

Pierwszy projekt zaliczeniowy

Statystyczna analiza danych 2020/2021

Joanna Kęczkowska

07.05.2021

Celem zadania jest statystyczna analiza danych znajdujących się w pliku `people.tab`. Dane: Są to dane symulowane; opisują wiek (zmienna `age`), wagę (`weight`), wzrost (`height`), płeć (`gender`), stan cywilny (`married`), liczbę dzieci (`number_of_kids`), posiadane zwierzę domowe (`pet`) oraz miesięczne wydatki (`expenses`) pewnych osób. We wszystkich zadaniach poniżej zmienna `expenses` jest zmienną objaśnianą (zależną), a pozostałe zmienne są zmiennymi objaśniającymi (niezależnymi).

Wczytywanie danych

1. Wczytaj dane, obejrzyj je i podsumuj w dwóch-trzech zdaniach. Pytania pomocnicze: ile jest obserwacji, ile zmiennych ilościowych, a ile jakościowych? Czy są zależności w zmiennych objaśniających (policz i zaprezentuj na wykresach korelacje pomiędzy zmiennymi ilościowymi, a także zbadaj zależność zmiennych jakościowych). Skomentuj wyniki. Czy występują jakieś braki danych?

```
df <- read.delim("people.tab", header = TRUE, sep = '\t')
sprintf("Dane zawierają %d obserwacji i %d cech", dim(df)[1], dim(df)[2])
```

```
## [1] "Dane zawierają 500 obserwacji i 8 cech"
```

```
summary(df)
```

```
##      age      weight      height      gender
## Min.   :17.00   Min.    : 19.40   Min.    :113.6   Length:500
## 1st Qu.:33.00   1st Qu.: 57.60   1st Qu.:155.6   Class  :character
## Median :39.00   Median : 66.60   Median :169.0   Mode   :character
## Mean   :39.48   Mean    : 66.39   Mean     :168.2
## 3rd Qu.:45.00   3rd Qu.: 75.30   3rd Qu.:180.1
## Max.    :72.00   Max.    :107.20   Max.     :235.2
## married number_of_kids      pet      expenses
## Mode :logical   Min.    :0.000   Length:500   Min.    :-685.68
## FALSE:327       1st Qu.:0.750   Class  :character   1st Qu.: 74.51
## TRUE :173       Median :1.000   Mode   :character   Median : 402.22
##                Mean    :1.558                Mean    : 478.60
##                3rd Qu.:2.000                3rd Qu.: 802.72
##                Max.    :6.000                Max.    :3503.90
```

Podsumowanie i faktoryzacja zmiennych jakościowych

Dane zawierają 500 obserwacji.

Zmienne ilościowe: 'age', 'weight', 'height', 'expenses'.

Zmienne jakościowe: 'gender', 'married', 'pet', 'number_of_kids'. Niepokojące są ujemne wartości w cesze 'expenses', jak również factor 'other' w cesze 'gender'. Wartość 'none' w cesze 'pet' interpretuję jako nieposiadanie zwierzęcia. W zmiennej 'gender' interpretuję factor 'other' jako brak informacji - 38 braków.

Faktoryzacja zmiennych jakościowych:

```
df$gender <- factor(df$gender)
df$pet <- factor(df$pet)
df$married <- factor(df$married)
df$number_of_kids <- factor(df$number_of_kids)
summary(df)
```

```
##      age      weight      height      gender      married
##  Min.   :17.00   Min.    : 19.40   Min.    :113.6   man   :223   FALSE:327
##  1st Qu.:33.00   1st Qu.  : 57.60   1st Qu. :155.6   other: 38   TRUE :173
##  Median :39.00   Median   : 66.60   Median   :169.0   woman:239
##  Mean    :39.48   Mean     : 66.39   Mean     :168.2
##  3rd Qu. :45.00   3rd Qu.  : 75.30   3rd Qu. :180.1
##  Max.    :72.00   Max.     :107.20   Max.     :235.2
##
##  number_of_kids      pet      expenses
##  0:125             cat      :105   Min.    : -685.68
##  1:161             dog       :100   1st Qu. :   74.51
##  2: 99             ferret    : 54   Median   : 402.22
##  3: 63             hedgehog: 54   Mean     : 478.60
##  4: 35             none     :187   3rd Qu. : 802.72
##  5: 11                                     Max.     :3503.90
##  6: 6
```

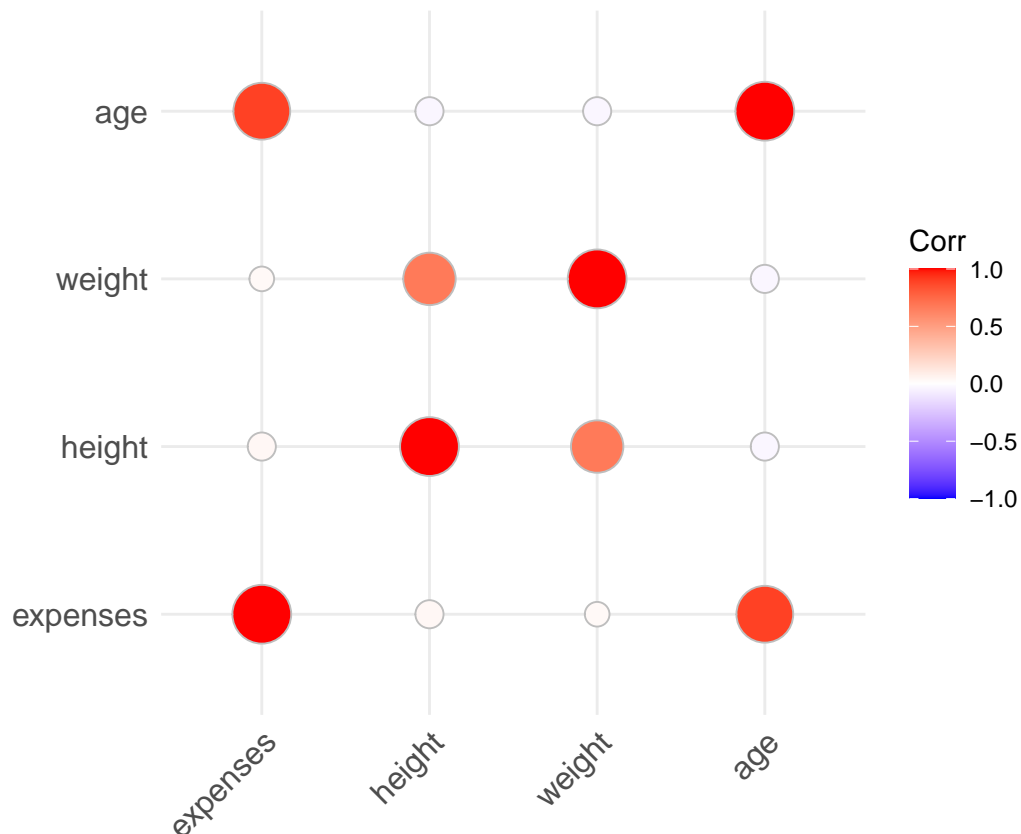
Korelacja zmiennych ilościowych

Współczynnik korelacji r jest liczbą pomiędzy -1 i 1 , która określa, w jakim stopniu dwie zmienne są współzależne. Wartość $r = 0$ oznacza, że nie ma żadnego powiązania, a wartość 1 lub -1 oznacza idealne powiązanie. Znak współczynnika korelacji wskazuje, czy zmienne są skorelowane dodatnio (większe wartości w jednej zmiennej pokrywają się z większymi wartościami w drugiej), czy też ujemnie (większe wartości w jednej zmiennej pokrywają się z mniejszymi wartościami w drugiej).

```
library(ggcorrplot)
numerical <- df[c("expenses", "height", "weight", "age")]
categorical <- df[c("married", "gender", "pet", "number_of_kids")]

corr <- round(cor(numerical), 2)

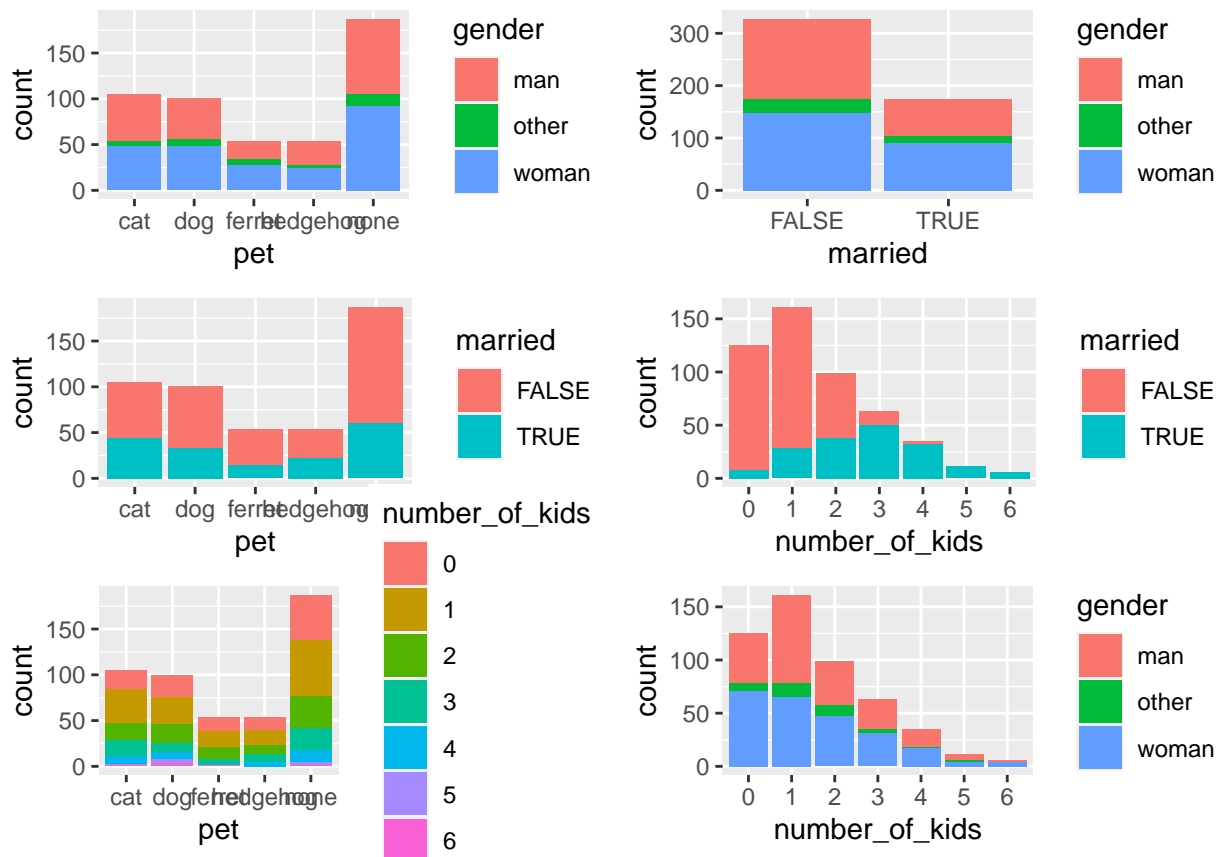
ggcorrplot(corr, method = "circle")
```



Zgodnie z intuicją wiek jest dodatnio skorelowany z wydatkami i wzrost jest dodatnio skorelowany z wagą.

Korelacja zmiennych jakościowych

```
p1 <- ggplot(df, aes(fill=gender, x=pet)) + geom_bar(position="stack")
p2 <- ggplot(df, aes(fill=gender, x=married)) + geom_bar(position="stack")
p3 <- ggplot(df, aes(fill=married, x=pet)) + geom_bar(position="stack")
p4 <- ggplot(df, aes(fill=married, x=number_of_kids)) + geom_bar(position="stack")
p5 <- ggplot(df, aes(fill=number_of_kids, x=pet)) + geom_bar(position="stack")
p6 <- ggplot(df, aes(fill=gender, x=number_of_kids)) + geom_bar(position="stack")
grid.arrange(p1, p2, p3, p4, p5, p6, nrow=3)
```



Liczba dzieci zależy od małżeństwa (rzędy 2 kolumna 2). W innych nie widać istotnych korelacji, ale można dla pewności przeprowadzić test:

Dla danej komórki wartość oczekiwana: $e = \frac{\text{row.sum} * \text{col.sum}}{\text{grand.total}}$

Chi-square statistic: $\chi^2 = \sum \frac{(o-e)^2}{e}$, gdzie o - obserwacja, e - wartość oczekiwana

Hipoteza zerowa H_0 : Zmienne są **niezależne**.

Hipoteza alternatywna H_1 : Zmienne są **zależne**.

```
#funkcja do testowania korelacji zmiennych jakościowych
#przyjmuje dwie kolumny zmiennych kategorycznych, które zamienia na tablicę wielozdzielczą

testchi <- function(feature1, feature2, sq = 20, t = ' ', alpha = 0.05) {
  TAB <- table(feature1, feature2)
  total <- sum(TAB)

  n <- nlevels(feature1)
  m <- nlevels(feature2)

  sumRows <- margin.table(TAB, 1) #rows
  sumCols <- margin.table(TAB, 2) #columns

  sumRows <- as.vector(sumRows)
  sumCols <- as.vector(sumCols)
```

```

exp <- matrix(rep(0, n * m), nrow = n, ncol = m)
exp[] <- 0L
for (i in 1:n) {
  exp[i,] <- sumRows[i] * sumCols / total
}

Tab <- data.frame(TAB)
obs <- matrix(Tab[["Freq"]], nrow = n, ncol = m)

chi_sq <- sum((obs - exp)^2 / exp) #test statistic
df <- (nrow(obs) - 1) * (ncol(obs) - 1) #deg of freedom
pval <- pchisq(chi_sq, df, lower.tail = FALSE) #right-tailed

quantile <- qchisq(alpha, df, lower.tail = FALSE) #quantile of chi-square distribution

x <- seq(0, sq, by = 0.1)
chi_dense <- dchisq(x, df)

plot(x, chi_dense, type = 'l', xlab = "x value",
      ylab = "Density", main = "Chi-square density")

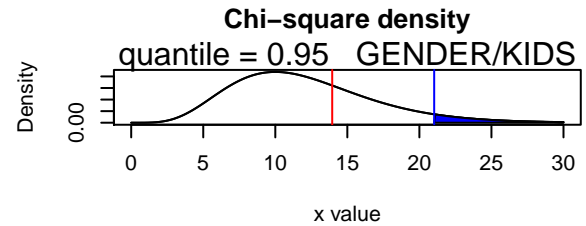
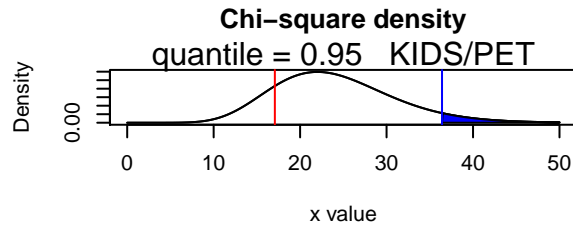
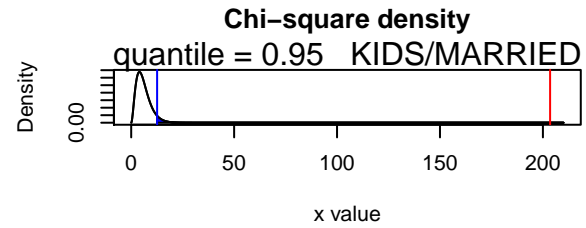
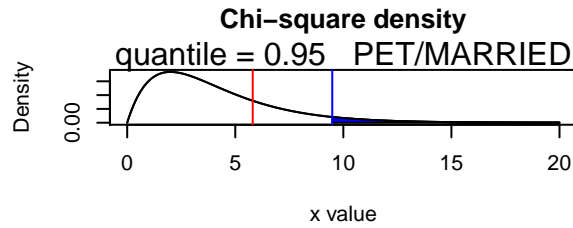
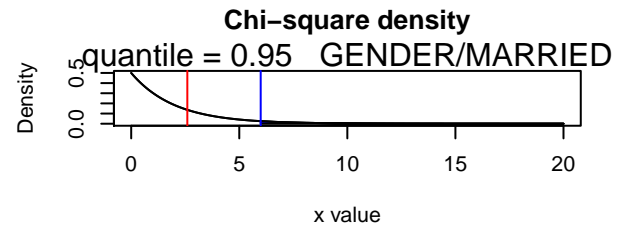
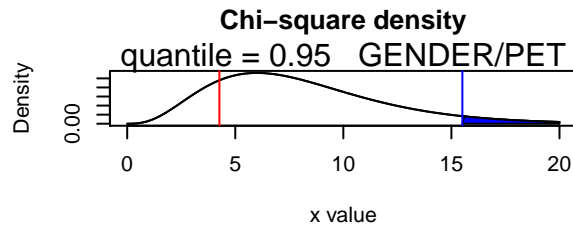
i <- x >= quantile
lines(x, chi_dense)
polygon(c(quantile, x[i], sq), c(0, chi_dense[i], 0), col = "blue")

area <- pchisq(quantile, df, lower.tail = TRUE)
result <- paste("quantile =", signif(area, digits = 3), " ", t)
mtext(result, 3)
abline(v = chi_sq, col = "red")
abline(v = quantile, col = "blue")

c <- list(chi_sq, pval, quantile)
return(c)
}

par(mfrow = c(3, 2))
gp <- testchi(df$gender, df$pet, t = "GENDER/PET")
gm <- testchi(df$gender, df$married, t = "GENDER/MARRIED")
pm <- testchi(df$pet, df$married, t = "PET/MARRIED")
nm <- testchi(df$number_of_kids, df$married, sq = 210, t = "KIDS/MARRIED")
np <- testchi(df$number_of_kids, df$pet, t = "KIDS/PET", sq = 50)
gn <- testchi(df$gender, df$number_of_kids, t = "GENDER/KIDS", sq = 30)

```



```

sprintf("GENDER/PET, test statistic = %f , p-value = %f, confidece interval = [-infinity, %f]", gp[[1]]

## [1] "GENDER/PET, test statistic = 4.264540 , p-value = 0.832502, confidece interval = [-infinity, 15

sprintf("GENDER/MARRIED, test statistic = %f , p-value = %f, confidece interval = [-infinity, %f]", gm[

## [1] "GENDER/MARRIED, test statistic = 2.597089 , p-value = 0.272929, confidece interval = [-infinity

sprintf("PET/MARRIED, test statistic = %f , p-value = %f, confidece interval = [-infinity, %f]", pm[[1]]

## [1] "PET/MARRIED, test statistic = 5.806971 , p-value = 0.214035, confidece interval = [-infinity, 9

sprintf("KIDS/MARRIED, test statistic = %f , p-value = %f, confidece interval = [-infinity, %f]", nm[[1]]

## [1] "KIDS/MARRIED, test statistic = 203.501380 , p-value = 0.000000, confidece interval = [-infinity

sprintf("KIDS/PET, test statistic = %f , p-value = %f, confidece interval = [-infinity, %f]", np[[1]], r

## [1] "KIDS/PET, test statistic = 17.076893 , p-value = 0.845364, confidece interval = [-infinity, 36.

