# R Notebook

ZADATAK 4. "Kakav je odnos izmedu BMI-a i krvnog tlaka te možemo li predvidjeti krvni tlak na temelju BMI-a, dobi i učestalosti tjelesne aktivnosti?"

Za početak učitajmo i okvirno pogledajmo podatke, gdje smo na krvni tlak primijenili šum koji smo opisali i obrazložili u prvom zadatku.

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.4.1

## Warning: package 'ggplot2' was built under R version 4.4.1

## Warning: package 'tibble' was built under R version 4.4.1

## Warning: package 'tidyr' was built under R version 4.4.1

## Warning: package 'readr' was built under R version 4.4.1

## Warning: package 'purrr' was built under R version 4.4.1

## Warning: package 'dplyr' was built under R version 4.4.1

## Warning: package 'stringr' was built under R version 4.4.1

## Warning: package 'forcats' was built under R version 4.4.1

## Warning: package 'lubridate' was built under R version 4.4.2

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```r
# Učitavanje podataka iz csv datoteke:
healthDATA.modif = read.csv("Health Screening Data.csv")

set.seed(906)
healthDATA.modif <- healthDATA.modif %>%
  mutate(
    ap_hi = ap_hi + runif(n(), min = -5, max = 5),
    ap_lo = ap_lo + runif(n(), min = -5, max = 5)
    )

#View(healthDATA.modif)
summary(healthDATA.modif)
```

```
##        X               id             age            gender         height
##  Min.   :    0   Min.   :    0   Min.   :10798   Min.   :1.00   Min.   : 55.0
##  1st Qu.:17497   1st Qu.:25002   1st Qu.:17665   1st Qu.:1.00   1st Qu.:159.0
##  Median :34995   Median :49994   Median :19703   Median :1.00   Median :165.0
##  Mean   :34998   Mean   :49970   Mean   :19469   Mean   :1.35   Mean   :164.4
##  3rd Qu.:52500   3rd Qu.:74890   3rd Qu.:21327   3rd Qu.:2.00   3rd Qu.:170.0
##  Max.   :69999   Max.   :99999   Max.   :23713   Max.   :2.00   Max.   :250.0
##      weight           ap_hi            ap_lo           cholesterol
##  Min.   : 10.00   Min.   :-151.1   Min.   :  -74.56   Min.   :1.000
##  1st Qu.: 65.00   1st Qu.: 116.6   1st Qu.:   76.09   1st Qu.:1.000
##  Median : 72.00   Median : 122.8   Median :   81.02   Median :1.000
##  Mean   : 74.21   Mean   : 126.7   Mean   :   96.64   Mean   :1.367
##  3rd Qu.: 82.00   3rd Qu.: 137.0   3rd Qu.:   87.19   3rd Qu.:2.000
##  Max.   :200.00   Max.   : 244.3   Max.   :10997.51   Max.   :3.000
##       gluc            smoke             alco             active
##  Min.   :1.000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:1.000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:1.0000
##  Median :1.000   Median :0.00000   Median :0.00000   Median :1.0000
##  Mean   :1.226   Mean   :0.08818   Mean   :0.05377   Mean   :0.8037
##  3rd Qu.:1.000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:1.0000
##  Max.   :3.000   Max.   :1.00000   Max.   :1.00000   Max.   :1.0000
##      cardio          AgeinYr           BMI            BMICat
##  Min.   :0.0000   Min.   :29.00   Min.   :  3.50   Length:69960
##  1st Qu.:0.0000   1st Qu.:48.00   1st Qu.: 23.90   Class :character
##  Median :0.0000   Median :53.00   Median : 26.40   Mode  :character
##  Mean   :0.4996   Mean   :52.84   Mean   : 27.56
##  3rd Qu.:1.0000   3rd Qu.:58.00   3rd Qu.: 30.20
##  Max.   :1.0000   Max.   :64.00   Max.   :298.70
##    AgeGroup
##  Length:69960
##  Class :character
##  Mode  :character
##
##
##
```

Prije nego što krenemo s analizom preko linearne regresije, želimo izbaciti outliere za relevantne veličine. Obratimo pažnju na sistolički krvni tlak "ap_hi", imamo:

$$IQR = 137 - 116.6 = 20.4$$

Pa dobivamo za donju i gornju granicu:

$$Q_1 - 1.5 \cdot IQR = 86$$

$$Q_3 + 1.5 \cdot IQR = 167.6$$

Slično za dijastolički "ap_lo" imamo:

$$IQR = 87.19 - 76.09 = 11.1$$

$$Q_1 - 1.5 \cdot IQR = 59.44$$

$$Q_3 + 1.5 \cdot IQR = 103.84$$

Isti postupak radimo za BMI:

$$IQR = 6.3$$

$$Q_1 - 1.5 \cdot IQR = 14.45$$

$$Q_3 + 1.5 \cdot IQR = 39.65$$

Dob ispitanika nismo odlučili filtrirati jer minimalna i maksimalna dob (29.5835 i 64.9671 godina) zdravorazumski ne predstavljaju "outliere". Prema ovim rezultatima ćemo filtrirati podatke.

```
library(tidyverse)

filtered_data <- subset(healthDATA.modif,
                        ap_hi > 86 & ap_hi < 167.6
                        & BMI > 14.45 & BMI < 39.65
                        & ap_lo > 59.44 & ap_lo < 103.84)
summary(filtered_data)
```

```
##        X               id             age            gender
##  Min.   :    0   Min.   :    0   Min.   :10798   Min.   :1.000
##  1st Qu.:17435   1st Qu.:24905   1st Qu.:17649   1st Qu.:1.000
##  Median :34989   Median :49983   Median :19696   Median :1.000
##  Mean   :34988   Mean   :49956   Mean   :19455   Mean   :1.355
##  3rd Qu.:52488   3rd Qu.:74871   3rd Qu.:21316   3rd Qu.:2.000
##  Max.   :69999   Max.   :99999   Max.   :23713   Max.   :2.000
##      height          weight          ap_hi           ap_lo
##  Min.   :120.0   Min.   : 28.00   Min.   : 86.04   Min.   : 59.44
##  1st Qu.:159.0   1st Qu.: 65.00   1st Qu.:116.62   1st Qu.: 76.14
##  Median :165.0   Median : 71.00   Median :122.41   Median : 80.74
##  Mean   :164.6   Mean   : 73.03   Mean   :125.49   Mean   : 81.03
##  3rd Qu.:170.0   3rd Qu.: 80.00   3rd Qu.:134.80   3rd Qu.: 85.69
##  Max.   :207.0   Max.   :135.00   Max.   :167.60   Max.   :103.84
##   cholesterol         gluc           smoke            alco
##  Min.   :1.000   Min.   :1.000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:0.00000
##  Median :1.000   Median :1.000   Median :0.0000   Median :0.00000
##  Mean   :1.351   Mean   :1.219   Mean   :0.0881   Mean   :0.05261
##  3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:0.0000   3rd Qu.:0.00000
##  Max.   :3.000   Max.   :3.000   Max.   :1.0000   Max.   :1.00000
##      active          cardio          AgeinYr          BMI
##  Min.   :0.0000   Min.   :0.0000   Min.   :29.0   Min.   :14.50
##  1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:48.0   1st Qu.:23.80
##  Median :1.0000   Median :0.0000   Median :53.0   Median :26.10
```

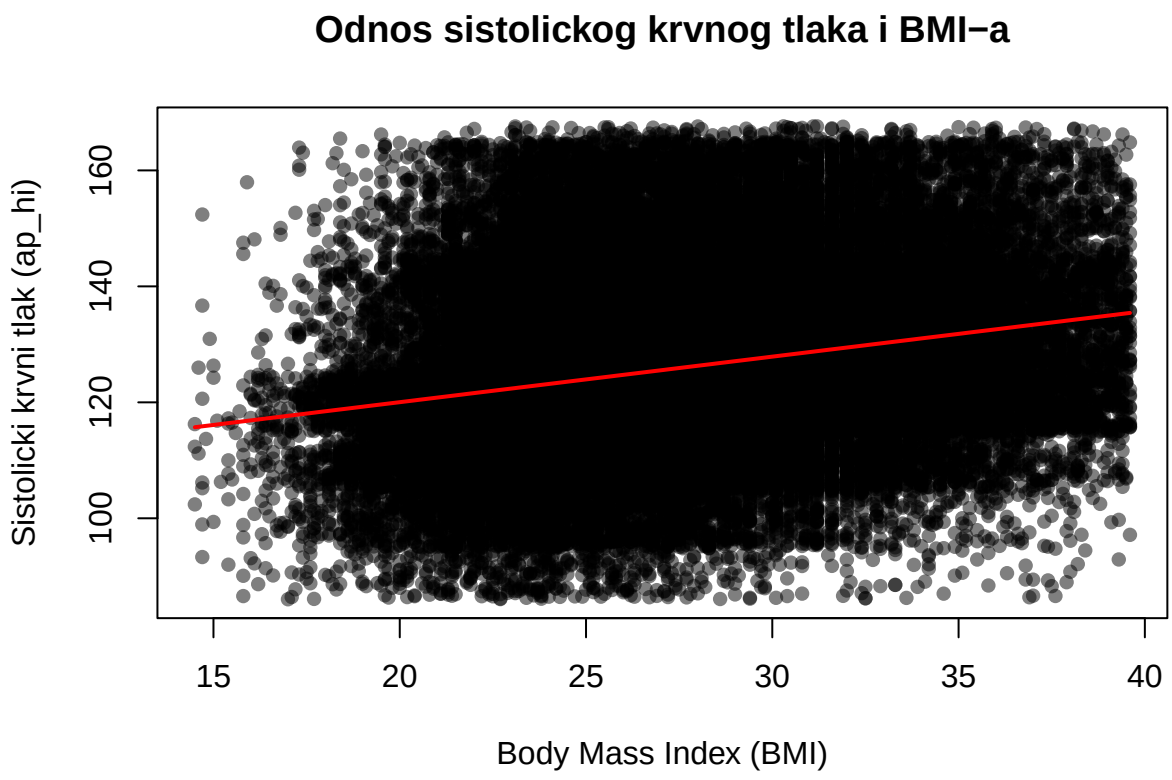
```
##  Mean   :0.8042   Mean   :0.4829   Mean   :52.8   Mean   :26.96
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:58.0   3rd Qu.:29.70
##  Max.   :1.0000   Max.   :1.0000   Max.   :64.0   Max.   :39.60
##     BMICat            AgeGroup
##  Length:63315       Length:63315
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

