SAP projekt - 2023/2024 - World happiness report - grupa 'Vjerojatnost i statistika'

Učitavanje podataka

Prvo učitavamo biblioteke za rukovanje skupom podataka.

```
if (!tinytex:::is_tinytex()) {
  tinytex::install_tinytex() # *LINUX* paket za knittanje PDF-a
}

if (!requireNamespace("lawstat", quietly = TRUE)) {
  install.packages("lawstat")
}

if (!requireNamespace("nortest", quietly = TRUE)) {
  install.packages("nortest")
}

library(nortest)
library(lawstat)
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Kako bi ispravno analizirali podatke potrebno je zamijeniti neprisutne vrijednosti i rukovati izuzetcima.

Ukoliko su podatci u stupcu normalno raspoređeni, neprisutne vrijednosti zamijenjujemo srednjom vrijednosti stupca. U slučaju nenormalnog rasporeda zamijenjujemo ih prosjekom.

Pišemo funkciju za pronalazak i zamjenu neprisutnih vrijednosti:

```
zamijeni_nedostajuće_vrijednosti <- function(stupac) {
  if (is.numeric(stupac)) { # Samo numerički stupci
    if (shapiro.test(stupac)$p.value > 0.05 && sum(!is.na(stupac)) > 3) {
      # Normalna distribucija prema shapiro testu, zamijeni nedostajuće podatke sa srednjom vrijednosću
      stupac[is.null(stupac)] <- mean(stupac, na.rm = TRUE)
      stupac[is.na(stupac)] <- mean(stupac, na.rm = TRUE)
    } else {
      # Distribucija nije normalna, zamijeni nedostajuće podatke sa prosjekom
      stupac[is.null(stupac)] <- median(stupac, na.rm = TRUE)
      stupac[is.na(stupac)] <- median(stupac, na.rm = TRUE)
    }
  }
  return(stupac)
}
```

Izuzetke definiramo usporedbom svake vrijednosti sa kvantilima. Q1 definira vrijednosti ispod 25%, a Q3 iznad 75% raspona svih podataka. Razliku Q3 - Q1 definiramo kao IRQ, mjerom statističke disperzije koja opisuje raspon središnjih 50% podataka.

Podatke ispod Q1 - 1.5 IRQ i iznad Q3 + 1.5 IRQ uklanjamo, zamijenjujemo vrijednošću granica, ili zamijenjujemo srednjom ili prosječnom vrijednošću ovisno o rasporedbi svih podataka. Uklanjanjem podataka gubimo preciznost pa smo odlučili mijenjati ih srednjom vrijedošću ili prosjekom.

Definiramo funckiju za zamjenu izuzetaka:

```r
# Pronađi i zamijeni izuzetke
zamijeni_izuzetke <- function(stupac) {
  if (is.numeric(stupac)) {
    # Kvantili i IQR
    Q1 <- quantile(stupac, 0.25, na.rm = TRUE)
    Q3 <- quantile(stupac, 0.75, na.rm = TRUE)
    IQR <- Q3 - Q1

    # Granice
    donja_granica <- Q1 - 1.5 * IQR
    gornja_granica <- Q3 + 1.5 * IQR


    zamjenska_varijabla <- NA

    # Petlja kroz svaki redak stupca
    for (i in 1:length(stupac)) {
      if (!is.null(stupac[i]) & !is.na(stupac[i])){ # Samo vrijednosti, inače greška

        # Ako je vrijednost izuzetak
        if(stupac[i] < donja_granica || stupac[i] > gornja_granica) {

          # Jednom računamo zamjensku varijablu, za svaki stupac
          if (is.na(zamjenska_varijabla)) {

            # Na osnovi srednje vrijednosti ili prosjeka vrijednosti stupca koje su unutar granica
            neizuzetci <- stupac[stupac >= donja_granica & stupac <= gornja_granica]