```

```
sprintf("GENDER/KIDS, test statistic = %f , p-value = %f, confidece interval = [-infinity, %f]", gn[[1],
```

```
## [1] "GENDER/KIDS, test statistic = 13.950817 , p-value = 0.303860, confidece interval = [-infinity, 1
```

Czerwoną linią zaznaczono test statystyczny, niebieską kwantyl na poziomie istotności 5%, a niebieski obszar to obszar krytyczny.

Jedyne dwie skorelowane zmienne jakościowe to 'number_of_kids' i 'married' - statystyka testowa wpada do obszaru krytycznego (wykres 2 rząd 2 kolumna). W przypadku pozostałych par zmiennych nie mamy podstawy do odrzucenia hipotezy zerowej. Żadne dwie inne zmienne nie są skorelowane.

Wykresy

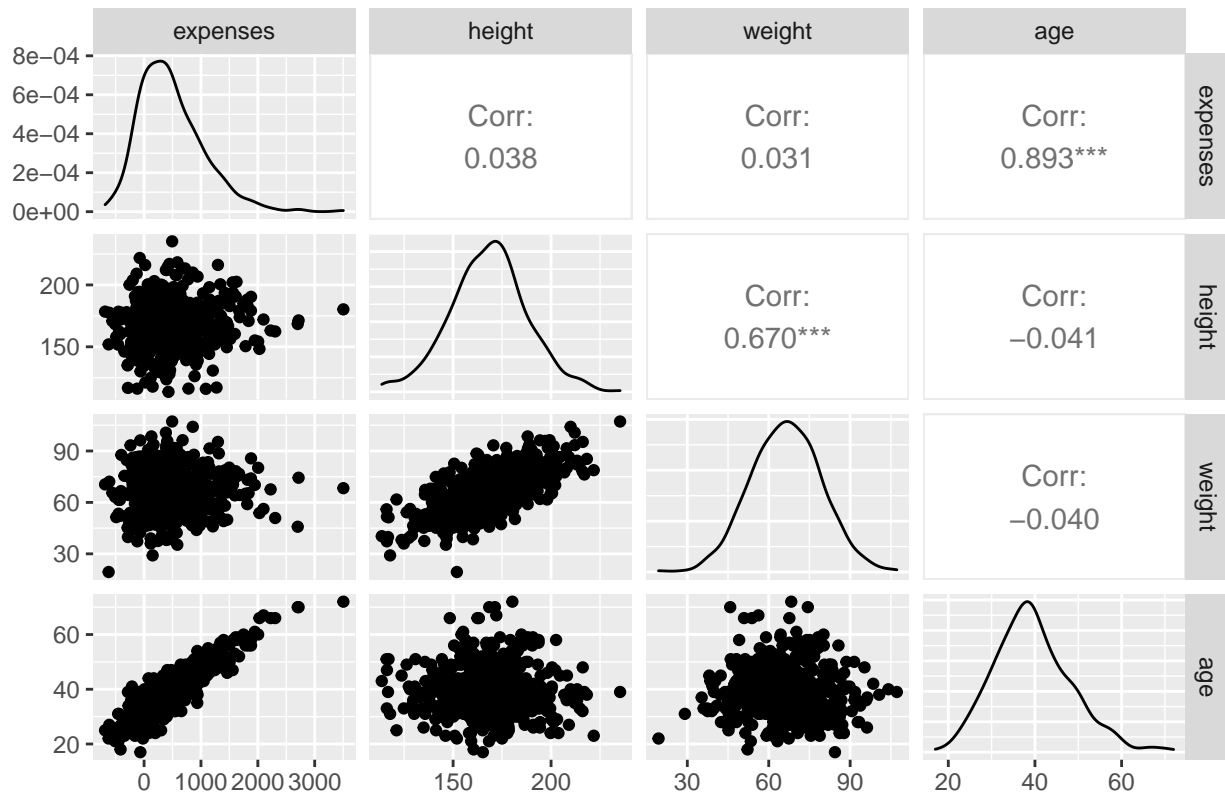
2. Podsumuj dane przynajmniej trzema różnymi wykresami. Należy przygotować: **a)** wykres typu scatter-plot (taki jak na wykładzie 6, slajd 3) dla wszystkich zmiennych objaśniających ilościowych i zmiennej objaśnianej. **b)** Wykresy typu pudełkowy (boxplot) dla jednej wybranej zmiennej ilościowej. **c)** Wykres typu słupkowy (barplot) dla jednej wybranej zmiennej jakościowej. Dodatkowe wykresy wg własnej inwencji (np. histogram, punktowy, liniowy, mapa ciepła...).

Scatter-plot dla wszystkich zmiennych ilościowych

```
library(GGally)

ggpairs(numerical, title = "Correlogram of numerical features")
```

Correlogram of numerical features



Trochę inny wykres od 'ggcorrplot(corr, method = "circle")', ale prowadzący do tych samych wniosków: dodatnia korelacja wzrostu z wagą (0.67) i dodatnia korelacja wieku z wydatkami (0.893).

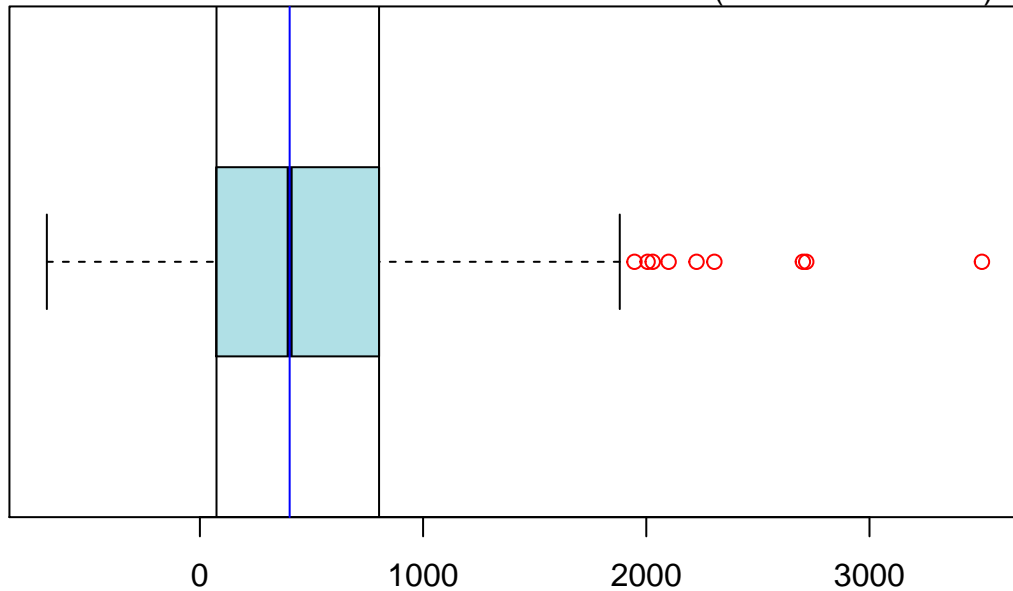
BoxPlot dla zmiennej 'expenses':

```
expenses <- df$expenses
quantiles <- unname(quantile(expenses))
boxplot(expenses, horizontal = TRUE, col = "powderblue", outcol = "red", main = "BoxPlot for expenses")

result <- paste("50% of values are between ", round(quantiles[2], 2), " and ", round(quantiles[4], 2),
mtext(result, 3)
abline(v = quantiles[4], col = "black") #quantile 75%
abline(v = quantiles[2], col = "black") #quantile 25%
abline(v = median(expenses), col="blue") #median
```


BoxPlot for expenses

50% of values are between 74.51 and 802.72 (black vertical lines)



```
sprintf("Kwantyl 3/4: %f, kwantyl 1/4: %f, mediana: %f", quantiles[4], quantiles[2], median(expenses))
```

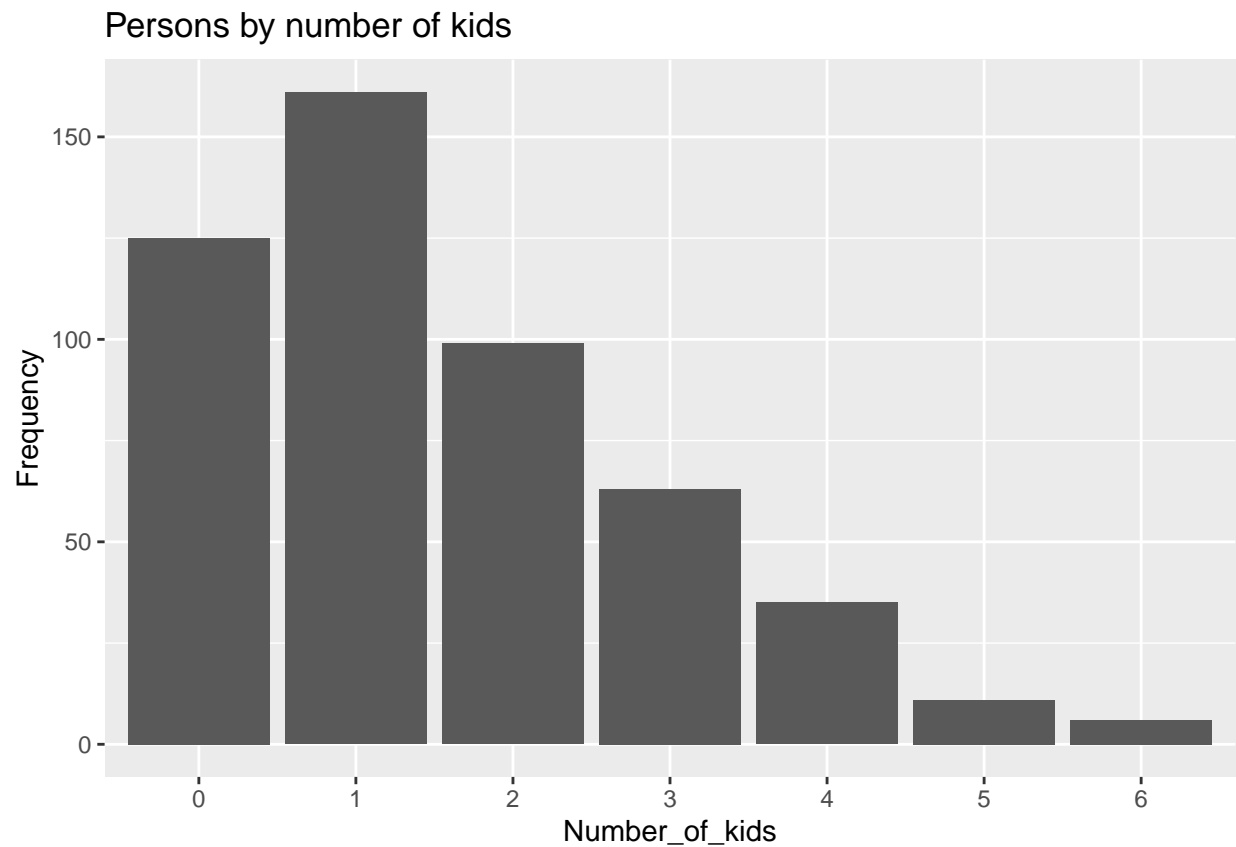
```
## [1] "Kwantyl 3/4: 802.721954, kwantyl 1/4: 74.514280, mediana: 402.219536"
```

Obserwacje odstające zaznaczono na czerwono.

Rozkład zmiennej 'number_of_kids' - wykres słupkowy:

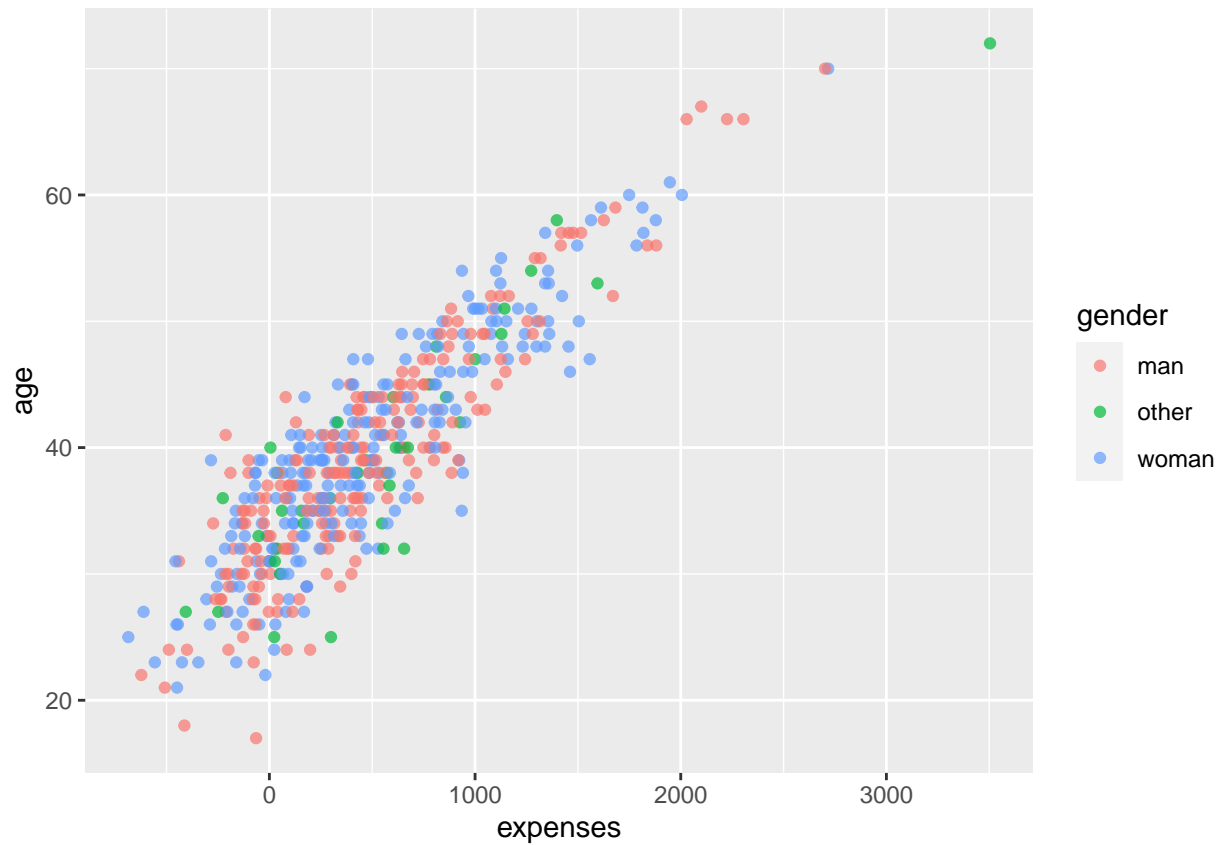
```
library(ggplot2)
```

```
ggplot(df, aes(x = number_of_kids)) +  
  geom_bar() +  
  labs(x = "Number_of_kids",  
       y = "Frequency",  
       title = "Persons by number of kids")
```



Wykres punktowy dla zmiennej 'expenses' i 'age' - dodatkowe wykresy:

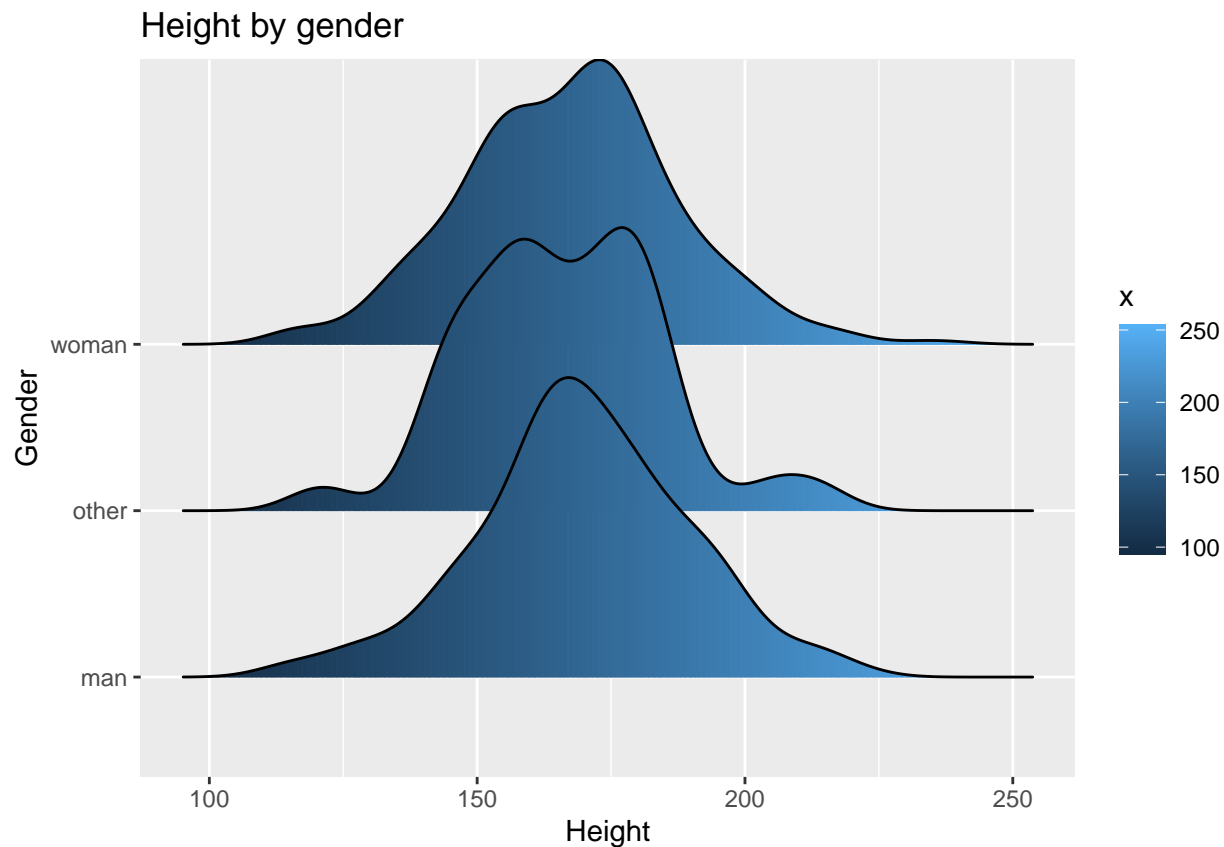
```
ggplot(df, aes(x = expenses, y = age, color = gender)) +  
  geom_point(alpha = 0.7) +  
  scale_size(range = c(1.4, 19))
```



Wykres punktowy zmiennej 'expenses' w zależności od 'age' z zaznaczoną kolorem 'gender'.

```
hW <- df[df$gender=="woman",]$height
hM <- df[df$gender=="man",]$height
ggplot(df, aes(x = height, y = gender, fill = stat(x))) +
  geom_density_ridges_gradient() +
  labs(x = "Height",
       y = "Gender",
       title = "Height by gender")
```

```
## Picking joint bandwidth of 6.16
```



Testowanie hipotez dla wartości średniej i mediany

3. Policz p-wartości dla hipotez o wartości średniej $m = 170$ i medianie $me = 165$ (cm) dla zmiennej wzrost. Wybierz statystykę testową dla alternatywy lewostronnej, **podaj założenia**, z jakich korzystałeś i skomentuj czy wydają Ci się uprawnione.

Jednopróbkowy test t dla średniej

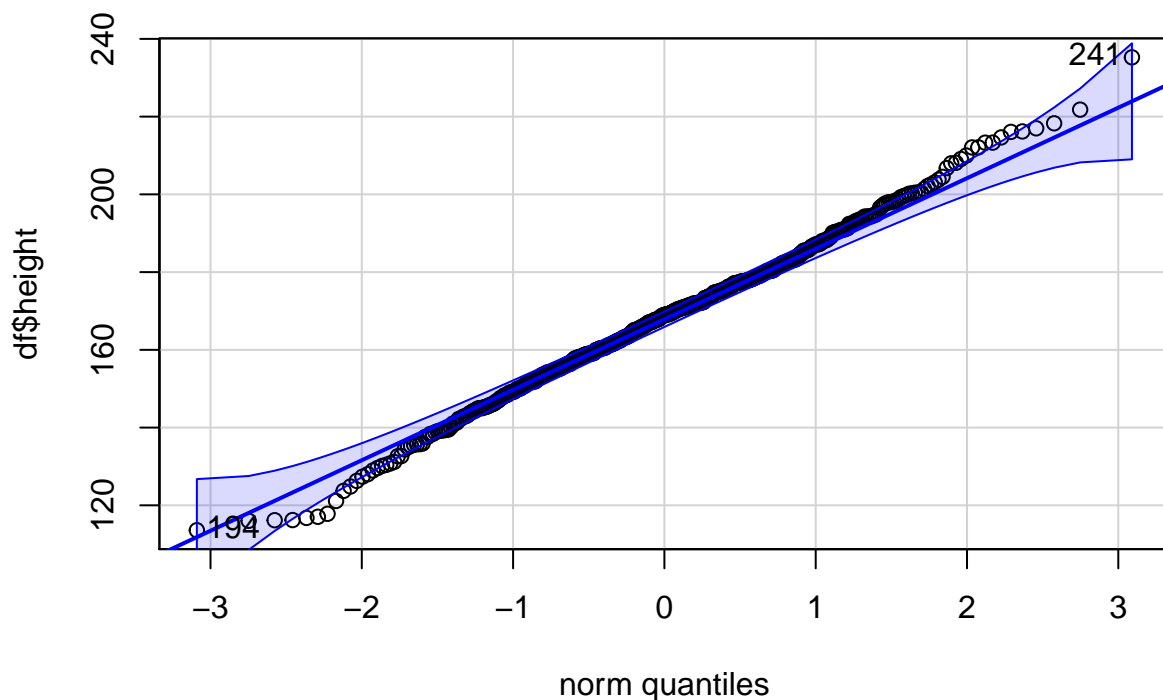
Jednopróbkowy test t: $t = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \cdot \sqrt{n}$, gdzie \bar{X} - średnia próby, n - liczba obserwacji, σ - odchylenie standardowe

Hipoteza zerowa $H_0: \mu = 170$

Hipoteza alternatywna $H_1: \mu < 170$ (alternatywa lewostronna)

Założenia: (zmienna ilościowa, rozkład normalny próby, niezależne obserwacje)

```
qqPlot(df$height)
```



```
## [1] 241 194
```

```
shaptest <- shapiro.test(df$height)
shaptest
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$height
## W = 0.99662, p-value = 0.3778
```

P-value jest większe niż 10% co w połączeniu z analizą wykresu qqPlot() implikuje, że zmienna nie różni się istotnie od rozkładu normalnego. Innymi słowy możemy założyć rozkład normalny. Oczywiście obserwacje są niezależne - jedna obserwacja nie dostarcza informacji o drugiej - obserwacje to różne osoby.

```
alpha <- 0.05
m <- 170
me <- 165
height <- df$height
n <- length(height)

test <- (mean(height) - m) / sd(height) * sqrt(n)
def <- n - 1
```

```

quantile <- qt(alpha, def) #left-tailed
pval <- pt(test, def)

x <- seq(-5, 5, by = 0.01)
t_dense <- dt(x, n - 1)

plot(x, t_dense, type = 'l', xlab = "x value",
      ylab = "Density", main = "Student's density")

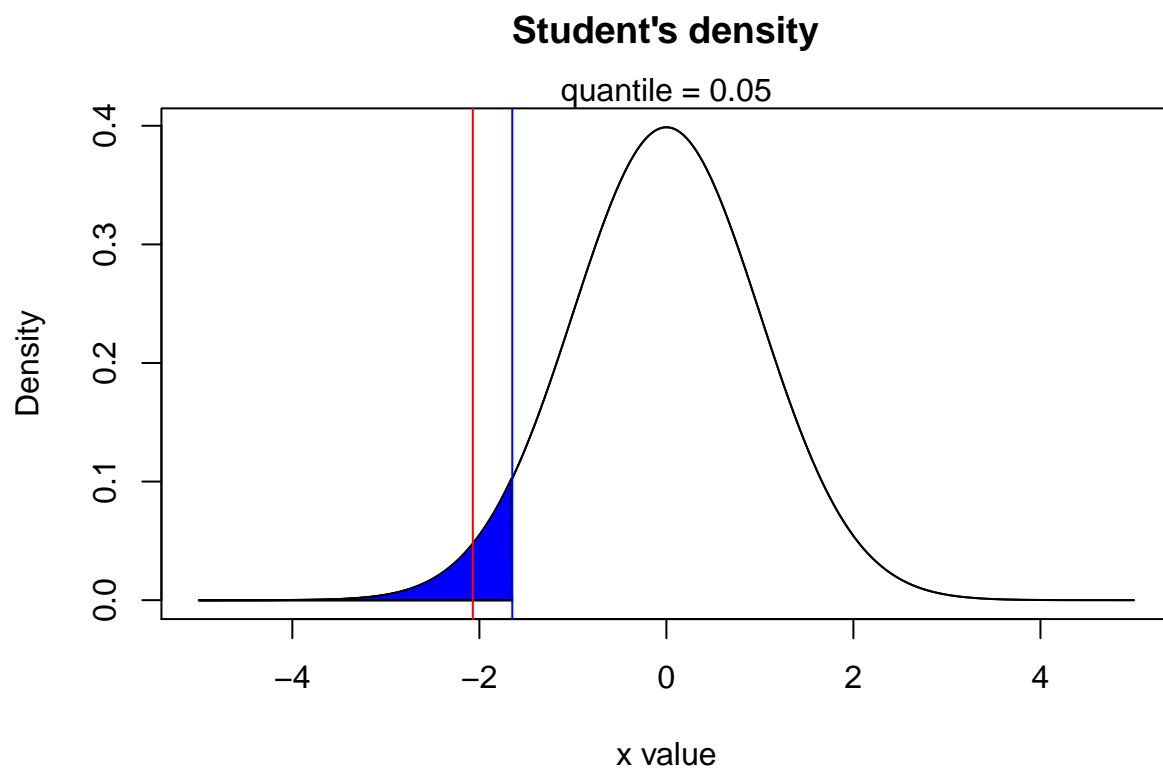
i <- x <= quantile
lines(x, t_dense)

polygon(c(-5, quantile, x[i]), c(0, t_dense[i], 0), col = "blue")
sprintf("test statistic = %f , p-value = %f, confidence interval = [%f, infinity]", test, pval, quantile)

## [1] "test statistic = -2.069917 , p-value = 0.019487, confidence interval = [-1.647913, infinity]"

area <- pt(quantile, def)
result <- paste("quantile =", signif(area, digits = 3))
mtext(result, 3)
abline(v = test, col = "red")
abline(v = quantile, col = "blue")

```



```

if (alpha > pval) {
  print("H0 rejected.")
}else {
  print("There is not enough evidence to reject H_0")
}

```

```
## [1] "H0 rejected."
```

Statystyka testowa (czerwona linia) wpada do obszaru krytycznego (zaznaczonego na niebiesko) dlatego odrzucamy hipotezę zerową na rzecz alternatywnej, czyli $\mu < 170$.

Test Wilcoxona dla mediany

Hipoteza zerowa H_0 : $m = 165$

Hipoteza alternatywna H_1 : $m < 165$ (alternatywa lewostronna)

Założenia: (zmienna ilościowa, niezależne obserwacje)

```

heightFrame <- as.data.frame(t(height))
rep <- as.data.frame(t(rep(165, length(height))))
heightFrame <- rbind(heightFrame, heightFrame - rep) #differences between 165
#heightFrame <- rbind(heightFrame, sign(heightFrame - rep))
heightFrame <- rbind(heightFrame, abs(heightFrame[2,]))

vec <- as.numeric(heightFrame[3,])
heightFrame <- rbind(heightFrame, rank(vec))

if (sum(heightFrame[2,] == 0) == 0) {
  n <- length(vec)
  posRanks <- sum(heightFrame[4, heightFrame[2,] > 0])
  negRanks <- sum(heightFrame[4, heightFrame[2,] < 0])
  if (posRanks + negRanks == n * (n+1)/2) {
    W <- min(posRanks, negRanks)
    m <- n * (n + 1) / 4
    sd <- n * (n + 1) * (2 * n + 1) / 24
    un <- as.numeric(heightFrame[4,])
    t <- length(un[un %in% un[duplicated(un)]] # = 0)
    tiedRanks <- (t^3 - t) / 48 # = 0
    zscore <- (W - m) / sqrt(sd - tiedRanks)
  }
}

quantile <- qnorm(alpha)
pval <- pnorm(zscore)

x <- seq(-5, 5, by = 0.01)
density <- dnorm(x)

plot(x, density, type = 'l', xlab = "x value",
      ylab = "Density", main = "Gaussian density")

i <- x <= quantile
lines(x, density)

```

```

polygon(c(-5, quantile, x[i]), c(0, density[i], 0), col = "blue")
sprintf("test statistic = %f , p-value = %f, confidece interval = [%f, infinity]", zscore, pval, quantile)

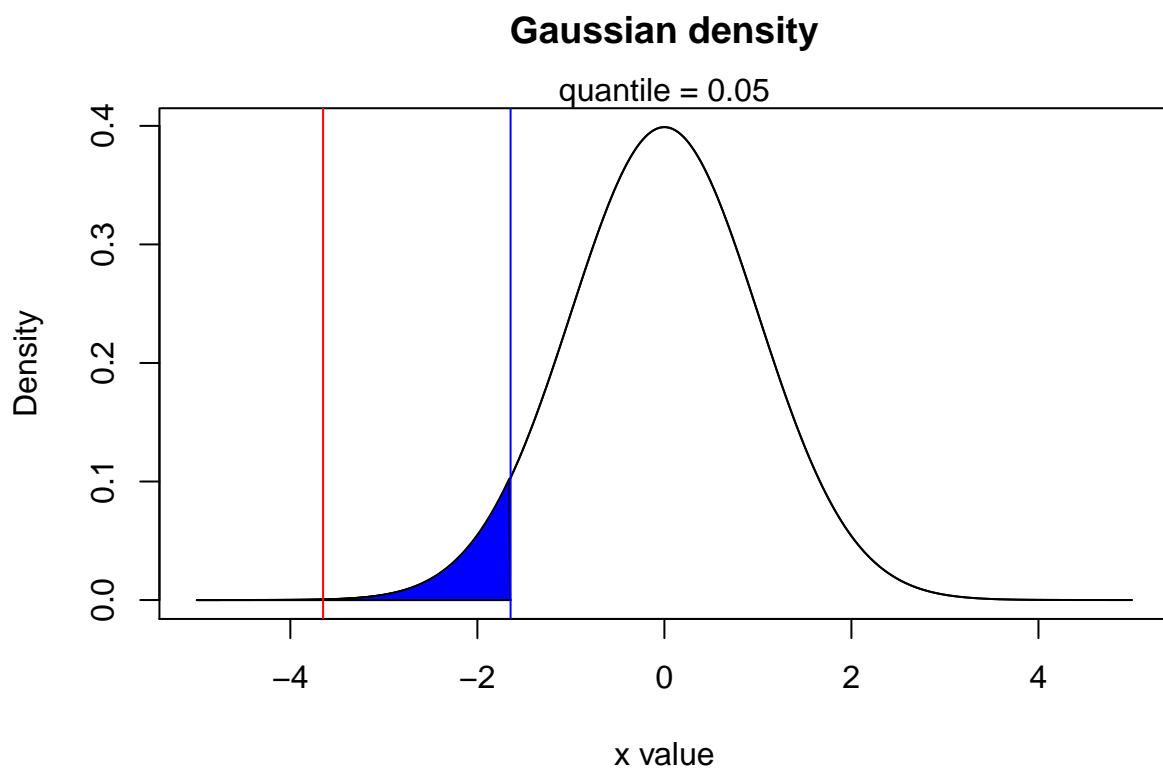
```

```
## [1] "test statistic = -3.649839 , p-value = 0.000131, confidece interval = [-1.644854, infinity]"
```

```

area <- pnorm(quantile)
result <- paste("quantile =", signif(area, digits = 3))
mtext(result, 3)
abline(v = zscore, col = "red")
abline(v = quantile, col = "blue")

