- sd(hamulce$V1))
```

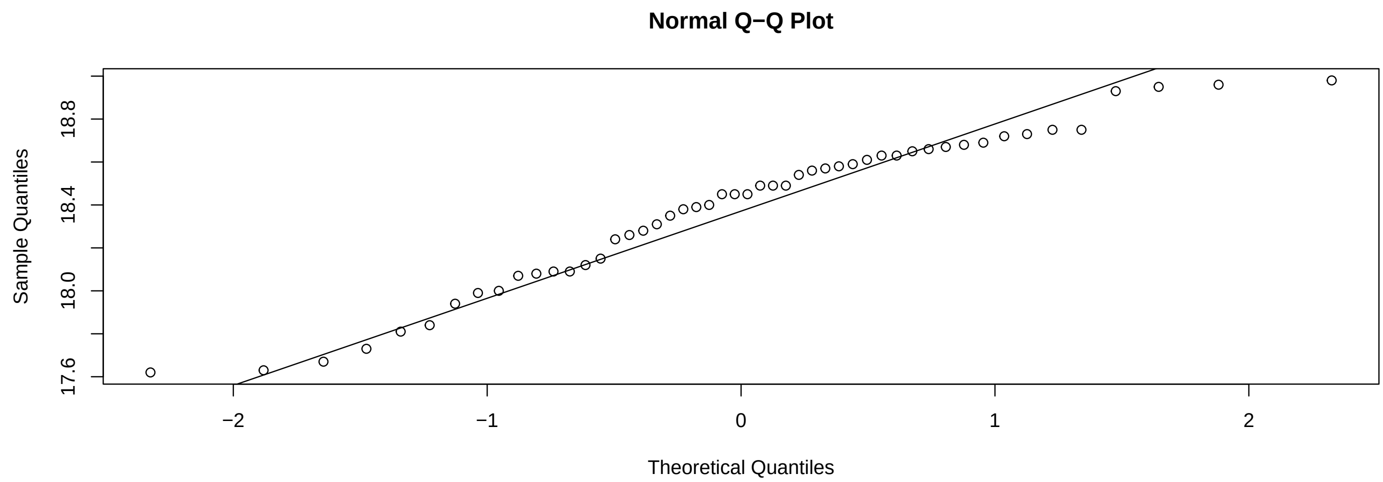
```
## [1] 0.3603439
```

```
# histogram z estymatorem jądrowym gęstości
hist(hamulce$V1,
     xlab = "Długość drogi hamowania",
```

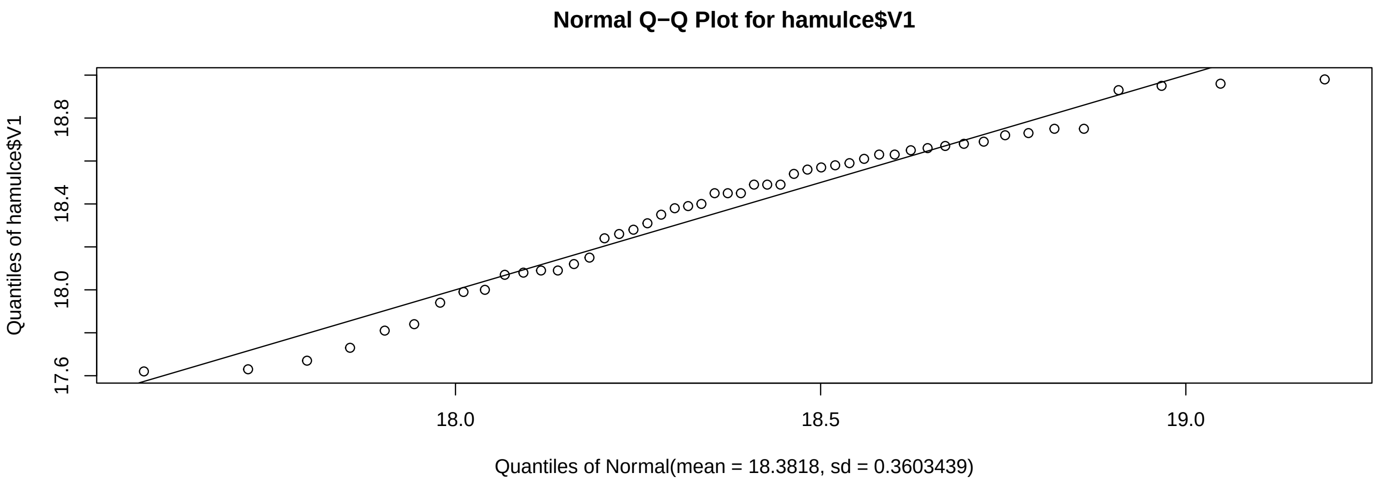
```r
     main = "Rozkłady empiryczny i teoretyczny długości drogi hamowania",
     probability = TRUE,
     col = "lightgreen")
lines(density(hamulce$V1), col = "red", lwd = 2)
curve(dnorm(x, mu_est, sigma_est),
      add = TRUE, col = "blue", lwd = 2)
legend("topleft", legend = c("empiryczny", "teoretyczny"), col = c("red", "blue"), lwd = 2)
```




```r
# wykres kwantyl-kwantyl
qqnorm(hamulce$V1)
qqline(hamulce$V1)
```




```r
# lub
library(EnvStats)
qqPlot(hamulce$V1,
       distribution = "norm",
       param.list = list(mean = mu_est, sd = sigma_est),
       add.line = TRUE)
```



- Empiryczne i teoretyczne prawdopodobieństwo, że droga hamowania jest większa niż 18,4, można obliczyć w następujący sposób:

```
# empirycznie
mean(hamulce$V1 > 18.4)
```

```
## [1] 0.54
```

```
# teoretycznie: X ~ N(mu_est, sigma_est) oraz P(X > 18.4) = 1 - P(X <= 18.4) = 1 - F(18.4)
1 - pnorm(18.4, mu_est, sigma_est)
```

```
## [1] 0.4798591
```

4.2 Przedziały ufności

- Niech $\mathbf{X} = (X_1, \dots, X_n)^\top$ będzie próbą prostą z populacji o rozkładzie P_θ, gdzie $\theta \in \Theta$ jest parametrem.
- Niech $g(\theta)$ będzie funkcją parametryczną.

Definicja. Przedział losowy

$$(T_1(\mathbf{X}), T_2(\mathbf{X}))$$

określony parą statystyk T_1 i T_2 takich, że

$$P_\theta(T_1 \leqslant T_2) = 1$$

dla każdego $\theta \in \Theta$ jest nazywany **przedziałem ufności** dla funkcji parametrycznej $g(\theta)$ na poziomie ufności $1 - \alpha$, $\alpha \in (0, 1)$, gdy

$$\forall \theta \in \Theta : P_\theta(T_1(\mathbf{X}) < g(\theta) < T_2(\mathbf{X})) \geqslant 1 - \alpha.$$

- Zazwyczaj $\alpha = 0{,}05$.

Przykłady przedziałów ufności dla wybranych rozkładów

- rozkład zero-jedynkowy $b(1, p)$, $p \in (0, 1)$

Dla parametru p, gdy próba jest duża (tj. $n \geqslant 100$)

$$\left(\bar{X} - \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} z\left(1 - \frac{\alpha}{2}\right), \bar{X} + \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} z\left(1 - \frac{\alpha}{2}\right) \right),$$

gdzie $z(\beta)$ oznacza kwantyl rzędu β z rozkładu normalnego $N(0, 1)$.

- rozkład Poissona $\pi(\lambda)$, $\lambda > 0$

Dla parametru λ, gdy próba jest duża (tj. $n \geqslant 100$)

$$\left(\bar{X} - \sqrt{\frac{\bar{X}}{n}} z\left(1 - \frac{\alpha}{2}\right), \bar{X} + \sqrt{\frac{\bar{X}}{n}} z\left(1 - \frac{\alpha}{2}\right)\right),$$

gdzie $z(\beta)$ oznacza kwantyl rzędu β z rozkładu normalnego $N(0, 1)$.

- rozkład jednostajny $U(0, \theta)$, $\theta > 0$

Dla parametru θ

$$\left(\frac{\max\{X_1, \dots, X_n\}}{\sqrt[n]{1 - \frac{\alpha}{2}}}, \frac{\max\{X_1, \dots, X_n\}}{\sqrt[n]{\frac{\alpha}{2}}}\right).$$

- rozkład normalny $N(\mu, \sigma)$, $\mu \in \mathbb{R}$, $\sigma > 0$

Dla parametru μ

$$\left(\bar{X} - \frac{S}{\sqrt{n}} t\left(1 - \frac{\alpha}{2}, n-1\right), \bar{X} + \frac{S}{\sqrt{n}} t\left(1 - \frac{\alpha}{2}, n-1\right)\right),$$

gdzie $t(\beta, m)$ oznacza kwantyl rzędu β z rozkładu t-Studenta $t(m)$ z m stopniami swobody.

Dla parametru σ^2

$$\left(\frac{(n-1)S^2}{\chi^2\left(1 - \frac{\alpha}{2}, n-1\right)}, \frac{(n-1)S^2}{\chi^2\left(\frac{\alpha}{2}, n-1\right)}\right),$$

gdzie $\chi^2(\beta, m)$ oznacza kwantyl rzędu β z rozkładu chi-kwadrat $\chi^2(m)$ z m stopniami swobody.

Dla parametru σ

$$\left(\frac{\sqrt{n-1}S}{\sqrt{\chi^2\left(1 - \frac{\alpha}{2}, n-1\right)}}, \frac{\sqrt{n-1}S}{\sqrt{\chi^2\left(\frac{\alpha}{2}, n-1\right)}}\right),$$

gdzie $\chi^2(\beta, m)$ oznacza kwantyl rzędu β z rozkładu chi-kwadrat $\chi^2(m)$ z m stopniami swobody.

- rozkład wykładniczy $Ex(\lambda)$, $\lambda > 0$

Dla parametru λ

$$\left(\frac{\chi^2\left(\frac{\alpha}{2}, 2n\right)}{2n\bar{X}}, \frac{\chi^2\left(1 - \frac{\alpha}{2}, 2n\right)}{2n\bar{X}}\right),$$

gdzie $\chi^2(\beta, m)$ oznacza kwantyl rzędu β z rozkładu chi-kwadrat $\chi^2(m)$ z m stopniami swobody.

Przykład 2 (cd.). Przeprowadzono 50 niezależnych eksperymentów obejmujących hamowanie pewnego typu samochodu (na suchym asfalcie, z prędkością 40km/h itp.). Notowano długość drogi hamowania w metrach z dokładnością do jednego centymetra. Uzyskane wyniki są zawarte w pliku hamulce.txt. Zmienna X to długość drogi hamowania. Jest to zmienna ilościowa ciągła.

- model: rozkład normalny
- $\mathcal{P} = \{N(\mu, \sigma) : \mu \in \mathbb{R}, \sigma > 0\}$
- $\Theta = \mathbb{R} \times (0, \infty)$ oraz $\theta = (\mu, \sigma)$

```
hamulce <- read.table("http://ls.home.amu.edu.pl/data_sets/hamulce.txt", dec = ",")
# estymatory
(mu_est <- mean(hamulce$V1))
```

```
## [1] 18.3818
```

```
(sigma_est <- sd(hamulce$V1))
```

```
## [1] 0.3603439
```

```r
# przedziały ufności
library(EnvStats)
enorm(hamulce$V1,
      method = "mvue",
      ci = TRUE, ci.type = "two-sided", conf.level = 0.95, ci.param = "mean")
```

```
## $distribution
## [1] "Normal"
##
## $sample.size
## [1] 50
##
## $parameters
##       mean         sd
## 18.3818000  0.3603439
##
## $n.param.est
## [1] 2
##
## $method
## [1] "mvue"
##
## $data.name
## [1] "hamulce$V1"
##
## $bad.obs
## [1] 0
##
## $interval
## $name
## [1] "Confidence"
##
## $parameter
## [1] "mean"
##
## $limits
##      LCL      UCL
## 18.27939 18.48421
##
## $type
## [1] "two-sided"
##
## $method
## [1] "Exact"
##
## $conf.level
## [1] 0.95
##
## $sample.size
## [1] 50
##
```

```
## $dof
## [1] 49
##
## attr(,"class")
## [1] "intervalEstimate"
##
## attr(,"class")
## [1] "estimate"
```

```
enorm(hamulce$V1,
      method = "mvue",
      ci = TRUE, ci.type = "two-sided", conf.level = 0.95, ci.param = "variance")
```

```
## $distribution
## [1] "Normal"
##
## $sample.size
## [1] 50
##
## $parameters
##       mean         sd
## 18.3818000  0.3603439
##
## $n.param.est
## [1] 2
##
## $method
## [1] "mvue"
##
## $data.name
## [1] "hamulce$V1"
##
## $bad.obs
## [1] 0
##
## $interval
## $name
## [1] "Confidence"
##
## $parameter
## [1] "variance"
##
## $limits
##        LCL        UCL
## 0.09060552 0.20163381
##
## $type
## [1] "two-sided"
##
## $method
## [1] "Exact"
##
```

```
## $conf.level
## [1] 0.95
##
## $sample.size
## [1] 50
##
## $dof
## [1] 49
##
## attr(,"class")
## [1] "intervalEstimate"
##
## attr(,"class")
## [1] "estimate"
```

```
# Uwaga. Powyższy estymator odchylenia standardowego nie jest ENMW!
# jednostronne przedziały ufności
enorm(hamulce$V1,
      method = "mvue",
      ci = TRUE, ci.type = "lower", conf.level = 0.95, ci.param = "mean")
```

```
## $distribution
## [1] "Normal"
##
## $sample.size
## [1] 50
##
## $parameters
##       mean         sd
## 18.3818000  0.3603439
##
## $n.param.est
## [1] 2
##
## $method
## [1] "mvue"
##
## $data.name
## [1] "hamulce$V1"
##
## $bad.obs
## [1] 0
##
## $interval
## $name
## [1] "Confidence"
##
## $parameter
## [1] "mean"
##
## $limits
##      LCL      UCL
```

```
## 18.29636      Inf
##
## $type
## [1] "lower"
##
## $method
## [1] "Exact"
##
## $conf.level
## [1] 0.95
##
## $sample.size
## [1] 50
##
## $dof
## [1] 49
##
## attr(,"class")
## [1] "intervalEstimate"
##
## attr(,"class")
## [1] "estimate"
```

```
enorm(hamulce$V1,
      method = "mvue",
      ci = TRUE, ci.type = "upper", conf.level = 0.95, ci.param = "mean")
```

```
## $distribution
## [1] "Normal"
##
## $sample.size
## [1] 50
##
## $parameters
##       mean        sd
## 18.3818000 0.3603439
##
## $n.param.est
## [1] 2
##
## $method
## [1] "mvue"
##
## $data.name
## [1] "hamulce$V1"
##
## $bad.obs
## [1] 0
##
## $interval
## $name
## [1] "Confidence"
```

```
##
## $parameter
## [1] "mean"
##
## $limits
##      LCL      UCL
##     -Inf 18.46724
##
## $type
## [1] "upper"
##
## $method
## [1] "Exact"
##
## $conf.level
## [1] 0.95
##
## $sample.size
## [1] 50
##
## $dof
## [1] 49
##
## attr(,"class")
## [1] "intervalEstimate"
##
## attr(,"class")
## [1] "estimate"
```

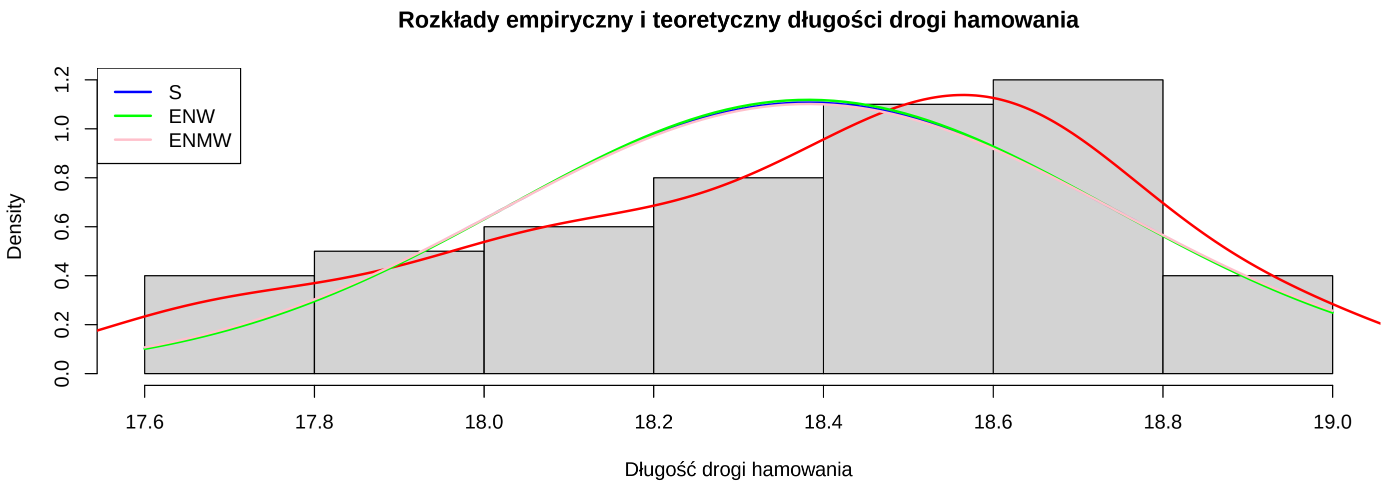
4.3 Zadania 4

Zadanie 1. Dla danych dotyczących długości drogi hamowania w przykładzie 2 z wykładu, biorąc pod uwagę przyjęty rozkład teoretyczny, obliczyć wartości trzech estymatorów odchylenia standardowego. Zilustruj otrzymane trzy teoretyczne funkcje gęstości na histogramie.

```
## [1] 0.3603439
```

```
## [1] 0.3567222
```

```
## [1] 0.3621869
```



Zadanie 2. Przebadano 200 losowo wybranych 5-sekundowych okresów pracy centrali telefonicznej. Rejestrowano liczbę zgłoszeń. Wyniki są zawarte w pliku Centrala.RData.

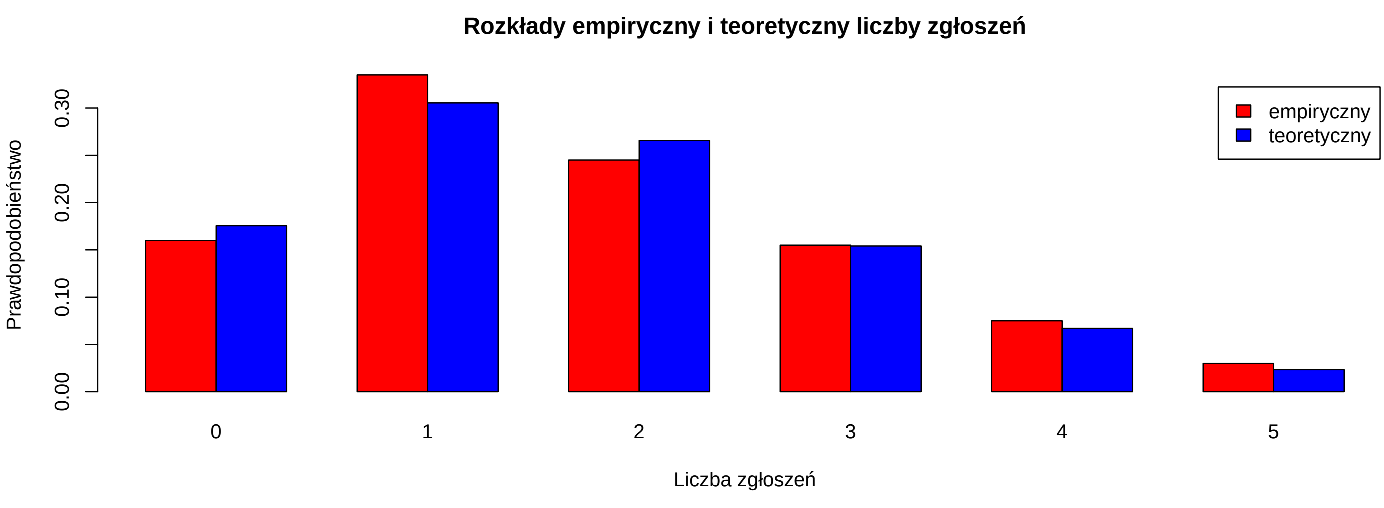
1. Zasugeruj rozkład teoretyczny badanej zmiennej.
2. Oblicz wartość estymatora parametru rozkładu teoretycznego.

```
## [1] 1.74
```

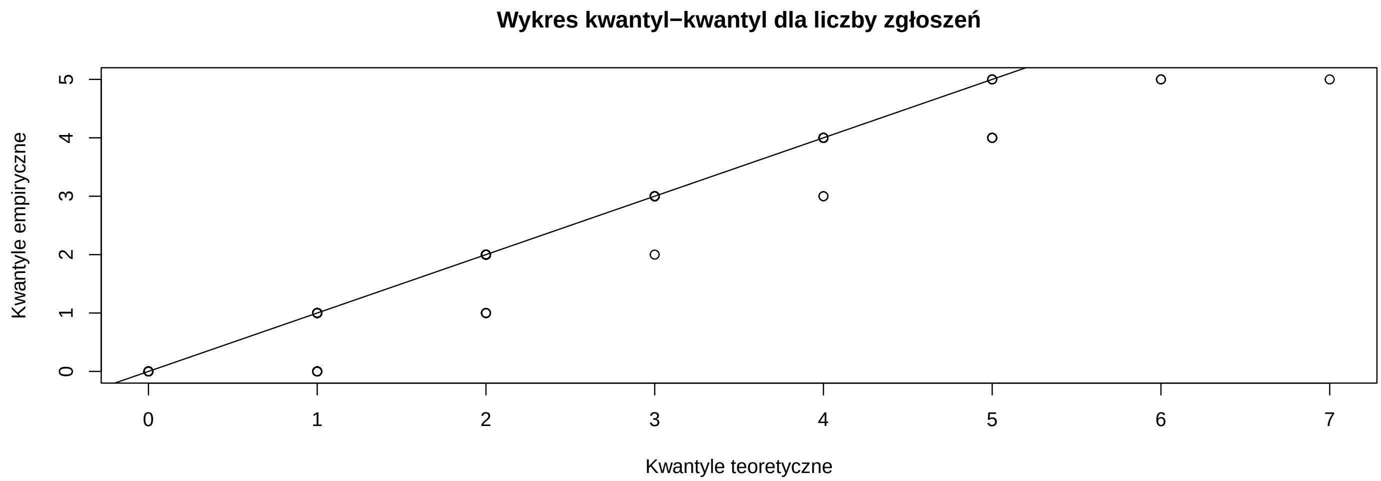
3. Porównaj empiryczne prawdopodobieństwa wystąpienia poszczególnych wartości liczby zgłoszeń w próbie z wartościami teoretycznymi uzyskanymi na podstawie rozkładu teoretycznego.

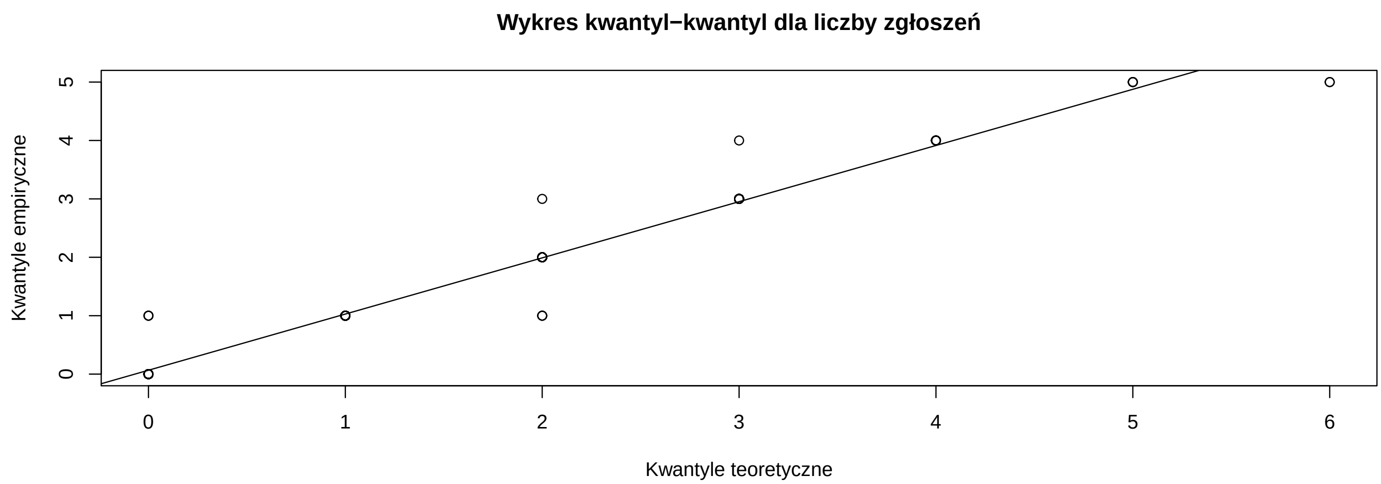
```
## [1] 0.9911019
```

```
##                     0         1         2         3          4          5
## empiryczny  0.1600000 0.3350000 0.2450000 0.1550000 0.07500000 0.03000000
## teoretyczny 0.1755204 0.3054055 0.2657028 0.1541076 0.06703681 0.02332881
```



4. Sprawdź dopasowanie rozkładu teoretycznego za pomocą wykresy kwantyl-kwantyl.





5. Czy na podstawie powyższych rozważań rozkład teoretyczny wydaje się odpowiedni?
6. Oblicz prawdopodobieństwo empiryczne i teoretyczne, że liczba zgłoszeń jest mniejsza niż 4.

```
## [1] 0.895
```

```
## [1] 0.9007363
```

7. Wyznacz (trzema metodami) przedział ufności dla parametru rozkładu teoretycznego.

```
##      LCL      UCL
## 1.561968 1.932765
```

```
##      LCL      UCL
## 1.561968 1.932765
```

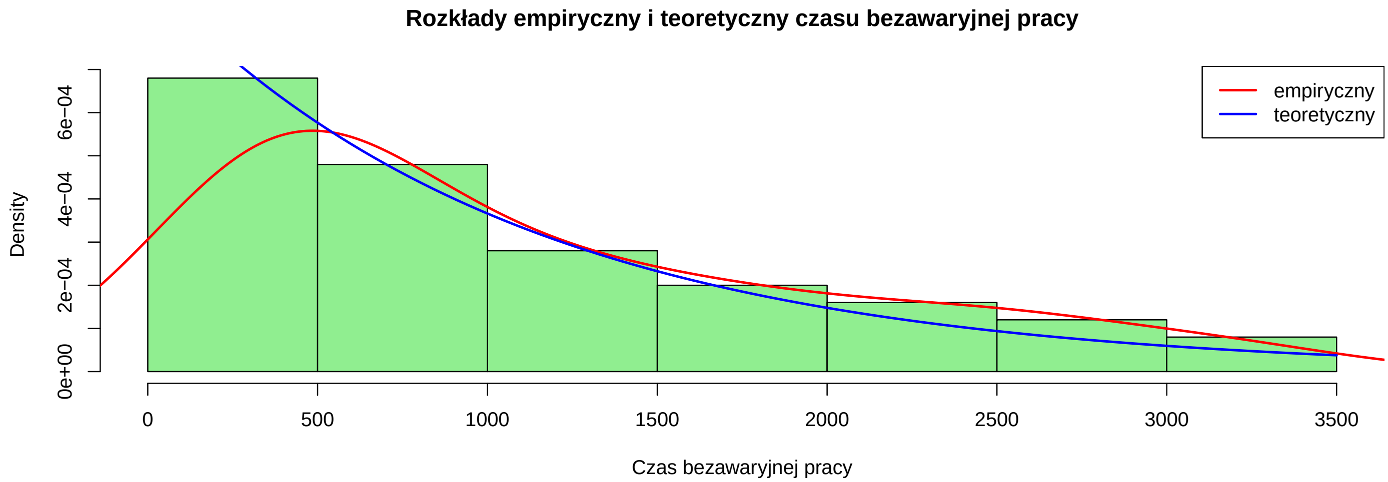
```
##      LCL      UCL
## 1.557187 1.922813
```

Zadanie 3. Zmienna w pliku awarie.txt opisuje wyniki 50 pomiarów czasu bezawaryjnej pracy danego urządzenia (w godzinach).

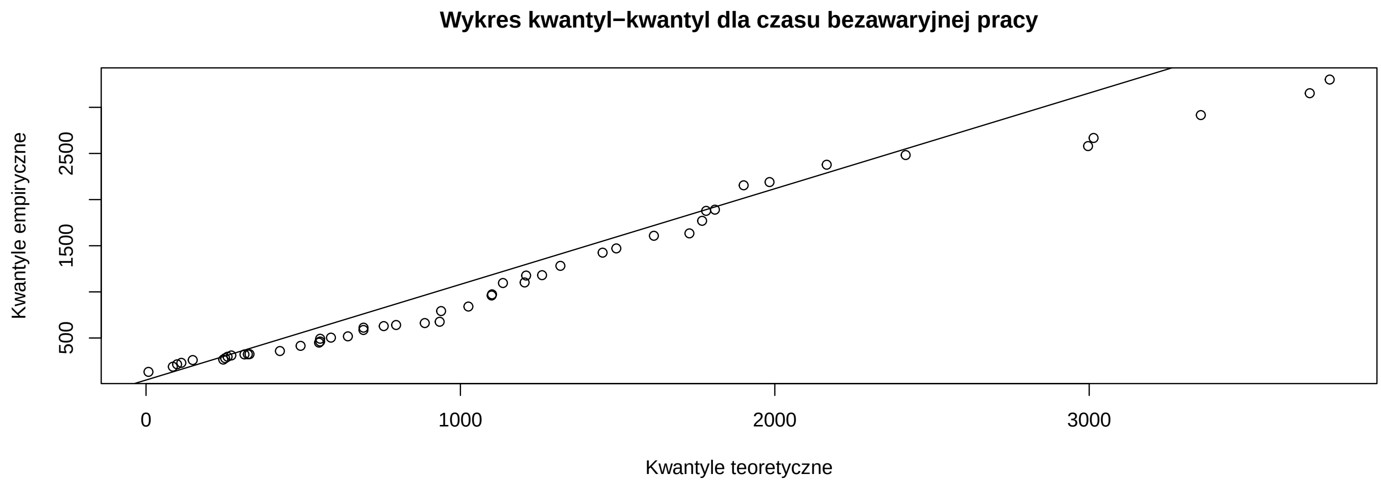
1. Zasugeruj rozkład teoretyczny badanej zmiennej.
2. Oblicz wartość ENW parametru rozkładu teoretycznego.

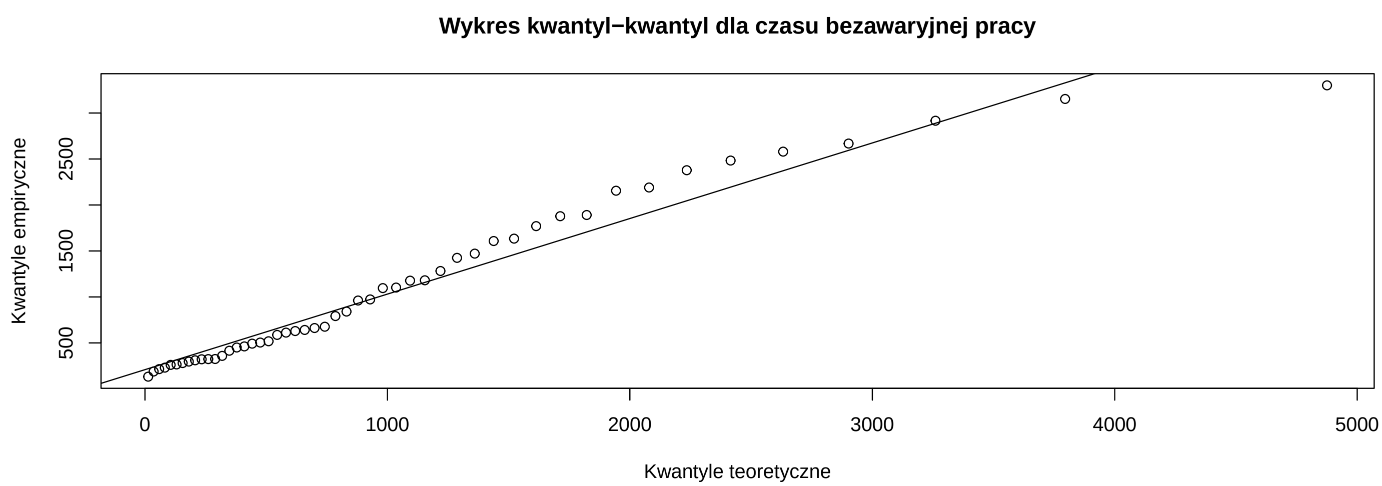
```
## [1] 0.0009079683
```

3. Porównaj rozkład empiryczny wystąpienia poszczególnych wartości czasu bezawaryjnej pracy w próbie z wartościami teoretycznymi uzyskanymi na podstawie rozkładu teoretycznego.



4. Sprawdź dopasowanie rozkładu teoretycznego za pomocą wykresy kwantyl-kwantyl.





5. Czy na podstawie powyższych rozważań rozkład teoretyczny wydaje się odpowiedni?
6. Oblicz empiryczne i teoretyczne prawdopodobieństwo, że czas bezawaryjnej pracy zawarty jest w przedziale $[1000, 1500]$.

```
## [1] 0.14
```

```
## [1] 0.1471827
```

7. Wyznacz przedział ufności dla parametru rozkładu teoretycznego.

```
##          LCL          UCL
## 0.0006739116 0.0011763746
```

8. Oblicz wartość ENW i granice przedziału ufności dla wartości oczekiwanej i wariancji rozkładu teoretycznego.

```
## [1] 1101.36
```

```
## [1] 1212994
```

```
##       UCL       LCL
##  850.0693 1483.8742
```

```
##       UCL       LCL
##  722617.9 2201882.5
```

Zadanie 4. Niech $\mathbf{X} = (X_1, \dots, X_n)^\top$ będzie próbą prostą z populacji o rozkładzie Rayleigha $R(\lambda)$, $\lambda > 0$.

1. Napisz funkcję `median_cint()`, która implementuje następujący przybliżony przedział ufności dla mediany $\sqrt{\lambda \ln 2}$ tego rozkładu:

$$\left(\sqrt{\ln(2)\frac{1}{n}\sum_{i=1}^{n} X_i^2 \left(1 - \frac{z(1-\alpha/2)}{\sqrt{n}}\right)}, \sqrt{\ln(2)\frac{1}{n}\sum_{i=1}^{n} X_i^2 \left(1 + \frac{z(1-\alpha/2)}{\sqrt{n}}\right)}\right),$$

   gdzie $z(\beta)$ oznacza kwantyl rzędu β z rozkładu normalnego $N(0,1)$. Funkcja ta powinna mieć dwa argumenty: `x` - wektor zawierający dane, `conf_level` - poziom ufności. Funkcja zwraca obiekt typu `list` klasy `confint` o następujących elementach: `title` - nazwa estymowanej funkcji parametrycznej, `est` - wartość ENW funkcji parametrycznej

$$ENW(\sqrt{\lambda \ln 2}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2 \ln 2},$$

   `l` - lewy kraniec przedziału ufności, `r` - prawy kraniec przedziału ufności, `conf_level` - poziom ufności.

2. Następujące dane to pomiary średniej szybkości wiatru w odstępach 15 minutowych odnotowane wokół nowo powstającej elektrowni wiatrowej:

0.9	6.2	2.1	4.1	7.3
1.0	4.6	6.4	3.8	5.0
2.7	9.2	5.9	7.4	3.0
4.9	8.2	5.0	1.2	10.1
12.2	2.8	5.9	8.2	0.5

   Teoretyczny rozkład średniej szybkości wiatru to rozkład Rayleigha $R(\lambda)$, $\lambda > 0$. Używając funkcji `median_cint()`, oblicz wartość ENW i krańce 95% przedziału ufności dla mediany średniej szybkości wiatru. **Wskazówka:** Przed wywołaniem funkcji `median_cint()`, najpierw załaduj następujące funkcje przeciążone `print()` i `summary()`:

```r
print.confint <- function(x) {
  cat(x$conf_level * 100, "percent confidence interval:", "\n")
  cat(x$l, " ", x$r, "\n")
}

summary.confint <- function(x) {
  cat("\n", "Confidence interval of", x$title, "\n", "\n")
  cat(x$conf_level * 100, "percent confidence interval:", "\n")
  cat(x$l, " ", x$r, "\n")
```

```
  cat("sample estimate", "\n")
  cat(x$est, "\n")
}
```

```
## 95 percent confidence interval:
## 3.863593   5.845955
```

```
##
##  Confidence interval of mediana
##
## 95 percent confidence interval:
## 3.863593   5.845955
## sample estimate
## 4.954924
```

Zadanie 5. Dla danego wektora obserwacji i poziomu ufności napisz funkcję określającą granice przedziału ufności na poziomie ufności $1-\alpha$, $\alpha \in (0,1)$ dla wartości oczekiwanej w rozkładzie normalnym. Domyślny poziom ufności powinien wynosić 0,95. Następnie przeprowadź symulacje (z liczbą powtórzeń `nr = 1000`) sprawdzając prawdopodobieństwo pokrycia tego przedziału ufności (tj. prawdopodobieństwo, że ten przedział ufności zawiera wartość oczekiwaną) dla rozkładów $N(1,3)$, $\chi^2(3)$ i $Ex(3)$ osobno. Rozważ liczby obserwacji $n = 10, 50, 100$. Zinterpretuj wyniki. **Wskazówka:** Symulacja powinna przebiegać według następujących kroków:

1. Przyjmij poziom istotności, n, `nr`, rozkład generowanych danych oraz `temp = 0`.
2. Wygeneruj n obserwacji z zadanego rozkładu.
3. Wyznacz granice przedziału ufności dla danych wygenerowanych w kroku 2.
4. Jeśli teoretyczna wartość oczekiwana należy do przedziału otrzymanego w kroku 3, zwiększ `temp` o jeden.
5. Powtórz kroki 2-4 `nr` razy.
6. Wyznacz `temp / nr`.

```
## n = 10
```

```
## [1] 0.959
```

```
## [1] 0.901
```

```
## [1] 0.899
```

```
## n = 50
```

```
## [1] 0.941
```

```
## [1] 0.944
```

```
## [1] 0.941
```

```
## n = 100
```

```
## [1] 0.946
```

```
## [1] 0.942
```

```
## [1] 0.946
```

5 Testowanie hipotez statystycznych

- W przypadku estymacji zaczynamy od wyników próby i wykorzystujemy je do formułowania wniosków na temat całej populacji. Z drugiej strony, testując hipotezy statystyczne, najpierw przyjmujemy pewne założenia dotyczące ogólnej populacji (hipotezy), a następnie sprawdzamy je na podstawie próby.

5.1 Hipotezy statystyczne

- Hipoteza statystyczna jest pewnym przypuszczeniem dotyczącym populacji wydanych bez szczegółowych badań i weryfikacji. Chcemy sprawdzić, czy wyniki uzyskane dla próbki można zastosować do całej populacji.
- Testowana hipoteza jest nazywana hipotezą zerową i oznaczamy ją symbolem H_0.
- Każda dopuszczalna hipoteza (za wyjątkiem hipotezy zerowej) jest nazywana hipotezą alternatywną i oznaczana przez H_1. Jest to hipoteza, którą chcielibyśmy przyjąć, gdy odrzucimy hipotezę zerową.
- Hipotezy mogą być związane z:
  1. postacią funkcyjną rozkładu (hipotezy nieparametryczne),
  2. wartościami parametrów rozkładów (hipotezy parametryczne).
- Niech $\mathbf{X} = (X_1, \dots, X_n)^\top$ będzie próbą z populacji o rozkładzie $P_\theta \in \mathcal{P} = \{P_\theta : \theta \in \Theta\}$.
  1. $H_0 : P_\theta \in \mathcal{P}_0 \subsetneq \mathcal{P}$ przeciwko $H_1 : P_\theta \in \mathcal{P}_1 \subsetneq \mathcal{P}$
  2. $H_0 : \theta \in \Theta_0 \subsetneq \Theta$ przeciwko $H_1 : \theta \in \Theta_1 \subsetneq \Theta$, $\Theta_0 \cap \Theta_1 = \emptyset$

5.2 Test statystyczny

- Weryfikujemy (testujemy) hipotezy statystyczne, porównując wyniki z próbki z tym co przedstawiają hipotezy. W tym celu stosuje się testy statystyczne.
- Test statystyczny jest pewną regułą lub procedurą związaną z próbą losową i podaje decyzję dotyczącą przyjęcia lub odrzucenia hipotezy zerowej.
- Formalnie test statystyczny jest statystyką postaci

$$\phi : \mathcal{X} \to \{0, 1\}$$

  taką, że

$$\phi(\mathbf{x}) = I_R(\mathbf{x}) = \begin{cases} 1, & \mathbf{x} \in R, \\ 0, & \mathbf{x} \notin R, \end{cases}$$

  gdzie R jest obszarem krytycznym oraz

  - 1 ($\mathbf{x} \in R$) oznacza decyzję „odrzucamy H_0”, tj. stwierdzamy występowanie statystycznie istotnych różnic,
  - 0 ($\mathbf{x} \notin R$) oznacza decyzję „brak podstaw do odrzucenia H_0”, tj. nie stwierdzamy występowania statystycznie istotnych różnic.

- Ponieważ decyzja jest oparta na próbie losowej, nie jest ona zawsze poprawna. Możemy popełnić dwa błędy:

  - błąd pierwszego rodzaju - odrzucamy H_0, gdy jest ona prawdziwa,
  - błąd drugiego rodzaju - nie odrzucamy H_0, gdy jest ona fałszywa.

	Hipoteza zerowa jest	
Decyzja	prawdziwa	fałszywa
brak podstaw do odrzucenia H0	decyzja poprawna	błąd drugiego rodzaju
odrzucamy H0	błąd pierwszego rodzaju	decyzja poprawna

Uwaga. Zmniejszenie prawdopodobieństwa wystąpienia błędu pierwszego rodzaju zwiększa prawdopodobieństwo wystąpienia błędu drugiego rodzaju i na odwrót.

- Zatem konstruując obszar krytyczny R:

  1. zakładamy, że prawdopodobieństwo popełnienia błędu pierwszego rodzaju nie może być większe niż zadany poziom istotności α $(0 < \alpha < 1)$ testu, tj.
  $$P_{H_0}(\mathbf{X} \in R) \leqslant \alpha.$$
  Zazwyczaj przyjmujemy $\alpha = 0{,}05$.
  2. staramy się aby prawdopodobieństwo popełnienia błędu drugiego rodzaju
  $$P_{H_1}(\mathbf{X} \notin R)$$
  było minimalne.

- Punkt drugi powyższej konstrukcji może być sformułowany równoważnie w następujący sposób:

  2. staramy się aby moc testu
  $$P_{H_1}(\mathbf{X} \in R) = 1 - P_{H_1}(\mathbf{X} \notin R)$$
  (tj. prawdopodobieństwo niepopełnienia błędu drugiego rodzaju) było maksymalne.

- Obszar krytyczny testu może mieć jedną z następujących postaci:

  (a) $R = \{\mathbf{x} \in \mathcal{X} : T(\mathbf{x}) \geqslant c_{1-\alpha}\}$
  (b) $R = \{\mathbf{x} \in \mathcal{X} : T(\mathbf{x}) \leqslant c_{\alpha}\}$
  (c) $R = \{\mathbf{x} \in \mathcal{X} : T(\mathbf{x}) \geqslant c_{1-\alpha/2} \textit{ lub } T(\mathbf{x}) \leqslant c_{\alpha/2}\}$

gdzie $T(\mathbf{X})$ jest statystyką testową, a $c_{\cdot}$ jest wartością krytyczną (stałą).

Definicja. p-wartość jest najmniejszym poziomem istotności testu, przy którym odrzucamy hipotezę zerową H_0.

- Jeżeli p-wartość jest mniejsza lub równa poziomowi istotności α, to odrzucamy H_0.
- Jeżeli p-wartość jest większa niż poziom istotności α, to nie mamy podstaw do odrzucenia H_0.
- p-wartość oblicza się według wzorów:

  (a) $P_{H_0}(T \geqslant T(\mathbf{x}))$
  (b) $P_{H_0}(T \leqslant T(\mathbf{x}))$
  (c) $2\min\{P_{H_0}(T \geqslant T(\mathbf{x})), P_{H_0}(T \leqslant T(\mathbf{x}))\}$

- Podsumowanie procedury testowej:

  1. Dane $\mathbf{x} = (x_1, \dots, x_n)^\top$.
  2. Obierz hipotezy.
  3. Wybierz test statystyczny i ustal poziom istotności.
  4. Oblicz p-wartość.
  5. Porównaj p-wartość z poziomem istotności.
  6. Podejmij decyzję.

5.3 Wybrane testy statystyczne

5.3.1 Test normalności Shapiro-Wilka

- Niech $\mathbf{X} = (X_1, \dots, X_n)^\top$ będzie próbą z populacji o rozkładzie $P_\theta \in \mathcal{P} = \{P_\theta : \theta \in \Theta\}$.

- Chcemy testować następującą hipotezę zerową:
$$H_0 : \mathcal{P} = \{N(\mu, \sigma) : \mu \in \mathbb{R}, \sigma > 0\}$$
przeciwko
$$H_1 : \neg H_0.$$
- Statystyka testowa jest postaci:
$$W(\mathbf{X}) = \frac{\left(\sum_{i=1}^{\lfloor n/2 \rfloor} a_{n-i+1}(X_{(n-i+1)} - X_{(i)})\right)^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\Bigg|_{H_0} \sim \mathcal{W},$$
gdzie a_{n-i+1} są tablicowanymi stałymi, $X_{(i)}$ statystykami porządkowymi oraz $\mathcal{W}$ jest rozkładem statystyki W przy prawdziwości hipotezy zerowej, który jest tablicowany.
- Obszar krytyczny testu normalności Shapiro-Wilka jest następujący:
$$R = \{\mathbf{x} \in \mathbb{R}^n : W(\mathbf{x}) \leqslant \mathcal{W}(\alpha, n)\},$$
gdzie $\mathcal{W}(\beta, m)$ oznacza kwantyl rzędu β z rozkładu $\mathcal{W}$ statystyki W przy prawdziwości hipotezy zerowej.

5.3.2 Testy t-Studenta

Test t-Studenta dla jednej próby

- Niech
$$\mathbf{X} = (X_1, \dots, X_n)^\top$$
będzie próbą z populacji o rozkładzie normalnym $N(\mu, \sigma)$, gdzie $\mu \in \mathbb{R}$ oraz $\sigma > 0$ są nieznanymi parametrami.
- Hipoteza zerowa jest postaci:
$$H_0 : \mu = \mu_0,$$
gdzie $\mu_0 \in \mathbb{R}$ jest ustalone.
- Możliwe hipotezy alternatywne są następujące:
$$H_1^{(1)} : \mu > \mu_0, \quad H_1^{(2)} : \mu < \mu_0, \quad H_1^{(3)} : \mu \neq \mu_0.$$
- Obszary krytyczne testów t-Studenta dla jednej próby mają postaci:
  1. dla $H_1^{(1)} : \mu > \mu_0$
  $$R = \left\{\mathbf{x} \in \mathbb{R}^n : \frac{\bar{x} - \mu_0}{s}\sqrt{n} \geqslant t(1 - \alpha, n - 1)\right\},$$
  2. dla $H_1^{(2)} : \mu < \mu_0$
  $$R = \left\{\mathbf{x} \in \mathbb{R}^n : \frac{\bar{x} - \mu_0}{s}\sqrt{n} \leqslant t(\alpha, n - 1)\right\},$$
  3. dla $H_1^{(3)} : \mu \neq \mu_0$
  $$R = \left\{\mathbf{x} \in \mathbb{R}^n : \frac{|\bar{x} - \mu_0|}{s}\sqrt{n} \geqslant t(1 - \alpha/2, n - 1)\right\},$$

gdzie
$$\frac{\bar{X} - \mu_0}{S}\sqrt{n}\Big|_{H_0} \sim t(n - 1)$$
jest statystyką testową, a $t(\beta, m)$ oznacza kwantyl rzędu β z rozkładu t-Studenta $t(m)$ z m stopniami swobody.

Przykład. Automat produkuje tabliczki czekolady o nominalnej wadze $250g$. Podczas kontroli technicznej pobrano 16-elementową próbę tabliczek czekolady otrzymując wyniki:

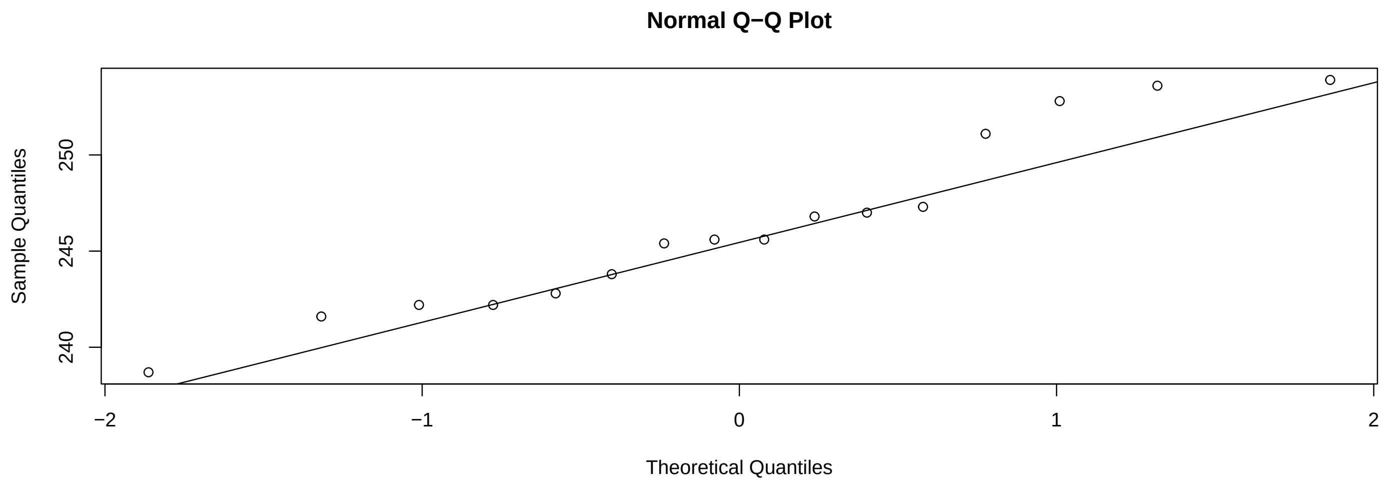
```
242.2 243.8 252.8 245.4 245.6 253.6 247.3 238.7 241.6 242.8 251.1 246.8 247 245.6 242.2 253.9
```

Na poziomie istotności $\alpha = 0{,}05$ zweryfikuj hipotezę, że automat rozlegulował się i produkuje tabliczki czekolady o istotnie różnej wadze od nominalnej wagi.

```r
x <- c(242.2, 243.8, 252.8, 245.4, 245.6, 253.6, 247.3, 238.7,
       241.6, 242.8, 251.1, 246.8, 247.0, 245.6, 242.2, 253.9)
shapiro.test(x)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.93622, p-value = 0.3052
```

```r
qqnorm(x)
qqline(x)
```




```r
mean(x)
```

```
## [1] 246.275
```

```r
t.test(x, mu = 250, alternative = "less")
```

```
##
##  One Sample t-test
##
## data:  x
## t = -3.2679, df = 15, p-value = 0.002595
## alternative hypothesis: true mean is less than 250
## 95 percent confidence interval:
##      -Inf 248.2732
## sample estimates:
## mean of x
##   246.275
```

Test t-Studenta dla dwóch prób niezależnych

- Niech

$$\mathbf{X}_1 = (X_{11}, \dots, X_{1n_1})^\top$$

  oraz

$$\mathbf{X}_2 = (X_{21}, \dots, X_{2n_2})^\top$$

$(n_1, n_2 > 1)$ będą niezależnymi próbami prostymi z populacji o rozkładach normalnych $N(\mu_1, \sigma)$ i $N(\mu_2, \sigma)$ odpowiednio, gdzie $\mu_1 \in \mathbb{R}$, $\mu_2 \in \mathbb{R}$ i $\sigma > 0$ są nieznanymi parametrami.

- Hipoteza zerowa jest postaci:
$$H_0 : \mu_1 = \mu_2.$$
- Możliwe hipotezy alternatywne są następujące:
$$H_1^{(1)} : \mu_1 > \mu_2, \quad H_1^{(2)} : \mu_1 < \mu_2, \quad H_1^{(3)} : \mu_1 \neq \mu_2.$$
- Obszary krytyczne testów t-Studenta dla dwóch prób niezależnych mają postaci:
  1. dla $H_1^{(1)} : \mu_1 > \mu_2$
  $$R = \left\{(\mathbf{x}_1, \mathbf{x}_2) : T(\mathbf{x}_1, \mathbf{x}_2) \geqslant t(1-\alpha, n_1+n_2-2)\right\},$$
  2. dla $H_1^{(2)} : \mu_1 < \mu_2$
  $$R = \left\{(\mathbf{x}_1, \mathbf{x}_2) : T(\mathbf{x}_1, \mathbf{x}_2) \leqslant t(\alpha, n_1+n_2-2)\right\},$$
  3. dla $H_1^{(3)} : \mu_1 \neq \mu_2$
  $$R = \left\{(\mathbf{x}_1, \mathbf{x}_2) : |T(\mathbf{x}_1, \mathbf{x}_2)| \geqslant t(1-\alpha/2, n_1+n_2-2)\right\},$$

gdzie
$$T(\mathbf{X}_1, \mathbf{X}_2) = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}} \sqrt{\frac{n_1 n_2}{n_1+n_2}}\Big|_{H_0} \sim t(n_1+n_2-2)$$
jest statystyką testową, a $t(\beta, m)$ oznacza kwantyl rzędu β z rozkładu t-Studenta $t(m)$ z m stopniami swobody.

Test F-Snedecora dla wariancji w dwóch próbach niezależnych

- Niech
$$\mathbf{X}_1 = (X_{11}, \dots, X_{1n_1})^\top$$
oraz
$$\mathbf{X}_2 = (X_{21}, \dots, X_{2n_2})^\top$$
$(n_1, n_2 > 1)$ będą niezależnymi próbami prostymi z populacji o rozkładach normalnych $N(\mu_1, \sigma_1)$ i $N(\mu_2, \sigma_2)$ odpowiednio, gdzie $\mu_1 \in \mathbb{R}$, $\mu_2 \in \mathbb{R}$, $\sigma_1 > 0$ i $\sigma_2 > 0$ są nieznanymi parametrami.
- Hipoteza zerowa jest postaci:
$$H_0 : \sigma_1^2 = \sigma_2^2.$$
- Możliwe hipotezy alternatywne są następujące:
$$H_1^{(1)} : \sigma_1^2 > \sigma_2^2, \quad H_1^{(2)} : \sigma_1^2 < \sigma_2^2, \quad H_1^{(3)} : \sigma_1^2 \neq \sigma_2^2.$$
- Obszary krytyczne testów F-Snedecora dla wariancji w dwóch próbach niezależnych mają postaci:
  1. dla $H_1^{(1)} : \sigma_1^2 > \sigma_2^2$
  $$R = \left\{(\mathbf{x}_1, \mathbf{x}_2) : F(\mathbf{x}_1, \mathbf{x}_2) \geqslant F(1-\alpha, n_1-1, n_2-1)\right\},$$
  2. dla $H_1^{(2)} : \sigma_1^2 < \sigma_2^2$
  $$R = \left\{(\mathbf{x}_1, \mathbf{x}_2) : F(\mathbf{x}_1, \mathbf{x}_2) \leqslant F(\alpha, n_1-1, n_2-1)\right\},$$
  3. dla $H_1^{(3)} : \sigma_1^2 \neq \sigma_2^2$
  $$R = \left\{(\mathbf{x}_1, \mathbf{x}_2) : \max\left\{F(\mathbf{x}_1, \mathbf{x}_2), \frac{1}{F(\mathbf{x}_1, \mathbf{x}_2)}\right\} \geqslant F\left(1-\frac{\alpha}{2}, n_L-1, n_M-1\right)\right\},$$

gdzie

$$F(\mathbf{X}_1, \mathbf{X}_2) = \frac{S_1^2}{S_2^2}\Big|_{H_0} \sim F(n_1 - 1, n_2 - 1)$$

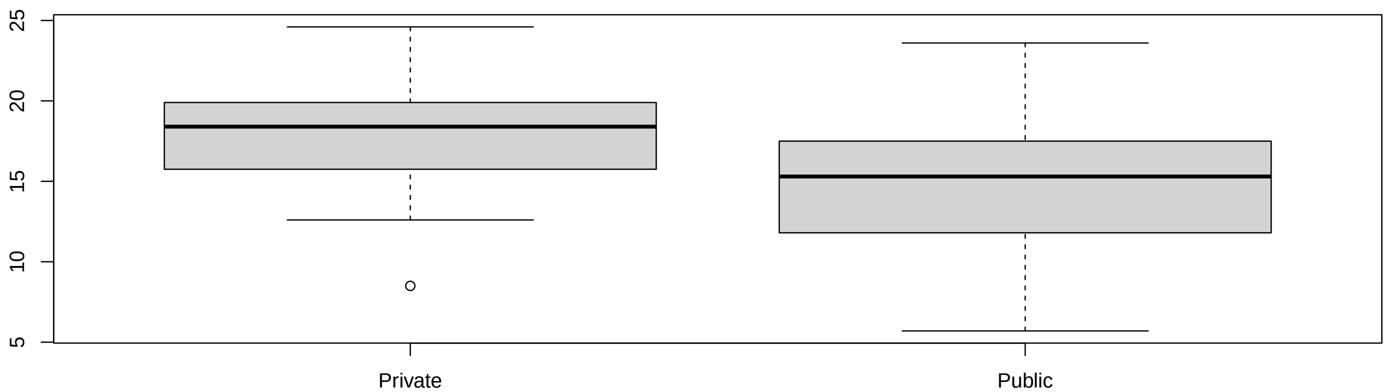
jest statystyką testową, a $F(\beta, m, n)$ oznacza kwantyl rzędu β z rozkładu F-Snedecora $F(m, n)$ z m i n stopniami swobody, oraz L i M oznaczają licznik i mianownik odpowiednio.

Przykład. Zbiór danych `homework` z pakietu `UsingR` zawiera informacje o ilości czasu poświęconego na odrabianie pracy domowej przez uczniów szkół prywatnych i publicznych. Naszym celem jest sprawdzenie, czy uczniowie obu typów szkół spędzają tyle samo czasu na odrabianiu zadań domowych.

```
library(UsingR)
head(homework)
```

```
##   Private Public
## 1    21.3   15.3
## 2    16.8   17.4
## 3     8.5   12.3
## 4    12.6   10.7
## 5    15.8   16.4
## 6    19.3   11.3
```

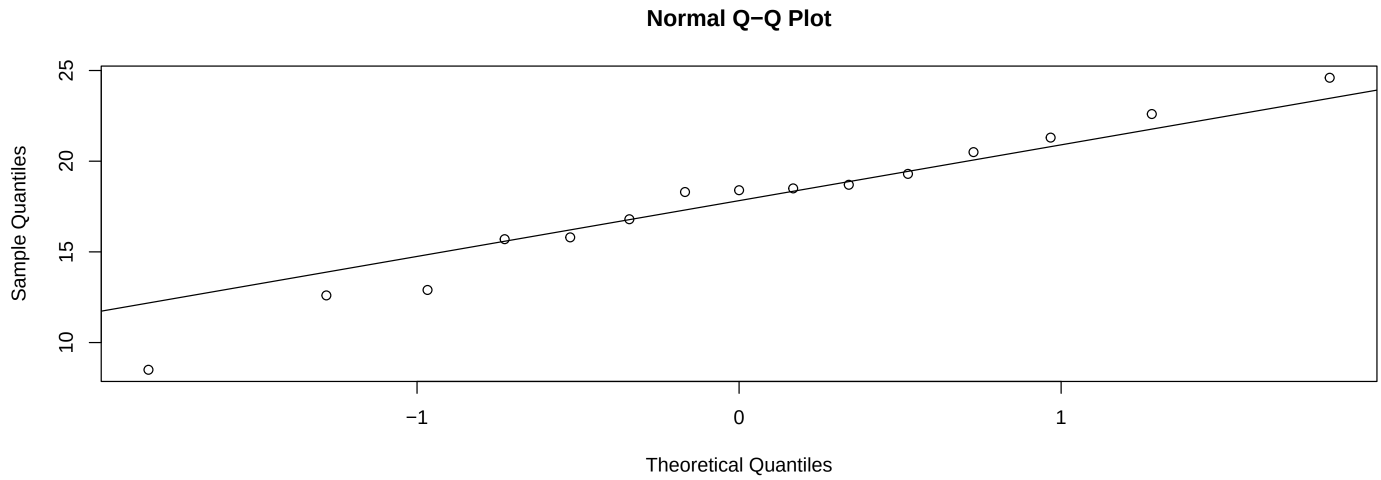
```
boxplot(homework)
```




```
shapiro.test(homework$Private)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  homework$Private
## W = 0.97017, p-value = 0.8606
```

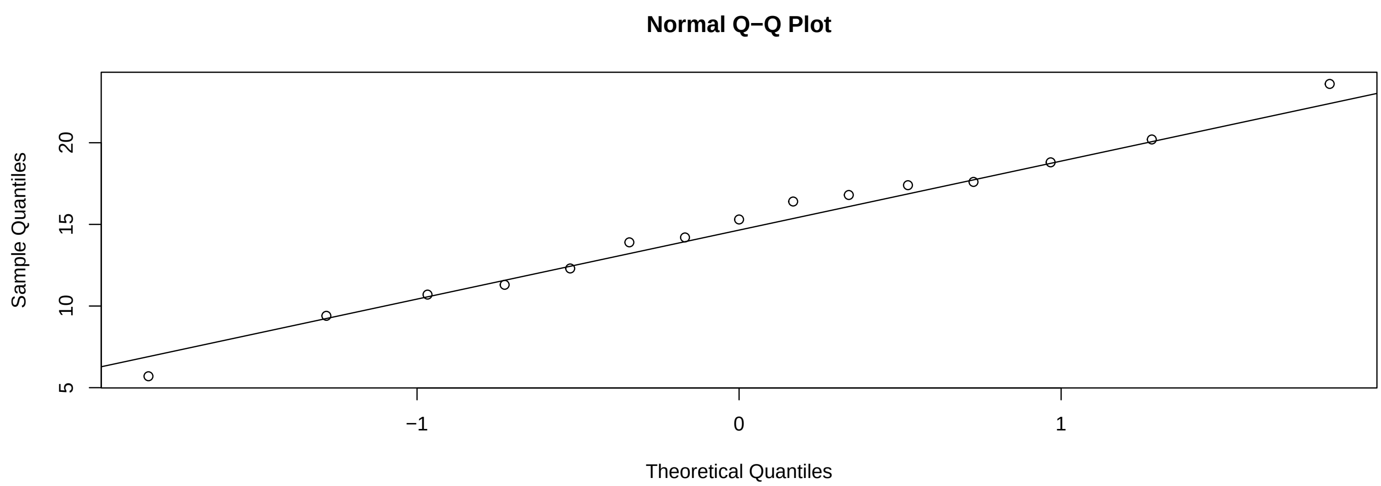
```
qqnorm(homework$Private)
qqline(homework$Private)
```




```r
shapiro.test(homework$Public)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  homework$Public
## W = 0.99275, p-value = 0.9999
```

```r
qqnorm(homework$Public)
qqline(homework$Public)
```




```r
var(homework$Private)
```

```
## [1] 17.1081
```

```r
var(homework$Public)
```

```
## [1] 20.87781
```

```r
var.test(homework$Private, homework$Public, alternative = "less")
```

```
##
##  F test to compare two variances
##
## data:  homework$Private and homework$Public
## F = 0.81944, num df = 14, denom df = 14, p-value = 0.3573
## alternative hypothesis: true ratio of variances is less than 1
```

```
## 95 percent confidence interval:
##  0.000000 2.035262
## sample estimates:
## ratio of variances
##          0.8194392
```

```r
mean(homework$Private)
```

```
## [1] 17.63333
```

```r
mean(homework$Public)
```

```
## [1] 14.90667
```

```r
t.test(homework$Private, homework$Public,
       var.equal = TRUE, alternative = 'greater')
```

```
##
##  Two Sample t-test
##
## data:  homework$Private and homework$Public
## t = 1.7134, df = 28, p-value = 0.04884
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.01957252        Inf
## sample estimates:
## mean of x mean of y
##  17.63333  14.90667
```

Test t-Studenta dla prób zależnych

- Niech
$$\mathbf{X}_1 = (X_{11}, \dots, X_{1n})^\top$$
oraz
$$\mathbf{X}_2 = (X_{21}, \dots, X_{2n})^\top$$
($n > 1$) będą dwiema zależnymi próbami z populacji o rozkładach normalnych $N(\mu_1, \sigma_1)$ i $N(\mu_2, \sigma_2)$ odpowiednio, gdzie $\mu_1 \in \mathbb{R}$, $\mu_2 \in \mathbb{R}$, $\sigma_1 > 0$ i $\sigma_2 > 0$ są nieznanymi parametrami.
- W tym problemie próby zależne oznaczają, że obserwacje zostały otrzymane poprzez przeprowadzenie tego samego eksperymentu na tych samych jednostkach eksperymentalnych, tj. X_{1j} i X_{2j} są zależnymi obserwacjami dla j-tej jednostki eksperymentalnej, $j = 1, \dots, n$.
- Na bazie prób $\mathbf{X}_1$ i $\mathbf{X}_2$, kontruujemy jedną próbę różnic obserwacji:
$$\mathbf{X} = (X_{11} - X_{21}, \dots, X_{1n} - X_{2n})^\top.$$
Wtedy problem testowania sprowadza się do problemu jednej próby z rozkładu normalnego, tj. odpowiedniego testu t-Studenta dla jednej próby.
- Hipoteza zerowa jest postaci:
$$H_0 : \mu_1 - \mu_2 = 0.$$
- Możliwe hipotezy alternatywne są następujące:
$$H_1^{(1)} : \mu_1 - \mu_2 > 0, \quad H_1^{(2)} : \mu_1 - \mu_2 < 0, \quad H_1^{(3)} : \mu_1 - \mu_2 \neq 0.$$

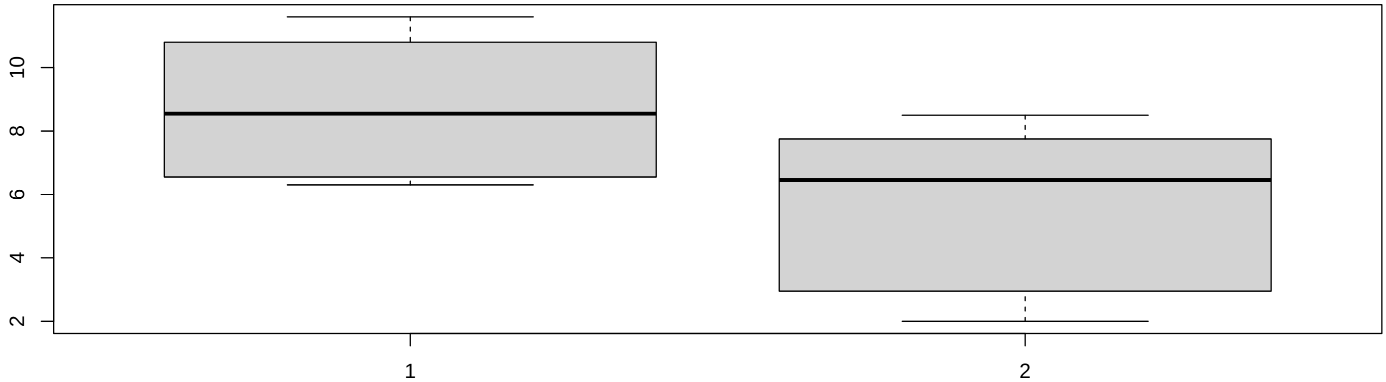
Przykład. Badano wpływ hipnozy na redukcję bólu. Notowano poziom odczuwalnego bólu:

- przed hipnozą: 6.6, 6.5, 9.0, 10.3, 11.3, 8.1, 6.3, 11.6,

- po hipnozie: 6.8, 2.5, 7.4, 8.5, 8.1, 6.1, 3.4, 2.0.

Czy na poziomie istotności 0,05 możemy stwierdzić, że hipnoza redukuje poziom odczuwalnego bólu?

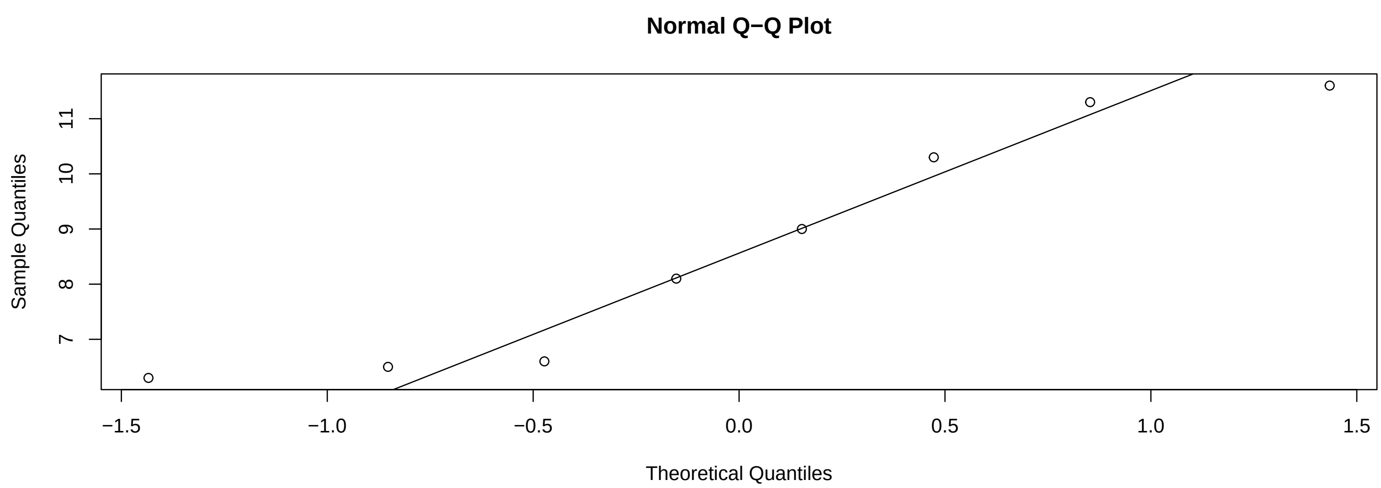
```
a <- c(6.6, 6.5, 9.0, 10.3, 11.3, 8.1, 6.3, 11.6)
b <- c(6.8, 2.5, 7.4, 8.5, 8.1, 6.1, 3.4, 2.0)
boxplot(a, b)
```




```
shapiro.test(a)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  a
## W = 0.88638, p-value = 0.2165
```

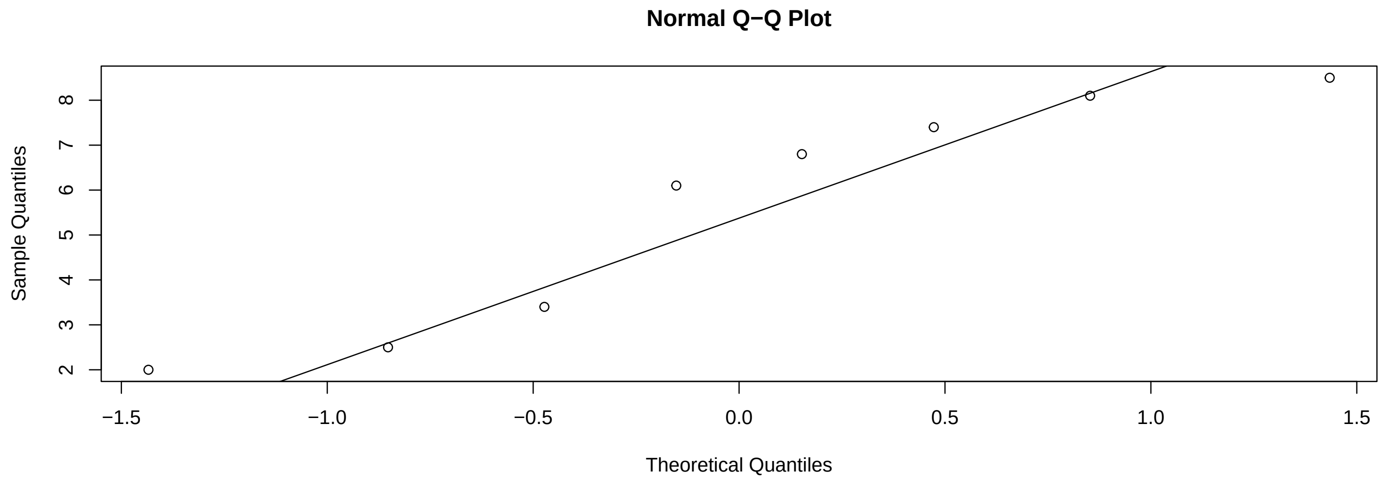
```
qqnorm(a)
qqline(a)
```




```
shapiro.test(b)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  b
## W = 0.88356, p-value = 0.2036
```

```
qqnorm(b)
qqline(b)
```




```r
mean(a)
```

```
## [1] 8.7125
```

```r
mean(b)
```

```
## [1] 5.6
```

```r
t.test(a, b, alternative = 'greater', paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  a and b
## t = 3.0285, df = 7, p-value = 0.009577
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  1.165386      Inf
## sample estimates:
## mean of the differences
##                  3.1125
```

5.3.3 Test Welcha

- Niech
$$\mathbf{X}_1 = (X_{11}, \dots, X_{1n_1})^\top$$
oraz
$$\mathbf{X}_2 = (X_{21}, \dots, X_{2n_2})^\top$$
$(n_1, n_2 > 1)$ będą dwiema niezależnymi próbami prostymi z populacji o rozkładach normalnych $N(\mu_1, \sigma_1)$ i $N(\mu_2, \sigma_2)$ odpowiednio, gdzie $\mu_1 \in \mathbb{R}$, $\mu_2 \in \mathbb{R}$, $\sigma_1 > 0$ i $\sigma_2 > 0$ są nieznanymi parametrami.
- Hipoteza zerowa jest postaci:
$$H_0 : \mu_1 = \mu_2.$$
- Możliwe hipotezy alternatywne są następujące:
$$H_1^{(1)} : \mu_1 > \mu_2, \quad H_1^{(2)} : \mu_1 < \mu_2, \quad H_1^{(3)} : \mu_1 \neq \mu_2.$$
- Obszary krytyczne testów Welcha są postaci:

1. dla $H_1^{(1)} : \mu_1 > \mu_2$
$$R = \{(\mathbf{x}_1, \mathbf{x}_2) : T(\mathbf{x}_1, \mathbf{x}_2) \geqslant t(1-\alpha, m)\},$$
2. dla $H_1^{(2)} : \mu_1 < \mu_2$
$$R = \{(\mathbf{x}_1, \mathbf{x}_2) : T(\mathbf{x}_1, \mathbf{x}_2) \leqslant t(\alpha, m)\},$$
3. dla $H_1^{(3)} : \mu_1 \neq \mu_2$
$$R = \{(\mathbf{x}_1, \mathbf{x}_2) : |T(\mathbf{x}_1, \mathbf{x}_2)| \geqslant t(1-\alpha/2, m)\},$$

gdzie
$$T(\mathbf{X}_1, \mathbf{X}_2) = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}\Big|_{H_0} \sim t(m) \textit{ granicznie}$$
jest statystyką testową,
$$m = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2-1}}$$
i $t(\beta, m)$ oznacza kwantyl rzędu β z rozkładu t-Studenta $t(m)$ z m stopniami swobody.

5.3.4 Testy Manna-Whitneya-Wilcoxona

- Niech
$$\mathbf{X} = (X_1, \dots, X_n)^\top$$
będzie próbą prostą z populacji o rozkładzie ciągłym z dystrybuantą F.
- Rangujemy próbę $\mathbf{X}$, tj. kontruujemy próbę rang
$$\mathbf{R} = (R_1, \dots, R_n)^\top,$$
gdzie
$$R_i = rank(X_i).$$

Przykład. Niech
$$\mathbf{x} = (4, 7, 1, 5)^\top.$$
Wtedy
$$x_1 = 4,\ x_2 = 7,\ x_3 = 1,\ x_4 = 5$$
oraz
$$x_{(1)} = 1,\ x_{(2)} = 4,\ x_{(3)} = 5,\ x_{(4)} = 7.$$
Zatem
$$\mathbf{r} = (2, 4, 1, 3)^\top.$$

- Z ciągłości rozkładu wynika, że obserwacje x_i, $i = 1, \dots, n$ powinny być wszystkie różne. Jednak, jeżeli $x_i = x_j$ i $i \neq j$, to $x_{(k)} = x_{(k+1)}$ dla pewnego k. W takiej sytuacji obserwacjom x_i i x_j przypisujemy równe rangi
$$\frac{k + (k+1)}{2} = k + \frac{1}{2}.$$

Lemat. Mamy
$$\bar{R} = \frac{1}{n}\sum_{i=1}^{n} R_i = \frac{n+1}{2}, \quad S_R^2 = \frac{1}{n-1}\sum_{i=1}^{n}(R_i - \bar{R})^2 = \frac{n(n+1)}{12}.$$

- Zatem $\bar{R}$ oraz S_R^2 są stałe, więc na ich podstawie nie można wnioskować.

- Niech
$$\mathbf{X}_1 = (X_{11}, \dots, X_{1n_1})^\top$$
oraz
$$\mathbf{X}_2 = (X_{21}, \dots, X_{2n_2})^\top$$
będą dwiem niezależnymi próbami prostymi z populacji o rozkładach ciągłych z dystrybuantami F_{μ_1} i F_{μ_2} odpowiednio, gdzie
$$F_\mu(x) = F(x - \mu)$$
dla pewnej ciągłej dystrybuanty F.

- Parametr μ nazywamy parametrem położenia. Przykładowo takim parametrem jest mediana w rozkładach normalnym, Laplace'a i Cauchy'ego.

- Hipoteza zerowa jest postaci:
$$H_0 : \mu_1 = \mu_2.$$

- Rangujemy próbę połączoną
$$(\mathbf{X}_1^\top, \mathbf{X}_2^\top)^\top = (X_{11}, \dots, X_{1n_1}, X_{21}, \dots, X_{2n_2})^\top,$$
a następnie otrzymujemy dwie próby rang:
$$\mathbf{R}_1 = (R_{11}, \dots, R_{1n_1})^\top, \quad \mathbf{R}_2 = (R_{21}, \dots, R_{2n_2})^\top.$$

- Statystyka testowa testu Manna-Whitneya-Wilcoxona jest postaci:
$$W(\mathbf{X}_1, \mathbf{X}_2) = \sum_{i=1}^{n_2} R_{2i}.$$

- Przy prawdziwości hipotezy zerowej, wszystkie układy rang są równo prawdopodobne. Oznacza to, że rozkład statystyki testowej W nie zależy od dystrybuanty F przy prawdziwości H_0.

- Obszary krytyczne testów Manna-Whitneya-Wilcoxona są następujące:
  1. dla $H_1^{(1)} : \mu_1 > \mu_2$
  $$R = \left\{(\mathbf{x}_1, \mathbf{x}_2) : W(\mathbf{x}_1, \mathbf{x}_2) \leqslant k(\alpha)\right\},$$
  2. dla $H_1^{(2)} : \mu_1 < \mu_2$
  $$R = \left\{(\mathbf{x}_1, \mathbf{x}_2) : W(\mathbf{x}_1, \mathbf{x}_2) \geqslant k(1-\alpha)\right\},$$

gdzie $k(\beta)$ jest wartością krytyczną otrzymaną z rozkładu statystyki testowej W przy prawdziwości hipotezy zerowej.

- Statystyka testowa U Manna-Whitneya ma postać:
$$U = \sum_{j=1}^{n_2} \sum_{i=1}^{n_1} I(X_{1i} < X_{2j}),$$
gdzie
$$I(x < y) = \begin{cases} 1, & \text{gdy } x < y, \\ 0, & \text{gdy } x \geqslant y. \end{cases}$$

Lemat. Mamy
$$U = W - \frac{1}{2}n_2(n_2+1).$$

Twierdzenie. Przy prawdziwości hipotezy zerowej
$$Z(\mathbf{X}_1, \mathbf{X}_2) = \frac{U - E_0(U)}{\sqrt{Var_0(U)}} \xrightarrow{d} N(0,1),$$
gdzie
$$E_0(U) = \frac{n_1 n_2}{2}, \quad Var_0(U) = \frac{n_1 n_2(n_1+n_2+1)}{12}.$$

- Na podstawie powyższego twierdzenia, otrzymujemy poniższe obszary krytyczne testów U-Manna-Whitneya dla odpowiednio dużych prób:
    1. dla $H_1^{(1)}: \mu_1 > \mu_2$
    $$R = \left\{(\mathbf{x}_1, \mathbf{x}_2) : Z(\mathbf{x}_1, \mathbf{x}_2) \leqslant -z(1-\alpha)\right\},$$
    2. dla $H_1^{(2)}: \mu_1 < \mu_2$
    $$R = \left\{(\mathbf{x}_1, \mathbf{x}_2) : Z(\mathbf{x}_1, \mathbf{x}_2) \geqslant z(1-\alpha)\right\},$$

gdzie $z(\beta)$ oznacza kwantyl rzędu β z rozkładu normalnego $N(0,1)$.

Przykład. (ten sam co dla testu t-Studenta dla dwóch prób niezależnych) Zbiór danych `homework` z pakietu `UsingR` zawiera informacje o ilości czasu poświęconego na odrabianie pracy domowej przez uczniów szkół prywatnych i publicznych. Naszym celem jest sprawdzenie, czy uczniowie obu typów szkół spędzają tyle samo czasu na odrabianiu zadań domowych.