Sada ćemo metodom najmanjih kvadrata pokušati uspostaviti vezu između BMI-a i krvnog tlaka.

```
fit.ap_hi <- lm(ap_hi ~ poly(BMI,1) , data = filtered_data)

plot(filtered_data$BMI, filtered_data$ap_hi,
     main = "Odnos sistoličkog krvnog tlaka i BMI-a",
     xlab = "Body Mass Index (BMI)",
     ylab = "Sistolički krvni tlak (ap_hi)",
     pch = 16, col = rgb(0, 0, 0, 0.5))


sorted_index <- order(filtered_data$BMI)
lines(filtered_data$BMI[sorted_index],
      fit.ap_hi$fitted.values[sorted_index],
      col = "red", lwd = 2)
```



```r
summary(fit.ap_hi)
```

```
##
## Call:
## lm(formula = ap_hi ~ poly(BMI, 1), data = filtered_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.244  -8.602  -2.391   8.199  46.721
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  125.49275    0.05548 2261.99   <2e-16 ***
## poly(BMI, 1) 867.90488   13.95989   62.17   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.96 on 63313 degrees of freedom
## Multiple R-squared:  0.05754,    Adjusted R-squared:  0.05752
## F-statistic:  3865 on 1 and 63313 DF,  p-value: < 2.2e-16
```

Iznad možemo vidjeti graf raspršenja između sistoličkog tlaka i BMI-a kao i pravac linearne regresije koji smo izračunali iz podataka. Pokušavali smo linearnu regresiju s polinomima viših stupnjeva, ali su svi stupnjevi bili veoma slični pravcima i nisu poboljšavali vrijednost $R^2$. Zbog toga smo dali prednost najjednostavnijem modelu, a to je naravno pravac. Vidimo blagi pozitivan trend, ali se iz p vrijednosti vidi da je značajnost regresora skoro pa zanemariva. Također, $R^2$ vrijednost je 0.0564 ($R^2_{adj}$ je 0.05638) što ukazuje na loš fit modela, no mi ćemo svakako sada nastaviti s analizom reziduala.

```r
standardized_residuals <- rstandard(fit.ap_hi)
ks_test <- ks.test(standardized_residuals, "pnorm")
print(ks_test)
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  standardized_residuals
## D = 0.092807, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```r
require(nortest)
```

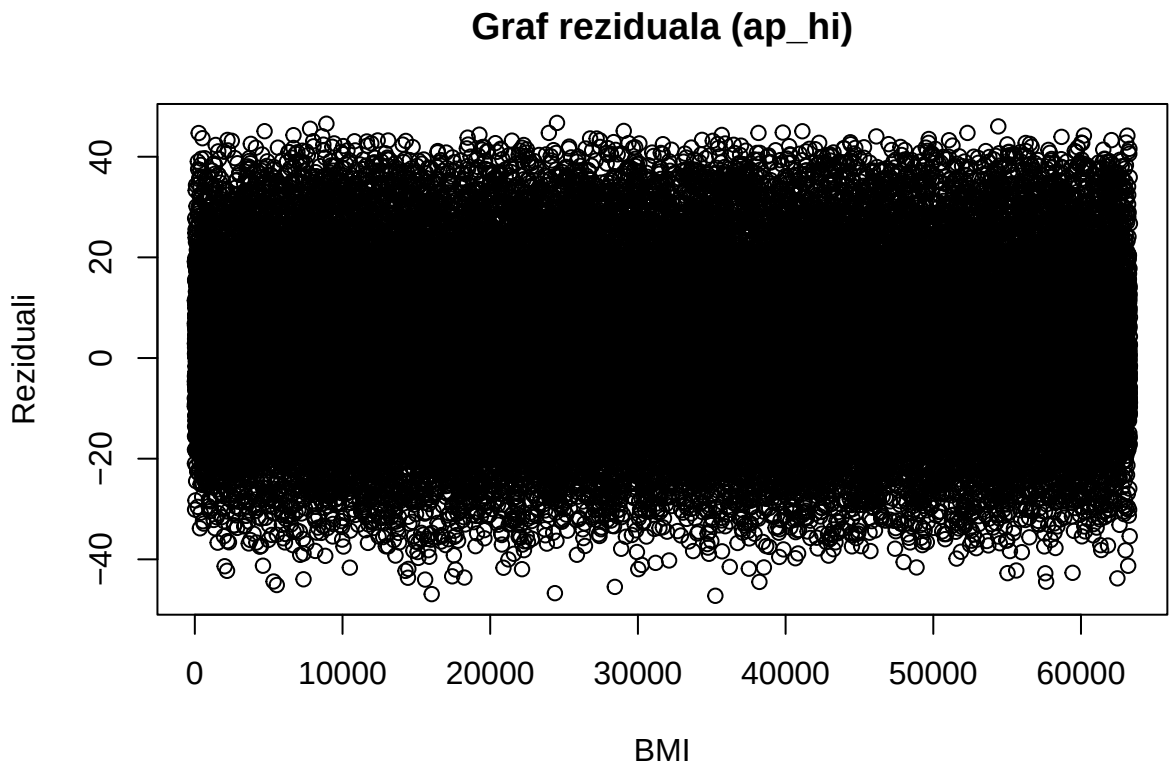
```
## Loading required package: nortest
```

```r
lilliefors_test <- lillie.test(standardized_residuals)
print(lilliefors_test)
```

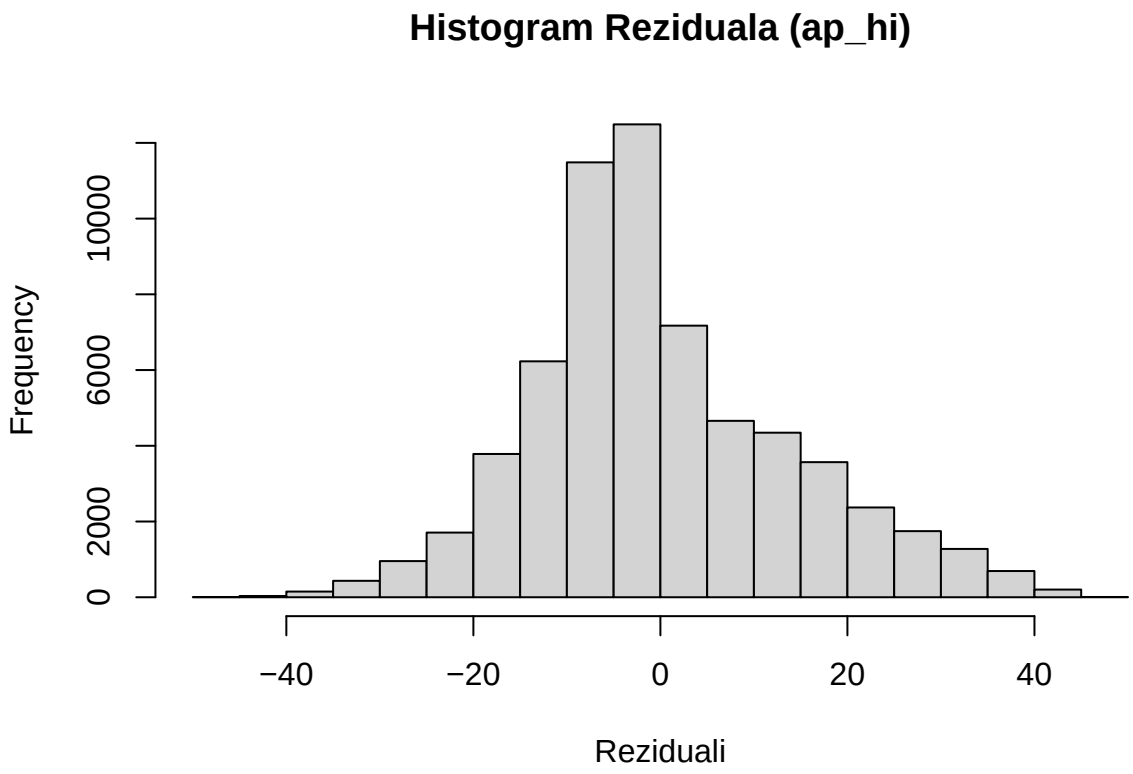
```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  standardized_residuals
## D = 0.092807, p-value < 2.2e-16
```