```



```

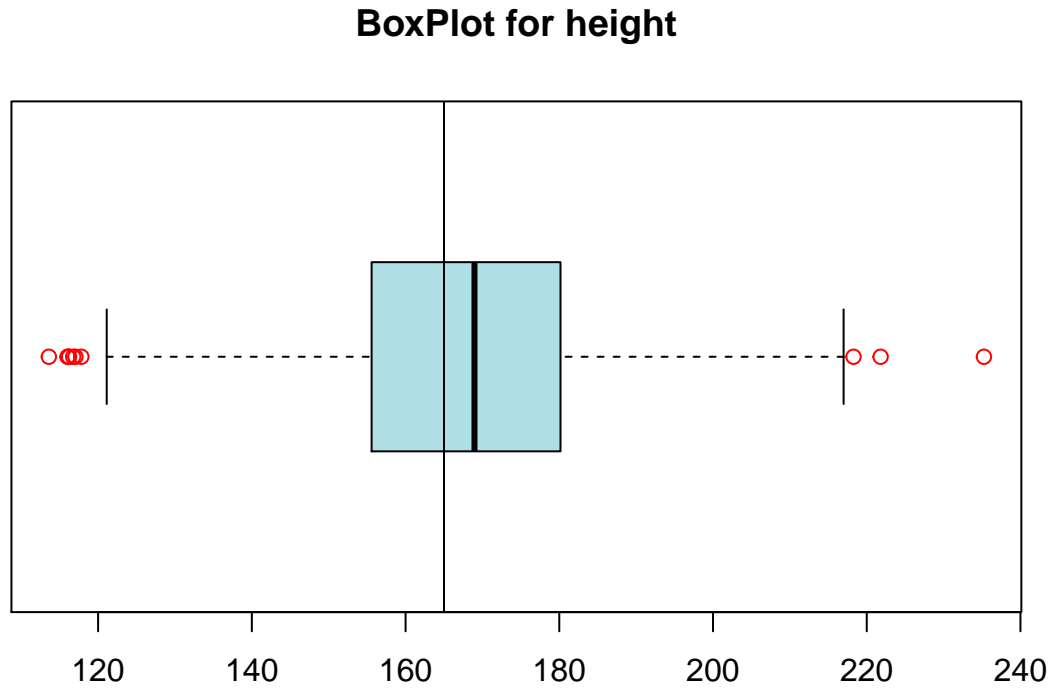
if (alpha > pval) {
  print("H0 rejected.")
}else {
  print("There is not enough evidence to reject H_0")
}

```

```
## [1] "H0 rejected."
```

Odrzucamy hipotezę zerową - statystyka testowa wpada do obszaru krytycznego (wykres) - na rzecz hipotezy alternatywnej, czyli $m < 165$.


```
boxplot(height, horizontal = TRUE, col = "powderblue", outcol = "red", main = "BoxPlot for height")
abline(v = 165)
```



Dwustronne przedziały ufności dla zmiennej wiek

4. Policz dwustronne przedziały ufności na poziomie 0.99 dla zmiennej wiek dla następujących parametrów rozkładu:

1. średnia i odchylenie standardowe;
2. kwantyle 1/4, 2/4 i 3/4.

Podaj założenia, z jakich korzystałeś i skomentuj czy wydają Ci się uprawnione.

Przedziały dla średniej i odchylenia standardowego

Studentyzowany przedział ufności:

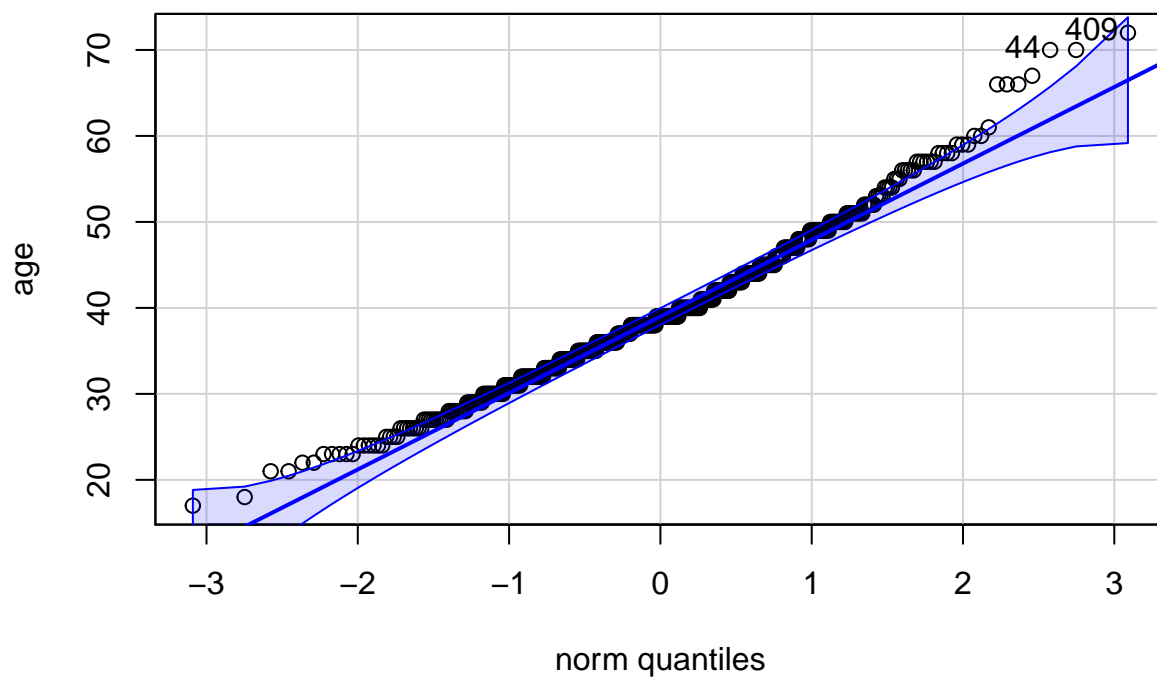
$\left(\bar{X} - \frac{t(1-\alpha/2, n-1)}{\sqrt{n}} \hat{S}, \bar{X} + \frac{t(1-\alpha/2, n-1)}{\sqrt{n}} \hat{S} \right)$ gdzie \bar{X} to średnia, \hat{S} to pierwiastek z *nieobciążonego* estymatora wariancji, a $t(1 - \alpha/2, n - 1)$ to kwantyl na poziomie $1 - \alpha/2$ dla rozkładu t Studenta o $n - 1$ stopniach swobody.

Asymptotyczny przedział ufności:

$\left(\bar{X} - \frac{q(1-\alpha/2)}{\sqrt{n}} \hat{S}, \bar{X} + \frac{q(1-\alpha/2)}{\sqrt{n}} \hat{S} \right)$ gdzie $q(1 - \alpha/2)$ jest kwantylem na poziomie $1 - \alpha/2$ ze standardowego rozkładu normalnego.

Założenia: (rozkład normalny)

```
alpha <- 0.01
#ocena czy zmienna age ma rozkład normalny
age <- df$age
qqPlot(age)
```



```
## [1] 409 44
```

```
n <- length(age)
shapiro.test(age)
```

```
##
## Shapiro-Wilk normality test
##
## data: age
## W = 0.98179, p-value = 6.589e-06
```

P-value 0.0163325 jest większe niż 1% co w połączeniu z analizą wykresu qqPlot() implikuje, że zmienna nie różni się istotnie od rozkładu normalnego. Innymi słowy możemy założyć rozkład normalny zmiennej age.

```
rightstud <- mean(age) + 1 / sqrt(n) *
  sd(age) *
  qt(1 - alpha / 2, (n - 1))
leftstud <- mean(age) - 1 / sqrt(n) *
  sd(age) *
```

```

qt(1 - alpha / 2, (n - 1))

rightasympt <- mean(age) + (qnorm(1 - alpha / 2)) / sqrt(n) * sd(age)
leftasympt <- mean(age) - (qnorm(1 - alpha / 2)) / sqrt(n) * sd(age)

sprintf("Studentyzowany: (%f, %f)", leftstud, rightstud)

## [1] "Studentyzowany: (38.446003, 40.521997)"

sprintf("Asymptotyczny: (%f, %f)", leftasympt, rightasympt)

## [1] "Asymptotyczny: (38.449973, 40.518027)"

```

Przedziały dla kwantyli 1/4, 1/2, 3/4

```

lv <- levels(quantcut(age, q = seq(0, 1, 1/4)))
lv

## [1] "[17,33]" "(33,39]" "(39,45]" "(45,72]"

jmuOutlier::quantileCI(age, c(0.25, 0.5, 0.75), 0.99) #przedziały ufności dla kwantyli

##      lower upper
## 0.25     32     35
## 0.5      38     40
## 0.75     43     47

```

Trzy hipotezy istotności

5. Przetestuj na poziomie istotności 0.01 trzy hipotezy istotności:

1. różnicy między średnią wartością wybranej zmiennej dla kobiet i dla mężczyzn;
2. zależności między dwiema zmiennymi ilościowymi;
3. zależności między dwiema zmiennymi jakościowymi.

Ponadto,

4. przetestuj hipotezę o zgodności z konkretnym rozkładem parametrycznym dla wybranej zmiennej (np. "zmienna A ma rozkład wykładniczy z parametrem 10"). Podaj założenia, z jakich korzystałeś i skomentuj czy wydają Ci się uprawnione.

Różnica między średnią wartością wzrostu dla kobiet i dla mężczyzn

Test t dla prób niezależnych: $t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}}$, gdzie \bar{X} , \bar{Y} - średnie arytmetyczne, s_X^2 , s_Y^2 - nieobciążone estymatory wariancji, n_X , n_Y - liczby obserwacji

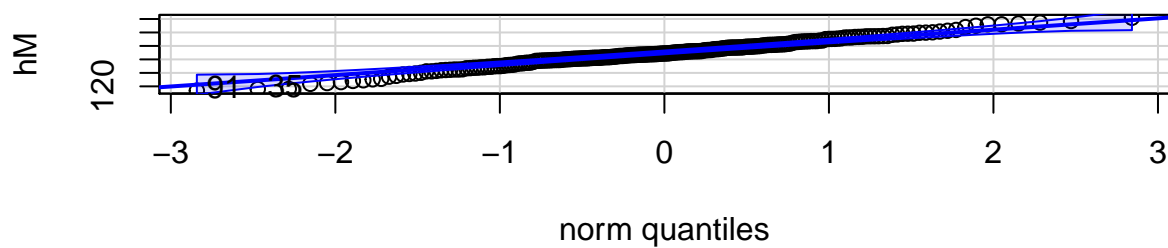
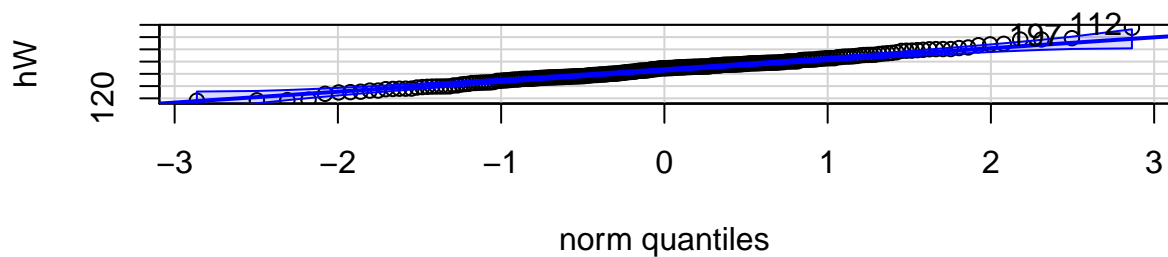
Hipoteza zerowa H_0 : Średni wzrost dla kobiet nie różni się od średniego wzrostu dla mężczyzn.
 Hipoteza alternatywna H_1 : Średni wzrost dla różnych płci różni się.

Jedyne założenie do sprawdzenia: czy próby pochodzą z rozkładu normalnego.

```
hW <- df[df$gender=="woman",]$height
hM <- df[df$gender=="man",]$height
par(mfrow=c(2, 1))
qqPlot(hW)
```

```
## [1] 112 197
```

```
qqPlot(hM)
```



```
## [1] 91 35
```

```
shapiro.test(hW)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  hW
## W = 0.99402, p-value = 0.4613
```

```
shapiro.test(hM)
```

```
##  
## Shapiro-Wilk normality test  
##  
## data: hM  
## W = 0.99298, p-value = 0.3721
```

P-value dla obu grup jest większe niż 10%, więc możemy założyć, że obie grupy dla zmiennej height mają rozkład normalny.

```
n <- length(hM)  
m <- length(hW)  
alpha <- 0.01  
#unbiased variance estimators  
unbiased_estX <- 1/(n-1)*sum((hM-mean(hM))^2)  
unbiased_estY <- 1/(m-1)*sum((hW-mean(hW))^2)  
  
a <- unbiased_estX/n + unbiased_estY/m  
t <- (mean(hM) - mean(hW))/sqrt(a) #test statistic  
  
def <- a^2/(1/(n-1)*(unbiased_estX/n)^2+1/(m-1)*(unbiased_estY/m)^2) #deg of freedom  
  
#two-tailed hypothesis  
pval <- 2*pt(t, n+m-2, lower.tail = FALSE)  
  
#confidence interval  
lowerBound <- qt(alpha/2, n+m-2)  
upperBound <- qt(1-alpha/2, n+m-2)  
  
if(alpha > pval) {  
  print("H0 rejected.")  
}else {  
  print("There is not enough evidence to reject H_0")  
}
```

```
## [1] "There is not enough evidence to reject H_0"
```

```
sprintf("test statistic = %f , p-value=%f, confidence interval = (%f, %f)", t, pval, lowerBound, upperBound)
```

```
## [1] "test statistic = 1.309895 , p-value=0.190885, confidence interval = (-2.586559, 2.586559)"
```

```
x <- seq(-3, 3, by = 0.01)  
t_dense <- dt(x, n+m-2)  
  
plot(x, t_dense, type='l', xlab="x value",  
      ylab="Density", main="Student's density")  
  
i <- x >= lowerBound & x <= upperBound  
lines(x, t_dense)
```

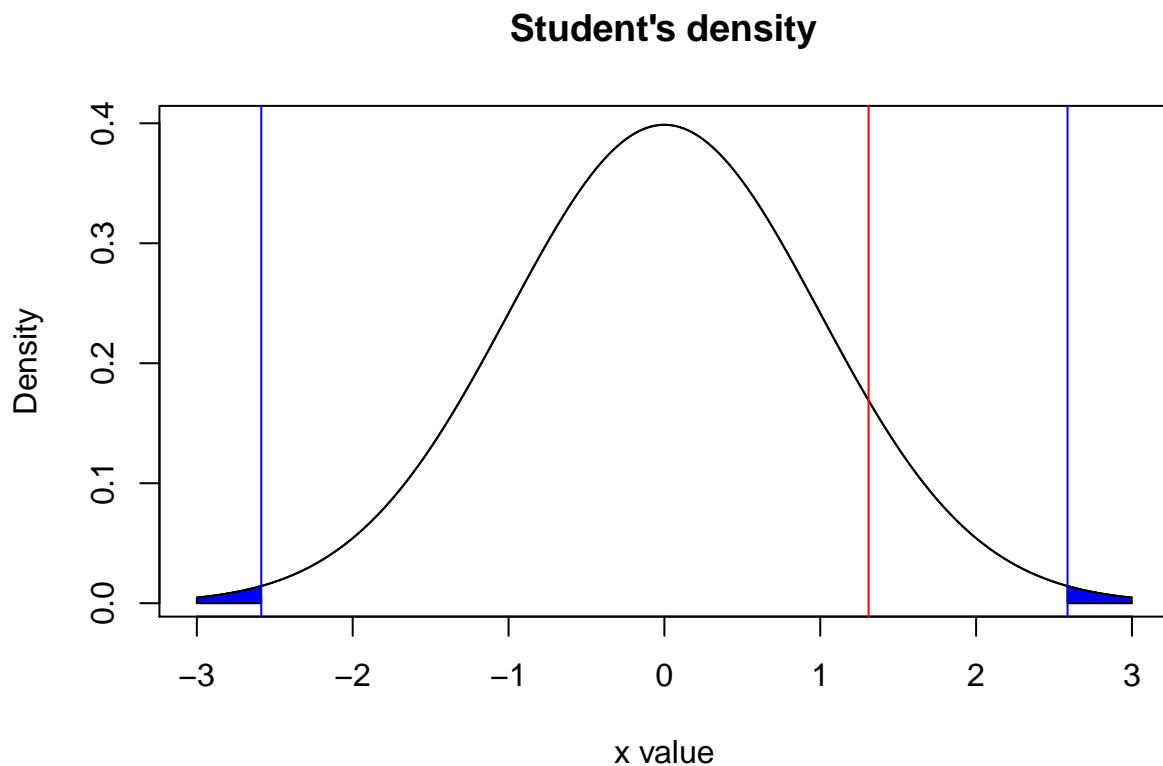
```

polygon(x = c(-3, seq(-3, lowerBound, 0.01), lowerBound),
        y = c(0, dt(seq(-3, lowerBound, 0.01), n+m-2), 0),
        col = 'blue')

polygon(x = c(upperBound, seq(upperBound, 3, 0.01), 3),
        y = c(0, dt(seq(upperBound, 3, 0.01), n+m-2), 0),
        col = 'blue')

area <- pt(upperBound, n+m-2) - pt(lowerBound, n+m-2)
abline(v=t, col="red")
abline(v=lowerBound, col="blue")
abline(v=upperBound, col="blue")

```



Nie ma powodów do odrzucenia hipotezy zerowej - statystyka testowa nie wpada do obszaru krytycznego. Średni wzrost dla kobiet nie różni się od średniego wzrostu dla mężczyzn. ## Zależność między dwiema zmiennymi ilościowymi.