```
library(UsingR)
head(homework)
```

```
##   Private Public
## 1    21.3   15.3
## 2    16.8   17.4
## 3     8.5   12.3
## 4    12.6   10.7
## 5    15.8   16.4
## 6    19.3   11.3
```

```
mean(homework$Private)
```

```
## [1] 17.63333
```

```
mean(homework$Public)
```

```
## [1] 14.90667
```

```
t.test(homework$Private, homework$Public,
       var.equal = TRUE, alternative = 'greater')
```

```
##
##  Two Sample t-test
##
## data:  homework$Private and homework$Public
## t = 1.7134, df = 28, p-value = 0.04884
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.01957252        Inf
```

```
## sample estimates:
## mean of x mean of y
##  17.63333  14.90667
```

```r
median(homework$Private)
```

```
## [1] 18.4
```

```r
median(homework$Public)
```

```
## [1] 15.3
```

```r
wilcox.test(homework$Private, homework$Public, alternative = 'greater')
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  homework$Private and homework$Public
## W = 154.5, p-value = 0.04258
## alternative hypothesis: true location shift is greater than 0
```

```r
# wilcox.test(x, y = NULL,
#             alternative = c("two.sided", "less", "greater"),
#             mu = 0, paired = FALSE, ...)
```

5.3.5 Testy istotności dla wskaźnika struktury

Test istotności dla wskaźnika struktury

- Test istotności dla wskaźnika struktury nazywany jest również testem istotności dla proporcji.
- Niech
$$\mathbf{X} = (X_1, \dots, X_n)^\top$$
będzie próbą prostą z populacji o rozkładzie zero-jedynkowym $b(1, p)$, gdzie $p \in (0, 1)$ jest nieznanym parametrem.
- Parametr p często reprezentuje procent obserwacji specjalnego rodzaju (wyróżnionych).
- Hipoteza zerowa jest postaci:
$$H_0 : p = p_0,$$
gdzie $p_0 \in (0, 1)$ jest ustalone.
- Możliwe hipotezy alternatywne są następujące:
$$H_1^{(1)} : p > p_0, \quad H_1^{(2)} : p < p_0, \quad H_1^{(3)} : p \neq p_0.$$
- Gdy $n \geqslant 100$, obszary krytyczne testów istotności dla wskaźnika struktury są następujące:
  1. dla $H_1^{(1)} : p > p_0$
  $$R = \left\{ \mathbf{x} \in \mathbb{R}^n : \frac{\bar{x} - p_0}{\sqrt{p_0(1 - p_0)}} \sqrt{n} \geqslant z(1 - \alpha) \right\},$$
  2. dla $H_1^{(2)} : p < p_0$
  $$R = \left\{ \mathbf{x} \in \mathbb{R}^n : \frac{\bar{x} - p_0}{\sqrt{p_0(1 - p_0)}} \sqrt{n} \leqslant z(\alpha) \right\},$$

3. dla $H_1^{(3)}: p \neq p_0$

$$R = \left\{ \mathbf{x} \in \mathbb{R}^n : \frac{|\bar{x} - p_0|}{\sqrt{p_0(1-p_0)}} \sqrt{n} \geqslant z(1-\alpha/2) \right\},$$

gdzie

$$\frac{\bar{X} - p_0}{\sqrt{p_0(1-p_0)}} \sqrt{n} \Big|_{H_0} \sim N(0,1) \text{ } granicznie$$

jest statystyką testową, a $z(\beta)$ oznacza kwantyl rzędu β z rozkładu normalnego $N(0,1)$.

Test dwumianowy

- Test dwumianowy jest dokładnym testem hipotezy testu istotności dla wskaźnika struktury.
- Niech $X \sim b(n, p_0)$ oraz k będzie liczbą elementów próby będącymi wyróżnionymi obserwacjami.
- p-wartości testów dwumianowych są następujące:

1. dla $H_1^{(1)}: p > p_0$, $P(X \geqslant k)$,
2. dla $H_1^{(2)}: p < p_0$, $P(X \leqslant k)$,
3. dla $H_1^{(3)}: p \neq p_0$, p-wartość jest bardziej skomplikowana, więc ją pomijamy.

Przykład. W mieszance nasiennej, według normy, udział żyta powinien wynosić 60%. Na podstawie 120 prób ustalono, że udział ten jest rzędu 50%. Zweryfikuj hipotezę, że ten wkład jest równy normie.

- Mamy $p_0 = 0.6$, $n = 120$ oraz $k = 0.5 \cdot 120 = 60$.

```
prop.test(x = 0.5 * 120, n = 120, p = 0.6, alternative = "less")
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  0.5 * 120 out of 120, null probability 0.6
## X-squared = 4.592, df = 1, p-value = 0.01606
## alternative hypothesis: true p is less than 0.6
## 95 percent confidence interval:
##  0.0000000 0.5783169
## sample estimates:
##   p
## 0.5
```

```
prop.test(x = 0.5 * 120, n = 120, p = 0.6, alternative = "less", correct = FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  0.5 * 120 out of 120, null probability 0.6
## X-squared = 5, df = 1, p-value = 0.01267
## alternative hypothesis: true p is less than 0.6
## 95 percent confidence interval:
##  0.0000000 0.5742447
## sample estimates:
##   p
## 0.5
```

```
binom.test(x = 0.5 * 120, n = 120, p = 0.6, alternative = "less")
```

```
##
##  Exact binomial test
##
## data:  0.5 * 120 and 120
## number of successes = 60, number of trials = 120, p-value = 0.01674
## alternative hypothesis: true probability of success is less than 0.6
## 95 percent confidence interval:
##  0.0000000 0.5785925
## sample estimates:
## probability of success
##                    0.5
```

```r
pbinom(0.5 * 120, 120, 0.6)
```

```
## [1] 0.01673614
```

Test istotności dla dwóch wskaźników struktury

- Niech
$$\mathbf{X}_1 = (X_{11}, \dots, X_{1n_1})^\top$$
oraz
$$\mathbf{X}_2 = (X_{21}, \dots, X_{2n_2})^\top$$
$(n_1, n_2 > 1)$ będą niezależnymi próbami prostymi z populacji o rozkładach zero-jedynkowych $b(1, p_1)$ i $b(1, p_2)$ odpowiednio, gdzie $p_1 \in (0, 1)$ i $p_2 \in (0, 1)$ są nieznanymi parametrami.

- Hipoteza zerowa jest postaci:
$$H_0 : p_1 = p_2.$$

- Możliwe hipotezy alternatywne są następujące:
$$H_1^{(1)} : p_1 > p_2, \quad H_1^{(2)} : p_1 < p_2, \quad H_1^{(3)} : p_1 \neq p_2.$$

- Gdy $n_1 > 100$ oraz $n_2 > 100$, obszary krytyczne testów istotności dla dwóch wskaźników struktury są następujące:
  1. dla $H_1^{(1)} : p_1 > p_2$
  $$R = \{(\mathbf{x}_1, \mathbf{x}_2) : U(\mathbf{x}_1, \mathbf{x}_2) \geqslant z(1-\alpha)\},$$
  2. dla $H_1^{(2)} : p_1 < p_2$
  $$R = \{(\mathbf{x}_1, \mathbf{x}_2) : U(\mathbf{x}_1, \mathbf{x}_2) \leqslant z(\alpha)\},$$
  3. dla $H_1^{(3)} : p_1 \neq p_2$
  $$R = \{(\mathbf{x}_1, \mathbf{x}_2) : |U(\mathbf{x}_1, \mathbf{x}_2)| \geqslant z(1-\alpha/2)\},$$

gdzie
$$U(\mathbf{X}_1, \mathbf{X}_2) = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1+n_2)\bar{X}_\bullet(1-\bar{X}_\bullet)}{n_1 n_2}}}\Bigg|_{H_0} \sim N(0,1) \textit{ granicznie}$$
jest statystyką testową,
$$\bar{X}_\bullet = \frac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1 + n_2},$$
a $z(\beta)$ oznacza kwantyl rzędu β z rozkładu normalnego $N(0, 1)$.

Przykład. Losowo wybrano 800 osób korzystających z transportu autobusowego i 800 osób korzystających z transportu kolejowego. Przeprowadzona ankieta wykazała, że 506 osób ma zastrzeżenia do komunikacji autobusowej, a 368 osób ma zastrzeżenia do kolei. Zweryfikuj hipotezę, że odsetek osób niezadowolonych z transportu autobusowego nie różni się znacząco od odsetka osób krytycznych wobec transportu kolejowego.

```
prop.test(c(506, 368), c(800, 800), alternative = "greater")
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(506, 368) out of c(800, 800)
## X-squared = 47.327, df = 1, p-value = 3.003e-12
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.1309241 1.0000000
## sample estimates:
## prop 1 prop 2
## 0.6325 0.4600
```

Test McNemary

- Test McNemary jest modyfikacją testu istotności dla dwóch wskaźników struktury w przypadku prób zależnych, podobnie jak test t-Studenta dla prób zależnych w stosunku do testu t-Studenta dla dwóch prób niezależnych.
- Niech
$$\mathbf{X}_1 = (X_{11}, \dots, X_{1n})^\top$$
oraz
$$\mathbf{X}_2 = (X_{21}, \dots, X_{2n})^\top$$
($n \geqslant 20$) będą dwiema zależnymi próbami prostymi z populacji o rozkładach zero-jedynkowych $b(1, p_1)$ i $b(1, p_2)$ odpowiednio, gdzie $p_1 \in (0, 1)$ i $p_2 \in (0, 1)$ są nieznanymi parametrami.
- Hipotezy zerowa i alternatywna są następujące:
$$H_0 : p_1 = p_2 \ przeciw \ H_1 : p_1 \neq p_2.$$
- Dane agreguje się w następującej tabeli:

	Próba II (po)	
Próba I (przed)	TAK	NIE
TAK	A	B
NIE	C	D

- Obszar krytyczny testu McNemary ma postać:
$$R = \left\{ (\mathbf{x}_1, \mathbf{x}_2) : \chi^2 = \frac{(|B - C| - 1)^2}{B + C} \geqslant \chi^2(1 - \alpha, 1) \right\},$$
gdzie $\chi(\beta, m)$ oznacza kwantyl rzędu β z rozkładu chi-kwadrat $\chi^2(m)$ z m stopniami swobody.

Przykład. 1319 dzieci w wieku 12 lat zapytano, czy miały objawy przeziębienia w ciągu ostatniego roku. 356 z nich powiedziało twierdząco. Badanie powtórzono po 2 latach, otrzymując 468 odpowiedzi twierdzących. Poniższa tabela zawiera pełne informacje na temat obu próbek.

	14 lat	
12 lat	TAK	NIE
TAK	212	144

	14 lat	
NIE	256	707

Czy można powiedzieć, że nastąpił znaczny wzrost liczby przeziębień?

```
matrix(c(212, 256, 144, 707), nrow = 2)
```

```
##      [,1] [,2]
## [1,]  212  144
## [2,]  256  707
```

```
mcnemar.test(matrix(c(212, 256, 144, 707), nrow = 2))
```

```
##
##  McNemar's Chi-squared test with continuity correction
##
## data:  matrix(c(212, 256, 144, 707), nrow = 2)
## McNemar's chi-squared = 30.802, df = 1, p-value = 2.857e-08
```

5.3.6 Testy χ^2-Pearsona

Test zgodności χ^2-Pearsona

- Test zgodności χ^2-Pearsona jest testem zgodności z wybranym rozkładem dyskretnym lub ciągłym. Jednak, my rozważymy tylko przypadek rozkładu dyskretnego.
- Niech
$$\mathbf{X} = (X_1, \dots, X_n)^\top$$
będzie próbą prostą z populacji o rozkładzie dyskretnym danym następująco: dla $i = 1, \dots, n$,
$$P(X_i = j) = p_j, \;\; j = 1, \dots, k.$$
- Hipotezy zerowa i alternatywna są następujące:
$$H_0 : \mathbf{p} = \mathbf{p}_0 \; przeciwko \; H_1 : \mathbf{p} \neq \mathbf{p}_0,$$
gdzie $\mathbf{p} = (p_1, \dots, p_k)$ i $\mathbf{p}_0 = (p_{01}, \dots, p_{0k})$ jest ustalonym wektorem prawdopodobieństw szczególnego rozkładu dyskretnego.
- Gdy $np_j \geqslant 5$, obszar krytyczny testu zgodności χ^2-Pearsona jest postaci:
$$R = \left\{ \mathbf{x} : \chi^2(\mathbf{x}) = \sum_{j=1}^{k} \frac{(n_j - np_{0j})^2}{np_{0j}} \geqslant \chi^2(1-\alpha, k-1) \right\},$$
gdzie
$$\chi^2(\mathbf{X})\Big|_{H_0} \sim \chi^2(k-1) \; granicznie$$
jest statystyką testową, n_j jest liczebnością j-tej wartości zmiennej w próbie, $j = 1, \dots, k$, a $\chi^2(\beta, m)$ oznacza kwantyl rzędu β z rozkładu chi-kwadrat $\chi^2(m)$ z m stopniami swobody.
- W przypadku ogólnym
$$H_0 : \mathbf{p} \in \{\mathbf{p}_0(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^s\}$$
statystyka testowa $\chi^2(\mathbf{X}) \sim \chi^2(k-s-1)$ granicznie, a $p_{0j}(\widehat{\boldsymbol{\theta}})$ jest wykorzystany zamiast p_{0j}, gdzie $\widehat{\boldsymbol{\theta}} = ENW(\boldsymbol{\theta})$.

Przykład. W pewnym banku zaobserwowano liczbę obsługiwanych klientów na minutę w ciągu 200 jednominutowych okresów w danym tygodniu.

Liczba klientów	0	1	2	3	4	5	6	7
Liczebność	14	31	47	41	29	21	10	7

Czy na podstawie tych danych można sądzić, że liczba obsługiwanych klientów ma rozkład Poissona?

- $H_0: X \sim \pi(\lambda)$, gdzie $\lambda > 0$ jest nieznanym parametrem.

```
x <- rep(0:7, c(14, 31, 47, 41, 29, 21, 10, 7))
lambda_est <- mean(x)
p0 <- c(dpois(0:6, lambda_est), 1 - ppois(6, lambda_est))
chisq.test(table(x), p = p0)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  table(x)
## X-squared = 2.1658, df = 7, p-value = 0.9501
```

```
# liczba stopni swobody = 8 - 1 - 1
1 - pchisq(2.1658, 6)
```

```
## [1] 0.9038357
```

Test χ^2-Pearsona dla dwóch prób

- Test χ^2-Pearsona dla dwóch prób jest testem służącym do porównania dwóch rozkładów. Może być on wykorzystany zarówno do rozkładów dyskretnych jak i ciągłych. Jednak, my rozważymy tylko przypadek rozkładu dyskretnego.
- Niech
$$\mathbf{X}_1 = (X_{11}, \dots, X_{1n_1})^\top$$
oraz
$$\mathbf{X}_2 = (X_{21}, \dots, X_{2n_2})^\top$$
będą dwiema niezależnymi próbami prostymi z populacji o rozkładach dyskretnych reprezentowanych następującymi funkcjami prawdopodobieństwa:
$$P(X_{1i} = j) = p_{1j},\ P(X_{2m} = j) = p_{2j},\ j = 1, \dots, k$$
odpowiednio, dla $i = 1, \dots, n_1$, $m = 1, \dots, n_2$.
- Hipotezy zerowa i alternatywna są następujące:
$$H_0 : \mathbf{p}_1 = \mathbf{p}_2 \textit{ przeciwko } H_1 : \mathbf{p}_1 \neq \mathbf{p}_2,$$
gdzie $\mathbf{p}_1 = (p_{11}, \dots, p_{1k})$ i $\mathbf{p}_2 = (p_{21}, \dots, p_{2k})$.
- Obszar krytyczny testu χ^2-Pearsona dla dwóch prób jest postaci:
$$R = \left\{ (\mathbf{x}_1, \mathbf{x}_2) : \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{(n_{ij} - E_{ij})^2}{E_{ij}} \geqslant \chi^2(1-\alpha, k-1) \right\},$$
gdzie n_{ij} jest liczebnością j-tej wartości zmiennej w i-tej próbie,
$$E_{ij} = \frac{n_{1j} + n_{2j}}{n_1 + n_2} n_i,$$
dla $i = 1, 2$, $j = 1, \dots, k$, a $\chi^2(\beta, m)$ oznacza kwantyl rzędu β z rozkładu chi-kwadrat $\chi^2(m)$ z m stopniami swobody.

Przykład. Losowo wybranym grupom gimnazjalistów zadano pytanie: Jak oceniasz sytuację materialną swojej rodziny? Dostępne były następujące odpowiedzi: dobra, średnia, zła. Uczniów podzielono na dwie grupy: dziewczynki i chłopców. Wyniki przedstawia poniższa tabela.

	Dziewczynki	Chłopcy
dobra	20	39
średnia	85	95
zła	5	6

Czy na podstawie uzyskanych odpowiedzi można stwierdzić, że istnieją istotne różnice w rozkładzie opinii na temat sytuacji materialnej rodzin wśród chłopców i dziewcząt?

```
matrix(c(20, 85, 5, 39, 95, 6), nrow = 3)
```

```
##      [,1] [,2]
## [1,]   20   39
## [2,]   85   95
## [3,]    5    6
```

```
chisq.test(matrix(c(20, 85, 5, 39, 95, 6), nrow = 3))
```

```
##
##  Pearson's Chi-squared test
##
## data:  matrix(c(20, 85, 5, 39, 95, 6), nrow = 3)
## X-squared = 3.2114, df = 2, p-value = 0.2008
```

5.3.7 Testy Kołmogorowa-Smirnowa

Test Kołmogorowa-Smirnowa dla jednej próby

- Test Kołmogorowa-Smirnowa dla jednej próby jest testem zgodności z wybranym rozkładem ciągłym.
- Niech
$$\mathbf{X} = (X_1, \ldots, X_n)^\top$$
będzie próbą z populacji o rozkładzie ciągłym z dystrybuantą F.
- Hipotezy zerowa i alternatywna są następujące:
$$H_0 : F = F_0 \textit{ przeciwko } H_1 : F \neq F_0,$$
gdzie F_0 jest ustaloną ciągłą dystrybuantą.

Definicja. Statystykę $F_n \equiv F_n(\cdot, \mathbf{X})$ daną wzorem

$$F_n(x) = \frac{\#\{k : X_k \leqslant x\}}{n}$$

nazywamy dystrybuantą empiryczną.

- Gdy $n > 100$, obszar krytyczny testu Kołmogorowa-Smirnowa dla jednej próby jest postaci:
$$R = \left\{\mathbf{x} : \sqrt{n} D_n \geqslant \lambda(1-\alpha)\right\},$$
gdzie
$$\sqrt{n} D_n(\mathbf{X}) = \sqrt{n} \sup_{-\infty<x<\infty} |F_n(x) - F_0(x)|\Big|_{H_0} \sim K \textit{ granicznie}$$

jest statystyką testową, K jest zmienną losową o rozkładzie Kołmogorowa, a $\lambda(\beta)$ oznacza kwantyl rzędu β z tego rozkładu (np. $\lambda(0.9) = 1.224$, $\lambda(0.95) = 1.354$, $\lambda(0.99) = 1.628$).

- Gdy $n \leqslant 100$, obszar krytyczny testu Kołmogorowa-Smirnowa dla jednej próby jest postaci:

$$R = \{\mathbf{x} : D_n \geqslant d_n(1-\alpha)\},$$

gdzie wartości $d_n(1-\alpha)$ są tablicowane.

Lemat. Przy prawdziwości hipotezy zerowej rozkład statystyki testowej $D_n(\mathbf{X})$ nie zależy od F_0.

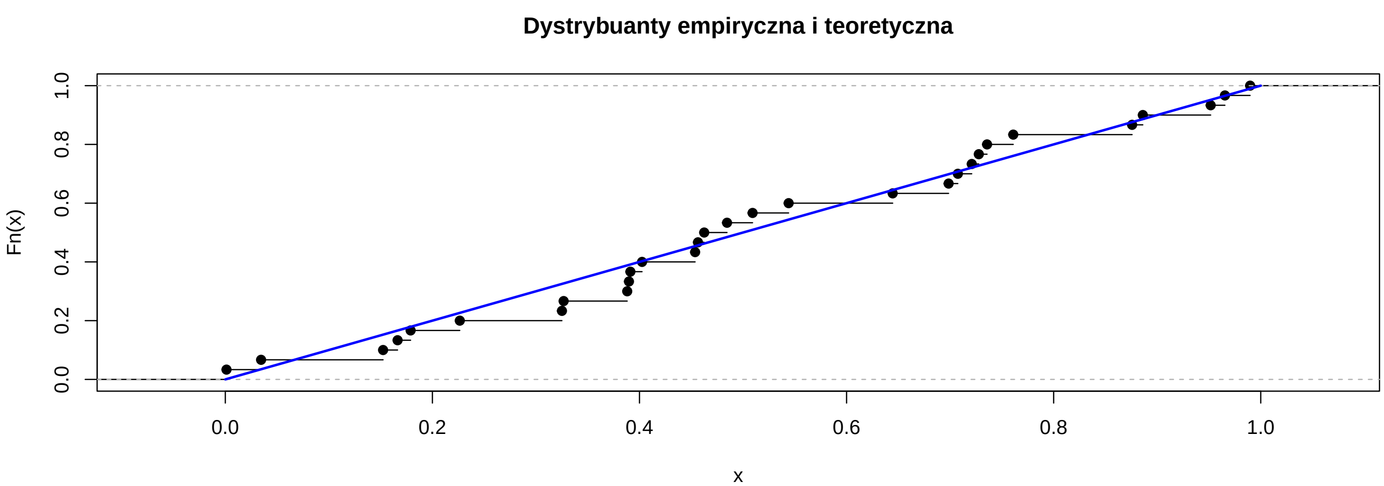
- Niestety test Kołmogorowa-Smirnowa dla jednej próby wykazuje zgodność z rozkładem normalnym, nawet w przypadku danych znacznie odbiegających od rozkładu normalnego.
- Ten test wymaga również znajomości parametrów rozkładu. Parametry nie mogą być estymowane.

Przykład. W celu weryfikacji poprawności generatora liczb pseudolosowych z rozkładu jednostajnego $U(0,1)$ w programie R, wygeneruj 30 liczb. Sprawdź hipotezę o poprawności generatora.

```
set.seed(12345)
x <- runif(30)
ks.test(x, "punif")
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  x
## D = 0.1251, p-value = 0.6895
## alternative hypothesis: two-sided
```

```
plot(ecdf(x), main = "Dystrybuanty empiryczna i teoretyczna")
curve(punif(x), from = 0, to = 1, col = "blue", add = TRUE, lwd = 2)
```



Test Kołmogorowa-Smirnowa dla dwóch prób

- Niech

$$\mathbf{X}_1 = (X_{11}, \dots, X_{1n_1})^\top$$

oraz

$$\mathbf{X}_2 = (X_{21}, \dots, X_{2n_2})^\top$$

będą niezależnymi próbami prostymi z populacji o rozkładach ciągłych z dystrybuantami F i G odpowiednio.

- Hipotezy zerowa i alternatywna są następujące:

$$H_0 : F = G \textit{ przeciwko } H_1 : F \neq G.$$

- Gdy $n_1, n_2 > 20$, obszar krytyczny testu Kołmogorowa-Smirnowa dla dwóch prób jest postaci:

$$R = \left\{ (\mathbf{x}, \mathbf{y}) : \sqrt{\frac{n_1 n_2}{n_1 + n_2}} D_{n_1,n_2} \geqslant \lambda(1-\alpha) \right\},$$

  gdzie
$$\sqrt{\frac{n_1 n_2}{n_1 + n_2}} D_{n_1,n_2} = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \sup_{-\infty < x < \infty} |F_{n_1}(x) - F_{n_2}(x)| \Big|_{H_0} \sim K \textit{ granicznie}$$
  jest statystyką testową, K jest zmienną losową o rozkładzie Kołmogorowa, a $\lambda(\beta)$ oznacza kwantyl rzędu β z tego rozkładu.
- Gdy $n_1, n_2 \leqslant 20$, obszar krytyczny testu Kołmogorowa-Smirnowa dla dwóch prób jest postaci:

$$R = \left\{ (\mathbf{x}, \mathbf{y}) : D_{n_1,n_2} \geqslant d(\alpha, n_1, n_2) \right\},$$

  gdzie wartości krytyczne $d(\alpha, n_1, n_2)$ są tablicowane.

Przykład. Przeprowadzono eksperyment na dwóch próbach świnek morskich. Zaobserwowana waga świń (w gramach) w pierwszej próbce wynosiła:

$$280, 325, 270, 385, 275, 290, 400, 330, 300, 345,$$

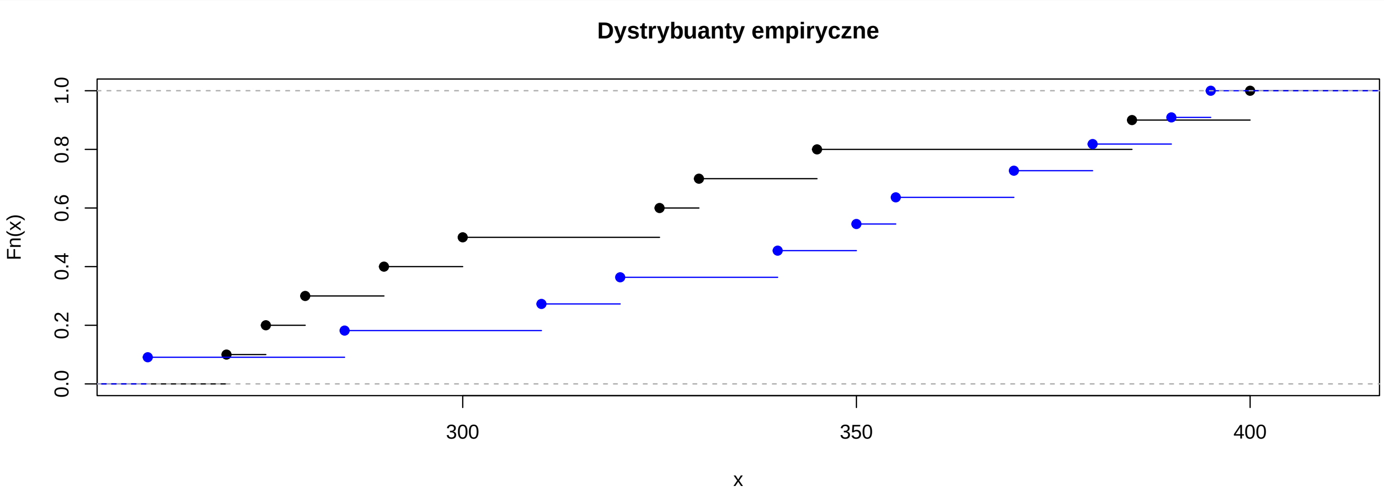
a w drugiej:
$$260, 380, 320, 350, 285, 395, 370, 340, 310, 390, 355.$$

Zweryfikuj hipotezę, że analizowane próby pochodzą z tej samej populacji.

```
x <- c(280, 325, 270, 385, 275, 290, 400, 330, 300, 345)
y <- c(260, 380, 320, 350, 285, 395, 370, 340, 310, 390, 355)
ks.test(x, y)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  x and y
## D = 0.34545, p-value = 0.4345
## alternative hypothesis: two-sided
```

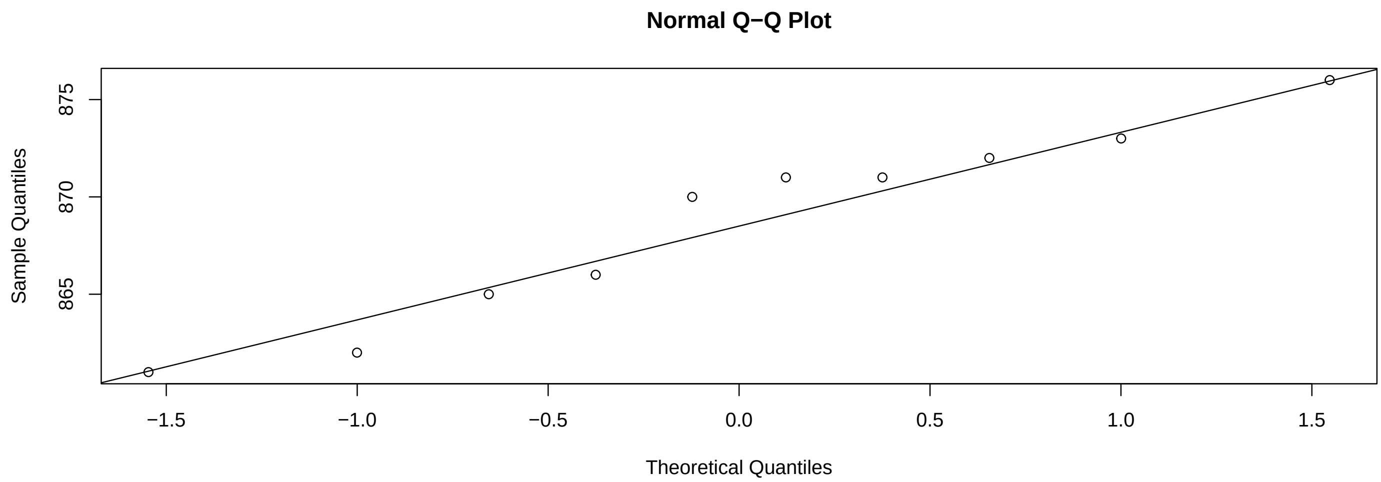
```
plot(ecdf(x), main = "Dystrybuanty empiryczne")
plot(ecdf(y), add = TRUE, col = "blue")
```



5.4 Zadania 5

Zadanie 1. W pewnym regionie wykonano dziesięć niezależnych pomiarów głębokości morza i uzyskano następujące wyniki: 862, 870, 876, 866, 871, 865, 861, 873, 871, 872. Na poziomie istotności $\alpha = 0{,}05$ zweryfikuj hipotezę, że średnia głębokość morza w tym regionie wynosi $870m$.

```
## [1] 0.545861
```




```
## [1] 868.7
```

```
## [1] 0.2136555
```

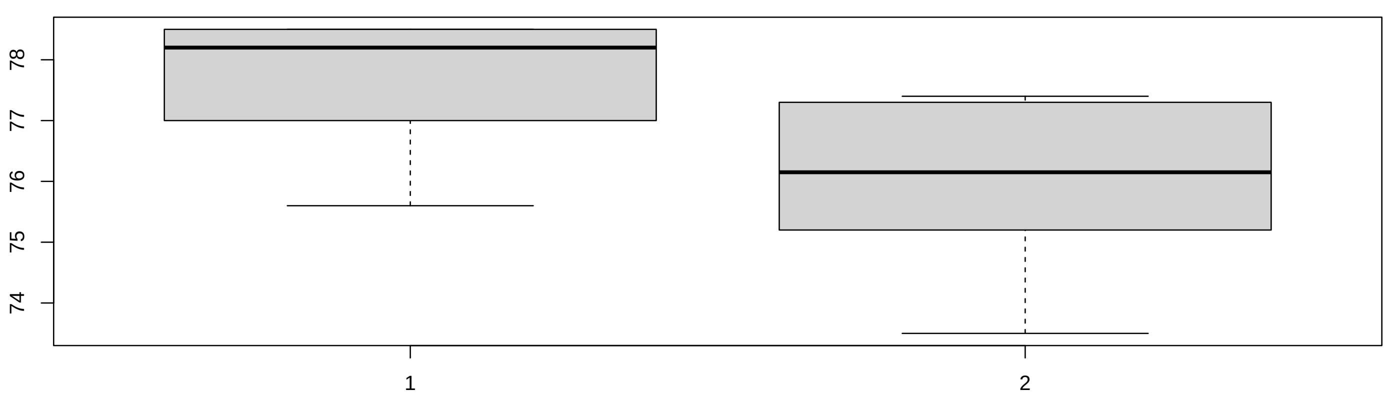
Zadanie 2. Producent proszku do prania A twierdzi, że jego produkt jest znacznie lepszy niż konkurencyjny proszek B. Aby zweryfikować to zapewnienie, CTA (Consumer Test Agency) przetestowało oba proszki do prania. W tym celu przeprowadzono pomiary stopnia wyprania 7 kawałków tkaniny z proszkiem A i uzyskano wyniki (w %):

$$78{,}2; 78{,}5; 75{,}6; 78{,}5; 78{,}5; 77{,}4; 76{,}6,$$

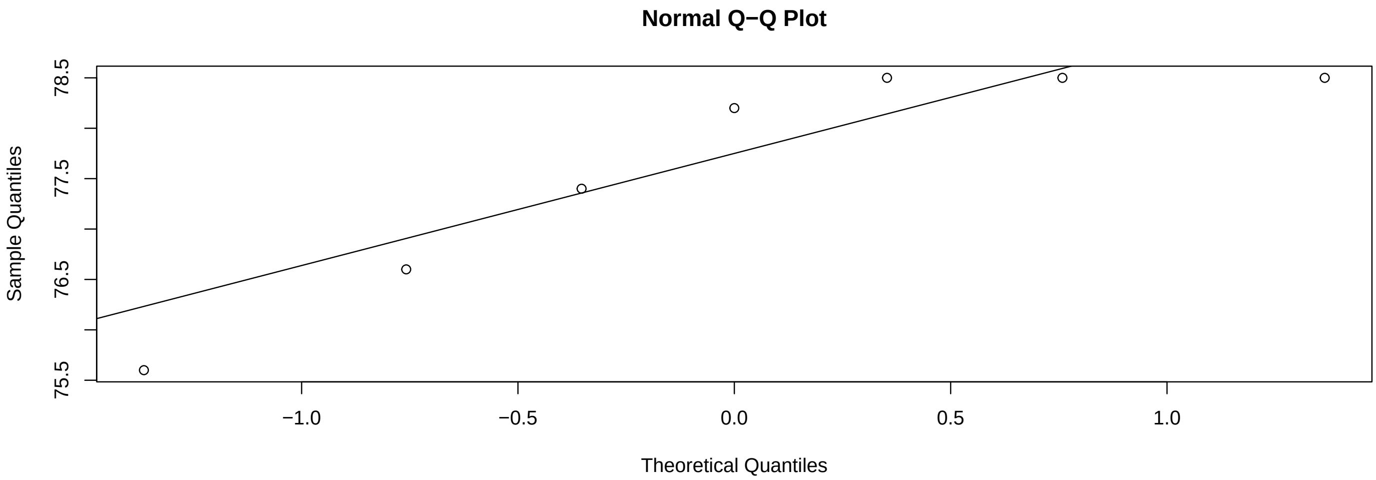
i 10 kawałków tkaniny z proszkiem B otrzymując wyniki (w %):

$$76{,}1; 75{,}2; 75{,}8; 77{,}3; 77{,}3; 77{,}0; 74{,}4; 76{,}2; 73{,}5; 77{,}4.$$

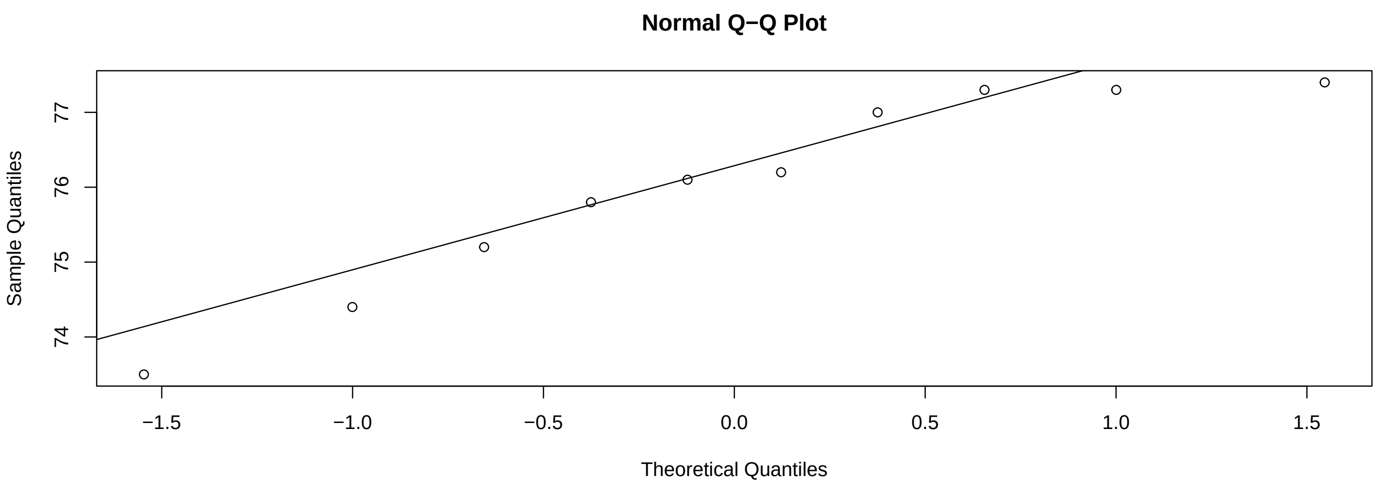
Jaki powinien być wniosek CTA na temat jakości tych proszków?




```
## [1] 0.06832755
```




```
## [1] 0.2558752
```




```
## [1] 1.304762
```

```
## [1] 1.764
```

```
## [1] 0.3683809
```

```
## [1] 77.61429
```

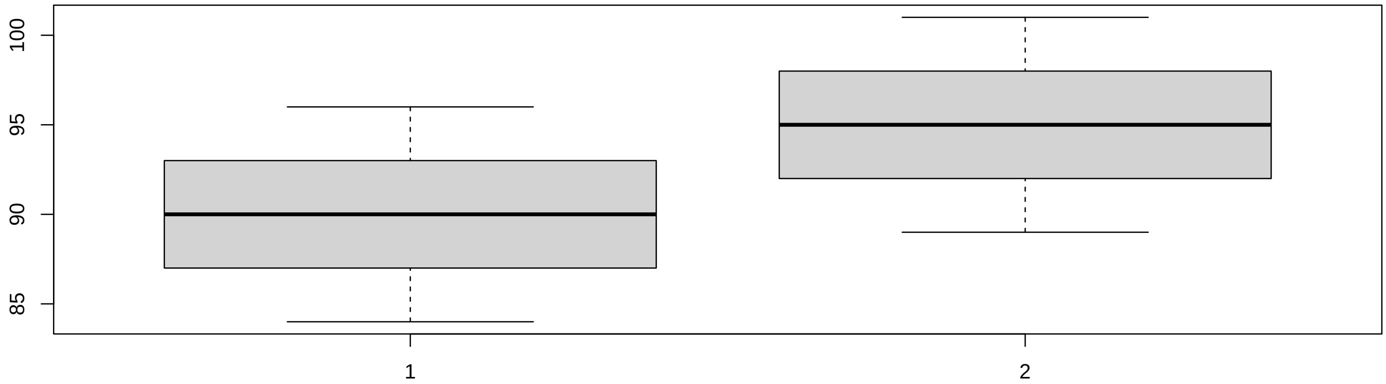
```
## [1] 76.02
```

```
## [1] 0.01059375
```

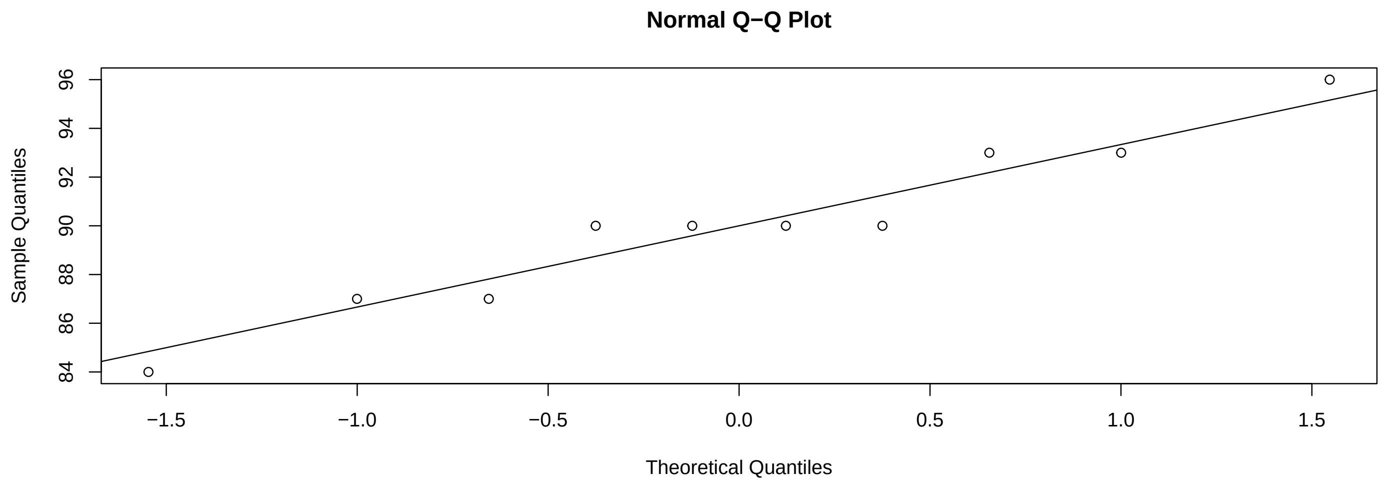
Zadanie 3. Grupa 10 osób została poddana badaniu mającemu na celu zbadanie stosunku do szkół publicznych. Następnie grupa obejrzała film edukacyjny mający na celu poprawę podejścia do tego typu szkół. Wyniki są następujące (wyższa wartość oznacza lepsze podejście):

- przed: 84, 87, 87, 90, 90, 90, 90, 93, 93, 96,
- po: 89, 92, 98, 95, 95, 92, 95, 92, 98, 101.

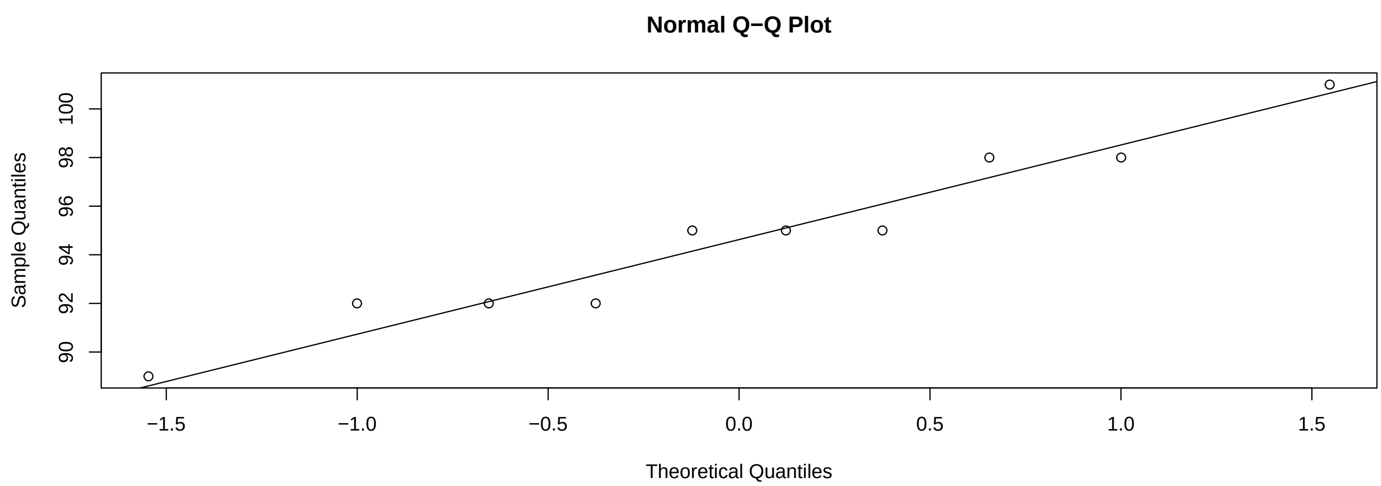
Zweryfikuj, czy film znacznie poprawia stosunek do szkół publicznych.




```
## [1] 0.7025892
```




```
## [1] 0.691489
```




```
## [1] 90
```

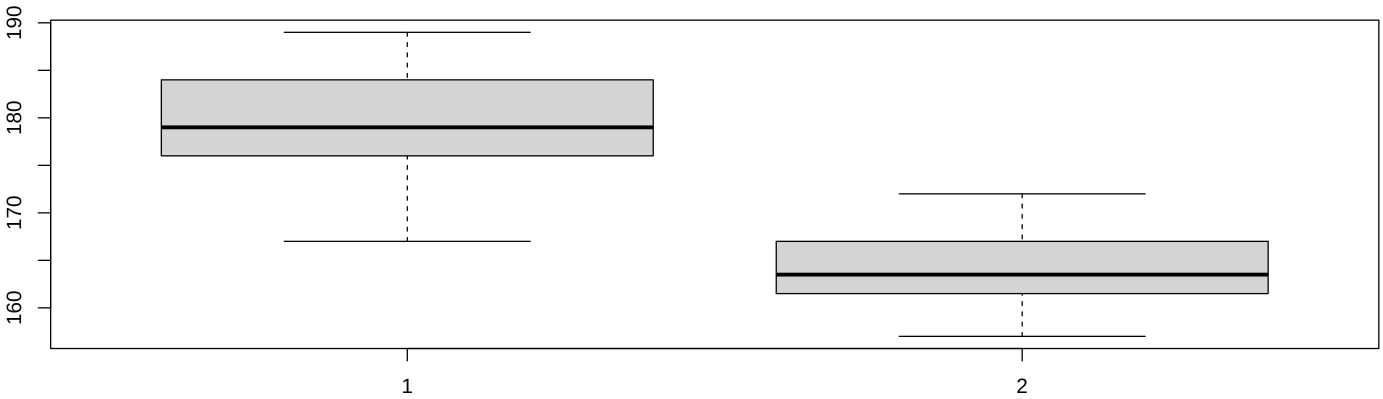
```
## [1] 94.7
```

```
## [1] 0.0003786878
```

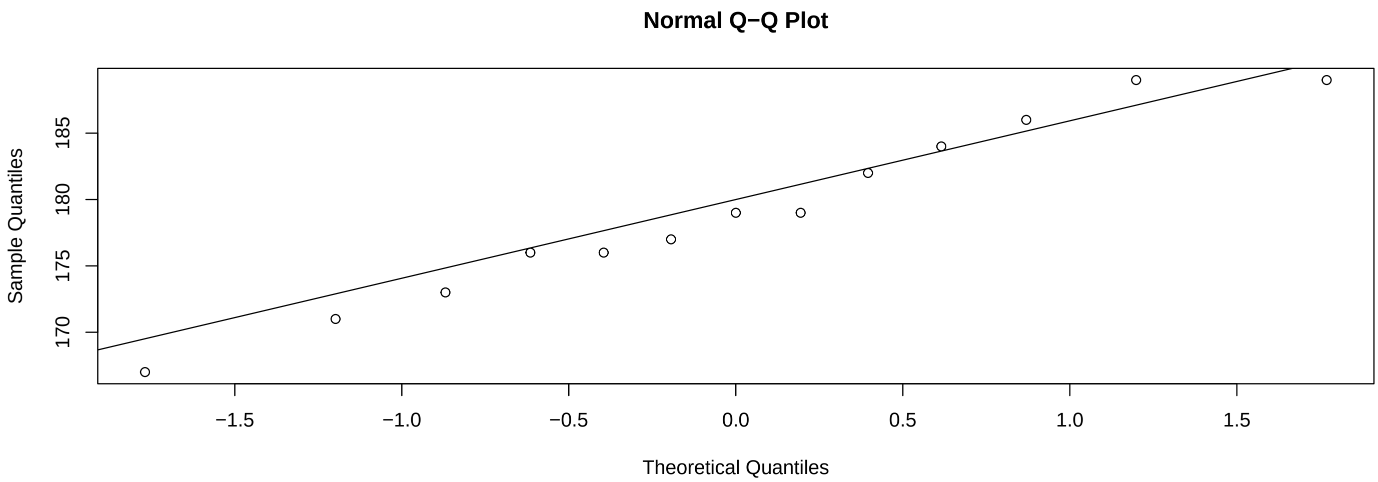
Zadanie 4. Zbadano wzrost 13 mężczyzn i 12 kobiet w pewnym centrum sportowym. Wyniki są następujące:

- mężczyźni: 171, 176, 179, 189, 176, 182, 173, 179, 184, 186, 189, 167, 177,
- kobiety: 161, 162, 163, 162, 166, 164, 168, 165, 168, 157, 161, 172.

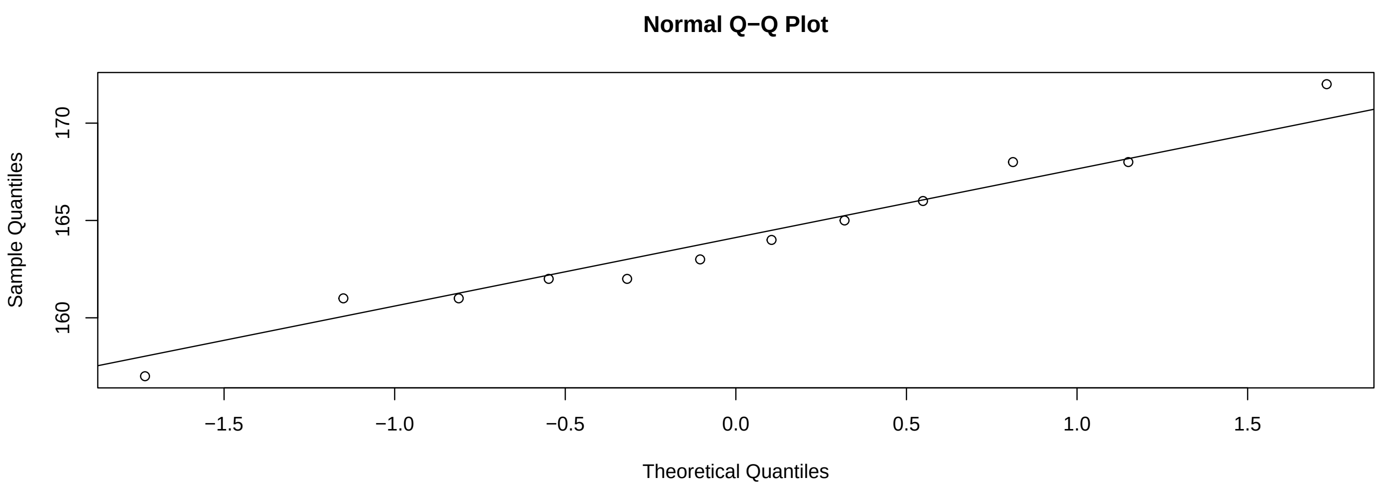
Czy możemy stwierdzić, że średni wzrost mężczyzn jest znacznie większy niż wzrost kobiet?




```
## [1] 0.8595396
```




```
## [1] 0.9447828
```




```
## [1] 45.74359
## [1] 16.08333
## [1] 0.04689163
## [1] 179.0769
## [1] 164.0833
## [1] 6.928802e-07
```

Zadanie 5.

(a) Napisz funkcję `w_test()` implementującą test χ^2 w modelu wykładniczym, który jest opisany we wskazówce. Funkcja ta powinna mieć trzy argumenty: `x` - wektor zawierający dane, `lambda_zero` - wartość λ_0 w hipotezie zerowej oraz `alternative` - typ hipotezy alternatywnej, która może mieć trzy możliwe wartości: `"two.sided"` (wartość domyślna), `"greater"`, `"less"`. Funkcja zwraca obiekt będący listą klasy `htest` o elementach: `statistic` - wartość statystyki testowej, `parameter` - liczba stopni swobody, `p.value` - p-wartość, `alternative` - wybrana hipoteza alternatywna, `method` - nazwa testu, `data.name` - nazwa zbioru danych (użyj `deparse(substitute(x))`). Dla obiektów klasy `htest` funkcja `print()` istnieje w programie R, więc nie trzeba jej tworzyć.

Wskazówka. Niech $\mathbf{X} = (X_1, \ldots, X_n)^\top$ będzie próbą prostą z populacji o rozkładzie wykładniczym $Ex(\lambda)$, gdzie $\lambda > 0$ jest nieznanym parametrem. Testy χ^2 w modelu wykładniczym weryfikują hipotezę zerową $H_0 : \lambda = \lambda_0$, gdzie $\lambda_0 > 0$ jest ustaloną liczbą. Ich obszary krytyczne są następujące:

1. dla $H_1^{(1)} : \lambda > \lambda_0$
$$R = \left\{\mathbf{x} : T(\mathbf{x}) \leqslant \chi^2(\alpha, 2n)\right\},$$
2. dla $H_1^{(2)} : \lambda < \lambda_0$
$$R = \left\{\mathbf{x} : T(\mathbf{x}) \geqslant \chi^2(1-\alpha, 2n)\right\},$$
3. dla $H_1^{(3)} : \lambda \neq \lambda_0$
$$R = \left\{\mathbf{x} : T(\mathbf{x}) \geqslant \chi^2(1-\alpha/2, 2n) \ or \ T(\mathbf{x}) \leqslant \chi^2(\alpha/2, 2n)\right\},$$

gdzie

$$T(\mathbf{X}) = 2\lambda_0 n\bar{X}\Big|_{H_0} \sim \chi^2(2n)$$

jest statystyką testową, a $\chi^2(\beta, m)$ oznacza kwantyl rzędu β z rozkładu chi-kwadrat $\chi^2(m)$ z m stopniami swobody.

(b) Wykorzystując funkcję `w_test()` zastosuj test χ^2 w modelu wykładniczym do danych dotyczących czasu bezawarynej pracy dostępnych w pliku awarie.txt i hipotezy zerowej $H_0 : \lambda = 0.001$.

```
## [1] 0.0009079683
```

```
##
##  Test chi-kwadrat w modelu wykładniczym
##
## data:  awarie$V1
## T = 110.14, num df = 100, p-value = 0.2295
## alternative hypothesis: less
```

Zadanie 6. Rozwiąż Przykład dla testu t-Studenta dla jednej próby i Zadanie 1 powyżej stosując odpowiedni test nieparametryczny.

- Przykład

```
## [1] 245.6
```

```
## [1] 0.004498527
```

- Zadanie 1

```
## [1] 870.5
```

```
## [1] 0.7615951
```

Zadanie 7. Rozwiąż Przykład dla testu t-Student dla prób zależnych i Zadania 2-4 powyżej stosując odpowiedni test nieparametryczny.

- Przykład

```
## [1] 8.55
```

```
## [1] 6.45
```

```
## [1] 0.0078125
```

- Zadanie 2

```
## [1] 78.2
```

```
## [1] 76.15
```

```
## [1] 0.01213373
```

- Zadanie 3

```
## [1] 90
```

```
## [1] 95
```

```
## [1] 0.002960434
```

- Zadanie 4

```
## [1] 179
```

```
## [1] 163.5
```

```
## [1] 3.133914e-05
```

Zadanie 8. W przypadku pewnego mikro RNA badacz chce przetestować hipotezę, że prawdopodobieństwo wystąpienia puryn wynosi 0,7. W przeprowadzonym eksperymencie mikro RNA o długości 22 zawierało 18 puryn. Zweryfikuj hipotezę badacza.

```
## [1] 0.8181818
```

```
## [1] 0.1642825
```

```
## [1] 0.1645488
```

Zadanie 9. Po porównaniu podobnych firm w dwóch różnych miastach A i B postawiono hipotezę, że odsetek firm korzystających z reklam w obu miastach jest znacząco różny. Aby sprawdzić tę hipotezę, wybrano 120 firm w mieście A, z czego 20 wykorzystało reklamę, a spośród 150 firm w mieście B 45 firm skorzystało z reklamy. Ustal, czy różnica między odsetkami firm korzystających z reklam w miastach A i B jest statystycznie istotna.

```
## [1] 0.01625468
```

Zadanie 10. W losowej próbie 1600 Amerykanów uprawnionych do głosowania 944 z nich pozytywnie oceniło działalność prezydenta. Po miesiącu ankieta została powtórzona, a 880 respondentów pozytywnie oceniło działalność prezydenta. Dokładne wyniki obu badań są następujące:

	Ankieta 2	
Ankieta 1	pozytywnie	negatywnie
pozytywnie	794	150
negatywnie	86	570

Sprawdź hipotezę o nieistotnej różnicy w odpowiedziach ankietowanych.

```
## [1] 4.114562e-05
```

Zadanie 11. Samochody określonej marki są produkowane w kolorze białym, niebieskim i czerwonym, a wielkość produkcji w poszczególnych kolorach jest ustalana w stosunku 2 : 5 : 3. Sprawdź, czy proporcje odpowiadają preferencjom klientów, jeśli spośród 150 wylosowanych potencjalnych nabywców: 38 osób wybrało biały, 72 osoby wybrało niebieski, 40 wybrało czerwony.

```
## [1] 0.2455034
```

Zadanie 12. Za pomocą odpowiedniego testu sprawdź poprawność modelu zaproponowanego w Zadaniu 2 w temacie 4.

```
## [1] 0.9252245
```

```
## [1] 0.8456537
```

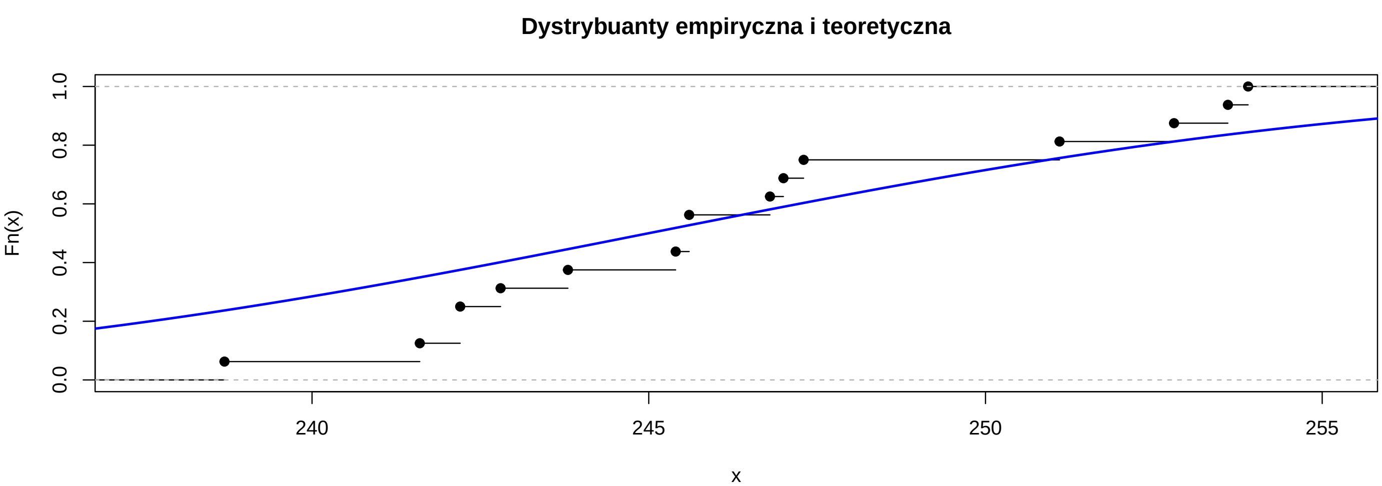
Zadanie 13. Pewien produkt można wytworzyć dwiema metodami. Postawiono hipotezę, że jakość produktu nie zależy od metody produkcji. Zweryfikuj tę hipotezę na podstawie następujących danych.

Jakość	Metoda 1	Metoda 2
I	50	90
II	20	50
III	10	6

```
## [1] 0.03740584
```

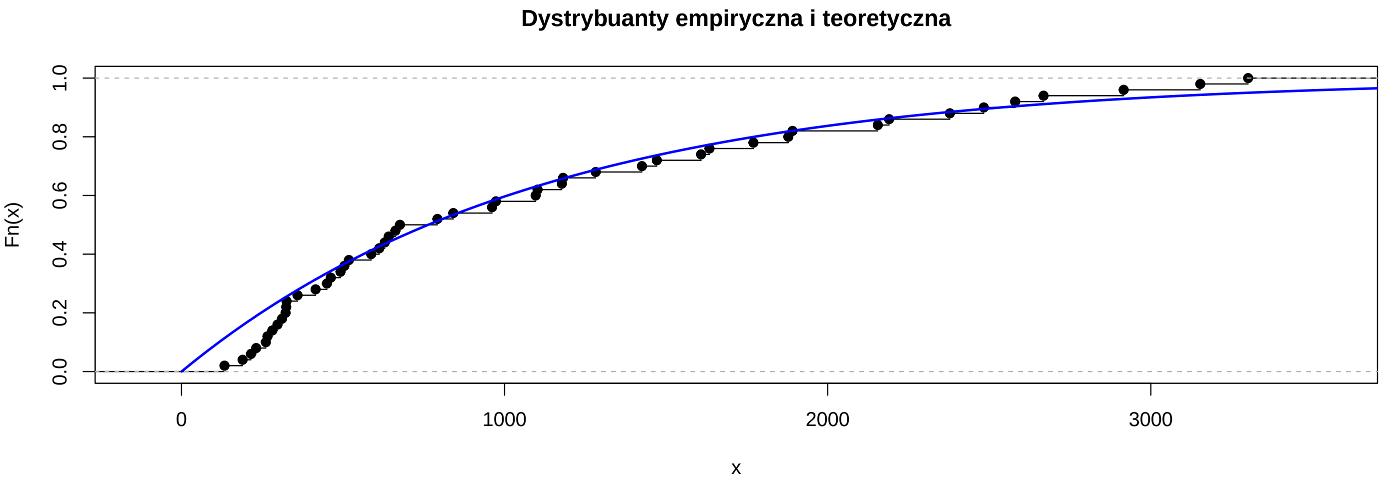
Zadanie 14. Zweryfikuj normalność zmiennej uwzględnionej w powyższym przykładzie dla testu t-Studenta dla jednej próby używając testu innego niż test Shapiro-Wilka.

```
## [1] 0.8008478
```



Zadanie 15. Za pomocą odpowiedniego testu sprawdź poprawność modelu zaproponowanego w Zadaniu 3 w temacie 4.

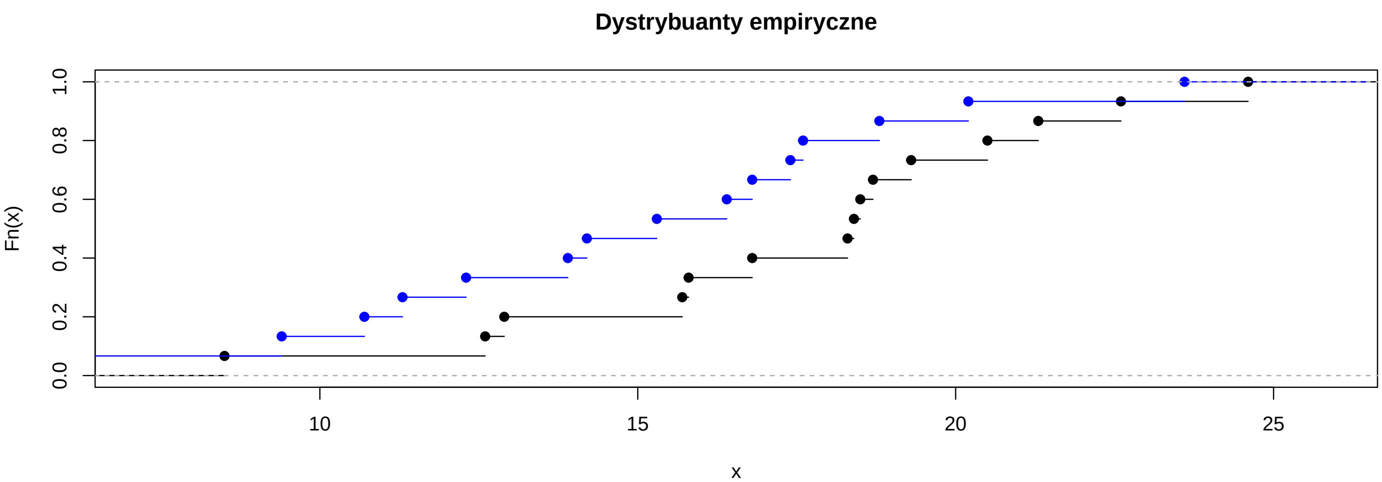
```
## [1] 0.2734308
```



Zadanie 16. W przypadku danych uwzględnionych w przykładzie dla testu t-Studenta dla dwóch prób niezależnych oraz w Zadaniach 2 i 4 powyżej zweryfikuj hipotezę, że rozważane próbki pochodzą z tej samej populacji.

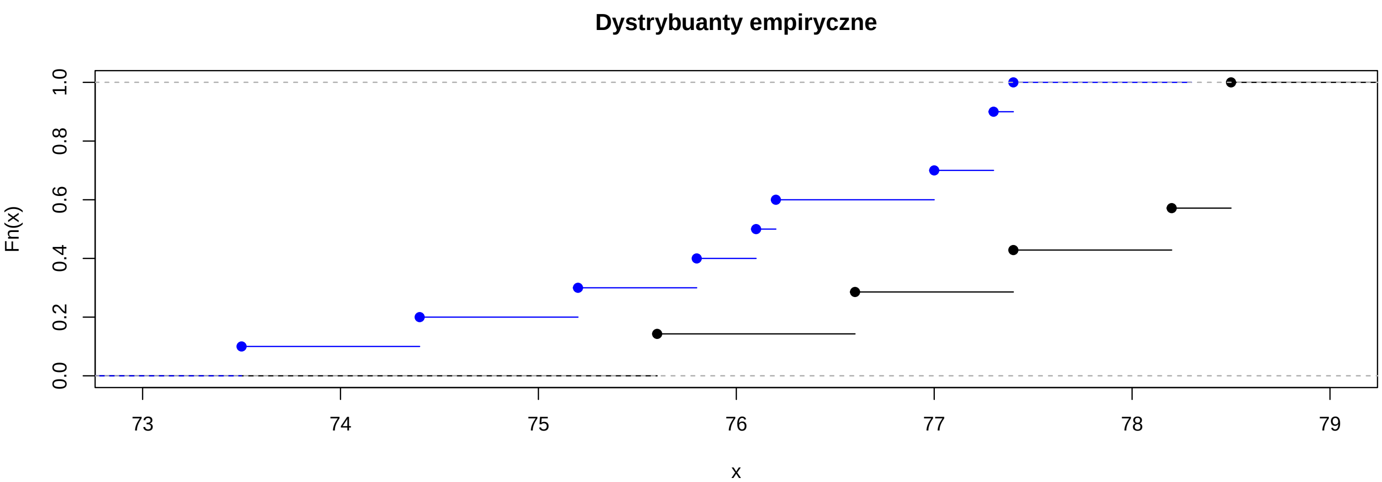
- Przykład

```
## [1] 0.1813004
```



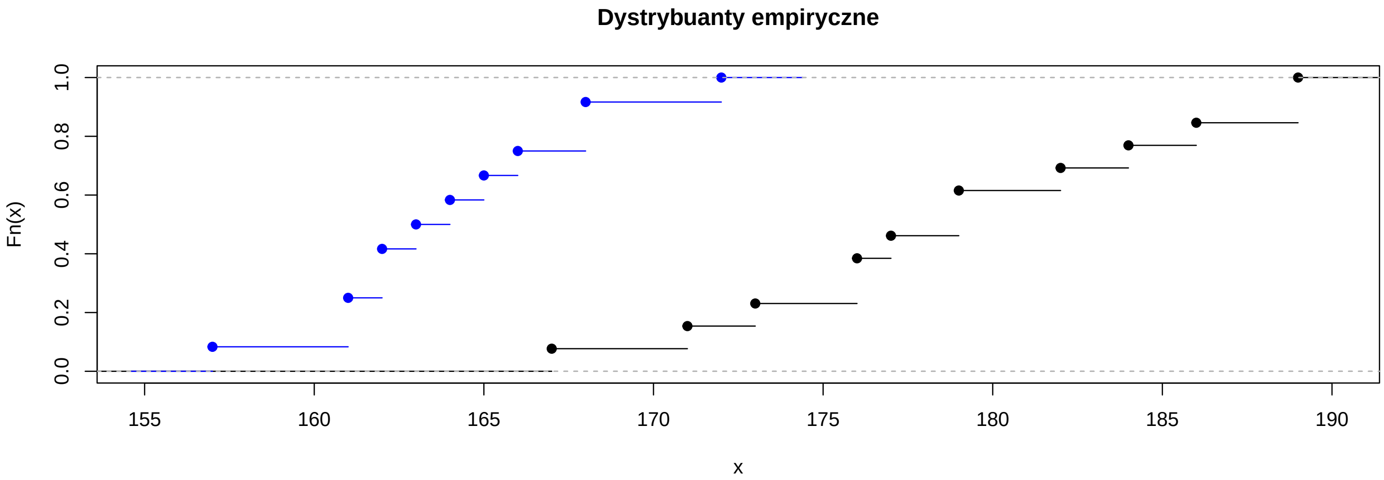
- Zadanie 2

```
## [1] 0.08941157
```



- Zadanie 4

```
## [1] 0.0002632945
```



6 Analiza wariancji

- Niech Y będzie zmienną losową o rozkładzie ciągłym oraz niech X będzie jakościową lub dyskretną zmienną losową o a wartościach (zwanych również obiektami).
- W jednoczynnikowej (jednokierunkowej) analizie wariancji (ANOVA) pytamy, czy wartość średnia badanej cechy Y różni się istotnie w zależności od wartości zmiennej X.
- Zmienną X nazywamy zmienną objaśniającą lub zmienną grupującą, ponieważ jej poziomy określają a grup obserwacji.

6.1 Model i hipotezy

- Model analizy wariancji zapisujemy następująco:

$$y_{ij} = \mu_i + \varepsilon_{ij},$$

  gdzie $i = 1, \dots, a$, $j = 1, \dots, n_i$, y_{ij} oznacza j-tą obserwację dotyczącą i-tego obiektu, μ_i jest średnią wartością zmiennej Y w grupie i, a ε_{ij} jest błędem losowym.
- O błędach losowych zakładamy, że są niezależnymi zmiennymi losowymi o jednakowym rozkładzie normalnym $N(0, \sigma^2)$, gdzie wariancja $\sigma^2 > 0$ jest nieznanym parametrem.
- Mamy

$$ENW(\mu_i) = \bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}, \; i = 1, 2 \dots, a,$$

  oraz

$$ENW(\sigma^2) = \frac{1}{n} \sum_{i=1}^{a} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2,$$

  gdzie $n = n_1 + \dots + n_a$.
- Hipotezy zerowa i alternatywna są następujące:

$$\begin{cases} H_0 : \mu_1 = \dots = \mu_a, \\ H_1 : \neg H_0. \end{cases}$$

6.2 Test statystyczny

- Test weryfikujący powyższy układu hipotez wyznacza się metodą analizy wariancji, która oparta jest o następującą zależność:

$$\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2=\sum_{i=1}^{a}n_i(\bar{y}_i-\bar{y})^2+\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2$$

zwaną zależnością analizy wariancji, gdzie

$$\bar{y}=\frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{n_i}y_{ij}.$$

- Suma po lewej stronie tej zależności jest sumą kwadratów odchyleń poszczególnych obserwacji od średniej ogólnej. Nazywamy ją sumą kwadratów dla całości (ang. *total sum of squares*) i oznaczamy przez SST. Dzieląc tę sumę kwadratów przez liczbę obserwacji, otrzymujemy wariancję, którą uważa się za miarę rozproszenia wszystkich pomiarów.
- Pierwsza suma po prawej stronie tej tożsamości określa sumę kwadratów odchyleń średnich $\bar{y}_i$, $i=1,\dots,a$ od średniej ogólnej, a nazywamy ją sumą kwadratów dla obiektów (ang. *treatment sum of squares*), ozn. $SSTR$.
- Drugą sumę po prawej stronie powyższej równości nazywamy sumą kwadratów dla błędu (ang. *error sum of squares*) i oznaczamy przez SSE. Dzieląc sumą kwadratów dla błędu przez n, uzyskujemy estymator największej wiarogodności wariancji σ^2.
- Tabela analizy wariancji

Źródło zmienności	Stopnie swobody (DF)	Suma kwadratów (SS)	Średni kwadrat (MS)
Obiekty	$a-1$	$SSTR$	$MSTR=SSTR/(a-1)$
Błąd	$n-a$	SSE	$MSE=SSE/(n-a)$
Całość	$n-1$	SST	

- Przy prawdziwości hipotezy zerowej:

$$SSTR\sim\chi^2(a-1),\quad SSE\sim\chi^2(n-a),\quad SST\sim\chi^2(n-1).$$

- Obszar krytyczny testu F jest postaci:

$$R=\left\{(y_{ij}):\frac{MSTR}{MSE}>F(1-\alpha,a-1,n-a)\right\},$$

gdzie

$$\left.\frac{MSTR}{MSE}\right|_{H_0}\sim F(a-1,n-a)$$

oraz $F(\beta,m,n)$ oznacza kwantyl rzędu β z rozkładu F-Snedecora $F(m,n)$ z m i n stopniami swobody.
- p-wartość ma postać:

$$P\left(F_{a-1,n-a}>\frac{MSTR}{MSE}\right)=1-F_{F_{a-1,n-a}}\left(\frac{MSTR}{MSE}\right).$$

Przykład. Zbiór danych `vaccination` z pakietu `PBImisc` zawiera dane opisujące reakcję organizmu na zwalczanie wirusa po podaniu określonej dawki leku. Problem praktyczny dotyczy ustalenia, jaką najmniejszą możliwą dawkę leku należy podać, aby wywołać pożądaną reakcję organizmu (zagadnienie najmniejszej dawki leku). Rozważane jest również zagadnienie maksymalnej bezpiecznej dawki, którego celem jest określenie, jaka maksymalna dawka może być przyjmowana bez dużego ryzyka wystąpienia efektów ubocznych.