Reziduali padaju na testu normalnosti (i KS i Lilliefors). Možemo također vidjeti kako se reziduali ponašaju grafički:

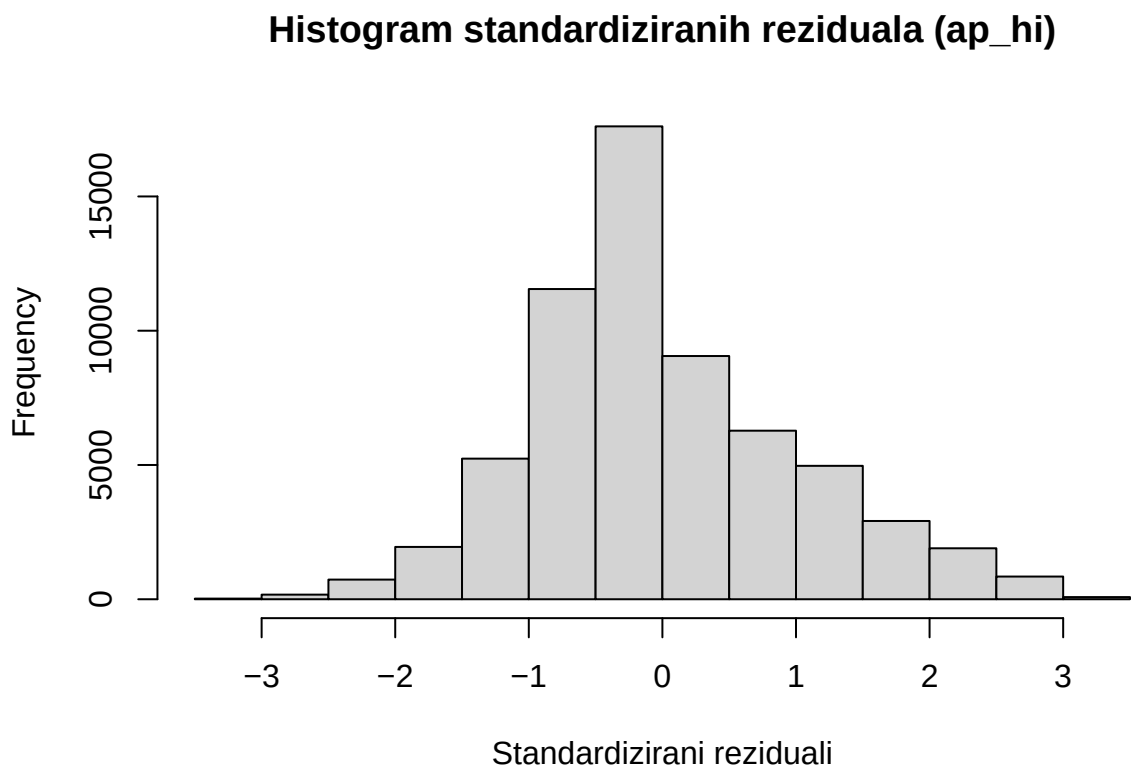
```
plot(fit.ap_hi$residuals,
     main = "Graf reziduala (ap_hi)",
     ylab = "Reziduali", xlab = "BMI")
```



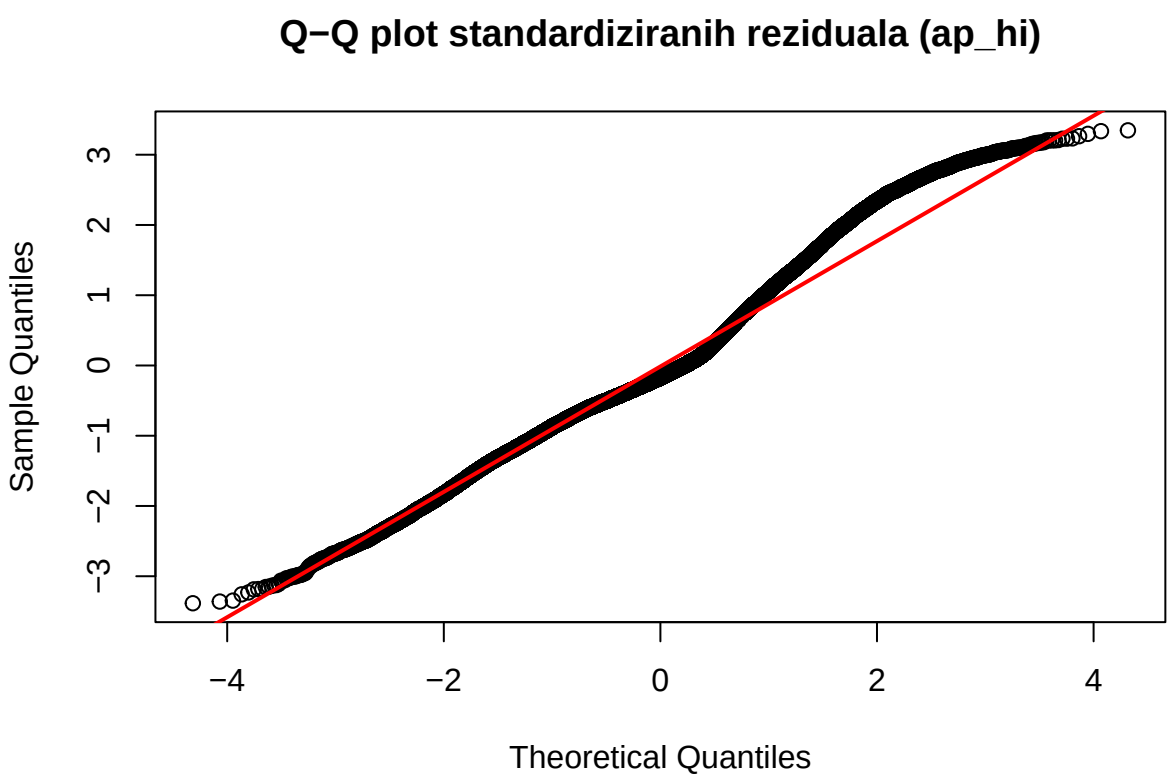

```
hist(fit.ap_hi$residuals,
     breaks = 20,
     main = "Histogram Reziduala (ap_hi)",
     xlab = "Reziduali")
```