Współczynnik korelacji Pearsona: (posłużę się gotową implementacją)

Hipoteza zerowa H_0 : Zmienne nie są skorelowane $\rho = 0$.

Hipoteza alternatywna H_1 : Zmienne są skorelowane $\rho \neq 0$.

Założenia: (zmienne mają rozkład normalny, zależność liniowa między zmiennymi)

```

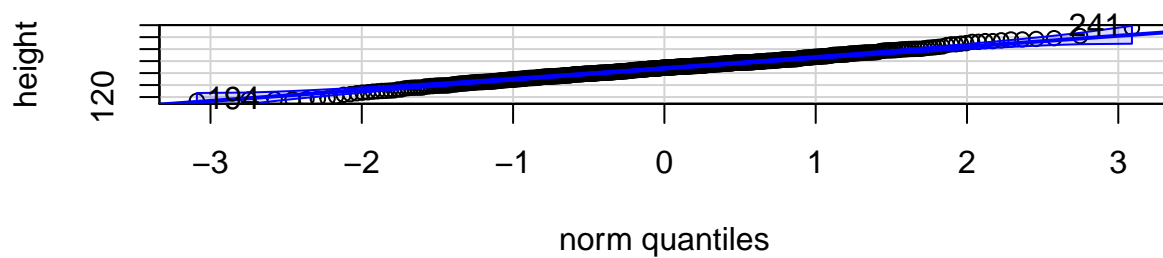
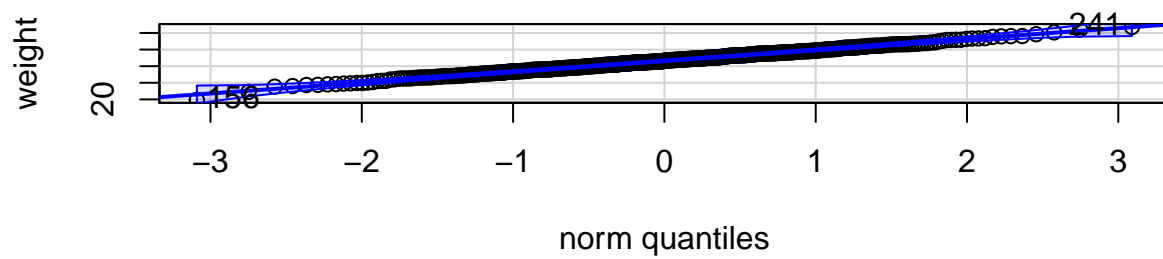
weight <- df$weight
height <- df$height

```

```
alpha <- 0.01
par(mfrow=c(2, 1))
qqPlot(weight)
```

```
## [1] 156 241
```

```
qqPlot(height)
```



```
## [1] 241 194
```

```
shapiro.test(weight)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  weight
## W = 0.99895, p-value = 0.9939
```

```
shapiro.test(height)
```

```
##
```

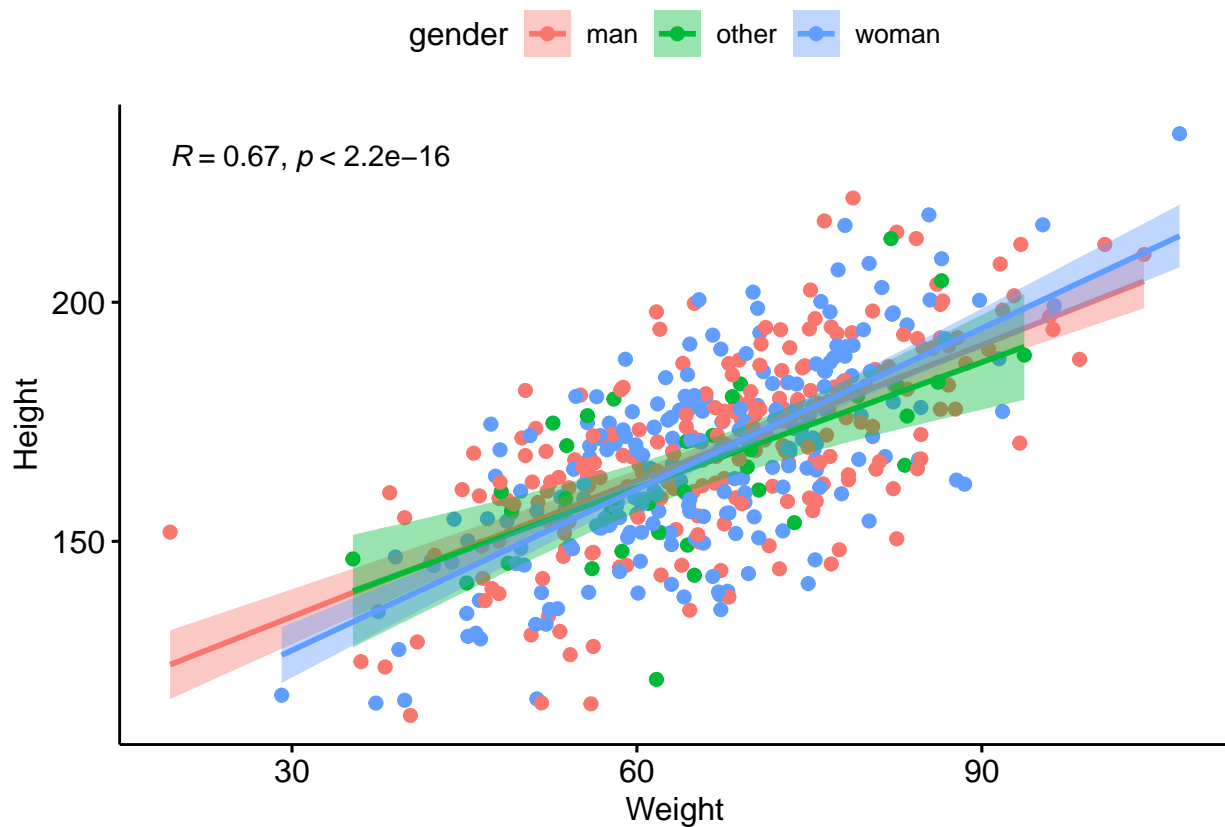
```
## Shapiro-Wilk normality test
##
## data: height
## W = 0.99662, p-value = 0.3778
```

Widzimy, że zmienne mają rozkład normalny (duże p-value z testu Shapiro-Wilka + wygląd wykresów qqPlot()).

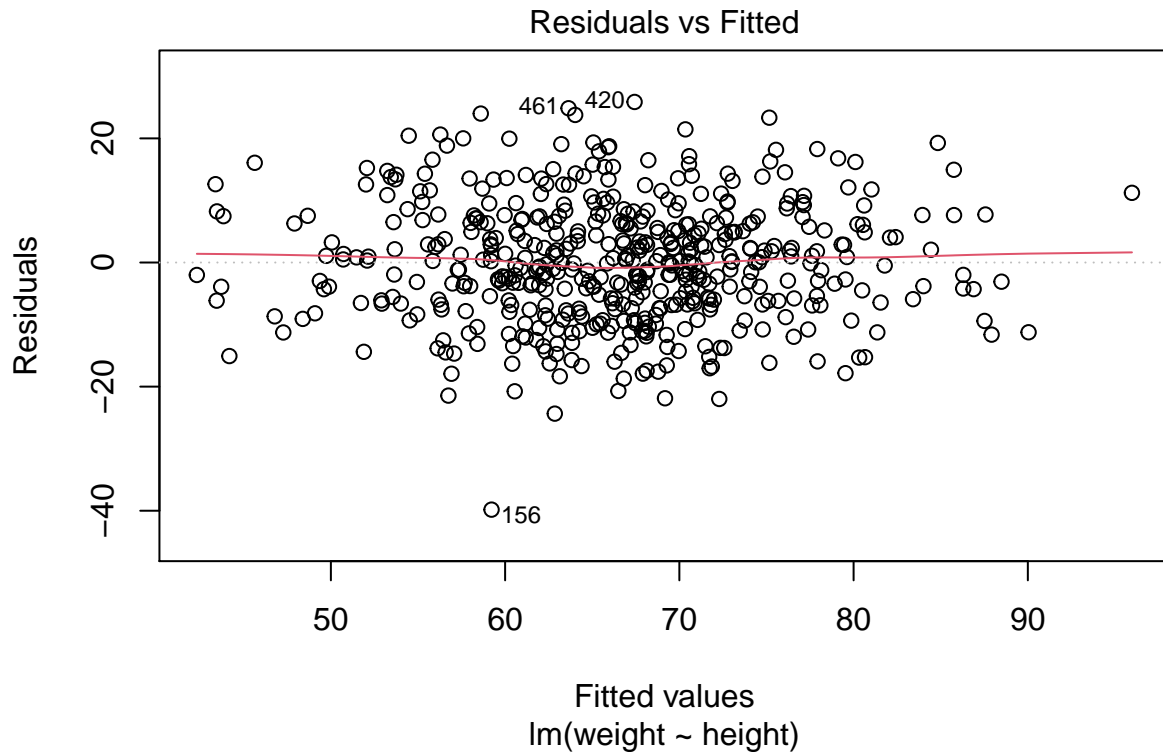
Scatterplot - zależność między zmiennymi:

```
p1 <- ggscatter(df, x = "weight", y = "height",
  add = "reg.line", conf.int = TRUE,
  cor.coef = TRUE, cor.method = "pearson", color="gender",
  xlab = "Weight", ylab = "Height")
p1
```

```
## 'geom_smooth()' using formula 'y ~ x'
```



```
linearity_test <- lm(weight~height, df)
plot(linearity_test, which=1)
```

```
#summary(linearity_test)
```

Liniowa zależność między zmiennymi weight i height - założenie spełnione.

```
test <- cor.test(x=weight, y=height, method="pearson", alternative = "two.sided", conf.level = 1-alpha)
sprintf("test statistic = %f , p-value=%f, confidence interval = (%f, %f), cor = %f", test$statistic, test$p.value,
```

```
## [1] "test statistic = 20.123937 , p-value=0.000000, confidence interval = (0.600957, 0.728596), cor = 0.999999"
```

```
if (alpha > test$p.value) {
  print("H0 rejected.")
}else {
  print("There is not enough evidence to reject H_0")
}
```

```
## [1] "H0 rejected."
```

P-value wynosi mniej niż ustalony poziom istotności 1% - odrzucamy hipotezę zerową. Możemy zatem wywnioskować, że zmienne weight i height **są istotnie skorelowane** co potwierdza wcześniejszy korelogram z pierwszego podpunktu.

Zależność między dwiema zmiennymi ilościowymi.

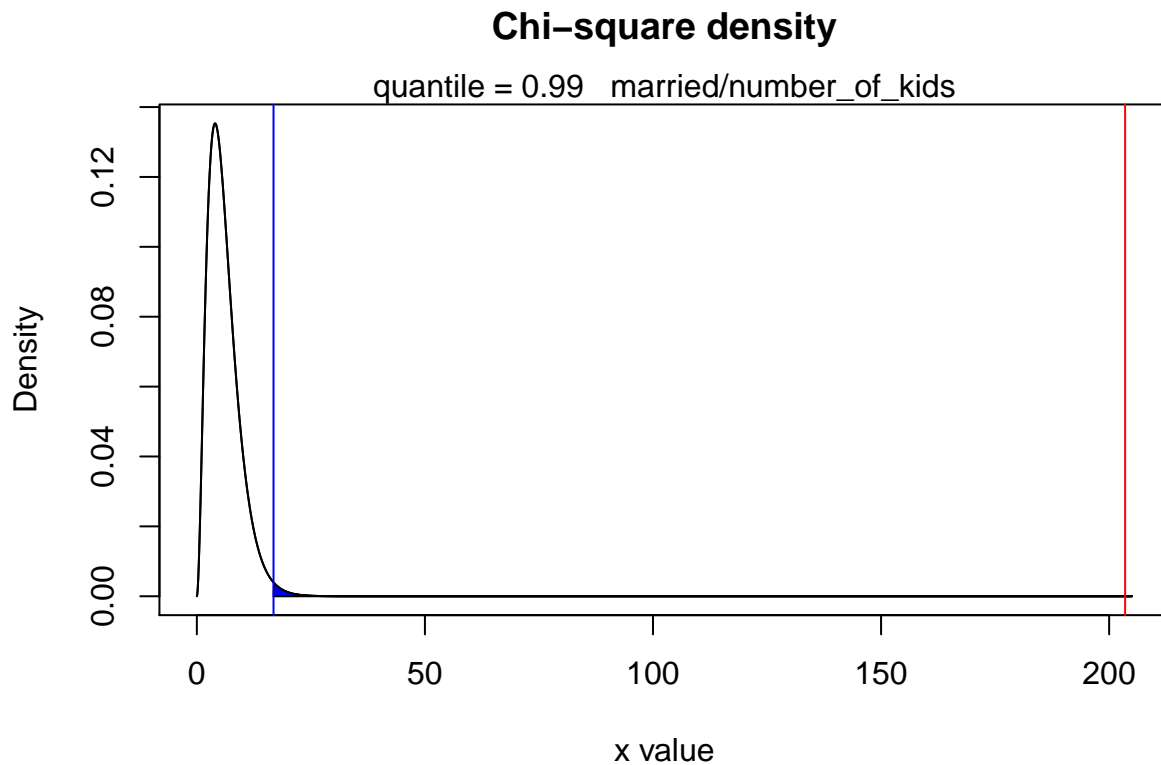
Wykorzystam funkcję napisaną wcześniej.

Hipoteza zerowa H_0 : Zmienne są **niezależne**.

Hipoteza alternatywna H_1 : Zmienne są **zależne**.

Założenia: zmienne wzajemnie się wykluczają - spełnione.

```
alpha <- 0.01
g <- testchi(df$married, df$number_of_kids, sq=205, t='married/number_of_kids', alpha = alpha)
```



```
sprintf("KIDS/MARRIED, test statistic = %f , p-value = %f, confidece interval = [-infinity, %f]", g[[1]
```

```
## [1] "KIDS/MARRIED, test statistic = 203.501380 , p-value = 0.000000, confidece interval = [-infinity
```

Odrzucamy hipotezę zerową, wartość statystyki testowej wpada do obszaru krytycznego. Zmienne married i number_of_kids są zależne.

Hipoteza o zgodności z konkretnym rozkładem parametrycznym

Przeprowadzę test dla zmiennej number_of_kids.

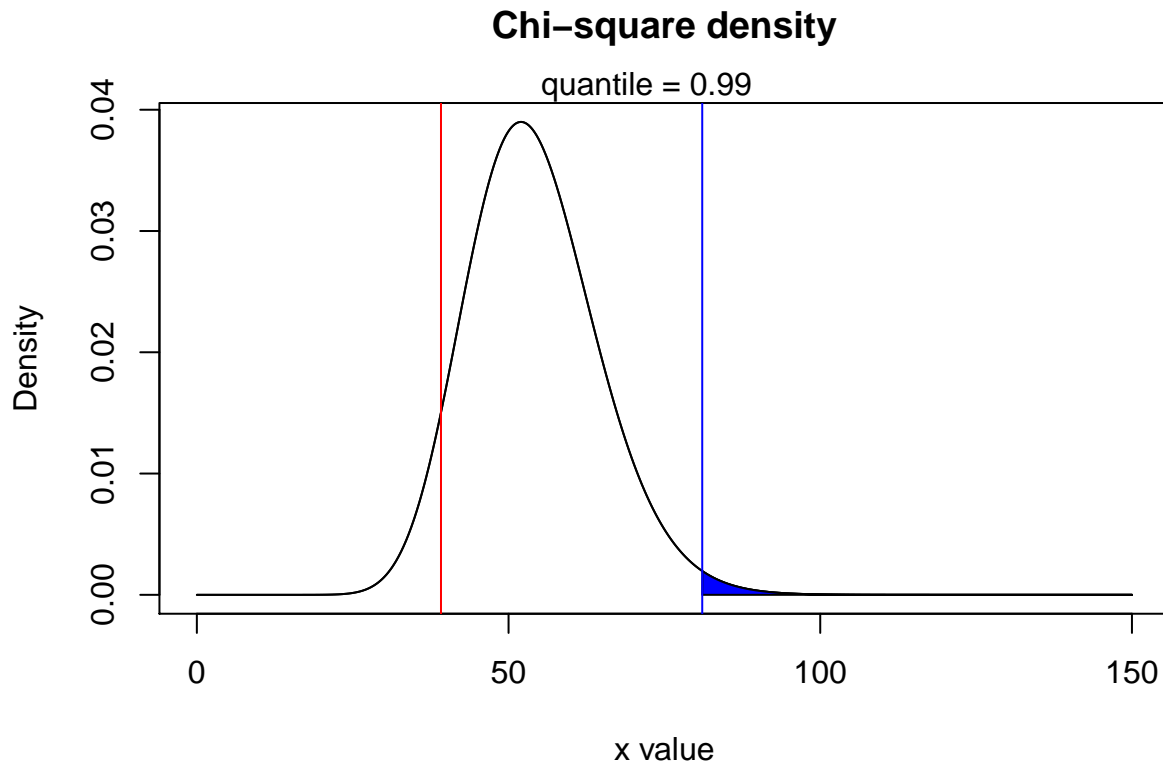
Niech zmienna X ma rozkład Pascala z parametrami $p = 0.7$ i $r = 5$. Przyjmijmy jedną zmienną jako wartość oczekiwaną a drugą jako obserwację dla testu chi2.