```
library(PBImisc)
head(vaccination)
```

```
##   response dose
## 1     88.9 0 ml
## 2    105.0 0 ml
## 3    138.4 0 ml
## 4     98.1 0 ml
## 5    107.2 0 ml
## 6     57.9 0 ml
```

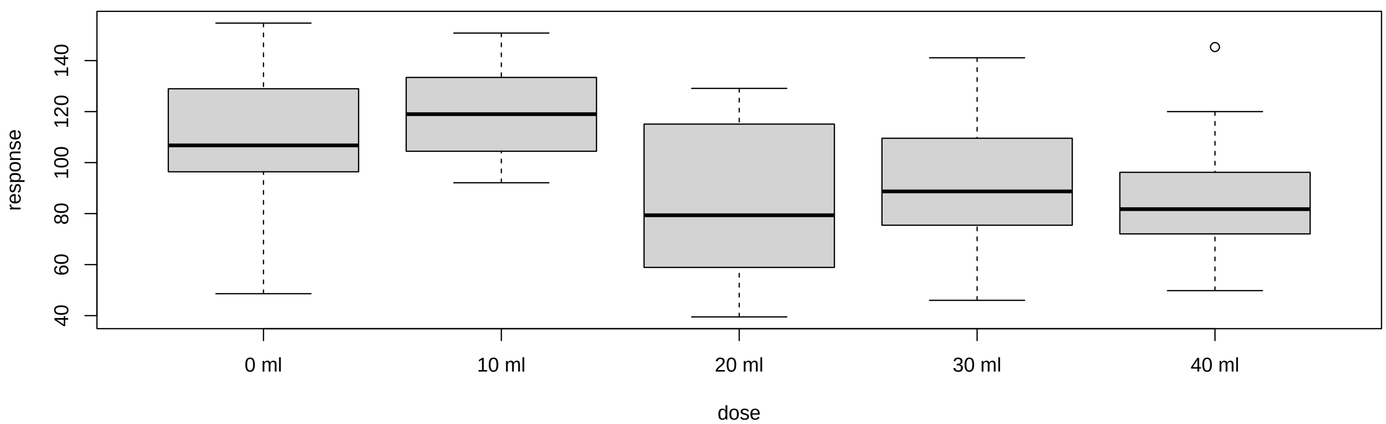
```
summary(vaccination)
```

```
##     response         dose
##  Min.   : 39.50   0 ml :20
##  1st Qu.: 77.30   10 ml:20
##  Median : 99.25   20 ml:20
##  Mean   : 97.89   30 ml:20
##  3rd Qu.:117.70   40 ml:20
##  Max.   :154.70
```

```
aggregate(vaccination$response,
          list(DOSE = vaccination$dose),
          FUN = mean)
```

```
##    DOSE       x
## 1  0 ml 108.570
## 2 10 ml 119.265
## 3 20 ml  84.025
## 4 30 ml  92.370
## 5 40 ml  85.220
```

```
boxplot(response ~ dose, data = vaccination)
```




```
summary(aov(response ~ dose, data = vaccination))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## dose         4  19084    4771   7.929 1.47e-05 ***
## Residuals   95  57164     602
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

6.3 Założenia

- Jednym z założeń modelu analizy wariancji była normalność błędów losowych. W celu jej zbadania wykonujemy test Shapiro-Wilka dla reszt, tj. $y_{ij} - \bar{y}_i$ dla $i = 1, \ldots, a$, $j = 1, \ldots, n_a$.
- Gdyby rozkład reszt był istotnie odległy od normalnego, to można by wykonać nieparametryczną jednokierunkową analizę wariancji, czyli test Kruskala-Wallisa. Test ten bada równość parametrów położenia w a populacjach.

Przykład (cd.).

```r
shapiro.test(lm(response ~ dose, data = vaccination)$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lm(response ~ dose, data = vaccination)$residuals
## W = 0.99244, p-value = 0.8524
```

- W analizie wariancji zakładamy również równość wariancji błędów losowych w poszczególnych grupach. W celu weryfikacji tego założenia wykorzystuje się testy jednorodności wariancji w grupach.
- Takie testy to:
  - test Bartletta,
  - test Flingera-Killeena,
  - test Levene'a,
  - test Browna-Forsytha (modyfikacja testu Levene'a, w której parametr położenia wyznaczany jest przez mediany a nie przez średnie).
- Wszystkie te testy weryfikują hipotezę zerową o równości wariancji w a populacjach.
- Często stosowanym testem jest test Bartletta, który jest alternatywą dla testu Levene'a.
- Test Bartletta ma wyższą moc, jeżeli dane pochodzą z rozkładu normalnego.
- Natomiast, test Levene'a jest bardziej odporny na brak normalności i w takich przypadkach daje bardziej wiarygodne wyniki.
- Choć metody analizy wariancji można stosować przy niewielkich odstępstwach od normalności, testu Bartletta nie powinno się stosować nawet dla małych odstępstw od normalności.

Przykład (cd.).

```r
bartlett.test(response ~ dose, data = vaccination)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  response by dose
## Bartlett's K-squared = 5.6387, df = 4, p-value = 0.2278
```

```r
fligner.test(response ~ dose, data = vaccination)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
```

```
## data:  response by dose
## Fligner-Killeen:med chi-squared = 4.8066, df = 4, p-value = 0.3077
```

```r
library(car)
leveneTest(response ~ dose, data = vaccination)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  4  1.3679 0.2509
##       95
```

```r
leveneTest(response ~ dose, data = vaccination, center = "mean")
```

```
## Levene's Test for Homogeneity of Variance (center = "mean")
##       Df F value Pr(>F)
## group  4  1.6203 0.1755
##       95
```

6.4 Analiza post hoc

- W przypadku odrzucenia hipotezy zerowej, przyjmujemy hipotezę alternatywną, że co najmniej dwie średnie się różnią. Jednak nie wiemy ani które średnie się różnią, ani która jest większa. Aby ocenić, które średnie się różnią, wykonuje się w drugim kroku analizy testy **post hoc** (po fakcie) porównujące wszystkie pary średnich:
$$\begin{cases} H_0 : \mu_i = \mu_j, \\ H_1 : \mu_i \neq \mu_j. \end{cases}$$
- Funkcja `pairwise.t.test()` przeprowadza test post hoc, który polega na wyznaczeniu p-wartości dla testu t-Studenta dla dwóch prób niezależnych i koryguje je, uwzględniając korektę Holma na liczbę hipotez. Jest to zagadnienie testowania zbioru hipotez.
- Test HSD Tukeya (ang. honestly significant differences, pol. uczciwie istotnych różnic) jest jednym z najpopularniejszych testów post hoc. Test ten jest konstruowany przy założeniu równych liczebności grup, tj. $n_1 = \cdots = n_a = m$, ale w praktyce niewielkie odstępstwa od tego założenia są dopuszczalne. Obszar krytyczny testu HSD Tukeya dla powyższego układu hipotez ma następującą postać:
$$R_{ij} = \left\{ (y_{ij}) : \sqrt{m}\frac{|\bar{y}_i - \bar{y}_j|}{\sqrt{MSE}} > q(1-\alpha, a, n-a) \right\},$$
gdzie $q(\beta, k, l)$ oznacza kwantyl rzędu α z rozkładu studentyzowanego rozstępu.
- Test Studenta-Newmana-Keulsa ma podobną konstrukcję do testu Tukeya, z jedną różnicą. Mianowicie w teście Tukeya dla każdej pary średnich stosuje się ten sam kwantyl studentyzowanego rozstępu dla a grup. Natomiast w teście Studenta-Newmana-Keulsa średnie w pierwszym kroku są sortowane, następnie jeżeli porównujemy średnią $\bar{y}_{1:a}$ (najmniejszą) z $\bar{y}_{a:a}$ (największą), to stosuje się rozkład studentyzowanego rozstępu dla a grup. Jednak dla innych średnich, np. porównując $\bar{y}_{i:a}$ z $\bar{y}_{j:a}$, stosuje się rozkład studentyzowany dla $|i-j|+1$ grup. Taka procedura testowa pozwala na znalezienie większej liczby różnic między średnimi, ale nie umożliwia kontroli łącznego błędu I rodzaju.
- Test LSD Fishera (ang. least significant differences, pol. test najmniejszych istotnych różnic) polega na wykonaniu $a(a-1)/2$ testów t-Studenta przez porównanie każdej pary średnich i zastosowaniu korekty na liczbę przeprowadzonych testów/wielokrotne testowanie. Przy czym w statystyce testowej testu t-Studenta za estymator wariancji przyjmuje się estymator skonstruowany na podstawie wszystkich prób, a nie tylko tych dwóch branych pod uwagę. Do korekty można wykorzystać poprawkę Bonferroniego,

Holma lub inne wymienione w wektorze `p.adjust.methods`. Test ten może być stosowany przy różnych liczebnościach grup.

- Test Scheffego to najbardziej konserwatywny test. Jest podobny do testu LSD Fishera, ale są tu uwzględnione wszystkie możliwe kontrasty (w sensie te liniowo niezależne). Z tego względu, mimo konserwatywności, jest on używany w sytuacji, gdy porównywane są „nieplanowane” kontrasty. Nie zakłada się tutaj równych liczebności grup.

Przykład (cd.).

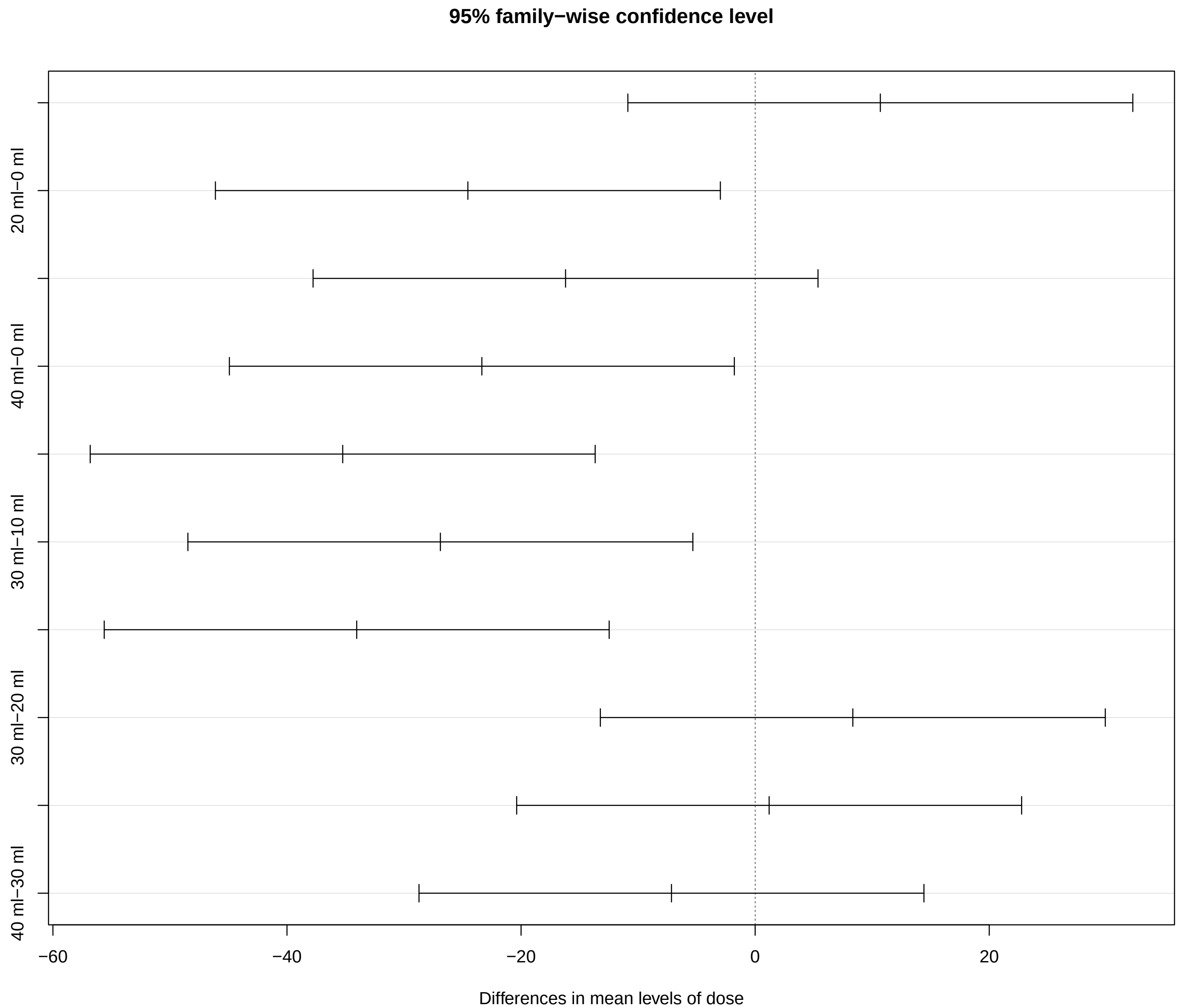
```
attach(vaccination)
pairwise.t.test(response, dose, data = vaccination)
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  response and dose
##
##       0 ml    10 ml   20 ml   30 ml
## 10 ml 0.68485 -       -       -
## 20 ml 0.01463 0.00016 -       -
## 30 ml 0.19718 0.00633 0.85424 -
## 40 ml 0.02007 0.00027 0.87790 0.85424
##
## P value adjustment method: holm
```

```
model_aov <- aov(response ~ dose, data = vaccination)
TukeyHSD(model_aov)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = response ~ dose, data = vaccination)
##
## $dose
##                diff       lwr        upr     p adj
## 10 ml-0 ml   10.695 -10.87643  32.266431 0.6426874
## 20 ml-0 ml  -24.545 -46.11643  -2.973569 0.0174170
## 30 ml-0 ml  -16.200 -37.77143   5.371431 0.2336465
## 40 ml-0 ml  -23.350 -44.92143  -1.778569 0.0270291
## 20 ml-10 ml -35.240 -56.81143 -13.668569 0.0001562
## 30 ml-10 ml -26.895 -48.46643  -5.323569 0.0069317
## 40 ml-10 ml -34.045 -55.61643 -12.473569 0.0002808
## 30 ml-20 ml   8.345 -13.22643  29.916431 0.8185005
## 40 ml-20 ml   1.195 -20.37643  22.766431 0.9998712
## 40 ml-30 ml  -7.150 -28.72143  14.421431 0.8878461
```

```
plot(TukeyHSD(model_aov))
```




```r
library(agricolae)
HSD.test(model_aov, "dose", console = TRUE)
```

```
##
## Study: model_aov ~ "dose"
##
## HSD Test for response
##
## Mean Square Error:  601.7253
##
## dose,  means
##
##       response      std  r  Min   Max
## 0 ml   108.570 25.91789 20 48.6 154.7
## 10 ml  119.265 17.64743 20 92.1 150.8
## 20 ml   84.025 30.42350 20 39.5 129.1
## 30 ml   92.370 24.27206 20 46.0 141.1
## 40 ml   85.220 22.59946 20 49.8 145.3
##
```

```
## Alpha: 0.05 ; DF Error: 95
## Critical Value of Studentized Range: 3.932736
##
## Minimun Significant Difference: 21.57143
##
## Treatments with the same letter are not significantly different.
##
##        response groups
## 10 ml  119.265      a
## 0 ml   108.570     ab
## 30 ml   92.370     bc
## 40 ml   85.220      c
## 20 ml   84.025      c
```

```
SNK.test(model_aov, "dose", console = TRUE)
```

```
##
## Study: model_aov ~ "dose"
##
## Student Newman Keuls Test
## for response
##
## Mean Square Error:  601.7253
##
## dose,  means
##
##        response      std  r  Min   Max
## 0 ml   108.570 25.91789 20 48.6 154.7
## 10 ml  119.265 17.64743 20 92.1 150.8
## 20 ml   84.025 30.42350 20 39.5 129.1
## 30 ml   92.370 24.27206 20 46.0 141.1
## 40 ml   85.220 22.59946 20 49.8 145.3
##
## Alpha: 0.05 ; DF Error: 95
##
## Critical Range
##        2        3        4        5
## 15.39978 18.46964 20.28552 21.57143
##
## Means with the same letter are not significantly different.
##
##        response groups
## 10 ml  119.265      a
## 0 ml   108.570      a
## 30 ml   92.370      b
## 40 ml   85.220      b
## 20 ml   84.025      b
```

```
LSD.test(model_aov, "dose", p.adj = "holm", console = TRUE)
```

```
##
## Study: model_aov ~ "dose"
##
```

```
## LSD t Test for response
## P value adjustment method: holm
##
## Mean Square Error:  601.7253
##
## dose,  means and individual ( 95 %) CI
##
##       response      std  r       LCL       UCL  Min   Max
## 0 ml   108.570 25.91789 20  97.68071 119.45929 48.6 154.7
## 10 ml  119.265 17.64743 20 108.37571 130.15429 92.1 150.8
## 20 ml   84.025 30.42350 20  73.13571  94.91429 39.5 129.1
## 30 ml   92.370 24.27206 20  81.48071 103.25929 46.0 141.1
## 40 ml   85.220 22.59946 20  74.33071  96.10929 49.8 145.3
##
## Alpha: 0.05 ; DF Error: 95
## Critical Value of t: 2.874073
##
## Minimum Significant Difference: 22.29446
##
## Treatments with the same letter are not significantly different.
##
##       response groups
## 10 ml  119.265      a
## 0 ml   108.570     ab
## 30 ml   92.370     bc
## 40 ml   85.220      c
## 20 ml   84.025      c
```

```r
scheffe.test(model_aov, "dose", console = TRUE)
```

```
##
## Study: model_aov ~ "dose"
##
## Scheffe Test for response
##
## Mean Square Error  : 601.7253
##
## dose,  means
##
##       response      std  r  Min   Max
## 0 ml   108.570 25.91789 20 48.6 154.7
## 10 ml  119.265 17.64743 20 92.1 150.8
## 20 ml   84.025 30.42350 20 39.5 129.1
## 30 ml   92.370 24.27206 20 46.0 141.1
## 40 ml   85.220 22.59946 20 49.8 145.3
##
## Alpha: 0.05 ; DF Error: 95
## Critical Value of F: 2.467494
##
## Minimum Significant Difference: 24.37009
##
## Means with the same letter are not significantly different.
```

```
##
##       response groups
## 10 ml  119.265      a
## 0 ml   108.570     ab
## 30 ml   92.370     bc
## 40 ml   85.220     bc
## 20 ml   84.025      c
```

6.5 Analiza kontrastów

- Nie zawsze jesteśmy zainteresowani porównywaniem wszystkich par średnich. W wielu sytuacjach chcemy porównać wybrane średni lub grupy średnich pomiędzy sobą.
- Do porównywania wybranych grup średnich służy analiza kontrastów.
- Kontrastem nazywamy liniową funkcję średnich μ_i, tj.

$$L = \sum_{i=1}^{a} c_i \mu_i,$$

przy czym $\sum_{i=1}^{a} c_i = 0$.

- Niech $L = \mathbf{c}^\top \boldsymbol{\mu}$ będzie kontrastem, gdzie $\mathbf{c} = (c_1, \dots, c_a)^\top$ oraz $\boldsymbol{\mu} = (\mu_1, \dots, \mu_a)^\top$, a ponadto niech

$$SSL = \frac{\left(\sum_{i=1}^{a} c_i \bar{y}_i\right)^2}{\sum_{i=1}^{a} \frac{c_i^2}{n_i}}.$$

- Weryfukujemy układ hipotez

$$H_0^L: L = 0, \qquad H_1^L: L \neq 0$$

testem o obszarze krytycznym postaci:

$$R = \left\{(y_{ij}): F_L = \frac{SSL}{MSE} > F(1-\alpha, 1, n-a)\right\}.$$

- Rozszerzona tabela analizy wariancji

Źródło zmienności	Stopnie swobody (DF)	Suma kwadratów (SS)	Średni kwadrat (MS)
Obiekty	$a-1$	$SSTR$	$MSTR = SSTR/(a-1)$
$L1$	1	$SSL1$	$SSL1$
…	…	…	…
$L(a-1)$	1	$SSL(a-1)$	$SSL(a-1)$
Błąd	$n-a$	SSE	$MSE = SSE/(n-a)$
Całość	$n-1$	SST	

- Przykładowe kontrasty wbudowane w programie R:

```
contr.helmert(5)
```

```
##   [,1] [,2] [,3] [,4]
## 1   -1   -1   -1   -1
## 2    1   -1   -1   -1
## 3    0    2   -1   -1
```

```
## 4    0    0    3   -1
## 5    0    0    0    4
```

```r
library(multcomp)
# kontrasty dla postepujących różnic
contr.sdif(5)
```

```
##    2-1  3-2  4-3  5-4
## 1 -0.8 -0.6 -0.4 -0.2
## 2  0.2 -0.6 -0.4 -0.2
## 3  0.2  0.4 -0.4 -0.2
## 4  0.2  0.4  0.6 -0.2
## 5  0.2  0.4  0.6  0.8
```

Przykład (cd.).

```r
model.1 <- aov(response ~ dose, data = vaccination)
summary(model.1)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## dose         4  19084    4771   7.929 1.47e-05 ***
## Residuals   95  57164     602                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
contrasts(vaccination$dose) <- contr.sdif(5)
vaccination$dose
```

```
##   [1] 0 ml  0 ml  0 ml  0 ml  0 ml  0 ml  0 ml  0 ml  0 ml  0 ml  0 ml  0 ml 
##  [13] 0 ml  0 ml  0 ml  0 ml  0 ml  0 ml  0 ml  0 ml  10 ml 10 ml 10 ml 10 ml
##  [25] 10 ml 10 ml 10 ml 10 ml 10 ml 10 ml 10 ml 10 ml 10 ml 10 ml 10 ml 10 ml
##  [37] 10 ml 10 ml 10 ml 10 ml 20 ml 20 ml 20 ml 20 ml 20 ml 20 ml 20 ml 20 ml
##  [49] 20 ml 20 ml 20 ml 20 ml 20 ml 20 ml 20 ml 20 ml 20 ml 20 ml 20 ml 20 ml
##  [61] 30 ml 30 ml 30 ml 30 ml 30 ml 30 ml 30 ml 30 ml 30 ml 30 ml 30 ml 30 ml
##  [73] 30 ml 30 ml 30 ml 30 ml 30 ml 30 ml 30 ml 30 ml 40 ml 40 ml 40 ml 40 ml
##  [85] 40 ml 40 ml 40 ml 40 ml 40 ml 40 ml 40 ml 40 ml 40 ml 40 ml 40 ml 40 ml
##  [97] 40 ml 40 ml 40 ml 40 ml
## attr(,"contrasts")
##        2-1  3-2  4-3  5-4
## 0 ml  -0.8 -0.6 -0.4 -0.2
## 10 ml  0.2 -0.6 -0.4 -0.2
## 20 ml  0.2  0.4 -0.4 -0.2
## 30 ml  0.2  0.4  0.6 -0.2
## 40 ml  0.2  0.4  0.6  0.8
## Levels: 0 ml 10 ml 20 ml 30 ml 40 ml
```

```r
model.2 <- aov(response ~ dose, data = vaccination)
summary(model.2,
        split = list(dose = list('C1' = 1, 'C2' = 2, 'C3' = 3, 'C4' = 4)))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## dose         4  19084    4771   7.929 1.47e-05 ***
##   dose: C1   1   2852    2852   4.739    0.032 *  
##   dose: C2   1  15418   15418  25.622 2.03e-06 ***
##   dose: C3   1    303     303   0.504    0.479
```

```
##   dose: C4   1    511     511   0.850    0.359
## Residuals   95  57164     602
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

6.6 Test Kruskala-Wallisa

- Nieparametryczną alternatywą dla jednoczynnikowej analizy wariancji jest test Kruskala-Wallisa (ang. Kruskal-Wallis test). Test ten jest uogólnieniem testu U-Manna-Whitneya na więcej niż dwie populacje. Wykorzystuje on rangi.

Przykład (cd.).

```
aggregate(vaccination$response,
          list(DOSE = vaccination$dose),
          FUN = median)
```

```
##    DOSE      x
## 1  0 ml 106.75
## 2 10 ml 119.00
## 3 20 ml  79.35
## 4 30 ml  88.70
## 5 40 ml  81.75
```

```
kruskal.test(response ~ dose, data = vaccination)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  response by dose
## Kruskal-Wallis chi-squared = 25.709, df = 4, p-value = 3.622e-05
```

```
# testy post hoc wykorzystujące test Manna-Whitneya dla dwóch prób
pairwise.wilcox.test(response, dose, data = vaccination)
```

```
## Warning in wilcox.test.default(xi, xj, paired = paired, ...): cannot compute
## exact p-value with ties

## Warning in wilcox.test.default(xi, xj, paired = paired, ...): cannot compute
## exact p-value with ties

## Warning in wilcox.test.default(xi, xj, paired = paired, ...): cannot compute
## exact p-value with ties

## Warning in wilcox.test.default(xi, xj, paired = paired, ...): cannot compute
## exact p-value with ties

## Warning in wilcox.test.default(xi, xj, paired = paired, ...): cannot compute
## exact p-value with ties

## Warning in wilcox.test.default(xi, xj, paired = paired, ...): cannot compute
## exact p-value with ties

##
```

```
##  Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  response and dose
##
##       0 ml   10 ml   20 ml  30 ml
## 10 ml 0.8337 -       -      -
## 20 ml 0.1336 0.0047  -      -
## 30 ml 0.1802 0.0042  0.9119 -
## 40 ml 0.0150 5.8e-05 0.9467 0.9119
##
## P value adjustment method: holm
```

```
# test Dunna
library(FSA)
dunnTest(response ~ dose, data = vaccination, method = "bh")
```

```
##       Comparison          Z      P.unadj        P.adj
## 1   0 ml - 10 ml -1.2290032 2.190706e-01 0.3129580268
## 2   0 ml - 20 ml  2.6078848 9.110362e-03 0.0182207235
## 3  10 ml - 20 ml  3.8368879 1.246033e-04 0.0006230165
## 4   0 ml - 30 ml  1.9702202 4.881314e-02 0.0813552412
## 5  10 ml - 30 ml  3.1992233 1.377984e-03 0.0045932794
## 6  20 ml - 30 ml -0.6376646 5.236920e-01 0.5818800276
## 7   0 ml - 40 ml  2.9185419 3.516726e-03 0.0087918156
## 8  10 ml - 40 ml  4.1475451 3.360594e-05 0.0003360594
## 9  20 ml - 40 ml  0.3106571 7.560613e-01 0.7560612988
## 10 30 ml - 40 ml  0.9483217 3.429657e-01 0.4287071123
```

6.7 Zadania 6

Zadanie 1. Zadanie to zostało opracowane na podstawie eksperymentu Smitha (1979). Głównym jego celem było pokazanie, że bycie w tym samym kontekście psychicznym w czasie nauki i podczas jej sprawdzania (test, egzamin) daje lepsze wyniki niż bycie w odmiennych kontekstach. Podczas fazy uczącej uczniowie uczyli się 80 słów w pokoju pomalowanym na pomarańczowo, ozdobionym plakatami, obrazami i dużą ilością dodatkowych akcesoriów. Pierwszy sprawdzian pamięci został przeprowadzony aby dać uczniom wrażenie, że eksperyment się zakończył. Następnego dnia, uczniowie zostali niespodziewanie poddani testowi ponownie. Mieli napisać wszystkie słowa, które zapamiętali. Test został przeprowadzony w 5 różnych warunkach. 50 uczniów zostało losowo podzielonych na 5 równolicznych grup:

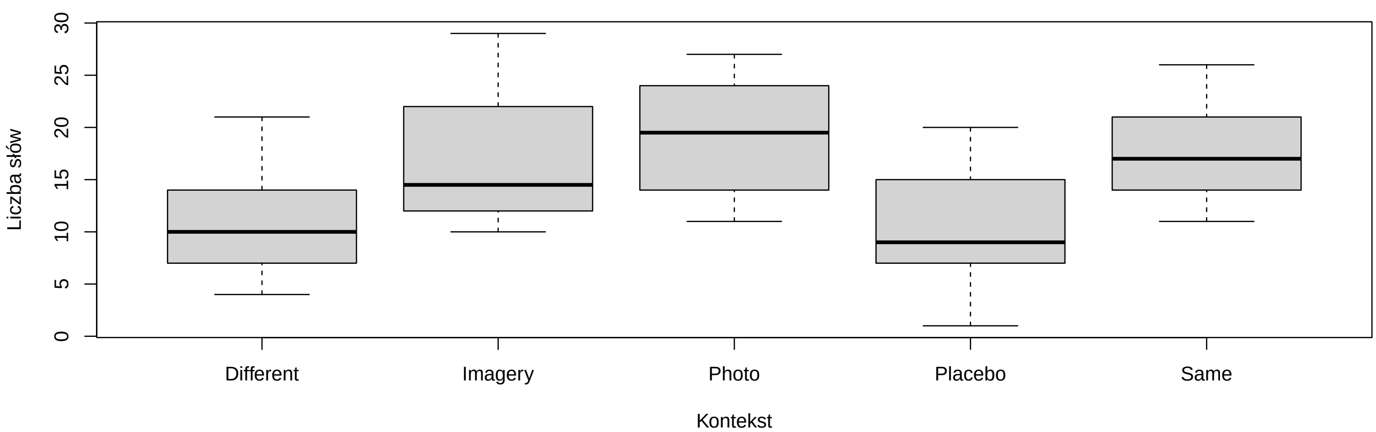
- „Same context” - test odbywał się w tym samym pokoju, w którym się uczyli.
- „Different context” - test odbywał się w bardzo odmiennym pomieszczeniu, w innej części kampusu, pomalowanym na szaro i wyglądającym bardzo surowo.
- „Imaginary context” - test odbywał się w tym samym pomieszczeniu, co w punkcie poprzednim. Dodatkowo, uczniowie mieli przypomnieć sobie pokój, w którym się uczyli. Aby im w tym pomóc badacz zadawał dodatkowe pytania o pokoju i jego wyposażeniu.
- „Photographed context” - test odbywał się w tych samych warunkach, co w punkcie poprzednim. Dodatkowo pokazano im zdjęcie pokoju, w którym się uczyli.
- „Placebo context” - test odbywał się w tym samych warunkach co grupy „Different context‘‘. Dodatkowo uczniowie wykonali ćwiczenia „rozgrzewające” (przypominanie sobie swojego salonu).

Liczba zapamiętanych słów została zawarta w poniższej tabeli.

Same	Different	Imagery	Photo	Placebo
25	11	14	25	8
26	21	15	15	20
17	9	29	23	10
15	6	10	21	7
14	7	12	18	15
17	14	22	24	7
14	12	14	14	1
20	4	20	27	17
11	7	22	12	11
21	19	12	11	4

(1) Wyznacz średnie liczb zapamiętanych słów w grupach. Ponadto, przedstaw otrzymane dane za pomocą wykresu ramkowego dla każdej grupy z osobna.

```
##     CONTEXT  x
## 1 Different 11
## 2   Imagery 17
## 3     Photo 19
## 4   Placebo 10
## 5      Same 18
```

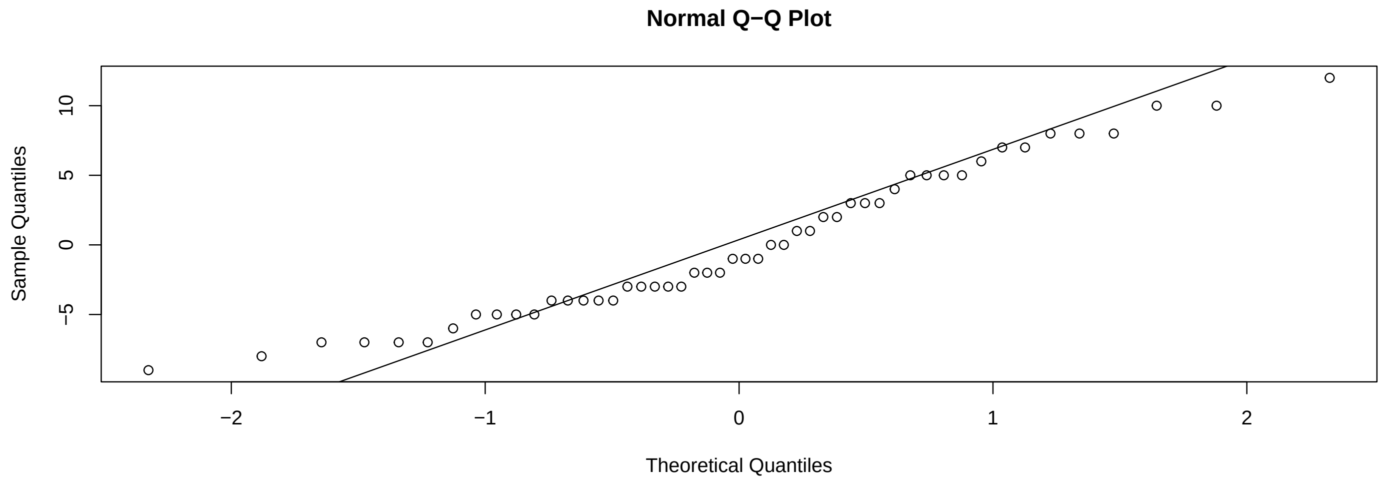


(2) Wykonaj test analizy wariancji w celu sprawdzenia, czy liczba zapamiętanych słów zależy od kontekstu sprawdzania wiedzy.

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## context      4    700     175   5.469 0.00112 **
## Residuals   45   1440      32
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(3) Sprawdź założenia modelu jednoczynnikowej analizy wariancji.

```
## [1] 0.05635956
```




```
## [1] 0.9817694
```

```
## [1] 0.9759731
```

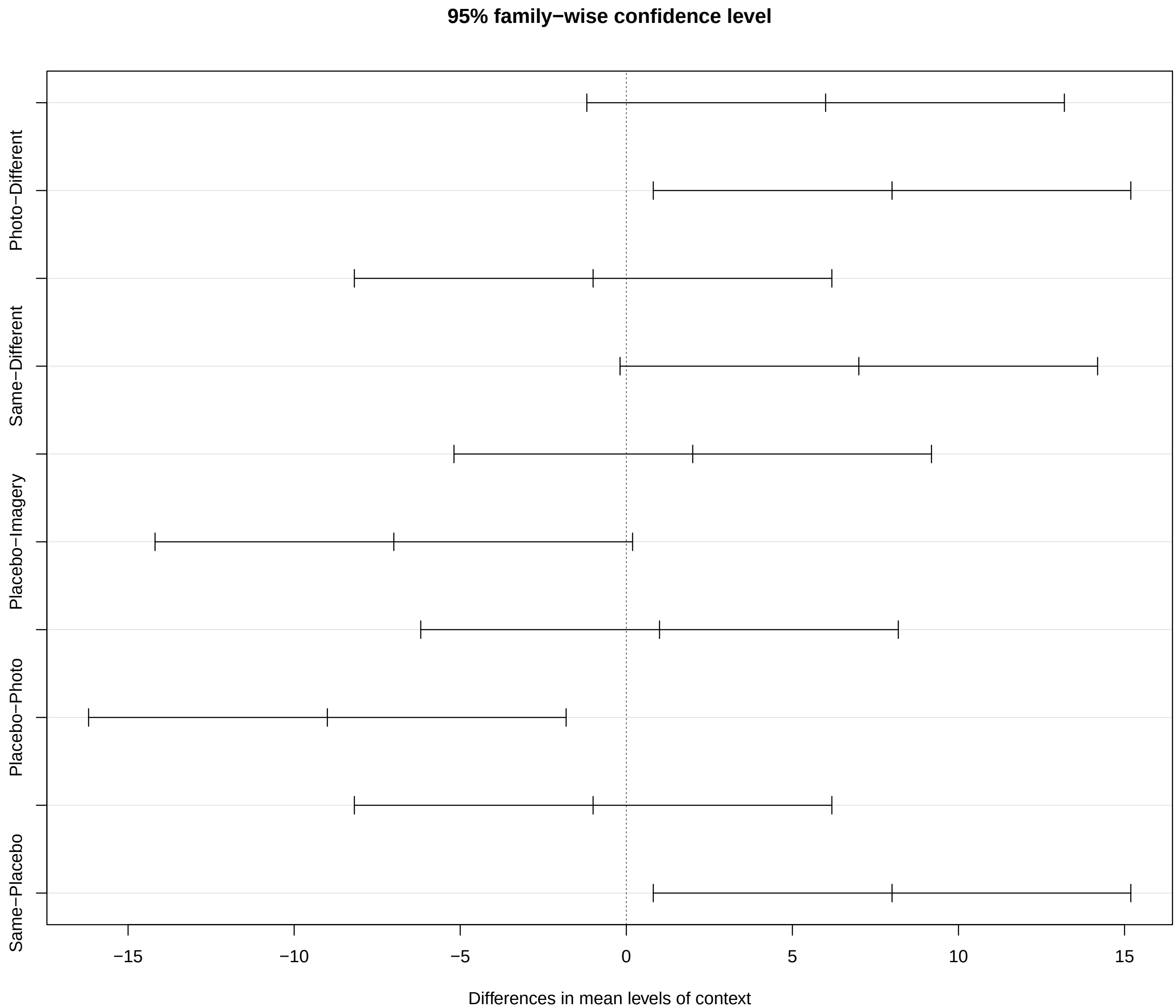
```
## [1] 0.9550502
```

```
## [1] 0.9281122
```

(4) Wykonaj testy post hoc w celu sprawdzenia, które konteksty sprawdzania wiedzy różnią się między sobą.

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  number and context
##
##         Different Imagery Photo Placebo
## Imagery 0.110     -       -     -
## Photo   0.025     1.000   -     -
## Placebo 1.000     0.057   0.009 -
## Same    0.057     1.000   1.000 0.025
##
## P value adjustment method: holm
```

```
##                   diff        lwr       upr      p adj
## Imagery-Different    6  -1.188363 13.188363 0.14198584
## Photo-Different      8   0.811637 15.188363 0.02232998
## Placebo-Different   -1  -8.188363  6.188363 0.99466042
## Same-Different       7  -0.188363 14.188363 0.05967870
## Photo-Imagery        2  -5.188363  9.188363 0.93203553
## Placebo-Imagery     -7 -14.188363  0.188363 0.05967870
## Same-Imagery         1  -6.188363  8.188363 0.99466042
## Placebo-Photo       -9 -16.188363 -1.811637 0.00759672
## Same-Photo          -1  -8.188363  6.188363 0.99466042
## Same-Placebo         8   0.811637 15.188363 0.02232998
```




```
##
## Study: model_aov ~ "context"
##
## HSD Test for number
##
## Mean Square Error:  32
##
## context,  means
##
##           number      std  r Min Max
## Different     11 5.617433 10   4  21
## Imagery       17 6.000000 10  10  29
## Photo         19 5.773503 10  11  27
## Placebo       10 5.906682 10   1  20
## Same          18 4.921608 10  11  26
##
## Alpha: 0.05 ; DF Error: 45
## Critical Value of Studentized Range: 4.018417
```

```
##
## Minimun Significant Difference: 7.188363
##
## Treatments with the same letter are not significantly different.
##
##           number groups
## Photo         19      a
## Same          18     ab
## Imagery       17    abc
## Different     11     bc
## Placebo       10      c

##
## Study: model_aov ~ "context"
##
## Student Newman Keuls Test
## for number
##
## Mean Square Error:  32
##
## context,  means
##
##           number      std  r Min Max
## Different     11 5.617433 10   4  21
## Imagery       17 6.000000 10  10  29
## Photo         19 5.773503 10  11  27
## Placebo       10 5.906682 10   1  20
## Same          18 4.921608 10  11  26
##
## Alpha: 0.05 ; DF Error: 45
##
## Critical Range
##        2        3        4        5
## 5.095323 6.131311 6.748805 7.188363
##
## Means with the same letter are not significantly different.
##
##           number groups
## Photo         19      a
## Same          18      a
## Imagery       17      a
## Different     11      b
## Placebo       10      b

##
## Study: model_aov ~ "context"
##
## LSD t Test for number
## P value adjustment method: holm
##
## Mean Square Error:  32
##
```

```
## context,  means and individual ( 95 %) CI
##
##           number      std  r       LCL      UCL Min Max
## Different     11 5.617433 10  7.397062 14.60294   4  21
## Imagery       17 6.000000 10 13.397062 20.60294  10  29
## Photo         19 5.773503 10 15.397062 22.60294  11  27
## Placebo       10 5.906682 10  6.397062 13.60294   1  20
## Same          18 4.921608 10 14.397062 21.60294  11  26
##
## Alpha: 0.05 ; DF Error: 45
## Critical Value of t: 2.952079
##
## Minimum Significant Difference: 7.468235
##
## Treatments with the same letter are not significantly different.
##
##           number groups
## Photo         19      a
## Same          18     ab
## Imagery       17    abc
## Different     11     bc
## Placebo       10      c
```

```
##
## Study: model_aov ~ "context"
##
## Scheffe Test for number
##
## Mean Square Error  : 32
##
## context,  means
##
##           number      std  r Min Max
## Different     11 5.617433 10   4  21
## Imagery       17 6.000000 10  10  29
## Photo         19 5.773503 10  11  27
## Placebo       10 5.906682 10   1  20
## Same          18 4.921608 10  11  26
##
## Alpha: 0.05 ; DF Error: 45
## Critical Value of F: 2.578739
##
## Minimum Significant Difference: 8.125006
##
## Means with the same letter are not significantly different.
##
##           number groups
## Photo         19      a
## Same          18     ab
## Imagery       17     ab
## Different     11     ab
## Placebo       10      b
```

(5) Chcemy przetestować następujące hipotezy szczegółowe:

  a. Grupy o takim samym kontekście podczas uczenia i testowania („Same” lub „Imaginary” lub „Photographed‘‘) wypadają lepiej od grup o różnym kontekście („Different” lub „Placebo”).
  b. Grupa „Same” różni się od grup „Imaginary” i „Photographed”.
  c. Grupa „Imaginary” różni się od grupy „Photographed”.
  d. Grupa „Different” różni się od grupy „Placebo”.

W tym celu wykonaj następujące polecenia:

- Zapisz odpowiednie hipotezy.
- Wyraź je za pomocą kontrastów.
- Czy ten układ kontrastów jest ortogonalny?
- Przetestuj zaproponowane kontrasty.

```
##                Df Sum Sq Mean Sq F value   Pr(>F)    
## context         4    700     175   5.469  0.00112 ** 
##   context: C1  1    675     675  21.094 3.52e-05 ***
##   context: C2  1      0       0   0.000  1.00000    
##   context: C3  1     20      20   0.625  0.43334    
##   context: C4  1      5       5   0.156  0.69450    
## Residuals      45   1440      32                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(6) Wykonaj polecenia 1, 2 i 4 wykorzystując odpowiednie metody nieparametryczne. Porównaj ich wyniki z wynikami metod parametrycznych.

```
##     CONTEXT    x
## 1 Different 10.0
## 2   Imagery 14.5
## 3     Photo 19.5
## 4   Placebo  9.0
## 5      Same 17.0
```

```
## [1] 0.002603633
```

```
## 
##  Pairwise comparisons using Wilcoxon rank sum test with continuity correction 
## 
## data:  number and context 
## 
##         Different Imagery Photo Placebo
## Imagery 0.138     -       -     -      
## Photo   0.081     1.000   -     -      
## Placebo 1.000     0.138   0.057 -      
## Same    0.096     1.000   1.000 0.081  
## 
## P value adjustment method: holm
```

```
##             Comparison          Z     P.unadj      P.adj
## 1  Different - Imagery -2.0753448 0.037954590 0.06325765
## 2    Different - Photo -2.8055587 0.005022943 0.02511471
## 3      Imagery - Photo -0.7302139 0.465259440 0.66465634
## 4  Different - Placebo  0.1844751 0.853640764 0.85364076
## 5    Imagery - Placebo  2.2598199 0.023832431 0.04766486
```

```
## 6      Photo - Placebo  2.9900338 0.002789466 0.02789466
## 7     Different - Same -2.5288461 0.011443820 0.02860955
## 8       Imagery - Same -0.4535013 0.650187828 0.81273479
## 9         Photo - Same  0.2767126 0.782000765 0.86888974
## 10      Placebo - Same -2.7133212 0.006661251 0.02220417
```

Zadanie 2. W 1974 roku Michael Eysenck opublikował w czasopiśmie Developmental Psychology wyniki badań dotyczących ubocznego uczenia werbalnego. W eksperymencie wzięło udział 100 osób, z czego połowę stanowili młodzi ludzie (w wieku studenckim), a drugą połowę osoby starsze (w wieku pięćdziesięciu i sześćdziesięciu lat). W obrębie każdej grupy wiekowej, pacjenci zostali przydzieleni do jednej z pięciu grup „Instrukcji". Następnie podano im listę słów i powiedziano, aby postępowali zgodnie z instrukcjami podanymi wcześniej. Instrukcje były następujące:

- Liczenie - liczenie liter w każdym wymienionym słowie,
- Rymowanie - pomyśleć o słowie, które rymuje się z wskazanym słowem,
- Przymiotnik - pomyśleć o przymiotniku, który mógłby zmodyfikować dane słowo,
- Wyobraźnia - wyobrazić sobie obraz obiektu opisanego przez wymienione słowo,
- Kontrola - pamiętaj wymienione słowa aby później je powtórzyć.

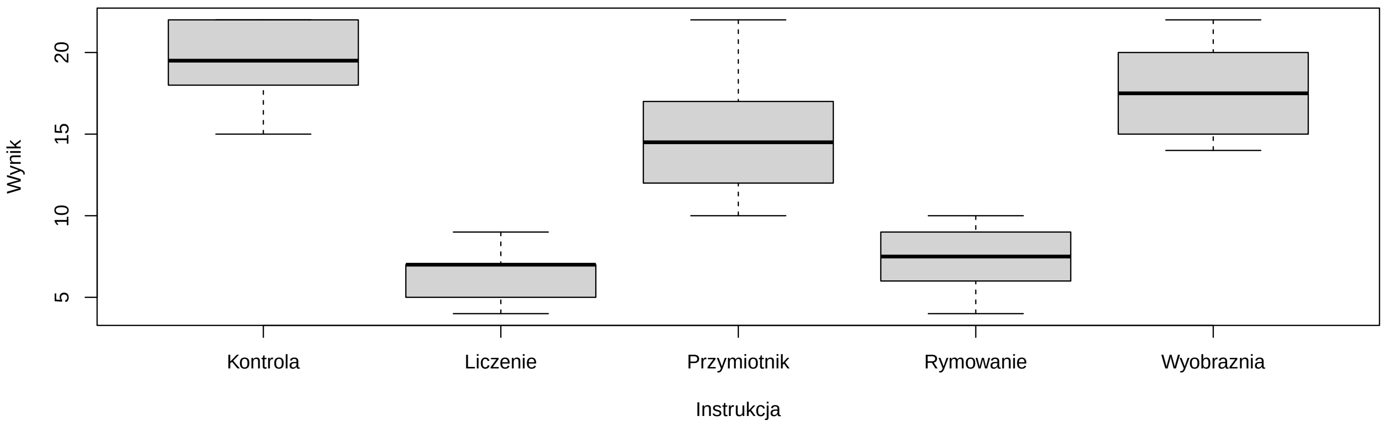
Każdy pacjent widział tę samą listę wyrazów trzy razy i powtarzał te instrukcje trzy razy. Instrukcje Liczenie i Rymowanie mają dać informację o powierzchownym poziomie przetwarzania semantycznego. Instrukcje Przymiotnik i Wyobraźnia mają informować o głębokim poziomie przetwarzania semantycznego, tj. liczenie i rymowanie nie wymagają od pacjenta znajomości sensu słów z listy, podczas gdy instrukcje Przymiotnik i Wyobraźnia wymagają znajomości znaczenia słów. Pacjenci w grupie kontrolnej mieli tylko zapamiętać słowa i ewentualnie później je powtórzyć. Dane zawarte w pliku Eysenck.txt dotyczą tylko pacjentów młodszych i zostały uzyskane w oparciu o średnie i błędy standardowe otrzymane w pracy Eysencka (1974).

(1) Załaduj zbiór danych do programu R. Następnie usuń zbędną kolumnę.

```
##   Wynik Instrukcja
## 1     7   Liczenie
## 2     9   Liczenie
## 3     7   Liczenie
## 4     7   Liczenie
## 5     5   Liczenie
## 6     7   Liczenie
```

(2) Wyznacz średnie wartości cechy zależnej w grupach. Ponadto, przedstaw otrzymane dane za pomocą wykresu ramkowego dla każdej grupy z osobna.

```
##    Instrukcja    x
## 1    Kontrola 19.3
## 2    Liczenie  6.5
## 3 Przymiotnik 14.8
## 4   Rymowanie  7.6
## 5  Wyobraznia 17.6
```

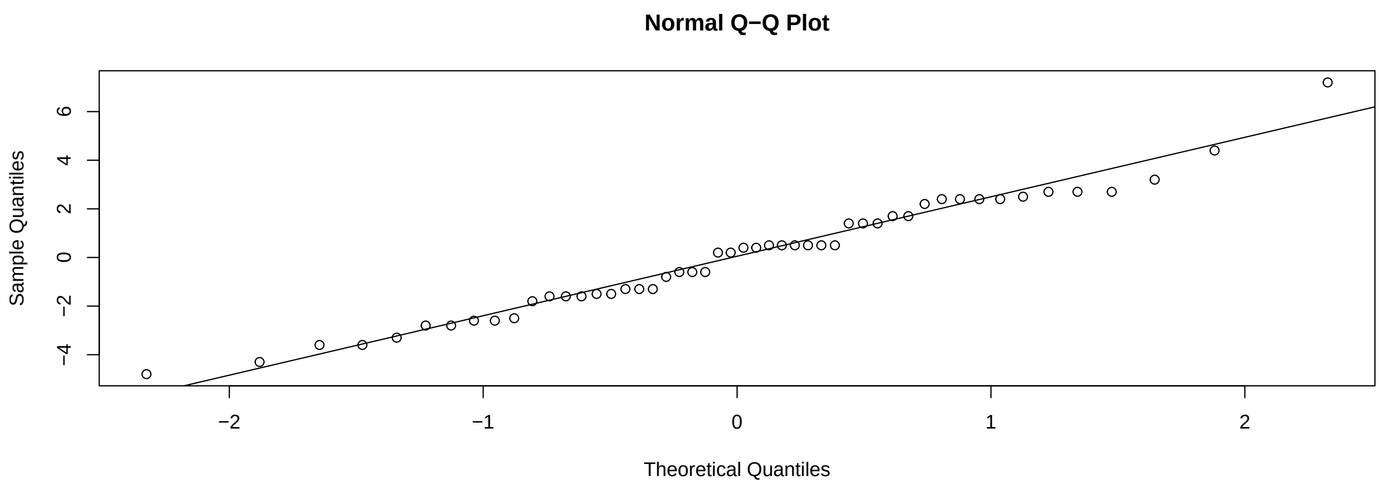


(3) Wykonaj test analizy wariancji w celu sprawdzenia, czy typ instrukcji ma istotny wpływ na badaną cechę zależną.

```
##             Df Sum Sq Mean Sq F value Pr(>F)    
## Instrukcja   4   1354   338.4   53.06 <2e-16 ***
## Residuals   45    287     6.4                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(4) Sprawdź założenia modelu jednoczynnikowej analizy wariancji.

```
## [1] 0.3756369
```




```
## [1] 0.1258206
```

```
## [1] 0.09922991
```

```
## [1] 0.07071935
```

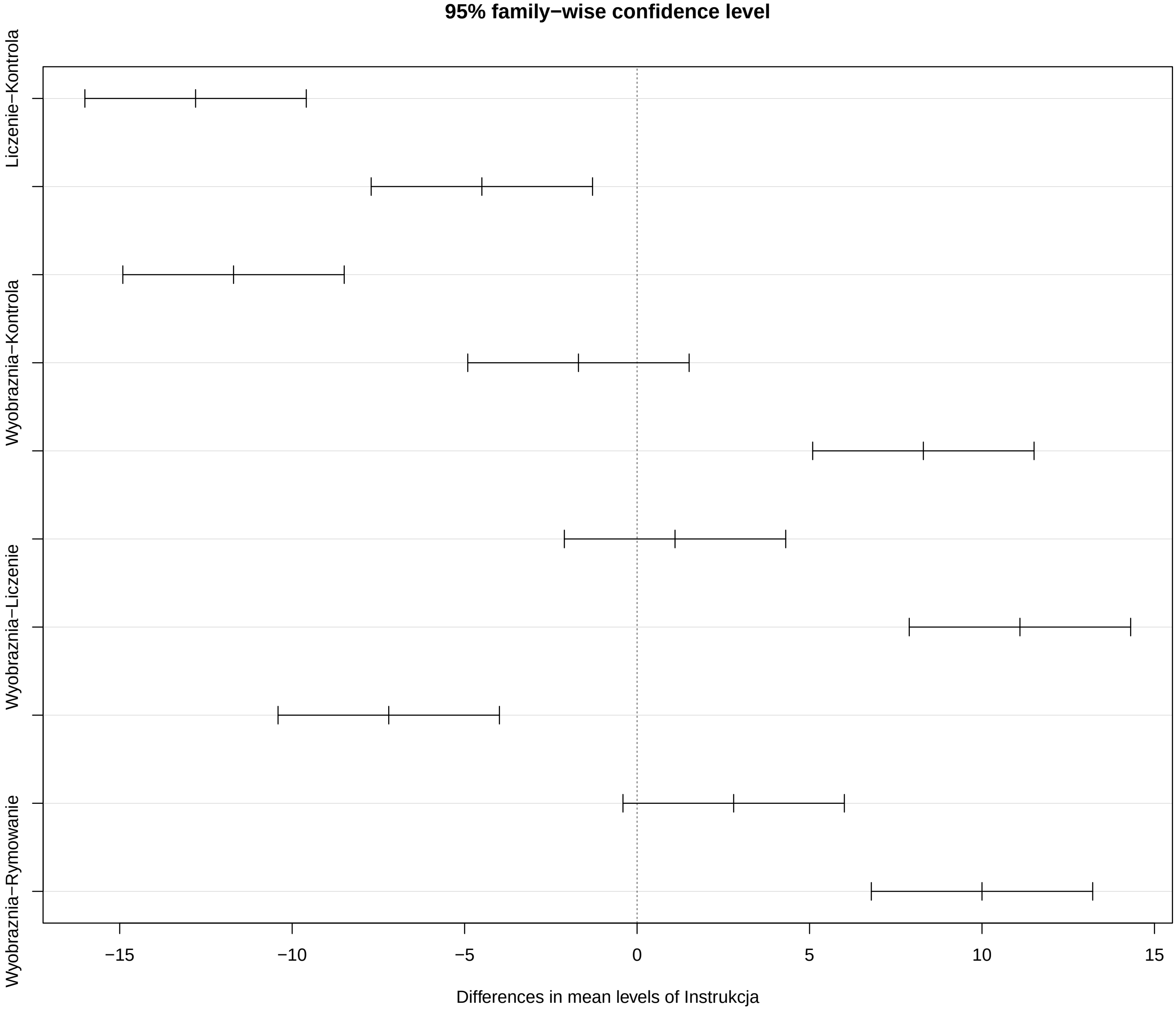
```
## [1] 0.1059926
```

(5) Wykonaj testy post hoc w celu sprawdzenia, które typy instrukcji różnią się między sobą.

```
## 
##  Pairwise comparisons using t tests with pooled SD 
## 
## data:  Wynik and Instrukcja 
## 
##             Kontrola Liczenie Przymiotnik Rymowanie
## Liczenie    8.9e-14  -        -           -        
```

```
## Przymiotnik 0.00098  1.9e-08  -          -
## Rymowanie   1.5e-12  0.33528  4.3e-07    -
## Wyobraznia  0.27851  7.1e-12  0.05094    1.4e-10
##
## P value adjustment method: holm
```

```
##                          diff         lwr       upr        p adj
## Liczenie-Kontrola       -12.8 -16.0091477 -9.590852 3.528289e-13
## Przymiotnik-Kontrola     -4.5  -7.7091477 -1.290852 2.177062e-03
## Rymowanie-Kontrola      -11.7 -14.9091477 -8.490852 1.968870e-12
## Wyobraznia-Kontrola      -1.7  -4.9091477  1.509148 5.645617e-01
## Przymiotnik-Liczenie      8.3   5.0908523 11.509148 3.057657e-08
## Rymowanie-Liczenie        1.1  -2.1091477  4.309148 8.654520e-01
## Wyobraznia-Liczenie      11.1   7.8908523 14.309148 9.156453e-12
## Rymowanie-Przymiotnik    -7.2 -10.4091477 -3.990852 8.442959e-07
## Wyobraznia-Przymiotnik    2.8  -0.4091477  6.009148 1.136213e-01
## Wyobraznia-Rymowanie     10.0   6.7908523 13.209148 2.024079e-10
```




```
##
```

```
## Study: model_aov ~ "Instrukcja"
##
## HSD Test for Wynik
##
## Mean Square Error:  6.377778
##
## Instrukcja,  means
##
##             Wynik      std  r Min Max
## Kontrola     19.3 2.626785 10  15  22
## Liczenie      6.5 1.433721 10   4   9
## Przymiotnik  14.8 3.489667 10  10  22
## Rymowanie     7.6 1.955050 10   4  10
## Wyobraznia   17.6 2.633122 10  14  22
##
## Alpha: 0.05 ; DF Error: 45
## Critical Value of Studentized Range: 4.018417
##
## Minimun Significant Difference: 3.209148
##
## Treatments with the same letter are not significantly different.
##
##             Wynik groups
## Kontrola     19.3      a
## Wyobraznia   17.6     ab
## Przymiotnik  14.8      b
## Rymowanie     7.6      c
## Liczenie      6.5      c

##
## Study: model_aov ~ "Instrukcja"
##
## Student Newman Keuls Test
## for Wynik
##
## Mean Square Error:  6.377778
##
## Instrukcja,  means
##
##             Wynik      std  r Min Max
## Kontrola     19.3 2.626785 10  15  22
## Liczenie      6.5 1.433721 10   4   9
## Przymiotnik  14.8 3.489667 10  10  22
## Rymowanie     7.6 1.955050 10   4  10
## Wyobraznia   17.6 2.633122 10  14  22
##
## Alpha: 0.05 ; DF Error: 45
##
## Critical Range
##        2        3        4        5
## 2.274738 2.737241 3.012913 3.209148
##
```

```
## Means with the same letter are not significantly different.
##
##              Wynik groups
## Kontrola      19.3      a
## Wyobraznia    17.6      a
## Przymiotnik   14.8      b
## Rymowanie      7.6      c
## Liczenie       6.5      c

##
## Study: model_aov ~ "Instrukcja"
##
## LSD t Test for Wynik
## P value adjustment method: holm
##
## Mean Square Error:  6.377778
##
## Instrukcja,  means and individual ( 95 %) CI
##
##              Wynik      std  r       LCL       UCL Min Max
## Kontrola      19.3 2.626785 10 17.691517 20.908483  15  22
## Liczenie       6.5 1.433721 10  4.891517  8.108483   4   9
## Przymiotnik   14.8 3.489667 10 13.191517 16.408483  10  22
## Rymowanie      7.6 1.955050 10  5.991517  9.208483   4  10
## Wyobraznia    17.6 2.633122 10 15.991517 19.208483  14  22
##
## Alpha: 0.05 ; DF Error: 45
## Critical Value of t: 2.952079
##
## Minimum Significant Difference: 3.334093
##
## Treatments with the same letter are not significantly different.
##
##              Wynik groups
## Kontrola      19.3      a
## Wyobraznia    17.6     ab
## Przymiotnik   14.8      b
## Rymowanie      7.6      c
## Liczenie       6.5      c

##
## Study: model_aov ~ "Instrukcja"
##
## Scheffe Test for Wynik
##
## Mean Square Error  : 6.377778
##
## Instrukcja,  means
##
##              Wynik      std  r Min Max
## Kontrola      19.3 2.626785 10  15  22
## Liczenie       6.5 1.433721 10   4   9
```

```
## Przymiotnik  14.8 3.489667 10  10  22
## Rymowanie     7.6 1.955050 10   4  10
## Wyobraznia   17.6 2.633122 10  14  22
##
## Alpha: 0.05 ; DF Error: 45
## Critical Value of F: 2.578739
##
## Minimum Significant Difference: 3.627299
##
## Means with the same letter are not significantly different.
##
##             Wynik groups
## Kontrola     19.3      a
## Wyobraznia   17.6     ab
## Przymiotnik  14.8      b
## Rymowanie     7.6      c
## Liczenie      6.5      c
```

(6) Przetestuj hipotezy szczegółowe związane z następującymi zagadnieniami:

- Porównaj dwie grupy powierzchownego uzyskiwania informacji z dwiema grupami głębokiego uzyskiwania informacji.
- Porównaj grupę kontrolną z pozostałymi czterema grupami.
- Porównaj dwie grupy powierzchownego uzyskiwania informacji między sobą.
- Porównaj dwie grupy głębokiego uzyskiwania informacji między sobą.

W tym celu wykonaj następujące polecenia:

- Zapisz odpowiednie hipotezy.
- Wyraź je za pomocą kontrastów.
- Czy ten układ kontrastów jest ortogonalny?
- Przetestuj zaproponowane kontrasty.

```
##                  Df Sum Sq Mean Sq F value   Pr(>F)
## Instrukcja        4 1353.7   338.4  53.064  < 2e-16 ***
##   Instrukcja: C1  1  837.2   837.2 131.272 6.19e-15 ***
##   Instrukcja: C2  1  471.2   471.2  73.889 4.76e-11 ***
##   Instrukcja: C3  1    6.1     6.1   0.949    0.335
##   Instrukcja: C4  1   39.2    39.2   6.146    0.017 *
## Residuals        45  287.0     6.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(7) Wykonaj polecenia 2, 3 i 5 wykorzystując odpowiednie metody nieparametryczne. Porównaj ich wyniki z wynikami metod parametrycznych.

```
##    Instrukcja    x
## 1    Kontrola 19.5
## 2    Liczenie  7.0
## 3 Przymiotnik 14.5
## 4   Rymowanie  7.5
## 5  Wyobraznia 17.5
```

```
## [1] 7.612601e-08
```

```
##
```

```
##  Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  Wynik and Instrukcja
##
##             Kontrola Liczenie Przymiotnik Rymowanie
## Liczenie    0.0014   -        -           -
## Przymiotnik 0.0338   0.0014   -           -
## Rymowanie   0.0014   0.3385   0.0014      -
## Wyobraznia  0.3385   0.0014   0.1574      0.0014
##
## P value adjustment method: holm
```

```
##                   Comparison          Z      P.unadj        P.adj
## 1       Kontrola - Liczenie  4.9276610 8.321985e-07 8.321985e-06
## 2    Kontrola - Przymiotnik  1.7785776 7.530903e-02 1.075843e-01
## 3    Liczenie - Przymiotnik -3.1490834 1.637835e-03 3.275669e-03
## 4      Kontrola - Rymowanie  4.4040970 1.062254e-05 5.311270e-05
## 5      Liczenie - Rymowanie -0.5235640 6.005818e-01 6.005818e-01
## 6   Przymiotnik - Rymowanie  2.6255194 8.651689e-03 1.441948e-02
## 7     Kontrola - Wyobraznia  0.7083513 4.787271e-01 5.319191e-01
## 8     Liczenie - Wyobraznia -4.2193097 2.450514e-05 8.168381e-05
## 9  Przymiotnik - Wyobraznia -1.0702264 2.845174e-01 3.556468e-01
## 10   Rymowanie - Wyobraznia -3.6957457 2.192423e-04 5.481058e-04
```

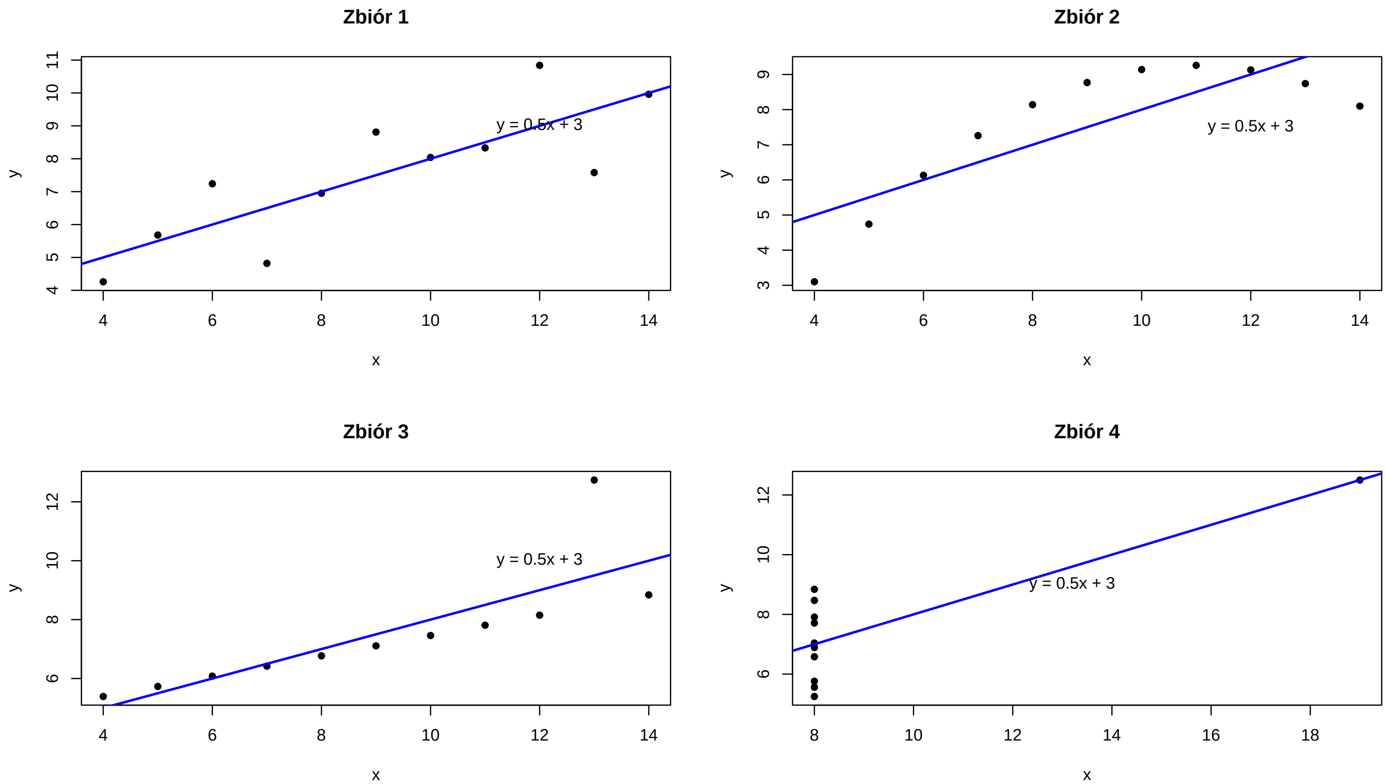
7 Regresja

- Termin regresja oznacza metodę, która pozwala badać związek między zmiennymi i wykorzystywać tę wiedzę do przewidywania nieznanych wartości jednej wielkości na podstawie innych.
- W praktyce poszukuje się związku między jedną (lub większą liczbą) zmienną objaśniającą (niezależną) X a zmienną objaśnianą (zależną) Y.
- Zależność ta może być dalej wykorzystana do przewidywania wartości zmiennej Y w zależności od zmiennej X.
- Jeśli badamy zależność zmiennej Y od wartości innej zmiennej, wartości zmiennej objaśniającej zostaną oznaczone przez x i potraktowane jako wartości deterministyczne zmiennej X, które wybieramy do obserwacji losowej zmienna Y.
- Zmienne X i Y są traktowane inaczej. Mianowicie, zmienna X jest uważana za w pełni kontrolowaną przez eksperymentatora, a zatem jest pozbawiona elementu losowości (w rzeczywistości jest traktowana jako liczba).
- Zatem chcemy odpowiedzieć na pytanie, jak zmienia się oczekiwana wartość zmiennej Y w zależności od wartości x zmiennej X, tj.
$$E(Y) = g(x),$$
gdzie $g(x)$ to funkcja regresji opisująca związek.
- Przyjmuje się również, że $Var(Y)$ jest stała dla wszystkich wartości x i równa σ^2 (jednorodność wariancji).
- Matematycznie regresję nazywa się dowolną metodę, która pozwala oszacować to równanie.
- Zależności regresyjnej poszukuje się w pewnej predefiniowanej klasie funkcji, zazwyczaj klasie wielomianów danego stopnia.

Wykres rozrzutu

- Wykres rozrzutu jest zwykle wykreślany w celu wstępnej oceny zależności.

- Jego znaczenie zostało doskonale podkreślone przez Anscombe (1973), który skonstruował 4 zbiory danych o identycznych podstawowych charakterystykach, ale ich wykresy rozrzutu były diametralnie różne. Średnia dla każdej zmiennej x_i wynosiła 9, zmiennej y_i wynosiła 10, wariancja dla x_i wynosiła 7,5, dla y_i 2,75, współczynnik korelacji liniowej wynosił 0,816 dla każdego zbioru, a prosta regresji wynosiła $y = 0{,}5x + 3$.



- Wykresy różnią się bardzo wyraźnie.
- Pierwszy wykres (lewy górny róg) sugeruje, że dane mają rozkład normalny, a prosta regresji i współczynnik korelacji są sensowne.
- Drugi wykres (prawy górny róg) pokazuje nieliniowy charakter związku, a zatem brak uzasadnienia dla prostej regresji i współczynnika korelacji.
- Trzeci wykres (lewy dolny róg) wskazuje znaczenie wartości odstającej, która jest przyczyną otrzymanej wartości współczynnika korelacji.
- Ostatni wykres (prawy dolny róg) pokazuje inne zjawisko, a mianowicie wpływową obserwację, która spowodowała, że współczynnik korelacji był wysoki, chociaż nie powinien.

7.1 Regresja liniowa

- W regresji liniowej zakładamy, że $g(x)$ jest funkcją liniową, tzn. otrzymujemy równanie regresji liniowej:

$$E(Y) = ax + b,$$

  gdzie $a, b \in \mathbb{R}$ i $a \neq 0$ są nieznanymi parametrami.
- W literaturze regresja liniowa tej postaci jest również nazywana prostą regresją liniową lub regresją prostą.
- Załóżmy, że mamy n obserwacji $x_1, \ldots, x_n$ zmiennej X i odpowiadające im obserwacje $Y_1, \ldots, Y_n$ zmiennej Y.
- W praktyce wygodnie jest stosować następujący model regresji liniowej:

$$Y_i = ax_i + b + e_i, \quad i = 1, \ldots, n,$$

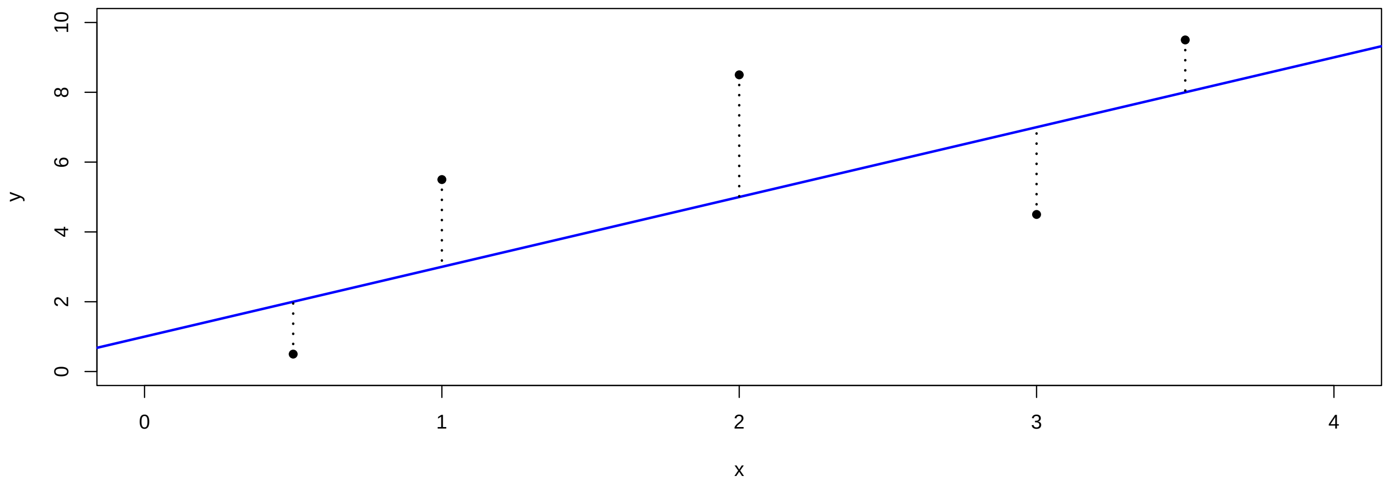
gdzie e_i są błędami losowymi, które są nieskorelowanymi zmiennymi losowymi o zerowej wartości oczekiwanej i tej samej nieznanej wariancji $\sigma^2 > 0$.

- Zakładamy również, że liczba obserwacji jest większa niż liczba parametrów, tj. $n > 2$.
- Zauważmy, że nie jest wymagane określenie rozkładu błędów losowych. Jednak w przypadku konstruowania przedziałów ufności i testów statystycznych w tym modelu przyjmuje się zwykle rozkład normalny.

Estymacja parametrów

- Estymacji parametrów a i b dokonujemy metodą najmniejszych kwadratów, która ma na celu minimalizacje sumy kwadratów błędów losowych, tj.

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - ax_i - b)^2.$$



- Estymatory parametrów a i b otrzymane metodą najmniejszych kwadratów mają postaci:

$$\hat{a} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \hat{b} = \bar{y} - \bar{x}\hat{a},$$

gdzie $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ oraz $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ są średnimi z prób odpowiadających zmiennym X i Y.

- Estymatory $\hat{a}$ i $\hat{b}$ są estymatorami nieobciążonymi parametrów a i b. Ponadto, są one najlepszymi estymatorami (ENMW) w klasie estymatorów liniowych parametrów a i b.

- Ponadto, estymator

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{a}x_i - \hat{b})^2$$

jest nieobciążonym estymatorem wariancji σ^2.

- Zatem mamy funkcję liniową postaci

$$y = \hat{a}x + \hat{b},$$

którą nazywa się prostą regresji opisującą zależność między zmiennymi X i Y.

- Współczynnik kierunkowy $\hat{a}$ jest nazywany współczynnikiem regresji liniowej.

- Określa on średni wzrost wartości zmiennej zależnej Y na jednostkę wzrostu wartości zmiennej niezależnej X. Jeśli $\hat{a} > 0$, to wraz ze wzrostem wartości zmiennej X rośnie również wartość zmiennej Y. Z drugiej strony, w przypadku $\hat{a} < 0$, zachodzi sytuacja odwrotna - wraz ze wzrostem wartości zmiennej X wartość zmiennej Y maleje.

- $\hat{b}$ jest nazywany wyrazem wolnym (ang. intercept).

- Przy założeniu normalności rozkładu, przedziały ufności dla parametrów a i b na poziomie ufności $1-\alpha$, $\alpha \in (0,1)$ są następujące:

$$\left(\hat{a} - t\left(1-\frac{\alpha}{2}, n-2\right)\sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{(n-2)\sum_{i=1}^n (x_i - \bar{x})^2}}, \hat{a} + t\left(1-\frac{\alpha}{2}, n-2\right)\sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{(n-2)\sum_{i=1}^n (x_i - \bar{x})^2}}\right),$$

$$\left(\hat{b} - t\left(1-\frac{\alpha}{2}, n-2\right)\sqrt{\frac{\sum_{i=1}^n x_i^2 \sum_{i=1}^n (\hat{y}_i - y_i)^2}{n(n-2)\sum_{i=1}^n (x_i - \bar{x})^2}}, \hat{b} + t\left(1-\frac{\alpha}{2}, n-2\right)\sqrt{\frac{\sum_{i=1}^n x_i^2 \sum_{i=1}^n (\hat{y}_i - y_i)^2}{n(n-2)\sum_{i=1}^n (x_i - \bar{x})^2}}\right)$$

odpowiednio, gdzie $t(\beta, m)$ oznacza kwantyl rzędu β z rozkładu t-Studenta $t(m)$ z m stopniami swobody.

Testy istotności dla współczynników regresji

- Testy istotności dla współczynników regresji są następujące:

1. Mamy

$$H_0^a : a = 0 \textit{ przeciw } H_1^a : a \neq 0.$$

Statystyka testowa jest postaci

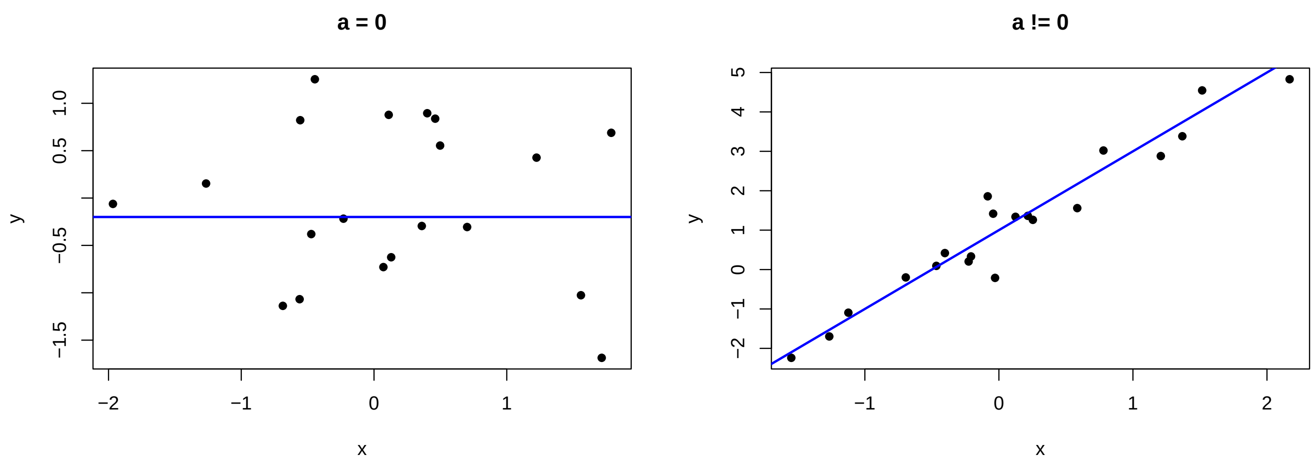
$$T_a = \frac{\hat{a}}{S_a},$$

gdzie

$$S_a^2 = \frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2/(n-2)}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \hat{y}_i = \hat{a}x_i + \hat{b}.$$

Przy założeniu normalności otrzymujemy

$$T_a\Big|_{H_0^a} \sim t(n-2).$$



2. Mamy

$$H_0^b : b = 0 \textit{ przeciw } H_1^b : b \neq 0.$$

Statystyka testowa jest postaci

$$T_b = \frac{\hat{b}}{S_b},$$

gdzie

$$S_b^2 = \frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n-2}\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right).$$

Przy założeniu normalności otrzymujemy

$$T_b\Big|_{H_0^b} \sim t(n-2).$$

Współczynnik determinacji R^2

- Miarą liczbową dopasowania regresji liniowej (z wyrazem wolnym) do danych empirycznych jest współczynnik determinacji.
- Współczynnik determinacji R^2 (zmiennej zależnej w modelu regresji liniowej) określa jaki procent wariancji zmiennej zależnej wyjaśnia model:
$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2},$$
gdzie $\hat{y}_i = \hat{a}x_i + \hat{b}$, $i = 1, \dots, n$. Różnice $y_i - \hat{y}_i$ są nazywane resztami lub residuami.
- Im wyższa wartość współczynnika $0 \leqslant R^2 \leqslant 1$, tym lepszy model (oczywiście w sensie tego kryterium). Należy również pamiętać, że ta ocena jakości modelu jest poprawna tylko wtedy, gdy model jest odpowiedni, tj. gdy założenia modelu są spełnione.
- Współczynnik determinacji nie ma sensu, jeśli wyraz wolny zostanie pominięty w modelu.
- Kiedy mamy model z wyrazem wolnym, R^2 porównuje ten model z modelem referencyjnym, który zawiera tylko wyraz wolny (tj. stały człon, średnia z próby).
- Gdy nie uwzględniamy wyrazu wolnego, oba modele są całkowicie niezależne od siebie, więc porównanie nie ma sensu. Zamiast tego
$$R_0^2 = 1 - \frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{\sum_{i=1}^n y_i^2}$$
jest obliczany dla modelu bez wyrazu wolnego, który domyślnie wykorzystuje model referencyjny odpowiadający tylko losowemu szumowi. Oczywiście nie jest to prawdziwy model (nic nie wyjaśnia) i jakiekolwiek porównanie z nim nie jest zbyt przydatne. W rzeczywistości nie ma sensownego modelu referencyjnego, którego moglibyśmy użyć do porównania modelu bez wyrazu wolnego. Zatem nie ma interpretowalnej wersji R^2 dla modeli z/bez wyrazu wolnego!
- Zauważmy, że możliwe jest uzyskanie ujemnej wartości R^2 dla modeli, które nie zawierają wyrazu wolnego. Ponieważ R^2 jest zdefiniowane jako proporcja wariancji wyjaśnionej przez dopasowany model, jeśli jest on w rzeczywistości gorszy niż tylko dopasowanie linii poziomej ($y = \bar{y}$), to R^2 jest ujemne. W takim przypadku R^2 nie może być interpretowane jako kwadrat korelacji. Wartości ujemne mogą wystąpić, gdy model zawiera zmienne, które nie pomóc w przewidywaniu wartości zmiennej zależnej.
- Kiedy należy uwzględnić lub usunąć wyraz wolny z modelu? Zasadniczo istnieje tylko jeden powód, aby przeprowadzić regresję bez wyrazu wolnego: gdy model jest używany do opisania procesu, o którym wiadomo, że ma zerowy wyraz wolny. Drugim rodzącym się problemem jest to, że estymator najmniejszych kwadratów dla współczynnika regresji (a) w modelu bez wyrazu wolnego jest obciążony (systematycznie przesuwany w kierunku większych lub mniejszych wartości). Usuwając wyraz wolny z modelu nakładamy ograniczenie, aby linia regresji przechodziła przez punkt $\mathbf{0}$. Prosta regresji bez wyrazu wolnego jest zwykle mocno „ściągnięta”, ponieważ musi przejść przez punkt $\mathbf{0}$.
- Modele bez wyrazu wolnego są rzadko stosowane w praktyce. Teoretycznie można użyć tego modelu, gdy wiemy, że prosta regresji musi przechodzić przez punkt $\mathbf{0}$. Na przykład, jeśli modelujemy PKB (Produkt Krajowy Brutto) w stosunku do liczby ludności, przypuszczalnie, gdy populacja wynosi 0, to PKB również wynosi 0. Model bez wyrazu wolnego miałby wtedy sens. Jednak modele regresji zwykle nie obejmują szerokiego zakresu wartości zmiennych niezależnych. Zatem w praktyce i tak używamy modelu z wyrazem wolnym.
- Dlatego zaleca się stosowanie modeli bez wyrazu wolnego tylko wtedy, gdy jest to teoretycznie uzasadnione!

Predykcja

- Regresja służy głównie do przewidywania (predykcji).
- Prognozowanie polega na znalezieniu możliwej wartości zmiennej zależnej Y dla wartości x_{nowa} zmiennej niezależnej X, która różni się od każdej obserwacji w próbie (tj. $x_{nowa} \notin \{x_1, \dots, x_n\}$).
- Na podstawie wybranego modelu regresji liniowej przewidujemy zmienną zależną Y dla wartości x_{nowa} zmiennej niezależnej X w następujący sposób:

$$y_p = \hat{a}x_{nowa} + \hat{b}.$$

- Jako ocenę jakości prognozy przyjmujemy oszacowanie standardowego odchylenia prognozy (średni błąd prognozy) postaci:

$$S_p = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n-2}\left(1 + \frac{1}{n} + \frac{(x_{nowa} - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)}.$$

- Przy założeniu normalności przedział ufności dla y_p na poziomie ufności $1 - \alpha$, $\alpha \in (0, 1)$ jest postaci:

$$\left(y_p - t\left(1 - \frac{\alpha}{2}, n-2\right) S_p, y_p + t\left(1 - \frac{\alpha}{2}, n-2\right) S_p\right),$$

  gdzie $t(\beta, m)$ oznacza kwantyl rzędu β z rozkładu t-Studenta $t(m)$ z m stopniami swobody.

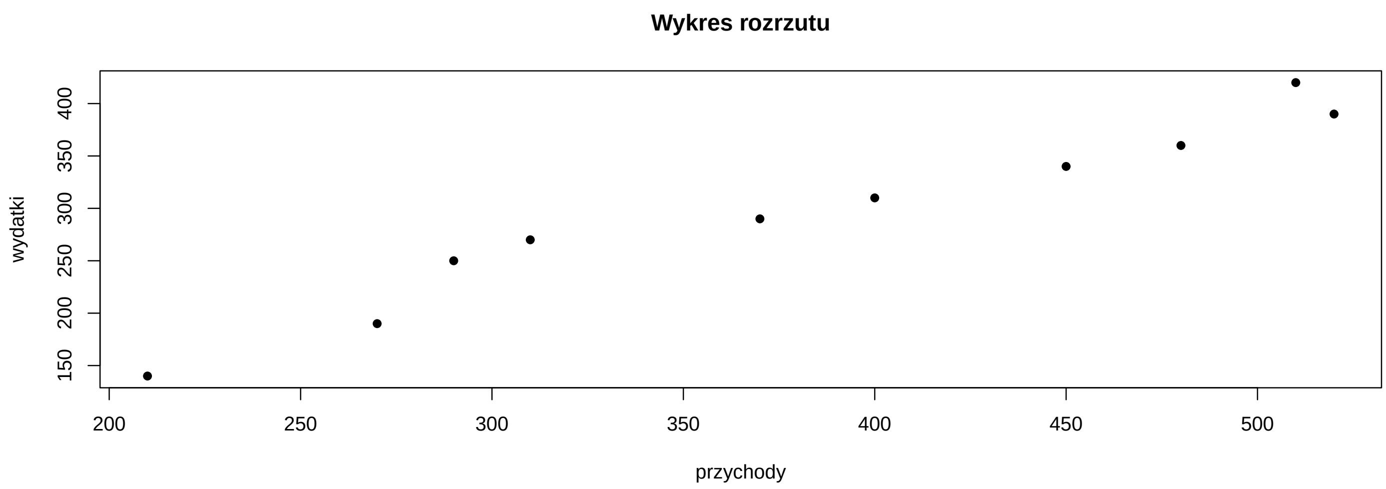
Przykład. Za pomocą regresji liniowej chcemy opisać związek między miesięcznym dochodem rodziny na jedną osobę a miesięczną wartością wydatków na jedną osobę. Dane dotyczące tych dwóch cech dla dziesięciu rodzin podano w poniższej tabeli.

rodzina	1	2	3	4	5	6	7	8	9	10
przychody	210	270	290	310	370	400	450	480	510	520
wydatki	140	190	250	270	290	310	340	360	420	390