            # Te na kojima odrađujemo test normalnosti
            if (length(neizuzetci) > 3 && shapiro.test(neizuzetci)$p.value > 0.05) {
              zamjenska_varijabla <- mean(neizuzetci, na.rm = TRUE)
            } else {
              zamjenska_varijabla <- median(neizuzetci, na.rm = TRUE)
            }
          }

          # Te zamijenimo vrijednost tom zamjenskom varijablom
          stupac[i] <- zamjenska_varijabla
        }
      }
    }
  }
  return(stupac)
}
```

Definirali smo funkcije za zamjenu izuzetaka i nedostajućih vrijednosti. Poredak kojim vršimo te funckije nad skupom podataka važne su u slučaju nenormalnih podataka. Ukoliko nepostojeće vrijednosti prvo namjestimo na prosjek svi izuzetci će biti namješteni na drugu vrijednost zbog novih podataka u drugoj funkciji (prije nepostojećih).

Iz tog razloga prvo uklanjamo izuzetke, zatim ispunjavamo nepostojeće vrijednosti.

Pišemo funckiju za učitavanje skupa podataka:

```r
# Učitavanje i preprocesiranje CSV-a
učitaj_csv <- function(file_path) {
  data <- read_csv(file_path)
  data <- type_convert(data) # Definira tip podataka u svakom stupcu

  data <- data %>%
    mutate(across(where(is.numeric), ~zamijeni_izuzetke(.))) %>% # Zamijeni nedostajuće vrijednosti s s

    mutate(across(where(is.numeric), ~zamijeni_nedostajuće_vrijednosti(.))) # Zamijeni nedostajuće vrij

  return(data)
}
```

Te učitavamo podatke za 2022. i 2023. godinu:

```r
# Učitaj podatke za 2022.
file_path1 <- "~/Desktop/SAP_projekt/WHR_2022.csv"
data_22 <- učitaj_csv(file_path1)

# Učitaj podatke za 2023.
file_path2 <- "~/Desktop/SAP_projekt/WHR_2023.csv"
data_23 <- učitaj_csv(file_path2)

# Prikaz podataka
print(data_22)
```

```
## # A tibble: 147 x 2
##    Country     `Happiness score`
##    <chr>                   <dbl>
##  1 Finland                  7.82
##  2 Denmark                  7.64
##  3 Iceland                  7.56
##  4 Switzerland              7.51
##  5 Netherlands              7.41
##  6 Luxembourg*              7.40
##  7 Sweden                   7.38
##  8 Norway                   7.37
##  9 Israel                   7.36
## 10 New Zealand              7.20
## # i 137 more rows
```

```r
print(data_23)
```

```
## # A tibble: 137 x 15
##    `Country name` `Regional indicator`         `Ladder score` `GDP per capita`
##    <chr>          <chr>                                 <dbl>            <dbl>
##  1 Finland        Western Europe                         7.80           48631.
##  2 Denmark        Western Europe                         7.59           57651.
##  3 Iceland        Western Europe                         7.53           53935.
##  4 Israel         Middle East and North Africa           7.47           41719.
##  5 Netherlands    Western Europe                         7.40           56516.
##  6 Sweden         Western Europe                         7.40           53254.
```