Hipoteza zerowa H_0 : Zmienne są niezależne.

Hipoteza alternatywna H_1 : Zmienne **nie** są niezależne.

Założenia: zmienne wzajemnie się wykluczają - spełnione.

```
kids <- df$number_of_kids
x <- rbinom(length(kids), 5, 0.7)
x <- factor(x)
yf <- testchi(kids, x, sq=150, t=' ', alpha=alpha)
```



```
sprintf("KIDS/MARRIED, test statistic = %f , p-value = %f, confidece interval = [-infinity, %f]", yf[[1],
```

```
## [1] "KIDS/MARRIED, test statistic = 39.162117 , p-value = 0.935591, confidece interval = [-infinity,
```

Nie mamy podstaw by odrzucić hipotezę zerową, więc zmienne `number_of_kids` i `X` są niezależne, więc `number_of_kids` nie ma rozkładu Pascala z parametrami $p = 0.7$ i $r = 5$.

Regresja liniowa

6. Oszacuj model regresji liniowej, przyjmując za zmienną zależną (y) wydatki domowe (expenses) a jako zmienne niezależne (x) przyjmując pozostałe zmienne.

Rozważ, czy konieczne są transformacje zmiennych lub zmiennej objaśnianej. Podaj **RSS**, **R^2** , **p-wartości** i **oszacowania współczynników w pełnym modelu** (w modelu zawierającym wszystkie zmienne). Następnie wybierz jedną zmienną objaśniającą, którą można by z pełnego modelu odrzucić (która najgorzej tłumaczy expenses). Aby dokonać wyboru takiej zmiennej, dla każdej ze zmiennych objaśniających sprawdź:

- Jaką ma p-wartość w pełnym modelu?
- O ile zmniejsza się R^2 , gdy ją usuniemy z pełnego modelu?
- O ile zwiększa się RSS, gdy ją usuniemy z pełnego modelu?

Opisz wnioski.

Oszacuj model ze zbiorem zmiennych objaśniających pomniejszonym o wybraną zmienną. Sprawdź czy w otrzymanym przez Ciebie modelu spełnione są założenia modelu liniowego i przedstaw na wykresach diagnostycznych: wykresie zależności reszt od zmiennej objaśnianej, na wykresie reszt studentyzowanych i na wykresie dźwigni i przedyskutuj, czy są spełnione.

Do tego zadania konieczne będzie zamiana zmiennych katerycznych funkcją `factor()`- zrobiłam to już wcześniej przy okazji wczytywania danych. W przeciwnym razie zmienna np. `number_of_kids` mogłaby zostać potraktowana przez funkcję `lm()` jako zmienna ilościowa.

Pełny model bez zmian

```
linear_regression <- lm( expenses ~ age + weight + height + gender + married + number_of_kids + pet, df)
summary(linear_regression)
```

```
##
## Call:
## lm(formula = expenses ~ age + weight + height + gender + married +
##     number_of_kids + pet, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -758.69 -119.55   3.06  128.17  885.29
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2299.7769   101.8435  -22.581 < 2e-16 ***
## age           57.5620     1.0753   53.529 < 2e-16 ***
## weight        1.3272     0.9963    1.332  0.18345
## height        2.0825     0.6572    3.169  0.00163 **
## genderother   48.1590    37.7395    1.276  0.20254
## genderwoman  -17.6926    20.1808   -0.877  0.38108
## marriedTRUE  -17.7722    26.3260   -0.675  0.49995
## number_of_kids1  14.9060    25.9026    0.575  0.56525
## number_of_kids2 -56.9082    30.1096   -1.890  0.05935 .
## number_of_kids3  25.9653    38.2434    0.679  0.49750
## number_of_kids4 -52.4638    46.6479   -1.125  0.26129
## number_of_kids5 -31.8876    72.4424   -0.440  0.66000
## number_of_kids6 -115.3335    93.2361   -1.237  0.21669
## petdog        34.8071    30.1051    1.156  0.24818
## petferret     413.5364    36.2367   11.412 < 2e-16 ***
## pethedgehog   244.8257    35.8244    6.834  2.5e-11 ***
## petnone       24.2985    26.1476    0.929  0.35321
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 212.9 on 483 degrees of freedom
## Multiple R-squared:  0.8637, Adjusted R-squared:  0.8592
## F-statistic: 191.3 on 16 and 483 DF,  p-value: < 2.2e-16
```

```
mean(linear_regression$residuals) #wartość średnia reszty
```

```
## [1] 8.637813e-16
```

Prawidłowość formy funkcyjnej

Prawidłowość formy funkcyjnej modelu weryfikujemy, wykorzystując test RESET (Regression Specification Error Test) Ramseya.

Test polega na tym, że do modelu przeprowadza się regresję pomocniczą z dodanymi iloczynami zm. objaśniających po czym testem F sprawdzamy łączną istotność tych dodatkowych członów.

Hipoteza zerowa $H_0: X\beta + \epsilon$.

Hipoteza alternatywna $H_1: f(X\beta) + \epsilon$.

```
resettest(linear_regression, power=2:6, type = c("fitted", "regressor", "princomp"))
```

```
##
```

```
## RESET test
```

```
##
```

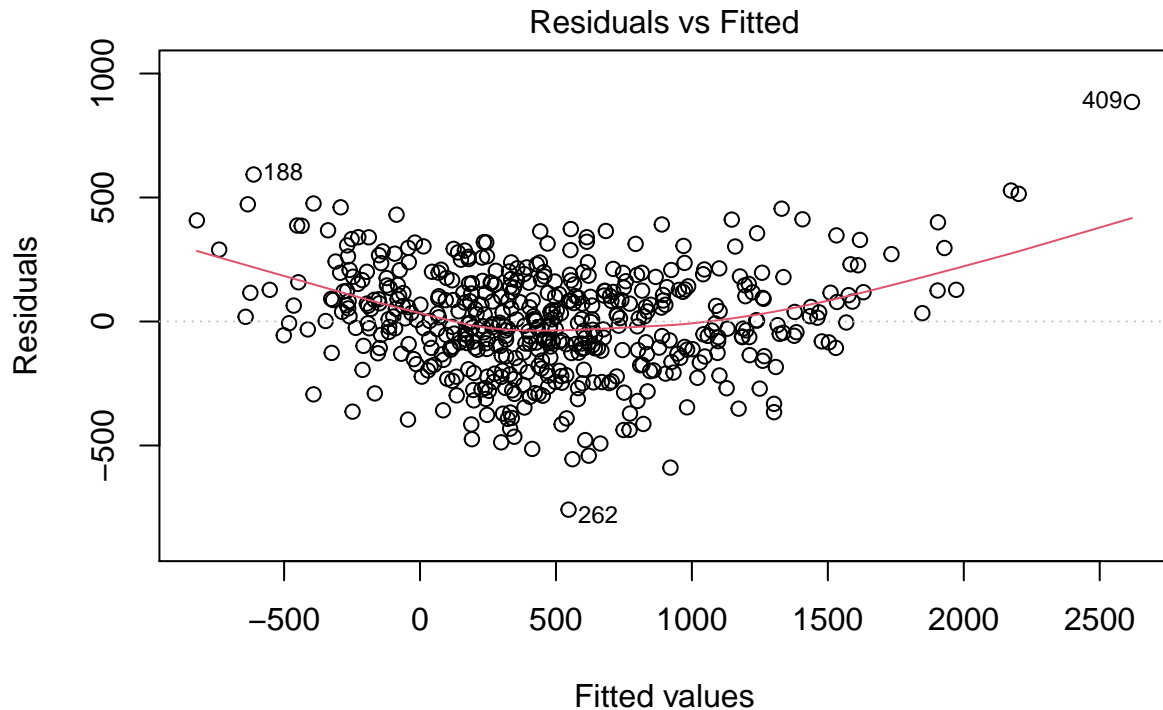
```
## data: linear_regression
```

```
## RESET = 19.441, df1 = 5, df2 = 478, p-value < 2.2e-16
```

Brak podstaw do odrzucenia hipotezy zerowej dla poziomu istotności 1%.

Nieliniowość w danych, residua nie są rozrzucone równo wokół 0:

```
plot(linear_regression, which = 1)
```



lm(expenses ~ age + weight + height + gender + married + number_of_kids + p ...

Przedziały ufności:

```
confint(linear_regression, level = 0.99)
```

##		0.5 %	99.5 %
##	(Intercept)	-2563.1487967	-2036.404917
##	age	54.7811353	60.342913
##	weight	-1.2493246	3.903807
##	height	0.3830433	3.781893
##	genderother	-49.4370254	145.754985
##	genderwoman	-69.8809813	34.495869
##	marriedTRUE	-85.8523275	50.307948
##	number_of_kids1	-52.0793817	81.891433
##	number_of_kids2	-134.7730157	20.956701
##	number_of_kids3	-72.9339491	124.864596
##	number_of_kids4	-173.0974669	68.169920
##	number_of_kids5	-219.2271101	155.451881
##	number_of_kids6	-356.4464843	125.779531
##	petdog	-43.0461111	112.660369
##	petferret	319.8266502	507.246079
##	pethedgehog	152.1820547	337.469348
##	petnone	-43.3205144	91.917542

Regresja liniowa na pełnym modelu: Oszacowania współczynników (Estimate)

```
linear_regression$coefficients
```

```
##      (Intercept)          age          weight          height  genderother
## -2299.776857      57.562024      1.327241      2.082468      48.158980
##      genderwoman marriedTRUE number_of_kids1 number_of_kids2 number_of_kids3
## -17.692556      -17.772190      14.906026      -56.908157      25.965324
## number_of_kids4 number_of_kids5 number_of_kids6          petdog          petferret
## -52.463774      -31.887615      -115.333476      34.807129      413.536364
##      pethedgehog          petnone
##      244.825701      24.298514
```

Residual sum of squares (RSS):

```
RSS <- sum(resid(linear_regression)^2)
print(RSS)
```

```
## [1] 21901205
```

Współczynnik determinacji (R^2):

```
R2 <- summary(linear_regression)$r.squared
print(round(R2, 3))
```

```
## [1] 0.864
```

P-wartości w pełnym modelu dla zmiennych:

```
summary(linear_regression)$coefficients[,4]
```

```
##      (Intercept)          age          weight          height  genderother
## 1.323336e-77  3.311685e-205  1.834474e-01  1.627050e-03  2.025367e-01
##      genderwoman marriedTRUE number_of_kids1 number_of_kids2 number_of_kids3
## 3.810836e-01  4.999464e-01  5.652457e-01  5.935204e-02  4.974956e-01
## number_of_kids4 number_of_kids5 number_of_kids6          petdog          petferret
## 2.612852e-01  6.600046e-01  2.166867e-01  2.481765e-01  7.146571e-27
##      pethedgehog          petnone
## 2.497126e-11  3.532076e-01
```

Variance inflation factor (VIF) - mierzy o ile wariancja oszacowanego współczynnika regresji jest zwiększona z powodu kolinearności.

Współczynnik VIF dla $\hat{\beta}_i$: $VIF_i = \frac{1}{1-R_i^2}$, gdzie R_i^2 - współczynnik determinacji dla X_i

```
vif(linear_regression)
```

```
##      GVIF Df GVIF^(1/(2*Df))
## age      1.025355 1      1.012598
## weight   1.831392 1      1.353289
## height   1.836310 1      1.355105
## gender    1.055675 2      1.013637
## married   1.729303 1      1.315030
## number_of_kids 1.871981 6      1.053639
## pet       1.070701 4      1.008576
```

```
mean(vif(linear_regression))
```

```
## [1] 1.644409
```

One-Hot Encoding

Funkcja `lm()` sama przekształca zmienne kateryczne. Jednak żeby usunąć, którąś z takich zakodowanych zmiennych należy wcześniej je samodzielnie zakodować. Funkcja `one_hot()` z biblioteki `mltools` przekształca zmienną kateryczną metodą One-Hot Encoding bez odrzucania żadnych kolumn bazowych. Do modelu `linear_regression2` nie wprowadzono poziomów `gender_woman`, `married_FALSE`, `number_of_kids_1`, `pet_none` (najczęściej występujące w swoich katerycznych zmiennych) - żeby nie wprowadzać zmiennych, które wiadomo, że będą skorelowane.

```
df_ohe <- one_hot(as.data.table(df), dropUnusedLevels=TRUE) #cols DEFAULT = "auto" encodes all unordere
head(df_ohe)
```

```
##      age weight height gender_man gender_other gender_woman married_FALSE
## 1:   25   61.7 121.12          0           1           0           1
## 2:   37   63.9 145.00          1           0           0           0
## 3:   41   50.2 145.03          0           0           1           0
## 4:   43   72.4 179.90          1           0           0           1
## 5:   26   78.4 163.91          1           0           0           1
## 6:   49   59.4 151.86          0           0           1           0
##      married_TRUE number_of_kids_0 number_of_kids_1 number_of_kids_2
## 1:              0              0              0              1
## 2:              1              0              0              0
## 3:              1              0              0              1
## 4:              0              0              1              0
## 5:              0              0              1              0
## 6:              1              0              0              1
##      number_of_kids_3 number_of_kids_4 number_of_kids_5 number_of_kids_6 pet_cat
## 1:                  0                  0                  0                  0      0
## 2:                  0                  0                  0                  1      0
## 3:                  0                  0                  0                  0      0
## 4:                  0                  0                  0                  0      0
## 5:                  0                  0                  0                  0      0
## 6:                  0                  0                  0                  0      0
##      pet_dog pet_ferret pet_hedgehog pet_none  expenses
## 1:         0         1         0         0    23.44299
## 2:         1         0         0         0    96.83683
## 3:         0         0         1         0   312.67693
## 4:         1         0         0         0   447.42838
## 5:         0         0         1         0  -78.22799
## 6:         0         1         0         0 1241.98263
```

```
linear_regression2 <- lm(expenses~.-gender_woman-married_FALSE-number_of_kids_1-pet_none, df_ohe)
summary(linear_regression2)
```

```
##
```

```
## Call:
```

```
## lm(formula = expenses ~ . - gender_woman - married_FALSE - number_of_kids_1 -
```



```
##      pet_none, data = df_ohe)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -758.69 -119.55    3.06  128.17  885.29
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2278.2649    97.5326  -23.359 < 2e-16 ***
## age             57.5620     1.0753   53.529 < 2e-16 ***
## weight         1.3272     0.9963    1.332 0.18345
## height         2.0825     0.6572    3.169 0.00163 **
## gender_man     17.6926    20.1808    0.877 0.38108
## gender_other   65.8515    37.5648    1.753 0.08023 .
## married_TRUE  -17.7722    26.3260   -0.675 0.49995
## number_of_kids_0 -14.9060    25.9026   -0.575 0.56525
## number_of_kids_2 -71.8142    27.8767   -2.576 0.01029 *
## number_of_kids_3  11.0593    35.5787    0.311 0.75606
## number_of_kids_4 -67.3698    44.3679   -1.518 0.12956
## number_of_kids_5 -46.7936    70.5569   -0.663 0.50752
## number_of_kids_6 -130.2395    92.0909   -1.414 0.15793
## pet_cat        -24.2985    26.1476   -0.929 0.35321
## pet_dog         10.5086    26.5640    0.396 0.69258
## pet_ferret      389.2379    33.1680   11.735 < 2e-16 ***
## pet_hedgehog    220.5272    33.0630    6.670 7.02e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 212.9 on 483 degrees of freedom
## Multiple R-squared:  0.8637, Adjusted R-squared:  0.8592
## F-statistic: 191.3 on 16 and 483 DF, p-value: < 2.2e-16
```

Duże pvalue pet_dog i number_of_kids_3 (nieistotne statystycznie - duże pvalue). Stała okazała się istotna statystycznie (pval < 2e-16) **Podobnie age** (< 2e-16), height(< 0.00163 **), number_of_kids_2 (< 0.01029), pet_ferret (< 2e-16), pet_hedgehog (< 7.02e-11 *)

Zmienne są łącznie istotne statystycznie - pval < 2.2e-16 w teście F Residual sum of squares (RSS):

Residual sum of squares (RSS):

```
RSS <- sum(resid(linear_regression2)^2)
print(RSS)
```

```
## [1] 21901205
```

Współczynnik determinacji (R²):

```
R2 <- summary(linear_regression2)$r.squared
print(round(R2, 3))
```

```
## [1] 0.864
```

```
summary(linear_regression2)$coefficients[,4]
```

```
##      (Intercept)          age          weight          height
## 2.557938e-81 3.311685e-205 1.834474e-01 1.627050e-03
##      gender_man    gender_other    married_TRUE    number_of_kids_0
## 3.810836e-01 8.023463e-02 4.999464e-01 5.652457e-01
## number_of_kids_2    number_of_kids_3    number_of_kids_4    number_of_kids_5
## 1.028706e-02 7.560560e-01 1.295589e-01 5.075161e-01
## number_of_kids_6          pet_cat          pet_dog          pet_ferret
## 1.579329e-01 3.532076e-01 6.925777e-01 3.739870e-28
##      pet_hedgehog
## 7.021587e-11
```

R² i RSS po usunięciu zmiennych

O ile zmniejsza się R², gdy usuniemy zmienną z pełnego modelu?