```r
# dane
przychody <- c(210, 270, 290, 310, 370, 400, 450, 480, 510, 520)
wydatki <- c(140, 190, 250, 270, 290, 310, 340, 360, 420, 390)
data_set <- data.frame(przychody = przychody, wydatki = wydatki)
head(data_set)
```

```
##   przychody wydatki
## 1       210     140
## 2       270     190
## 3       290     250
## 4       310     270
## 5       370     290
## 6       400     310
```

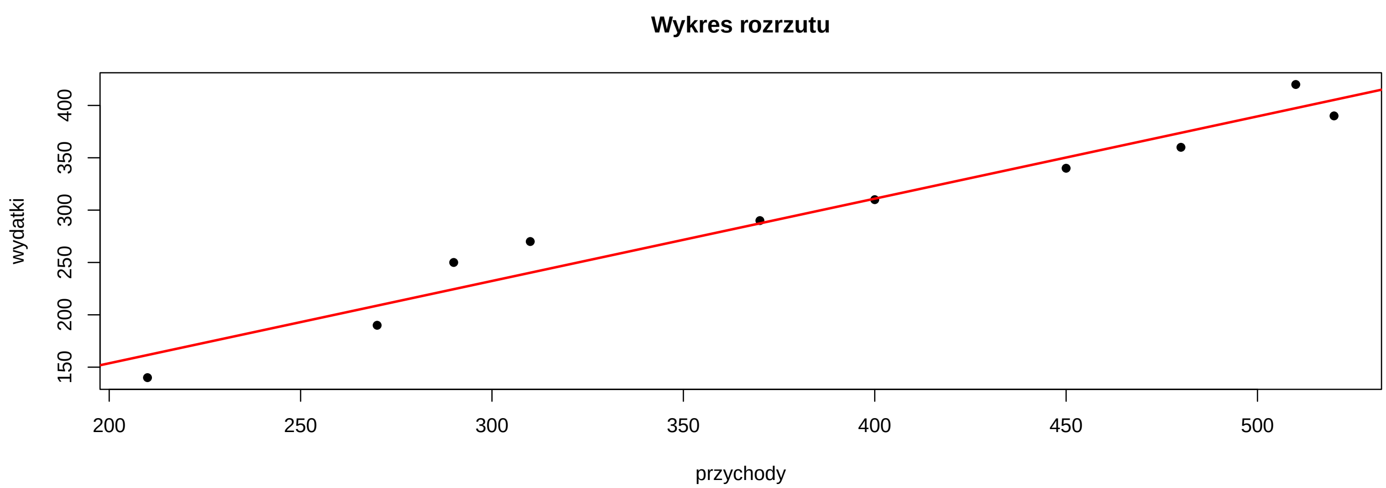
```r
# Wykres rozrzutu
plot(data_set, main = "Wykres rozrzutu", pch = 16)
```




```r
# model
model <- lm(wydatki ~ przychody, data = data_set)
model
```

```
##
## Call:
## lm(formula = wydatki ~ przychody, data = data_set)
##
## Coefficients:
## (Intercept)    przychody
##     -3.5036       0.7861
```

```r
plot(data_set, main = "Wykres rozrzutu", pch = 16)
abline(model, col = "red", lwd = 2)
```




```r
# estymacja parametrów
coef(model)
```

```
## (Intercept)   przychody
##  -3.5036358   0.7860988
```

```r
confint(model)
```

```
##                   2.5 %     97.5 %
## (Intercept) -61.2027257 54.1954540
## przychody     0.6398962  0.9323013
```

```r
# podsumowanie modelu
# tj. reszty, estymacja punktowa, testy istotności dla współczynników regresji,
# R^2, test istnotności modelu
summary(model)
```

```
## 
## Call:
## lm(formula = wydatki ~ przychody, data = data_set)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -21.577 -14.907  -5.588  17.607  29.813 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  -3.5036    25.0212   -0.14    0.892    
## przychody     0.7861     0.0634   12.40 1.67e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 20.63 on 8 degrees of freedom
## Multiple R-squared:  0.9505, Adjusted R-squared:  0.9444 
## F-statistic: 153.7 on 1 and 8 DF,  p-value: 1.67e-06
```

```r
# wartości dopasowane przez model
fitted(model)
```

```
##        1        2        3        4        5        6        7        8 
## 161.5771 208.7430 224.4650 240.1870 287.3529 310.9359 350.2408 373.8238 
##        9       10 
## 397.4067 405.2677
```

```r
# reszty
residuals(model)
```

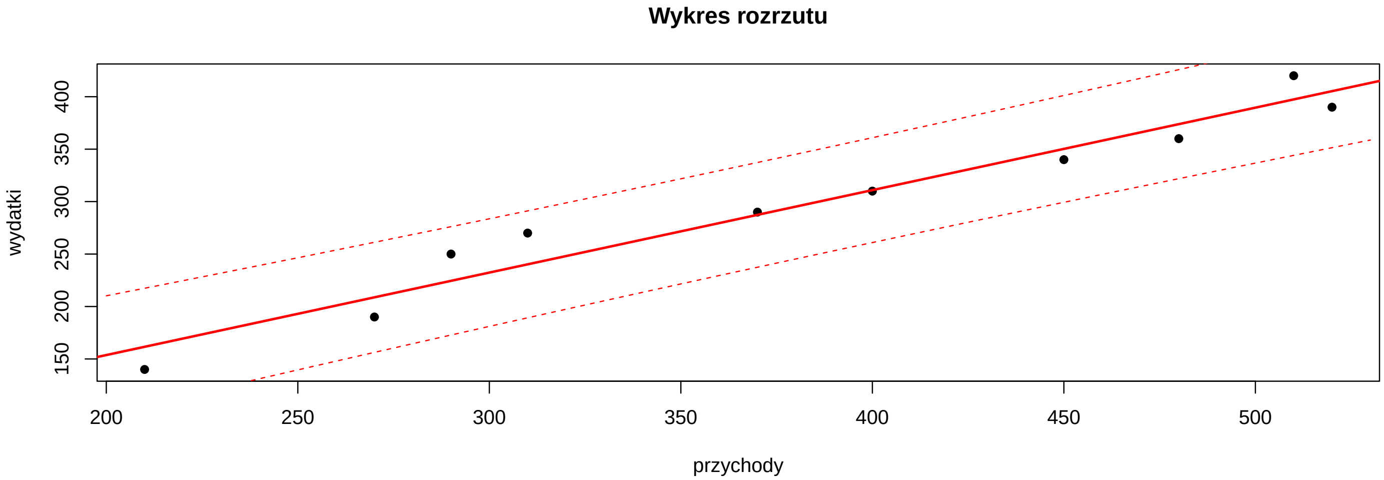
```
##           1           2           3           4           5           6 
## -21.5771083 -18.7430352  25.5349891  29.8130135   2.6470866  -0.9358769 
##           7           8           9          10 
## -10.2408159 -13.8237794  22.5932572 -15.2677307
```

```r
# sprawdzenie
wydatki - fitted(model)
```

```
##           1           2           3           4           5           6 
## -21.5771083 -18.7430352  25.5349891  29.8130135   2.6470866  -0.9358769 
##           7           8           9          10 
## -10.2408159 -13.8237794  22.5932572 -15.2677307
```

```r
# przedziały ufności dla predykcji
temp_przychody <- data.frame(przychody = seq(min(data_set$przychody) - 10,
                                             max(data_set$przychody) + 10,
                                             length = 100))
pred <- stats::predict(model, temp_przychody, interval = "prediction")
plot(data_set, main = "Wykres rozrzutu", pch = 16)
```

```
abline(model, col = "red", lwd = 2)
lines(temp_przychody$przychody, pred[, 2], lty = 2, col = "red")
lines(temp_przychody$przychody, pred[, 3], lty = 2, col = "red")
```




```
# predykcja wydatków dla przychodu = 350
nowy_przychod <- data.frame(przychody = 350)
stats::predict(model, nowy_przychod, interval = 'prediction')
```

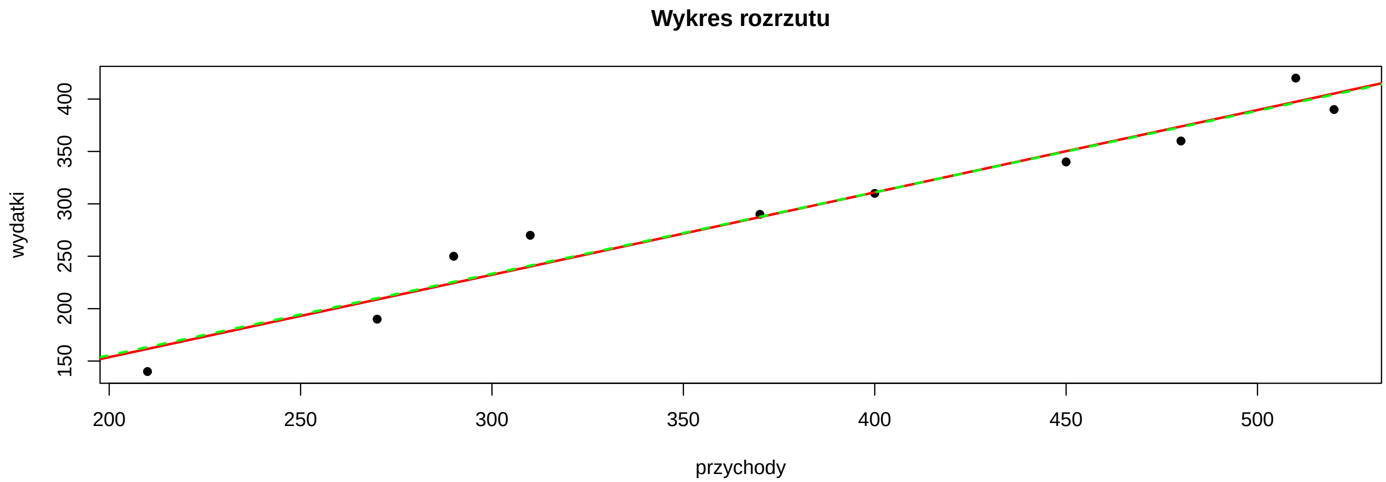
```
##        fit     lwr      upr
## 1 271.6309 221.528 321.7338
```

- model bez wyrazu wolnego

```
model_bez_ww <- lm(wydatki ~ przychody - 1, data = data_set)
model_bez_ww
```

```
##
## Call:
## lm(formula = wydatki ~ przychody - 1, data = data_set)
##
## Coefficients:
## przychody
##    0.7775
```

```
plot(data_set, main = "Wykres rozrzutu", pch = 16)
abline(model, col = "red", lwd = 2)
abline(model_bez_ww, col = "green", lwd = 2, lty = 2)
```




```r
coef(model_bez_ww)
```

```
## przychody
## 0.7775281
```

```r
confint(model_bez_ww)
```

```
##               2.5 %   97.5 %
## przychody 0.7422271 0.812829
```

```r
summary(model_bez_ww)
```

```
##
## Call:
## lm(formula = wydatki ~ przychody - 1, data = data_set)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.281 -14.039  -5.449  18.174  28.966
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## przychody   0.7775     0.0156   49.83 2.65e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.48 on 9 degrees of freedom
## Multiple R-squared:  0.9964, Adjusted R-squared:  0.996
## F-statistic:  2483 on 1 and 9 DF,  p-value: 2.651e-12
```

```r
fitted(model_bez_ww)
```

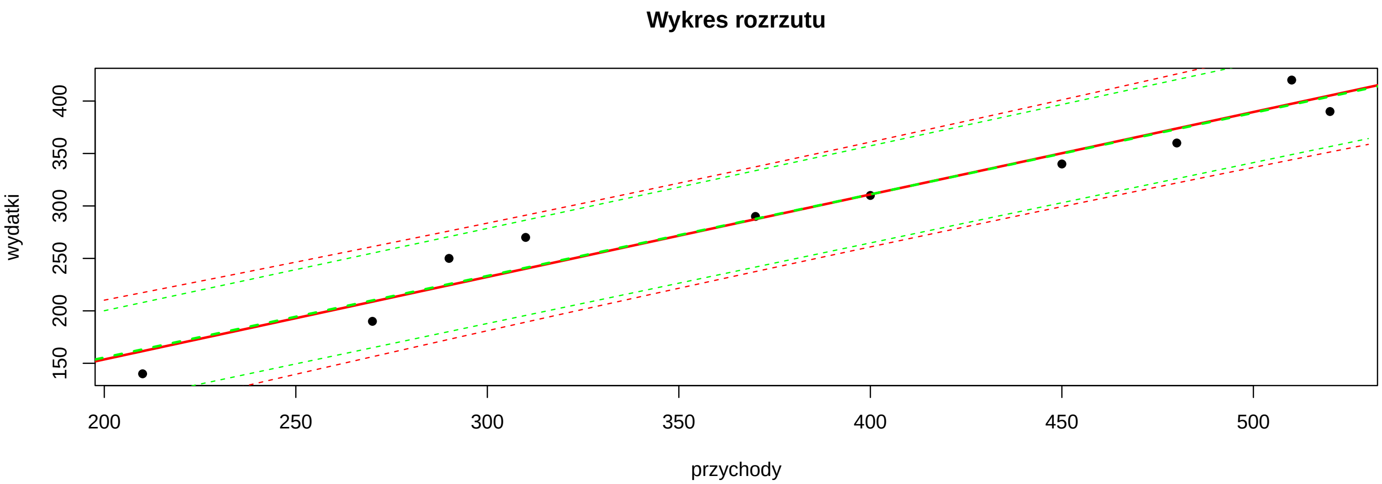
```
##        1        2        3        4        5        6        7        8
## 163.2809 209.9326 225.4831 241.0337 287.6854 311.0112 349.8876 373.2135
##        9       10
## 396.5393 404.3146
```

```r
residuals(model_bez_ww)
```

```
##          1          2          3          4          5          6          7
## -23.280899 -19.932584  24.516854  28.966292   2.314607  -1.011236  -9.887640
```

```
##          8          9         10
## -13.213483  23.460674 -14.314607
```

```r
temp_przychody <- data.frame(przychody = seq(min(data_set$przychody) - 10,
                                             max(data_set$przychody) + 10,
                                             length = 100))
pred1 <- stats::predict(model_bez_ww, temp_przychody, interval = "prediction")
plot(data_set, main = "Wykres rozrzutu", pch = 16)
abline(model, col = "red", lwd = 2)
abline(model_bez_ww, col = "green", lwd = 2, lty = 2)
lines(temp_przychody$przychody, pred[, 2], lty = 2, col = "red")
lines(temp_przychody$przychody, pred[, 3], lty = 2, col = "red")
lines(temp_przychody$przychody, pred1[, 2], lty = 2, col = "green")
lines(temp_przychody$przychody, pred1[, 3], lty = 2, col = "green")
```




```r
nowy_przychod <- data.frame(przychody = 350)
stats::predict(model_bez_ww, nowy_przychod, interval = 'prediction')
```

```
##        fit      lwr      upr
## 1 272.1348 226.3796 317.8901
```

7.2 Regresja wielokrotna

- Często badane zjawisko zależy nie tylko od jednej zmiennej, ale od wielu zmiennych i możemy nie wiedzieć, która cecha ma główny wpływ.
- Uogólnieniem prostej (liniowej) regresji jest regresja wielokrotna, która uwzględnia wpływ wielu zmiennych niezależnych, powiedzmy $X_1, \dots, X_p$, na wybraną zmienną zależną Y.
- Model regresji wielokrotnej jest następujący:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + e_i = \mathbf{x}_i^\top \boldsymbol{\beta} + e_i,$$

  gdzie $i = 1, \dots, n$,

$$\mathbf{x}_i^\top = (1, x_{i1}, x_{i2}, \dots, x_{ip})$$

  są wektorami obserwacji zmienny $X_1, \dots, X_p$ oraz

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^\top$$

  jest wektorem nieznanych parametrów.

- W notacji macierzowej powyższy model może być zapisany następująco:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

gdzie

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}, \quad \mathbf{e} = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}.$$

- Założenia są takie same jak w modelu regresji liniowej, np. liczba obserwacji jest większa niż liczba parametrów, tj. $n > p + 1$. Dodatkowo zakładamy, że nie ma liniowej zależności między wektorami obserwacji zmiennych objaśniających, tj. $rank(\mathbf{X}) = p + 1$.

Estymacja parametrów

- Estymacji parametrów w wektorze $\boldsymbol{\beta}$ dokonujemy metodą najmniejszych kwadratów, która ma na celu minimalizacje sumy kwadratów błędów losowych, tj.

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2.$$

- Estymator metody najmniejszych kwadratów wektora $\boldsymbol{\beta}$ jest postaci:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}.$$

- Estymator $\widehat{\boldsymbol{\beta}}$ jest estymatorem nieobciążonym wektora $\boldsymbol{\beta}$.

Twierdzenie. Niech $\mathbf{A}^\top \boldsymbol{\beta} = a_0\beta_0 + \dots + a_p\beta_p$, gdzie $\mathbf{A}^\top = (a_0, \dots, a_p)$, będzie funkcją parametryczną. Wtedy $\mathbf{A}^\top \widehat{\boldsymbol{\beta}}$ jest ENMW funkcji parametrycznej $\mathbf{A}^\top \boldsymbol{\beta}$ wśród wszystkich liniowych estymatorów tej funkcji.

- Ponadto,

$$\widehat{\sigma}^2 = \frac{1}{n - p - 1} \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$$

jest nieobciążonym estymatorem wariancji σ^2, gdzie $\widehat{y}_i = \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}$, $i = 1, \dots, n$.

Testy istotności dla współczynników regresji

- Weryfikujemy układ hipotez

$$H_0^j : \beta_j = 0 \ przeciw \ H_1^j : \beta_j \neq 0$$

dla $j = 0, \dots, p$.
- Statystyka testowa jest postaci

$$T_j = \frac{\widehat{\beta}_j}{S_j},$$

gdzie

$$S_j^2 = d_{jj} \sum_{i=1}^{n} (\widehat{y}_i - y_i)^2 / (n - p - 1), \quad \widehat{y}_i = \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}$$

oraz d_{jj} jest j-tym elementem diagonalnym macierzy $(\mathbf{X}^\top \mathbf{X})^{-1}$.
- Przy założeniu normalności otrzymujemy

$$T_j \Big|_{H_0^j} \sim t(n - p - 1).$$

Test analizy wariancji w modelu regresji

- Zamiast testować istotność każdej zmiennej objaśniającej osobno, możemy przetestować istotność modelu jako całości.
- Mianowicie, weryfikujemy układ hipotez

$$H_0 : \beta_1 = \cdots = \beta_p = 0 \ przeciw \ H_1 : \neg H_0.$$

- W ten sposób testujemy jednocześnie trzy równoważne hipotezy:
  - istotność współczynników regresji,
  - istotność liniowej zależności między zmiennymi,
  - istotność związku między zmiennymi.
- Test weryfikujący powyższe hipotezy konstruuje się metodą analizy wariancji, która opiera się na następującej zależności:

$$SST = SSR + SSE,$$

  gdzie

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2, \quad SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2, \quad SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

  są sumą kwadratów dla całości, sumą kwadratów dla regresji i sumą kwadratów dla błędu odpowiednio, $\hat{y}_i = \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}$ i $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.
- Tabel analizy wariancji w regresji

Źródło zmienności	Stopnie swobody (DF)	Suma kwadratów (SS)	Średni kwadrat (MS)
Regresja	p	SSR	$MSR = SSR/p$
Błąd	$n-p-1$	SSE	$MSE = SSE/(n-p-1)$
Całość	$n-1$	SST	

- Przy prawdziwości hipotezy zerowej i założeniu normalności:

$$SSR \sim \chi^2(p), \quad SSE \sim \chi^2(n-p-1), \quad SST \sim \chi^2(n-1).$$

- Obszar krytyczny testu analizy wariancji w regresji jest postaci:

$$R = \left\{(y_i) : \frac{MSR}{MSE} > F(1-\alpha, p, n-p-1)\right\},$$

  gdzie

$$\left.\frac{MSR}{MSE}\right|_{H_0} \sim F(p, n-p-1)$$

  and $F(\beta, m, n)$ oznacza kwantyl rzędu β z rozkładu F-Snedecora $F(m,n)$ z m i n stopniami swobody.
- p-wartość ma postać:

$$P\left(F_{p,n-p-1} > \frac{MSR}{MSE}\right) = 1 - F_{F_{p,n-p-1}}\left(\frac{MSR}{MSE}\right).$$

- W przypadku modelu regresji prostej test ten daje takie same wyniki jak test istotności współczynnika regresji.

Poprawiony współczynnik determinacji R^2_{adj}

- Dodanie nowej zmiennej niezależnej do modelu zawsze zwiększa R^2.
- Dlatego w praktyce wykorzystuje się głównie tak zwany poprawiony współczynnik determinacji R^2_{adj}.
- Bierze on pod uwagę, że R^2 jest obliczany na podstawie próby i jest nieco „zbyt dobry”, jeśli uogólnimy nasze wyniki na populację.

- Poprawiony współczynnik determinacji R^2_{adj} jest zawsze mniejszy niż R^2.
- Poprawiony współczynnik determinacji R^2_{adj} (zmiennej zależnej w modelu regresji wielokrotnej) określa, jaki procent wariancji zmiennej zależnej wyjaśnia model:

$$R^2_{adj} = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)} = 1 - \frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2/(n-p-1)}{\sum_{i=1}^n (y_i - \bar{y})^2/(n-1)} = 1 - \frac{(1-R^2)(n-1)}{n-p-1},$$

  gdzie $\hat{y}_i = \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}$, $i = 1, \ldots, n$. Różnice $y_i - \hat{y}_i$ są nazywane resztami lub residuami.
- Im wyższa wartość współczynnika $0 \leqslant R^2_{adj} \leqslant 1$, tym lepszy model (oczywiście w sensie tego kryterium). Należy również pamiętać, że ta ocena jakości modelu jest poprawna tylko wtedy, gdy model jest odpowiedni, tj. gdy założenia modelu są spełnione.

Predykcja

- Niech

$$\mathbf{x}_{nowa}^\top = (1, x_{nowy,1} \ldots, x_{nowy,p})$$

  będzie wektorem wartości zmiennych niezależnych $X_1, \ldots, X_p$, dla których chcemy przewidzieć wartość zmiennej zależnej Y. W oparciu o model regresji wielokrotnej predykcja jest następująca:

$$y_p = \mathbf{x}_{nowa}^\top \hat{\boldsymbol{\beta}}.$$

  Jako ocenę jakości prognozy przyjmujemy oszacowanie standardowego odchylenia prognozy (średni błąd prognozy) postaci:

$$S_p = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n-p-1} (1 + \mathbf{x}_{nowa}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_{nowa})}.$$

  Przy założeniu normalności przedział ufności dla y_p na poziomie ufności $1-\alpha$, $\alpha \in (0,1)$ jest postaci:

$$\left(y_p - t\left(1 - \frac{\alpha}{2}, n-p-1\right) S_p, y_p + t\left(1 - \frac{\alpha}{2}, n-p-1\right) S_p\right),$$

  gdzie $t(\beta, m)$ oznacza kwantyl rzędu β z rozkładu t-Studenta $t(m)$ z m stopniami swobody.

Stymulanty i destymulanty

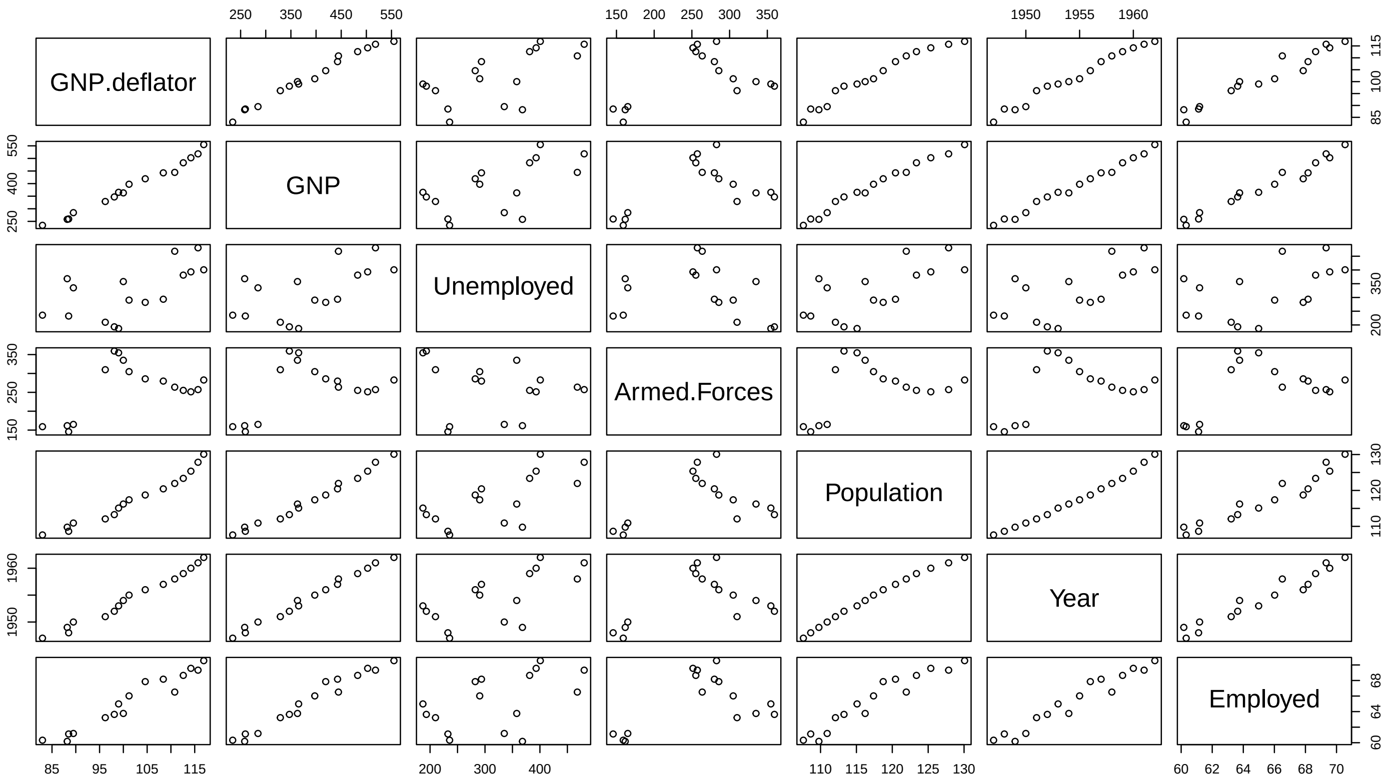
- Gdy stosowana jest regresja wielokrotna, często jesteśmy bardziej zainteresowani, które zmienne wpływają na badane zjawisko w sposób stymulujący, a które je hamują, niż prognozowaniem.
- Pierwsza z tych zmiennych nazywa się stymulantami, a drugie destymulantami.
- Oczywiście stymulanty to zmienne, które mają wartości parametrów regresji dodatniej w oszacowanym modelu regresji.
- Destymulanty to zmienne o ujemnych parametrach.
- Można również określić neutralne (nieistotne) zmienne, tj. zmienne, które nie mają wpływu na badane zjawisko.

Przykład. Zbiór danych `longley` zawiera 7 zmiennych makroekonomicznych. Chcemy modelować liczbę zatrudnionych za pomocą innych (niekoniecznie wszystkich) zmiennych przy użyciu modelu regresji wielokrotnej.

```r
head(longley)
```

```
##      GNP.deflator     GNP Unemployed Armed.Forces Population Year Employed
## 1947         83.0 234.289      235.6        159.0    107.608 1947   60.323
## 1948         88.5 259.426      232.5        145.6    108.632 1948   61.122
## 1949         88.2 258.054      368.2        161.6    109.773 1949   60.171
## 1950         89.5 284.599      335.1        165.0    110.929 1950   61.187
## 1951         96.2 328.975      209.9        309.9    112.075 1951   63.221
## 1952         98.1 346.999      193.2        359.4    113.270 1952   63.639
```

```r
pairs(longley)
```




```r
# model pełny
model_1 <- lm(Employed ~ ., data = longley)
# model_1 <- lm(Employed ~ GNP.deflator + GNP + Unemployed +
#                          Armed.Forces + Population + Year,
#                data = longley)
model_1
```

```
##
## Call:
## lm(formula = Employed ~ ., data = longley)
##
## Coefficients:
##  (Intercept)  GNP.deflator           GNP    Unemployed  Armed.Forces
##   -3.482e+03     1.506e-02    -3.582e-02    -2.020e-02    -1.033e-02
##   Population          Year
##   -5.110e-02     1.829e+00
```

```r
# estymacja parametrów
coef(model_1)
```

```
##   (Intercept)  GNP.deflator           GNP    Unemployed  Armed.Forces
## -3.482259e+03  1.506187e-02 -3.581918e-02 -2.020230e-02 -1.033227e-02
##    Population          Year
## -5.110411e-02  1.829151e+00
```

```r
confint(model_1)
```

```
##                      2.5 %        97.5 %
## (Intercept)  -5.496529e+03 -1.467988e+03
```

```
## GNP.deflator -1.770290e-01  2.071528e-01
## GNP          -1.115811e-01  3.994274e-02
## Unemployed   -3.125067e-02 -9.153930e-03
## Armed.Forces -1.517949e-02 -5.485050e-03
## Population   -5.625172e-01  4.603090e-01
## Year          7.987875e-01  2.859515e+00
```

```
# podsumowanie modelu
# tj. reszty, estymacja punktowa, testy istotności dla współczynników regresji,
# R_adj^2, test istnotności modelu (test analizy wariancji w regresji)
summary(model_1)
```

```
##
## Call:
## lm(formula = Employed ~ ., data = longley)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41011 -0.15767 -0.02816  0.10155  0.45539
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.482e+03  8.904e+02  -3.911 0.003560 **
## GNP.deflator  1.506e-02  8.492e-02   0.177 0.863141
## GNP          -3.582e-02  3.349e-02  -1.070 0.312681
## Unemployed   -2.020e-02  4.884e-03  -4.136 0.002535 **
## Armed.Forces -1.033e-02  2.143e-03  -4.822 0.000944 ***
## Population   -5.110e-02  2.261e-01  -0.226 0.826212
## Year          1.829e+00  4.555e-01   4.016 0.003037 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3049 on 9 degrees of freedom
## Multiple R-squared:  0.9955, Adjusted R-squared:  0.9925
## F-statistic: 330.3 on 6 and 9 DF,  p-value: 4.984e-10
```

```
# wartości dopasowane przez model
fitted(model_1)
```

```
##     1947     1948     1949     1950     1951     1952     1953     1954
## 60.05566 61.21601 60.12471 61.59711 62.91129 63.88831 65.15305 63.77418
##     1955     1956     1957     1958     1959     1960     1961     1962
## 66.00470 67.40161 68.18627 66.55206 68.81055 69.64967 68.98907 70.75776
```

```
# reszty
residuals(model_1)
```

```
##        1947        1948        1949        1950        1951        1952
##  0.26734003 -0.09401394  0.04628717 -0.41011462  0.30971459 -0.24931122
##        1953        1954        1955        1956        1957        1958
## -0.16404896 -0.01318036  0.01430477  0.45539409 -0.01726893 -0.03905504
##        1959        1960        1961        1962
## -0.15554997 -0.08567131  0.34193151 -0.20675783
```

```r
# predykcja
new_data <- data.frame(GNP.deflator = 115.4,
                       GNP = 518.163,
                       Unemployed = 480.3,
                       Armed.Forces = 257.4,
                       Population = 127.857,
                       Year = 1963)
stats::predict(model_1, new_data, interval = "prediction")
```

```
##        fit      lwr      upr
## 1 72.64695 70.55039 74.74351
```

```r
# redukcja modelu pełnego
summary(model_1)
```

```
##
## Call:
## lm(formula = Employed ~ ., data = longley)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41011 -0.15767 -0.02816  0.10155  0.45539
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.482e+03  8.904e+02  -3.911 0.003560 **
## GNP.deflator  1.506e-02  8.492e-02   0.177 0.863141
## GNP          -3.582e-02  3.349e-02  -1.070 0.312681
## Unemployed   -2.020e-02  4.884e-03  -4.136 0.002535 **
## Armed.Forces -1.033e-02  2.143e-03  -4.822 0.000944 ***
## Population   -5.110e-02  2.261e-01  -0.226 0.826212
## Year          1.829e+00  4.555e-01   4.016 0.003037 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3049 on 9 degrees of freedom
## Multiple R-squared:  0.9955, Adjusted R-squared:  0.9925
## F-statistic: 330.3 on 6 and 9 DF,  p-value: 4.984e-10
```

```r
model_2 <- lm(Employed ~ Unemployed + Armed.Forces + Year, data = longley)
# model_2 <- update(model_1, . ~ . - GNP.deflator - GNP - Population)
summary(model_2)
```

```
##
## Call:
## lm(formula = Employed ~ Unemployed + Armed.Forces + Year, data = longley)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.57285 -0.11989  0.04087  0.13979  0.75303
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  -1.797e+03  6.864e+01 -26.183 5.89e-12 ***
## Unemployed   -1.470e-02  1.671e-03  -8.793 1.41e-06 ***
## Armed.Forces -7.723e-03  1.837e-03  -4.204  0.00122 **
## Year          9.564e-01  3.553e-02  26.921 4.24e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3321 on 12 degrees of freedom
## Multiple R-squared:  0.9928, Adjusted R-squared:  0.9911
## F-statistic: 555.2 on 3 and 12 DF,  p-value: 3.916e-13
```

7.3 Regresja krokowa

- Istnieje również inna metoda konstrukcji modeli z dużą liczbą zmiennych objaśniających niż konstruowanie pełnego modelu i szacowanie jego parametrów (tak jak robimy to w regresji wielokrotnej).
- Jest to procedura regresji krokowej, w której możemy odrzucić lub dodać zmienną na każdym kroku.
- Powiedzmy, że zaczynamy od modelu zawierającego tylko stałą - „regresja w przód” (możemy też zacząć od pełnego modelu - „regresja w tył”). W następnym kroku dodajemy najlepszą zmienną w sensie kryterium (np. test istotności, AIC, BIC). W następnym dodamy ponownie, ale możemy również sprawdzić co się dzieje, jakbyśmy usunęli z modelu zmienną dodaną w poprzednim kroku, itd.
- Jakość modelu jest zwykle oceniana przy użyciu kryterium informacyjnego Akaike (AIC).
- Wartość tego kryterium zależy nie tylko od sumy kwadratów reszt, ale także od liczby zmiennych w modelu.
- Zatem zwiększając liczbę parametrów w modelu, chociaż suma kwadratów reszt zawsze maleje, od pewnego momentu AIC zacznie rosnąć.
- AIC zwykle wybiera model o zbyt wielu parametrach. Istnieją zatem również modyfikacje i alternatywy dla AIC.
- Jeśli bardziej zależy nam na jakości prognozy powinniśmy użyć AIC, a jeśli priorytetem jest jakość dopasowania modelu wybieramy BIC (bayesowskiej kryterium informacyjne).
- Mamy
$$AIC = -2\ln(L) + 2(p+1), \quad BIC = -2\ln(L) + \ln(n)(p+1),$$
gdzie L jest maksimum funkcji wiarogodności oraz ln jest logarytmem naturalnym.
- W przypadku estymacji parametrów metodą najmniejszych kwadratów i przy założeniu normalności błędów mamy:
  - bez wyrazu stałego:
$$AIC = n\ln\left(\frac{1}{n}\sum_{i=1}^{n} e_i^2\right) + 2(p+1),$$
  - z wyrazem stałym:
$$AIC = n\ln\left(\frac{1}{n}\sum_{i=1}^{n} e_i^2\right) + 2(p+1) + n\ln(2\pi) + n + 2.$$
- Im mniejsza wartość AIC lub BIC, tym lepszy model.

Przykład (cd.).

```r
model_1 <- lm(Employed ~ ., data = longley)
model_2 <- lm(Employed ~ Unemployed + Armed.Forces + Year, data = longley)
# AIC (z wyrazem stałym)
AIC(model_1, model_2)
```

```
##         df      AIC
## model_1  8 14.18670
## model_2  5 15.52741
```

```r
n <- nrow(longley)
p <- 6
n * log(mean(model_1$residuals^2)) + 2 * (p + 1) + n * log(2 * pi) + n + 2
```

```
## [1] 14.1867
```

```r
p <- 3
n * log(mean(model_2$residuals^2)) + 2 * (p + 1) + n * log(2 * pi) + n + 2
```

```
## [1] 15.52741
```

```r
# BIC (z wyrazem stałym)
AIC(model_1, model_2, k = log(nrow(longley)))
```

```
##         df      AIC
## model_1  8 20.36741
## model_2  5 19.39035
```

```r
# AIC i BIC (bez wyrazu stałego)
extractAIC(model_1)[2]
```

```
## [1] -33.21933
```

```r
extractAIC(model_1, k = log(n))[2]
```

```
## [1] -27.81121
```

```r
p <- 6
n * log(mean(model_1$residuals^2)) + 2 * (p + 1)
```

```
## [1] -33.21933
```

```r
n * log(mean(model_1$residuals^2)) + log(n) * (p + 1)
```

```
## [1] -27.81121
```

```r
extractAIC(model_2)[2]
```

```
## [1] -31.87863
```

```r
extractAIC(model_2, k = log(n))[2]
```

```
## [1] -28.78827
```

```r
p <- 3
n * log(mean(model_2$residuals^2)) + 2 * (p + 1)
```

```
## [1] -31.87863
```

```r
n * log(mean(model_2$residuals^2)) + log(n) * (p + 1)
```

```
## [1] -28.78827
```

```r
# regresja krokowa
step(model_1)
```

```
## Start:  AIC=-33.22
## Employed ~ GNP.deflator + GNP + Unemployed + Armed.Forces + Population +
##     Year
##
##                Df Sum of Sq     RSS     AIC
## - GNP.deflator  1   0.00292 0.83935 -35.163
## - Population    1   0.00475 0.84117 -35.129
## - GNP           1   0.10631 0.94273 -33.305
## <none>                      0.83642 -33.219
## - Year          1   1.49881 2.33524 -18.792
## - Unemployed    1   1.59014 2.42656 -18.178
## - Armed.Forces  1   2.16091 2.99733 -14.798
##
## Step:  AIC=-35.16
## Employed ~ GNP + Unemployed + Armed.Forces + Population + Year
##
##                Df Sum of Sq    RSS     AIC
## - Population    1   0.01933 0.8587 -36.799
## <none>                      0.8393 -35.163
## - GNP           1   0.14637 0.9857 -34.592
## - Year          1   1.52725 2.3666 -20.578
## - Unemployed    1   2.18989 3.0292 -16.628
## - Armed.Forces  1   2.39752 3.2369 -15.568
##
## Step:  AIC=-36.8
## Employed ~ GNP + Unemployed + Armed.Forces + Year
##
##                Df Sum of Sq    RSS     AIC
## <none>                      0.8587 -36.799
## - GNP           1    0.4647 1.3234 -31.879
## - Year          1    1.8980 2.7567 -20.137
## - Armed.Forces  1    2.3806 3.2393 -17.556
## - Unemployed    1    4.0491 4.9077 -10.908

##
## Call:
## lm(formula = Employed ~ GNP + Unemployed + Armed.Forces + Year,
##     data = longley)
##
## Coefficients:
##  (Intercept)          GNP    Unemployed  Armed.Forces          Year
##   -3.599e+03   -4.019e-02    -2.088e-02    -1.015e-02     1.887e+00
```

```r
# step(model_1, direction = "backward")
step(model_1, k = log(nrow(longley)))
```

```
## Start:  AIC=-27.81
## Employed ~ GNP.deflator + GNP + Unemployed + Armed.Forces + Population +
##     Year
```

```
##
##                Df Sum of Sq     RSS     AIC
## - GNP.deflator  1   0.00292 0.83935 -30.528
## - Population    1   0.00475 0.84117 -30.493
## - GNP           1   0.10631 0.94273 -28.669
## <none>                      0.83642 -27.811
## - Year          1   1.49881 2.33524 -14.156
## - Unemployed    1   1.59014 2.42656 -13.542
## - Armed.Forces  1   2.16091 2.99733 -10.162
##
## Step:  AIC=-30.53
## Employed ~ GNP + Unemployed + Armed.Forces + Population + Year
##
##                Df Sum of Sq    RSS     AIC
## - Population    1   0.01933 0.8587 -32.936
## - GNP           1   0.14637 0.9857 -30.729
## <none>                      0.8393 -30.528
## - Year          1   1.52725 2.3666 -16.715
## - Unemployed    1   2.18989 3.0292 -12.765
## - Armed.Forces  1   2.39752 3.2369 -11.705
##
## Step:  AIC=-32.94
## Employed ~ GNP + Unemployed + Armed.Forces + Year
##
##                Df Sum of Sq    RSS     AIC
## <none>                      0.8587 -32.936
## - GNP           1    0.4647 1.3234 -28.788
## - Year          1    1.8980 2.7567 -17.046
## - Armed.Forces  1    2.3806 3.2393 -14.466
## - Unemployed    1    4.0491 4.9077  -7.818
```

```
##
## Call:
## lm(formula = Employed ~ GNP + Unemployed + Armed.Forces + Year,
##     data = longley)
##
## Coefficients:
##  (Intercept)           GNP    Unemployed  Armed.Forces          Year
##   -3.599e+03    -4.019e-02    -2.088e-02    -1.015e-02     1.887e+00
```

```r
model_0 <- lm(Employed ~ 1, data = longley)
step(model_0, direction = "forward", scope = formula(model_1))
```

```
## Start:  AIC=41.17
## Employed ~ 1
##
##                Df Sum of Sq     RSS     AIC
## + GNP           1   178.973   6.036 -11.597
## + Year          1   174.552  10.457  -2.806
## + GNP.deflator  1   174.397  10.611  -2.571
## + Population    1   170.643  14.366   2.276
## + Unemployed    1    46.716 138.293  38.509
## + Armed.Forces  1    38.691 146.318  39.411
```

```
## <none>                       185.009  41.165
##
## Step:  AIC=-11.6
## Employed ~ GNP
##
##                Df Sum of Sq    RSS      AIC
## + Unemployed    1   2.45708 3.5791 -17.9598
## + Population    1   2.16178 3.8744 -16.6913
## + Year          1   1.12520 4.9109 -12.8980
## <none>                      6.0361 -11.5972
## + GNP.deflator  1   0.21194 5.8242 -10.1691
## + Armed.Forces  1   0.07665 5.9595  -9.8017
##
## Step:  AIC=-17.96
## Employed ~ GNP + Unemployed
##
##                Df Sum of Sq    RSS     AIC
## + Armed.Forces  1   0.82235 2.7567 -20.137
## <none>                      3.5791 -17.960
## + Year          1   0.33980 3.2393 -17.556
## + Population    1   0.09682 3.4822 -16.399
## + GNP.deflator  1   0.01884 3.5602 -16.044
##
## Step:  AIC=-20.14
## Employed ~ GNP + Unemployed + Armed.Forces
##
##                Df Sum of Sq     RSS     AIC
## + Year          1   1.89803 0.85868 -36.799
## + Population    1   0.39011 2.36660 -20.578
## <none>                      2.75671 -20.137
## + GNP.deflator  1   0.07288 2.68383 -18.566
##
## Step:  AIC=-36.8
## Employed ~ GNP + Unemployed + Armed.Forces + Year
##
##                Df Sum of Sq     RSS     AIC
## <none>                       0.85868 -36.799
## + Population    1  0.019332 0.83935 -35.163
## + GNP.deflator  1  0.017507 0.84117 -35.129

##
## Call:
## lm(formula = Employed ~ GNP + Unemployed + Armed.Forces + Year,
##     data = longley)
##
## Coefficients:
##  (Intercept)           GNP    Unemployed  Armed.Forces          Year
##   -3.599e+03    -4.019e-02    -2.088e-02    -1.015e-02     1.887e+00
```

```
step(model_0, direction = "forward", scope = formula(model_1), k = log(nrow(longley)))
```

```
## Start:  AIC=41.94
## Employed ~ 1
```

```
##
##                 Df Sum of Sq     RSS     AIC
## + GNP           1   178.973   6.036 -10.052
## + Year          1   174.552  10.457  -1.261
## + GNP.deflator  1   174.397  10.611  -1.025
## + Population    1   170.643  14.366   3.822
## + Unemployed    1    46.716 138.293  40.054
## + Armed.Forces  1    38.691 146.318  40.956
## <none>                     185.009  41.938
##
## Step:  AIC=-10.05
## Employed ~ GNP
##
##                 Df Sum of Sq    RSS      AIC
## + Unemployed    1   2.45708 3.5791 -15.6420
## + Population    1   2.16178 3.8744 -14.3736
## + Year          1   1.12520 4.9109 -10.5802
## <none>                      6.0361 -10.0520
## + GNP.deflator  1   0.21194 5.8242  -7.8513
## + Armed.Forces  1   0.07665 5.9595  -7.4839
##
## Step:  AIC=-15.64
## Employed ~ GNP + Unemployed
##
##                 Df Sum of Sq    RSS     AIC
## + Armed.Forces  1   0.82235 2.7567 -17.046
## <none>                      3.5791 -15.642
## + Year          1   0.33980 3.2393 -14.466
## + Population    1   0.09682 3.4822 -13.308
## + GNP.deflator  1   0.01884 3.5602 -12.954
##
## Step:  AIC=-17.05
## Employed ~ GNP + Unemployed + Armed.Forces
##
##                 Df Sum of Sq     RSS     AIC
## + Year          1   1.89803 0.85868 -32.936
## <none>                      2.75671 -17.046
## + Population    1   0.39011 2.36660 -16.715
## + GNP.deflator  1   0.07288 2.68383 -14.703
##
## Step:  AIC=-32.94
## Employed ~ GNP + Unemployed + Armed.Forces + Year
##
##                 Df Sum of Sq     RSS     AIC
## <none>                      0.85868 -32.936
## + Population    1  0.019332 0.83935 -30.528
## + GNP.deflator  1  0.017507 0.84117 -30.493

##
## Call:
## lm(formula = Employed ~ GNP + Unemployed + Armed.Forces + Year,
##     data = longley)
```

```
##
## Coefficients:
##  (Intercept)          GNP    Unemployed  Armed.Forces          Year
##   -3.599e+03   -4.019e-02   -2.088e-02    -1.015e-02     1.887e+00
```

7.4 Uogólniony model liniowy

- W wielu sytuacjach nie możemy założyć, że zmienna zależna jest ciągła.
- W tej sytuacji powinniśmy zastosować uogólnione modele liniowe (GLM), w których zakłada się pewien rozkład zmiennej zależnej (dopuszczalne są rozkłady z tak zwanej wykładniczej rodziny rozkładów, np. rozkład normalny, wykładniczy, gamma, Poissona, dwumianowy, geometryczny i wielomianowy).
- Ponadto, aby uwzględnić nieliniowy charakter zależności, bierze się pod uwagę tak zwaną funkcję wiążącą h, która ma następującą własność:

$$E(Y) = h^{-1}(\mathbf{X}\boldsymbol{\beta}).$$

- Zauważmy, że jeśli funkcją wiążąca jest tożsamościowa ($h(x) = x$), a zmienna zależna ma rozkład normalny, model ten jest modelem regresji liniowej ($E(Y) = \mathbf{X}\boldsymbol{\beta}$).
- Inne przykłady są następujące:
  - rozkład Y: zero-jedynkowy, $h(x) = \ln\left(\frac{x}{1-x}\right)$, $E(Y) = \frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1+\exp(\mathbf{X}\boldsymbol{\beta})}$ - regresja logistyczna (h to funkcja logitowa lub logistyczna)
  - rozkład Y: -, $h(x) = \Phi^{-1}(x)$, $E(Y) = \Phi(\mathbf{X}\boldsymbol{\beta})$ - regresja probitowa (Φ jest dystrybuantą rozkładu normalnego $N(0, 1)$. Φ^{-1} to funkcja probitowa.)
  - rozkład Y: Poisson, $h(x) = \ln(x)$, $E(Y) = \exp(\mathbf{X}\boldsymbol{\beta})$ - regresja Poissona
  - rozkład Y: gamma, $h(x) = \frac{1}{x}$, $E(Y) = \frac{1}{\mathbf{X}\boldsymbol{\beta}}$

7.4.1 Regresja logistyczna

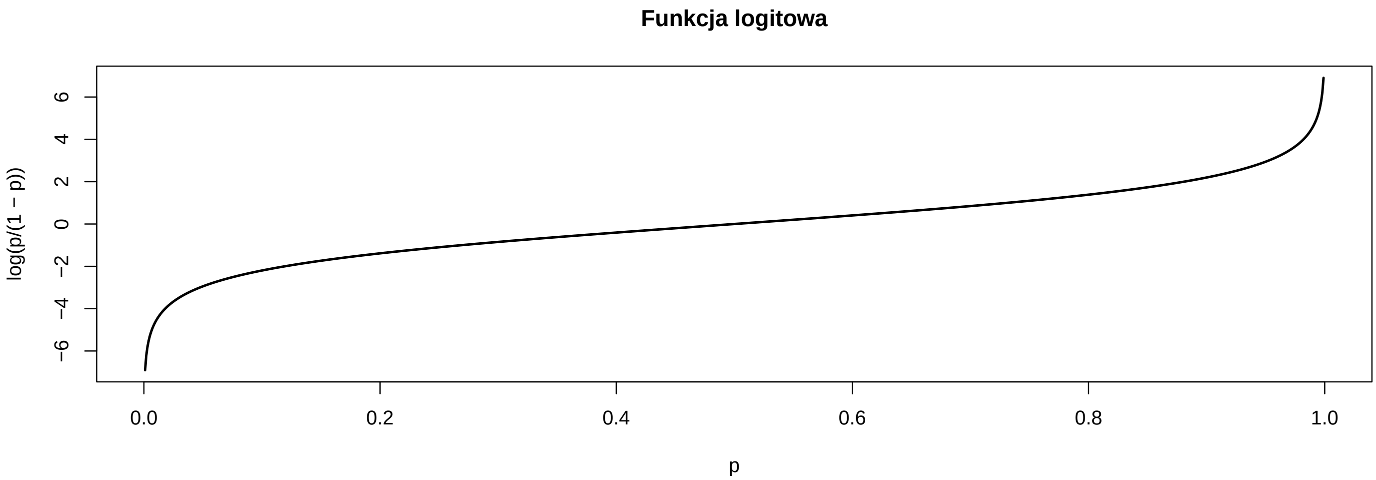
- Regresja logistyczna jest szczególnym i ważnym przykładem uogólnionego modelu liniowego.
- Formalnie w tym przypadku zakładamy, że

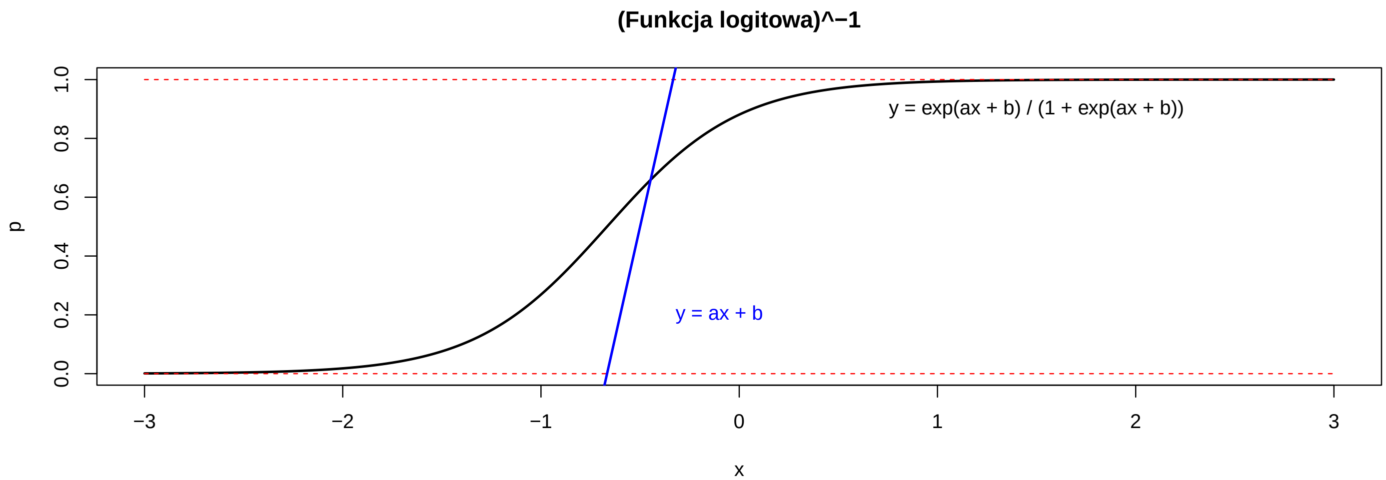
$$Y \sim b(p).$$

- Oznacza to, że zmienna objaśniana ma tylko dwie wartości (najczęściej jest to zmienna binarna).
- Modelujemy prawdopodobieństwo sukcesu p.
- Funkcją łączącą jest funkcja logitowa:

$$h(p) = logit(p) = \ln\left(\frac{p}{1-p}\right).$$

- Prawdopodobieńtwo p jest wtedy szaconane następująco: $p = E(Y) = \frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1+\exp(\mathbf{X}\boldsymbol{\beta})}$.





- Nie można użyć metody najmniejszych kwadratów do oszacowania parametrów w wektorze $\boldsymbol{\beta}$.
- Wymaga to założenia o jednorodności wariancji, która w przypadku zmiennej binarnej nie jest spełniona.
- W przypadku regresji logistycznej stosujemy metodę największej wiarogodności, stosując iteracyjny algorytm ważonej metody najmniejszych kwadratów.
- Wartości oszacowanych współczynników nie podlegają interpretacji.
- Interpretacja podlega ilorazowi szans (OR, ang. odds ratio), który można wyrazić jako

$$OR = \frac{p}{1-p} = e^{\mathbf{X}\boldsymbol{\beta}} = e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p} = e^{\beta_0} e^{\beta_1 X_1} \cdots e^{\beta_p X_p}.$$

- Jeżeli $e^{\beta_j} > 1$, to zmienna X_j działa stymulująco możliwość wystąpienia badanego zjawiska, a w przeciwnym razie działa ograniczająco (jeżeli $e^{\beta_j} = 1$, to zmienna X_j nie ma wpływu na badane zjawisko).
- Jakość dopasowania można przetestować jak poprzednio dla regresji wielokrotnej przy użyciu kryterium informacyjnego, ale w przypadku regresji logistycznej bardziej efektywne są inne kryteria.
- Jednym z nich są krzywe charakterystyczne.
- Model regresji logistycznej można traktować jako model służący do diagnozowania dwóch stanów: dobrego lub złego. Model oblicza prawdopodobieństwo „dobrego” stanu. Wybieramy pewien próg $0 < t < 1$, jeśli prawdopodobieństwo uzyskane z modelu przekracza t, diagnozujemy stan jako „dobry‘‘, w przeciwnym razie jako „zły”. Mamy więc cztery opcje:
  - TP (true positive) - model przewidział „dobry” oraz zaobserwowano „dobry”,
  - TN (true negative) - model przewidział „zły” oraz zaobserwowano „zły”,

- FP (false positive) - model przewidział „dobry” oraz zaobserwowano „zły”,
- FN (false negative) - model przewidział „zły” oraz zaobserwowano „dobry”,

		zaobserwowano	
		dobry	zły
przewidziano	dobry	TP	FP
	zły	FN	TN

- Jeżeli przez n_g oznaczymy liczbę zaobserwowanych „dobry‘‘, a przez n_b „zły”, to

$$TPR = \frac{TP}{n_g}, \quad TNR = \frac{TN}{n_b}, \quad FPR = 1 - TNR, \quad FNR = 1 - TPR.$$

- Krzywa charakterystyka (ROC, ang. receiver operating characteristic) jest wykresem współczynnika TPR na osi pionowej przeciwko współczynnikowi FPR na osi poziomej dla wszystkich wartości progowych t. Krzywa ROC jest zatem rodziną punktów (FPR, TPR) ilustrującą związek między zdolnością do rozróżniania przypadków pozytywnych i negatywnych dla różnych parametrów modelu.
- Aby teraz zmierzyć jakość modelu, liczy się pole pod krzywą ROC (AUC).
- AUC jest powszechnie znaną miarą przyjmująca wartościw przedziale $[0, 1]$. Wyższa wartość oznacza lepszą wydajność.
- Im wartość AUC jest bliższa 1, tym lepsza jest zdolność modelu do przewidywania „dobrego” stanu.
- Idealny model poprawnie przewidujący wszystkie przypadki ma wartość AUC $=$ 1, podczas gdy model całkowicie losowy (słaby) ma AUC $=$ 0,5.
- Ale co oznacza wynik 0,8 lub 0,95? Sensowna interpretacja wyniku jest następująca: jeśli weźmiemy losowy przypadek pozytywny i losowy wynik negatywny, to AUC pokazuje prawdopodobieństwo, że model przypisuje wyższy wynik przypadkowi pozytywnemu niż negatywnemu.
- Chronologicznie wcześniejszym modelem był model probitowy niż model logistyczny. Obecnie, ze względu na prostszą interpretację, stosuje się model logistyczny. Mianowicie, w modelu probitowym nie można nic powiedzieć o wpływie zmiennych na prawdopodobieństwo, podczas gdy w modelu logistycznym, jeśli współczynnik obok zmiennej jest dodatni (ujemny) oznacza to, że zmienna ma pozytywny (negatywny) wpływ na szansę sukcesu. Dla obu modeli uzyskuje się zwykle bardzo podobne wyniki.

Przykład. Rozważmy przykład dotyczący badania szansy ponownego ataku serca w ciągu roku od pierwszego ataku, w zależności od *treatment of anger* oraz *trait anxiety*. Zmienna zależna ma wartość 1, jeśli nastąpił ponowny atak, a 0 w przeciwnym razie.

```
y <- c(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
x1 <- c(1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0)
x2 <- c(70, 80, 50, 60, 40, 65, 75, 80, 70, 60, 65, 50, 45, 35, 40, 50, 55, 45, 50, 60)
data_set <- data.frame(y, x1, x2)
head(data_set)
```

```
##   y x1 x2
## 1 1  1 70
## 2 1  1 80
## 3 1  1 50
## 4 1  0 60
## 5 1  0 40
## 6 1  0 65
```

```r
# model logistyczny
model_1 <- glm(y ~ x1 + x2, data = data_set, family = 'binomial')
model_1
```

```
##
## Call:  glm(formula = y ~ x1 + x2, family = "binomial", data = data_set)
##
## Coefficients:
## (Intercept)           x1           x2
##      -6.363       -1.024        0.119
##
## Degrees of Freedom: 19 Total (i.e. Null);  17 Residual
## Null Deviance:       27.73
## Residual Deviance: 18.82     AIC: 24.82
```

```r
# podsumowanie modelu
# tj. reszty, estymacja punktowa, testy istotności dla współczynników regresji, AIC
summary(model_1)
```

```
##
## Call:
## glm(formula = y ~ x1 + x2, family = "binomial", data = data_set)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.52106  -0.68746   0.00424   0.70625   1.88960
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.36347    3.21362  -1.980   0.0477 *
## x1          -1.02411    1.17101  -0.875   0.3818
## x2           0.11904    0.05497   2.165   0.0304 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 27.726  on 19  degrees of freedom
## Residual deviance: 18.820  on 17  degrees of freedom
## AIC: 24.82
##
## Number of Fisher Scoring iterations: 4
```

```r
# zredukowany model logistyczny
model_2 <- glm(y ~ x2, data = data_set, family = 'binomial')
summary(model_2)
```

```
##
## Call:
## glm(formula = y ~ x2, family = "binomial", data = data_set)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
```

```
## -1.62461  -0.83983  -0.01232   0.64540   2.10801
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -7.0925     3.1709  -2.237   0.0253 *
## x2            0.1246     0.0553   2.254   0.0242 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 27.726  on 19  degrees of freedom
## Residual deviance: 19.601  on 18  degrees of freedom
## AIC: 23.601
##
## Number of Fisher Scoring iterations: 4
```

```r
# regresja krokowa
AIC(model_1, model_2)
```

```
##         df      AIC
## model_1  3 24.82037
## model_2  2 23.60052
```

```r
step(model_1)
```

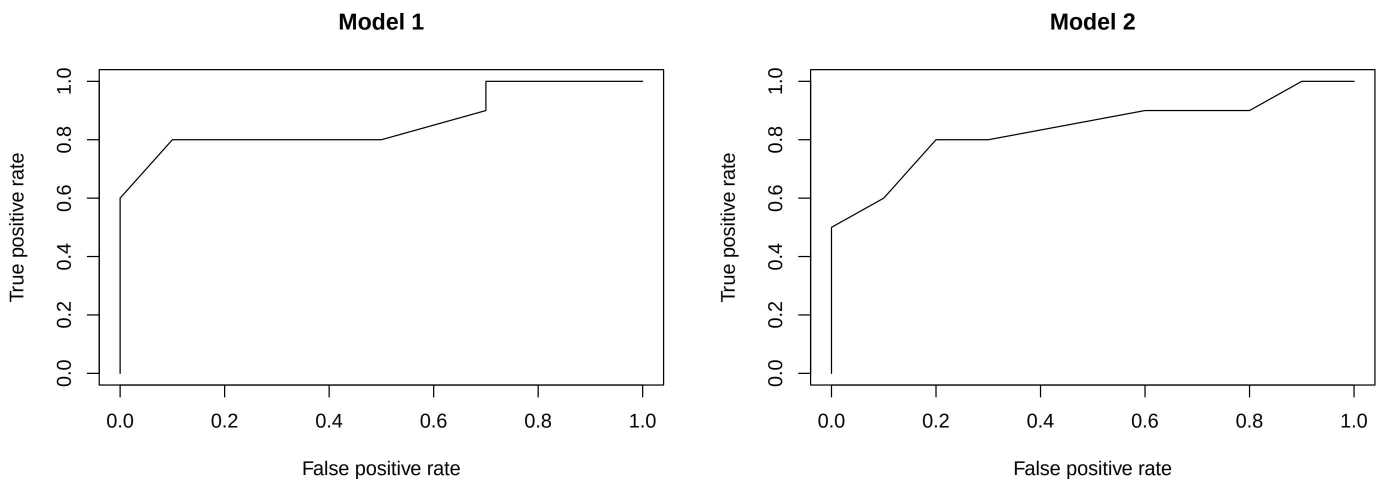
```
## Start:  AIC=24.82
## y ~ x1 + x2
##
##        Df Deviance    AIC
## - x1    1   19.601 23.601
## <none>      18.820 24.820
## - x2    1   25.878 29.878
##
## Step:  AIC=23.6
## y ~ x2
##
##        Df Deviance    AIC
## <none>      19.601 23.601
## - x2    1   27.726 29.726
```

```
##
## Call:  glm(formula = y ~ x2, family = "binomial", data = data_set)
##
## Coefficients:
## (Intercept)           x2
##     -7.0925       0.1246
##
## Degrees of Freedom: 19 Total (i.e. Null);  18 Residual
## Null Deviance:      27.73
## Residual Deviance: 19.6  AIC: 23.6
```

```r
# iloraz szans (ręcznie)
exp(coef(model_2)[2])
```

```
##       x2
## 1.132734
```

```
# Wartość ta oznacza, że wraz ze wzrostem wartości zmiennej x2 o jedną jednostkę,
# przewidywane ryzyko (w sensie ilorazu szans p / (1 - p))
# ponownego zawału serca wzrasta o 13%.
# do krzywych ROC
library(ROCR)
pred_1 <- prediction(model_1$fitted, y)
pred_2 <- prediction(model_2$fitted, y)
# krzywe ROC
par(mfrow = c(1, 2))
plot(performance(pred_1, 'tpr', 'fpr'), main = "Model 1")
plot(performance(pred_2, 'tpr', 'fpr'), main = "Model 2")
```




```
par(mfrow = c(1, 1))
# AUC
performance(pred_1, 'auc')@y.values
```

```
## [[1]]
## [1] 0.86
```

```
performance(pred_2, 'auc')@y.values
```

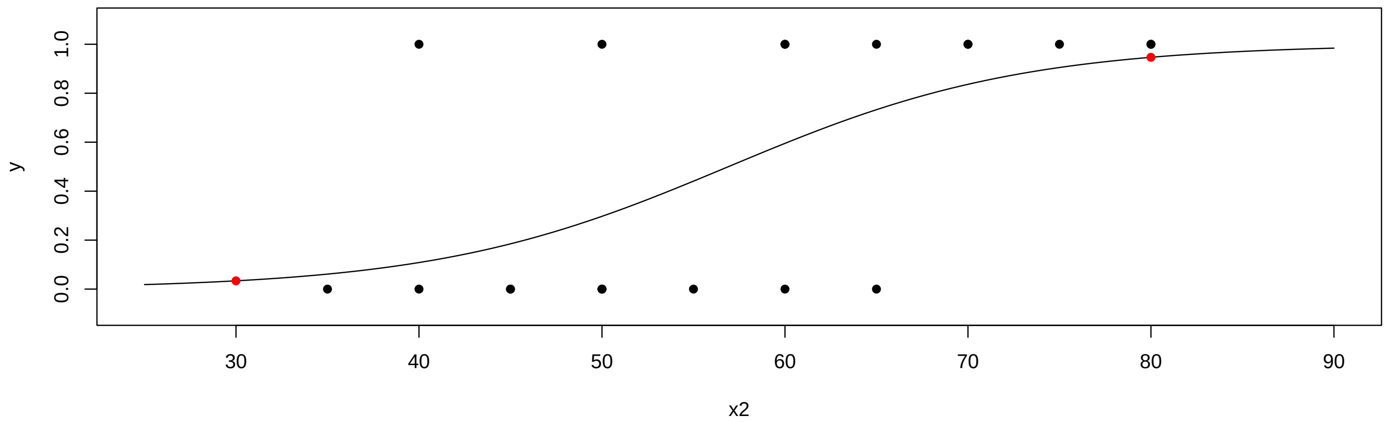
```
## [[1]]
## [1] 0.835
```

```
# predykcja
(predict_glm <- stats::predict(model_2,
                        data.frame(x2 = c(30, 80)),
                        type = 'response'))
```

```
##          1          2
## 0.03378247 0.94676209
```

```
# Uwzględniamy argument type = 'response' w celu uzyskania przewidywanego prawdopodobieństwa, że y
# Domyślne przewidywane są zlogarytmowane ilorazy szans (prawdopodobieństwa w skali logitowej).
x_temp <- seq(min(x2) - 10, max(x2) + 10, length.out = 100)
y_temp <- exp(coef(model_2)[1] + coef(model_2)[2] * x_temp) /
  (1 + exp(coef(model_2)[1] + coef(model_2)[2] * x_temp))
plot(x_temp, y_temp, type = "l", xlab = "x2", ylab = "y", ylim = c(-0.1, 1.1))
```

```r
points(x2, y, pch = 16)
points(c(30, 80), predict_glm, pch = 16, col = "red")
```



7.4.2 Regresja Poissona

Nie zawsze interesuje nas prawdopodobieństwo sukcesu. Dość często jesteśmy zainteresowani liczbą sukcesów (ogólnie liczebnościami). W tej sytuacji najbardziej popularny jest model Poissona, który zakłada, że zmienna zależna ma rozkład Poissona i

$$h(x) = \ln(x), \quad E(Y) = \exp(\mathbf{X}\boldsymbol{\beta}).$$

Przykład. W zbiorze danych student_award.RData, zmienna `num_awards` podaje liczbę nagród zdobytych przez uczniów szkoły średniej przez rok, zmienna `math` jest zmienną ciągłą i reprezentuje wyniki uczniów na końcowym egzaminie z matematyki, a zmienna `prog` jest zmienną jakościową z trzema poziomami wskazującymi rodzaj programu, ma który uczniowie byli zapisani ("General" - ogólny, "Academic" - akademicki, "Vocational" - zawodowy). Chcemy opisać związek między liczbą nagród a wynikiem egzaminu z matematyki i programem.