```
##  7 Norway         Western Europe                        7.32          65364.
##  8 Switzerland    Western Europe                        7.24          70546.
##  9 Luxembourg     Western Europe                        7.23          13680.
## 10 New Zealand    North America and ANZ                 7.12          42696.
## # i 127 more rows
## # i 11 more variables: ‘Social support‘ <dbl>, ‘Healthy life expectancy‘ <dbl>,
## #   ‘Freedom to make life choices‘ <dbl>, Generosity <dbl>,
## #   ‘Perceptions of corruption‘ <dbl>,
## #   ‘Alcohol consumption Both Sexes (L/year)‘ <dbl>,
## #   ‘Alcohol consumption Male (L/year)‘ <dbl>,
## #   ‘Alcohol consumption Female (L/year)‘ <dbl>, ...
```

S time smo dovršili obradu skupa podataka.

1. Je li razina sreće u publikaciji za 2023. veća ili manja u usporedbi s istraživanjem provedenim godinu ranije?

Planiramo provesti t-test za nezavisne uzorke kako bismo usporedili razinu sreće u publikaciji za 2023. s istraživanjem provedenim godinu ranije. Prije provođenja testa, moramo ispuniti određene pretpostavke, uključujući normalnost podataka i homogenost varijanci. Želimo osigurati da oba uzorka imaju približno normalnu distribuciju i slične varijance kako bismo osigurali valjanost rezultata.

Prvi potrebni korak je obrada i usklađivanje podataka o zemljama iz dva različita vremenska razdoblja 2022. i 2023. godine, istraživanje promjena u nazivima zemalja te identifikacija zemalja koje su prisutne ili odsutne u samo jednom od skupova podataka.

```
# Uklonanje '*' s kraja naziva zemalja u data_2022
data_22$'Modified Country' <- gsub("\\*$", "", data_22$Country)

# Zamjena "Turkey" s "Turkiye"
data_22_cleaned <- data_22 %>%
  mutate(`Modified Country` = ifelse(`Modified Country` == "Turkey", "Turkiye", `Modified Country`))

# Zamjena "Palestinian Territories" s "State of Palestine"
data_22_cleaned <- data_22_cleaned %>%
  mutate(`Modified Country` = ifelse(`Modified Country` == "Palestinian Territories", "State of Palesti

# Pronalaženje zemalja koje su samo u jednom skupu podataka
unique_countries_22 <- unique(data_22_cleaned$`Modified Country`)
unique_countries_23 <- unique(data_23$`Country name`)

countries_only_in_22 <- setdiff(unique_countries_22, unique_countries_23)
countries_only_in_23 <- setdiff(unique_countries_23, unique_countries_22)

# Uklananje zemalja koje su samo u jednom skupu podataka
data_22_filtered <- data_22_cleaned[data_22_cleaned$`Modified Country` %in% intersect(unique_countries_
data_23_filtered <- data_23[data_23$`Country name` %in% intersect(unique_countries_22, unique_countries

# Ispis zemlji koje su samo u jednom skupu podataka
cat("Zemlje samo u 2022. godini:", toString(countries_only_in_22), "\n")
```

```
## Zemlje samo u 2022. godini: Kuwait, Belarus, Turkmenistan, North Cyprus, Libya, Azerbaijan, Congo, Es
```

```r
cat("Zemlje samo u 2023. godini:", toString(countries_only_in_23), "\n")
```