O ile zwiększa się RSS, gdy usuniemy zmienną z pełnego modelu?

Bedziemy odrzucać po jednej zmiennej z pełnego modelu i patrzeć jak bardzo zmienia się R² i RSS:

```
df_ohe <- data.table(df_ohe)

features <- c("age", "weight", "height", "gender_other", "gender_man", "married_TRUE", "number_of_kids_0",
             "number_of_kids_3", "number_of_kids_4", "number_of_kids_5", "pet_cat", "pet_dog", "pet_ferret")
df_features <- df_ohe[, c("age", "weight", "height", "gender_other", "gender_man", "married_TRUE", "number_of_kids_0",
                        "number_of_kids_3", "number_of_kids_4", "number_of_kids_5", "pet_cat", "pet_dog", "pet_ferret")]

rsq <- NULL
rsss <- NULL
i<-1
for(c in features) {
  fo <- as.formula(paste("expenses", "~.-", c))
  r <- do.call("lm", list(fo, quote(df_features)))
  rsq[i] <- summary(r)$r.squared
  rsss[i] <- sum(resid(r)^2)
  i <- i+1
}

R <- data.frame(features, rsq, rep(R2, length(features))-rsq, rsss, rsss-rep(RSS, length(features)))
colnames(R) <- c("deleted_column", "R_2", "difference_R2", "RSS", "difference_RSS")
R2ord <- R[order(R$difference_R2), ]
R2ord
```

```
##      deleted_column      R_2 difference_R2      RSS difference_RSS
## 10 number_of_kids_3 0.86367537 2.726562e-05 21905586 4.381229e+03
## 14          pet_dog 0.86365847 4.416154e-05 21908301 7.096182e+03
## 8  number_of_kids_0 0.86360918 9.344936e-05 21916221 1.501609e+04
## 12 number_of_kids_5 0.86357852 1.241181e-04 21921149 1.994415e+04
## 6   married_TRUE 0.86357403 1.286039e-04 21921870 2.066496e+04
## 5      gender_man 0.86348574 2.168931e-04 21936057 3.485189e+04
## 13      pet_cat 0.86345895 2.436882e-04 21940363 3.915751e+04
## 2      weight 0.86320187 5.007606e-04 21981671 8.046569e+04
## 7  number_of_kids_6 0.86313823 5.644070e-04 21991898 9.069283e+04
```

```
## 11 number_of_kids_4 0.86305201 6.506286e-04 22005753 1.045475e+05
## 4      gender_other 0.86283545 8.671800e-04 22040550 1.393445e+05
## 9  number_of_kids_2 0.86182989 1.872746e-03 22202131 3.009257e+05
## 3      height 0.86086886 2.833778e-03 22356556 4.553510e+05
## 16    pet_hedgehog 0.85114867 1.255397e-02 23918463 2.017258e+06
## 15      pet_ferret 0.82484011 3.886253e-02 28145905 6.244700e+06
## 1          age 0.05513236 8.085703e-01 151827880 1.299267e+08
```

```
RSSord <- R[order(R$difference_RSS), ]
RSSord
```

```
##      deleted_column      R_2 difference_R2      RSS difference_RSS
## 10 number_of_kids_3 0.86367537 2.726562e-05 21905586 4.381229e+03
## 14      pet_dog 0.86365847 4.416154e-05 21908301 7.096182e+03
## 8  number_of_kids_0 0.86360918 9.344936e-05 21916221 1.501609e+04
## 12 number_of_kids_5 0.86357852 1.241181e-04 21921149 1.994415e+04
## 6  married_TRUE 0.86357403 1.286039e-04 21921870 2.066496e+04
## 5  gender_man 0.86348574 2.168931e-04 21936057 3.485189e+04
## 13      pet_cat 0.86345895 2.436882e-04 21940363 3.915751e+04
## 2  weight 0.86320187 5.007606e-04 21981671 8.046569e+04
## 7  number_of_kids_6 0.86313823 5.644070e-04 21991898 9.069283e+04
## 11 number_of_kids_4 0.86305201 6.506286e-04 22005753 1.045475e+05
## 4  gender_other 0.86283545 8.671800e-04 22040550 1.393445e+05
## 9  number_of_kids_2 0.86182989 1.872746e-03 22202131 3.009257e+05
## 3  height 0.86086886 2.833778e-03 22356556 4.553510e+05
## 16    pet_hedgehog 0.85114867 1.255397e-02 23918463 2.017258e+06
## 15      pet_ferret 0.82484011 3.886253e-02 28145905 6.244700e+06
## 1          age 0.05513236 8.085703e-01 151827880 1.299267e+08
```

Najmniejsza różnica w R^2 i RSS (3 i 5 kolumna) dla number_of_kids_3.

Usunięcie zmiennej najgorzej tłumaczącą expenses

```
#wyrzucam dodatkowo number_of_kids_3
linear_regression3 <- lm(expenses ~.-number_of_kids_1-married_TRUE-gender_woman-pet_none-number_of_kids_3, data = df_ohe)
summary(linear_regression3)
```

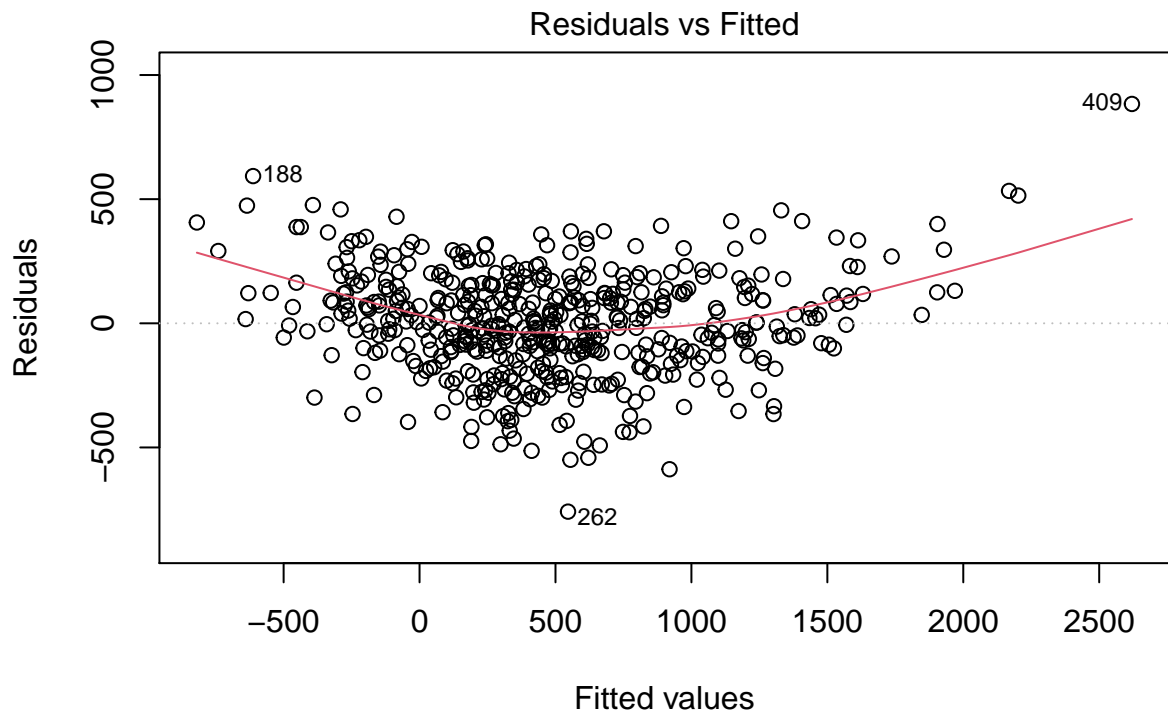
```
##
## Call:
## lm(formula = expenses ~ . - number_of_kids_1 - married_TRUE -
##      gender_woman - pet_none - number_of_kids_3, data = df_ohe)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -758.47 -119.27    2.59  129.27  883.45
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2289.6498    97.5110  -23.481  < 2e-16 ***
## age              57.5719     1.0739   53.612  < 2e-16 ***
## weight          1.3274      0.9954    1.334  0.18298
```

```
## height          2.0749      0.6561      3.163 0.00166 **
## gender_man      17.7275     20.1616      0.879 0.37969
## gender_other    65.6605     37.5247      1.750 0.08079 .
## married_FALSE   14.1053     23.5129      0.600 0.54886
## number_of_kids_0 -16.9776     25.0072     -0.679 0.49752
## number_of_kids_2 -75.0262     25.8665     -2.901 0.00390 **
## number_of_kids_4 -72.5561     41.0727     -1.767 0.07794 .
## number_of_kids_5 -52.3030     68.2307     -0.767 0.44372
## number_of_kids_6 -135.8840    90.1985     -1.506 0.13259
## pet_cat         -24.3437     26.1228     -0.932 0.35186
## pet_dog          10.5283     26.5391      0.397 0.69176
## pet_ferret       389.1340     33.1354    11.744 < 2e-16 ***
## pet_hedgehog     220.4780     33.0317      6.675 6.8e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 212.7 on 484 degrees of freedom
## Multiple R-squared:  0.8637, Adjusted R-squared:  0.8595
## F-statistic: 204.4 on 15 and 484 DF,  p-value: < 2.2e-16
```

Sprawdzanie założeń KMLR

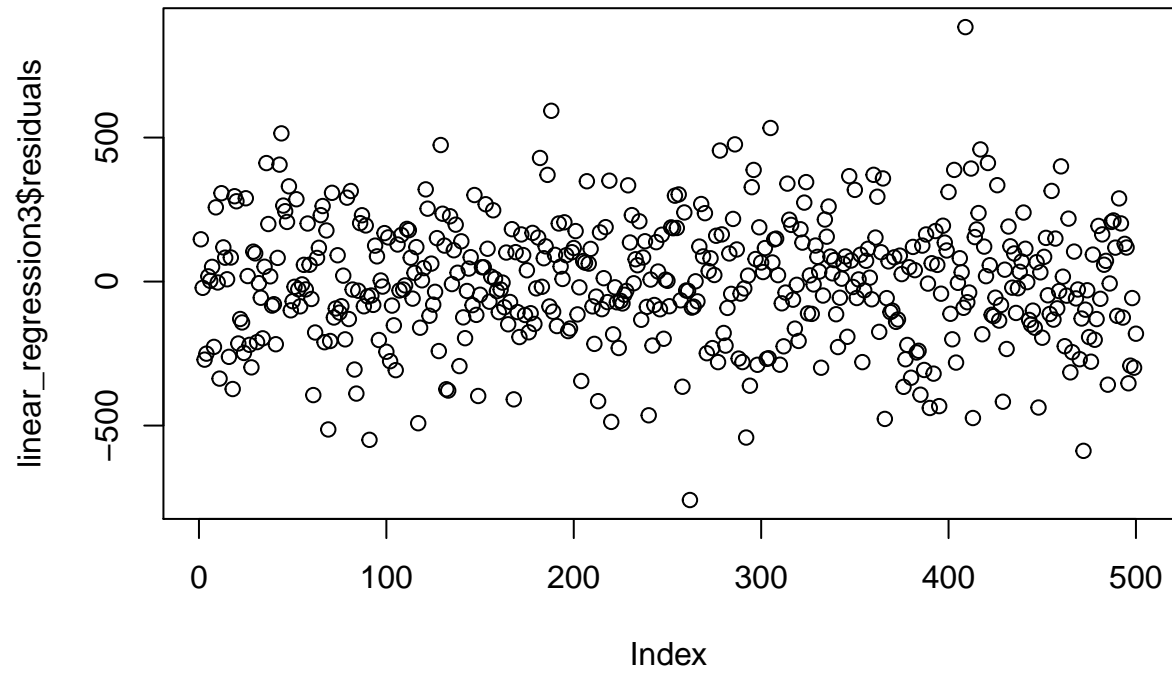
W modelu jest stała, założenie o wartości oczekiwanej błędu losowego nie musi być sprawdzane.

```
plot(linear_regression3, which = 1)
```

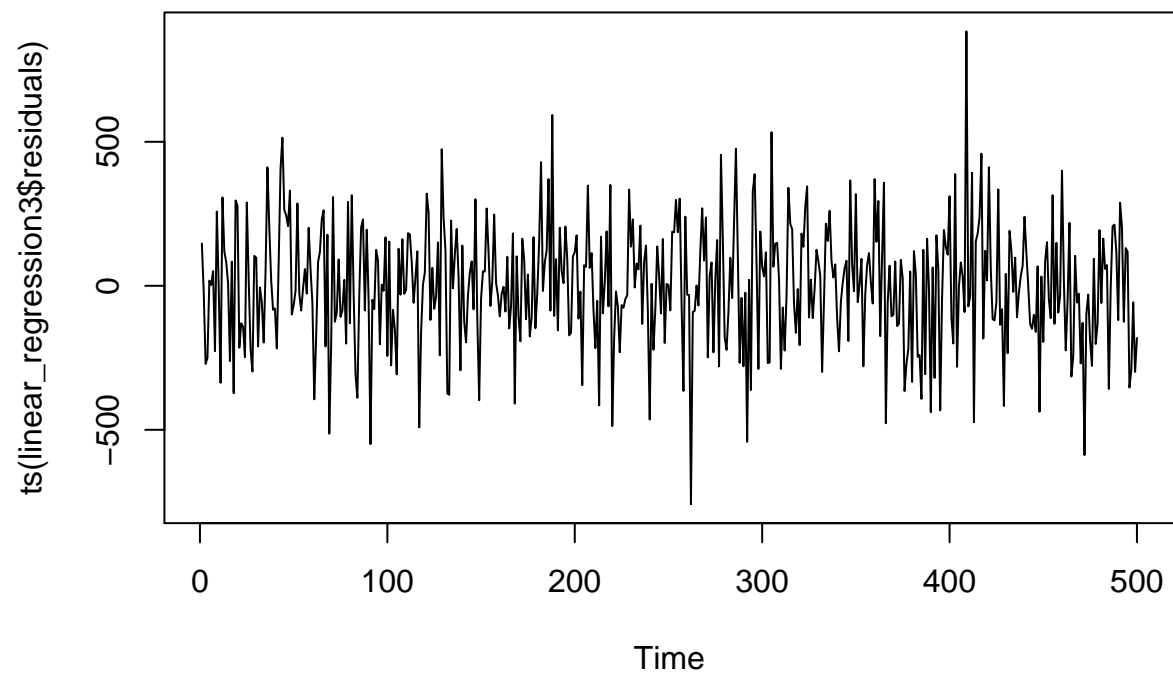


lm(expenses ~ . - number_of_kids_1 - married_TRUE - gender_woman - pet_none

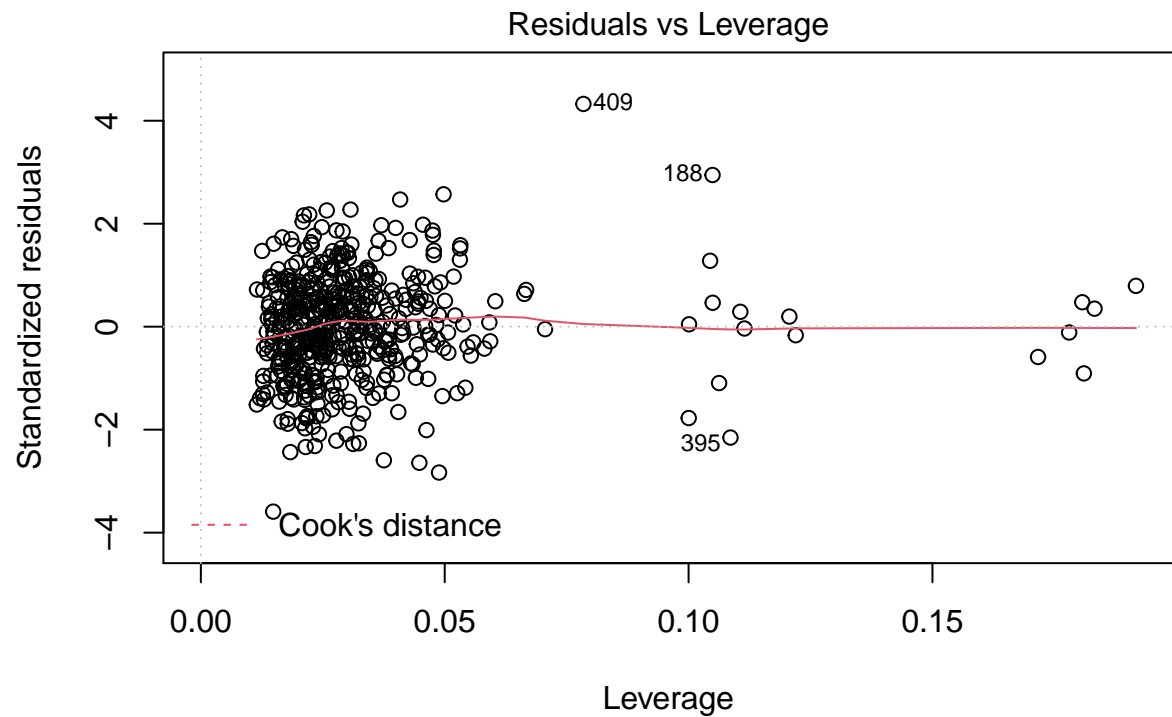
```
plot(linear_regression3$residuals)
```