```r
load(url("http://ls.home.amu.edu.pl/data_sets/student_award.RData"))
head(student_award)
```

```
##   num_awards math       prog
## 1          0   41 Vocational
## 2          0   41    General
## 3          0   44 Vocational
## 4          0   42 Vocational
## 5          0   40 Vocational
## 6          0   42    General
```

```r
model_1 <- glm(num_awards ~ math + prog, data = student_award, family = "poisson")
model_1
```

```
##
## Call:  glm(formula = num_awards ~ math + prog, family = "poisson", data = student_award)
##
## Coefficients:
##    (Intercept)            math    progAcademic  progVocational
##       -5.24712         0.07015         1.08386         0.36981
##
## Degrees of Freedom: 199 Total (i.e. Null);  196 Residual
## Null Deviance:       287.7
```

```
## Residual Deviance: 189.4     AIC: 373.5
```

```r
summary(model_1)
```

```
##
## Call:
## glm(formula = num_awards ~ math + prog, family = "poisson", data = student_award)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2043  -0.8436  -0.5106   0.2558   2.6796
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -5.24712    0.65845  -7.969 1.60e-15 ***
## math            0.07015    0.01060   6.619 3.63e-11 ***
## progAcademic    1.08386    0.35825   3.025  0.00248 **
## progVocational  0.36981    0.44107   0.838  0.40179
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 287.67  on 199  degrees of freedom
## Residual deviance: 189.45  on 196  degrees of freedom
## AIC: 373.5
##
## Number of Fisher Scoring iterations: 6
```

```r
# Możemy również przetestować ogólny efekt programu, porównując pełny model
# z modelem bez zmiennej program. Test chi-kwadrat wskazuje, że program,
# jest statystycznie istotnym predyktorem liczby nagród.
model_2 <- update(model_1, . ~ . - prog)
anova(model_1, model_2, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: num_awards ~ math + prog
## Model 2: num_awards ~ math
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       196     189.45
## 2       198     204.02 -2  -14.572 0.0006852 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
AIC(model_1, model_2)
```

```
##         df      AIC
## model_1  4 373.5045
## model_2  2 384.0762
```

```r
step(model_1)
```

```
## Start:  AIC=373.5
## num_awards ~ math + prog
```

```
##
##        Df Deviance    AIC
## <none>      189.45 373.50
## - prog  2   204.02 384.08
## - math  1   234.46 416.51
```

```
##
## Call:  glm(formula = num_awards ~ math + prog, family = "poisson", data = student_award)
##
## Coefficients:
##    (Intercept)           math    progAcademic  progVocational
##       -5.24712        0.07015         1.08386         0.36981
##
## Degrees of Freedom: 199 Total (i.e. Null);  196 Residual
## Null Deviance:       287.7
## Residual Deviance: 189.4     AIC: 373.5
```

```
(data_new <- data.frame(math = mean(student_award$math),
                        prog = factor(1:3, levels = 1:3,
                                      labels = levels(student_award$prog))))
```

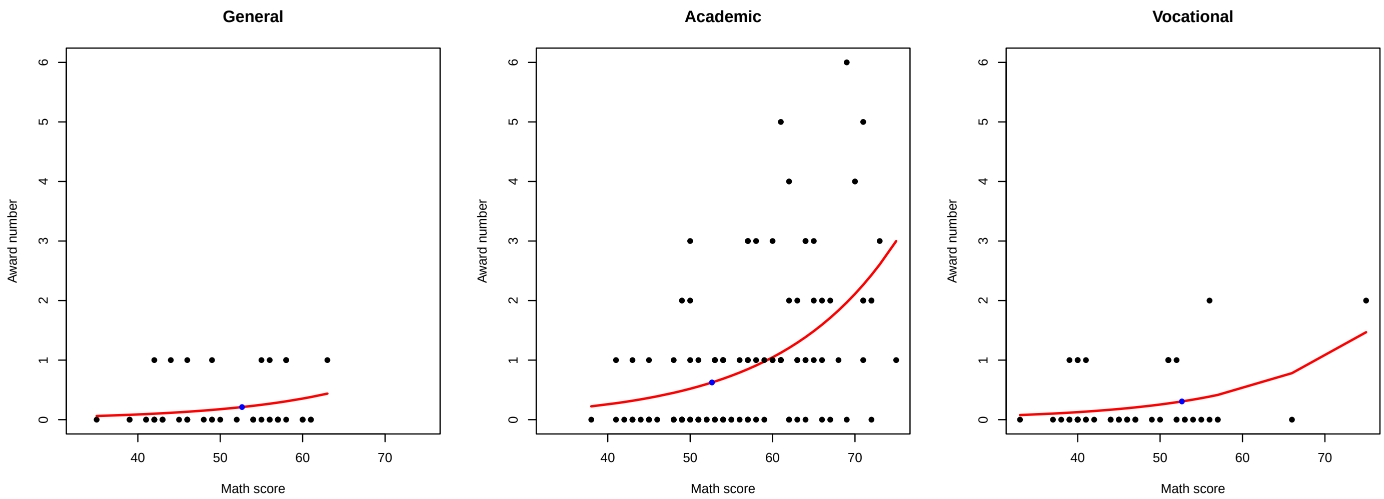
```
##     math       prog
## 1 52.645    General
## 2 52.645   Academic
## 3 52.645 Vocational
```

```
(pred <- stats::predict(model_1, data_new, type = "response"))
```

```
##         1         2         3
## 0.2114109 0.6249446 0.3060086
```

```
student_award$num_award_hat <- stats::predict(model_1, type = "response")
# sortowanie według programu, a następnie według wyniku z matematyki
student_award <- student_award[with(student_award, order(prog, math)), ]
par(mfrow = c(1, 3))
plot(student_award$math[student_award$prog == "General"],
     student_award$num_award_hat[student_award$prog == "General"],
     type = "l", lwd = 2, col = "red",
     xlim = c(min(student_award$math), max(student_award$math)), ylim = c(0, 6),
     xlab = "Math score", ylab = "Award number", main = "General")
points(student_award$math[student_award$prog == "General"],
       student_award$num_awards[student_award$prog == "General"], pch = 16)
points(mean(student_award$math), pred[1], pch = 16, col = "blue", lwd = 4)
plot(student_award$math[student_award$prog == "Academic"],
     student_award$num_award_hat[student_award$prog == "Academic"],
     type = "l", lwd = 2, col = "red",
     xlim = c(min(student_award$math), max(student_award$math)), ylim = c(0, 6),
     xlab = "Math score", ylab = "Award number", main = "Academic")
points(student_award$math[student_award$prog == "Academic"],
       student_award$num_awards[student_award$prog == "Academic"], pch = 16)
points(mean(student_award$math), pred[2], pch = 16, col = "blue", lwd = 4)
plot(student_award$math[student_award$prog == "Vocational"],
     student_award$num_award_hat[student_award$prog == "Vocational"],
     type = "l", lwd = 2, col = "red",
```

```r
     xlim = c(min(student_award$math), max(student_award$math)), ylim = c(0, 6),
     xlab = "Math score", ylab = "Award number", main = "Vocational")
points(student_award$math[student_award$prog == "Vocational"],
       student_award$num_awards[student_award$prog == "Vocational"], pch = 16)
points(mean(student_award$math), pred[3], pch = 16, col = "blue", lwd = 4)
```



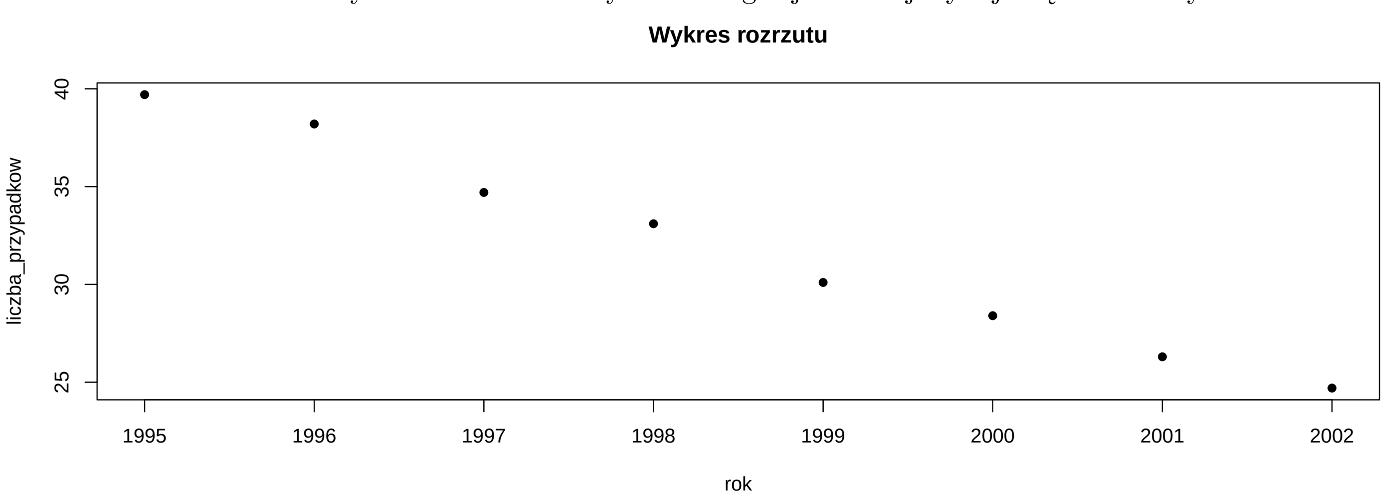

```r
par(mfrow = c(1, 1))
```

7.5 Zadania 7

Zadanie 1. Poniższa tabela przedstawia liczbę przypadków gruźlicy układu oddechowego w latach 1995-2002. Podano liczbę przypadków na 100.000 ludności. Zakładając liniową zależność między rokiem a liczbą przypadków, wykonaj kompleksową analizę regresji.

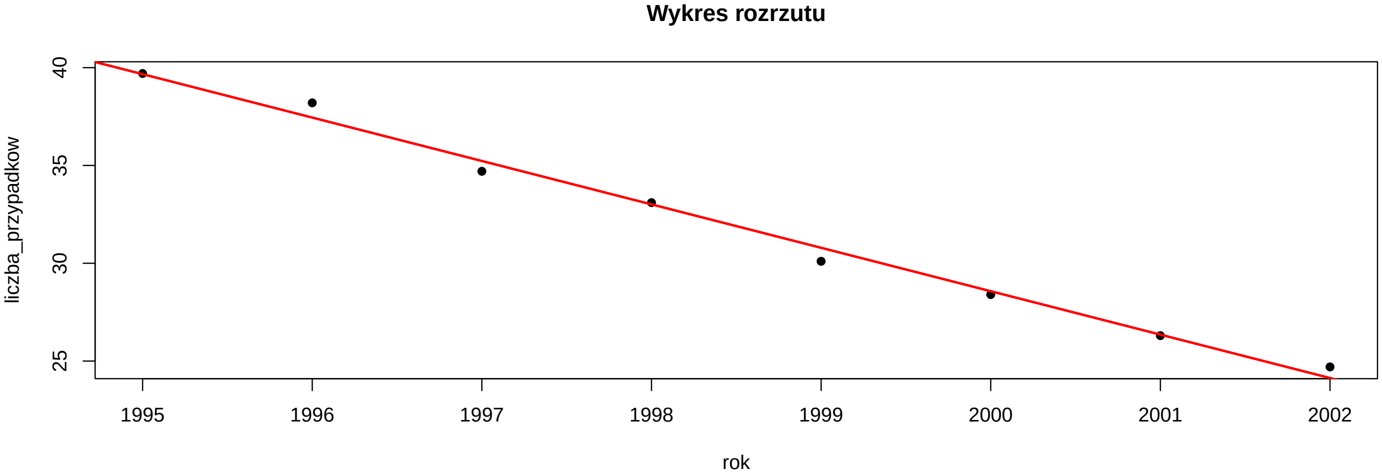
Dane								
rok	1995	1996	1997	1998	1999	2000	2001	2002
liczba przypadków	39.7	38.2	34.7	33.1	30.1	28.4	26.3	24.7

1. Przedstaw dane na wykresie rozrzutu. Czy model regresji liniowej wydaje się adekwatny?



2. Dopasuj model regresji liniowej do tych danych. Jakie są wartości estymatorów współczynników regresji i przedziały ufności? Narysuj uzyskaną prostą regresji na wykresie rozrzutu.

```
## (Intercept)         rok
## 4466.666667   -2.219048
```




```
## (Intercept)         rok
## 4466.666667   -2.219048
```

```
##                  2.5 %      97.5 %
## (Intercept) 4066.82158 4866.511749
## rok           -2.41912   -2.018975
```

3. Które współczynniki są istotne statystycznie w skonstruowanym modelu? Jakie jest dopasowanie modelu?

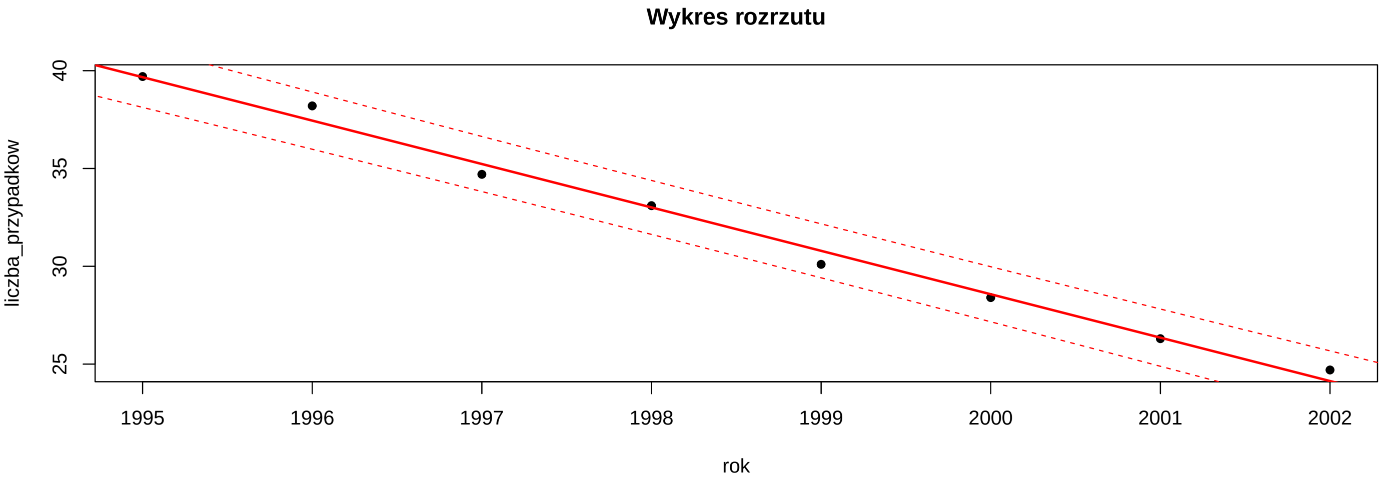
```
##
## Call:
## lm(formula = liczba_przypadkow ~ rok, data = data_set)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.69048 -0.26071 -0.00952  0.20952  0.75238
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4466.66667  163.40805   27.33 1.58e-07 ***
## rok           -2.21905    0.08177  -27.14 1.65e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5299 on 6 degrees of freedom
## Multiple R-squared:  0.9919, Adjusted R-squared:  0.9906
## F-statistic: 736.5 on 1 and 6 DF,  p-value: 1.654e-07
```

4. Oblicz wartości dopasowane przez model, a także reszty.

```
##        1        2        3        4        5        6        7        8
## 39.66667 37.44762 35.22857 33.00952 30.79048 28.57143 26.35238 24.13333
```

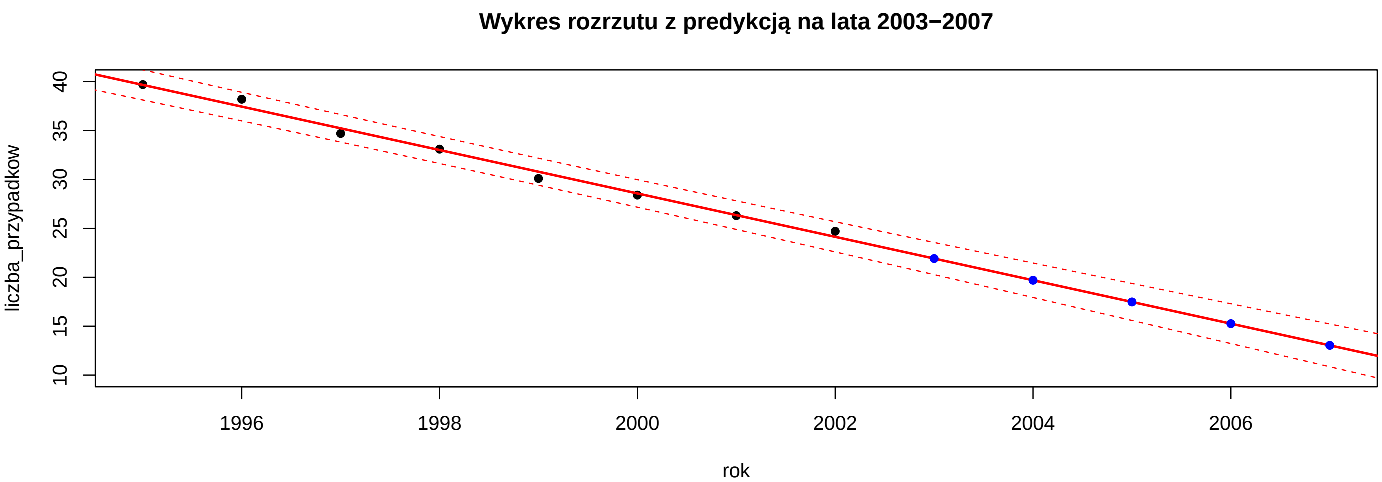
```
##           1           2           3           4           5           6
##  0.03333333  0.75238095 -0.52857143  0.09047619 -0.69047619 -0.17142857
##           7           8
## -0.05238095  0.56666667
```

5. Na wykresie rozrzutu przedstaw granice przedziału prognozy 95%.



6. Dokonaj predykcji liczby przypadków gruźlicy układu oddechowego w latach 2003-2007. Zilustruj wyniki na wykresie rozrzutu.

```
##        fit      lwr      upr
## 1 21.91429 20.27052 23.55805
## 2 19.69524 17.93392 21.45656
## 3 17.47619 15.58342 19.36896
## 4 15.25714 13.22171 17.29258
## 5 13.03810 10.85098 15.22521
```

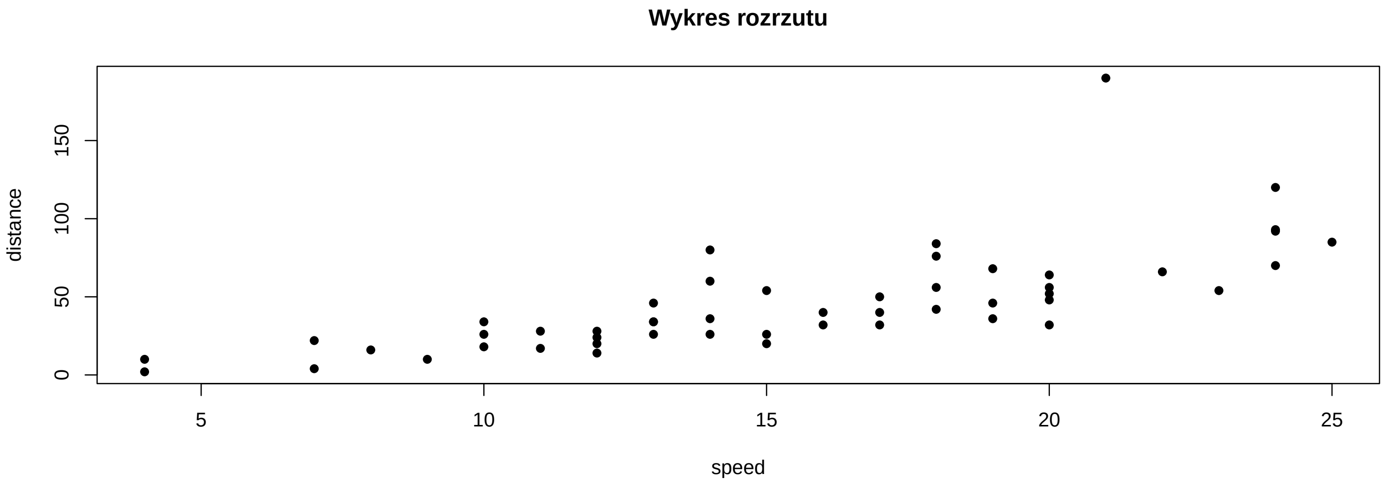


7. Czy miałoby sens usunięcie wyrazu wolnego z modelu? Jeśli tak, wykonaj powyższe polecenia dla modelu regresji liniowej bez wyrazu losowego.

Zadanie 2. Zbiór danych zawarty w pliku braking.RData zawiera informacje o długości drogi hamowania przy danej prędkości określonego modelu samochodu. W tym zbiorze danych występuje obserwacja odstająca. Zidentyfikuj ją za pomocą wykresu rozrzutu. Korzystając z modelu regresji liniowej, opisz związek między długością drogi hamowania a prędkością przy użyciu pełnych danych i danych bez obserwacji odstającej. Jakie są wyniki dla obu modeli? Który model jest lepszy? Dokładniej, wykonaj polecenia 2-7 Zadania 1 dla każdego modelu osobno. W punkcie 6 przeprowadź predykcję długości drogi hamowania dla prędkości 30, 31, ..., 50.

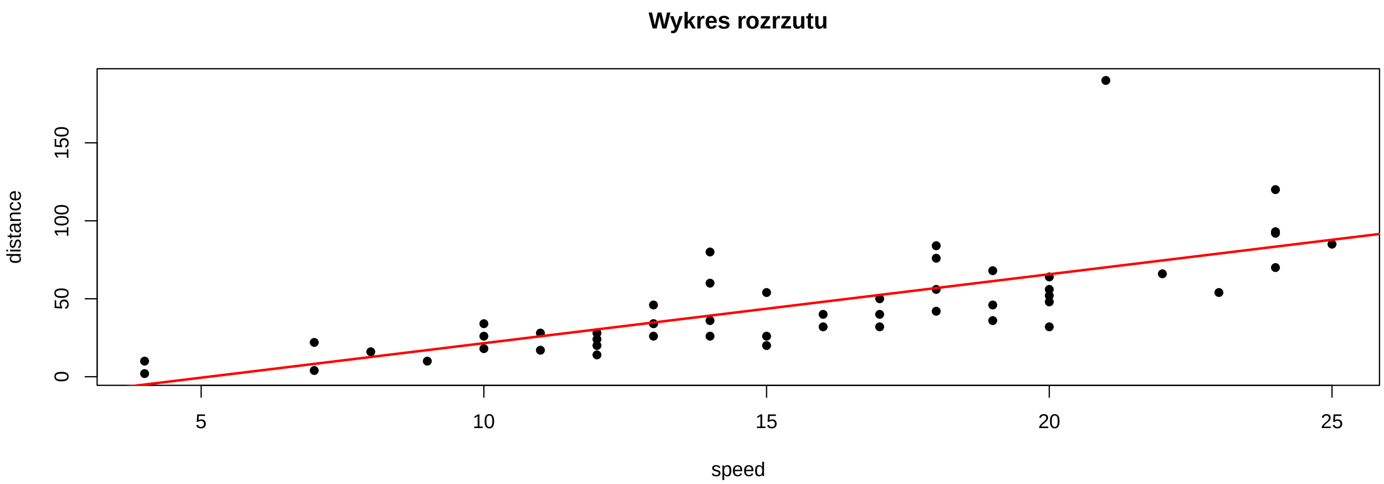
```
##   speed distance
## 1     4        2
## 2     4       10
## 3     7        4
## 4     7       22
```

```
## 5      8       16
## 6      9       10
```



Model dla pełnych danych

2.




```
## (Intercept)       speed
##  -22.726854    4.422338

##                   2.5 %    97.5 %
## (Intercept) -43.105778 -2.347930
## speed         3.177543  5.667134
```

3.

```
##
## Call:
## lm(formula = distance ~ speed, data = braking)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.720 -13.298  -3.186   7.814 119.858
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -22.7269    10.1409  -2.241   0.0296 *
```

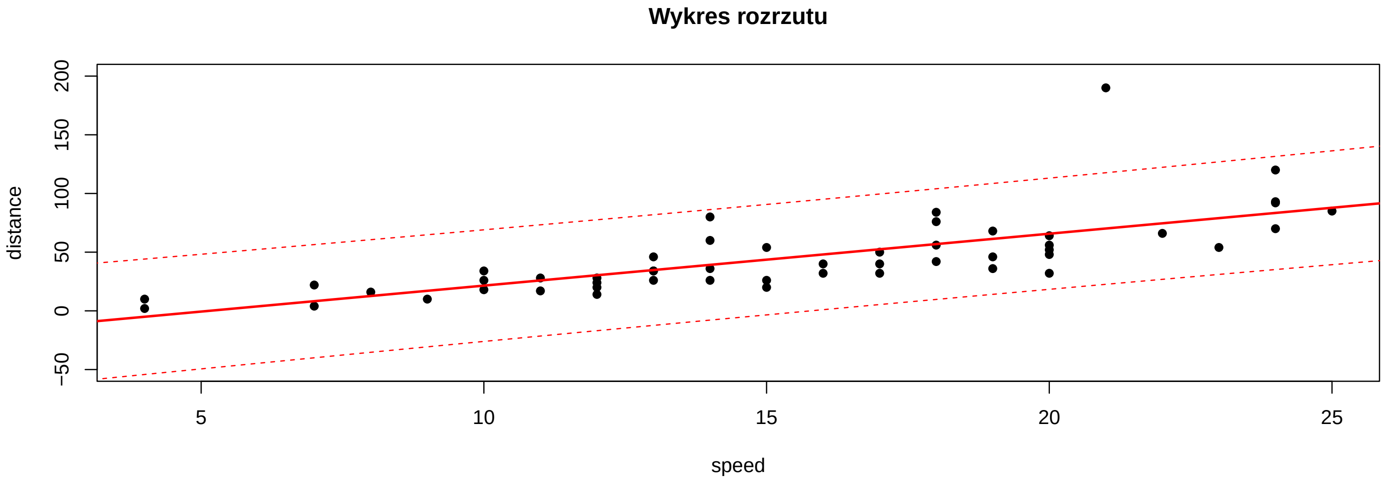
```
## speed         4.4223     0.6194   7.139 4.04e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.18 on 49 degrees of freedom
## Multiple R-squared:  0.5099, Adjusted R-squared:  0.4999
## F-statistic: 50.97 on 1 and 49 DF,  p-value: 4.037e-09
```

4.

```
##         1         2         3         4         5         6         7         8
## -5.037501 -5.037501  8.229514  8.229514 12.651852 17.074190 21.496528 21.496528
##         9        10        11        12        13        14        15        16
## 21.496528 25.918867 25.918867 30.341205 30.341205 30.341205 30.341205 34.763543
##        17        18        19        20        21        22        23        24
## 34.763543 34.763543 34.763543 39.185881 39.185881 39.185881 39.185881 43.608220
##        25        26        27        28        29        30        31        32
## 43.608220 43.608220 70.142249 48.030558 48.030558 52.452896 52.452896 52.452896
##        33        34        35        36        37        38        39        40
## 56.875234 56.875234 56.875234 56.875234 61.297573 61.297573 61.297573 65.719911
##        41        42        43        44        45        46        47        48
## 65.719911 65.719911 65.719911 65.719911 74.564587 78.986926 83.409264 83.409264
##        49        50        51
## 83.409264 83.409264 87.831602
```

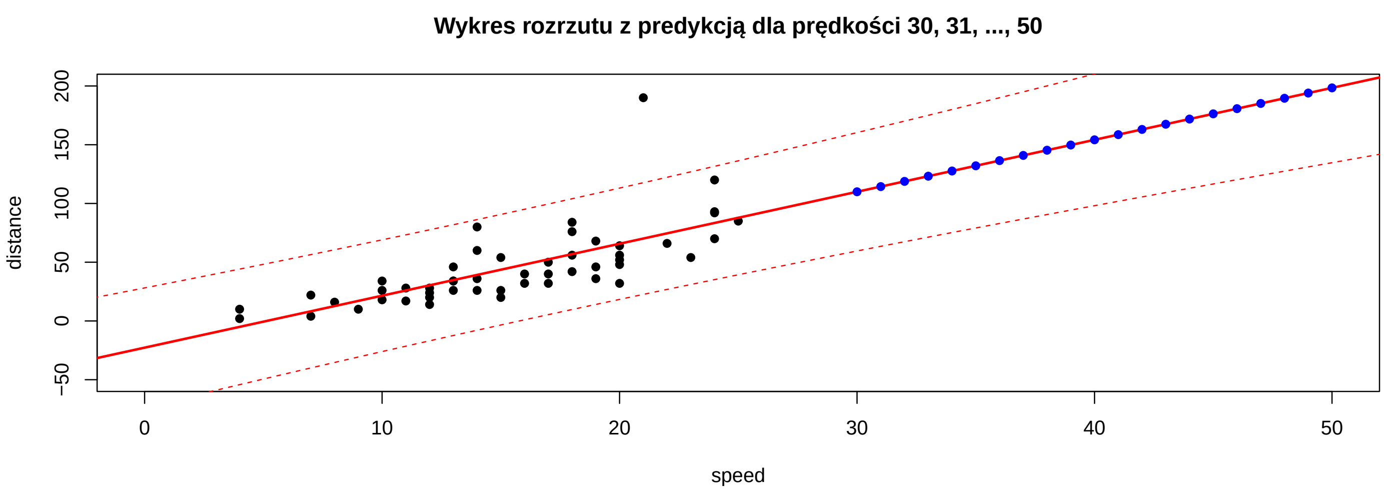
```
##           1           2           3           4           5           6
##   7.0375010  15.0375010  -4.2295137  13.7704863   3.3481480  -7.0741902
##           7           8           9          10          11          12
##  -3.4965285   4.5034715  12.5034715  -8.9188667   2.0811333 -16.3412050
##          13          14          15          16          17          18
## -10.3412050  -6.3412050  -2.3412050  -8.7635432  -0.7635432  -0.7635432
##          19          20          21          22          23          24
##  11.2364568 -13.1858815  -3.1858815  20.8141185  40.8141185 -23.6082197
##          25          26          27          28          29          30
## -17.6082197  10.3917803 119.8577508 -16.0305580  -8.0305580 -20.4528962
##          31          32          33          34          35          36
## -12.4528962  -2.4528962 -14.8752345  -0.8752345  19.1247655  27.1247655
##          37          38          39          40          41          42
## -25.2975727 -15.2975727   6.7024273 -33.7199110 -17.7199110 -13.7199110
##          43          44          45          46          47          48
##  -9.7199110  -1.7199110  -8.5645875 -24.9869257 -13.4092640   8.5907360
##          49          50          51
##   9.5907360  36.5907360  -2.8316022
```

5.

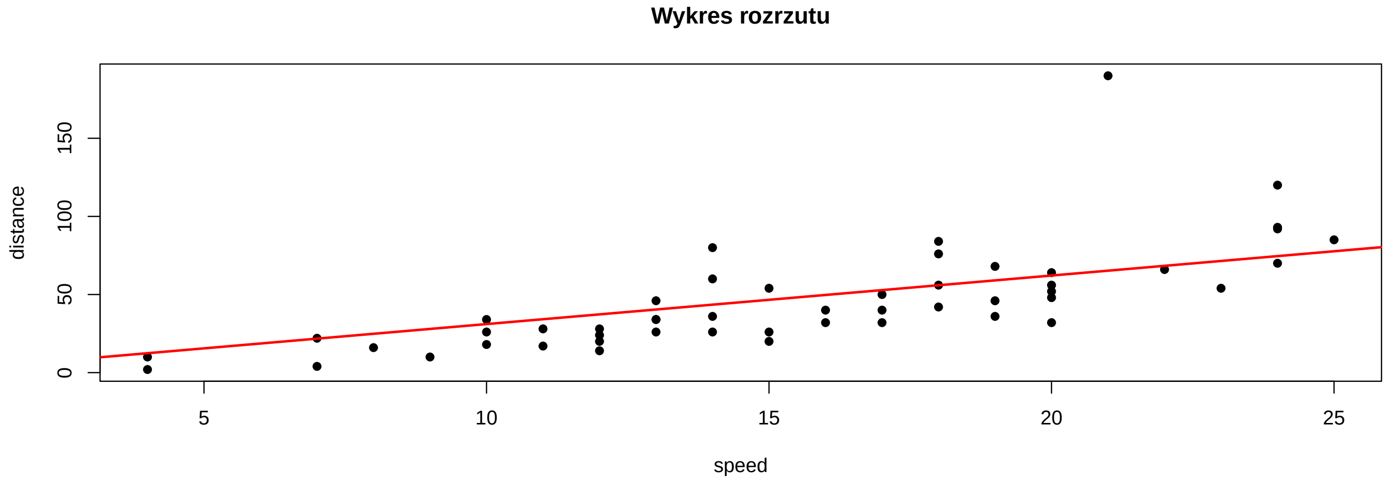


6.

```
##          fit       lwr      upr
## 1  109.9433  59.56096 160.3256
## 2  114.3656  63.52436 165.2069
## 3  118.7880  67.46167 170.1143
## 4  123.2103  71.37362 175.0470
## 5  127.6326  75.26095 180.0043
## 6  132.0550  79.12441 184.9856
## 7  136.4773  82.96475 189.9899
## 8  140.8997  86.78270 195.0166
## 9  145.3220  90.57902 200.0650
## 10 149.7443  94.35444 205.1342
## 11 154.1667  98.10968 210.2237
## 12 158.5890 101.84545 215.3326
## 13 163.0114 105.56245 220.4603
## 14 167.4337 109.26136 225.6060
## 15 171.8560 112.94285 230.7692
## 16 176.2784 116.60757 235.9492
## 17 180.7007 120.25614 241.1453
## 18 185.1230 123.88919 246.3569
## 19 189.5454 127.50730 251.5835
## 20 193.9677 131.11105 256.8244
## 21 198.3901 134.70099 262.0791
```



7.



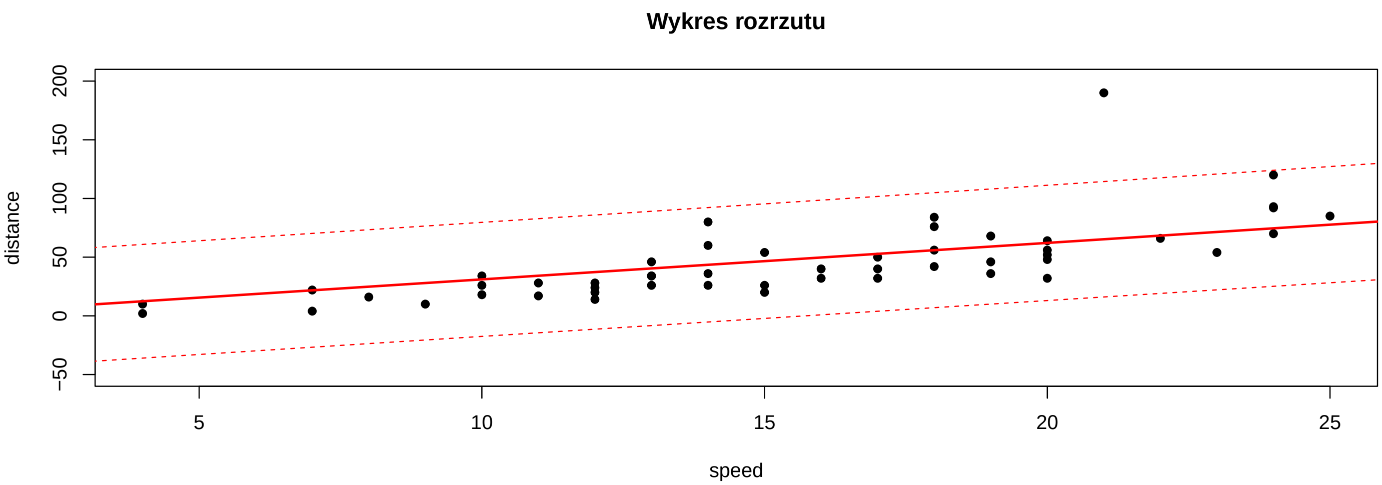

```
##    speed
## 3.107177
```

```
##          2.5 %   97.5 %
## speed 2.693185 3.521169
```

```
##
## Call:
## lm(formula = distance ~ speed - 1, data = braking)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.144 -15.786  -7.500   2.392 124.749
##
## Coefficients:
##       Estimate Std. Error t value Pr(>|t|)
## speed   3.1072     0.2061   15.07   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.1 on 50 degrees of freedom
## Multiple R-squared:  0.8197, Adjusted R-squared:  0.8161
## F-statistic: 227.3 on 1 and 50 DF,  p-value: < 2.2e-16
```

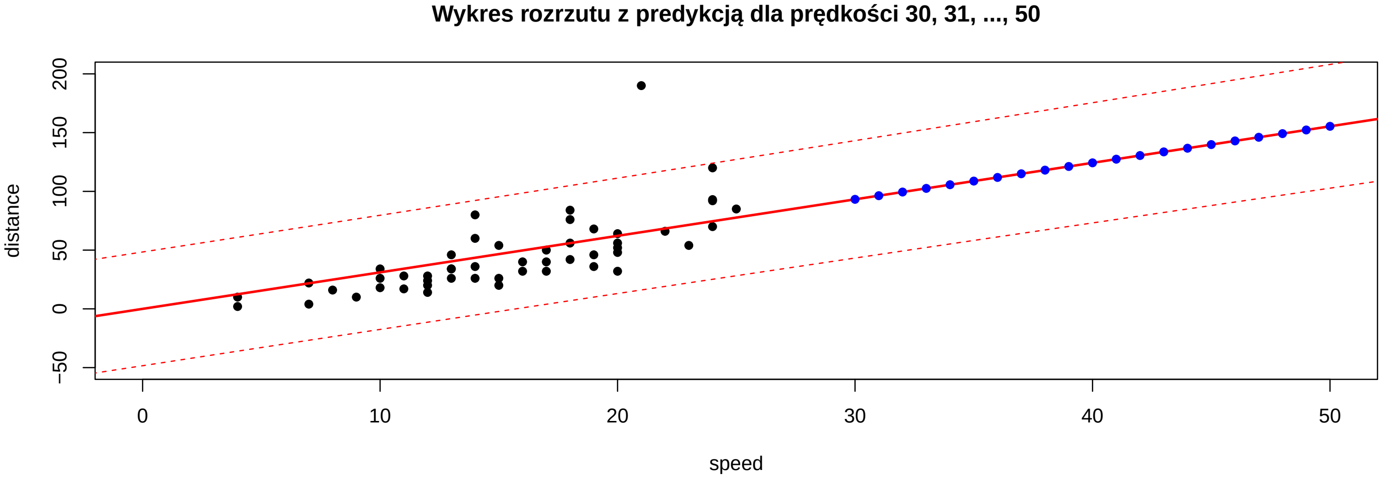
```
##        1        2        3        4        5        6        7        8
## 12.42871 12.42871 21.75024 21.75024 24.85741 27.96459 31.07177 31.07177
##        9       10       11       12       13       14       15       16
## 31.07177 34.17895 34.17895 37.28612 37.28612 37.28612 37.28612 40.39330
##       17       18       19       20       21       22       23       24
## 40.39330 40.39330 40.39330 43.50048 43.50048 43.50048 43.50048 46.60765
##       25       26       27       28       29       30       31       32
## 46.60765 46.60765 65.25071 49.71483 49.71483 52.82201 52.82201 52.82201
##       33       34       35       36       37       38       39       40
## 55.92918 55.92918 55.92918 55.92918 59.03636 59.03636 59.03636 62.14354
##       41       42       43       44       45       46       47       48
## 62.14354 62.14354 62.14354 62.14354 68.35789 71.46507 74.57224 74.57224
##       49       50       51
## 74.57224 74.57224 77.67942
```

```
##            1            2            3            4            5            6
## -10.42870729  -2.42870729 -17.75023776   0.24976224  -8.85741459 -17.96459141
##            7            8            9           10           11           12
## -13.07176823  -5.07176823   2.92823177 -17.17894506  -6.17894506 -23.28612188
##           13           14           15           16           17           18
## -17.28612188 -13.28612188  -9.28612188 -14.39329871  -6.39329871  -6.39329871
##           19           20           21           22           23           24
##   5.60670129 -17.50047553  -7.50047553  16.49952447  36.49952447 -26.60765235
##           25           26           27           28           29           30
## -20.60765235   7.39234765 124.74928671 -17.71482918  -9.71482918 -20.82200600
##           31           32           33           34           35           36
## -12.82200600  -2.82200600 -13.92918282   0.07081718  20.07081718  28.07081718
##           37           38           39           40           41           42
## -23.03635965 -13.03635965   8.96364035 -30.14353647 -14.14353647 -10.14353647
##           43           44           45           46           47           48
##  -6.14353647   1.85646353  -2.35789012 -17.46506694  -4.57224376  17.42775624
##           49           50           51
##  18.42775624  45.42775624   7.32057941
```



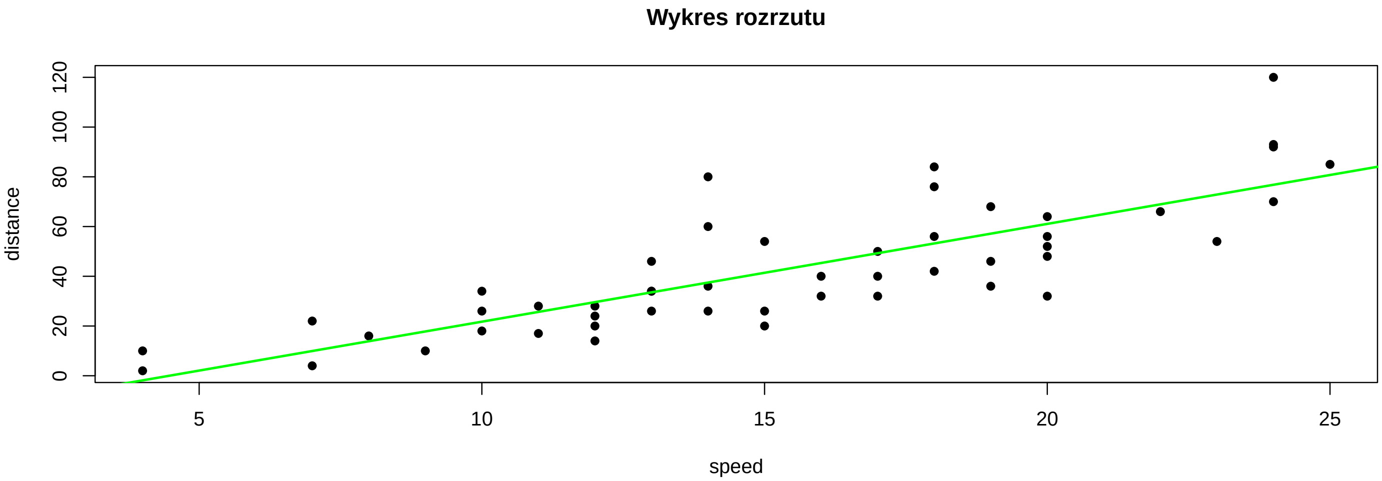

```
##          fit      lwr      upr
## 1   93.21530 43.24558 143.1850
## 2   96.32248 46.24826 146.3967
## 3   99.42966 49.24774 149.6116
## 4  102.53684 52.24404 152.8296
## 5  105.64401 55.23718 156.0508
## 6  108.75119 58.22719 159.2752
## 7  111.85837 61.21408 162.5026
## 8  114.96554 64.19789 165.7332
## 9  118.07272 67.17862 168.9668
## 10 121.17990 70.15631 172.2035
## 11 124.28707 73.13098 175.4432
## 12 127.39425 76.10265 178.6858
## 13 130.50143 79.07134 181.9315
## 14 133.60860 82.03708 185.1801
## 15 136.71578 84.99990 188.4317
## 16 139.82296 87.95981 191.6861
## 17 142.93013 90.91684 194.9434
## 18 146.03731 93.87102 198.2036
```

```
## 19 149.14449  96.82237 201.4666
## 20 152.25166  99.77092 204.7324
## 21 155.35884 102.71669 208.0010
```



Model dla zbioru danych bez obserwacji odstającej

2.




```
## (Intercept)       speed
##  -17.579095    3.932409
```

```
##                   2.5 %    97.5 %
## (Intercept) -31.167850 -3.990340
## speed         3.096964  4.767853
```

3.

```
##
## Call:
## lm(formula = distance ~ speed, data = braking_1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.069  -9.525  -2.272   9.215  43.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

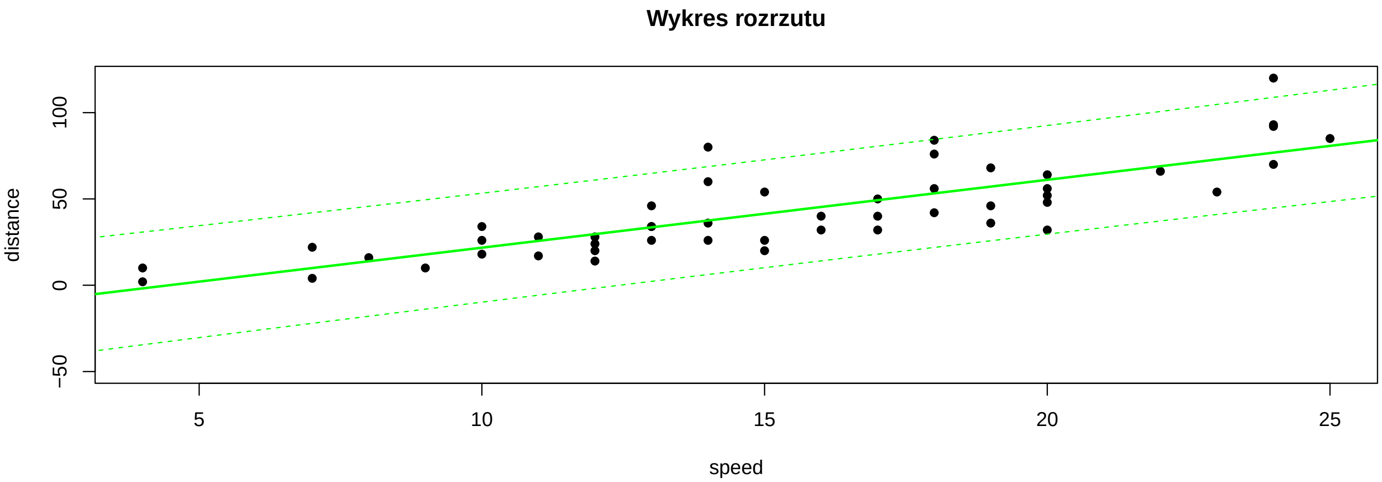
```
## (Intercept) -17.5791     6.7584  -2.601   0.0123 *
## speed         3.9324     0.4155   9.464 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.38 on 48 degrees of freedom
## Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

4.

```
##         1         2         3         4         5         6         7         8
## -1.849460 -1.849460  9.947766  9.947766 13.880175 17.812584 21.744993 21.744993
##         9        10        11        12        13        14        15        16
## 21.744993 25.677401 25.677401 29.609810 29.609810 29.609810 29.609810 33.542219
##        17        18        19        20        21        22        23        24
## 33.542219 33.542219 33.542219 37.474628 37.474628 37.474628 37.474628 41.407036
##        25        26        28        29        30        31        32        33
## 41.407036 41.407036 45.339445 45.339445 49.271854 49.271854 49.271854 53.204263
##        34        35        36        37        38        39        40        41
## 53.204263 53.204263 53.204263 57.136672 57.136672 57.136672 61.069080 61.069080
##        42        43        44        45        46        47        48        49
## 61.069080 61.069080 61.069080 68.933898 72.866307 76.798715 76.798715 76.798715
##        50        51
## 76.798715 80.731124
```

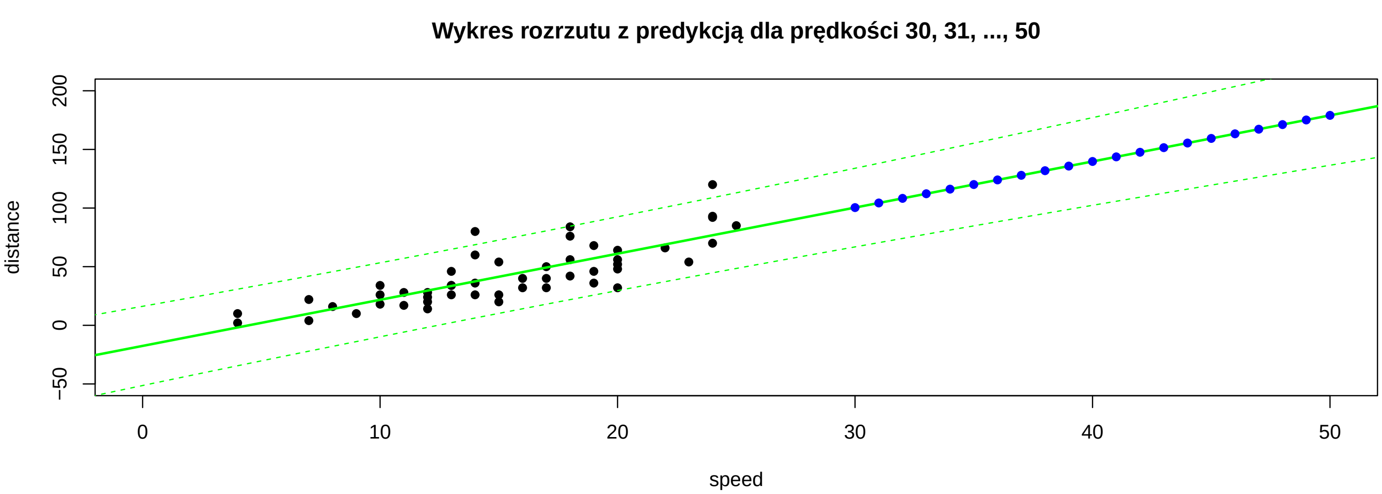
```
##          1          2          3          4          5          6          7
##   3.849460  11.849460  -5.947766  12.052234   2.119825  -7.812584  -3.744993
##          8          9         10         11         12         13         14
##   4.255007  12.255007  -8.677401   2.322599 -15.609810  -9.609810  -5.609810
##         15         16         17         18         19         20         21
##  -1.609810  -7.542219   0.457781   0.457781  12.457781 -11.474628  -1.474628
##         22         23         24         25         26         28         29
##  22.525372  42.525372 -21.407036 -15.407036  12.592964 -13.339445  -5.339445
##         30         31         32         33         34         35         36
## -17.271854  -9.271854   0.728146 -11.204263   2.795737  22.795737  30.795737
##         37         38         39         40         41         42         43
## -21.136672 -11.136672  10.863328 -29.069080 -13.069080  -9.069080  -5.069080
##         44         45         46         47         48         49         50
##   2.930920  -2.933898 -18.866307  -6.798715  15.201285  16.201285  43.201285
##         51
##   4.268876
```

5.

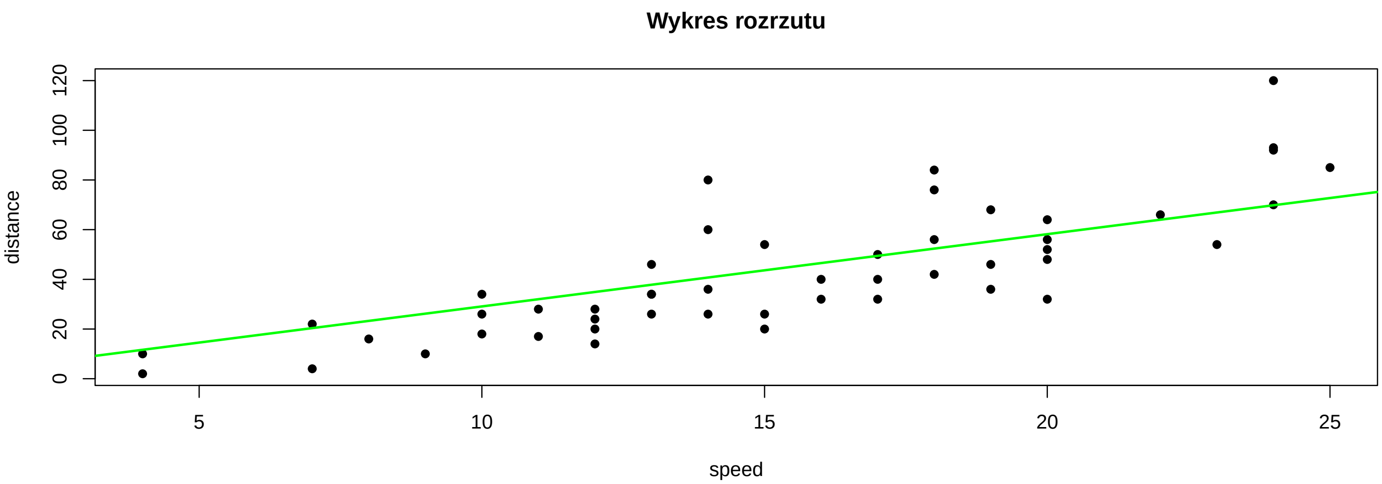


6.

```
##         fit       lwr      upr
## 1  100.3932  66.86529 133.9210
## 2  104.3256  70.48482 138.1663
## 3  108.2580  74.08678 142.4292
## 4  112.1904  77.67167 146.7091
## 5  116.1228  81.24002 151.0056
## 6  120.0552  84.79233 155.3181
## 7  123.9876  88.32911 159.6461
## 8  127.9200  91.85088 163.9892
## 9  131.8524  95.35814 168.3467
## 10 135.7848  98.85140 172.7183
## 11 139.7173 102.33114 177.1034
## 12 143.6497 105.79785 181.5015
## 13 147.5821 109.25201 185.9121
## 14 151.5145 112.69408 190.3349
## 15 155.4469 116.12452 194.7693
## 16 159.3793 119.54375 199.2148
## 17 163.3117 122.95222 203.6712
## 18 167.2441 126.35033 208.1379
## 19 171.1765 129.73848 212.6146
## 20 175.1089 133.11707 217.1008
## 21 179.0413 136.48646 221.5962
```

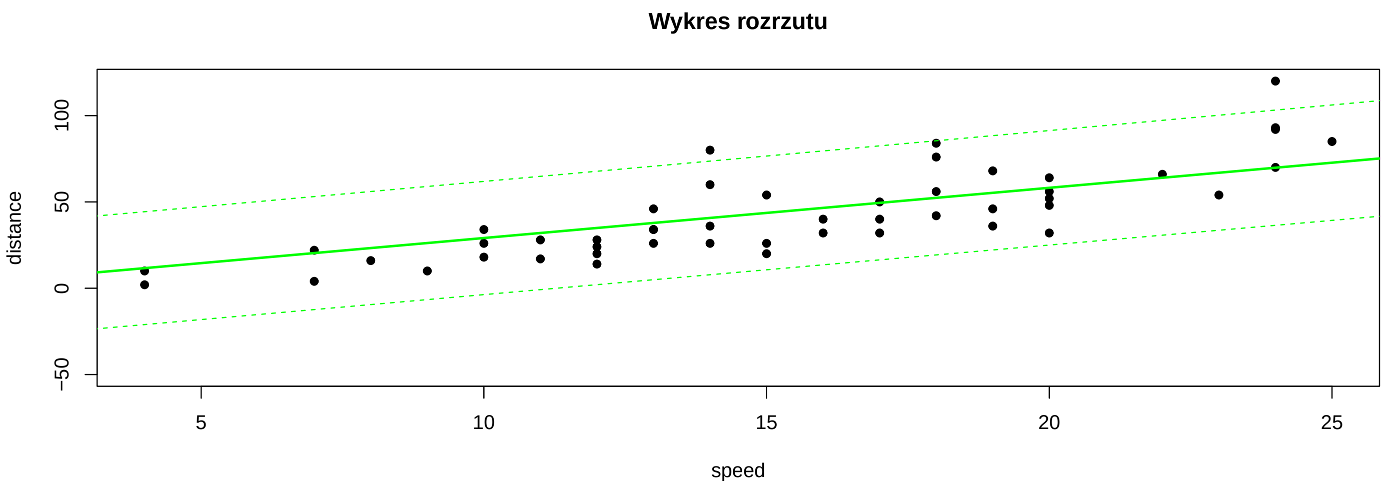


7.



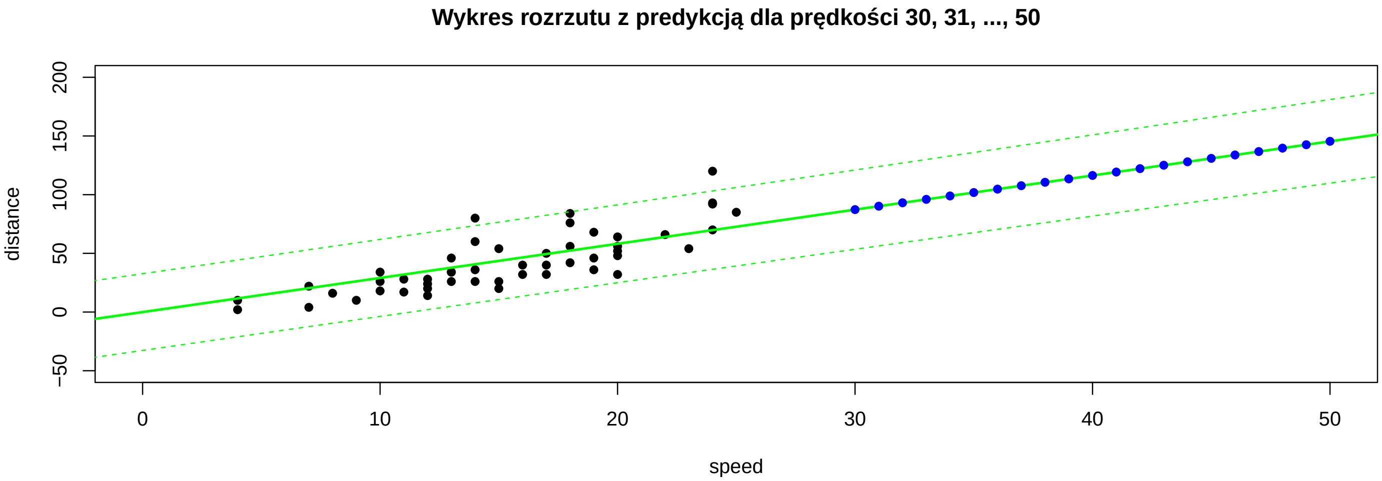

```
##    speed
## 2.909132

##          2.5 %   97.5 %
## speed 2.625041 3.193223

##
## Call:
## lm(formula = distance ~ speed - 1, data = braking_1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.183 -12.637  -5.455   4.590  50.181
##
## Coefficients:
##       Estimate Std. Error t value Pr(>|t|)
## speed   2.9091     0.1414   20.58   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.26 on 49 degrees of freedom
## Multiple R-squared:  0.8963, Adjusted R-squared:  0.8942
## F-statistic: 423.5 on 1 and 49 DF,  p-value: < 2.2e-16

##        1        2        3        4        5        6        7        8
## 11.63653 11.63653 20.36393 20.36393 23.27306 26.18219 29.09132 29.09132
##        9       10       11       12       13       14       15       16
## 29.09132 32.00045 32.00045 34.90959 34.90959 34.90959 34.90959 37.81872
##       17       18       19       20       21       22       23       24
## 37.81872 37.81872 37.81872 40.72785 40.72785 40.72785 40.72785 43.63698
##       25       26       28       29       30       31       32       33
## 43.63698 43.63698 46.54611 46.54611 49.45525 49.45525 49.45525 52.36438
##       34       35       36       37       38       39       40       41
## 52.36438 52.36438 52.36438 55.27351 55.27351 55.27351 58.18264 58.18264
##       42       43       44       45       46       47       48       49
## 58.18264 58.18264 58.18264 64.00091 66.91004 69.81917 69.81917 69.81917
##       50       51
## 69.81917 72.72830
```

```
##          1          2          3          4          5          6
##  -9.6365286  -1.6365286 -16.3639250   1.6360750  -7.2730572 -16.1821893
##          7          8          9         10         11         12
## -11.0913214  -3.0913214   4.9086786 -15.0004536  -4.0004536 -20.9095857
##         13         14         15         16         17         18
## -14.9095857 -10.9095857  -6.9095857 -11.8187179  -3.8187179  -3.8187179
##         19         20         21         22         23         24
##   8.1812821 -14.7278500  -4.7278500  19.2721500  39.2721500 -23.6369822
##         25         26         28         29         30         31
## -17.6369822  10.3630178 -14.5461143  -6.5461143 -17.4552464  -9.4552464
##         32         33         34         35         36         37
##   0.5447536 -10.3643786   3.6356214  23.6356214  31.6356214 -19.2735107
##         38         39         40         41         42         43
##  -9.2735107  12.7264893 -26.1826429 -10.1826429  -6.1826429  -2.1826429
##         44         45         46         47         48         49
##   5.8173571   1.9990928 -12.9100393   0.1808285  22.1808285  23.1808285
##         50         51
##  50.1808285  12.2716964
```




```
##          fit       lwr      upr
## 1   87.27396  53.50656 121.0414
## 2   90.18310  56.34287 124.0233
## 3   93.09223  59.17696 127.0075
## 4   96.00136  62.00884 129.9939
## 5   98.91049  64.83853 132.9825
## 6  101.81963  67.66604 135.9732
## 7  104.72876  70.49138 138.9661
## 8  107.63789  73.31458 141.9612
## 9  110.54702  76.13565 144.9584
## 10 113.45615  78.95460 147.9577
## 11 116.36529  81.77146 150.9591
## 12 119.27442  84.58623 153.9626
## 13 122.18355  87.39894 156.9682
## 14 125.09268  90.20960 159.9758
## 15 128.00181  93.01824 162.9854
## 16 130.91095  95.82486 165.9970
## 17 133.82008  98.62948 169.0107
## 18 136.72921 101.43213 172.0263
```

```
## 19 139.63834 104.23282 175.0439
## 20 142.54748 107.03157 178.0634
## 21 145.45661 109.82839 181.0848
```



Zadanie 3. Zbiór danych w pliku Automobile.csv zawiera dane charakteryzujące różne typy samochodów.

```
##   symboling normalized.losses        make fuel.type aspiration num.of.doors
## 1         3                NA alfa-romero       gas        std            2
## 2         3                NA alfa-romero       gas        std            2
## 3         1                NA alfa-romero       gas        std            2
## 4         2               164        audi       gas        std            4
## 5         2               164        audi       gas        std            4
## 6         2                NA        audi       gas        std            2
##    body.style drive.wheels engine.location wheel.base length width height
## 1 convertible          rwd           front       88.6  168.8  64.1   48.8
## 2 convertible          rwd           front       88.6  168.8  64.1   48.8
## 3   hatchback          rwd           front       94.5  171.2  65.5   52.4
## 4       sedan          fwd           front       99.8  176.6  66.2   54.3
## 5       sedan          4wd           front       99.4  176.6  66.4   54.3
## 6       sedan          fwd           front       99.8  177.3  66.3   53.1
##   curb.weight engine.type num.of.cylinders engine.size fuel.system bore stroke
## 1        2548        dohc             four         130        mpfi 3.47   2.68
## 2        2548        dohc             four         130        mpfi 3.47   2.68
## 3        2823        ohcv              six         152        mpfi 2.68   3.47
## 4        2337         ohc             four         109        mpfi 3.19   3.40
## 5        2824         ohc             five         136        mpfi 3.19   3.40
## 6        2507         ohc             five         136        mpfi 3.19   3.40
##   compression.ratio horsepower peak.rpm city.mpg highway.mpg price
## 1               9.0        111     5000       21          27 13495
## 2               9.0        111     5000       21          27 16500
## 3               9.0        154     5000       19          26 16500
## 4              10.0        102     5500       24          30 13950
## 5               8.0        115     5500       18          22 17450
## 6               8.5        110     5500       19          25 15250
```

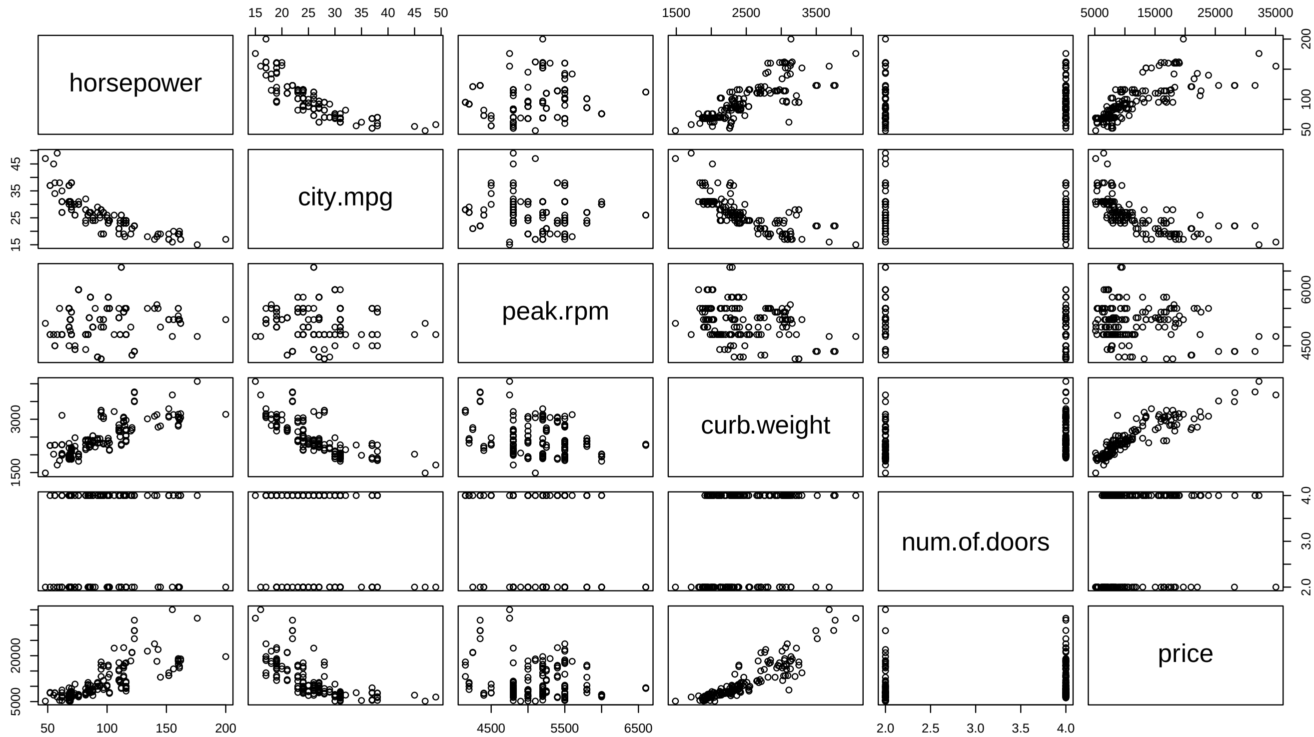
1. W tym zestawie danych występuja braki danych. Usuń wszystkie obserwacje, dla których nie mamy pełnych informacji o wszystkich zmiennych zawartych w zbiorze danych, używając funkcji `na.omit()`.

```
## wymiar nowych danych
```

```
## [1] 159  26
```

2. Interesuje nas zbudowanie modelu opisującego cenę samochodów w zależności od pewnych ich cech. Weźmy pod uwagę następujące zmienne: `horsepower`, `city.mpg`, `peak.rpm`, `curb.weight` i `num.of.doors` jako zmienne niezależne.

    - Dopasuj model regresji liniowej do tych danych.




```
##
## Call:
## lm(formula = price ~ horsepower + city.mpg + peak.rpm + curb.weight +
##     num.of.doors, data = auto_wna)
##
## Coefficients:
##  (Intercept)    horsepower      city.mpg      peak.rpm   curb.weight
##   -2.075e+04     2.743e+01     7.733e+01     4.847e-01     1.053e+01
## num.of.doors
##   -2.758e+02
```

    - Jakie są wartości estymatorów współczynników regresji i przedziały ufności? Które zmienne są stymulantami a które destymulantami?

```
##   (Intercept)    horsepower      city.mpg      peak.rpm   curb.weight
## -2.074964e+04  2.742792e+01  7.732533e+01  4.847128e-01  1.053105e+01
##  num.of.doors
## -2.757982e+02
```

```
##                       2.5 %        97.5 %
## (Intercept)  -3.063522e+04 -10864.058946
## horsepower   -2.484962e+00     57.340811
## city.mpg     -5.460472e+01    209.255385
## peak.rpm     -5.605008e-01      1.529926
## curb.weight   8.731667e+00     12.330436
## num.of.doors -7.313926e+02    179.796209
```

- Które współczynniki są statystycznie istotne w skontruowanym modelu? Jakie jest dopasowanie modelu?

```
##
## Call:
## lm(formula = price ~ horsepower + city.mpg + peak.rpm + curb.weight +
##     num.of.doors, data = auto_wna)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8235.7 -1413.0   -89.7   937.4  9759.4
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.075e+04  5.004e+03  -4.147 5.57e-05 ***
## horsepower    2.743e+01  1.514e+01   1.811    0.072 .
## city.mpg      7.733e+01  6.678e+01   1.158    0.249
## peak.rpm      4.847e-01  5.291e-01   0.916    0.361
## curb.weight   1.053e+01  9.108e-01  11.562  < 2e-16 ***
## num.of.doors -2.758e+02  2.306e+02  -1.196    0.234
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2597 on 153 degrees of freedom
## Multiple R-squared:  0.8109, Adjusted R-squared:  0.8048
## F-statistic: 131.2 on 5 and 153 DF,  p-value: < 2.2e-16
```

- Oblicz wartości dopasowane przez model oraz wartości reszt.

```
##         4         5         7         9        11        12        13        14
## 10077.611 15098.844 15249.651 18466.353 11280.669 10729.073 14240.554 14268.165
##        19        20        21        22        23        24        25        26
##  1791.835  5909.721  5726.711  5847.073  5383.121  8428.218  5789.850  6021.533
##        27        29        30        31        32        33        34        35
##  6021.533 11536.412 16171.344  4444.837  5244.631  5294.264  6441.563  6610.060
##        36        37        38        39        40        41        42        43
##  6627.140  6774.575  9504.115 10062.260  9668.630 10384.741 11543.572 10188.311
##        48        51        52        53        54        55        60        61
## 29256.003  5210.874  5393.510  5446.165  5315.811  5368.466 10456.347 10168.027
##        62        63        65        66        68        69        70        71
## 10456.347 10168.027 10325.993 13449.171 22347.106 24821.903 22688.081 25032.524
##        73        77        78        79        80        81        82        86
## 25296.608  6289.377  6099.232  6731.096  8607.246 11283.398  9985.406  9823.459
##        87        88        89        90        91        92        93        94
## 10244.701 11079.326 11079.326  5402.039  7254.692  5707.439  5366.464  6272.134
##        95        96        97        98        99       100       101       102
##  6054.964  6865.855  5713.989  6409.038  6655.234  9890.130  9658.447 18744.853
##       103       104       105       106       107       108       109       112
## 20861.595 18530.917 19417.779 21076.356 20133.890 16504.197 18597.259 17028.549
##       113       116       117       118       119       120       121       122
## 19176.467 17083.405 19176.467 19110.371  6289.377  8428.218  5789.850  6021.533
##       123       124       126       133       134       135       136       137
##  8148.806 11536.412 16011.320 13875.944 13713.997 14391.966 14377.453 16793.526
```

```
##       138       139       140       141       142       143       144       145 
## 16652.641  6952.124  7170.026  8433.753  7786.395  7757.106  9899.018  9695.244 
##       146       147       148       149       150       151       152       153 
## 11807.036  9004.096 11032.764  9986.506 13204.058  6336.440  6606.347  5791.474 
##       154       155       156       157       158       159       160       161 
##  8582.203  8378.212 17013.674  6628.622  6923.491  8451.542  8760.844  7384.128 
##       162       163       164       165       166       167       168       169 
##  6905.744  7095.303  8029.625  8398.212 10833.086 11201.673 12811.703 12769.579 
##       170       171       172       173       174       175       176       177 
## 12927.545 14275.519 14644.106 17392.711  9443.991 10767.381 10216.073 10216.073 
##       178       179       180       181       183       184       185       186 
## 10679.439 18522.082 18865.999 19465.659  9123.382  8925.756  8603.379  8405.753 
##       187       188       189       191       195       196       197       198 
##  9069.209  9476.020  9787.757  9078.471 16336.304 17621.093 16655.844 17782.666 
##       199       200       201       202       203       204       205 
## 18444.109 19623.587 16757.546 18682.970 17599.813 19270.000 17606.661
```

```
##           4           5           7           9          11          12 
##  3872.38860  2351.15555  2460.34881  5408.64732  5149.33069  6195.92704 
##          13          14          19          20          21          22 
##  6729.44647  6836.83500  3359.16525   385.27925   848.28880  -275.07295 
##          23          24          25          26          27          29 
##   993.87903  -471.21802   439.14971   670.46658  1587.46658 -2615.41226 
##          30          31          32          33          34          35 
## -3207.34381  2034.16262  1610.36919   104.73611    87.43730   518.94048 
##          36          37          38          39          40          41 
##   667.86006   520.42534 -1609.11460  -967.26032  -823.62974   -89.74124 
##          42          43          48          51          52          53 
##  1401.42812   156.68902  2993.99694   -15.87397   701.49018  1348.83492 
##          54          55          60          61          62          63 
##  1379.18922  2026.53396 -1611.34731 -1673.02725   138.65269    76.97275 
##          65          66          68          69          70          71 
##   919.00698  4830.82889  3204.89401  3426.09694  5487.91869  6567.47591 
##          73          77          78          79          80          81 
##  9759.39224  -900.37711    89.76754   -62.09554  -918.24589 -1324.39806 
##          82          86          87          88          89          90 
## -1486.40630 -2834.45885 -2055.70091 -1800.32640 -1800.32640    96.96127 
##          91          92          93          94          95          96 
##  -155.69190   941.56078  1482.53610  1076.86568  1244.03608   933.14513 
##          97          98          99         100         101         102 
##  1785.01141  1589.96202  1593.76616  -941.13028  -109.44715 -5245.85340 
##         103         104         105         106         107         108 
## -6462.59473 -5031.91727 -2218.77858 -1377.35637 -1734.89007 -4604.19683 
##         109         112         113         116         117         118 
## -5397.25921 -1448.54881 -2276.46704  -453.40466 -1226.46704  -960.37140 
##         119         120         121         122         123         124 
##  -717.37711  -471.21802   439.14971   670.46658  -539.80580 -2615.41226 
##         126         133         134         135         136         137 
##  6006.68036 -2025.94445 -1543.99700   648.03403  1132.54676  1356.47410 
##         138         139         140         141         142         143 
##  1967.35945 -1834.12417  -117.02643  -830.75259  -660.39477    17.89435 
##         144         145         146         147         148         149 
```

```
    ##    60.98200  -462.24445  -548.03567 -1541.09590  -834.76357 -1973.50592
    ##         150         151         152         153         154         155
    ## -1510.05754  -988.44041  -268.34691   696.52572 -1664.20289  -480.21208
    ##         156         157         158         159         160         161
    ## -8235.67421   309.37827   274.50883  -553.54226  -972.84358   353.87195
    ##         162         163         164         165         166         167
    ##  1452.25582  2162.69690    28.37473  -160.21207 -1535.08600 -1663.67279
    ##         168         169         170         171         172         173
    ## -4362.70319 -3130.57898 -2938.54475 -3076.51933 -3095.10613   276.28947
    ##         174         175         176         177         178         179
    ##  -495.99066   -69.38117  -228.07252   681.92748   568.56122 -1964.08195
    ##         180         181         183         184         185         186
    ## -2867.99868 -3775.65894 -1348.38201  -950.75627  -608.37882  -210.75307
    ##         187         188         189         191         195         196
    ##  -574.20931    18.98040   207.24268   901.52930 -3396.30442 -4206.09269
    ##         197         198         199         200         201         202
    ##  -670.84394 -1267.66643   -24.10880  -673.58655    87.45352   362.02963
    ##         203         204         205
    ##  3885.18734  3199.99996  5018.33919
```

3. Spróbuj zredukować model korzystając z regresji krokowej ("backward", "forward", AIC, BIC).