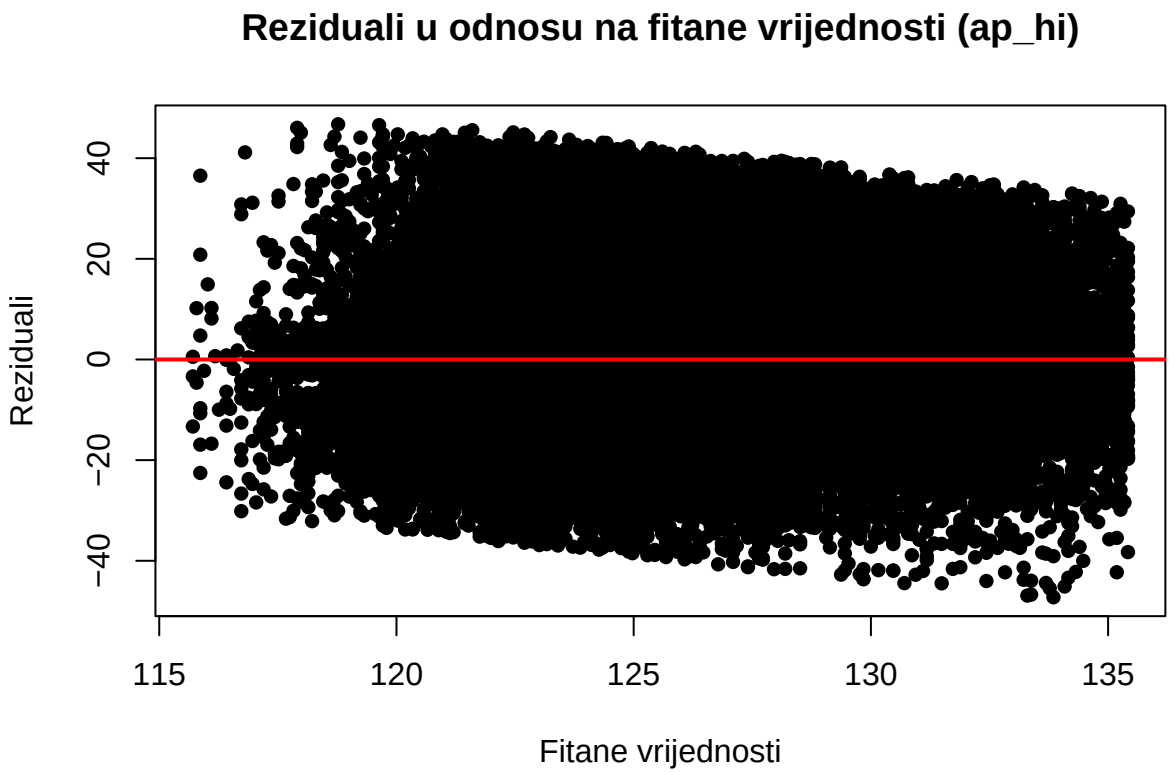

```
hist(rstandard(fit.ap_hi),
     breaks = 20,
     main = "Histogram standardiziranih reziduala (ap_hi)",
     xlab = "Standardizirani reziduali")
```




```
qqnorm(rstandard(fit.ap_hi),
       main = "Q-Q plot standardiziranih reziduala (ap_hi)")
qqline(rstandard(fit.ap_hi), col = "red", lwd = 2)
```




```
plot(fit.ap_hi$fitted.values, fit.ap_hi$residuals,
     main = "Reziduali u odnosu na fitane vrijednosti (ap_hi)",
     xlab = "Fitane vrijednosti", ylab = "Reziduali", pch = 16)
abline(h = 0, col = "red", lwd = 2)
```

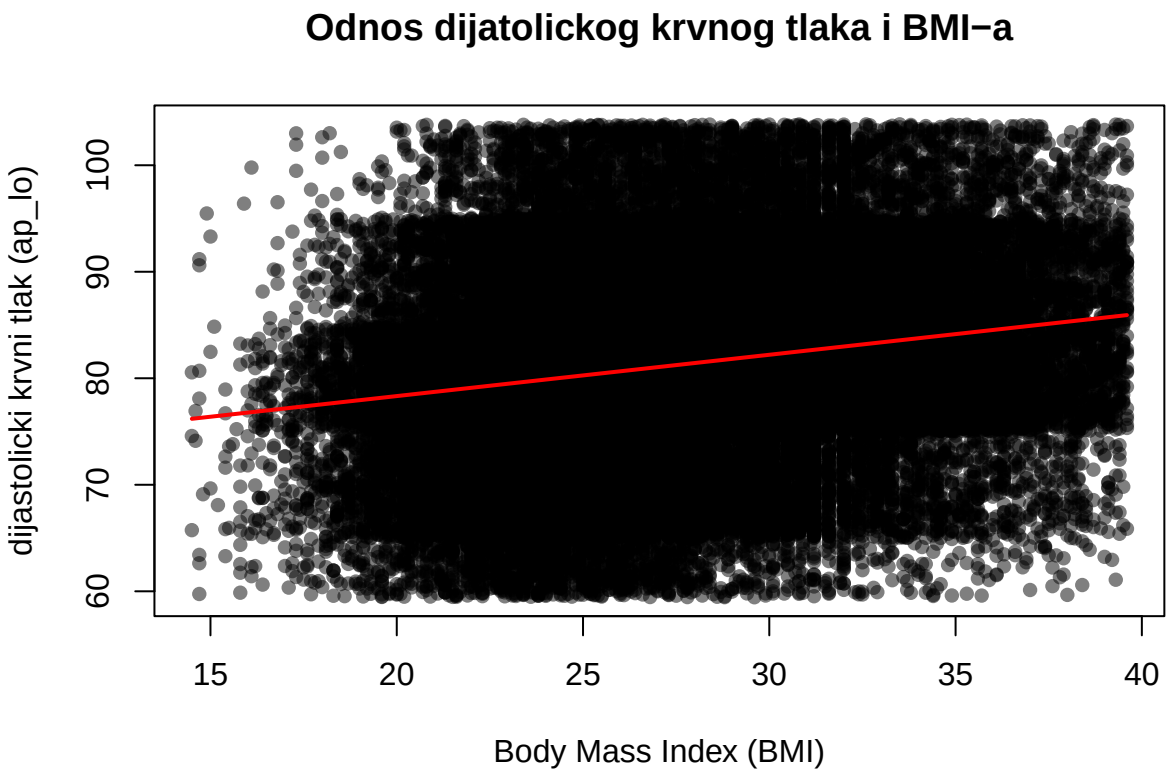



Q-Q plot nam govori da ova razdioba ima lakše repove od normalne, ali ovo svakako nije normalna distribucija. Sada možemo zaključiti da je nemoguće predvidjeti sistolički krvni tlak iz BMI-a (iz ovih podataka).

Za dijastolički krvni tlak ponavljamo isti postupak.

```
fit.ap_lo <- lm(ap_lo ~ poly(BMI, 1) , data = filtered_data)
plot(filtered_data$BMI, filtered_data$ap_lo,
     main = "Odnos dijatoličkog krvnog tlaka i BMI-a",
     xlab = "Body Mass Index (BMI)",
     ylab = "dijastolički krvni tlak (ap_lo)",
     pch = 16, col = rgb(0, 0, 0, 0.5))


sorted_index <- order(filtered_data$BMI)
lines(filtered_data$BMI[sorted_index],
      fit.ap_lo$fitted.values[sorted_index],
      col = "red", lwd = 2)
```




```
summary(fit.ap_lo)
```

```
##
## Call:
## lm(formula = ap_lo ~ poly(BMI, 1), data = filtered_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.6478  -4.8973  -0.1736   4.7408  25.7154
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   81.02753    0.03263  2483.1   <2e-16 ***
## poly(BMI, 1) 428.64235    8.21098    52.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.211 on 63313 degrees of freedom
## Multiple R-squared:  0.04127,    Adjusted R-squared:  0.04125
## F-statistic:  2725 on 1 and 63313 DF,  p-value: < 2.2e-16
```

Zadržat ćemo model pravca iz istog razloga kao i za sistolički tlak. Vidi se blagi pozitivan trend, ali vidimo (iz p vrijednosti) da regresor ima jako malenu značajnost. Također, $R^2$ vrijednost je sada 0.04008 ($R^2_{adj}$ je 0.04007) što opet ukazuje na loš fit modela, no mi ćemo svakako opet nastaviti s analizom reziduala. Analiza reziduala

```r
standardized_residuals <- rstandard(fit.ap_hi)

ks_test <- ks.test(standardized_residuals, "pnorm")
print(ks_test)
```