```
## Zemlje samo u 2023. godini: Congo (Brazzaville), Congo (Kinshasa)
```

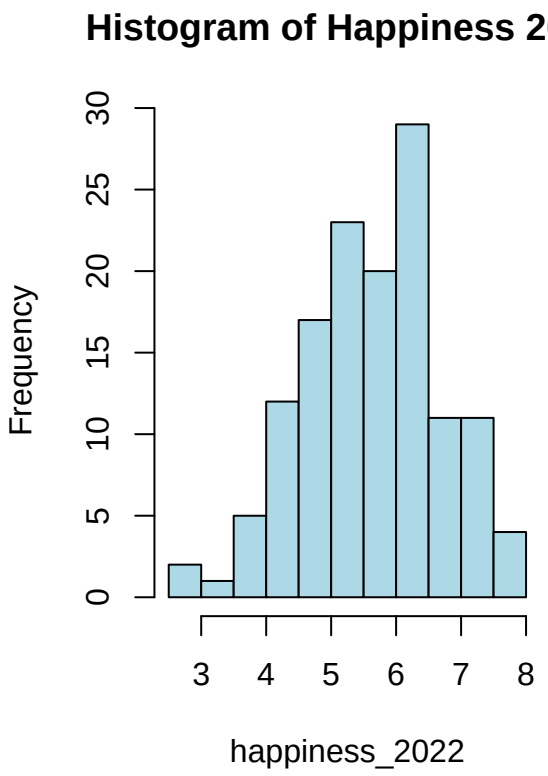
Izdvajanje relevantnih podataka o sreći iz različitih godina, što olakšava daljnju analizu podataka između te dvije različite godine.

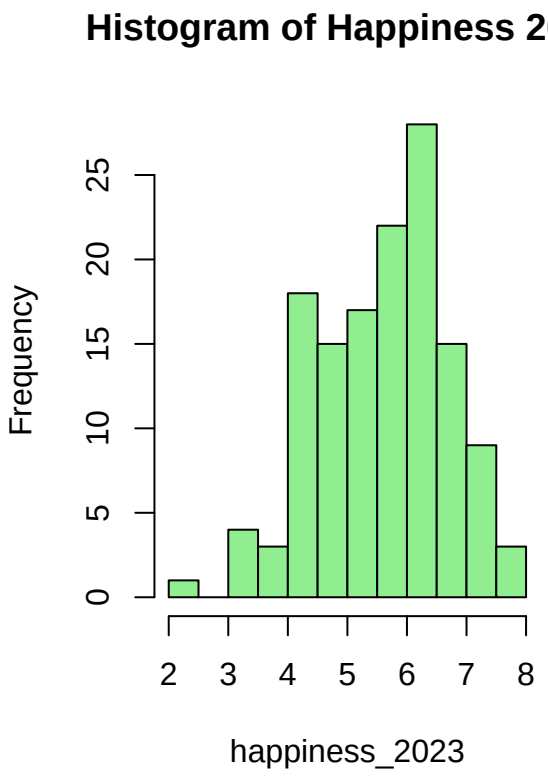
```r
happiness_2022 <- data_22_filtered$'Happiness score'
happiness_2023 <- data_23_filtered$'Ladder score'
```

Daljnji korak je crtanje podataka. Crtanjem grafova ne možemo jednoznačno potvrditi pretpostavke nužne za provođenje našeg testa, no vizualna analiza može nam pomoći u donošenju zaključka o istima.

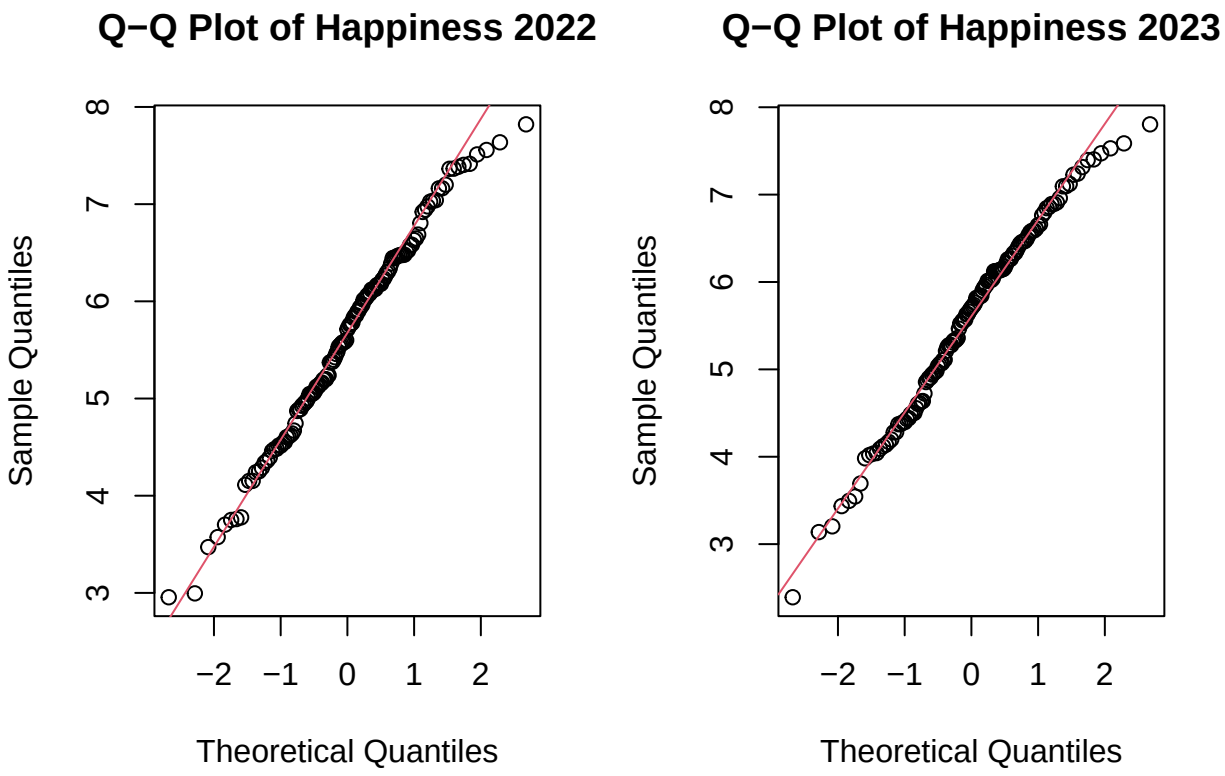
```r
# Crtanje Boxplot za varijable happiness_2022 i happiness_2023
boxplot(happiness_2022, happiness_2023, names = c("2022", "2023"), main = "Boxplot of Happiness Score")
```