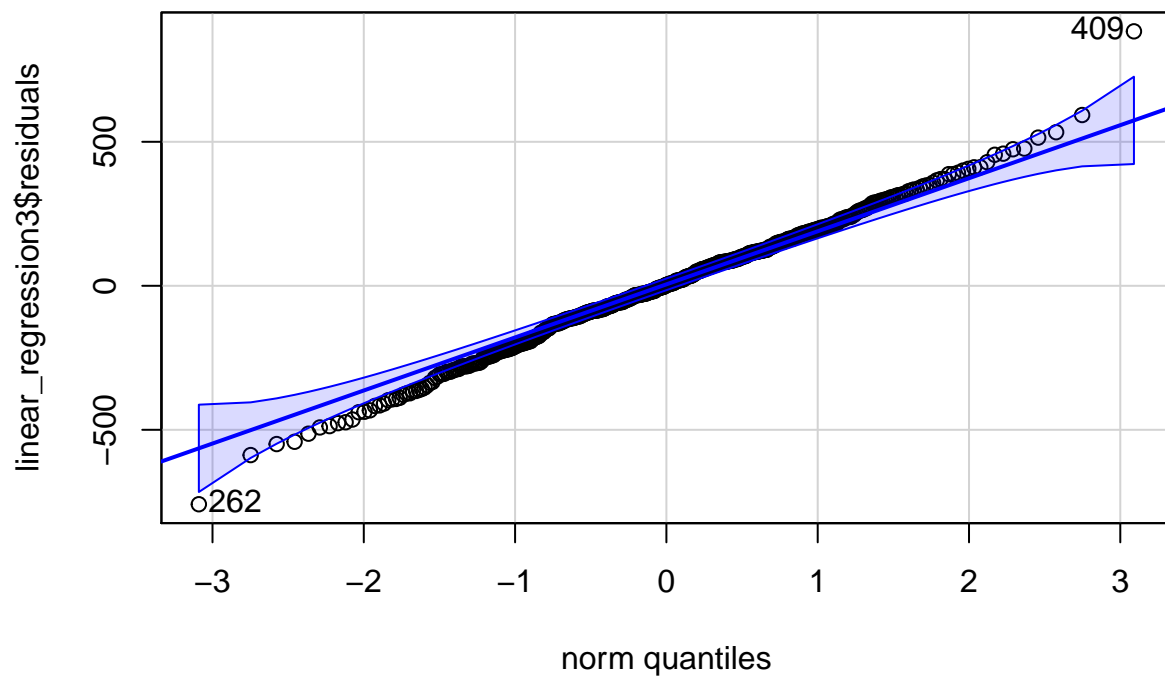
```
plot(ts(linear_regression3$residuals))
```



```
plot(linear_regression3, which=5)
```



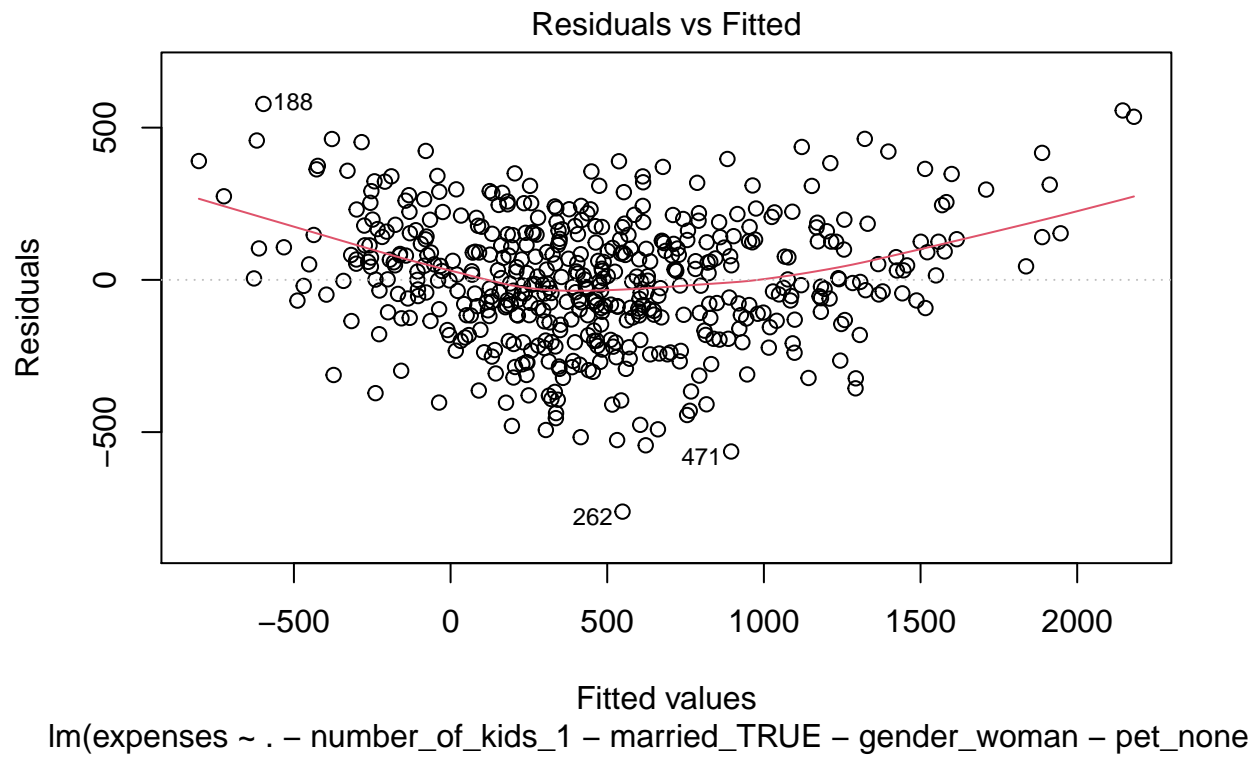
```
qqPlot(linear_regression3$residuals)
```

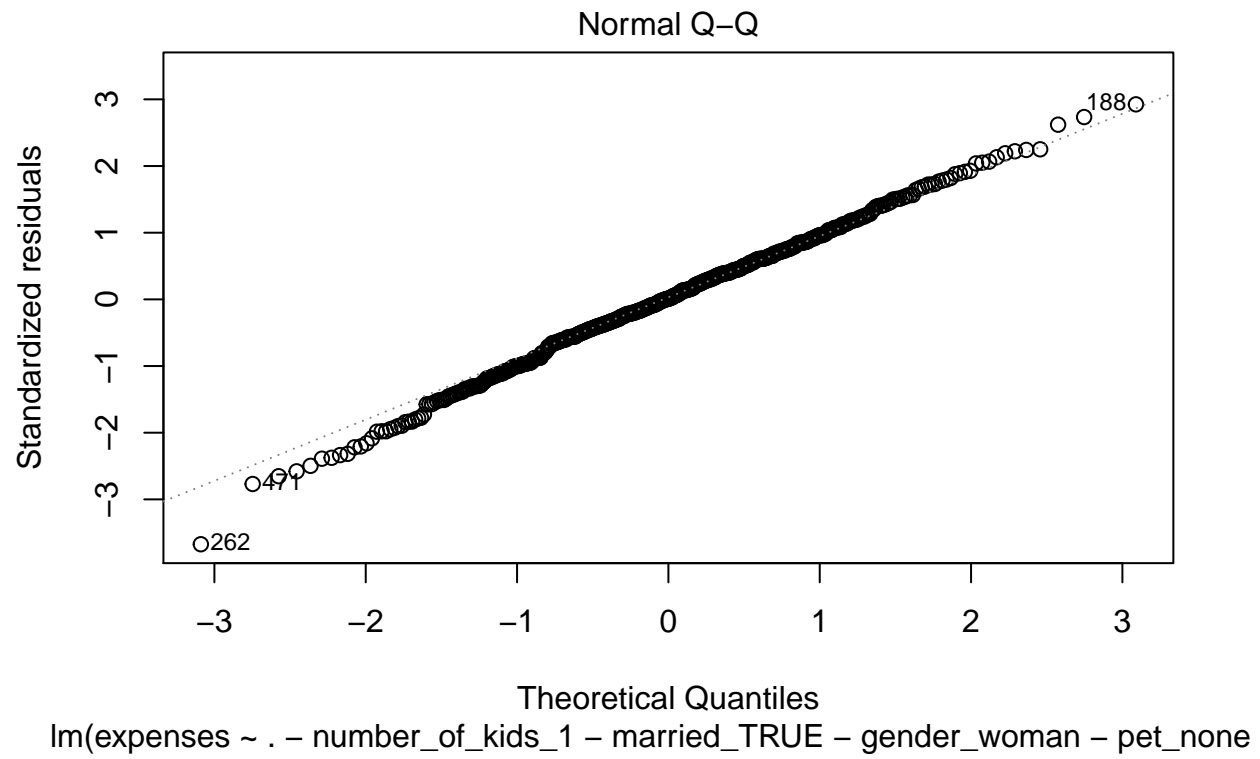


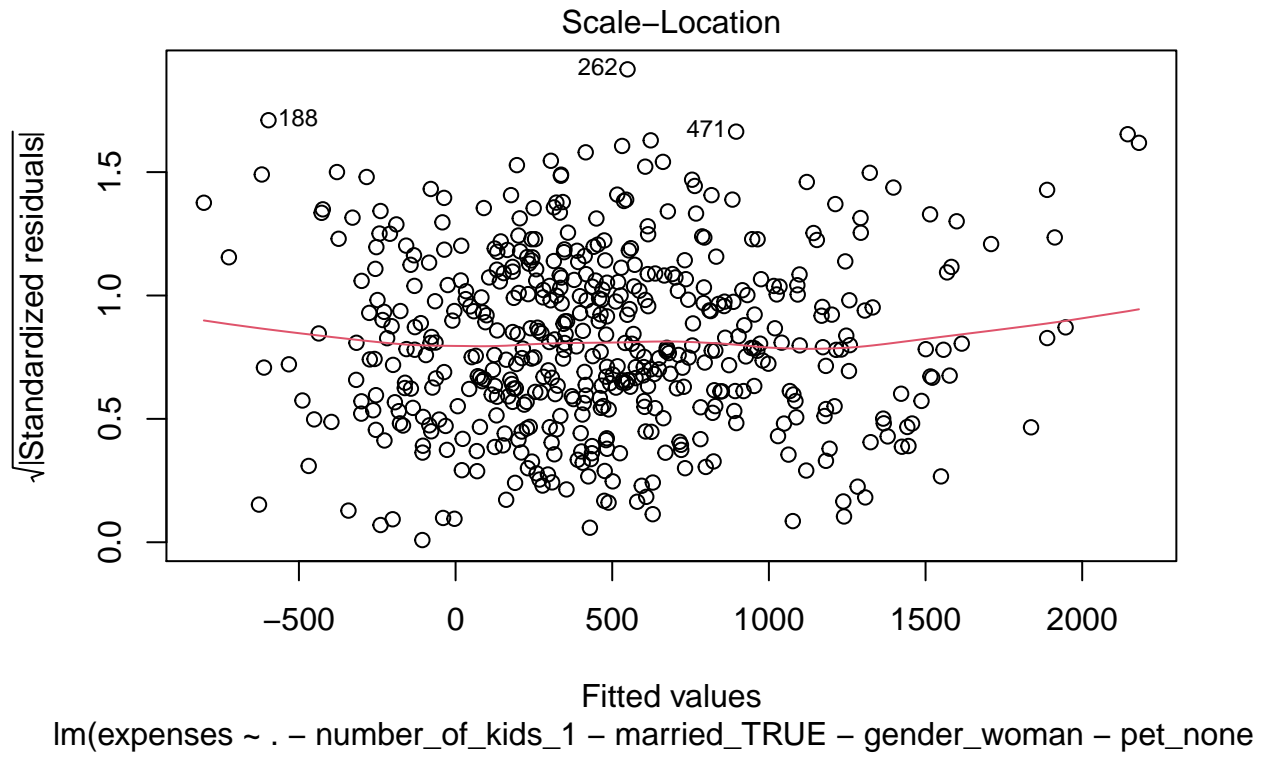
```
## [1] 409 262
```

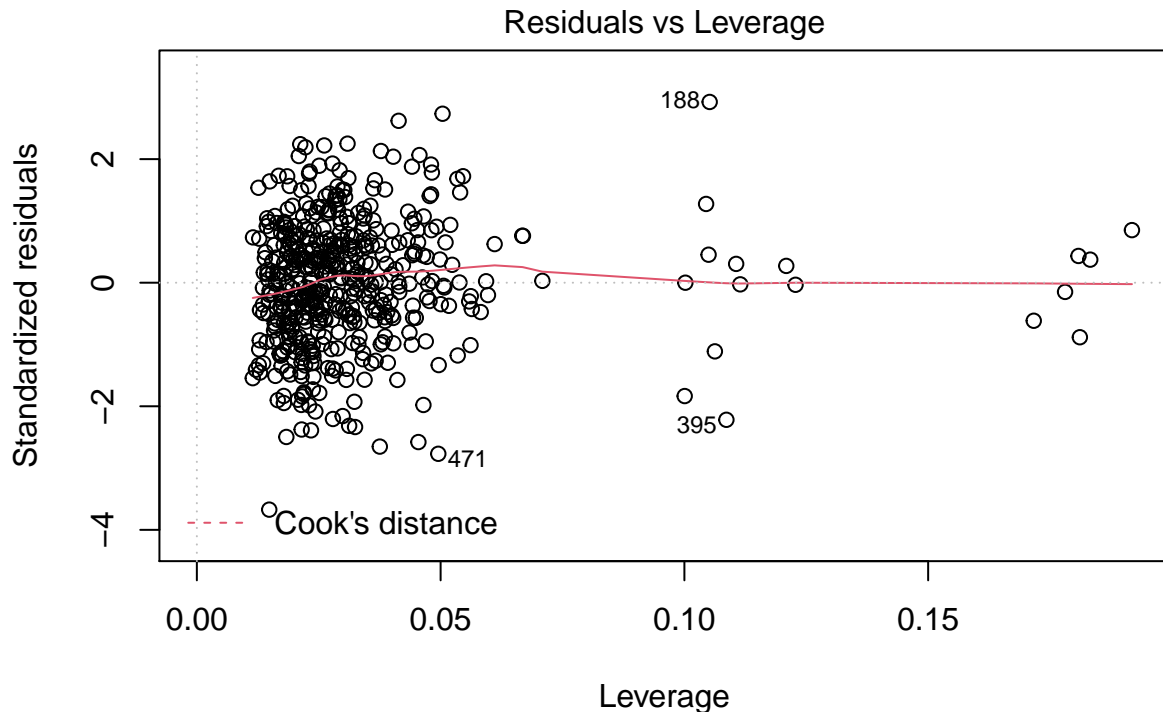
Występuje heteroskedastyczność, brak autokorelacji. 406 obserwacja odstająca. Rozkład wygląda na normalny chociaż jeden ogon zachowuje się dziwnie.

```
l <- lm(expenses ~.-number_of_kids_1-married_TRUE-gender_woman-pet_none-number_of_kids_3, df_ohc[-409,])
plot(l)
```







`lm(expenses ~ . - number_of_kids_1 - married_TRUE - gender_woman - pet_none`

Testy statystyczne

Normalność rozkładu reszt:

```
shapiro.test(linear_regression3$residuals)
```

```
##
## Shapiro-Wilk normality test
##
## data: linear_regression3$residuals
## W = 0.99554, p-value = 0.1642
```

Pvalue większe niż 0.1 poziom istotności. Rozkład normalny reszt. Stałość wariancji (studentized Breusch-Pagan test):

```
bptest(linear_regression3)
```

```
##
## studentized Breusch-Pagan test
##
## data: linear_regression3
## BP = 18.568, df = 15, p-value = 0.234
```

Pvalue 0.234 nie daje podstawy do odrzucenia hipotezy zerowej o braku autokorelacji na poziomie istotności 0.1.