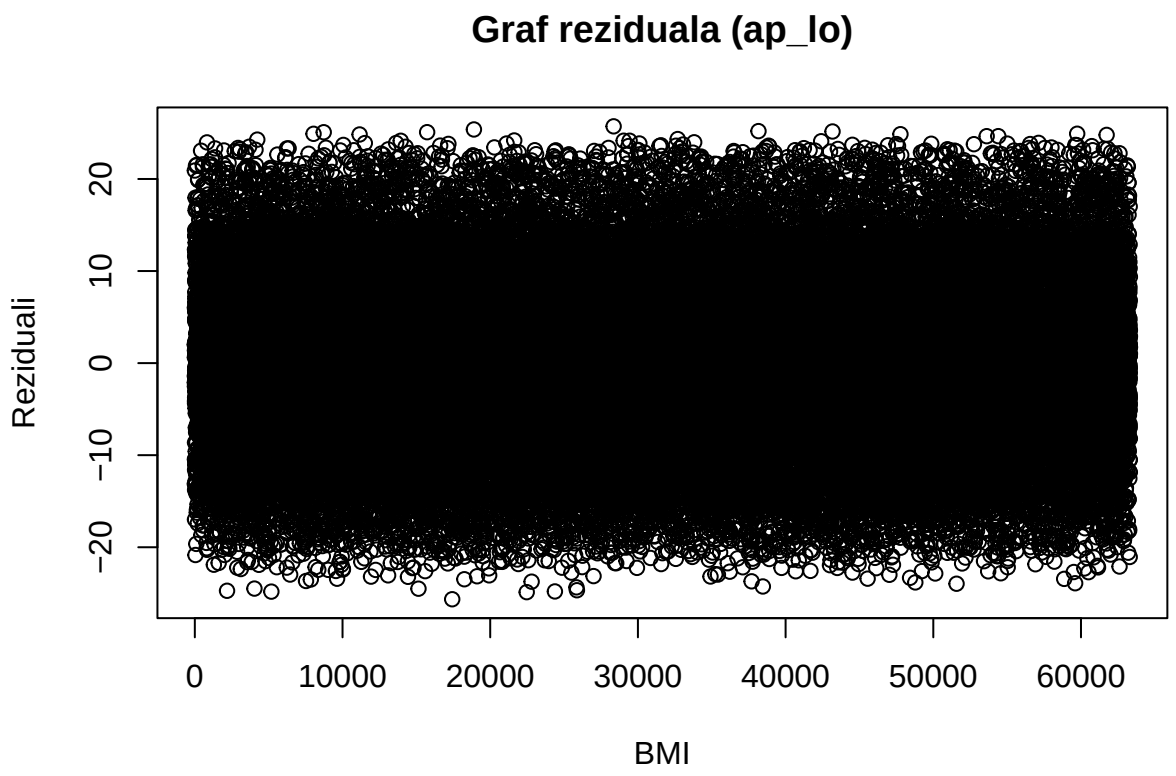
```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  standardized_residuals
## D = 0.092807, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```r
lilliefors_test <- lillie.test(standardized_residuals)
print(lilliefors_test)
```

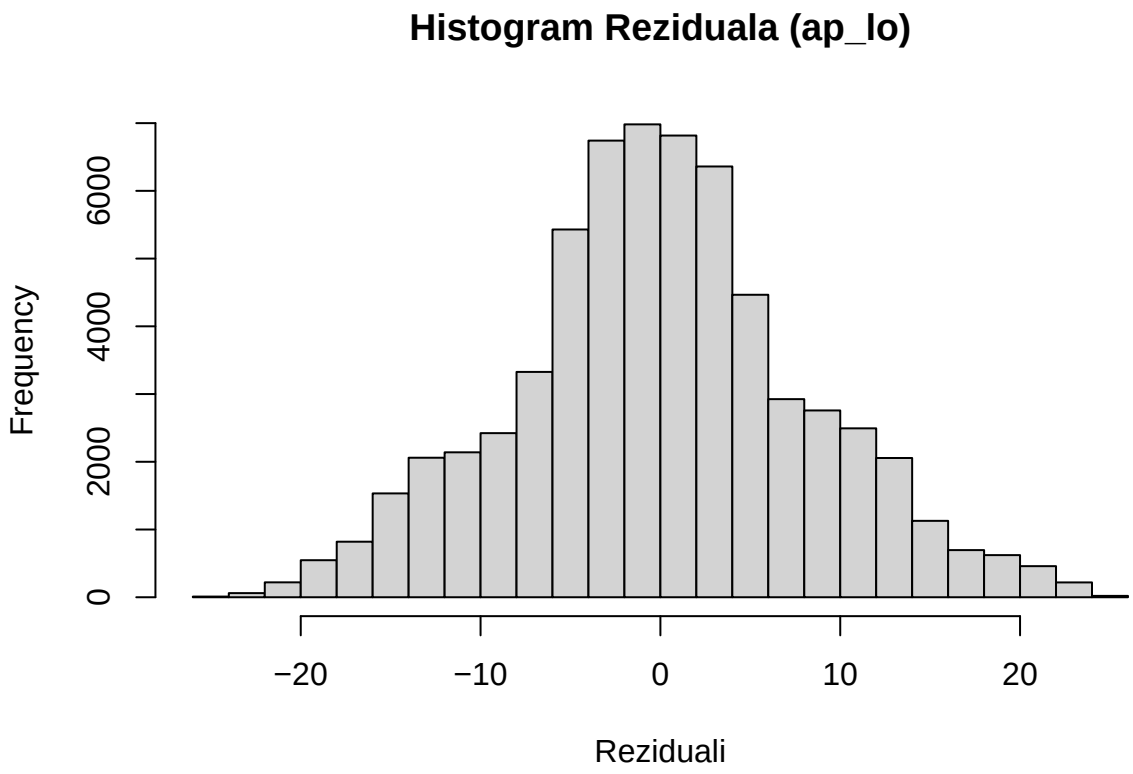
```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  standardized_residuals
## D = 0.092807, p-value < 2.2e-16
```

Reziduali padaju na testu normalnosti (i KS i Lilliefors). Možemo također vidjeti kako se reziduali ponašaju grafički:

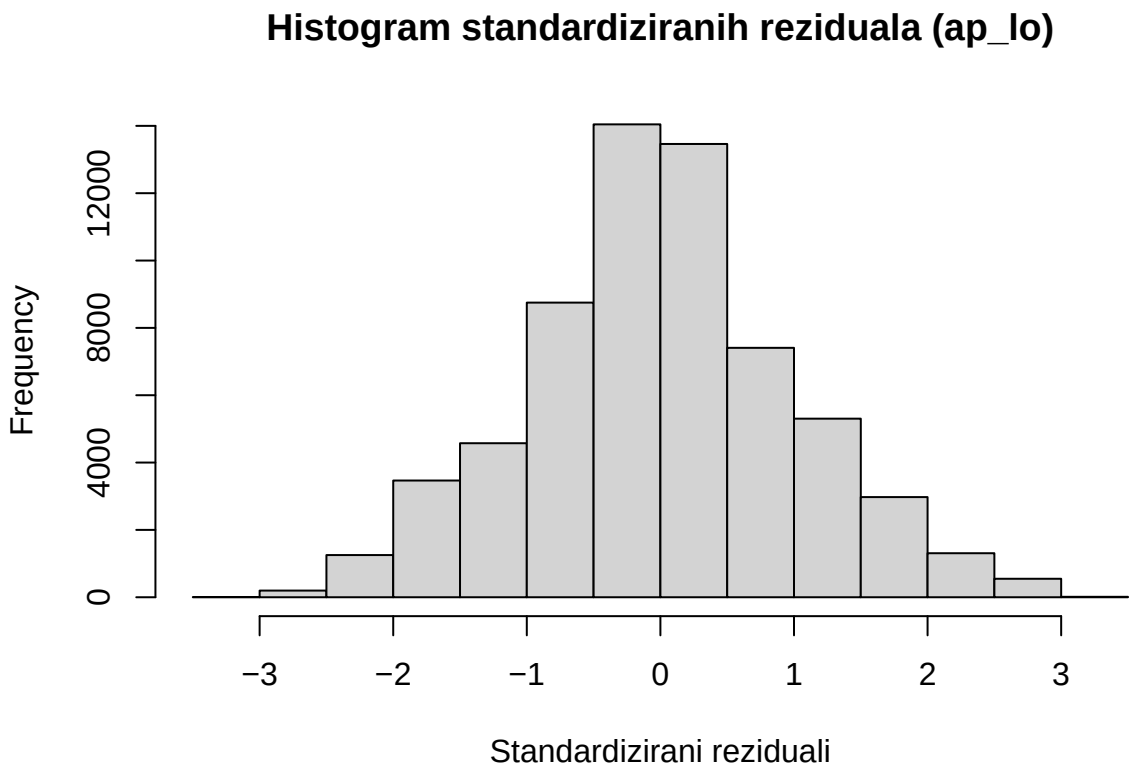
```r
plot(fit.ap_lo$residuals,
     main = "Graf reziduala (ap_lo)",
     ylab = "Reziduali", xlab = "BMI")
```