```
## Start:  AIC=2506.06
## price ~ horsepower + city.mpg + peak.rpm + curb.weight + num.of.doors
##
##                Df Sum of Sq        RSS    AIC
## - peak.rpm      1   5661937 1037719493 2504.9
## - city.mpg      1   9044038 1041101594 2505.4
## - num.of.doors  1   9647889 1041705445 2505.5
## <none>                      1032057556 2506.1
## - horsepower    1  22134795 1054192350 2507.4
## - curb.weight   1 901782660 1933840216 2603.9
##
## Step:  AIC=2504.93
## price ~ horsepower + city.mpg + curb.weight + num.of.doors
##
##                Df  Sum of Sq        RSS    AIC
## - city.mpg      1    6994707 1044714200 2504.0
## - num.of.doors  1    9518068 1047237561 2504.4
## <none>                       1037719493 2504.9
## - horsepower    1   32461892 1070181386 2507.8
## - curb.weight   1 1136974423 2174693916 2620.6
##
## Step:  AIC=2504
## price ~ horsepower + curb.weight + num.of.doors
##
##                Df  Sum of Sq        RSS    AIC
## - num.of.doors  1   12661847 1057376047 2503.9
## <none>                       1044714200 2504.0
## - horsepower    1   26482698 1071196898 2506.0
## - curb.weight   1 1155965636 2200679836 2620.5
##
```

```
## Step:  AIC=2503.91
## price ~ horsepower + curb.weight
##
##               Df  Sum of Sq        RSS    AIC
## <none>                      1057376047 2503.9
## - horsepower   1   42071205 1099447251 2508.1
## - curb.weight  1 1249455315 2306831362 2625.9
```

```
##
## Call:
## lm(formula = price ~ horsepower + curb.weight, data = auto_wna)
##
## Coefficients:
## (Intercept)   horsepower  curb.weight
##  -14608.000       27.404        9.519
```

```
## Start:  AIC=2524.47
## price ~ horsepower + city.mpg + peak.rpm + curb.weight + num.of.doors
##
##                Df Sum of Sq        RSS    AIC
## - peak.rpm      1   5661937 1037719493 2520.3
## - city.mpg      1   9044038 1041101594 2520.8
## - num.of.doors  1   9647889 1041705445 2520.9
## - horsepower    1  22134795 1054192350 2522.8
## <none>                      1032057556 2524.5
## - curb.weight   1 901782660 1933840216 2619.2
##
## Step:  AIC=2520.28
## price ~ horsepower + city.mpg + curb.weight + num.of.doors
##
##                Df  Sum of Sq        RSS    AIC
## - city.mpg      1    6994707 1044714200 2516.3
## - num.of.doors  1    9518068 1047237561 2516.7
## - horsepower    1   32461892 1070181386 2520.1
## <none>                       1037719493 2520.3
## - curb.weight   1 1136974423 2174693916 2632.8
##
## Step:  AIC=2516.27
## price ~ horsepower + curb.weight + num.of.doors
##
##                Df  Sum of Sq        RSS    AIC
## - num.of.doors  1   12661847 1057376047 2513.1
## - horsepower    1   26482698 1071196898 2515.2
## <none>                       1044714200 2516.3
## - curb.weight   1 1155965636 2200679836 2629.7
##
## Step:  AIC=2513.12
## price ~ horsepower + curb.weight
##
##               Df  Sum of Sq        RSS    AIC
## <none>                      1057376047 2513.1
## - horsepower   1   42071205 1099447251 2514.3
```

```
## - curb.weight  1 1249455315 2306831362 2632.1
##
## Call:
## lm(formula = price ~ horsepower + curb.weight, data = auto_wna)
##
## Coefficients:
## (Intercept)   horsepower  curb.weight
##  -14608.000       27.404        9.519
## Start:  AIC=2760.9
## price ~ 1
##
##                Df  Sum of Sq        RSS    AIC
## + curb.weight   1 4359325314 1099447251 2508.1
## + horsepower    1 3151941203 2306831362 2625.9
## + city.mpg      1 2616073039 2842699526 2659.2
## + peak.rpm      1  161334765 5297437800 2758.1
## + num.of.doors  1  143528709 5315243857 2758.7
## <none>                       5458772565 2760.9
##
## Step:  AIC=2508.12
## price ~ curb.weight
##
##                Df Sum of Sq        RSS    AIC
## + horsepower    1  42071205 1057376047 2503.9
## + num.of.doors  1  28250353 1071196898 2506.0
## + peak.rpm      1  21371766 1078075485 2507.0
## <none>                      1099447251 2508.1
## + city.mpg      1   1628352 1097818899 2509.9
##
## Step:  AIC=2503.91
## price ~ curb.weight + horsepower
##
##                Df Sum of Sq        RSS    AIC
## <none>                      1057376047 2503.9
## + num.of.doors  1  12661847 1044714200 2504.0
## + city.mpg      1  10138486 1047237561 2504.4
## + peak.rpm      1   3133537 1054242509 2505.4
##
## Call:
## lm(formula = price ~ curb.weight + horsepower, data = auto_wna)
##
## Coefficients:
## (Intercept)  curb.weight   horsepower
##  -14608.000        9.519       27.404
## Start:  AIC=2763.97
## price ~ 1
##
##                Df  Sum of Sq        RSS    AIC
## + curb.weight   1 4359325314 1099447251 2514.3
```

```
## + horsepower    1 3151941203 2306831362 2632.1
## + city.mpg      1 2616073039 2842699526 2665.3
## <none>                       5458772565 2764.0
## + peak.rpm      1  161334765 5297437800 2764.3
## + num.of.doors  1  143528709 5315243857 2764.8
##
## Step:  AIC=2514.26
## price ~ curb.weight
##
##                Df Sum of Sq        RSS    AIC
## + horsepower    1  42071205 1057376047 2513.1
## <none>                      1099447251 2514.3
## + num.of.doors  1  28250353 1071196898 2515.2
## + peak.rpm      1  21371766 1078075485 2516.2
## + city.mpg      1   1628352 1097818899 2519.1
##
## Step:  AIC=2513.12
## price ~ curb.weight + horsepower
##
##                Df Sum of Sq        RSS    AIC
## <none>                      1057376047 2513.1
## + num.of.doors  1  12661847 1044714200 2516.3
## + city.mpg      1  10138486 1047237561 2516.7
## + peak.rpm      1   3133537 1054242509 2517.7
```

```
##
## Call:
## lm(formula = price ~ curb.weight + horsepower, data = auto_wna)
##
## Coefficients:
## (Intercept)  curb.weight   horsepower
##  -14608.000        9.519       27.404
```

4. Dokonaj redukcji modelu metodą eliminacji wstecznej, tak aby w kolejnych krokach z pełnego modelu stopniowo usuwać najmniej istotną zmienną niezależną, aż otrzymamy model ze wszystkimi istotnymi zmiennymi niezależnymi. Jakie było zachowanie odpowiedniego współczynnika determinacji w kolejnych modelach?

```
##                  Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept)  -17339.01129 3341.9262712 -5.188328 6.597139e-07
## horsepower       31.64419   14.4173920  2.194862 2.967071e-02
## city.mpg         67.03355   65.7941206  1.018838 3.098778e-01
## curb.weight      10.09671    0.7772917 12.989598 1.600643e-26
## num.of.doors   -273.92557  230.4824166 -1.188488 2.364708e-01
```

```
## [1] 0.8049611
```

```
##                   Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept)  -14148.487819 1167.1823639 -12.121917 3.283606e-24
## horsepower       22.757081   11.4806990   1.982203 4.922435e-02
## curb.weight       9.917956    0.7573256  13.096027 7.371627e-27
## num.of.doors   -311.804092  227.4920297  -1.370615 1.724765e-01
```

```
## [1] 0.8049132
```

```
##                 Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept) -14607.99973 1121.1401301 -13.02959 1.000781e-26
## horsepower      27.40398   10.9995177   2.49138 1.377104e-02
## curb.weight      9.51894    0.7011011  13.57713 3.237394e-28
```

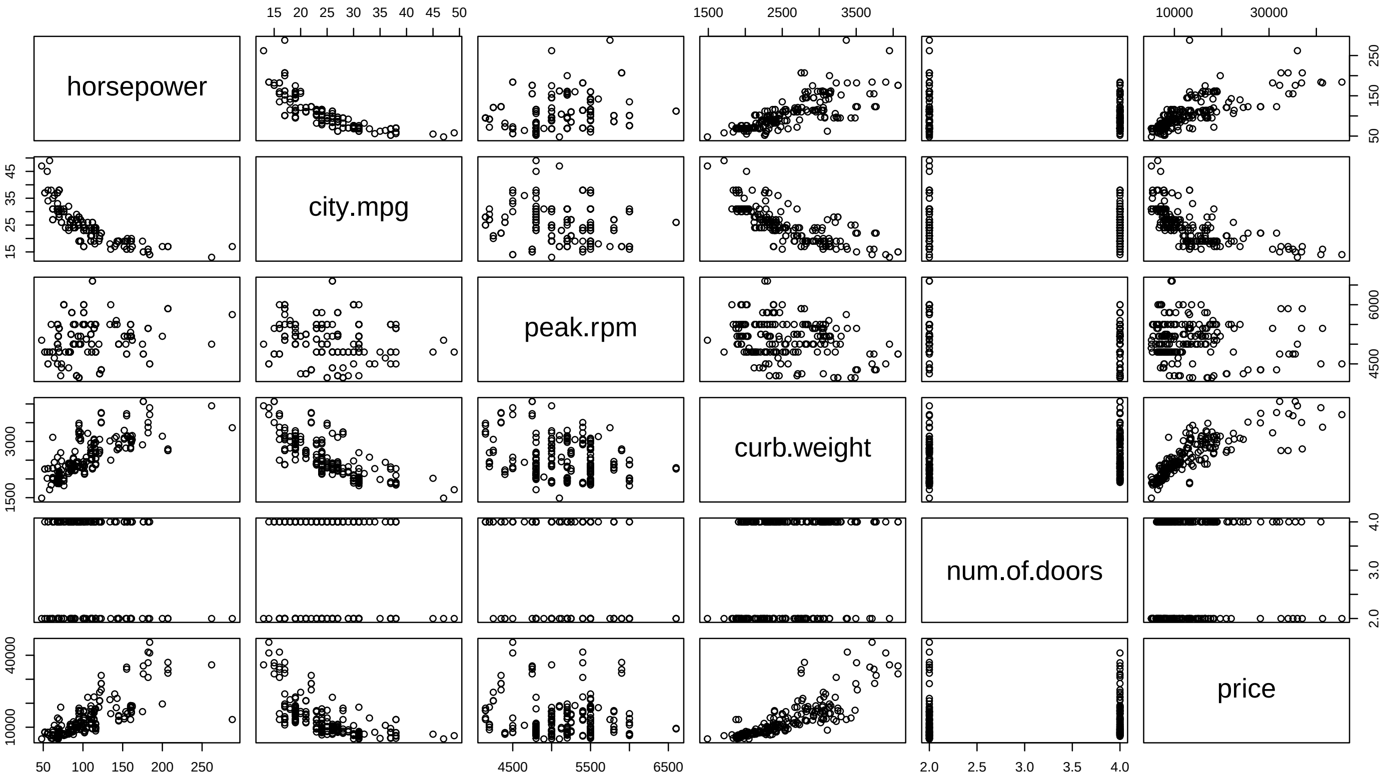
```
## [1] 0.8038145
```

5. Zamiast usuwać obserwacje z brakującymi danymi, jak to zrobiliśmy w punkcie 1, uzupełnij je za pomocą średniej i mediany sąsiednich wartości dla zmiennych ilościowych i porządkowych, odpowiednio. Aby to zrobić, użyj funkcji `impute()` dostępnej w pakiecie `Hmisc`. W przypadku takich danych postępuj zgodnie z instrukcjami w punktach 2-4.

```
##    horsepower      city.mpg        peak.rpm     curb.weight    num.of.doors
##  Min.   : 48.0   Min.   :13.00   Min.   :4150   Min.   :1488   Min.   :2.000
##  1st Qu.: 70.0   1st Qu.:19.00   1st Qu.:4800   1st Qu.:2145   1st Qu.:2.000
##  Median : 95.0   Median :24.00   Median :5200   Median :2414   Median :4.000
##  Mean   :104.3   Mean   :25.22   Mean   :5125   Mean   :2556   Mean   :3.123
##  3rd Qu.:116.0   3rd Qu.:30.00   3rd Qu.:5500   3rd Qu.:2935   3rd Qu.:4.000
##  Max.   :288.0   Max.   :49.00   Max.   :6600   Max.   :4066   Max.   :4.000
##  NA's   :2                       NA's   :2                     NA's   :2
##      price
##  Min.   : 5118
##  1st Qu.: 7775
##  Median :10295
##  Mean   :13207
##  3rd Qu.:16500
##  Max.   :45400
##  NA's   :4
```

```
##    horsepower      city.mpg        peak.rpm     curb.weight    num.of.doors
##  Min.   : 48.0   Min.   :13.00   Min.   :4150   Min.   :1488   Min.   :2.000
##  1st Qu.: 70.0   1st Qu.:19.00   1st Qu.:4800   1st Qu.:2145   1st Qu.:2.000
##  Median : 95.0   Median :24.00   Median :5200   Median :2414   Median :4.000
##  Mean   :104.3   Mean   :25.22   Mean   :5125   Mean   :2556   Mean   :3.132
##  3rd Qu.:116.0   3rd Qu.:30.00   3rd Qu.:5500   3rd Qu.:2935   3rd Qu.:4.000
##  Max.   :288.0   Max.   :49.00   Max.   :6600   Max.   :4066   Max.   :4.000
##      price
##  Min.   : 5118
##  1st Qu.: 7788
##  Median :10595
##  Mean   :13207
##  3rd Qu.:16500
##  Max.   :45400
```

```
## 2.
```




```
##
## Call:
## lm(formula = price ~ horsepower + city.mpg + peak.rpm + curb.weight +
##     num.of.doors, data = auto_sel)
##
## Coefficients:
##  (Intercept)    horsepower      city.mpg      peak.rpm   curb.weight
##   -2.593e+04     6.722e+01     1.413e+02     6.572e-01     1.017e+01
## num.of.doors
##   -2.525e+02
```

```
##   (Intercept)    horsepower      city.mpg      peak.rpm   curb.weight
## -2.593379e+04  6.721715e+01  1.413170e+02  6.572019e-01  1.017053e+01
##  num.of.doors
## -2.524809e+02
```

```
##                      2.5 %        97.5 %
## (Intercept)  -4.024060e+04 -11626.968553
## horsepower    3.717412e+01     97.260183
## city.mpg     -2.781769e+01    310.451628
## peak.rpm     -8.865075e-01      2.200911
## curb.weight   7.723371e+00     12.617692
## num.of.doors -9.046288e+02    399.666961
```

```
##
## Call:
## lm(formula = price ~ horsepower + city.mpg + peak.rpm + curb.weight +
##     num.of.doors, data = auto_sel)
##
## Residuals:
##    Min     1Q Median     3Q    Max
```

```
## -20128  -2083   -138   1379  16751
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.593e+04  7.255e+03  -3.575  0.00044 ***
## horsepower    6.722e+01  1.524e+01   4.412 1.68e-05 ***
## city.mpg      1.413e+02  8.577e+01   1.648  0.10101
## peak.rpm      6.572e-01  7.828e-01   0.840  0.40219
## curb.weight   1.017e+01  1.241e+00   8.196 3.00e-14 ***
## num.of.doors -2.525e+02  3.307e+02  -0.763  0.44610
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4182 on 199 degrees of freedom
## Multiple R-squared:  0.7245, Adjusted R-squared:  0.7176
## F-statistic: 104.7 on 5 and 199 DF,  p-value: < 2.2e-16
```

```
##         1         2         3         4         5         6         7         8
## 13190.535 13190.535 18595.135 10687.189 15666.159 12752.293 15674.800 16793.559
##         9        10        11        12        13        14        15        16
## 19869.950 21242.310 11770.668 11265.706 15017.431 15071.849 17879.986 23950.589
##        17        18        19        20        21        22        23        24
## 25981.131 26606.168  1915.053  6244.963  6095.970  6055.273  5207.372  9066.510
##        25        26        27        28        29        30        31        32
##  5627.928  5851.680  5851.680  9202.291 11431.313 17868.134  4961.068  5493.989
##        33        34        35        36        37        38        39        40
##  5262.202  6583.307  6746.035  6790.282  6932.669  9710.564 10249.602  9897.198
##        41        42        43        44        45        46        47        48
## 10588.794 12118.960 10751.530  8613.936  6244.963  6095.970 14094.645 31481.352
##        49        50        51        52        53        54        55        56
## 31481.352 36468.874  4879.841  5122.863  5173.716  5075.575  5126.428 10901.649
##        57        58        59        60        61        62        63        64
## 10901.649 10952.501 14266.179 10293.020 10042.321 10293.020 10042.321 10348.195
##        65        66        67        68        69        70        71        72
## 10194.879 14248.699 12497.433 23041.219 25431.293 23342.770 25634.704 26895.516
##        73        74        75        76        77        78        79        80
## 26841.098 30025.164 28648.577 20891.531  6482.436  5898.968  6509.199  9239.409
##        81        82        83        84        85        86        87        88
## 12327.501  9972.292 18091.886 18986.893 19037.745  9843.640 10250.461 12158.167
##        89        90        91        92        93        94        95        96
## 12158.167  5209.645  7285.990  5504.590  5203.039  6077.705  5840.218  6623.349
##        97        98        99       100       101       102       103       104
##  5538.667  6209.922  6419.938 10445.677 10221.925 20570.930 22615.206 20497.595
##       105       106       107       108       109       110       111       112
## 21652.170 24749.818 22343.766 16262.390 18641.371 18398.202 20587.154 16687.335
##       113       114       115       116       117       118       119       120
## 19200.750 18823.146 21146.533 16821.769 19200.750 20658.924  6482.436  9066.510
##       121       122       123       124       125       126       127       128
##  5627.928  5851.680  7906.127 11431.313 17939.328 17726.673 21785.066 21785.066
##       129       130       131       132       133       134       135       136
## 22232.570 33335.099 12912.585 12207.254 14406.377 14277.725 14904.733 14918.468
##       137       138       139       140       141       142       143       144
```

```
## 19174.481 19066.170  6649.940  6595.560  7816.024  8060.598  7690.124 10265.437
##       145       146       147       148       149       150       151       152
##  9370.990 12591.604  8970.057 11293.731  9585.642 13874.161  6017.883  6011.994
##       153       154       155       156       157       158       159       160
##  5252.769  7947.960  7484.397 15824.233  6320.445  6605.219  7720.595  8285.863
##       161       162       163       164       165       166       167       168
##  7583.197  6454.802  6637.872  7579.096  7935.065 12137.600 12493.569 13737.767
##       169       170       171       172       173       174       175       176
## 13697.085 13849.643 15151.471 15507.440 18161.948  9755.364 10382.977 10367.737
##       177       178       179       180       181       182       183       184
## 10367.737 10815.240 20894.504 21160.008 21629.888 21691.982  8435.412  9007.282
##       185       186       187       188       189       190       191       192
##  7960.962  8532.831  9173.575  9398.655 10459.079  9541.391  9205.763 13813.593
##       193       194       195       196       197       198       199       200
## 11477.725 12186.006 17134.813 18375.618 17510.052 18598.299 20668.854 21807.954
##       201       202       203       204       205
## 17541.634 20989.177 18855.344 19728.717 18095.125

##             1            2            3            4            5            6
##     304.46471   3309.46471  -2095.13493   3262.81115   1783.84136   2497.70662
##             7            8            9           10           11           12
##    2035.19953   2126.44112   4005.05039  -8035.18068   4659.33202   5659.29390
##            13           14           15           16           17           18
##    5952.56858   6033.15126   6685.01422   6809.41069  15333.86915  10273.83162
##            19           20           21           22           23           24
##    3235.94685     50.03666    479.02995   -483.27332   1169.62848  -1109.50973
##            25           26           27           28           29           30
##     601.07204    840.32035   1757.32035   -644.29131  -2510.31292  -4904.13419
##            31           32           33           34           35           36
##    1517.93247   1361.01059    136.79764    -54.30670    382.96480    504.71800
##            37           38           39           40           41           42
##     362.33057  -1815.56413  -1154.60228  -1052.19836   -293.79448    826.03976
##            43           44           45           46           47           48
##    -406.53002  -1828.93587   6962.16601   7111.15930  -3046.64478    768.64785
##            49           50           51           52           53           54
##    4068.64785   -468.87404    315.15897    972.13669   1621.28403   1619.42467
##            55           56           57           58           59           60
##    2268.57202     43.35142    943.35142   2692.49876   1378.82149  -1448.02008
##            61           62           63           64           65           66
##   -1547.32148    301.97992    202.67852    446.80464   1050.12055   4031.30137
##            67           68           69           70           71           72
##    5846.56664   2510.78150   2816.70670   4833.23024   5965.29608   7288.48419
##            73           74           75           76           77           78
##    8214.90152  10934.83623  16751.42259  -4388.53133  -1093.43563    290.03237
##            79           80           81           82           83           84
##     159.80051  -1550.40876  -2368.50141  -1473.29184  -5462.88588  -4117.89261
##            85           86           87           88           89           90
##   -4548.74526  -2854.63960  -2061.46085  -2879.16706  -2879.16706    289.35500
##            91           92           93           94           95           96
##    -186.98964   1144.40960   1645.96086   1271.29519   1458.78207   1175.65118
##            97           98           99          100          101          102
##    1960.33333   1789.07828   1829.06180  -1496.67652   -672.92483  -7071.92964
```

```
##         103         104         105         106         107         108
##  -8216.20638 -6998.59499 -4453.16993 -5050.81818 -3944.76604 -4362.39001
##         109         110         111         112         113         114
##  -5441.37080 -5958.20153 -6727.15363 -1107.33491 -2300.75000 -2128.14643
##         115         116         117         118         119         120
##  -4071.53284  -191.76922 -1250.75000 -2508.92444  -910.43563 -1109.50973
##         121         122         123         124         125         126
##    601.07204   840.32035  -297.12692 -2510.31292 -5175.32791  4291.32669
##         127         128         129         130         131         132
##  10742.93382 12242.93382 14795.43045 -20127.96980 -3617.58498 -2312.25366
##         133         134         135         136         137         138
##  -2556.37703 -2107.72480   135.26695   591.53174 -1024.48083  -446.16966
##         139         140         141         142         143         144
##  -1531.93993   457.44008  -213.02365  -934.59825    84.87648  -305.43659
##         145         146         147         148         149         150
##   -137.98996 -1332.60375 -1507.05739 -1095.73069 -1572.64158 -2180.16114
##         151         152         153         154         155         156
##   -669.88303   326.00563  1235.23079 -1029.95994   413.60262 -7046.23285
##         157         158         159         160         161         162
##    617.55549   592.78062   177.40529  -497.86258   154.80285  1903.19766
##         163         164         165         166         167         168
##   2620.12810   478.90385   302.93526 -2839.60004 -2955.56863 -5288.76733
##         169         170         171         172         173         174
##  -4058.08520 -3860.64317 -3952.47114 -3958.43973  -492.94834  -807.36385
##         175         176         177         178         179         180
##    315.02272  -379.73665   530.26335   432.75999 -4336.50360 -5162.00787
##         181         182         183         184         185         186
##  -5939.88827 -5941.98192  -660.41202 -1032.28159    34.03827  -337.83131
##         187         188         189         190         191         192
##   -678.57476    96.34520  -464.07883  2053.60918   774.23670  -518.59328
##         193         194         195         196         197         198
##   2367.27500   103.99429 -4194.81287 -4960.61766 -1525.05205 -2083.29888
##         199         200         201         202         203         204
##  -2248.85442 -2857.95390  -696.63411 -1944.17655  2629.65563  2741.28261
##         205
##   4529.87534
```

```
## 3.
```

```
## Start:  AIC=3424.69
## price ~ horsepower + city.mpg + peak.rpm + curb.weight + num.of.doors
##
##                Df  Sum of Sq        RSS    AIC
## - num.of.doors  1   10192607 3490187999 3423.3
## - peak.rpm      1   12324997 3492320389 3423.4
## <none>                       3479995392 3424.7
## - city.mpg      1   47472671 3527468063 3425.5
## - horsepower    1  340402702 3820398094 3441.8
## - curb.weight   1 1174580109 4654575501 3482.3
##
## Step:  AIC=3423.29
## price ~ horsepower + city.mpg + peak.rpm + curb.weight
```

```
##
##               Df  Sum of Sq        RSS    AIC
## - peak.rpm     1   12030940 3502218939 3422.0
## <none>                      3490187999 3423.3
## - city.mpg     1   48682445 3538870444 3424.1
## - horsepower   1  440974262 3931162261 3445.7
## - curb.weight  1 1240381716 4730569715 3483.6
##
## Step:  AIC=3422
## price ~ horsepower + city.mpg + curb.weight
##
##               Df  Sum of Sq        RSS    AIC
## <none>                      3502218939 3422.0
## - city.mpg     1   38636782 3540855721 3422.2
## - horsepower   1  556659511 4058878450 3450.2
## - curb.weight  1 1750422882 5252641821 3503.1
```

```
##
## Call:
## lm(formula = price ~ horsepower + city.mpg + curb.weight, data = auto_sel)
##
## Coefficients:
## (Intercept)   horsepower     city.mpg  curb.weight
##  -21384.432       75.415      121.380        9.261
```

```
## Start:  AIC=3444.63
## price ~ horsepower + city.mpg + peak.rpm + curb.weight + num.of.doors
##
##                Df  Sum of Sq        RSS    AIC
## - num.of.doors  1   10192607 3490187999 3439.9
## - peak.rpm      1   12324997 3492320389 3440.0
## - city.mpg      1   47472671 3527468063 3442.1
## <none>                       3479995392 3444.6
## - horsepower    1  340402702 3820398094 3458.4
## - curb.weight   1 1174580109 4654575501 3498.9
##
## Step:  AIC=3439.91
## price ~ horsepower + city.mpg + peak.rpm + curb.weight
##
##               Df  Sum of Sq        RSS    AIC
## - peak.rpm     1   12030940 3502218939 3435.3
## - city.mpg     1   48682445 3538870444 3437.4
## <none>                      3490187999 3439.9
## - horsepower   1  440974262 3931162261 3459.0
## - curb.weight  1 1240381716 4730569715 3496.9
##
## Step:  AIC=3435.29
## price ~ horsepower + city.mpg + curb.weight
##
##               Df  Sum of Sq        RSS    AIC
## - city.mpg     1   38636782 3540855721 3432.2
## <none>                      3502218939 3435.3
```

```
## - horsepower   1  556659511 4058878450 3460.2
## - curb.weight  1 1750422882 5252641821 3513.1
##
## Step:  AIC=3432.22
## price ~ horsepower + curb.weight
##
##               Df  Sum of Sq        RSS    AIC
## <none>                      3540855721 3432.2
## - horsepower   1  580023407 4120879128 3458.0
## - curb.weight  1 1834490017 5375345738 3512.5

##
## Call:
## lm(formula = price ~ horsepower + curb.weight, data = auto_sel)
##
## Coefficients:
## (Intercept)   horsepower  curb.weight
##  -15818.459       64.615        8.722

## Start:  AIC=3678.97
## price ~ 1
##
##                Df  Sum of Sq        RSS    AIC
## + curb.weight   1 8510293560 4.1209e+09 3451.3
## + horsepower    1 7255826951 5.3753e+09 3505.8
## + city.mpg      1 5627042447 7.0041e+09 3560.1
## + peak.rpm      1  128478511 1.2503e+10 3678.9
## <none>                       1.2631e+10 3679.0
## + num.of.doors  1   22223129 1.2609e+10 3680.6
##
## Step:  AIC=3451.35
## price ~ curb.weight
##
##                Df Sum of Sq        RSS    AIC
## + horsepower    1 580023407 3540855721 3422.2
## + peak.rpm      1 188393930 3932485198 3443.8
## + num.of.doors  1 172156795 3948722333 3444.6
## + city.mpg      1  62000678 4058878450 3450.2
## <none>                      4120879128 3451.3
##
## Step:  AIC=3422.25
## price ~ curb.weight + horsepower
##
##                Df Sum of Sq        RSS    AIC
## + city.mpg      1  38636782 3502218939 3422.0
## <none>                      3540855721 3422.2
## + num.of.doors  1  11184104 3529671617 3423.6
## + peak.rpm      1   1985277 3538870444 3424.1
##
## Step:  AIC=3422
## price ~ curb.weight + horsepower + city.mpg
##
```

```
##                Df Sum of Sq        RSS    AIC
## <none>                      3502218939 3422.0
## + peak.rpm      1  12030940 3490187999 3423.3
## + num.of.doors  1   9898550 3492320389 3423.4
```

```
##
## Call:
## lm(formula = price ~ curb.weight + horsepower + city.mpg, data = auto_sel)
##
## Coefficients:
## (Intercept)  curb.weight   horsepower     city.mpg
##  -21384.432        9.261       75.415      121.380
```

```
## Start:  AIC=3682.29
## price ~ 1
##
##                Df  Sum of Sq        RSS    AIC
## + curb.weight   1 8510293560 4.1209e+09 3458.0
## + horsepower    1 7255826951 5.3753e+09 3512.5
## + city.mpg      1 5627042447 7.0041e+09 3566.7
## <none>                       1.2631e+10 3682.3
## + peak.rpm      1  128478511 1.2503e+10 3685.5
## + num.of.doors  1   22223129 1.2609e+10 3687.3
##
## Step:  AIC=3457.99
## price ~ curb.weight
##
##                Df Sum of Sq        RSS    AIC
## + horsepower    1 580023407 3540855721 3432.2
## + peak.rpm      1 188393930 3932485198 3453.7
## + num.of.doors  1 172156795 3948722333 3454.6
## <none>                      4120879128 3458.0
## + city.mpg      1  62000678 4058878450 3460.2
##
## Step:  AIC=3432.22
## price ~ curb.weight + horsepower
##
##                Df Sum of Sq        RSS    AIC
## <none>                      3540855721 3432.2
## + city.mpg      1  38636782 3502218939 3435.3
## + num.of.doors  1  11184104 3529671617 3436.9
## + peak.rpm      1   1985277 3538870444 3437.4
```

```
##
## Call:
## lm(formula = price ~ curb.weight + horsepower, data = auto_sel)
##
## Coefficients:
## (Intercept)  curb.weight   horsepower
##  -15818.459        8.722       64.615
```

```
## 4.
```

```
##                  Estimate   Std. Error    t value     Pr(>|t|)
```

```
## (Intercept) -2.635706e+04 7226.3754586 -3.6473412 3.379537e-04
## horsepower   7.139625e+01   14.2029379  5.0268648 1.104343e-06
## city.mpg     1.430558e+02   85.6502655  1.6702319 9.643773e-02
## curb.weight  9.855041e+00    1.1689345  8.4307899 6.749078e-15
## peak.rpm     6.492573e-01    0.7819454  0.8303102 4.073535e-01
```

```
## [1] 0.7181583
```

```
##                 Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept) -21384.431936 4040.8655146 -5.292042 3.151748e-07
## horsepower      75.415001   13.3424794  5.652248 5.374600e-08
## city.mpg       121.380259   81.5119348  1.489110 1.380257e-01
## curb.weight      9.261327    0.9240072 10.023003 1.962328e-19
```

```
## [1] 0.7185938
```

```
##                Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept) -15818.45917 1540.0560178 -10.271353 3.508074e-20
## horsepower      64.61495   11.2328166   5.752337 3.223308e-08
## curb.weight      8.72178    0.8525618  10.230085 4.645579e-20
```

```
## [1] 0.7168977
```

6. Korzystając z ostatecznych modeli uzyskanych dla obu zbiorów danych, wykonaj prognozę ceny samochodu, dla którego zmienne `curb.weight` i `horsepower` są równe 2823 i 154, odpowiednio. Który model daje lepszą prognozę, gdyby cena tego samochodu wynosiła 1650? Jak możemy to wyjaśnić?

```
##        fit      lwr      upr
## 1 16484.18 11243.94 21724.42
```

```
##        fit      lwr      upr
## 1 18753.83 10437.85 27069.81
```

```
## [1] 0.8038145
```

```
## [1] 0.7168977
```

Zadanie 4. W jednym badaniu klinicznym oceniono wpływ poziomów enzymu LDH i zmian poziomów bilirubiny na zdrowie pacjentów z przewlekłym zapaleniem wątroby. Uzyskane wyniki są zawarte w pliku liver_data.RData. Zmienne to: `bilirubin` - zmiana stężenia bilirubiny we krwi, `ldh` - stężenie enzymu LDH w ciele pacjenta, `condition` - zmiana stanu pacjenta (`Yes` - pogorszenie, `No` - brak pogorszenia).

```
##   bilirubin ldh condition
## 1       0.9  75        No
## 2       0.8 150        No
## 3       0.6 250        No
## 4       0.8 375       Yes
## 5       3.2 160       Yes
## 6       1.7 106        No
```

1. Dopasuj model regresji logistycznej do tych danych. Jakie są wartości estymatorów współczynników regresji?

```
##
## Call:  glm(formula = condition ~ bilirubin + ldh, family = "binomial",
##     data = liver_data)
##
## Coefficients:
```

```
## (Intercept)    bilirubin          ldh
##    -8.13113      2.88050      0.02464
##
## Degrees of Freedom: 38 Total (i.e. Null);  36 Residual
## Null Deviance:       54.04
## Residual Deviance: 33.11     AIC: 39.11
```

2. Które współczynniki są statystycznie istotne w skonstruowanym modelu? Jakie jest dopasowanie modelu?

```
##
## Call:
## glm(formula = condition ~ bilirubin + ldh, family = "binomial",
##     data = liver_data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.05593  -0.79191   0.04353   0.57765   2.11829
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.131132   2.639959  -3.080  0.00207 **
## bilirubin    2.880497   1.105836   2.605  0.00919 **
## ldh          0.024635   0.008764   2.811  0.00494 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 54.040  on 38  degrees of freedom
## Residual deviance: 33.114  on 36  degrees of freedom
## AIC: 39.114
##
## Number of Fisher Scoring iterations: 6
```

3. Czy model ten może być zredukowany za pomocą regresji krokowej?

```
## Start:  AIC=39.11
## condition ~ bilirubin + ldh
##
##             Df Deviance    AIC
## <none>           33.114 39.114
## - ldh        1   46.989 50.989
## - bilirubin  1   48.726 52.726
```

```
##
## Call:  glm(formula = condition ~ bilirubin + ldh, family = "binomial",
##     data = liver_data)
##
## Coefficients:
## (Intercept)    bilirubin          ldh
##    -8.13113      2.88050      0.02464
##
## Degrees of Freedom: 38 Total (i.e. Null);  36 Residual
```

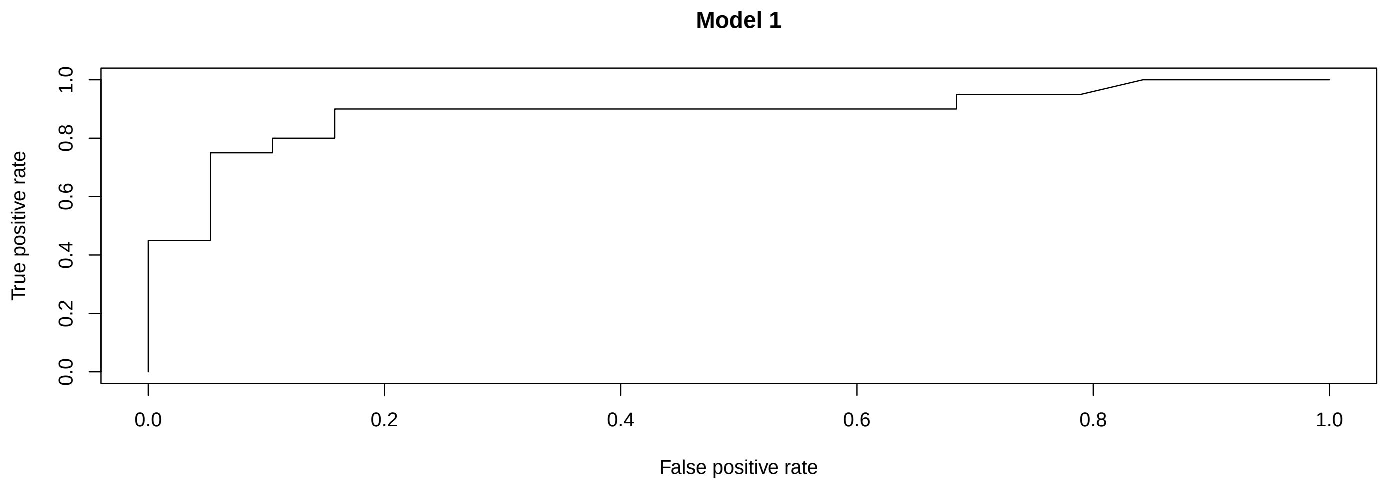
```
## Null Deviance:       54.04
## Residual Deviance: 33.11     AIC: 39.11
```

4. Zinterpretuj współczynniki modelu.

```
## bilirubin
##  17.82313
```

```
##      ldh
## 1.024941
```

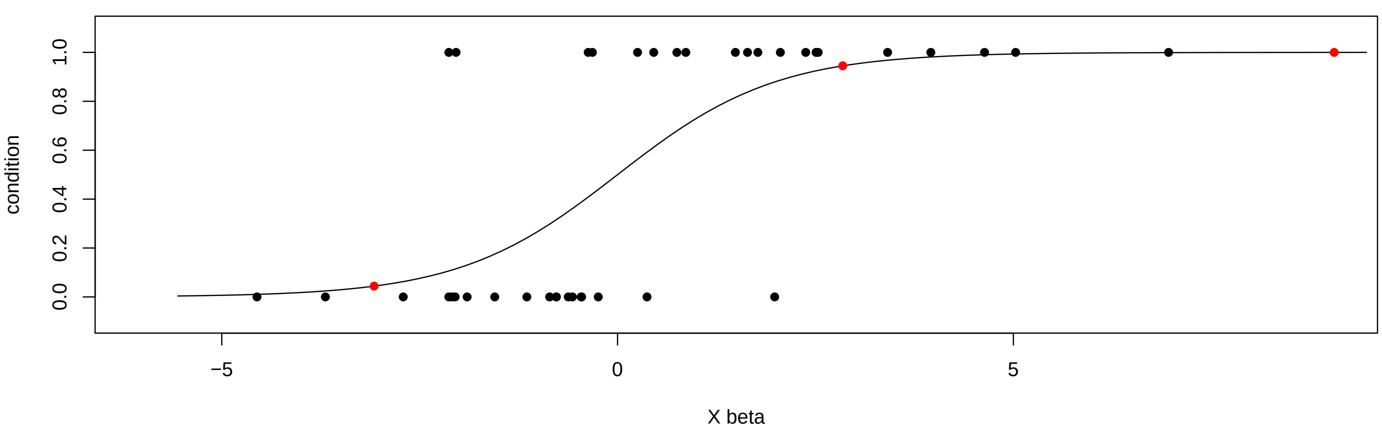
5. Narysuj krzywą ROC i oblicz AUC dla modelu.




```
## [[1]]
## [1] 0.8881579
```

6. Dokonaj predykcji zmiennej `condition` dla trzech pacjentów scharakteryzowanych następująco: (`bilirubin`, `ldh`) = (0.9, 100), (2.1, 200), (3.4, 300). Zilustruj wyniki na wykresie.

```
##          1          2          3
## 0.04414365 0.94505776 0.99988299
```



7. Powyższy wykres pokazuje, że istnieją dwie obserwacje odstające dla pacjentów z pogorszeniem i jedna obserwacja odstająca dla pacjentów bez pogorszenia. Zidentyfikuj je i wykonaj powyższą analizę dla danych bez tych trzech wartości odstających. Jak zmieniają się wyniki?

   1.

```
##
## Call:  glm(formula = condition ~ bilirubin + ldh, family = "binomial",
##     data = liver_data_wo)
```

```
##
## Coefficients:
## (Intercept)    bilirubin          ldh
##    -72.7256      30.2781       0.1947
##
## Degrees of Freedom: 35 Total (i.e. Null);  33 Residual
## Null Deviance:       49.91
## Residual Deviance: 6.207     AIC: 12.21
```

2.

```
##
## Call:
## glm(formula = condition ~ bilirubin + ldh, family = "binomial",
##     data = liver_data_wo)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.93161  -0.01879   0.00000   0.00047   1.89807
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -72.7256    45.3298  -1.604    0.109
## bilirubin    30.2781    18.9417   1.598    0.110
## ldh           0.1947     0.1235   1.577    0.115
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 49.9066  on 35  degrees of freedom
## Residual deviance:  6.2068  on 33  degrees of freedom
## AIC: 12.207
##
## Number of Fisher Scoring iterations: 10
```

3.

```
## Start:  AIC=12.21
## condition ~ bilirubin + ldh
##
##             Df Deviance    AIC
## <none>            6.207 12.207
## - ldh        1   38.422 42.422
## - bilirubin  1   44.216 48.216
```

```
##
## Call:  glm(formula = condition ~ bilirubin + ldh, family = "binomial",
##     data = liver_data_wo)
##
## Coefficients:
## (Intercept)    bilirubin          ldh
##    -72.7256      30.2781       0.1947
##
## Degrees of Freedom: 35 Total (i.e. Null);  33 Residual
## Null Deviance:       49.91
```

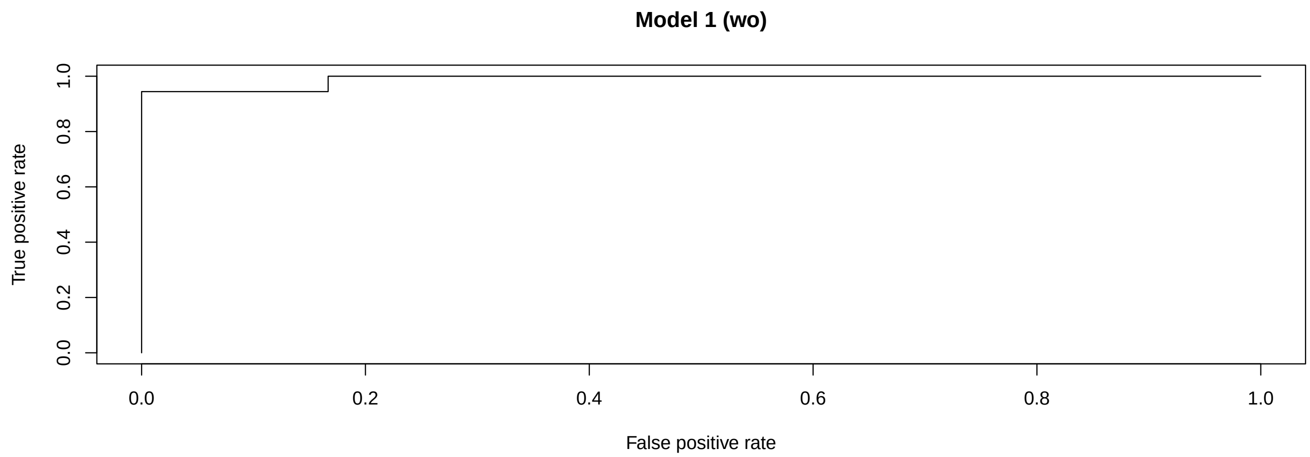
```
## Residual Deviance: 6.207     AIC: 12.21
```

4.

```
##    bilirubin
## 1.411294e+13
```

```
##      ldh
## 1.214999
```

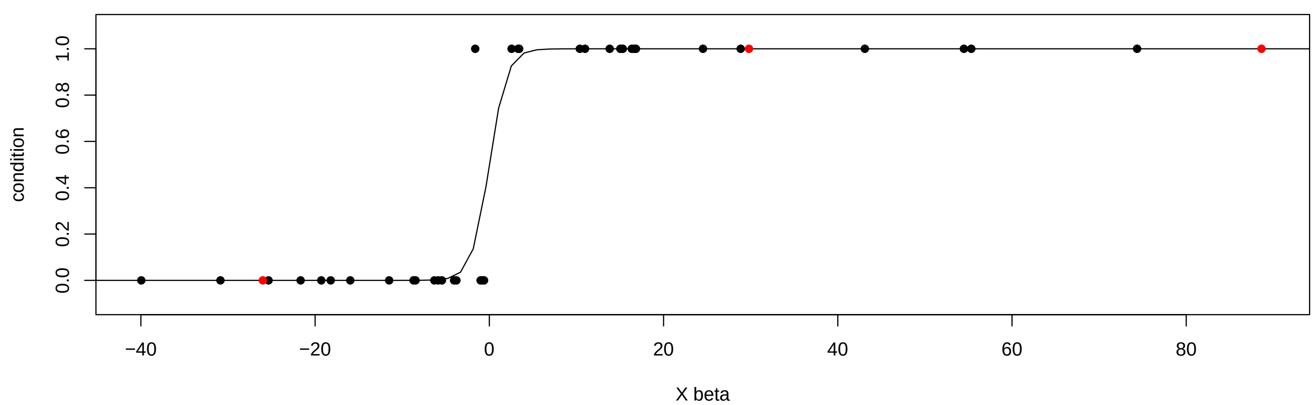
5.




```
## [[1]]
## [1] 0.9907407
```

6.

```
##            1            2            3
## 5.104082e-12 1.000000e+00 1.000000e+00
```



Zadanie 5. Użyj modelu regresji Poissona do zestawu danych `moths` (wpływ siedliska na liczbę moli) z pakietu `DAAG`. Użyj zlogarytmowanej zmiennej `meters` jako zmiennej objaśniającej, a liczby moli `A` jako zmiennej objaśnianej.

```
##   meters A  P   habitat
## 1     25 9  8    NWsoak
## 2     37 3 20    SWsoak
## 3    109 7  9 Lowerside
## 4     10 0  2 Lowerside
## 5    133 9  1 Upperside
## 6     26 3 18 Disturbed
```

1. Dopasuj model regresji Poissona do tych danych. Jakie są wartości estymatorów współczynników regresji?

```
##
## Call:  glm(formula = A ~ log(meters), family = "poisson", data = moths)
##
## Coefficients:
## (Intercept)  log(meters)
##      1.2058       0.1506
##
## Degrees of Freedom: 40 Total (i.e. Null);  39 Residual
## Null Deviance:       257.1
## Residual Deviance: 248.3     AIC: 367
```

2. Które współczynniki są statystycznie istotne w skontruowanym modelu? Jakie jest dopasowanie modelu?

```
##
## Call:
## glm(formula = A ~ log(meters), family = "poisson", data = moths)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4366  -1.7754  -1.1501   0.7331   9.2711
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.20577    0.17814   6.769  1.3e-11 ***
## log(meters)  0.15065    0.05068   2.972  0.00295 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 257.11  on 40  degrees of freedom
## Residual deviance: 248.25  on 39  degrees of freedom
## AIC: 366.97
##
## Number of Fisher Scoring iterations: 6
```

3. Czy model ten może być zredukowany za pomocą regresji krokowej?

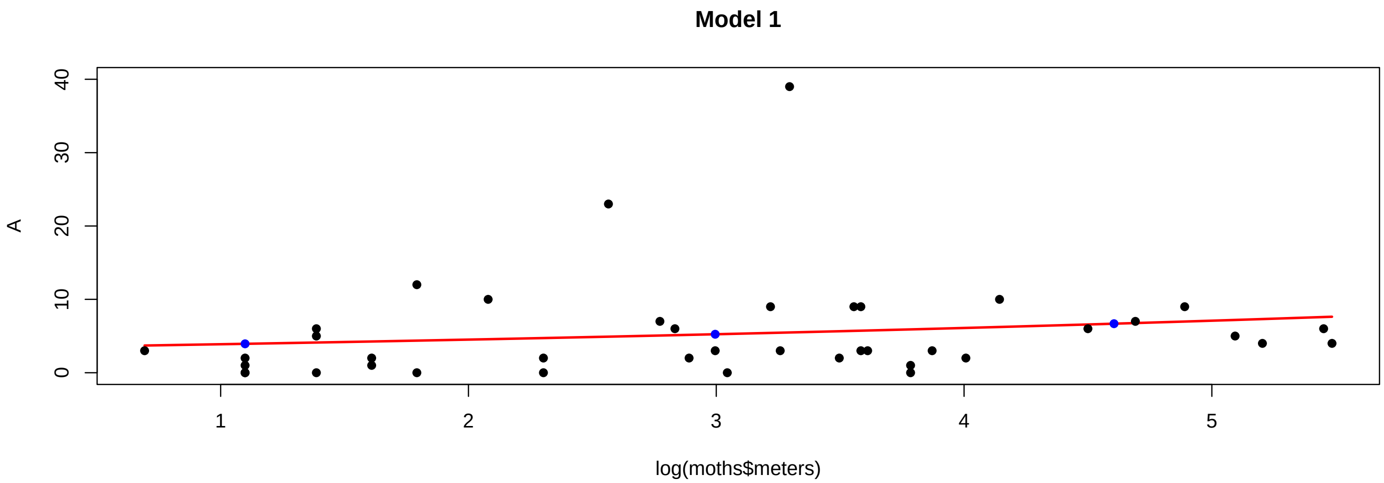
```
## Start:  AIC=366.97
## A ~ log(meters)
##
##               Df Deviance    AIC
## <none>             248.25 366.97
## - log(meters)  1   257.11 373.83
```

```
##
## Call:  glm(formula = A ~ log(meters), family = "poisson", data = moths)
##
## Coefficients:
## (Intercept)  log(meters)
##      1.2058       0.1506
```

```
##
## Degrees of Freedom: 40 Total (i.e. Null);  39 Residual
## Null Deviance:       257.1
## Residual Deviance: 248.3     AIC: 367
```

4. Dokonaj predykcji zmiennej `A` dla `meters = 3, 20, 100`. Zilustruj wyniki na wykresie.

```
##        1        2        3
## 3.940363 5.243913 6.682717
```



5. Wykonaj powyższą analizę dla zmiennej `P` jako zmiennej zależnej.

```
## 1.
```

```
##
## Call:  glm(formula = P ~ log(meters), family = "poisson", data = moths)
##
## Coefficients:
## (Intercept)  log(meters)
##      0.8643       0.1372
##
## Degrees of Freedom: 40 Total (i.e. Null);  39 Residual
## Null Deviance:       217.8
## Residual Deviance: 212.8     AIC: 309
```

```
## 2.
```

```
##
## Call:
## glm(formula = P ~ log(meters), family = "poisson", data = moths)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8679  -2.3492  -1.1408   0.6247   5.7649
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.8643     0.2145   4.030 5.58e-05 ***
## log(meters)   0.1372     0.0614   2.234   0.0255 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 217.82  on 40  degrees of freedom
## Residual deviance: 212.82  on 39  degrees of freedom
## AIC: 309.05
##
## Number of Fisher Scoring iterations: 6
```

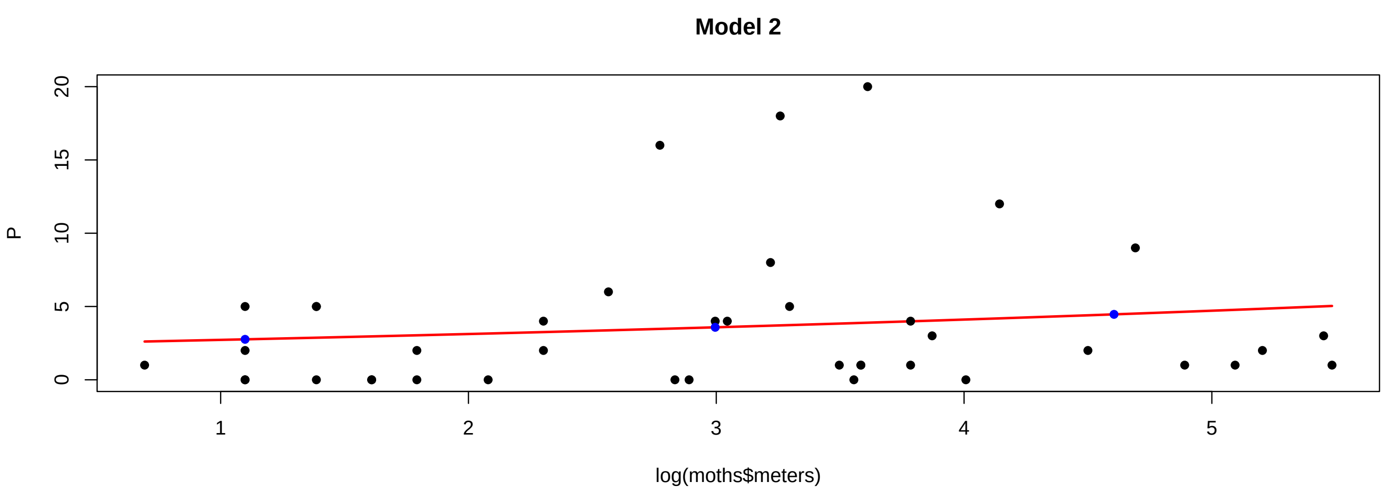
```
## 3.
```

```
## Start:  AIC=309.05
## P ~ log(meters)
##
##               Df Deviance    AIC
## <none>             212.82 309.05
## - log(meters)  1   217.82 312.04
```

```
##
## Call:  glm(formula = P ~ log(meters), family = "poisson", data = moths)
##
## Coefficients:
## (Intercept)  log(meters)
##      0.8643       0.1372
##
## Degrees of Freedom: 40 Total (i.e. Null);  39 Residual
## Null Deviance:       217.8
## Residual Deviance: 212.8     AIC: 309
```

```
## 4.
```

```
##        1        2        3
## 2.759453 3.579565 4.463761
```



8 Analiza składowych głównych

- Analiza składowych głównych jest techniką redukcji wymiaru.
- Składowe główne zostały po raz pierwszy zaproponowane przez Pearsona (1901), a następnie rozwinięte przez Hotellinga (1933).

- Analiza składowych głównych jest zaliczana do systemów uczących się bez nadzoru, a więc każdy element zbioru uczącego składa się jedynie z wektora cech (zmiennych). Zadaniem systemu uczącego się bez nadzoru jest opisanie obserwowanych danych na podstawie wyłącznie nich samych. Można je określić jako zadanie wykrycia wewnętrznej struktury zbioru danych lub współzależności między tymi danymi.
- Celem badacza może być redukcja danych, a dokładniej - liczby zmiennych. Polega ona na poszukiwaniu takiego zbioru zmiennych, mniej licznego od zbioru zmiennych oryginalnych, na których podstawie można z pewnym, ale możliwie najmniejszym błędem, odtworzyć wartości zmiennych oryginalnych. Aby taka redukcja była możliwa między zmiennymi oryginalnymi muszą zachodzić zależności statystyczne.
- Nowe zmienne (składowe główne) są liniowymi funkcjami zmiennych oryginalnych.
- Metoda składowych głównych ma głównie charakter ekploracyjny i umożliwia redukcję danych w przypadku zbioru skorelowanych ze sobą zmiennych.
- Zmienne te są traktowane w jednakowy sposób, tj. nie są one dzielone - tak jak w przypadku analizy regresji - na zmienne zależne i niezależne.
- Metoda ta przekształca oryginalne, skorelowane zmienne w nowe, nieskorelowane zmienne, tzw. składowe główne, które wyjaśniają w maksymalnym stopniu całkowitą wariancję z próby.
- Każda nowa zmienna jest liniową funkcją oryginalnych zmiennych.
- Składowe główne są uporządkowane według udziału w redukcji wspólnego zróżnicowania oryginalnych zmiennych (wielkości całkowitej wariancji). Pierwsza składowa główna redukuje największą część tego zróżnicowania. Druga - kolejną największą część tego zróżnicowania, którego nie redukowała pierwsza składowa główna, itd.
- Badacz może więc zredukować liczbę zmiennych, ograniczając się do kilku pierwszych składowych głównych, z możliwie małą stratą informacji.
- Oceną ograniczenia się tylko do kilku składowych głównych jest udział zredukowanej przez nie wariancji w wielkości całkowitej wariancji.
- W sytuacji gdy oryginalne zmienne nie są skorelowane, zastosowanie metody składowych głównych nie zapewnia możliwości redukcji danych przy ograniczonej stracie informacji.
- Wyznaczanie wartości składowych głównych dla badanych obiektów nie przedstawia żadnej trudności i nie wymaga przyjmowania dodatkowych założeń.
- Analiza składowych głównych jest często stosowana. Przekształcenie licznego zbioru skorelowanych zmiennych oryginalnych w kilka nieskorelowanych składowych głównych stanowi na ogół pierwszy etap dla zastosowania innych metod wielowymiarowej analizy danych, np. analizy skupień, regresji czy też analizy dyskryminacyjnej.

8.1 Konstrukcja składowych głównych

- Pierwsza składowa główna jest definiowana jako unormowana kombinacja liniowa mająca maksymalną wariancję z próby spośród wszystkich unormowanych kombinacji liniowych zmiennych pierwotnych $X_1, \dots, X_p$.
- Dokładniej, dla wektora obserwacji $\mathbf{X} = (X_1, X_2, \dots, X_p)'$ w próbie poszukujemy kombinacji liniowej

$$Z_1 = a_{11}X_1 + a_{12}X_2 + \dots + a_{1p}X_p = \mathbf{a}_1'\mathbf{X},$$

której wariancja z próby

$$s_{Z_1}^2 = \mathbf{a}_1'\mathbf{S}\mathbf{a}_1$$

jest maksymalna, gdzie

$$\mathbf{S} = \frac{1}{n-1}\sum_{j=1}^{n}(\mathbf{X}_j - \bar{\mathbf{X}})(\mathbf{X}_j - \bar{\mathbf{X}})'$$

jest macierzą kowariancji z próby $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ oraz $\bar{\mathbf{X}} = \frac{1}{n}\sum_{j=1}^{n}\mathbf{X}_j$, natomiast wektor $\mathbf{a}_1$ spełnia warunek $\mathbf{a}_1'\mathbf{a}_1 = 1$, tj. kwadrat jego długości jest równy jeden. Warunek ten wprowadzony jest po to, by zapewnić jednoznaczność (z wyjątkiem znaku) składowej głównej.

- Wektor $\mathbf{a}_1$, który maksymalizuje wariancję $s^2_{Z_1}$, przy dodatkowym warunku $\mathbf{a}_1'\mathbf{a}_1 = 1$, jest wektorem charakterystycznym odpowiadającym największej wartości własnej λ_1 macierzy $\mathbf{S}$, lub inaczej największemu pierwiastkowi równania

$$|\mathbf{S} - \lambda\mathbf{I}| = 0.$$

- Wariancja składowej głównej Z_1 jest zatem największym pierwiastkiem tego równania.
- W celu wyznaczenia drugiej składowej głównej, konstruujemy kombinację liniową

$$Z_2 = \mathbf{a}_2'\mathbf{X}$$

taką, że jest ona nieskorelowana z Z_1, ma maksymalną wariancję i spełnia warunek $\mathbf{a}_2'\mathbf{a}_2 = 1$. Wariancja z próby Z_2 jest równa

$$s^2_{Z_2} = \mathbf{a}_2'\mathbf{S}\mathbf{a}_2.$$

- Stąd poszukujemy wektora $\mathbf{a}_2$ maksymalizującego $s^2_{Z_2}$ przy dodatkowym warunkach $\mathbf{a}_2'\mathbf{a}_2 = 1$ i $\mathbf{a}_2'\mathbf{a}_1 = 0$.
- Wektor $\mathbf{a}_2$ jest wektorem własnym macierzy $\mathbf{S}$ odpowiadającym drugiej wartości własnej $\lambda_2 < \lambda_1$ ortogonalnym do wektora $\mathbf{a}_1$ i unormowanym tak, by kwadrat jego długości był równy jedności ($\mathbf{a}_2'\mathbf{a}_2 = 1$).
- Ponieważ macierz $\mathbf{S}$ ma p wartości własnych, to otrzymujemy p składowych głównych:

$$\begin{aligned} Z_1 &= \mathbf{a}_1'\mathbf{X}, \\ Z_2 &= \mathbf{a}_2'\mathbf{X}, \\ &\ldots \\ Z_p &= \mathbf{a}_p'\mathbf{X}. \end{aligned}$$

- Składowe główne $Z_1, Z_2, \ldots, Z_p$ można zapisać w postaci

$$\mathbf{Z} = \mathbf{A}\mathbf{X},$$

gdzie

$$\mathbf{Z} = \begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_p \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} \mathbf{a}_1' \\ \mathbf{a}_2' \\ \vdots \\ \mathbf{a}_p' \end{bmatrix}.$$

- W rezultacie otrzymujemy tyle składowych głównych ile było zmiennych wejściowych, ale najczęściej jedynie kilka z nich wyjaśnia prawie całą zmienność oryginalnych danych.

8.2 Własności

- Jeżeli wektor własny $\mathbf{a}_1$ macierzy kowariancji z próby $\mathbf{S}$ jest wyskalowany tak, by $\mathbf{a}_1'\mathbf{a}_1 = 1$, to wariancja z próby pierwszej składowej głównej Z_1 jest równa
$$s_{Z_1}^2 = \mathbf{a}_1'\mathbf{S}\mathbf{a}_1 = \lambda_1.$$
- Stąd wartość własna λ_1 macierzy $\mathbf{S}$ jest równa wariancji z próby pierwszej składowej głównej $Z_1 = \mathbf{a}_1'\mathbf{X}$.
- Podobnie, wariancja z próby każdej innej składowej głównej jest równa odpowiedniej wartości własnej:
$$s_{Z_j}^2 = \mathbf{a}_j'\mathbf{S}\mathbf{a}_j = \lambda_j, \quad j = 2, 3, \dots, p.$$
- Składowa główna Z_1 ma maksymalną wariancję λ_1, natomiast składowa główna Z_p ma najmniejszą wariancję λ_p, gdzie $\lambda_1 > \lambda_2 > \cdots > \lambda_p$ są wartościami własnymi macierzy kowariancji z próby $\mathbf{S}$.
- Składowe główne są wzajemnie ortogonalne, tj. $\mathbf{a}_j'\mathbf{a}_k = 0$, dla wszystkich $j \neq k$. Ortogonalność składowych głównych pociąga za sobą własność ich nieskorelowania.
- Suma wariancji z próby składowych głównych jest równa sumie wariancji z próby zmiennych pierwotnych:
$$\sum_{j=1}^{p} s_{Z_j}^2 = \sum_{j=1}^{p} \lambda_j = \mathrm{tr}(\mathbf{S}) = \sum_{j=1}^{p} s_{X_j}^2.$$
- W analizie składowych głównych oczekujemy, że dla pewnego małego k, suma
$$\lambda_1 + \lambda_2 + \cdots + \lambda_k$$
będzie bliska
$$\mathrm{tr}(\mathbf{S}) = \lambda_1 + \lambda_2 + \cdots + \lambda_p.$$
Jeśli tak jest, to k pierwszych składowych głównych wyjaśnia dobrze zmienność wektora $\mathbf{X} = (X_1, X_2, \dots, X_p)'$ i pozostałe $p-k$ składowe główne wnoszą niewiele, ponieważ mają one małe wariancje z próby.
- Wskaźnik
$$\frac{\lambda_1 + \cdots + \lambda_k}{\lambda_1 + \cdots + \lambda_p}\, 100\%$$
jest procentową miarą wyjaśniania zmienności wektora $\mathbf{X}$ przez pierwszych k składowych głównych.
- Składowe główne nie są niezmiennicze względem zmiany skali zmiennych pierwotnych. Oznacza to, że przeskalowanie danych zmienia wyniki analizy metodą składowych głównych.
- Z tego względu składowe główne uzyskane z macierzy kowariancji oraz korelacji różnią się. Zaleca się wykorzystywać te uzyskane z macierzy kowariancji. W przypadku jednak dużych różnic w wariancjach lub cech mierzonych na różnych skalach należy wpierw przeskalować dane.

8.3 Ładunki i wyniki

- Jako wynik otrzymujemy najczęściej dwa typy parametrów: **ładunki** (ang. *loadings*) oraz **wyniki** (ang. *scores*).
- Ładunki to współczynniki pokazujące wkład poszczególnych zmiennych bazowych w tworzeniu składowych głównych. Im wartość bezwzględna z ładunku większa tym zmienna ma większy wkład w budowę składowej głównej.
- Wyniki nie są niczym innym jak współrzędnymi obserwacji w nowym układzie współrzędnych utworzonym przez składowe główne, to one najczęściej podlegają wizualizacji.
- Niestety przy większej liczbie pierwotnych zmiennych występują problemy z interpretacją ładunków.

8.4 Metody pomijania składowych głównych

- Jeśli chcemy zredukować wymiar danych musimy się zastanowić ile składowych wybrać do dalszej analizy.
- Najczęściej decyzję tę podejmuje się bazując na **wykresie osypiska**, zwanym też **wykresem piargowym** (ang. *scree plot*). Wartości własne numerujemy w porządku malejącym. Na osi odciętych zaznaczamy numer wartości własnych, na osi rzędnych zaznaczamy wielkości wartości własnych i wielkości te łączymy odcinkami. Jako optymalną liczbę czynników wybieramy tę, gdzie wykres się znacząco spłaszcza. Kryterium osypiska prowadzi niekiedy do odrzucenia zbyt wielu czynników, ale w typowych sytuacjach (niezbyt dużo czynników i sporo obserwacji) radzi sobie całkiem dobrze.
- Drugim popularnym kryterium jest ustalenie pewnego poziomu wariancji jaki muszą wyjaśnić składowe główne (najczęściej 90%).
- W trzecim podejściu, pomijamy te składowe główne, których wartości własne są mniejsze od średniej

$$\bar{\lambda} = \frac{1}{p} \sum_{j=1}^{p} \lambda_j.$$

- Jest to zarazem średnia wariancja zmiennych pierwotnych, ponieważ $\sum_{j=1}^{p} \lambda_j = \mathrm{tr}(\mathbf{S})$.

8.5 Wizualizacja

- Na koniec możemy zwizualizować nowe dane na jednym wykresie, na którym jako punkty będą przedstawione poszczególne obserwacje w nowym układzie dwóch pierwszych składowych głównych, natomiast wektory oznaczać będą cechy.
- Kierunek wektorów pokazuje wpływ tych cech odpowiednio na pierwszą i drugą składową. Kąt przecięcia strzałek jest proporcjonalny do zależności pomiędzy cechami (dokładnie iloczyn skalarny odpowiednich wektorów wyznacza korelację), a ich długość odzwierciedla odchylenie standardowe.
- Tego typu wykres nazywa się **biplotem** (ang. *biplot*).
- Żeby stwierdzić, czy taki wykres jest adekwatnym odzwierciedleniem położenia oryginalnych punktów, można na niego nanieść **minimalne drzewo rozpinające (MST)** (ang. *minimum spanning tree*). MST to graf, którego wierzchołkami są obserwacje, dwa punkty połączone są dokładnie jedną ścieżką, a suma krawędzi jest minimalna. Punkty połączone krawędziami powinny być blisko siebie na wykresie.

8.6 Zastosowanie

- Analiza składowych głównych ma szerokie zastosowanie.
- Jej dwa popularne zastosowania to regresja składowych głównych (PCR) i regresja częściowych najmniejszych kwadratów (PLSR).
- Pierwsza z nich polega na zastąpieniu oryginalnych zmiennych przez pewną liczbę składowych głównych.
- Metoda PLSR jest wariantem metody składowych głównych, w której szukamy pewnej liczby ortogonalnych do siebie kombinacji liniowych predyktorów dobrze prognozujących zmienną objaśnianą.
- Przewaga PCR/PLSR nad metodą najmniejszych kwadratów jest najczęściej widoczna w sytuacji, gdy liczba zmiennych objaśniających jest duża w stosunku do liczby obserwacji.

8.7 Przykład 8

Przykład. Zbiór danych `USArrests` zawiera informacje dotyczące liczby morderstw, napadów, gwałtów przypadających na 100,000 osób w poszczególnych stanach USA w roku 1973 oraz procent ludności mieszkającej w miastach. Chcielibyśmy się dowiedzieć, czy stany są do siebie w pewien sposób zbliżone oraz spróbować zwizualizować je na płaszczyźnie.

```r
head(USArrests)
```

```
##            Murder Assault UrbanPop Rape
## Alabama      13.2     236       58 21.2
## Alaska       10.0     263       48 44.5
## Arizona       8.1     294       80 31.0
## Arkansas      8.8     190       50 19.5
## California    9.0     276       91 40.6
## Colorado      7.9     204       78 38.7
```

```r
dim(USArrests)
```

```
## [1] 50  4
```

- przygotowanie danych do analizy składowych głównych

```r
# sprawdzamy czy wariancje (,,zmienności'') zmiennych są bardzo zróżnicowane
var(USArrests)
```

```
##              Murder   Assault   UrbanPop      Rape
## Murder    18.970465  291.0624   4.386204  22.99141
## Assault  291.062367 6945.1657 312.275102 519.26906
## UrbanPop   4.386204  312.2751 209.518776  55.76808
## Rape      22.991412  519.2691  55.768082  87.72916
```

```r
# tak są, więc centrujemy i skalujemy funkcją scale
USArrests_scale <- scale(USArrests)
var(USArrests_scale)
```

```
##              Murder   Assault   UrbanPop      Rape
## Murder   1.00000000 0.8018733 0.06957262 0.5635788
## Assault  0.80187331 1.0000000 0.25887170 0.6652412
## UrbanPop 0.06957262 0.2588717 1.00000000 0.4113412
## Rape     0.56357883 0.6652412 0.41134124 1.0000000
```

- model analizy składowych głównych w R i procent wyjaśnianej wariancji zmiennych oryginalnych przez poszczególne składowe główne

```r
pca <- prcomp(USArrests, scale = TRUE)
# lub
# pca <- prcomp(USArrests_scale)
pca
```

```
## Standard deviations (1, .., p=4):
## [1] 1.5748783 0.9948694 0.5971291 0.4164494
##
## Rotation (n x k) = (4 x 4):
##                 PC1        PC2        PC3         PC4
## Murder   -0.5358995  0.4181809 -0.3412327  0.64922780
## Assault  -0.5831836  0.1879856 -0.2681484 -0.74340748
## UrbanPop -0.2781909 -0.8728062 -0.3780158  0.13387773
## Rape     -0.5434321 -0.1673186  0.8177779  0.08902432
```

```r
# bez skalowania
prcomp(USArrests)
```

```
## Standard deviations (1, .., p=4):
```

```
## [1] 83.732400 14.212402  6.489426  2.482790
##
## Rotation (n x k) = (4 x 4):
##                 PC1         PC2         PC3         PC4
## Murder   0.04170432 -0.04482166  0.07989066 -0.99492173
## Assault  0.99522128 -0.05876003 -0.06756974  0.03893830
## UrbanPop 0.04633575  0.97685748 -0.20054629 -0.05816914
## Rape     0.07515550  0.20071807  0.97408059  0.07232502
```

```r
summary(pca)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4
## Standard deviation     1.5749 0.9949 0.59713 0.41645
## Proportion of Variance 0.6201 0.2474 0.08914 0.04336
## Cumulative Proportion  0.6201 0.8675 0.95664 1.00000
```

- wyniki

```r
head(pca$x)
```

```
##                   PC1        PC2         PC3          PC4
## Alabama    -0.9756604  1.1220012 -0.43980366  0.154696581
## Alaska     -1.9305379  1.0624269  2.01950027 -0.434175454
## Arizona    -1.7454429 -0.7384595  0.05423025 -0.826264240
## Arkansas    0.1399989  1.1085423  0.11342217 -0.180973554
## California -2.4986128 -1.5274267  0.59254100 -0.338559240
## Colorado   -1.4993407 -0.9776297  1.08400162  0.001450164
```

- ładunki

```r
pca$rotation
```

```
##                 PC1        PC2        PC3         PC4
## Murder   -0.5358995  0.4181809 -0.3412327  0.64922780
## Assault  -0.5831836  0.1879856 -0.2681484 -0.74340748
## UrbanPop -0.2781909 -0.8728062 -0.3780158  0.13387773
## Rape     -0.5434321 -0.1673186  0.8177779  0.08902432
```

- wykres osypiska (piargowy)

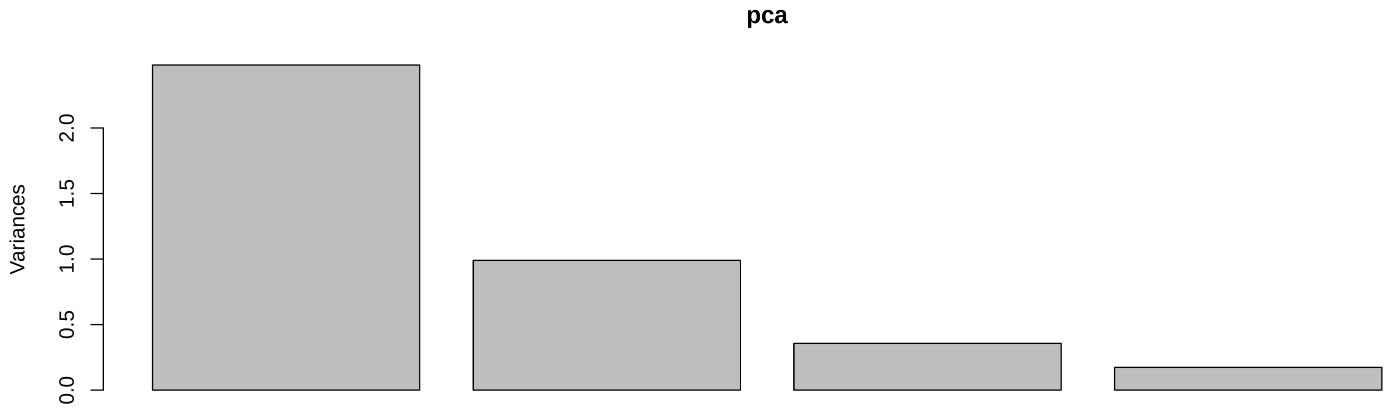
```r
pca$sdev^2
```

```
## [1] 2.4802416 0.9897652 0.3565632 0.1734301
```

```r
apply(pca$x, 2, var)
```

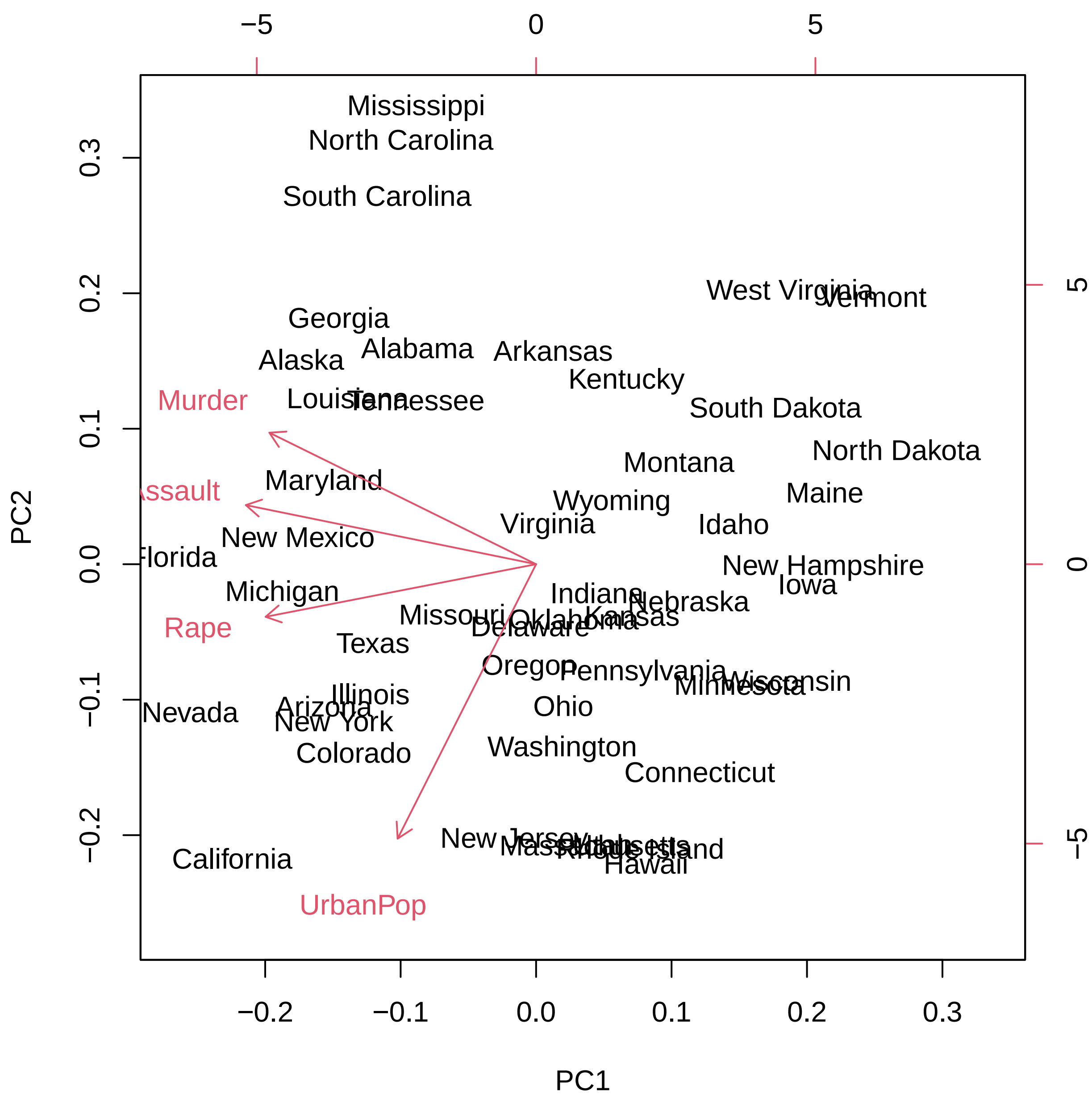
```
##       PC1       PC2       PC3       PC4
## 2.4802416 0.9897652 0.3565632 0.1734301
```

```r
plot(pca)
```



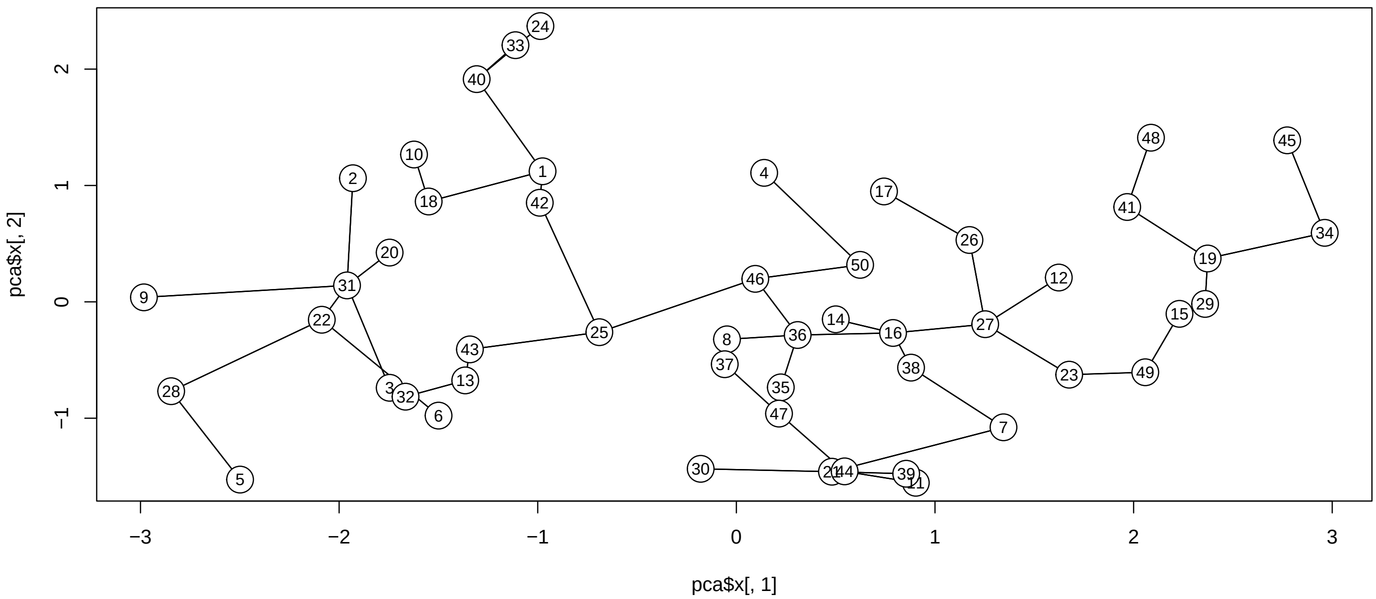
- biplot

```
biplot(pca)
```



- minimalne drzewo rozpinające (MST)

```r
library(ape)
plot(mst(dist(USArrests_scale)), x1 = pca$x[, 1], x2 = pca$x[, 2])
```



```r
# odczytywanie nazw obserwacji
row.names(USArrests_scale[c(24, 33),])
```

```
## [1] "Mississippi"    "North Carolina"
```

8.8 Zadania 8

Zadanie 1. W powyższym przykładzie do analizy składowych głównych zostały wykorzystane wszystkie zmienne. Jednak jedna z nich jest bardzo słabo skorelowana z pozostałymi. Ustal tę zmienną, a następnie wykonaj poniższe polecenia bez jej uwzględnienia:

1. Dokonaj analizy składowych głównych.