```r
# Crtanje histograma za varijable happiness_2022 i happiness_2023
par(mfrow = c(1, 2))
hist(happiness_2022, main = "Histogram of Happiness 2022", col = "lightblue")
hist(happiness_2023, main = "Histogram of Happiness 2023", col = "lightgreen")
```






```r
# Cratanje Q-Q za varijable happiness_2022 i happiness_2023
par(mfrow = c(1, 2))
qqnorm(happiness_2022, main = "Q-Q Plot of Happiness 2022")
qqline(happiness_2022, col = 2)
qqnorm(happiness_2023, main = "Q-Q Plot of Happiness 2023")
qqline(happiness_2023, col = 2)
```



Kako bismo potvrdili normalnost podataka, izvršit ćemo Lilliefors test, koji predstavlja varijaciju Kolmogorov-Smirnov testa za ispitivanje normalnosti distribucije.

```
# Testiranje normalnosti za sreću u 2022. i 2023. godini koristeći Lilliefors test
install.packages("nortest")
```

```
## Installing package into '/home/feliks/R/x86_64-pc-linux-gnu-library/4.1'
## (as 'lib' is unspecified)
```

```
library(nortest)
lillie.test(happiness_2022)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  happiness_2022
## D = 0.050881, p-value = 0.5324
```

```
lillie.test(happiness_2023)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  happiness_2023
## D = 0.060214, p-value = 0.2694
```

S obzirom na alfa od 5%, možemo zaključiti da nema dovoljno statističkih dokaza za odbacivanje hipoteze da podaci dolaze iz normalne distribucije za obje godine. Pretpostavljamo normalnost podataka te nastavljamo dalje s ispitivanjem.

Preostaje nam još provjeriti homogenost varijanci koristeći Bartlettov test, koji ispituje hipotezu H0 da su sve varijance u populacijama jednake.

```
# Test homogenosti varijanci 2022. i 2023. godini koristeći Bartlett test
bartlett.test(list(happiness_2022, happiness_2023))
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  list(happiness_2022, happiness_2023)
## Bartlett's K-squared = 0.15002, df = 1, p-value = 0.6985
```

S obzirom na visoku p-vrijednost, nemamo dovoljno statističkih dokaza za odbacivanje nul hipoteze. Dakle, možemo pretpostaviti homogenost varijanci između tih skupova podataka, što znači da se varijance podataka u obje godine smatraju sličnima na razini značajnosti od 5%.