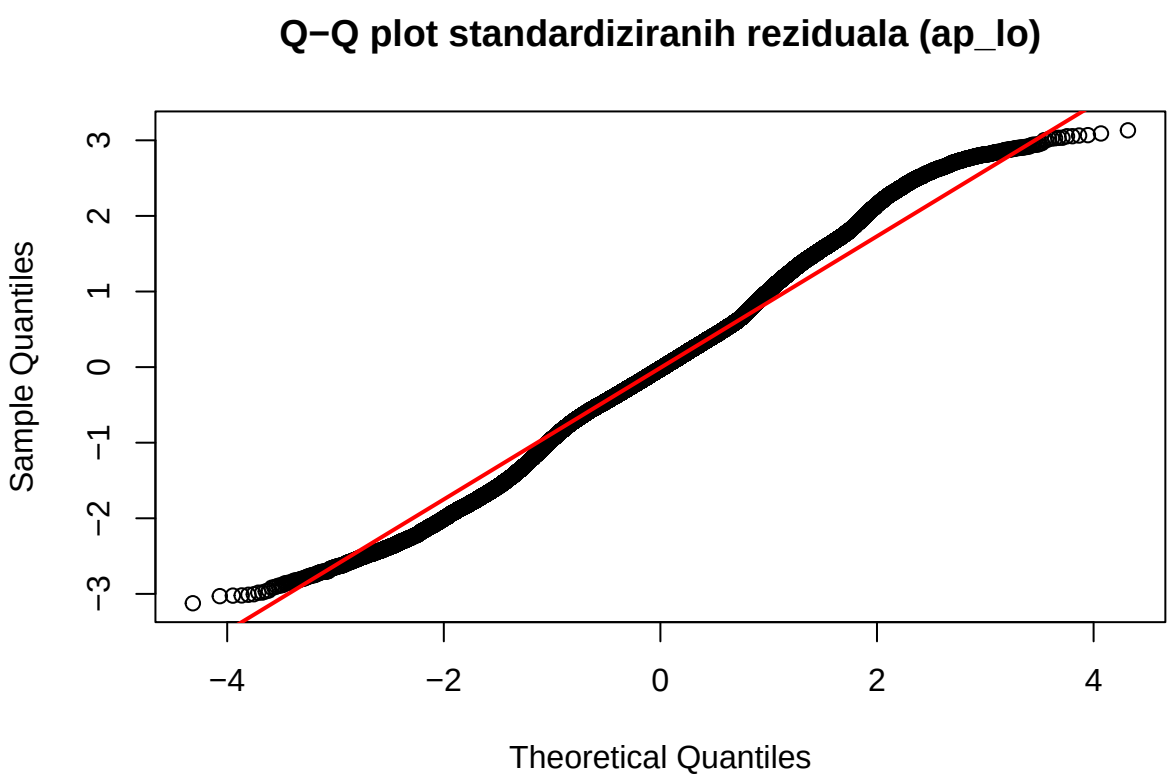

```r
hist(fit.ap_lo$residuals,
     breaks = 20,
     main = "Histogram Reziduala (ap_lo)",
     xlab = "Reziduali")
```



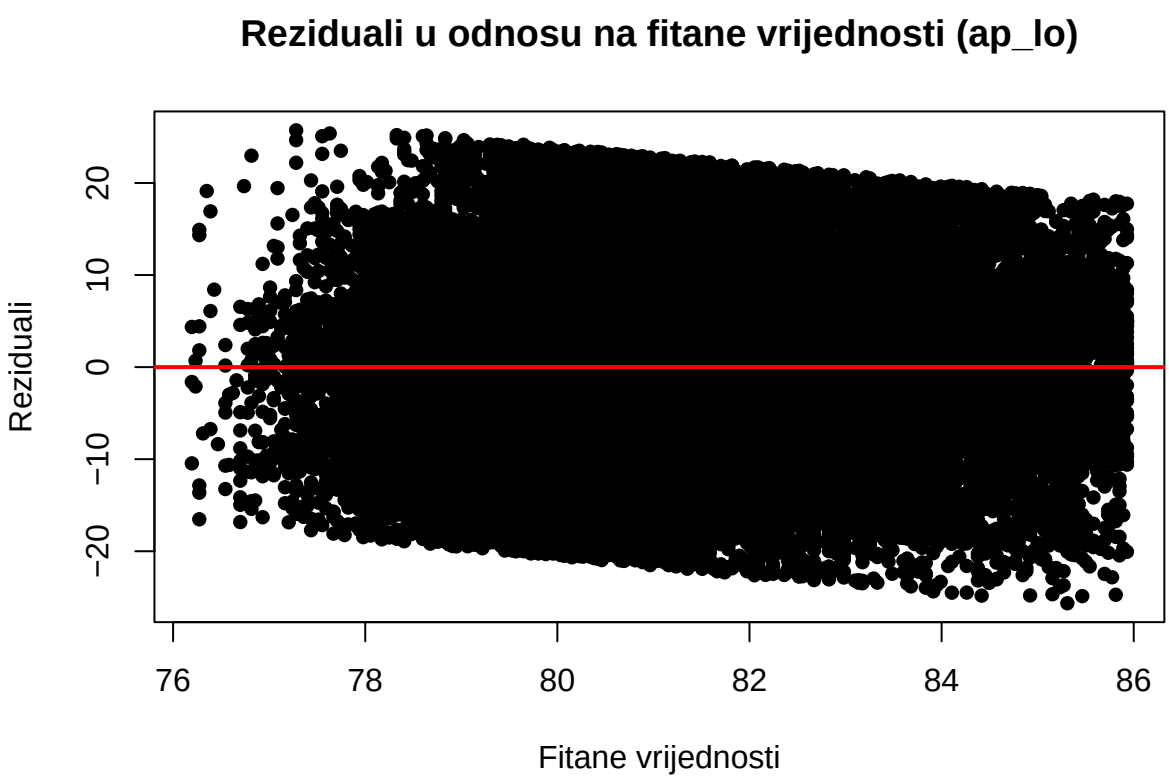

```
hist(rstandard(fit.ap_lo),
     breaks = 20,
     main = "Histogram standardiziranih reziduala (ap_lo)",
     xlab = "Standardizirani reziduali")
```




```
qqnorm(rstandard(fit.ap_lo),
       main = "Q-Q plot standardiziranih reziduala (ap_lo)")
qqline(rstandard(fit.ap_lo), col = "red", lwd = 2)
```




```
plot(fit.ap_lo$fitted.values, fit.ap_lo$residuals,
     main = "Reziduali u odnosu na fitane vrijednosti (ap_lo)",
     xlab = "Fitane vrijednosti", ylab = "Reziduali", pch = 16)
abline(h = 0, col = "red", lwd = 2)
```




Grafički možemo reći da reziduali imaju teže repove, ali se ne ponašaju baš pravino. Nemoguće je (na temelju ovih podataka) predvidjeti dijastolički krvni tlak iz BMI-a.

Obratimo sada pažnju na drugi dio problema: "Možemo li predvidjeti krvni tlak na temelju BMI-a, dobi i učestalosti tjelesne aktivnosti?"

Kada radimo na višestrukoj regresiji želimo da nam regresori budu međusobno "dovoljno" nezavisni, inače ne možemo interpretirati rezultate. Stoga računamo kovarijancu za sve parove od BMI, starosti i tjelesne aktivnosti. NAPOMENA: S obzirom da je tjelesna aktivnost binarna kategorijska varijabla nije loša ideja staviti ju u model višestruke regresije.

```
cor(cbind(filtered_data$active,filtered_data$BMI,filtered_data$AgeinYr))
```

```
##              [,1]         [,2]        [,3]
## [1,]  1.000000000 -0.008373687 -0.01032664
## [2,] -0.008373687  1.000000000  0.10301095
## [3,] -0.010326644  0.103010955  1.00000000
```

Iz kovarijanci možemo zaključiti da su varijable "dovoljno" nezavisne. Veću zavisnost vidimo između BMI i starosti, što ima smisla jer kako starimo naša visina se toliko ne mijenja koliko naša masa, pa je normalno da će BMI ovisiti o starosti, no svakako možemo pretpostaviti nezavisnost i zbog toga što je najstarija osoba u uzorku ima 64 godina, što nije dovoljno staro da krene značajno odumiranje mišićnog tkiva.