```
## Standard deviations (1, .., p=3):
## [1] 1.5357670 0.6767949 0.4282154
##
## Rotation (n x k) = (3 x 3):
##                 PC1        PC2        PC3
## Murder   -0.5826006  0.5339532 -0.6127565
## Assault -0.6079818  0.2140236  0.7645600
## Rape     -0.5393836 -0.8179779 -0.1999436
```

2. Jaki procent wariancji tłumaczony jest przez poszczególne składowe?

```
## Importance of components:
##                           PC1    PC2     PC3
## Standard deviation     1.5358 0.6768 0.42822
## Proportion of Variance 0.7862 0.1527 0.06112
## Cumulative Proportion  0.7862 0.9389 1.00000
```

3. Wyznacz współrzędne obserwacji w nowym układzie współrzędnych utworzonym przez składowe główne.

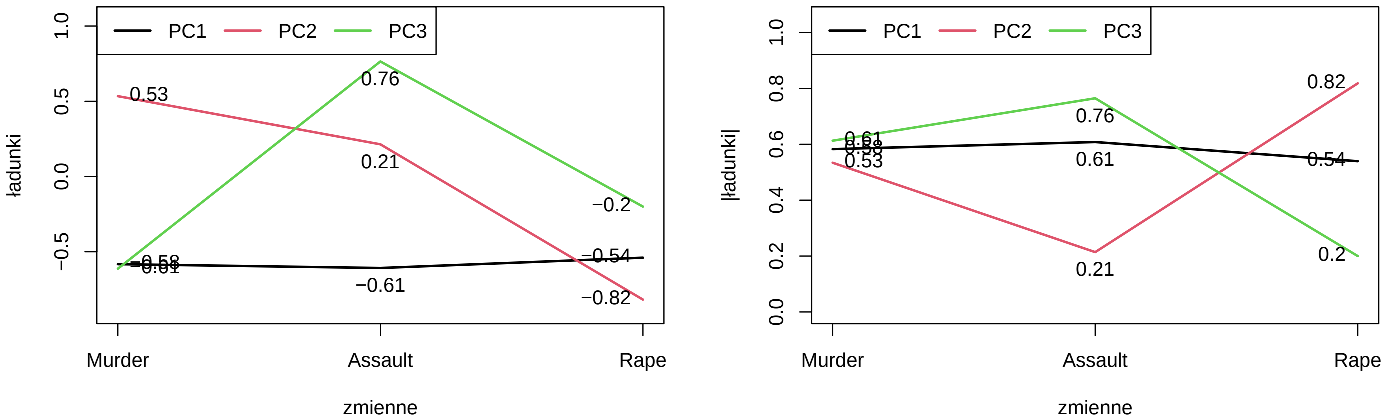
```
##                  PC1        PC2         PC3
## Alabama    -1.1980278  0.8338118 -0.16217848
```

```
## Alaska     -2.3087473 -1.5239622  0.03833574
## Arizona    -1.5033307 -0.4983038  0.87822311
## Arkansas   -0.1759894  0.3247326  0.07111174
## California -2.0452358 -1.2725770  0.38153933
## Colorado   -1.2634133 -1.4264063 -0.08369314
```

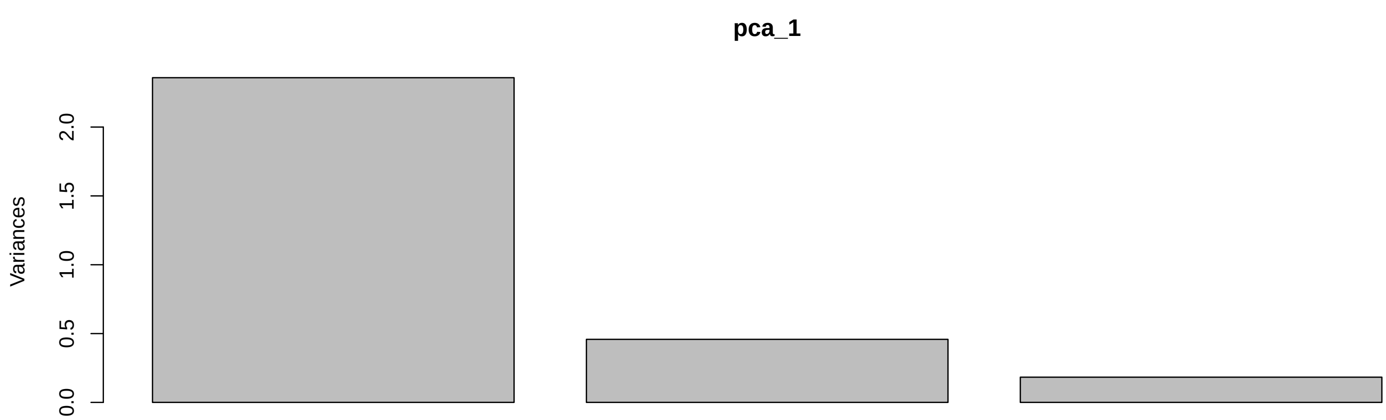
```
## ...
```

4. Dokonaj interpretacji ładunków i zilustruj je na wykresie.

```
##                PC1        PC2        PC3
## Murder  -0.5826006  0.5339532 -0.6127565
## Assault -0.6079818  0.2140236  0.7645600
## Rape    -0.5393836 -0.8179779 -0.1999436
```

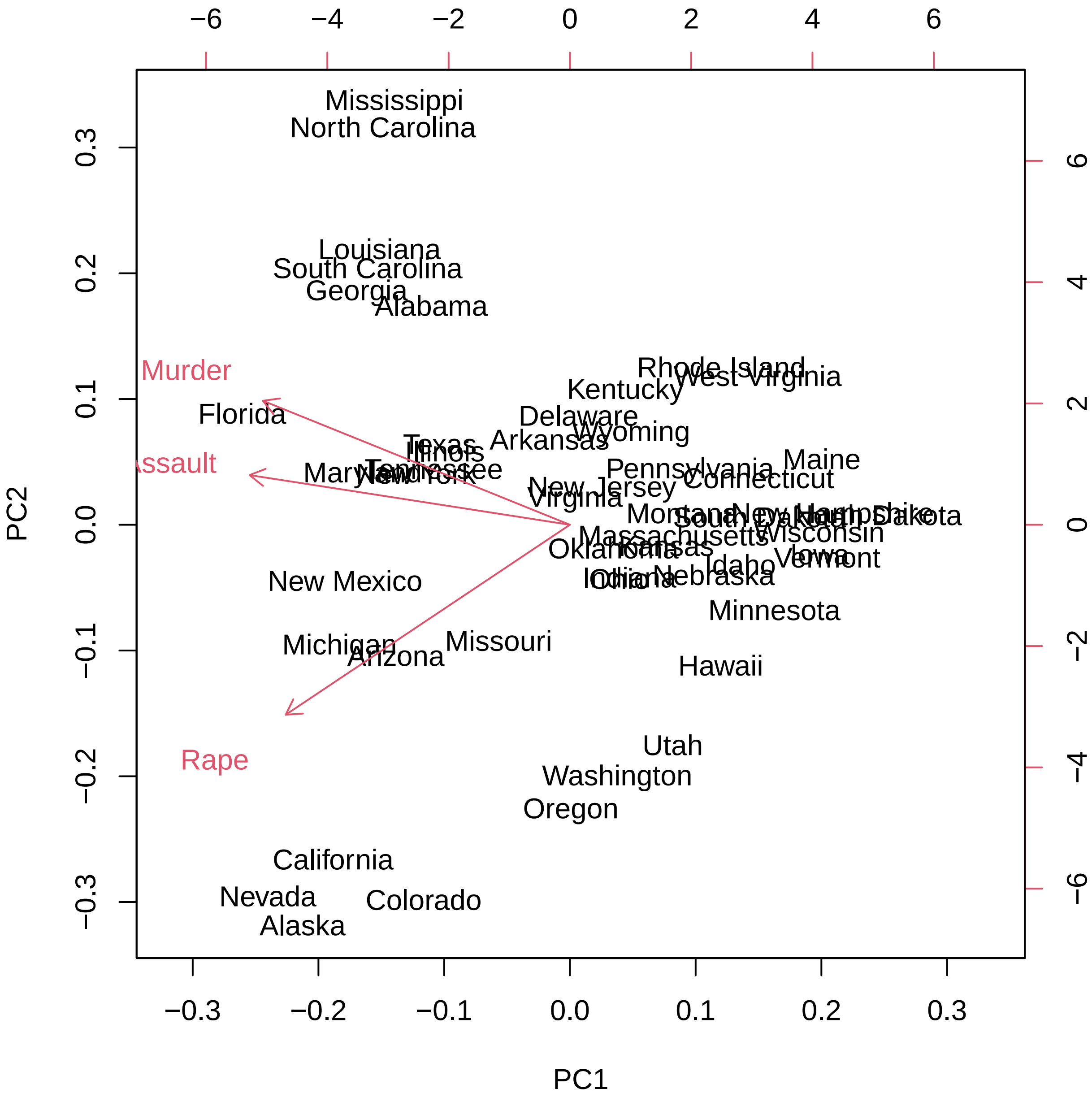


5. Narysuj wykres osypiska i zaproponuj optymalną liczbę składowych głównych w oparciu o trzy kryteria.

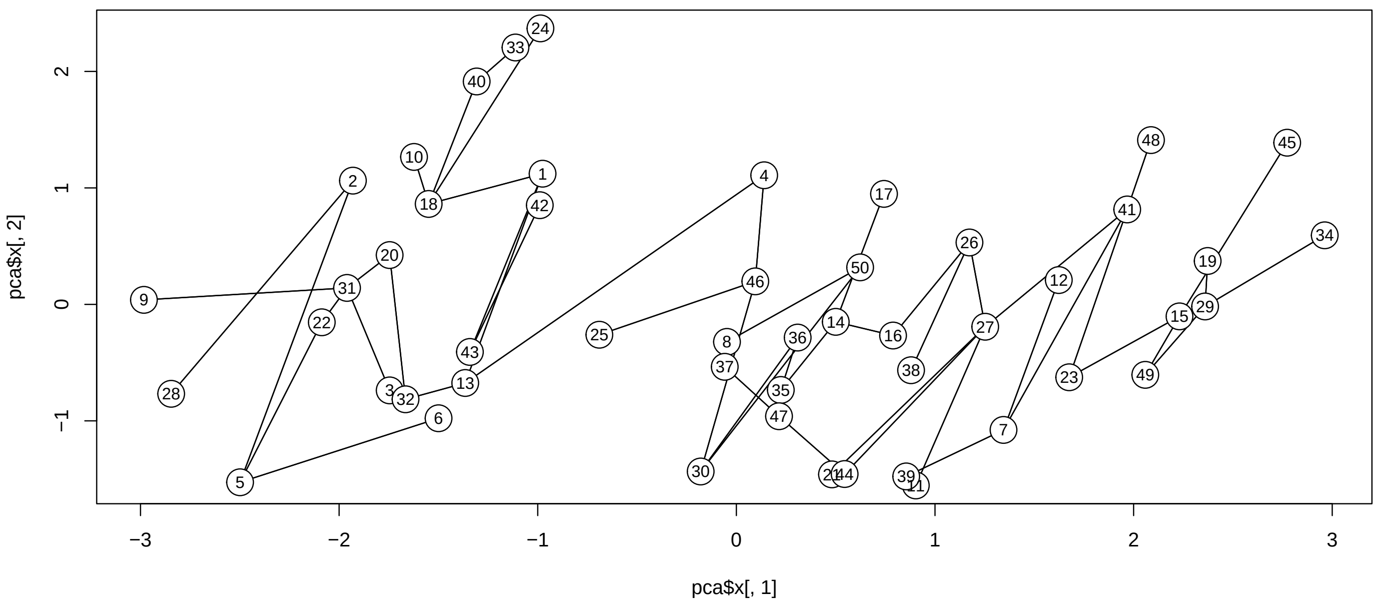



```
## 1 lub 2
```

6. Przedstaw stany w układzie dwóch pierwszych składowych głównych (dokładniej narysuj biplot i dokonaj jego interpretacji).



7. Przedstaw stany za pomocą minimalnego drzewa rozpinającego.



Zadanie 2. Zbiór danych `mtcars` zawiera informacje na temat 32 samochodów z roku 1974.

1. Dokonaj analizy składowych głównych biorąc pod uwagę cechy: `mpg`, `disp`, `hp`, `drat`, `wt`, `qsec`.

```
## Standard deviations (1, .., p=6):
## [1] 2.0463129 1.0714999 0.5773705 0.3928874 0.3532648 0.2279872
##
## Rotation (n x k) = (6 x 6):
##             PC1         PC2         PC3        PC4        PC5         PC6
## mpg  -0.4586835  0.05867609 -0.19479235 0.78205878 -0.1111533 -0.35249327
## disp  0.4660354 -0.06065296  0.09688406 0.60001871  0.2946297  0.56825752
## hp    0.4258534  0.36147576  0.14613554 0.12301873 -0.8057408 -0.04771555
## drat -0.3670963  0.43652537  0.80049152 0.02259258  0.1437714  0.11277675
## wt    0.4386179 -0.29953457  0.41776208 0.10438337  0.2301541 -0.69246040
## qsec -0.2528320 -0.76284877  0.34059066 0.04268124 -0.4218755  0.24152663
```

2. Jaki procent wariancji tłumaczony jest przez poszczególne składowe?

```
## Importance of components:
##                           PC1    PC2     PC3     PC4    PC5     PC6
## Standard deviation     2.0463 1.0715 0.57737 0.39289 0.3533 0.22799
## Proportion of Variance 0.6979 0.1913 0.05556 0.02573 0.0208 0.00866
## Cumulative Proportion  0.6979 0.8892 0.94481 0.97054 0.9913 1.00000
```

3. Wyznacz współrzędne obserwacji w nowym układzie współrzędnych utworzonym przez składowe główne.

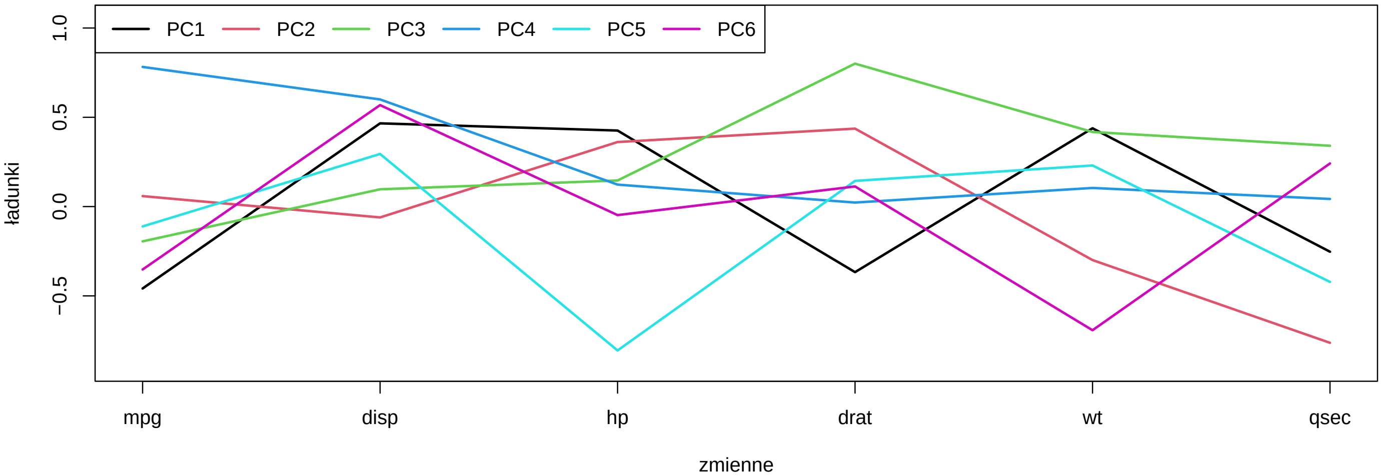
```
##                          PC1          PC2        PC3        PC4         PC5
## Mazda RX4         -0.8425806  0.873469391 -0.2282783 -0.3742725  0.51522641
## Mazda RX4 Wag     -0.8075041  0.556341552 -0.0126678 -0.3336931  0.44299870
## Datsun 710        -1.6850448 -0.040006569 -0.1564937 -0.4057157 -0.03340433
## Hornet 4 Drive    -0.0964443 -1.294377904 -0.5702297  0.2520788 -0.04326023
## Hornet Sportabout  1.2915096 -0.006516693 -0.5250741  0.4813192  0.12822104
## Valiant            0.2187309 -2.005957905 -0.7258399 -0.3136170 -0.21465335
##                          PC6
## Mazda RX4         -0.05293884
## Mazda RX4 Wag     -0.15771326
## Datsun 710         0.10756126
```

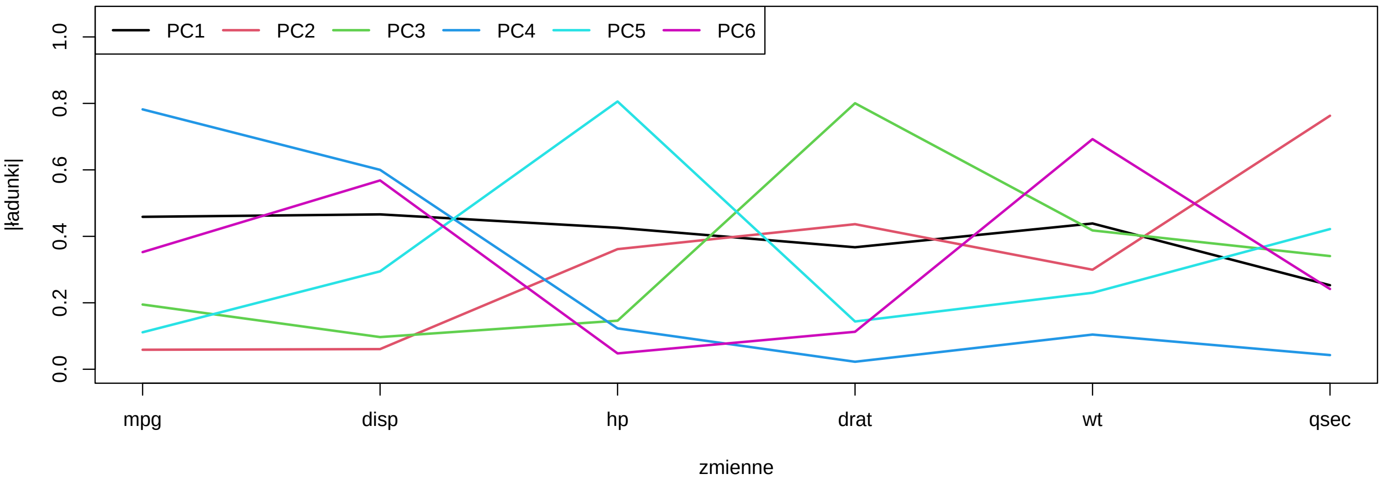
```
## Hornet 4 Drive      0.18173489
## Hornet Sportabout  0.29051949
## Valiant            0.09145688
```

```
## ...
```

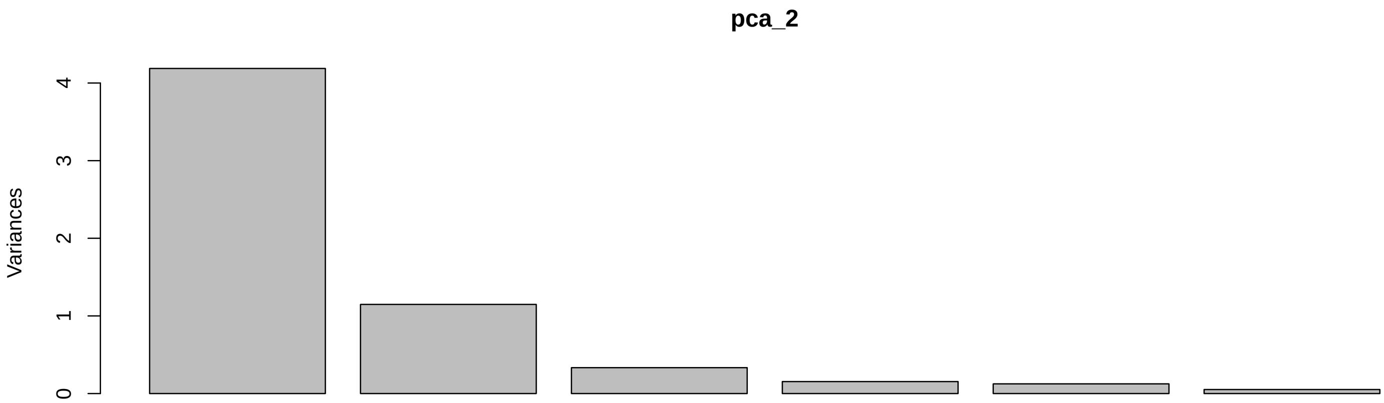
4. Dokonaj interpretacji ładunków i zilustruj je na wykresie.

```
##              PC1         PC2         PC3        PC4        PC5         PC6
## mpg  -0.4586835  0.05867609 -0.19479235 0.78205878 -0.1111533 -0.35249327
## disp  0.4660354 -0.06065296  0.09688406 0.60001871  0.2946297  0.56825752
## hp    0.4258534  0.36147576  0.14613554 0.12301873 -0.8057408 -0.04771555
## drat -0.3670963  0.43652537  0.80049152 0.02259258  0.1437714  0.11277675
## wt    0.4386179 -0.29953457  0.41776208 0.10438337  0.2301541 -0.69246040
## qsec -0.2528320 -0.76284877  0.34059066 0.04268124 -0.4218755  0.24152663
```



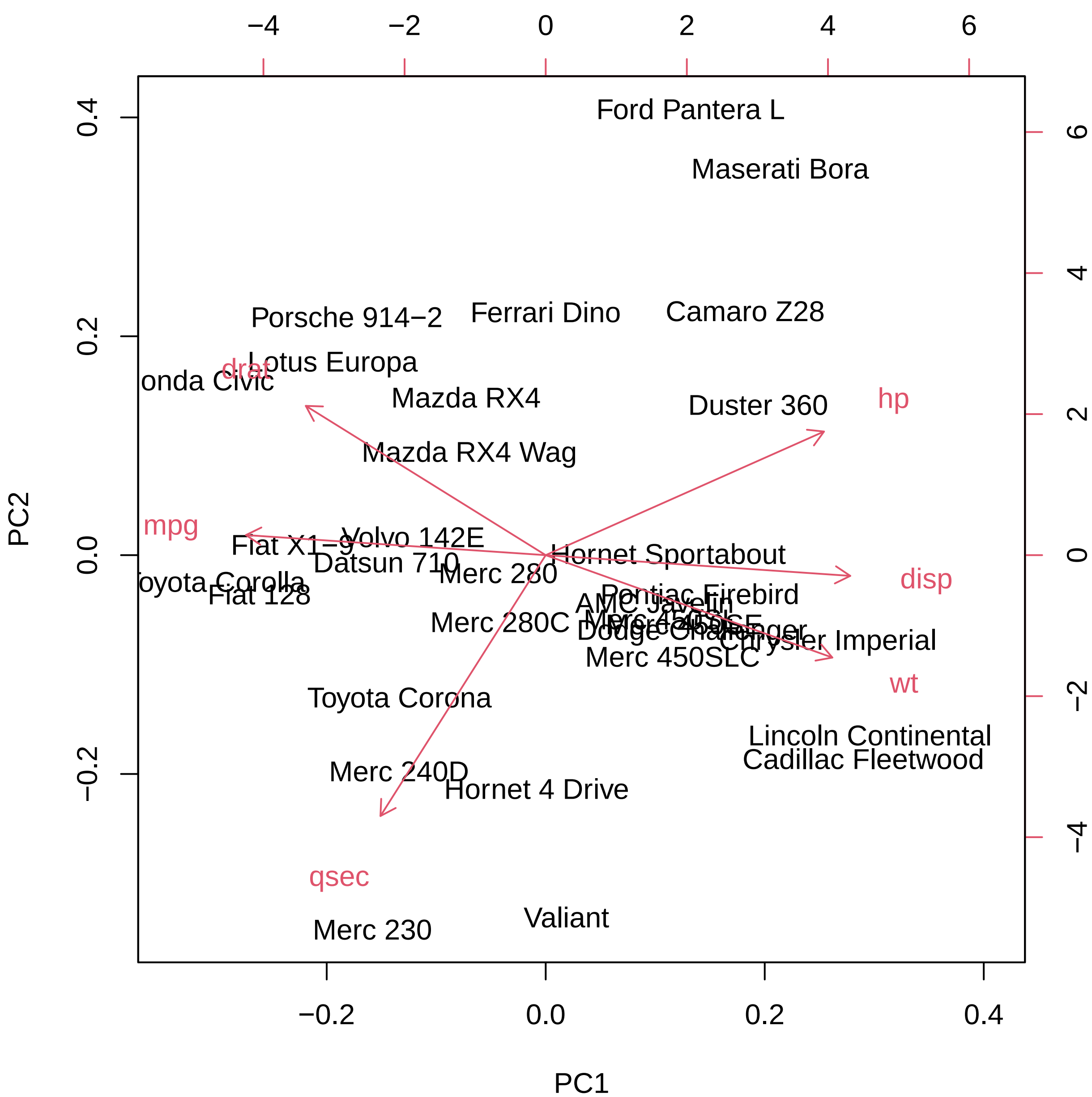


5. Narysuj wykres osypiska i zaproponuj optymalną liczbę składowych głównych w oparciu o trzy kryteria.

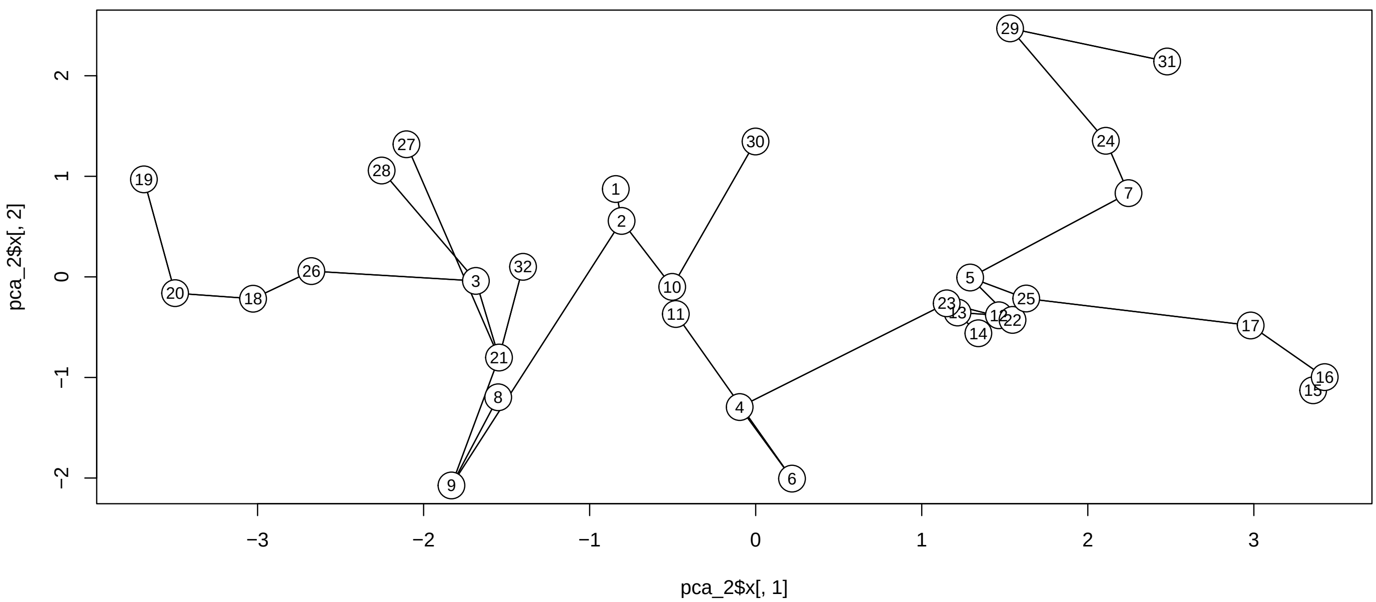



```
## 2 lub 3
```

6. Przedstaw samochody w układzie dwóch pierwszych składowych głównych (dokładniej narysuj biplot i dokonaj jego interpretacji).



7. Przedstaw samochody za pomocą minimalnego drzewa rozpinającego.



8. Jak bardzo będą różniły się wyniki, jeśli nie wykonamy skalowania danych?

Ad. 1.

```
## Standard deviations (1, .., p=6):
## [1] 310.0207637  40.8471739  15.7168252   2.1068823   0.3894500   0.2969505
##
## Rotation (n x k) = (6 x 6):
##              PC1          PC2         PC3          PC4          PC5
## mpg  -0.05193468  0.121255352 -0.82446804  0.540735371 -0.064362234
## disp -0.85253108 -0.522102198 -0.00915689  0.022137483  0.001587345
## hp   -0.51734213  0.841835388  0.15361995 -0.004990023 -0.006795464
## drat -0.01010286  0.021298587 -0.10869056 -0.033506518  0.982931599
## wt   -0.01067910  0.001369032 -0.04162846 -0.192177061  0.129755288
## qsec -0.05132793  0.059700171 -0.53199901 -0.817945952 -0.113215907
##                PC6
## mpg   0.0794678281
## disp -0.0048593900
## hp   -0.0003699391
## drat -0.1426655136
## wt    0.9717935462
## qsec -0.1700734209
```

Ad. 2.

```
## Importance of components:
##                             PC1      PC2      PC3     PC4    PC5   PC6
## Standard deviation     310.0208 40.84717 15.71683 2.10688 0.3895 0.297
## Proportion of Variance   0.9804  0.01702  0.00252 0.00005 0.0000 0.000
## Cumulative Proportion    0.9804  0.99743  0.99995 1.00000 1.0000 1.000
```

Ad. 3.

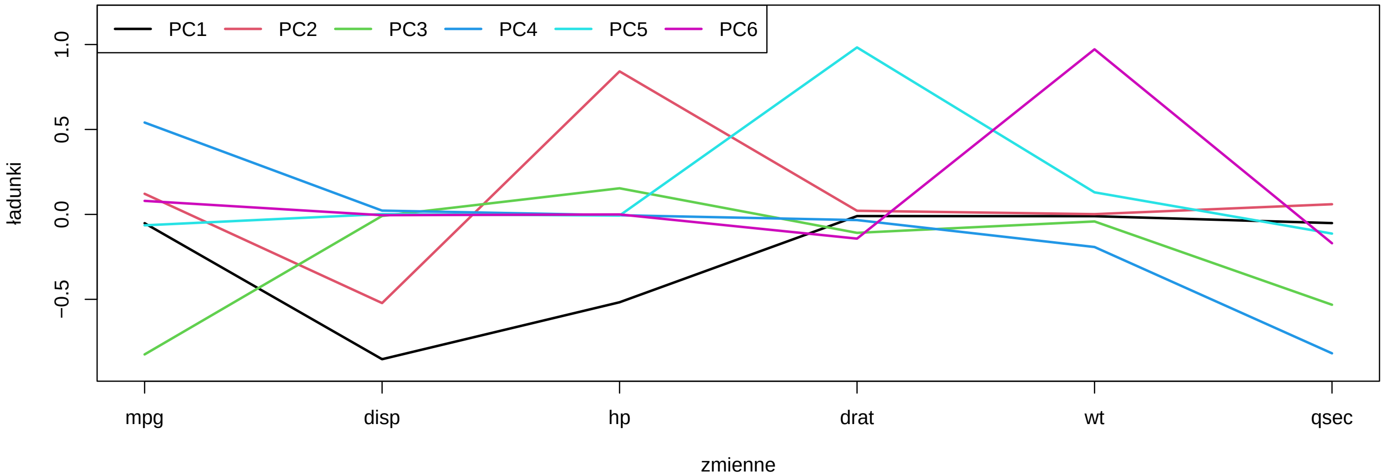
```
##                          PC1       PC2        PC3        PC4         PC5
## Mazda RX4          -195.3155  12.68122 -11.170400  0.2509678  0.46472555
## Mazda RX4 Wag      -195.3469  12.71500 -11.478935 -0.2560871  0.43441224
## Datsun 710         -142.3892  25.86447 -15.915699 -1.5412826  0.05036709
```

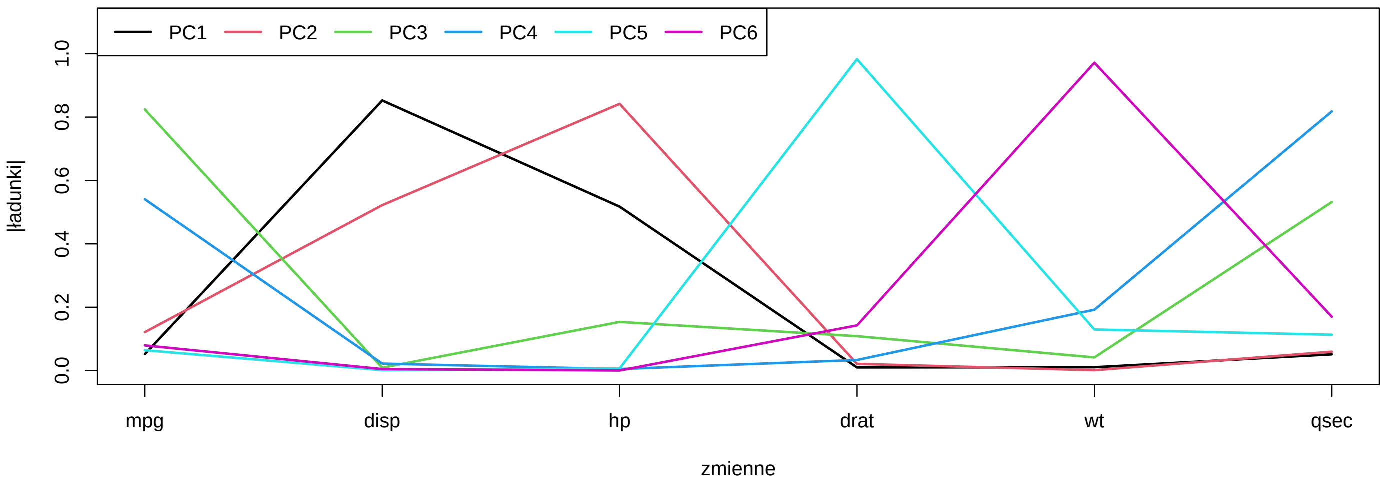
```
## Hornet 4 Drive    -279.0353 -38.27504 -13.918563  0.1123864 -0.47164240
## Hornet Sportabout -399.3594 -37.28023  -1.370742  2.5199166 -0.20567766
## Valiant           -248.1831 -25.61490 -12.054118 -3.0519863 -0.64870858
##                          PC6
## Mazda RX4          0.04092377
## Mazda RX4 Wag      0.19349001
## Datsun 710        -0.20711953
## Hornet 4 Drive    -0.21512523
## Hornet Sportabout -0.32914754
## Valiant           -0.16407438
```

```
## ...
```

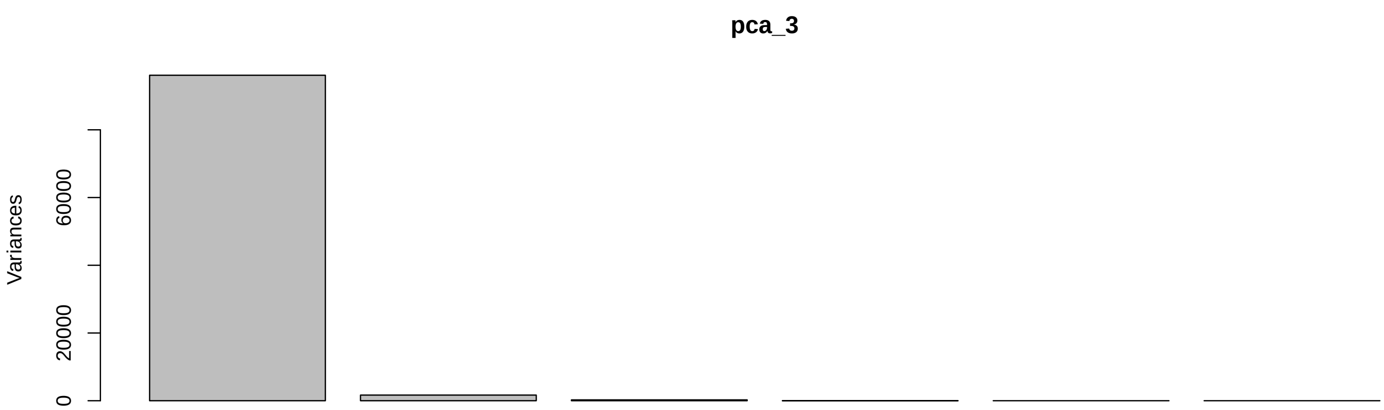
Ad. 4.

```
##              PC1          PC2         PC3          PC4          PC5
## mpg  -0.05193468  0.121255352 -0.82446804  0.540735371 -0.064362234
## disp -0.85253108 -0.522102198 -0.00915689  0.022137483  0.001587345
## hp   -0.51734213  0.841835388  0.15361995 -0.004990023 -0.006795464
## drat -0.01010286  0.021298587 -0.10869056 -0.033506518  0.982931599
## wt   -0.01067910  0.001369032 -0.04162846 -0.192177061  0.129755288
## qsec -0.05132793  0.059700171 -0.53199901 -0.817945952 -0.113215907
##                PC6
## mpg   0.0794678281
## disp -0.0048593900
## hp   -0.0003699391
## drat -0.1426655136
## wt    0.9717935462
## qsec -0.1700734209
```


PC1
PC2
PC3
PC4
PC5
PC6
ładunki
1.0
0.5
0.0
−0.5
mpg
disp
hp
drat
wt
qsec
zmienne

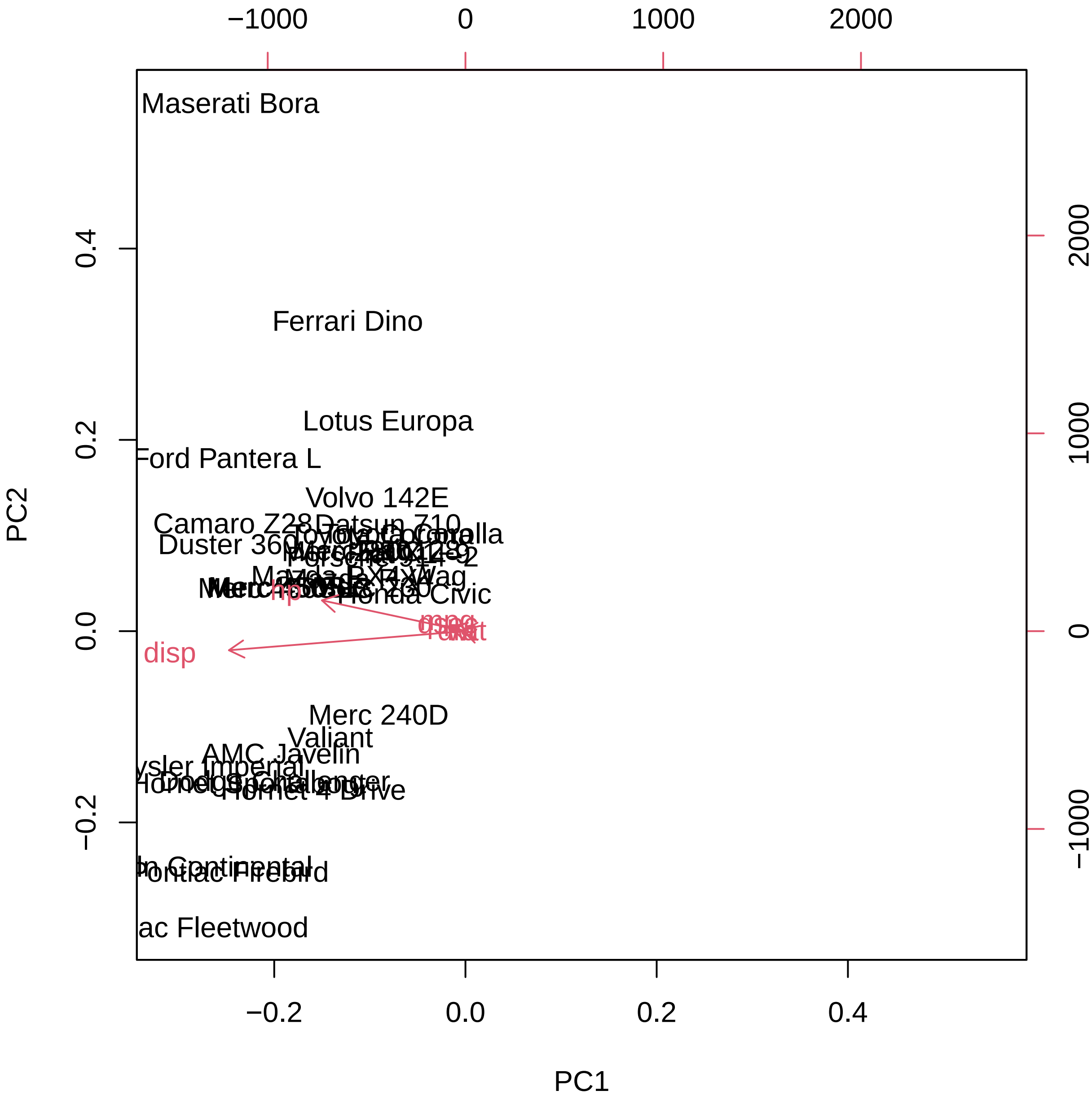

PC1
PC2
PC3
PC4
PC5
PC6
|ładunki|
1.0
0.8
0.6
0.4
0.2
0.0
mpg
disp
hp
drat
wt
qsec
zmienne

Ad. 5.


pca_3
Variances
60000
20000
0

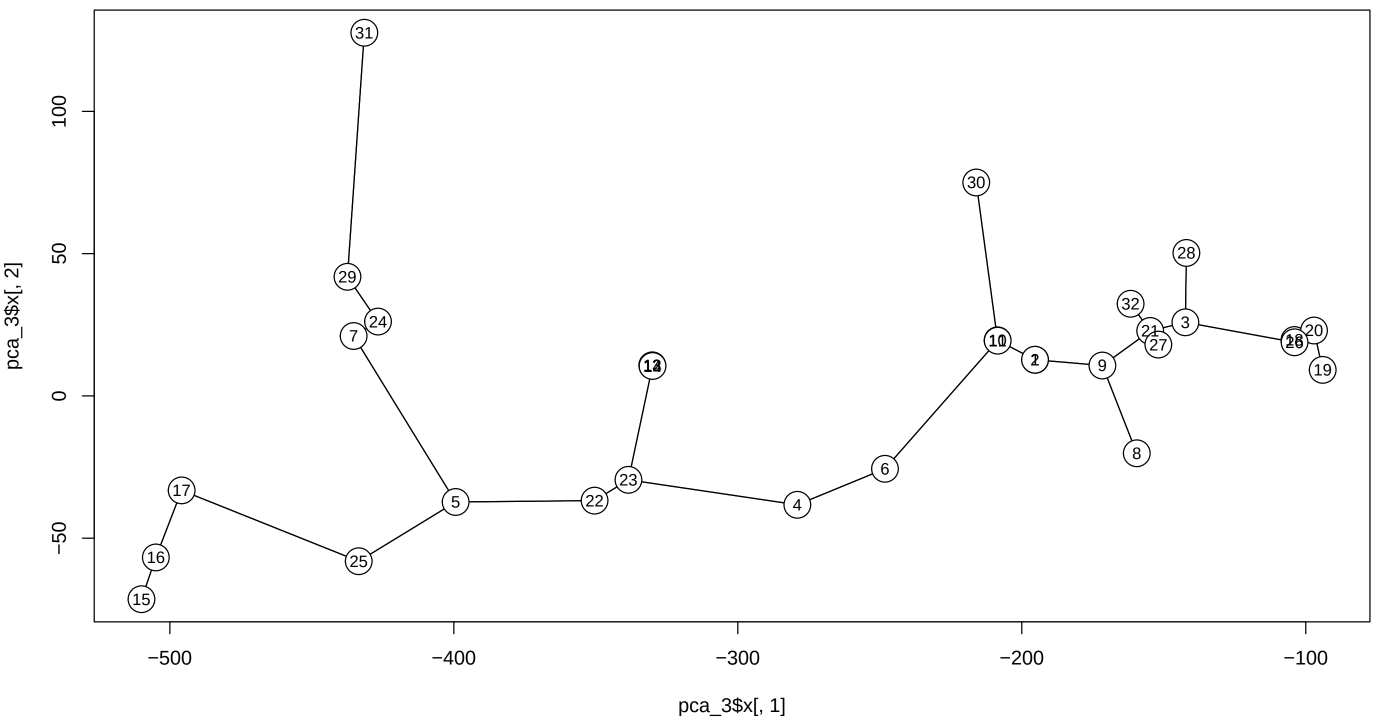
```
## 1
```

Ad. 6.



−1000
0
1000
2000
Maserati Bora
Ferrari Dino
Lotus Europa
Ford Pantera L
Volvo 142E
Honda Civic
disp
hp
Merc 240D
Valiant
AMC Javelin
Hornet 4 Drive
Pontiac Firebird
0.4
0.2
0.0
−0.2
−1000
PC2
−0.2
0.0
0.2
0.4
PC1

Ad. 7.



9 Analiza skupień

- Analiza skupień jest narzędziem analizy danych służącym do grupowania n obiektów, opisanych za pomocą wektora p-cech, w K niepustych, rozłącznych i możliwie „jednorodnych” grup (skupień).
- Obiekty należące do danego skupienia powinny być „podobne” do siebie, a obiekty należące do różnych skupień powinny być z kolei możliwie mocno „niepodobne” do siebie.
- Głównym celem tej analizy jest wykrycie w zbiorze danych, tzw. „naturalnych” skupień, czyli skupień, które dają się w sensowny sposób interpretować.

9.1 Algorytm zachłanny

- Zwróćmy uwagę, że pod tym terminem kryje się szereg różnych algorytmów.
- Koncepcyjnie, najprostszym byłby następujący: Ustalamy liczbę skupień K oraz kryterium optymalnego podziału obiektów. Przeszukujemy wszystkie możliwe podziały n obiektów na K skupień, wybierając najlepszy podział ze względu na przyjęte kryterium optymalności.
- Bezpośrednie sprawdzenie wszystkich możliwych podziałów jest jednak, nawet przy niewielkim n, praktycznie niemożliwe. Ich liczba bowiem jest równa

$$\frac{1}{K!}\sum_{k=1}^{K}(-1)^{K-k}\binom{K}{k}k^n$$

  i np. dla $n = 100$ obiektów i $K = 4$ skupień jest rzędu 10^{58}.
- Dodatkowym problemem jest wybór końcowej liczby skupień. Często bardzo pomocne są w tym przypadku metody wizualizacji danych. W sytuacji, gdy liczba cech jest większa niż trzy, zmuszeni jesteśmy dodatkowo do redukcji wymiaru danych. W tym celu korzystamy zazwyczaj z techniki analizy składowych głównych.
- Istnieją również inne, bardziej automatyczne, kryteria wyboru końcowej liczby skupień.

9.2 Algorytmy hierarchiczne

- Najprostszą i zarazem najczęściej używaną metodą analizy skupień jest **metoda hierarchiczna**.
- Wspólną cechą krokowych algorytmów tej metody jest wyznaczanie skupień poprzez łączenie (aglomerację) powstałych, w poprzednich krokach algorytmu, mniejszych skupień.
- Inne wersje tej metody zamiast idei łączenia skupień, bazują na pomyśle ich dzielenia.
- Podstawą wszystkich algorytmów tej metody jest odpowiednie określenie miary niepodobieństwa obiektów. Miary niepodobieństwa, to semi-metryki (a często również metryki) na przestrzeni próby $\mathcal{X}$.

Definicja. Funkcję $\rho : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ nazywamy **miarą niepodobieństwa** jeśli:

1. $\rho(\mathbf{x}, \mathbf{y}) \geq 0$,
2. $\rho(\mathbf{x}, \mathbf{y}) = 0$ wtedy i tyko wtedy, gdy $\mathbf{x} = \mathbf{y}$,
3. $\rho(\mathbf{y}, \mathbf{x}) = \rho(\mathbf{x}, \mathbf{y})$.

- Określona w ten sposób miara jest semi-metryką na przestrzeni próby.
- Jak widać nie musi ona być (choć często jest) metryką, tzn. nie musi spełniać warunku trójkąta:

$$\rho(\mathbf{x}, \mathbf{y}) \leq \rho(\mathbf{x}, \mathbf{z}) + \rho(\mathbf{z}, \mathbf{y}).$$

Nierówność trójkąta nie jest nam potrzebna do określenia kolejności odległości punktów od $\mathbf{x}$, ponieważ nie interesują nas odległości pomiędzy pozostałymi punktami.

- Wybór miary niepodobieństwa obiektów jest arbitralny i zależy głównie od charakteru danych.
- Dla danych ilościowych, jako miarę niepodobieństwa pomiędzy obiektami używa się często zwykłą **odległość (metrykę) euklidesową**

$$\rho_1(\mathbf{x}, \mathbf{y}) = ((\mathbf{x} - \mathbf{y})'(\mathbf{x} - \mathbf{y}))^{1/2} = \left(\sum_{i=1}^{p} (x_i - y_i)^2 \right)^{1/2}$$

lub jej kwadrat

$$\rho_2(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})'(\mathbf{x} - \mathbf{y}) = \sum_{i=1}^{p} (x_i - y_i)^2.$$

- Zwróćmy uwagę, że druga miara nie jest metryką, ponieważ nie jest dla niej spełniony warunek trójkąta.
- Jeżeli cechy opisujące obiekty wyrażone są w różnych jednostkach, to w celu zniwelowania ich wpływu możemy zastosować **ważoną odległość euklidesową**

$$\rho_3(\mathbf{x}, \mathbf{y}) = ((\mathbf{x} - \mathbf{y})'\mathbf{W}^{-1}(\mathbf{x} - \mathbf{y}))^{1/2} = \left(\sum_{i=1}^{p} \frac{1}{w_i^2} (x_i - y_i)^2 \right)^{1/2},$$

gdzie $\mathbf{W} = \mathrm{diag}\{w_1^2, \dots, w_p^2\}$, a wagi w_i są odchyleniami standardowymi poszczególnych cech.

- Aby miara uwzględniała również korelacje pomiędzy cechami stosujemy jako miarę niepodobieństwa **odległość Mahalanobisa**

$$\rho_4(\mathbf{x}, \mathbf{y}) = ((\mathbf{x} - \mathbf{y})'\mathbf{S}^{-1}(\mathbf{x} - \mathbf{y}))^{1/2},$$

gdzie $\mathbf{S}$ jest estymatorem macierzy kowariancji.

- Rzadziej stosuje się również inne miary niepodobieństwa:
  - **Odległość miejska (taksówkowa, manhatańska)**

$$\rho_5(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{p} |x_i - y_i|.$$

- Odległość ta, tak samo jak odległość euklidesowa, jest szczególnym przypadkiem odległości Minkowskiego w przestrzeni $\mathbb{R}^p$ danej wzorem:

$$\rho(\mathbf{x},\mathbf{y}) = \left(\sum_{i=1}^{p} |x_i - y_i|^q\right)^{1/q}.$$

- W przypadku danych jakościowych, możemy w naturalny sposób zdefiniować miarę niepodobieństwa obiektów jako

$$\rho_{11}(\mathbf{x},\mathbf{y}) = \frac{1}{p}\sum_{i=1}^{p} I(x_i \neq y_i).$$

Miara ta nazywana jest **współczynnikiem Sneatha**.

Algorytm aglomeracyjny

1. W pierwszym kroku każdy z obiektów tworzy oddzielne skupienie. Zatem skupień tych jest n.
2. W kroku drugim w jedno skupienie połączone zostają dwa najbardziej podobne do siebie obiekty (w sensie wybranej miary niepodobieństwa obiektów). Otrzymujemy zatem $n-1$ skupień.
3. Postępując analogicznie, tzn. łącząc (wiążąc) ze sobą skupienia złożone z najbardziej podobnych do siebie obiektów, w każdym następnym kroku, liczba skupień maleje o jeden.
4. Obliczenia prowadzimy do momentu uzyskania zadeklarowanej, końcowej liczby skupień K lub do połączenia wszystkich obiektów w jedno skupienie.

Dendrogram

- Graficzną ilustracją algorytmu jest **dendrogram**, czyli drzewo binarne, którego węzły reprezentują skupienia, a liście obiekty. Liście są na poziomie zerowym, a węzły na wysokości odpowiadającej mierze niepodobieństwa pomiędzy skupieniami reprezentowanymi przez węzły potomki.

Metody wiązania skupień

- Algorytm ten wykorzystuje nie tylko miary niepodobieństwa pomiędzy obiektami, potrzebne są nam również metody wiązania skupień.
- Niech R i S oznaczają skupienia, a $\rho(R,S)$ oznacza miarę niepodobieństwa pomiędzy nimi.
- Poniżej podano trzy najczęściej wykorzystywane sposoby jej określenia.
  - Metoda pojedynczego wiązania (najbliższego sąsiedztwa) - miara niepodobieństwa pomiędzy dwoma skupieniami jest określona jako najmniejsza miara niepodobieństwa między dwoma obiektami należącymi do różnych skupień, tzn.

$$\rho(R,S) = \min_{i\in R, j\in S} \rho(\mathbf{X}_i, \mathbf{X}_j).$$

    Zastosowanie tego typu odległości prowadzi do tworzenia wydłużonych skupień, tzw. łańcuchów. Pozwala na wykrycie obserwacji odstających, nie należących do żadnej z grup, i warto przeprowadzić klasyfikację za jej pomocą na samym początku, aby wyeliminować takie obserwacje i przejść bez nich do właściwej części analizy.
  - Metoda pełnego wiązania (najdalszego sąsiedztwa) - miara niepodobieństwa pomiędzy dwoma skupieniami jest określona jako największa miara niepodobieństwa między dwoma obiektami należącymi do różnych skupień, tzn.

$$\rho(R,S) = \max_{i\in R, j\in S} \rho(\mathbf{X}_i, \mathbf{X}_j).$$

    Metoda ta jest przeciwieństwem metody pojedynczego wiązania. Jej zastosowanie prowadzi do tworzenia zwartych skupień o małej średnicy.

– Metoda średniego wiązania - miara niepodobieństwa pomiędzy dwoma skupieniami jest określona jako średnia miara niepodobieństwa między wszystkimi parami obiektów należących do różnych skupień, tzn.

$$\rho(R,S) = \frac{1}{n_R n_S} \sum_{i \in R} \sum_{j \in S} \rho(\mathbf{X}_i, \mathbf{X}_j),$$

gdzie n_R i n_S są liczbami obiektów wchodzących w skład skupień R i S odpowiednio. Metoda ta jest swoistym kompromisem pomiędzy metodami pojedynczego i pełnego wiązania. Ma ona jednak zasadniczą wadę. W odróżnieniu od dwóch poprzednich wykorzystywana w niej miara niepodobieństwa nie jest niezmiennicza ze względu na monotoniczne przekształcenia miar niepodobieństwa pomiędzy obiektami.

- Inne metody wiązania skupień
  – W przypadku gdy liczebności skupień są zdecydowanie różne, zamiast metodą średniego wiązania możemy posługiwać się jej ważonym odpowiednikiem. Wagami są wtedy liczebności poszczególnych skupień.
  – Inna popularna metoda wiązania skupień pochodzi od Warda (1963). Do obliczania miary niepodobieństwa pomiędzy skupieniami wykorzystuje on podejście analizy wariancji (minimalizacja sumy kwadratów odchyleń dowolnych dwóch skupień (wariancji wewnątrz grupowej), które mogą zostać uformowane na każdym etapie). Metoda daje bardzo dobre wyniki (grupy bardzo homogeniczne), jednak ma skłonność do tworzenia skupień o małej wielkości i o podobnych rozmiarach. Często nie jest też w stanie zidentyfikować grup o szerokim zakresie zmienności poszczególnych cech oraz niewielkich grup.
- Algorytm aglomeracyjny jest bardzo szybki i uniwersalny w tym sensie, że może być on stosowany zarówno do danych ilościowych jak i jakościowych. Wykorzystuje on jedynie miary niepodobieństwa pomiędzy obiektami oraz pomiędzy skupieniami.
- Należy podkreślić zasadniczy wpływ wybranej miary niepodobieństwa na uzyskane w końcowym efekcie skupienia.
- Do ustalenia końcowej liczby skupień wykorzystać możemy wykresy rozrzutu (przy wielu wymiarach w układzie dwóch pierwszych składowych głównych). Pomocny może być także dendrogram. Ustalamy wtedy progową wartość miary niepodobieństwa pomiędzy skupieniami, po przekroczeniu której zatrzymany zostaje proces ich dalszego łączenia.

9.3 Metoda K-średnich

- Najbardziej popularnym, niehierarchicznym algorytmem analizy skupień jest **algorytm K-średnich**.
- Przyporządkowanie n obiektów do zadanej liczby skupień K, odbywa się niezależnie dla każdej wartości K (nie bazując na wyznaczonych wcześniej mniejszych lub większych skupieniach).
- Niech C_K oznacza funkcję, która każdemu obiektowi (dokładnie jego numerowi), przyporządkowuje numer skupienia do którego jest on przyporządkowany (przy podziale na K skupień).
- Zakładamy, że wszystkie cechy są ilościowe o wartościach rzeczywistych (przestrzeń próby to $\mathbb{R}^p$).
- Główną ideą metody K-średnich jest taka alokacja obiektów, która minimalizuje zmienność wewnątrz powstałych skupień, a co za tym idzie maksymalizuje zmienność pomiędzy skupieniami.
- Dla ustalonej funkcji C_K, przez $W(C_K)$ i $B(C_K)$ oznaczmy macierze zmienności odpowiednio wewnątrz i pomiędzy skupieniami.
- Niech $\bar{\mathbf{X}}_k = \frac{1}{n_k} \sum_{C_K(i)=k} \mathbf{X}_i$ oznacza wektor średnich k-tego skupienia, $\bar{\mathbf{X}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i$ średnią ogólną,

a n_k jest liczebnością k-tego skupienia.

$$W(C_K) = \sum_{k=1}^{K} \sum_{C_K(i)=k} (\mathbf{X}_i - \bar{\mathbf{X}}_k)(\mathbf{X}_i - \bar{\mathbf{X}}_k)',$$

$$B(C_K) = \sum_{k=1}^{K} n_k(\bar{\mathbf{X}}_k - \bar{\mathbf{X}})(\bar{\mathbf{X}}_k - \bar{\mathbf{X}})'.$$

- Znana z analizy wariancji, zależność opisuje związek pomiędzy tymi macierzami:

$$T = W(C_K) + B(C_K),$$

  gdzie

$$T = \sum_{i=1}^{n} (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})'$$

  jest niezależną od dokonanego podziału na skupienia macierzą zmienności całkowitej.
- Powszechnie stosowane algorytmy metody K-średnich minimalizują ślad macierzy $W(C_K)$.

Algorytm metody K-średnich

1. W losowy sposób rozmieszczamy n obiektów w K skupieniach. Niech funkcja $C_K^{(1)}$ opisuje to rozmieszczenie.
2. Dla każdego z K skupień obliczamy wektory średnich $\bar{\mathbf{X}}_k$, $k = 1, 2, \ldots, K$.
3. Rozmieszczamy ponownie obiekty w K skupieniach, w taki sposób że

$$C_K^{(l)}(i) = \arg \min_{1 \le k \le K} \rho_2(\mathbf{X}_i, \bar{\mathbf{X}}_k).$$

4. Powtarzamy kroki drugi i trzeci aż do momentu, gdy przyporządkowanie obiektów do skupień pozostanie niezmienione, tzn. aż do momentu, gdy $C_K^{(l)} = C_K^{(l-1)}$.

- Istnieje wiele modyfikacji powyższego algorytmu. Przykładowo, losowe rozmieszczenie elementów w skupieniach (krok pierwszy algorytmu) zastąpione zostaje narzuconym podziałem, mającym na celu szybsze ustabilizowanie się algorytmu.
- Wszystkie wersje algorytmu K-średnich są zbieżne. Nie gwarantują one jednak zbieżności do optymalnego rozwiązania C_K^*. Niestety, w zależności od początkowego podziału, algorytm zbiega do zazwyczaj różnych lokalnie optymalnych rozwiązań. W związku z tym, aby uzyskać najlepszy podział, zaleca się często wielokrotne stosowanie tego algorytmu z różnymi, wstępnymi rozmieszczeniami obiektów.

Wybór K

- W literaturze znaleźć można wiele pomysłów na automatyczne wyznaczania końcowej liczby skupień. Jedna z nich zasługują na szczególną uwagę.
- Caliński i Harabasz (1974) zaproponowali aby końcową liczbę skupień wybierać w oparciu o wartości pseudo-statystyki F postaci:

$$CH(K) = \frac{\mathrm{tr}(B(C_K))/(K-1)}{\mathrm{tr}(W(C_K))/(n-K)}.$$

- Optymalną wartość K dobieramy tak, aby ją zmaksymalizować.

9.4 Metoda hierarchiczna, a niehierarchiczna

- Obie metody mają swoje wady i zalety.
- W przypadku metod hierarchicznych istnieje wiele algorytmów dających różne wyniki, z których nie jesteśmy w stanie określić, które rozwiązanie jest najlepsze. Poza tym nie ma możliwości korekty rozwiązania, obiekt raz przydzielony do klasy już w niej pozostaje. Ostatecznie metody hierarchiczne są mało wydajne w przypadku dużych zbiorów danych (duża czaso- i pamięciożerność).

- Główną wadą metod optymalizacyjnych jest konieczność zadania liczby klas z góry. Dodatkowo bardzo duże znaczenie ma wybór początkowych środków ciężkości.
- W praktyce często metoda hierarchiczna służy do wstępnej obróbki danych i wyznaczenia punktów startowych dla metody K-średnich (np. jako średnie w skupieniach).
- Analiza skupień nie jest odporna na zmiany skali, czyli jeśli różne zmienne mają różne skale, to te największe mogą zdominować odległości. Oczywiście może to być celowe (pewne zmienne są „ważniejsze” od innych). W ogólności jednak warto wykonać wpierw skalowanie danych.

9.5 Przykład 9

Przykład. Zbiór danych `USArrests` zawiera informacje dotyczące liczby morderstw, napadów, gwałtów przypadających na 100,000 osób w poszczególnych stanach USA w roku 1973 oraz procent ludności mieszkającej w miastach. Chcielibyśmy się dowiedzieć, czy i które stany są do siebie w pewien sposób zbliżone.

```
head(USArrests)
```

```
##            Murder Assault UrbanPop Rape
## Alabama      13.2     236       58 21.2
## Alaska       10.0     263       48 44.5
## Arizona       8.1     294       80 31.0
## Arkansas      8.8     190       50 19.5
## California    9.0     276       91 40.6
## Colorado      7.9     204       78 38.7
```

```
dim(USArrests)
```

```
## [1] 50  4
```

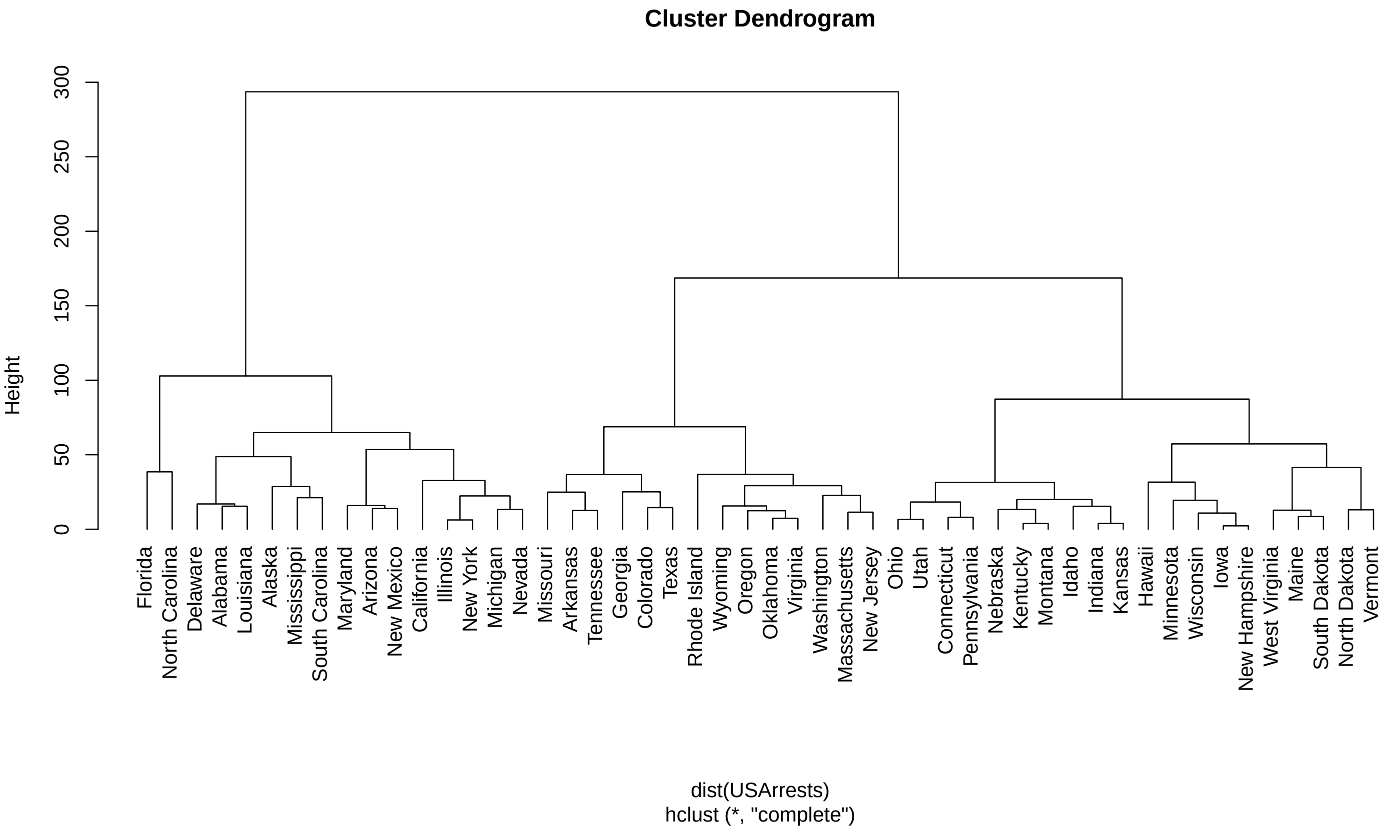
1. metoda hierarchiczna

```
(skupienia_1 <- hclust(dist(USArrests)))
```

```
##
## Call:
## hclust(d = dist(USArrests))
##
## Cluster method   : complete
## Distance         : euclidean
## Number of objects: 50
```

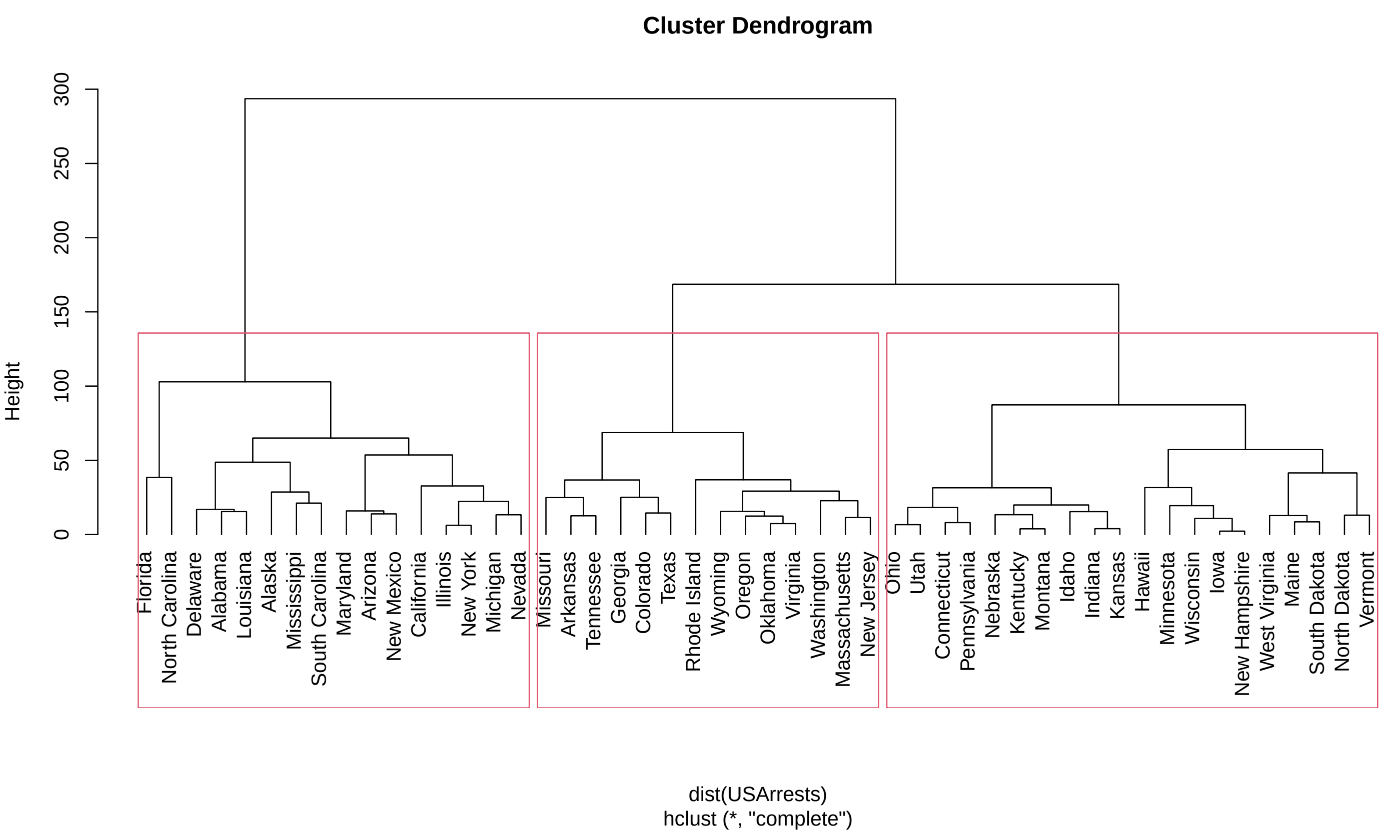
- dendrogram

```
plot(skupienia_1, hang = -1)
```



- automatyczny podział na skupienia i nanoszenie ich na dendrogram

```
plot(skupienia_1, hang = -1)
(podzial_1 <- rect.hclust(skupienia_1, k = 3))
```




```
## [[1]]
```

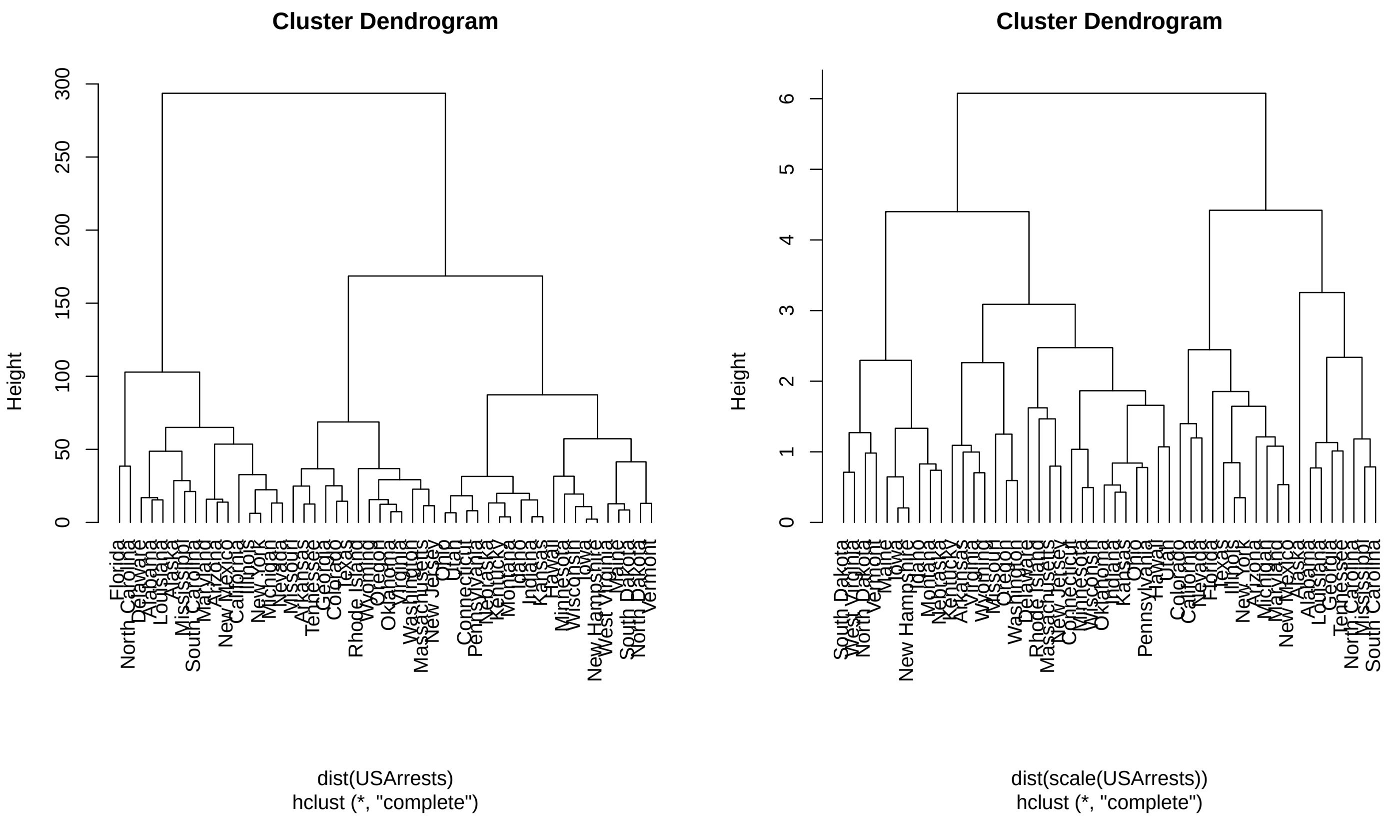
```
##        Alabama         Alaska        Arizona     California       Delaware
##              1              2              3              5              8
##        Florida       Illinois      Louisiana       Maryland       Michigan
##              9             13             18             20             22
##    Mississippi         Nevada     New Mexico       New York North Carolina
##             24             28             31             32             33
## South Carolina
##             40
##
## [[2]]
##      Arkansas      Colorado       Georgia Massachusetts      Missouri
##             4             6            10            21            25
##    New Jersey      Oklahoma        Oregon  Rhode Island     Tennessee
##            30            36            37            39            42
##         Texas      Virginia    Washington       Wyoming
##            43            46            47            50
##
## [[3]]
##   Connecticut        Hawaii         Idaho       Indiana          Iowa
##             7            11            12            14            15
##        Kansas      Kentucky         Maine     Minnesota       Montana
##            16            17            19            23            26
##      Nebraska New Hampshire  North Dakota          Ohio  Pennsylvania
##            27            29            34            35            38
##  South Dakota          Utah       Vermont West Virginia     Wisconsin
##            41            44            45            48            49
```

```
(podzial_2 <- cutree(skupienia_1, k = 3))
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              1              1              2              1
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              2              3              1              1              2
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              3              3              1              3              3
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              3              3              1              3              1
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              2              1              3              1              2
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              3              3              1              3              2
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              1              1              1              3              3
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              2              2              3              2              1
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              3              2              2              3              3
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              2              2              3              3              2
```

- zmiana skali ma wpływ na analizę skupień

```r
par(mfrow = c(1, 2))
plot(hclust(dist(USArrests)), hang = -1)
plot(hclust(dist(scale(USArrests))), hang = -1)
```