```
# Dvostrani t test
t_test_result <- t.test(happiness_2022, happiness_2023, paired = TRUE)

# Print the result
print(t_test_result)
```

```
##
##  Paired t-test
##
## data:  happiness_2022 and happiness_2023
## t = 5.8125, df = 134, p-value = 4.279e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.0292879 0.0594999
## sample estimates:
## mean of the differences
##               0.0443939
```

Zaključak: : Pareni t-test pokazuje statistički značajnu razliku u razini sreće između 2022. i 2023. godine. Budući da je p-vrijednost znatno manja od alfa razine od 5%, odbacujemo nul hipotezu o jednakosti srednjih vrijednosti.

Stoga, na temelju ovog testa, možemo zaključiti da **postoji statistički značajna razlika u razini sreće između 2022. i 2023. godine**, pri čemu je srednja vrijednost sreće u 2023. godini prosječno viša za 0.0443939.

2.Možemo li temeljem drugih dostupnih varijabli predvidjeti konzumaciju alkohola po zemljama?

##Jednostavna linearna regresija

###Utjecaj nezavisne varijable na zavisnu

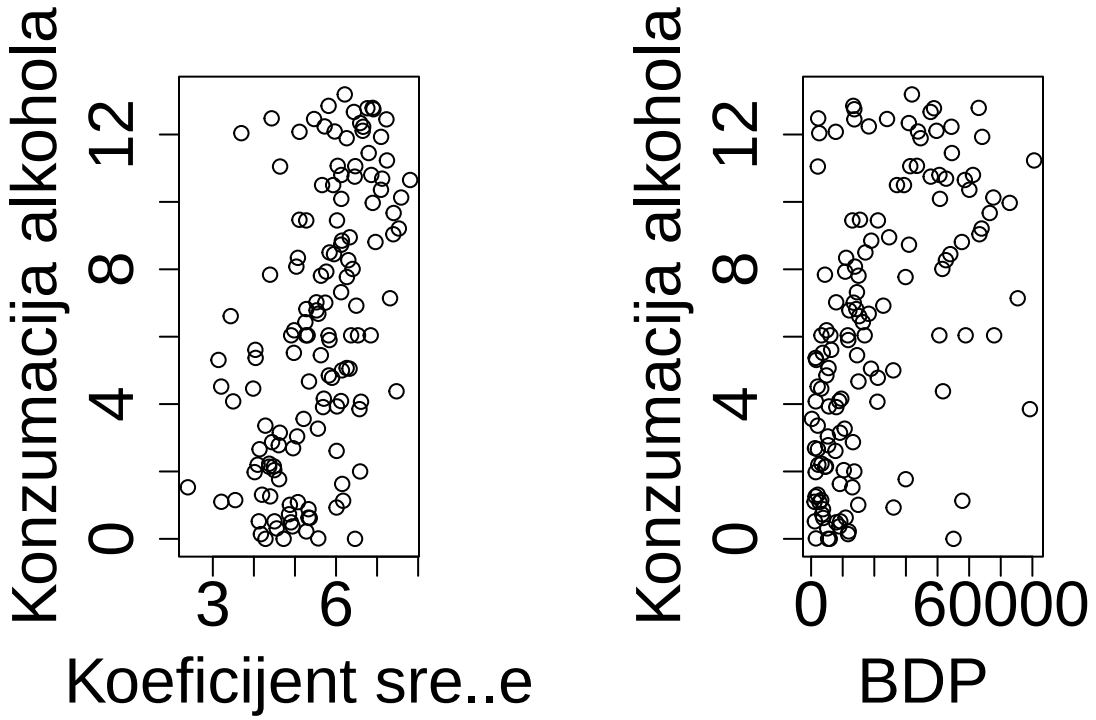
Ispituju se određene mjere i promatra se njihov "utjecaj" na konzumaciju alkohola po zemljama. Odabrane su sljedeće mjere:koeficijent sreće, bdp po glavi stanovnika, socijalna skrb, očekivanje zdravog života, darežljivost, percepcija korupcije, indeks stope kriminala i Gini koeficijent od Svjeske banke. Utjecaj pojedine nezavisne varijable na zavisnu varijablu (konzumacija alkohola) prikazan je pomoću scatterplot-a.