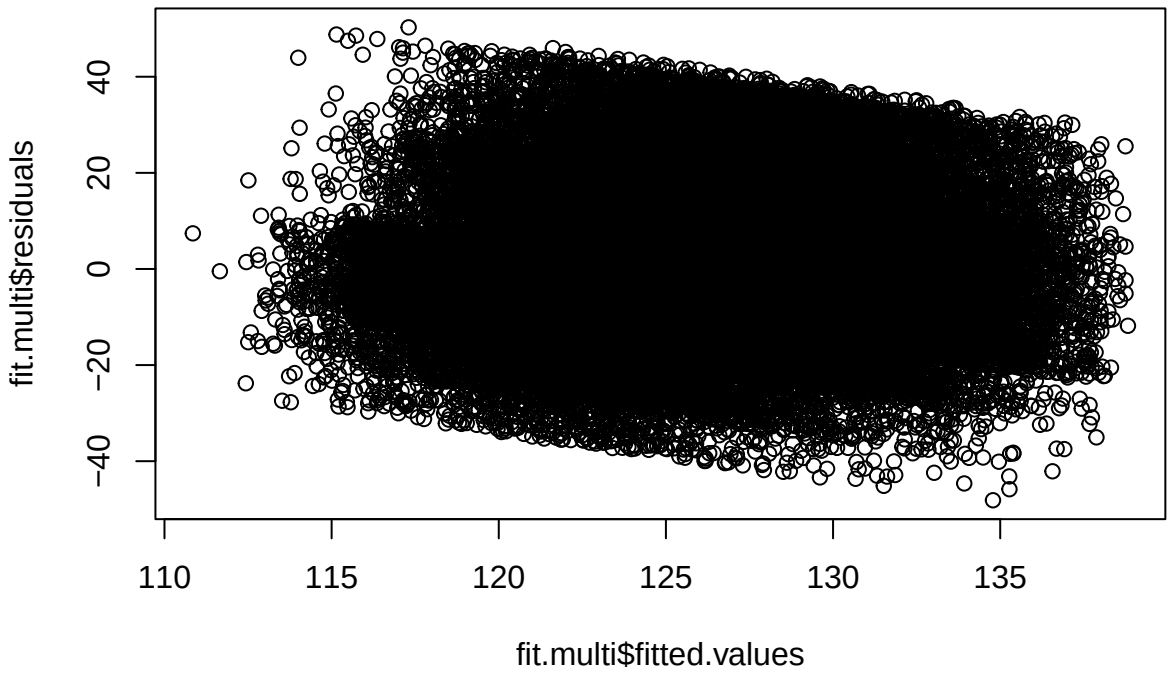
```
fit.multi <- lm(ap_hi ~ BMI + active + AgeinYr, filtered_data) #ako maknete regresore koji su manje zna
#fit.multi = lm(ap_hi ~ AgeinYr + active, filtered_data)
summary(fit.multi)
```

```
##
## Call:
## lm(formula = ap_hi ~ BMI + active + AgeinYr, data = filtered_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.131  -8.727  -2.233   7.765  50.278
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 85.625551   0.532206 160.888   <2e-16 ***
## BMI          0.724585   0.012480  58.058   <2e-16 ***
## active       0.129298   0.137419   0.941    0.347
## AgeinYr      0.383084   0.008102  47.282   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.72 on 63311 degrees of freedom
## Multiple R-squared:  0.08969,    Adjusted R-squared:  0.08964
## F-statistic:  2079 on 3 and 63311 DF,  p-value: < 2.2e-16
```

Vidimo da je jedini značajan regresor mjera tjelesne aktivnosti, no s trenutnim odabirom regresora dobivamo najbolju $R^2$ vrijednost tako da smo ih odlučili zadržati.

Nastavimo s analizom reziduala. Prvo testiramo normalnost:

```
plot(fit.multi$fitted.values,fit.multi$residuals)
```



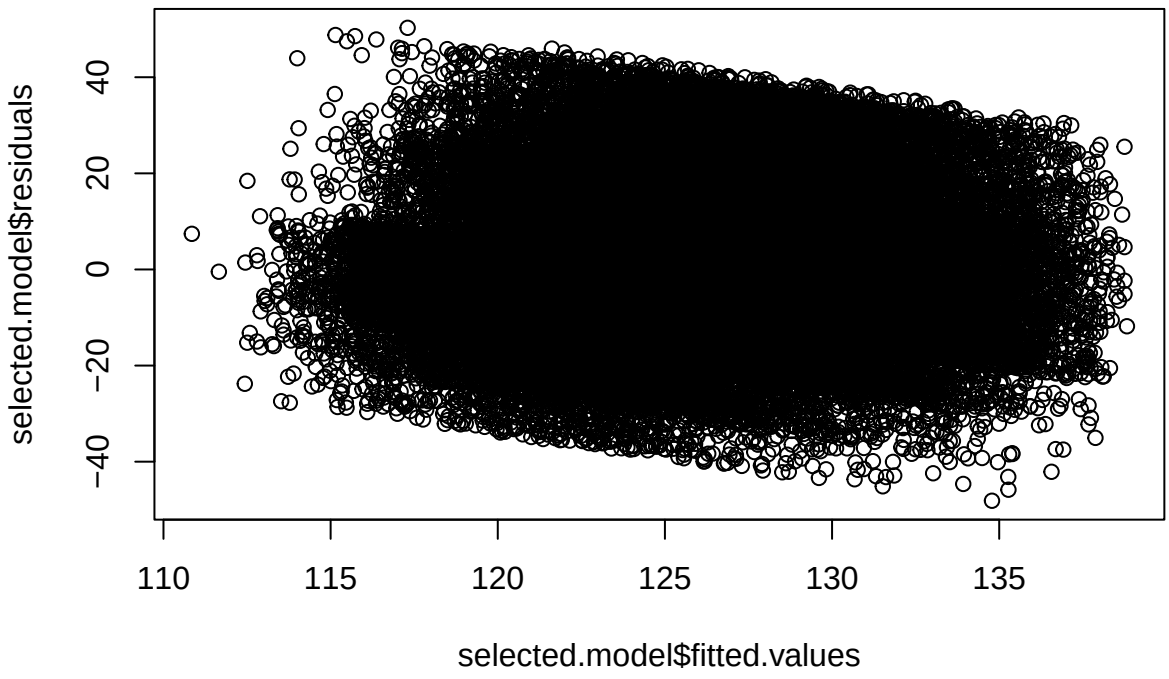
```r
#KS test na normalnost
ks.test(rstandard(fit.ap_hi),'pnorm')
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  rstandard(fit.ap_hi)
## D = 0.092807, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

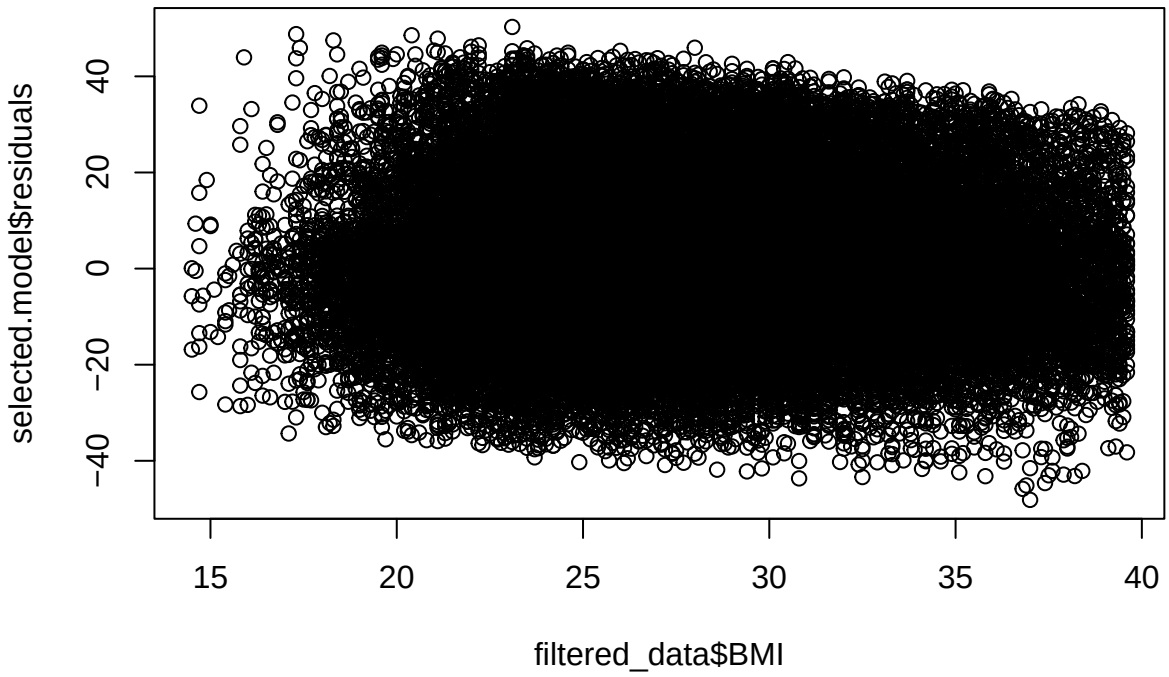
```r
require(nortest)
lillie.test(rstandard(fit.ap_hi))
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rstandard(fit.ap_hi)
## D = 0.092807, p-value < 2.2e-16
```

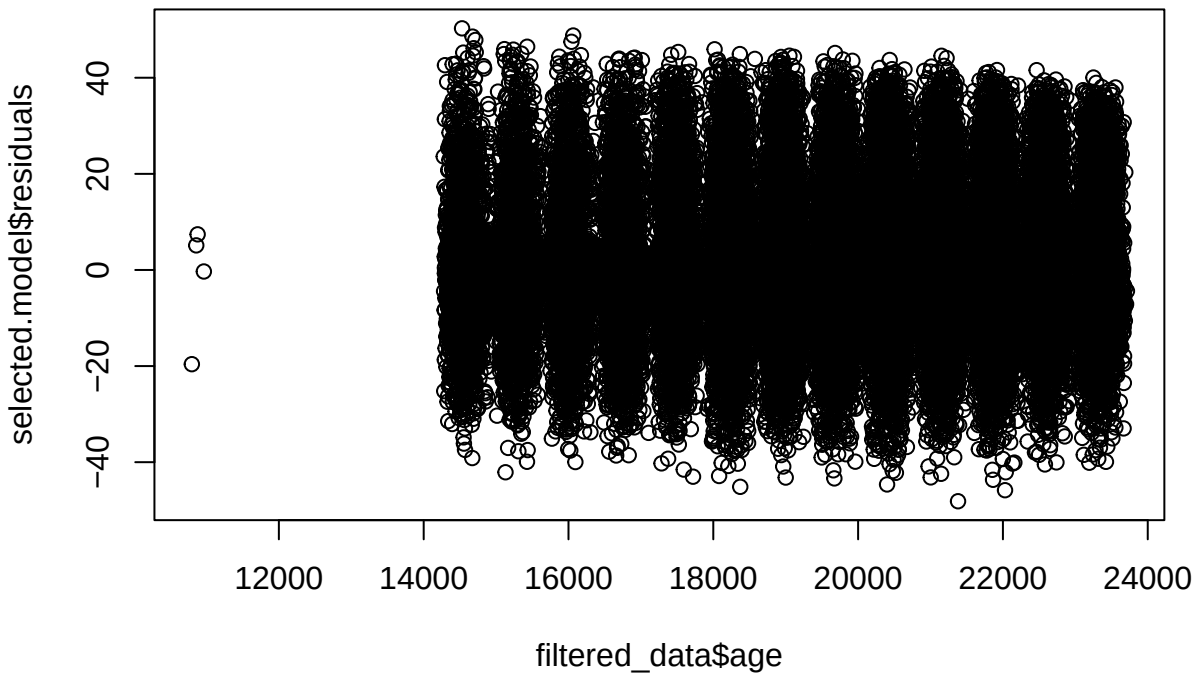
```r
selected.model = fit.multi
plot(selected.model$fitted.values,selected.model$residuals)
```