```r
par(mfrow = c(1, 1))
```

- parametry metody hierarchicznej

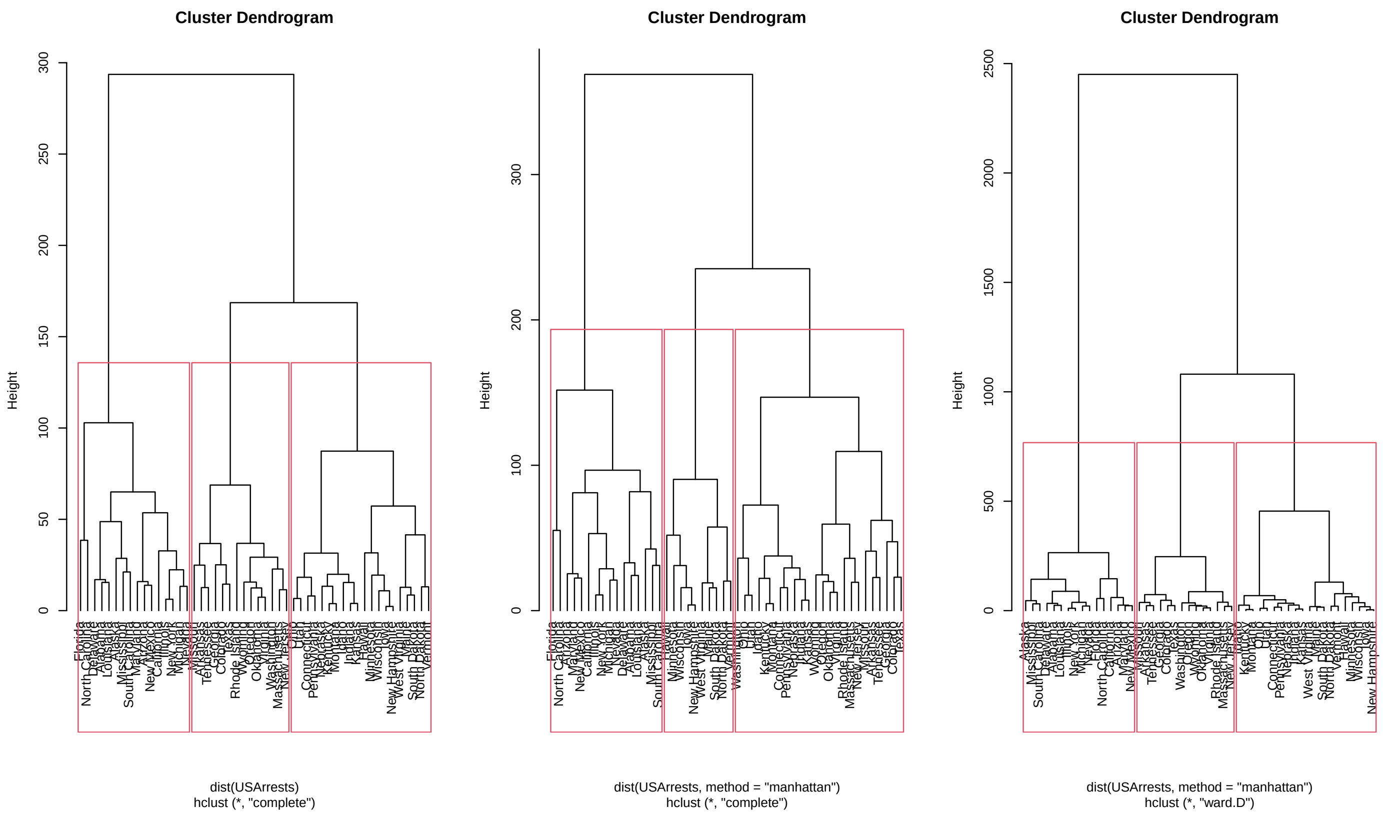
```r
# inna miara niepodobieństwa
(skupienia_2 <- hclust(dist(USArrests, method = 'manhattan')))
```

```
##
## Call:
## hclust(d = dist(USArrests, method = "manhattan"))
##
## Cluster method   : complete
## Distance         : manhattan
## Number of objects: 50
```

```r
# inna miara niepodobieństwa i inna metoda wiązania skupień
(skupienia_3 <- hclust(dist(USArrests, method = 'manhattan'), 'ward.D'))
```

```
##
## Call:
## hclust(d = dist(USArrests, method = "manhattan"), method = "ward.D")
##
## Cluster method   : ward.D
## Distance         : manhattan
## Number of objects: 50
```

```r
# porównianie dendrogramów
par(mfrow = c(1, 3))
plot(skupienia_1, hang = -1)
rect.hclust(skupienia_1, k = 3)
plot(skupienia_2, hang = -1)
rect.hclust(skupienia_2, k = 3)
plot(skupienia_3, hang = -1)
rect.hclust(skupienia_3, k = 3)
```




```r
par(mfrow = c(1, 1))
```

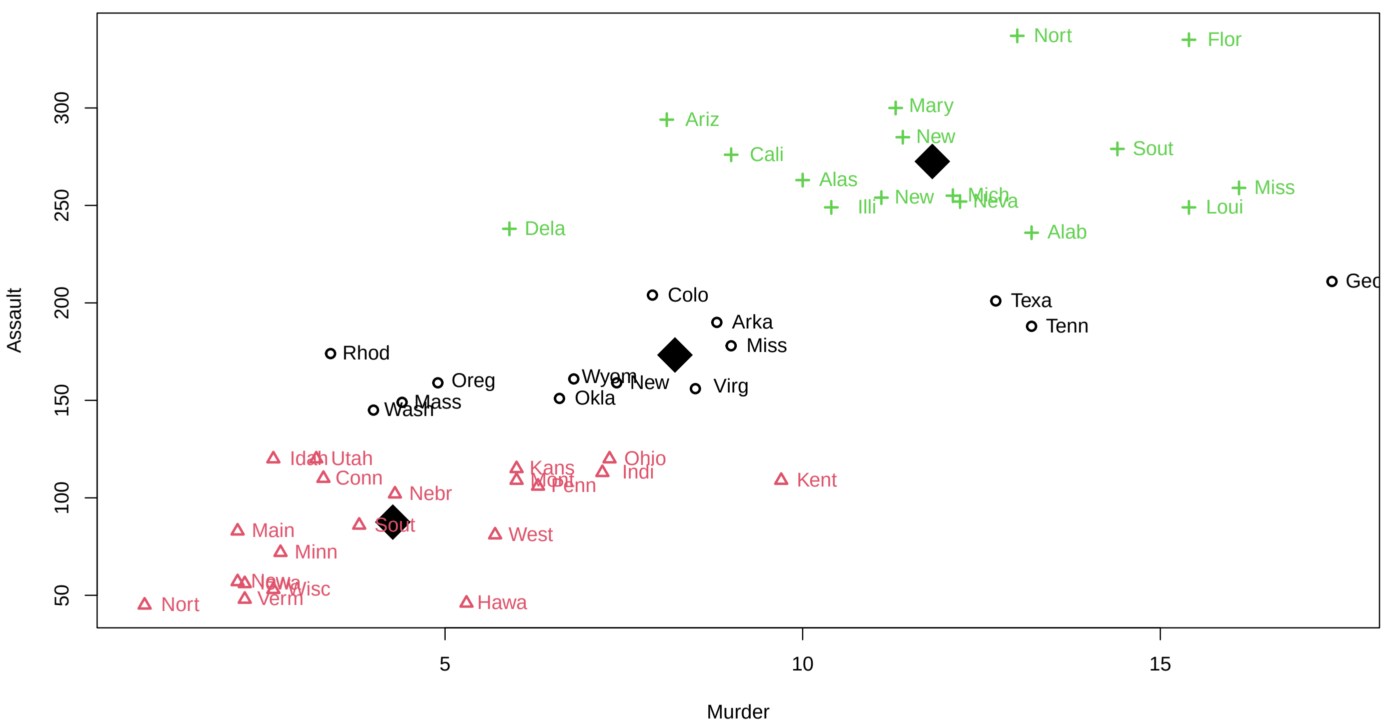
2. metoda K-średnich

```r
set.seed(1234)
(skupienia_4 <- kmeans(USArrests, centers = 3, nstart = 1000))
```

```
## K-means clustering with 3 clusters of sizes 14, 20, 16
##
## Cluster means:
##      Murder  Assault UrbanPop     Rape
## 1  8.214286 173.2857 70.64286 22.84286
## 2  4.270000  87.5500 59.75000 14.39000
## 3 11.812500 272.5625 68.31250 28.37500
##
## Clustering vector:
##        Alabama         Alaska        Arizona       Arkansas     California
##              3              3              3              1              3
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              1              2              3              3              1
##         Hawaii          Idaho       Illinois        Indiana           Iowa
```

```
##              2              2              3              2              2
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              2              2              3              2              3
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              1              3              2              3              1
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              2              2              3              2              1
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              3              3              3              2              2
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              1              1              2              1              3
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              2              1              1              2              2
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              1              1              2              2              1
##
## Within cluster sum of squares by cluster:
## [1]  9136.643 19263.760 19563.863
##  (between_SS / total_SS =  86.5 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
# wykres danych w układzie Murder-Assault z podziałem na
# otrzymane skupienia i centrami skupień
plot(USArrests[, 1:2], pch = skupienia_4$cluster,
     col = skupienia_4$cluster, lwd = 2)
points(skupienia_4$centers, pch = 18, cex = 4)
text(USArrests[, 1:2] + 0.5, substring(row.names(USArrests), 1, 4),
     col = skupienia_4$cluster)
```



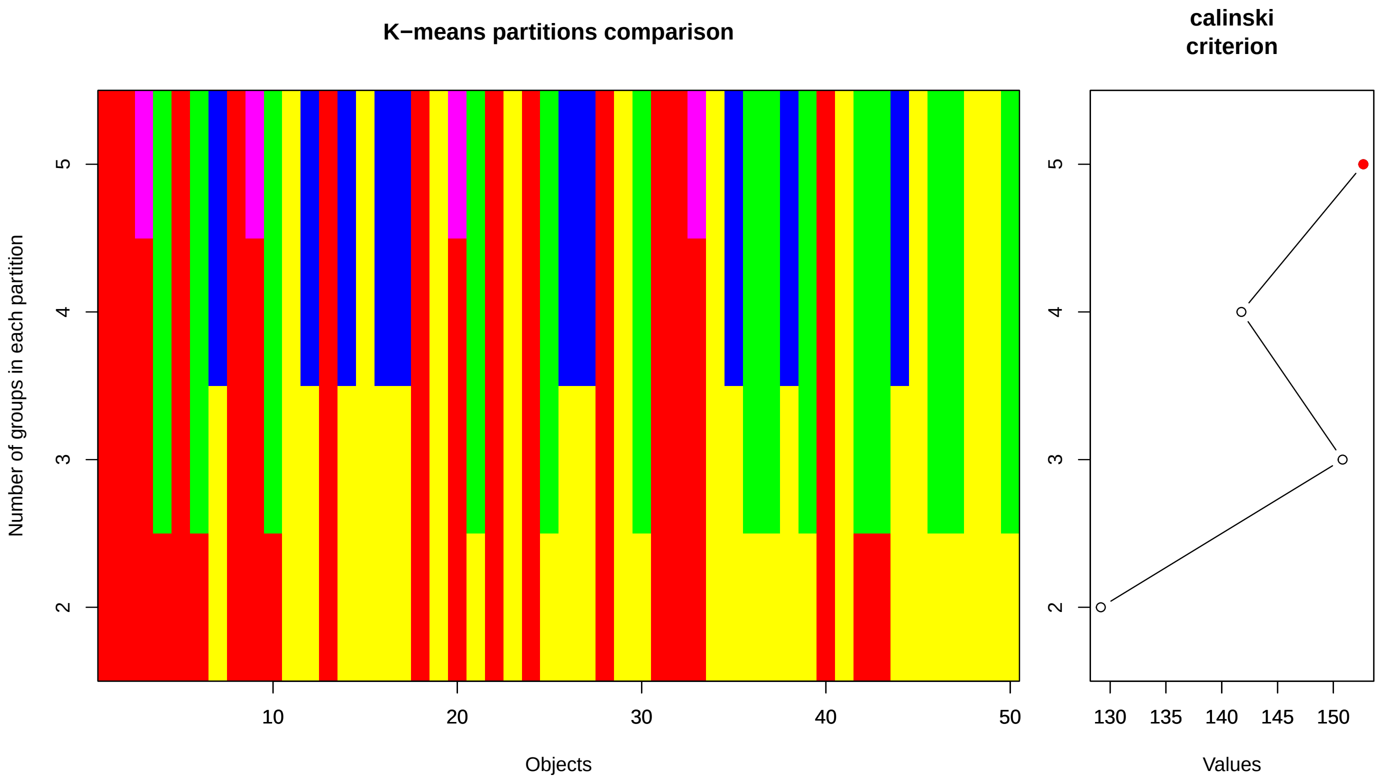
- metoda K-średnich z wyborem optymalnej liczby skupień poprzez indeks Calińskiego-Harabasza

```
library(vegan)
set.seed(1234)
(model <- cascadeKM(USArrests, 2, 5))
```

```
## $partition
##                2 groups 3 groups 4 groups 5 groups
## Alabama               1        1        2        1
## Alaska                1        1        2        1
## Arizona               1        1        2        3
## Arkansas              1        3        3        4
## California            1        1        2        1
## Colorado              1        3        3        4
## Connecticut           2        2        4        5
## Delaware              1        1        2        1
## Florida               1        1        2        3
## Georgia               1        3        3        4
## Hawaii                2        2        1        2
## Idaho                 2        2        4        5
## Illinois              1        1        2        1
## Indiana               2        2        4        5
## Iowa                  2        2        1        2
## Kansas                2        2        4        5
## Kentucky              2        2        4        5
## Louisiana             1        1        2        1
## Maine                 2        2        1        2
## Maryland              1        1        2        3
## Massachusetts         2        3        3        4
## Michigan              1        1        2        1
## Minnesota             2        2        1        2
## Mississippi           1        1        2        1
## Missouri              2        3        3        4
## Montana               2        2        4        5
## Nebraska              2        2        4        5
## Nevada                1        1        2        1
## New Hampshire         2        2        1        2
## New Jersey            2        3        3        4
## New Mexico            1        1        2        1
## New York              1        1        2        1
## North Carolina        1        1        2        3
## North Dakota          2        2        1        2
## Ohio                  2        2        4        5
## Oklahoma              2        3        3        4
## Oregon                2        3        3        4
## Pennsylvania          2        2        4        5
## Rhode Island          2        3        3        4
## South Carolina        1        1        2        1
## South Dakota          2        2        1        2
## Tennessee             1        3        3        4
## Texas                 1        3        3        4
## Utah                  2        2        4        5
```

```
## Vermont                  2        2        1        2
## Virginia                 2        3        3        4
## Washington               2        3        3        4
## West Virginia            2        2        1        2
## Wisconsin                2        2        1        2
## Wyoming                  2        3        3        4
##
## $results
##            2 groups   3 groups   4 groups   5 groups
## SSE      96399.0281 47964.2654 34728.6294 24417.0235
## calinski   129.1675   150.8274   141.7624   152.6864
##
## $criterion
## [1] "calinski"
##
## $size
##         2 groups 3 groups 4 groups 5 groups
## Group 1       21       16       10       12
## Group 2       29       20       16       10
## Group 3       NA       14       14        4
## Group 4       NA       NA       10       14
## Group 5       NA       NA       NA       10
##
## attr(,"class")
## [1] "cascadeKM"
```

```r
# wykres podziału na grupy
# (na osi x obserwacje, na osi y liczba skupień, kolory oznaczają skupienia)
# oraz wykres wartości indeksu Calińskiego-Harabasza dla
# poszczególnych liczb skupień (czerwona kropka oznacza
# optymalną liczbę skupień według tego kryterium)
plot(model)
```



9.6 Zadania 9

Zadanie 1. Plik wojewodztwa.txt zawiera dane dotyczące następujących cech województw w Polsce: współczynnik aktywności zawodowej (w %), wskaźnik zatrudnienia (w %), stopa bezrobocia rejestrowanego (w %), śmiertelność niemowląt (na 1000 urodzeń żywych), oczekiwana dalsza długość życia w momencie narodzin, gęstość zaludnienia (osoby na 1 km kwadratowy), produkt krajowy brutto na mieszkańca. Celem badania jest wyznaczenie podobieństw w województwach Polski.

```
##         wojewodztwo wspaktzaw wskzatr bezrobrej smniemowl lifeexp gestzaludn
## 1      Dolnoslaskie      54.3    41.5      20.6       6.9    74.6      144.8
## 2 KujawskoPomorskie      56.2    45.1      22.3       6.6    74.8      115.1
## 3         Lubelskie      56.1    48.7      17.0       7.3    74.9       86.8
## 4          Lubuskie      53.2    42.2      23.0       6.2    74.6       72.1
## 5           Lodzkie      55.9    45.7      17.9       6.1    73.5      141.5
## 6       Malopolskie      54.8    46.1      13.8       5.8    76.2      215.0
##   pkbcap
## 1  26620
## 2  22474
## 3  17591
## 4  23241
## 5  23666
## 6  21989
```

```
## ...
```

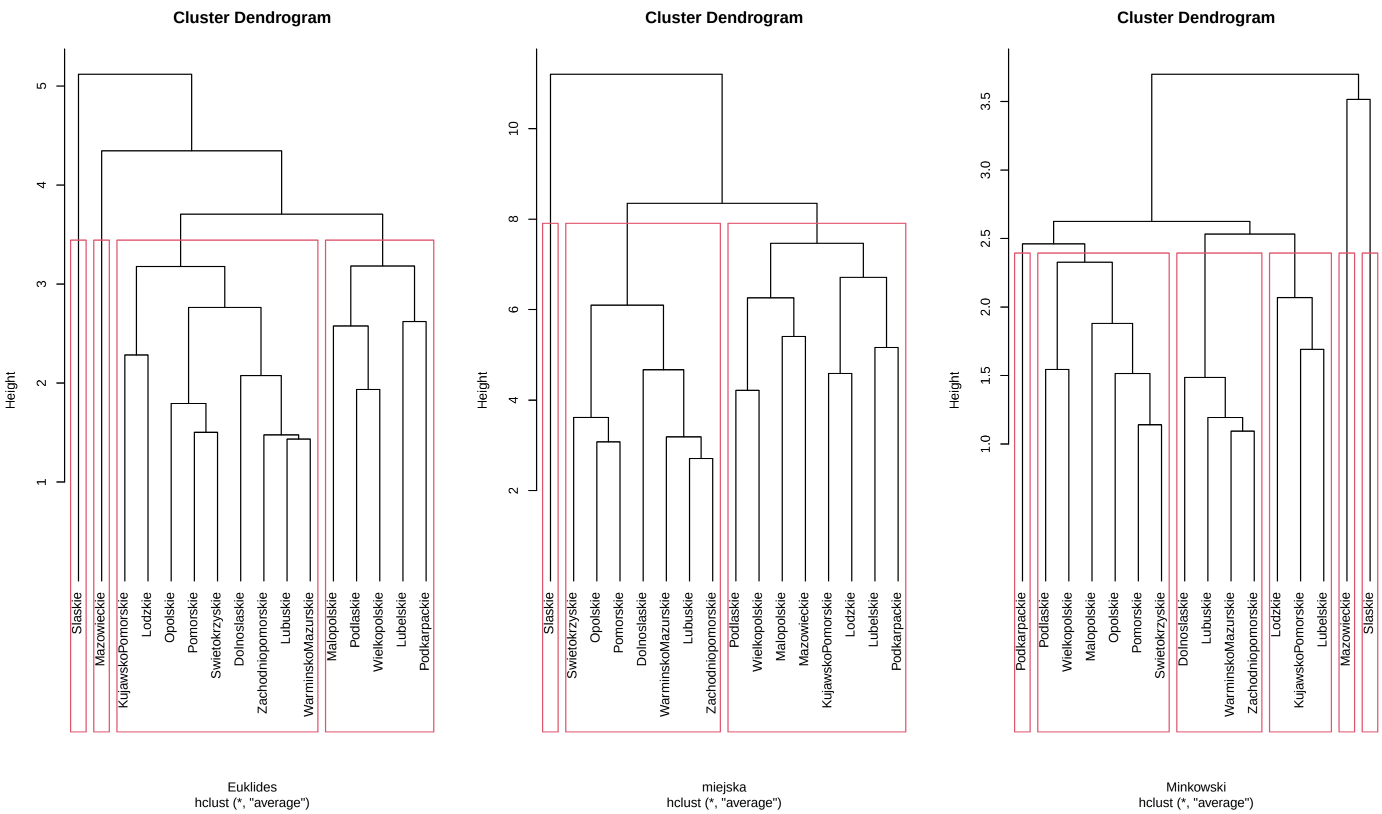
```
## [1] 16  8
```

1. Zauważmy, że jedna ze zmiennych przyjmuje znacznie większe wartości niż pozostałe zmienne. Czy w takim przypadku powinniśmy dokonać standaryzacji wszystkich zmiennych?

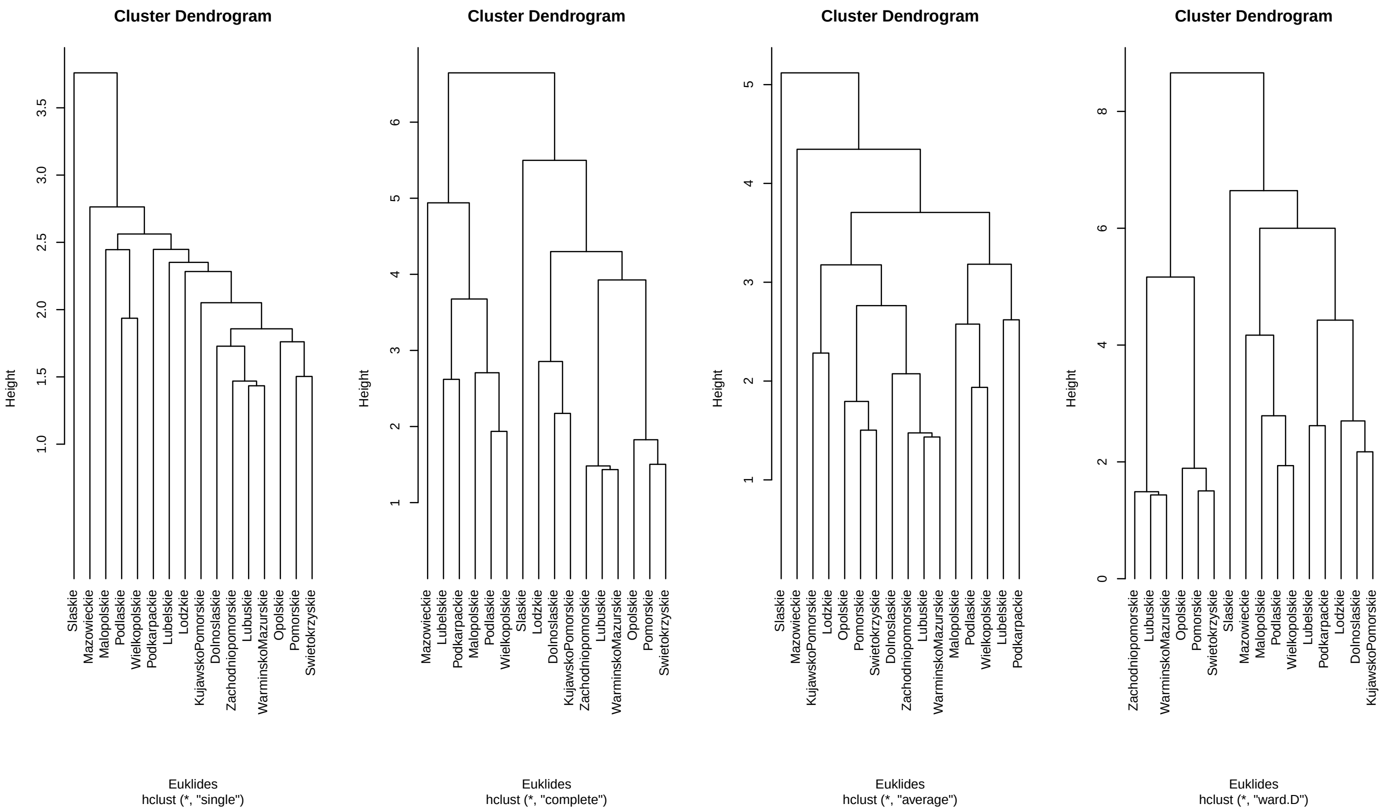
```
##         wojewodztwo   wspaktzaw   wskzatr  bezrobrej  smniemowl    lifeexp
```

```
## 1      Dolnoslaskie -0.08925698 -1.1278399  0.3969181  0.7935138 -0.7398300
## 2 KujawskoPomorskie  1.20284403  0.2602707  0.8170027  0.3703064 -0.4613058
## 3         Lubelskie  1.13483871  1.6483813 -0.4926727  1.3577903 -0.3220437
## 4          Lubuskie -0.83731545 -0.8579295  0.9899787 -0.1939700 -0.7398300
## 5           Lodzkie  0.99882808  0.4916225 -0.2702750 -0.3350392 -2.2717134
## 6        Malopolskie  0.25076960  0.6458570 -1.2834201 -0.7582465  1.4883640
##   gestzaludn      pkbcap
## 1  0.2031423  0.530348302
## 2 -0.1799261 -0.206399582
## 3 -0.5449373 -1.074113021
## 4 -0.7345368 -0.070103001
## 5  0.1605792  0.005419877
## 6  1.1085766 -0.292584513

## ...
```

2. Wykorzystując odległości euklidesową, miejską i Minkowskiego z potęgą cztery jako miary niepodobieństwa oraz metodę średniego wiązania skupień wykonaj hierarchiczną analizę skupień. Narysuj dendrogramy. Przy ich pomocy określ jaka liczba skupień wydaje się najbardziej sensowna. Zaznacz te skupienia na wykresie.

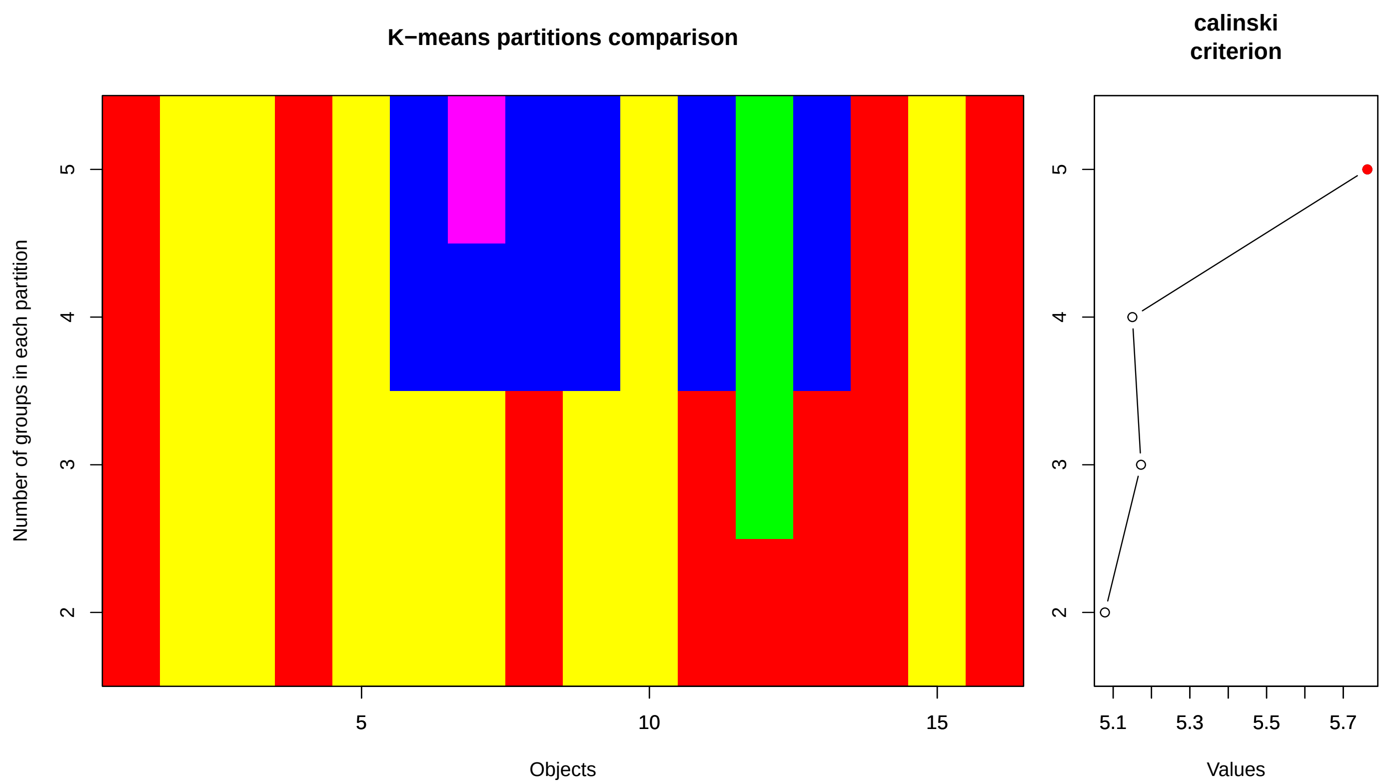


3. Wykorzystując odległość euklidesową jako miarę niepodobieństwa oraz metody pojedynczego, kompletnego, średniego wiązania skupień oraz metodę Warda łączenia skupień wykonaj hierarchiczną analizę skupień. Narysuj dendrogramy.



4. Jaką optymalną liczbę skupień proponuje indeks Calińskiego-Harabasza? Rozważ $K = 2, 3, 4, 5$.

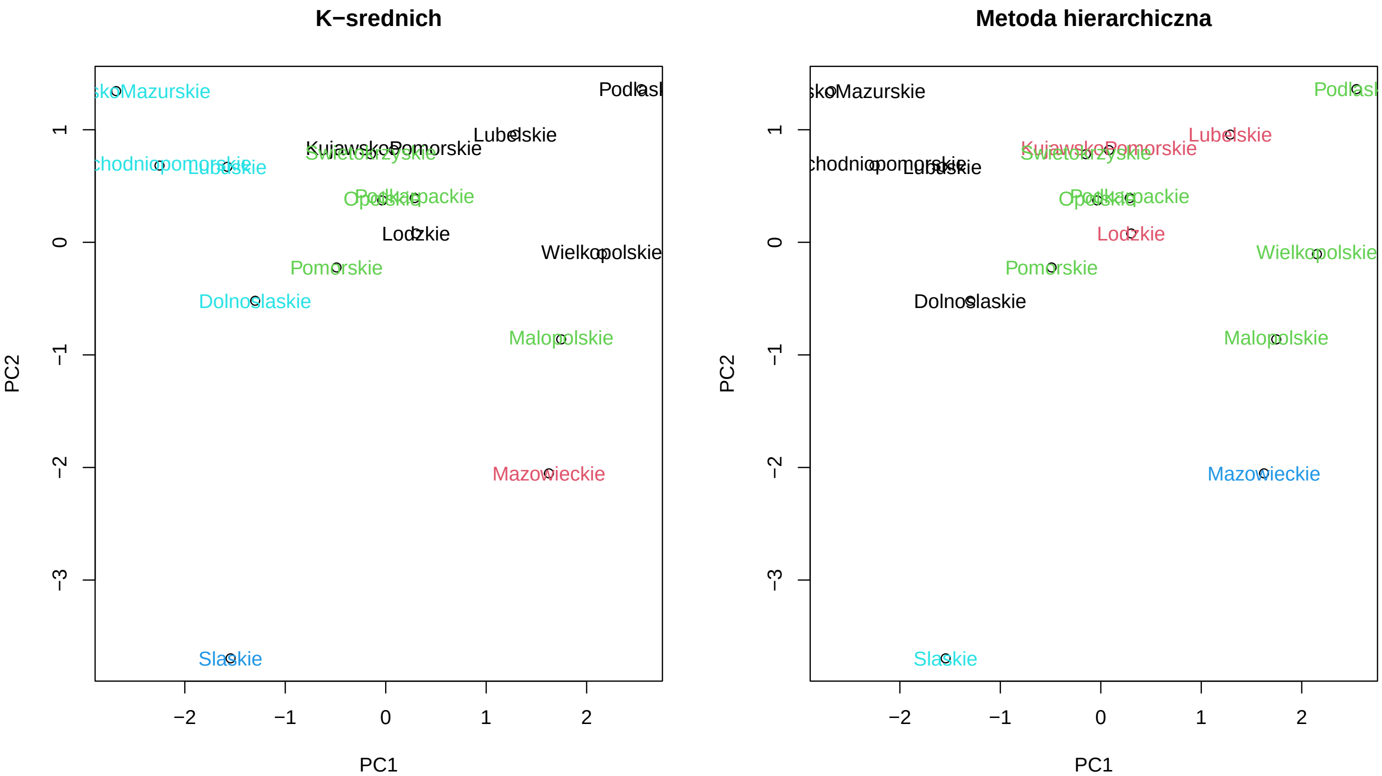
```
##            2 groups  3 groups  4 groups  5 groups
## SSE      77.049773 58.469889 45.900765 33.919492
## calinski  5.078577  5.172675  5.150174  5.762804
```



5. Wykonaj analizę skupień korzystając z metody K-średnich oraz hierarchicznej analizy skupień (odległość Minkowskiego z potęgą cztery, metoda średniego wiązania skupień) dla liczby skupień wyznaczonej

przez indeks Calińskiego-Harabasza. Przedstaw obserwacje w układzie dwóch pierwszych składowych głównych z podziałem na otrzymane skupienia.

```
## K-średnich

## [1] "KujawskoPomorskie" "Lubelskie"         "Lodzkie"
## [4] "Podlaskie"         "Wielkopolskie"

## [1] "Mazowieckie"

## [1] "Malopolskie"    "Opolskie"       "Podkarpackie"   "Pomorskie"
## [5] "Swietokrzyskie"

## [1] "Slaskie"

## [1] "Dolnoslaskie"       "Lubuskie"           "WarminskoMazurskie"
## [4] "Zachodniopomorskie"

## metoda hierarchiczna

## [1] "Dolnoslaskie"       "Lubuskie"           "WarminskoMazurskie"
## [4] "Zachodniopomorskie"

## [1] "KujawskoPomorskie" "Lubelskie"         "Lodzkie"

## [1] "Malopolskie"    "Opolskie"       "Podkarpackie"   "Podlaskie"
## [5] "Pomorskie"      "Swietokrzyskie" "Wielkopolskie"

## [1] "Mazowieckie"

## [1] "Slaskie"

## Importance of components:
##                           PC1    PC2    PC3    PC4     PC5     PC6     PC7
## Standard deviation     1.5860 1.3168 1.1021 0.9503 0.70586 0.31884 0.18202
## Proportion of Variance 0.3593 0.2477 0.1735 0.1290 0.07118 0.01452 0.00473
## Cumulative Proportion  0.3593 0.6070 0.7805 0.9096 0.98074 0.99527 1.00000
```



Zadanie 2. W pliku wina.txt zawarto informację o trzynastu cechach różnych gatunków win. Co więcej obserwacje podzielone są na trzy grupy.

```
##      V1   V2   V3   V4  V5   V6   V7   V8   V9  V10  V11  V12  V13 V14
## 1 14.23 1.71 2.43 15.6 127 2.80 3.06 0.28 2.29 5.64 1.04 3.92 1065   1
## 2 13.20 1.78 2.14 11.2 100 2.65 2.76 0.26 1.28 4.38 1.05 3.40 1050   1
## 3 13.16 2.36 2.67 18.6 101 2.80 3.24 0.30 2.81 5.68 1.03 3.17 1185   1
## 4 14.37 1.95 2.50 16.8 113 3.85 3.49 0.24 2.18 7.80 0.86 3.45 1480   1
## 5 13.24 2.59 2.87 21.0 118 2.80 2.69 0.39 1.82 4.32 1.04 2.93  735   1
## 6 14.20 1.76 2.45 15.2 112 3.27 3.39 0.34 1.97 6.75 1.05 2.85 1450   1
```

```
## ...
```

```
## [1] 178  14
```

```
##
##  1  2  3
## 59 71 48
```

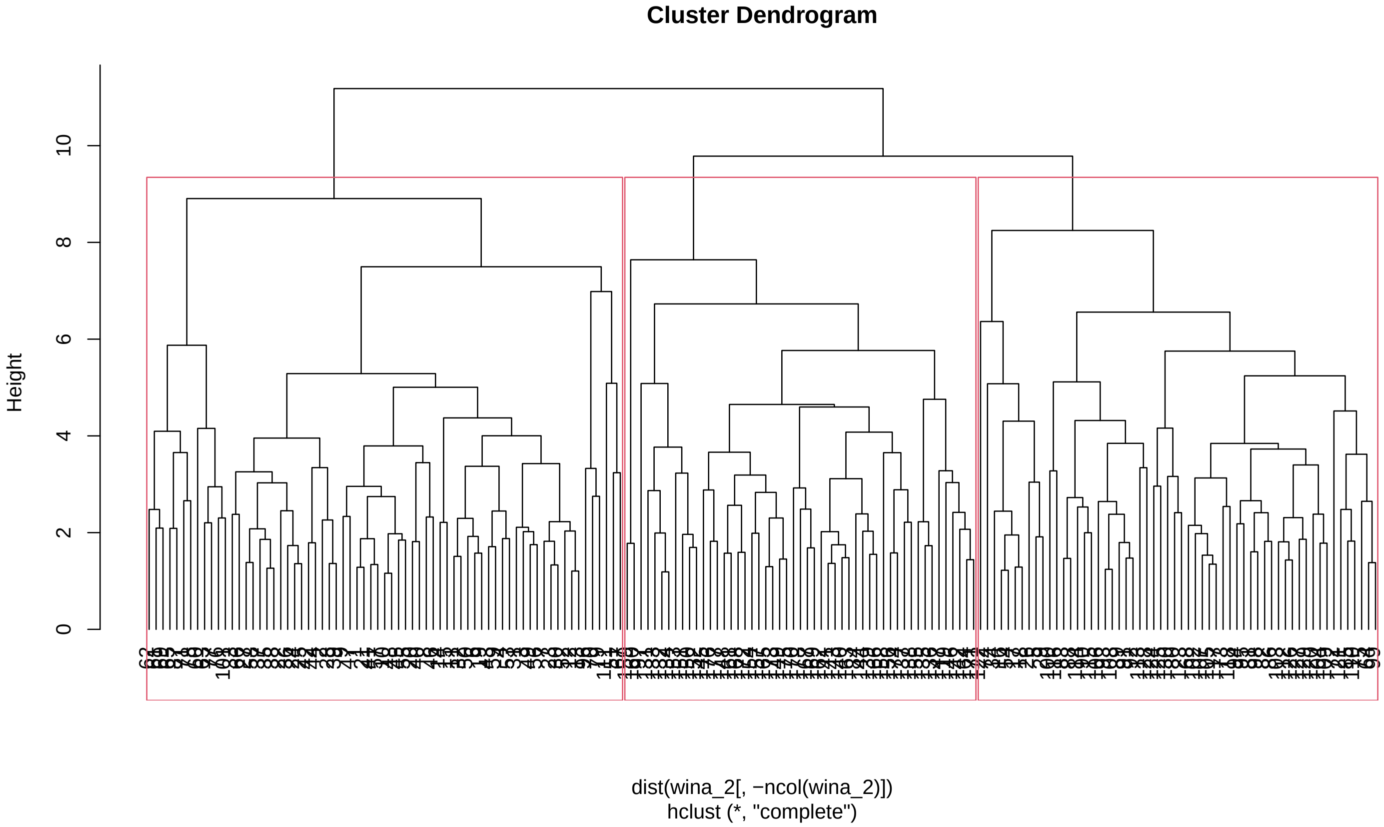
1. Czy powinniśmy dokonać standaryzacji zmiennych?

```
##          V1          V2         V3         V4         V5        V6        V7
## 1 1.5143408 -0.56066822  0.2313998 -1.1663032 1.90852151 0.8067217 1.0319081
## 2 0.2455968 -0.49800856 -0.8256672 -2.4838405 0.01809398 0.5670481 0.7315653
## 3 0.1963252  0.02117152  1.1062139 -0.2679823 0.08810981 0.8067217 1.2121137
## 4 1.6867914 -0.34583508  0.4865539 -0.8069748 0.92829983 2.4844372 1.4623994
## 5 0.2948684  0.22705328  1.8352256  0.4506745 1.27837900 0.8067217 0.6614853
## 6 1.4773871 -0.51591132  0.3043010 -1.2860793 0.85828399 1.5576991 1.3622851
##           V8         V9        V10        V11       V12         V13 V14
## 1 -0.6577078  1.2214385  0.2510088  0.3611585 1.8427215  1.01015939   1
## 2 -0.8184106 -0.5431887 -0.2924962  0.4049085 1.1103172  0.96252635   1
## 3 -0.4970050  2.1299594  0.2682629  0.3174085 0.7863692  1.39122370   1
## 4 -0.9791134  1.0292513  1.1827317 -0.4263410 1.1807407  2.32800680   1
```

```
## 5  0.2261576  0.4002753 -0.3183774  0.3611585 0.4483365 -0.03776747   1
## 6 -0.1755994  0.6623487  0.7298108  0.4049085 0.3356589  2.23274072   1
```

```
## ...
```

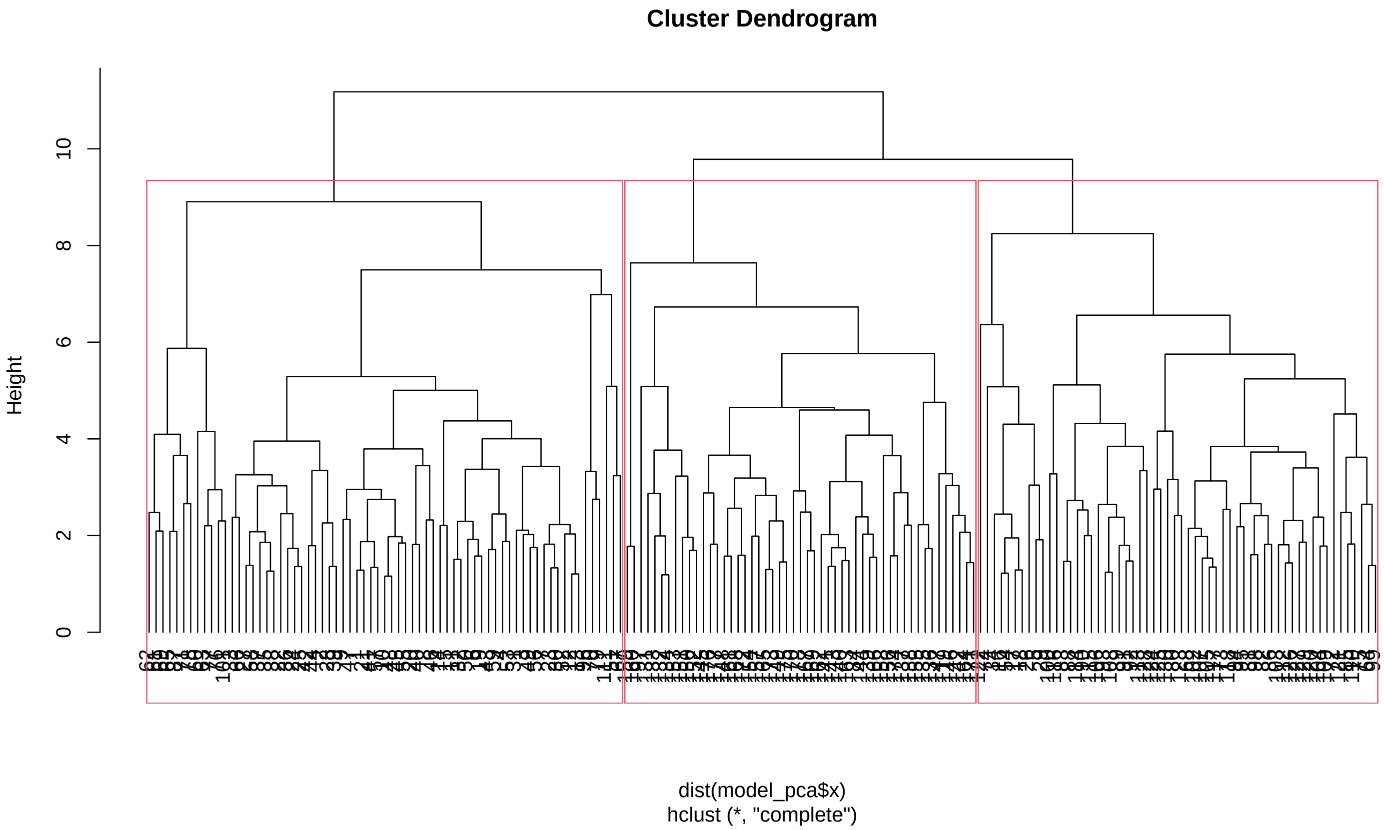
2. Wykonaj hierarchiczną analizę skupień. Narysuje dendrogram z podziałem na skupienia w liczbie równej liczbie grup wyszczególnionych w danych. Jaki jest błąd otrzymanego podziału?




```
##   [1] 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 2 2 2 1 1 1 1 1 1 1 2 1 1 2 1 1 1 1 2 1 1 1
##  [38] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 2 1 1 1 2 1 1 1 1 2 2 2
##  [75] 2 1 1 1 1 2 2 2 2 3 2 2 1 2 2 2 2 2 2 2 2 1 3 2 2 2 1 2 2 2 2 2 2 2 2 2 2 1
## [112] 2 2 2 2 2 2 3 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
## [149] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
```

```
## [1] 0.1629213
```

3. Wykonaj polecenie 2 tylko, że na składowych głównych. Co obserwujemy i dlaczego?




```
##   [1] 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 2 2 2 1 1 1 1 1 1 1 2 1 1 2 1 1 1 1 2 1 1 1
##  [38] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 2 1 1 1 2 1 1 1 2 2 2
##  [75] 2 1 1 1 1 2 2 2 2 3 2 2 1 2 2 2 2 2 2 2 2 1 3 2 2 2 1 2 2 2 2 2 2 2 2 2 1
## [112] 2 2 2 2 2 2 2 3 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
## [149] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
```

```
## [1] 0.1629213
```

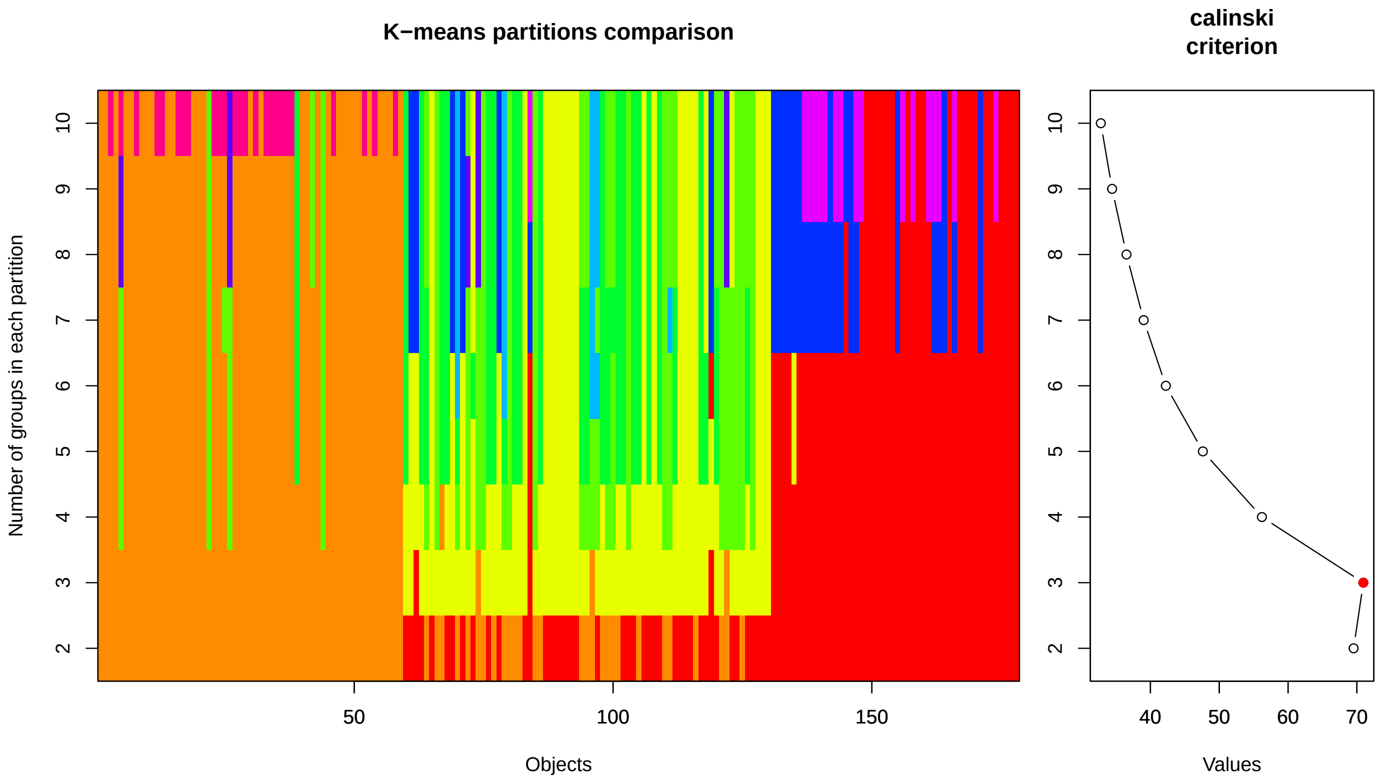
4. Wykonaj analizę skupień korzystając z metody K-średnich dla K równego liczbie grup wyszczególnionych w danych. Jaki jest błąd otrzymanego podziału?

```
##   [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
##  [38] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 1 3 3 3 3 3 3 3 3 3 3 3 2
##  [75] 3 3 3 3 3 3 3 3 3 1 3 3 3 3 3 3 3 3 3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3
## [112] 3 3 3 3 3 3 3 1 3 3 2 3 3 3 3 3 3 3 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [149] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

```
## [1] 0.03370787
```

5. Jaką optymalną liczbę skupień proponuje indeks Calińskiego-Harabasza? Rozważ $K = 2, 3, \dots, 10$.

```
##            2 groups   3 groups   4 groups   5 groups   6 groups  7 groups
## SSE      1649.43998 1270.74912 1168.61434 1095.15295 1032.79520 971.23335
## calinski   69.52333   70.94001   56.20192   47.62155   42.24094  39.02085
##           8 groups  9 groups 10 groups
## SSE      918.95440 874.89052 834.66749
## calinski  36.52408  34.43467  32.79335
```



10 Klasyfikacja

- **Uczenie się pod nadzorem** lub **uczenie się z przykładów** jest procesem budowy (konstrukcji), na bazie dostępnych danych wejściowych $\mathbf{X}_i$ oraz wyjściowych Y_i, $i = 1, 2, \ldots, n$, reguły klasyfikacyjnej zwanej inaczej **klasyfikatorem**, służącej do predykcji etykiety Y grupy, do której należy obserwacja $\mathbf{X}$.

- Załóżmy, że dysponujemy K niezależnymi, prostymi próbami losowymi o liczebnościach, odpowiednio, $n_1, n_2, \ldots, n_K$, pobranymi z K różnych populacji (klas, grup):

$$\begin{aligned} \mathbf{X}_{11}, \mathbf{X}_{12}, \ldots, \mathbf{X}_{1n_1} &- \ \textit{z populacji } 1 \\ \mathbf{X}_{21}, \mathbf{X}_{22}, \ldots, \mathbf{X}_{2n_2} &- \ \textit{z populacji } 2 \\ &\ldots \\ \mathbf{X}_{K1}, \mathbf{X}_{K2}, \ldots, \mathbf{X}_{Kn_K} &- \ \textit{z populacji } K \end{aligned}$$

  gdzie $\mathbf{X}_{ij} = (X_{ij1}, X_{ij2}, \ldots, X_{ijp})'$ jest j-tą obserwacją z i-tej populacji zawierającą p obserwowanych cech, $i = 1, 2, \ldots, K$, $j = 1, 2, \ldots, n_i$.

- Powyższe dane można wygodniej zapisać w innej postaci, a mianowicie w postaci jednego ciągu n uporządkowanych par losowych

$$(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \ldots, (\mathbf{X}_n, Y_n),$$

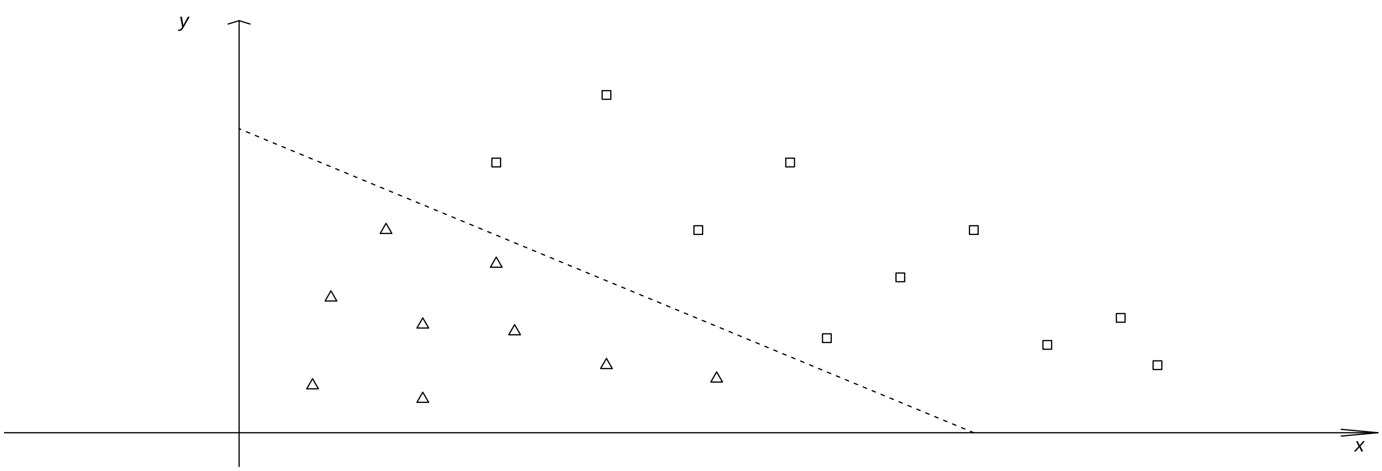
  gdzie $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})' \in \mathcal{X} \subset \mathbb{R}^p$ jest i-tą obserwacją, natomiast Y_i jest etykietą populacji, do której ta obserwacja należy, przyjmującą wartości w pewnym skończonym zbiorze $\mathcal{Y}$, $i = 1, 2, \ldots, n$, $n = n_1 + n_2 + \cdots + n_K$.

- Zbiór $\mathcal{Y}$ nazywamy **przestrzenią etykiet**.

- Składowe wektora $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})'$ nazywać będziemy **cechami**, **zmiennymi** lub **atrybutami**.

- Próbę
$$\mathcal{L}_n = \{(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \dots, (\mathbf{X}_n, Y_n)\}$$
nazywać będziemy **próbą uczącą**.

- Interesuje nas problem predykcji etykiety Y na podstawie wektora cech $\mathbf{X}$.
- Problem ten nazywany jest **klasyfikacją**, **dyskryminacją**, **uczeniem się pod nadzorem** lub **rozpoznawaniem wzorców**.
- Reguła klasyfikacyjna, zwana krótko **klasyfikatorem**, jest funkcją
$$d: \ \mathcal{X} \to \mathcal{Y}.$$
- Gdy obserwujemy nowy wektor $\mathbf{X}$, to prognozą etykiety Y jest $d(\mathbf{X})$.
- Na poniższym rysunku pokazanych jest 20 punktów. Wektor cech $\mathbf{X} = (X_1, X_2)'$ jest dwuwymiarowy a etykieta $Y \in \mathcal{Y} = \{1, 0\}$.
- Wartości cechy $\mathbf{X}$ dla $Y = 0$ reprezentowane są przez trójkąty, a dla $Y = 1$ przez kwadraty.
- Linia przerywana reprezentuje liniową regułę klasyfikacyjną postaci
$$d(\mathbf{x}) = \begin{cases} 1, & \text{jeżeli } a + b_1x_1 + b_2x_2 > 0, \\ 0, & \text{poza tym.} \end{cases}$$
- Każdy punkt leżący poniżej tej linii klasyfikowany jest do grupy o etykiecie 0 oraz każdy punkt leżący powyżej tej linii klasyfikowany jest do grupy o etykiecie 1.



10.1 Błąd klasyfikacji

- Naszym celem jest znalezienie takiego klasyfikatora $d: \ \mathcal{X} \to \mathcal{Y}$, który daje dokładną predykcję.
- Miarą jakości klasyfikatora jest jego **rzeczywisty poziom błędu** równy
$$e(d) = P(d(\mathbf{X}) \neq Y).$$

10.2 Klasyfikator bayesowski

- Załóżmy, że $Y \in \mathcal{Y} = \{1, 2, \dots, K\}$.

- Prawdopodobieństwa

$$\pi_1 = P(Y = 1), \pi_2 = P(Y = 2), \dots, \pi_K = P(Y = K)$$

  nazywamy prawdopodobieństwami **a priori** (przed doświadczeniem).
- Prawdopodobieństwa

$$\begin{aligned} p_1(\mathbf{x}) &= P(Y = 1|\mathbf{X} = \mathbf{x}), \\ p_2(\mathbf{x}) &= P(Y = 2|\mathbf{X} = \mathbf{x}), \\ &\dots \\ p_K(\mathbf{x}) &= P(Y = K|\mathbf{X} = \mathbf{x}) \end{aligned}$$

  nazywamy prawdopodobieństwami **a posteriori** (po doświadczeniu).
- Ze wzoru Bayesa mamy

$$p_k(\mathbf{x}) = \frac{\pi_k f_k(\mathbf{x})}{\sum_{i=1}^K \pi_i f_i(\mathbf{x})}, \ k = 1, 2, \dots, K,$$

  gdzie $f_k(\mathbf{x})$ oznacza gęstość rozkładu wektora $\mathbf{X}$ w k-tej klasie.

Definicja. Klasyfikator postaci

$$d_B(\mathbf{x}) = \arg\max_k p_k(\mathbf{x}) = \arg\max_k \pi_k f_k(\mathbf{x})$$

nazywamy **klasyfikatorem bayesowskim**, gdzie $\arg\max_k$ oznacza tę wartość k, która maksymalizuje dane wyrażenie.

Twierdzenie. Klasyfikator bayesowski d_B jest optymalny, tj. jeżeli d jest jakimkolwiek innym klasyfikatorem, to $e(d_B) \leq e(d)$, gdzie $e(d)$ jest rzeczywistym poziomem błędu klasyfikatora d.

Klasyfikatory gaussowskie

- Najprostszym podejściem do zagadnienia klasyfikacji jest przyjęcie modelu parametrycznego dla gęstości oraz wykorzystanie jej estymatora, tj. przyjęcie założenia, że znana jest postać gęstości z wyjątkiem tkwiących w niej parametrów.
- Załóżmy, że $f_k(\mathbf{x})$ są gęstościami p-wymiarowego rozkładu normalnego, $k = 1, 2, \dots, K$. Dokładniej $\mathbf{X}|Y = k \sim N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.
- Mówimy, że wektor losowy $\mathbf{X} = \left(X_1, X_2, \dots, X_p\right)'$ ma p-wymiarowy rozkład normalny $N_p\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ z parametrami $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_p)'$ i $\boldsymbol{\Sigma} = (\sigma_{ij})_{i,j=1}^p > 0$, jeżeli jego gęstość jest postaci

$$f(\mathbf{x}) = (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\left(\mathbf{x} - \boldsymbol{\mu}\right)' \boldsymbol{\Sigma}^{-1} \left(\mathbf{x} - \boldsymbol{\mu}\right)\right].$$

Twierdzenie. Załóżmy, że $Y \in \{1, 2, \dots, K\}$. Jeżeli $f_k(\mathbf{x})$ jest gęstością p-wymiarowego rozkładu normalnego (gaussowskiego) $N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, to klasyfikator bayesowski ma postać

$$d_B(\mathbf{x}) = \arg\max_k \delta_k(\mathbf{x}),$$

gdzie

$$\delta_k(\mathbf{x}) = -\frac{1}{2}\ln|\boldsymbol{\Sigma}_k| - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k) + \ln\pi_k.$$

Procedura klasyfikacji oparta na tej funkcji nosi nazwę **kwadratowej analizy dyskryminacyjnej (QDA)**. Jeżeli ponadto wszystkie macierze kowariancji są sobie równe i równe macierzy $\boldsymbol{\Sigma}$, to

$$\delta_k(\mathbf{x}) = \mathbf{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \ln\pi_k.$$

Procedura klasyfikacji oparta na tej funkcji nosi nazwę **liniowej analizy dyskryminacyjnej (LDA)**.

- Występujące w powyższych wzorach parametry nie są zazwyczaj znane i w praktyce należy zastąpić je ich estymatorami z próby uczącej.
- Jeżeli próba ucząca zawiera n_i obserwacji z i-tej grupy, $n_1 + n_2 + \cdots + n_K = n$ oraz $\mathbf{X}_{ij}$ jest j-tą obserwacją z i-tej grupy, to estymatory nieznanych parametrów są równe:

$$\hat{\pi}_k = \frac{n_k}{n},$$

$$\hat{\boldsymbol{\mu}}_k = \bar{\mathbf{X}}_k = \frac{1}{n_k}\sum_{j=1}^{n_k}\mathbf{X}_{kj},$$

$$\hat{\boldsymbol{\Sigma}}_k = \mathbf{S}_k = \frac{1}{n_k - 1}\sum_{j=1}^{n_k}(\mathbf{X}_{kj} - \bar{\mathbf{X}}_k)(\mathbf{X}_{kj} - \bar{\mathbf{X}}_k)',$$

$$\hat{\boldsymbol{\Sigma}} = \mathbf{S} = \frac{1}{n - K}\sum_{k=1}^{K}\sum_{j=1}^{n_k}(\mathbf{X}_{kj} - \bar{\mathbf{X}}_k)(\mathbf{X}_{kj} - \bar{\mathbf{X}}_k)'.$$

10.3 Estymacja błędu klasyfikacji

- Jakość klasyfikatora $\hat{d}$ mierzona jest za pomocą warunkowego prawdopodobieństwa błędu

$$e(\hat{d}) = P(\hat{d}(\mathbf{X}) \neq Y|\mathcal{L}_n),$$

  gdzie para losowa $(\mathbf{X}, Y)$ jest niezależna od próby uczącej $\mathcal{L}_n$.
- Wielkość $e(\hat{d})$ nazywamy **aktualnym poziomem błędu** klasyfikatora.
- Chcemy znaleźć taki klasyfikator $\hat{d}$, dla którego $e(\hat{d})$ jest bliskie $e(d_B)$. Jednakże $e(\hat{d})$ jest zmienną losową, ponieważ zależy od losowej próby uczącej $\mathcal{L}_n$.
- Niech $\hat{d}(\mathbf{x}) = \hat{d}(\mathbf{x}; \mathcal{L}_n)$ oznacza klasyfikator skonstruowany przy pomocy próby uczącej $\mathcal{L}_n$. Ponadto, niech $\hat{e} \equiv \hat{e}(\hat{d})$ oznacza ocenę aktualnego poziomu błędu klasyfikatora $\hat{d}$.
- Ocenę $\hat{e}$ nazywać będziemy **błędem klasyfikacji**.
- W sytuacjach, kiedy na populacje nie narzuca się żadnej konkretnej rodziny rozkładów, jedyną drogą oceny prawdopodobieństwa $e(\hat{d})$ jest użycie metod estymacji nieparamerycznej.
- W najlepszej sytuacji jesteśmy wtedy, gdy dysponujemy m-elementową **próbą testową** (ang. *test sample*) $\mathcal{T}_m$ niezależną od próby uczącej $\mathcal{L}_n$. Niech zatem

$$\mathcal{T}_m = \{(\mathbf{X}_1^t, Y_1^t), (\mathbf{X}_2^t, Y_2^t), \ldots, (\mathbf{X}_m^t, Y_m^t)\}.$$

  Wtedy za estymator aktualnego poziomu błędu klasyfikatora $\hat{d}$ przyjmujemy:

$$\hat{e}_{\mathcal{T}} = \frac{1}{m}\sum_{j=1}^{m} I(\hat{d}(\mathbf{X}_j^t; \mathcal{L}_n) \neq Y_j^t).$$

- W przypadku, gdy nie dysponujemy niezależną próbą testową, do estymacji używamy jedynie próby uczącej.
- Naturalną oceną aktualnego poziomu błędu jest wtedy wartość **estymatora ponownego podstawiania (resubstytucji)** (ang. *resubstitution error*)

$$\hat{e}_R = \frac{1}{n}\sum_{j=1}^{n} I(\hat{d}(\mathbf{X}_j; \mathcal{L}_n) \neq Y_j).$$

- Wartość tego estymatora uzyskuje się poprzez klasyfikację regułą $\hat{d}$ tych samych obserwacji, które służyły do jej konstrukcji. Oznacza to, iż próba ucząca jest zarazem próbą testową.
- Estymator ten jest więc obciążonym estymatorem wielkości $e(\hat{d})$ i zaniża jej rzeczywistą wartość. Uwidacznia się to szczególnie w przypadku złożonych klasyfikatorów opartych na relatywnie małych próbach uczących.

- Redukcję obciążenia można uzyskać stosując poniższe metody estymacji.
- Jednym ze sposobów redukcji obciążenia estymatora $\hat{e}_R$ jest tzw. metoda podziału próby na dwa podzbiory: próbę uczącą i próbę testową. Wówczas klasyfikator konstruuje się za pomocą pierwszego z nich, drugi natomiast służy do konstrukcji estymatora.
- Wykorzystanie tylko części informacji w celu uzyskania reguły klasyfikacyjnej prowadzi jednak często do zawyżenia wartości estymatora błędu. Rozwiązaniem tego problemu jest **metoda sprawdzania krzyżowego** (ang. *cross validation*, *leave-one-out*).
  - Oznaczmy przez $\mathcal{L}_n^{(-j)}$ próbę uczącą $\mathcal{L}_n$, z której usunięto obserwację $\mathbf{Z}_j = (\mathbf{X}_j, Y_j)$. Klasyfikator konstruuje się wykorzystując próbę $\mathcal{L}_n^{(-j)}$, a następnie testuje się go na pojedynczej obserwacji $\mathbf{Z}_j$. Czynność tę powtarza się n razy, dla każdej obserwacji $\mathbf{Z}_j$ z osobna. Odpowiedni estymator ma postać:
$$\hat{e}_{CV} = \frac{1}{n}\sum_{j=1}^{n} I(\hat{d}(\mathbf{X}_j; \mathcal{L}_n^{(-j)}) \neq Y_j).$$
  - Procedura ta w każdym z n etapów jest w rzeczywistości metodą podziału próby dla przypadku jednoelementowego zbioru testowego. Każda obserwacja próby jest użyta do konstrukcji klasyfikatora $\hat{d}$. Każda z nich jest też (dokładnie jeden raz) elementem testowym.
  - Estymator ten, choć granicznie nieobciążony, ma większą wariancję. Ponadto wymaga on konstrukcji n klasyfikatorów, co dla dużych n oznacza znaczący wzrost obliczeń.
- Rozwiązaniem pośrednim jest **metoda rotacyjna**, zwana często **v-krokową metodą sprawdzania krzyżowego** (ang. *v-fold cross validation*). Polega ona na losowym podziale próby na v podzbiorów, przy czym $v-1$ z nich tworzy próbę uczącą, natomiast pozostały - próbę testową. Procedurę tę powtarza się v razy, dla każdego podzbioru rozpatrywanego kolejno jako zbiór testowy.
  - Odpowiedni estymator jest postaci:
$$\hat{e}_{vCV} = \frac{1}{n}\sum_{i=1}^{v}\sum_{j=1}^{n} I(\mathbf{Z}_j \in \tilde{\mathcal{L}}_n^{(i)}) I(\hat{d}(\mathbf{X}_j; \tilde{\mathcal{L}}_n^{(-i)}) \neq Y_j),$$
    gdzie $\tilde{\mathcal{L}}_n^{(1)}, \tilde{\mathcal{L}}_n^{(2)}, \ldots, \tilde{\mathcal{L}}_n^{(v)}$ jest losowym v-podziałem próby $\mathcal{L}_n$ na równoliczne podzbiory, a $\tilde{\mathcal{L}}_n^{(-i)} = \mathcal{L}_n \backslash \tilde{\mathcal{L}}_n^{(i)}$, $i = 1, 2, \ldots, v$.
  - Metoda ta daje mniejsze obciążenie błędu niż metoda podziału próby i wymaga mniejszej liczby obliczeń w porównaniu ze sprawdzaniem krzyżowym (jeśli tylko $v < n$).
  - W zagadnieniu estymacji aktualnego poziomu błędu zalecane jest obranie wartości $v = 10$.
  - Metoda sprawdzania krzyżowego jest powszechnie wykorzystywana w zagadnieniu wyboru modelu. Z rodziny klasyfikatorów opisanej parametrycznie wybieramy wtedy klasyfikator, dla którego błąd klasyfikacji ma wartość najmniejszą.
- **Próbą bootstrapową** nazywamy próbę n-elementową pobraną z n-elementowej próby uczącej w procesie n-krotnego losowania pojedynczych obserwacji ze zwracaniem.
  - Niech $\mathcal{L}_n^{*1}, \mathcal{L}_n^{*2}, \ldots, \mathcal{L}_n^{*B}$ będzie ciągiem kolejno pobranych B prób bootstrapowych.

- **Bootstrapowa ocena aktualnego poziomu błędu** (ang. *bootstrap error*) ma postać

$$\hat{e}_B = \frac{1}{B}\sum_{b=1}^{B}\frac{\sum_{j=1}^{n} I(\mathbf{Z}_j \notin \mathcal{L}_n^{*b}) I(\hat{d}(\mathbf{X}_j; \mathcal{L}_n^{*b}) \neq Y_j)}{\sum_{j=1}^{n} I(\mathbf{Z}_j \notin \mathcal{L}_n^{*b})}.$$

- Widać, że powyższa ocena aktualnego poziomu błędu jest uzyskana metodą sprawdzania krzyżowego zastosowaną do prób bootstrapowych.

10.4 Przykład 10

Przykład. Zbiór danych `iris` zawiera informacje na temat czterech cech trzech gatunków irysa.

```
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

```
dim(iris)
```

```
## [1] 150   5
```

```
table(iris$Species)
```

```
##
##     setosa versicolor  virginica
##         50         50         50
```

Na przykładzie tego zbioru danych przedstawimy liniową analizę dyskryminacyjną (LDA).

- model liniowej analizy dyskryminacyjnej w R

```
library(MASS)
(model_lda <- lda(Species ~ ., data = iris))
```

```
## Call:
## lda(Species ~ ., data = iris)
##
## Prior probabilities of groups:
##     setosa versicolor  virginica
##  0.3333333  0.3333333  0.3333333
##
## Group means:
##            Sepal.Length Sepal.Width Petal.Length Petal.Width
## setosa            5.006       3.428        1.462       0.246
## versicolor        5.936       2.770        4.260       1.326
## virginica         6.588       2.974        5.552       2.026
##
## Coefficients of linear discriminants:
##                     LD1         LD2
```

```
## Sepal.Length  0.8293776  0.02410215
## Sepal.Width   1.5344731  2.16452123
## Petal.Length -2.2012117 -0.93192121
## Petal.Width  -2.8104603  2.83918785
##
## Proportion of trace:
##    LD1    LD2
## 0.9912 0.0088
```

```
# lub
# model_lda <- lda(iris[, 1:4], grouping = iris$Species)
```

- tablica kontyngencji

```
head(stats::predict(model_lda)$posterior)
```

```
##   setosa   versicolor    virginica
## 1      1 3.896358e-22 2.611168e-42
## 2      1 7.217970e-18 5.042143e-37
## 3      1 1.463849e-19 4.675932e-39
## 4      1 1.268536e-16 3.566610e-35
## 5      1 1.637387e-22 1.082605e-42
## 6      1 3.883282e-21 4.566540e-40
```

```
head(stats::predict(model_lda)$class)
```

```
## [1] setosa setosa setosa setosa setosa setosa
## Levels: setosa versicolor virginica
```

```
(conf_matrix <- table(stats::predict(model_lda)$class, iris$Species))
```

```
##
##              setosa versicolor virginica
##   setosa         50          0         0
##   versicolor      0         48         1
##   virginica       0          2        49
```

- błąd klasyfikacji metodą ponownego podstawiania

```
(1 - sum(diag(conf_matrix)) / nrow(iris))
```

```
## [1] 0.02
```

- błąd klasyfikacji metodą sprawdzania krzyżowego z $v = 1$ (1-CV, LOO, ang. *leave one out*)

```
pred_loo <- numeric(nrow(iris))
for (i in 1:nrow(iris)) {
  model_lda_i <- lda(Species ~ ., data = iris[-i, ])
  pred_loo[i] <- stats::predict(model_lda_i, iris[i, ])$class
}
table(iris$Species, pred_loo)
```

```
##             pred_loo
##               1  2  3
##   setosa     50  0  0
##   versicolor  0 48  2
##   virginica   0  1 49
```

```
(1 - sum(diag(table(iris$Species, pred_loo))) / nrow(iris))
```

```
## [1] 0.02
```

- predykcja

```
new_data <- data.frame(Sepal.Length = 5.1,
                       Sepal.Width = 3.5,
                       Petal.Length = 1.3,
                       Petal.Width = 0.3)
stats::predict(model_lda, new_data)
```

```
## $class
## [1] setosa
## Levels: setosa versicolor virginica
##
## $posterior
##   setosa   versicolor    virginica
## 1      1 4.850575e-22 6.605032e-42
##
## $x
##        LD1       LD2
## 1 8.000875 0.6775315
```

10.5 Zadania 10

Zadanie 1. Kontynuujemy przykład dotyczący zbioru danych `iris`.

1. Wyznacz błąd klasyfikacji liniowej analizy dyskryminacyjnej metodą sprawdzania krzyżowego z $v = 10$ (10-CV).

```
## [1] 0.02
```

2. Błąd klasyfikacji można oszacować również następującą metodą bootstrapową.

   - Przyjmijmy, że zbiór danych ma n obserwacji.
   - Krok 1. Losujemy ze zwracaniem n obserwacji ze zbioru danych tworzących próbę bootstrapową.
   - Krok 2. Konstruujemy klasyfikator na bazie próby bootstrapowej.
   - Krok 3. Liczymy błąd klasyfikatora wyznaczonego w kroku 2 dla obserwacji, które nie znalazły się w próbie bootstrapowej.
   - Krok 4. Powtarzamy kroki 1-3 `n_boot` razy, otrzymując błędy $b_1, \dots, b_{n\_boot}$.
   - Krok 5. Obliczamy błąd klasyfikacji metodą bootstrapową według wzoru

$$\frac{1}{n\_boot} \sum_{i=1}^{n\_boot} b_i.$$

Wyznacz błąd klasyfikacji liniowej analizy dyskryminacyjnej metodą bootstrapową. Przyjmij `n_boot = 100`.

```
## [1] 0.0259815
```

Zadanie 2. W pliku wina.txt zawarto informację o trzynastu cechach różnych gatunków win. Co więcej obserwacje podzielone są na trzy grupy.

```
##      V1   V2   V3   V4  V5   V6   V7   V8   V9  V10  V11  V12  V13 V14
## 1 14.23 1.71 2.43 15.6 127 2.80 3.06 0.28 2.29 5.64 1.04 3.92 1065   1
## 2 13.20 1.78 2.14 11.2 100 2.65 2.76 0.26 1.28 4.38 1.05 3.40 1050   1
```

```
## 3 13.16 2.36 2.67 18.6 101 2.80 3.24 0.30 2.81 5.68 1.03 3.17 1185   1
## 4 14.37 1.95 2.50 16.8 113 3.85 3.49 0.24 2.18 7.80 0.86 3.45 1480   1
## 5 13.24 2.59 2.87 21.0 118 2.80 2.69 0.39 1.82 4.32 1.04 2.93  735   1
## 6 14.20 1.76 2.45 15.2 112 3.27 3.39 0.34 1.97 6.75 1.05 2.85 1450   1
```

```
## ...
```

1. Jaki jest wymiar tych danych? Jakie są etykiety klas i ich liczebności?

```
## [1] 178  14
```

```
##
##  1  2  3
## 59 71 48
```

2. Wykonaj liniową analizę dyskryminacyjną bazując na trzech pierwszych zmiennych w tym zbiorze danych.

```
##         1         2         3
## 0.3314607 0.3988764 0.2696629
```

```
##         V1       V2       V3
## 1 13.74475 2.010678 2.455593
## 2 12.27873 1.932676 2.244789
## 3 13.15375 3.333750 2.437083
```

```
##           LD1        LD2
## V1 -1.8725417 -0.2943580
## V2 -0.0862327  1.0473192
## V3 -1.4493443  0.1419408
```

3. Wyznacz oceny prawdopodobieństw a posteriori i przewidywaną przynależność do klas obserwacji oraz tablicę kontyngencji otrzymanego klasyfikatora.

```
##           1            2          3
## 1 0.9705550 0.0006735689 0.02877140
## 2 0.3933512 0.3924750849 0.21417373
## 3 0.5316537 0.0682685490 0.40007778
## 4 0.9723331 0.0002235964 0.02744332
## 5 0.5798070 0.0197639349 0.40042907
## 6 0.9668517 0.0007345077 0.03241381
```

```
## [1] 1 1 1 1 1 1
## Levels: 1 2 3
```

```
##
##      1  2  3
##   1 51  5  7
##   2  4 62  8
##   3  4  4 33
```

4. Wyznacz błąd klasyfikacji metodą ponownego podstawiania.

```
## [1] 0.1797753
```

5. Wyznacz błąd klasyfikacji metodą sprawdzania krzyżowego z $v = 1$.

```
##    pred_loo
##      1  2  3
```

```
##    1 49  5  5
##    2  5 61  5
##    3 10  8 30
```

```
## [1] 0.2134831
```

6. Wyznacz błąd klasyfikacji metodą sprawdzania krzyżowego z $v = 10$.

```
## [1] 0.2078652
```

7. Wyznacz błąd klasyfikacji metodą bootstrapową. Przyjmij `n_boot = 100`.

```
## [1] 0.2111531
```

8. Do których klas i z jakimi prawdopodobieństwami a posteriori należy zaklasyfikować poniższe nowe obserwacje?

V1	V2	V3
13.64	3.10	2.56
13.94	1.73	2.27
13.08	3.90	2.36
12.29	3.17	2.21

```
## $class
## [1] 1 1 3 2
## Levels: 1 2 3
##
## $posterior
##             1           2          3
## 1 0.531302523 0.007133455 0.46156402
## 2 0.924346812 0.007006399 0.06864679
## 3 0.061216479 0.054434582 0.88434894
## 4 0.005015639 0.810915785 0.18406858
##
## $x
##          LD1        LD2
## 1 -1.5435449  0.6390430
## 2 -1.5668588 -0.9252545
## 3 -0.2740389  1.6133507
## 4  1.4856206  1.0600594
```