```
par(mfrow=c(1,2),mai=c(1,1,1,1))
plot(data_23$"Ladder score",data_23$"Alcohol consumption Both Sexes (L/year)", xlab="Koeficijent sreće"
```

```
## Warning in title(...): conversion failure on 'Koeficijent sreće' in
## 'mbcsToSbcs': dot substituted for <c4>
```

```
## Warning in title(...): conversion failure on 'Koeficijent sreće' in
## 'mbcsToSbcs': dot substituted for <87>
```

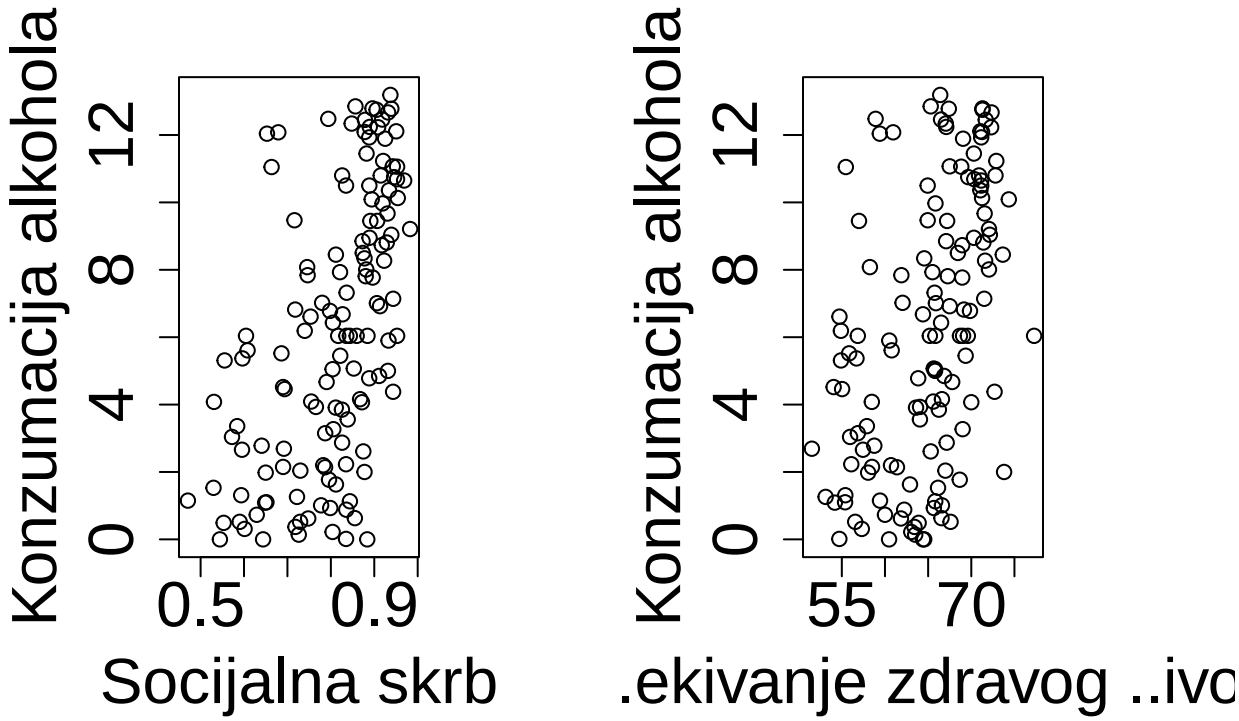
```
plot(data_23$"GDP per capita",data_23$"Alcohol consumption Both Sexes (L/year)", xlab="BDP",ylab="Konzum
cex.main=2,cex.lab=2,cex.axis=2)#gdp/alcohol
```