```r
plot(filtered_data$BMI, selected.model$residuals)
```




```r
plot(filtered_data$age, selected.model$residuals)
```



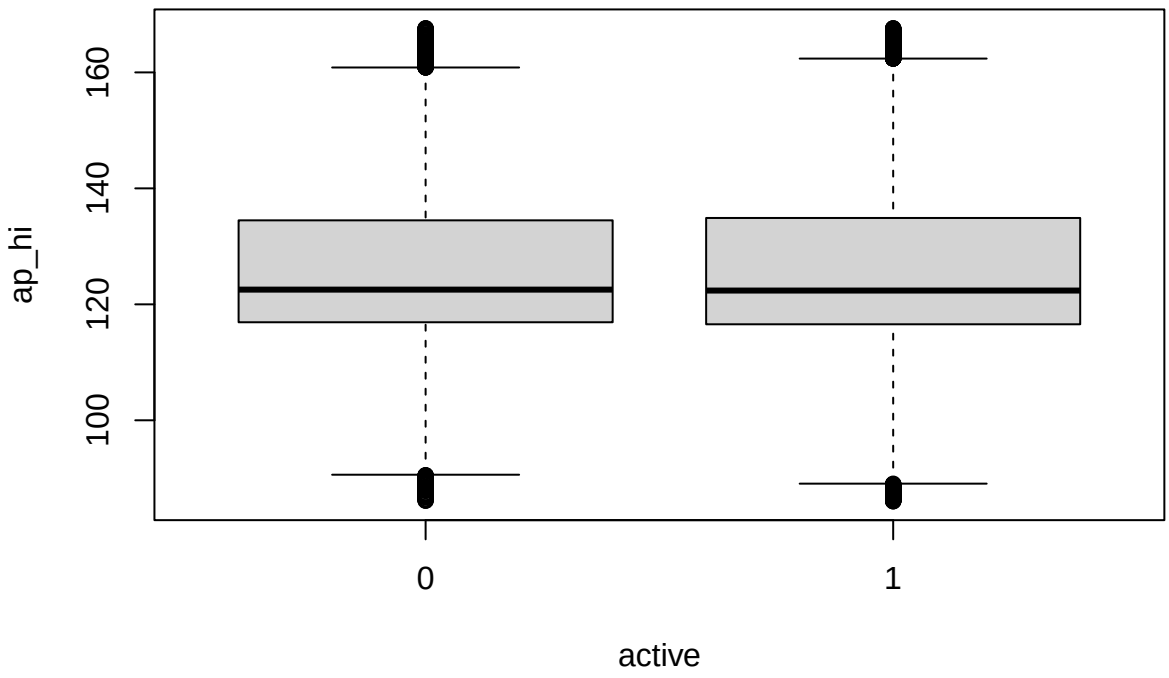

```
ks.test(rstandard(fit.multi),'pnorm')
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  rstandard(fit.multi)
## D = 0.081072, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

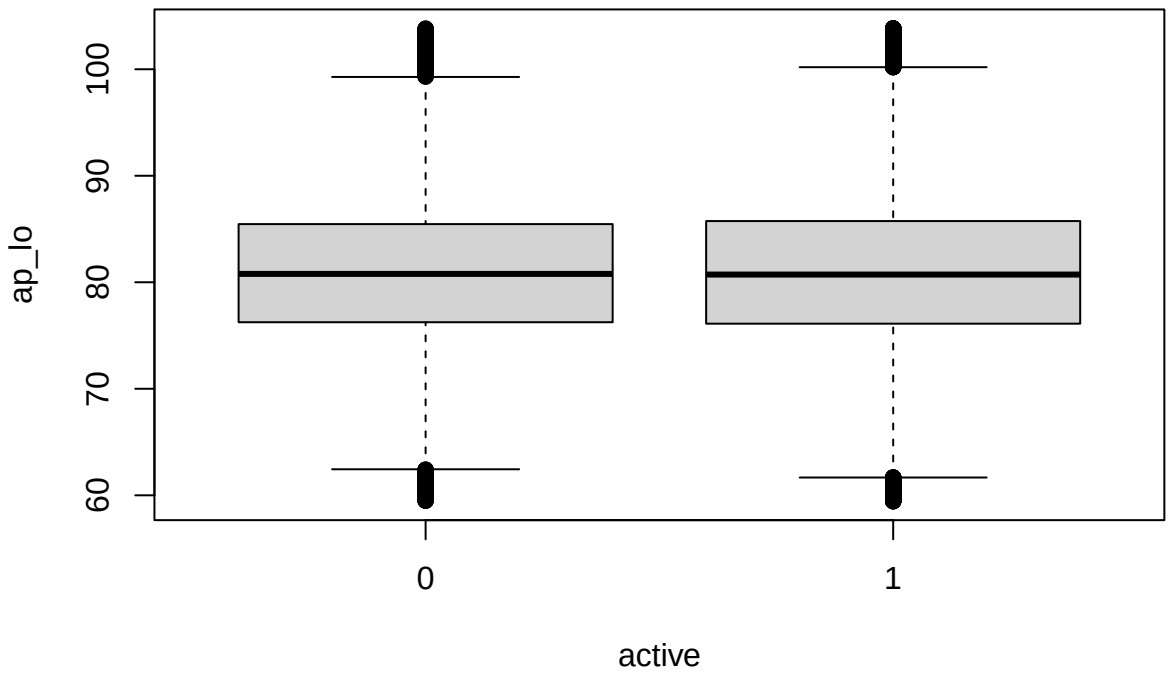
```
require(nortest)
lillie.test(rstandard(fit.multi))
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rstandard(fit.multi)
## D = 0.081073, p-value < 2.2e-16
```

```
boxplot(ap_hi~active,data=filtered_data)
```




```
boxplot(ap_lo~active,data=filtered_data)
```




```
notA <- subset(filtered_data, active == 0)
A <- subset(filtered_data, active == 1)
mean(notA$ap_hi)
```

```
## [1] 125.497
```

```
mean(A$ap_hi)
```

```
## [1] 125.4917
```

```
mean(notA$ap_lo)
```

```
## [1] 81.11432
```

```
mean(A$ap_lo)
```

```
## [1] 81.0064
```

Iz grafova gore se ne čini kao da tjelesna aktivnost uopće utječe na krvni tlak.