```
plot(data_23$"Social support",data_23$"Alcohol consumption Both Sexes (L/year)", xlab="Socijalna skrb",y
cex.main=2,cex.lab=2,cex.axis=2)#socsup/alcohol
plot(data_23$"Healthy life expectancy",data_23$"Alcohol consumption Both Sexes (L/year)", xlab="Očekivan
```

```
## Warning in title(...): conversion failure on 'Očekivanje zdravog života' in
## 'mbcsToSbcs': dot substituted for <c4>
```

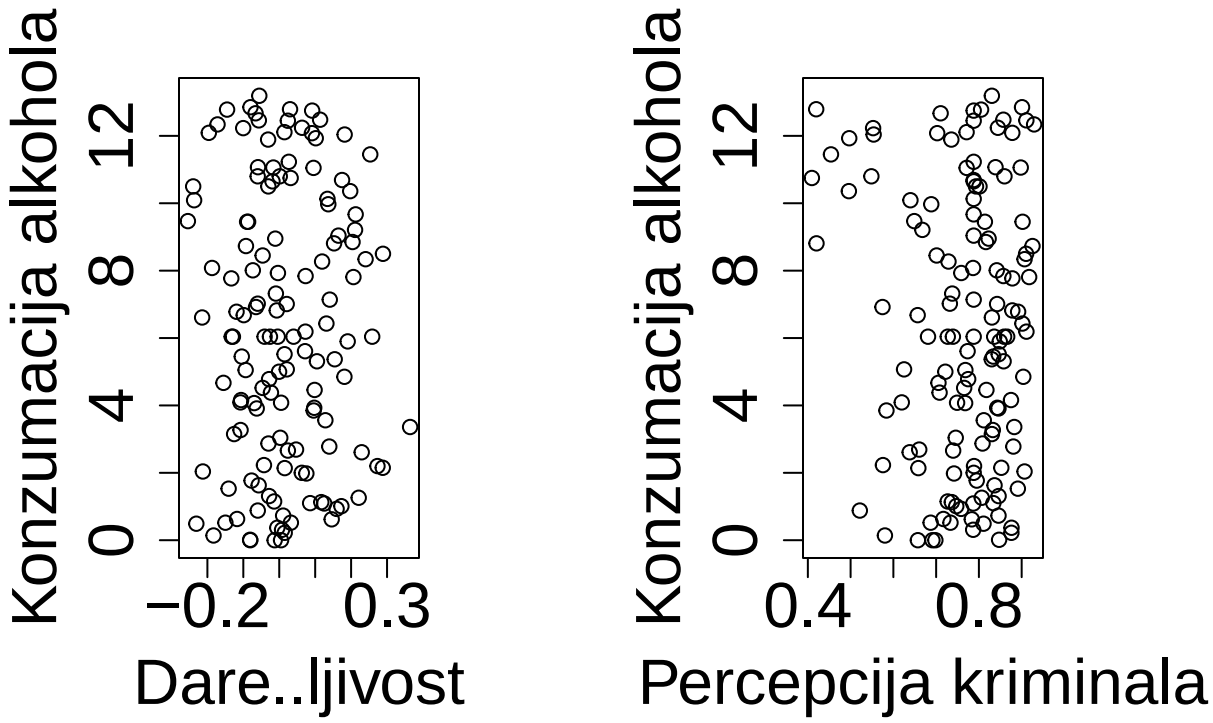
```
## Warning in title(...): conversion failure on 'Očekivanje zdravog života' in
## 'mbcsToSbcs': dot substituted for <8d>

## Warning in title(...): conversion failure on 'Očekivanje zdravog života' in
## 'mbcsToSbcs': dot substituted for <c5>

## Warning in title(...): conversion failure on 'Očekivanje zdravog života' in
## 'mbcsToSbcs': dot substituted for <be>
```



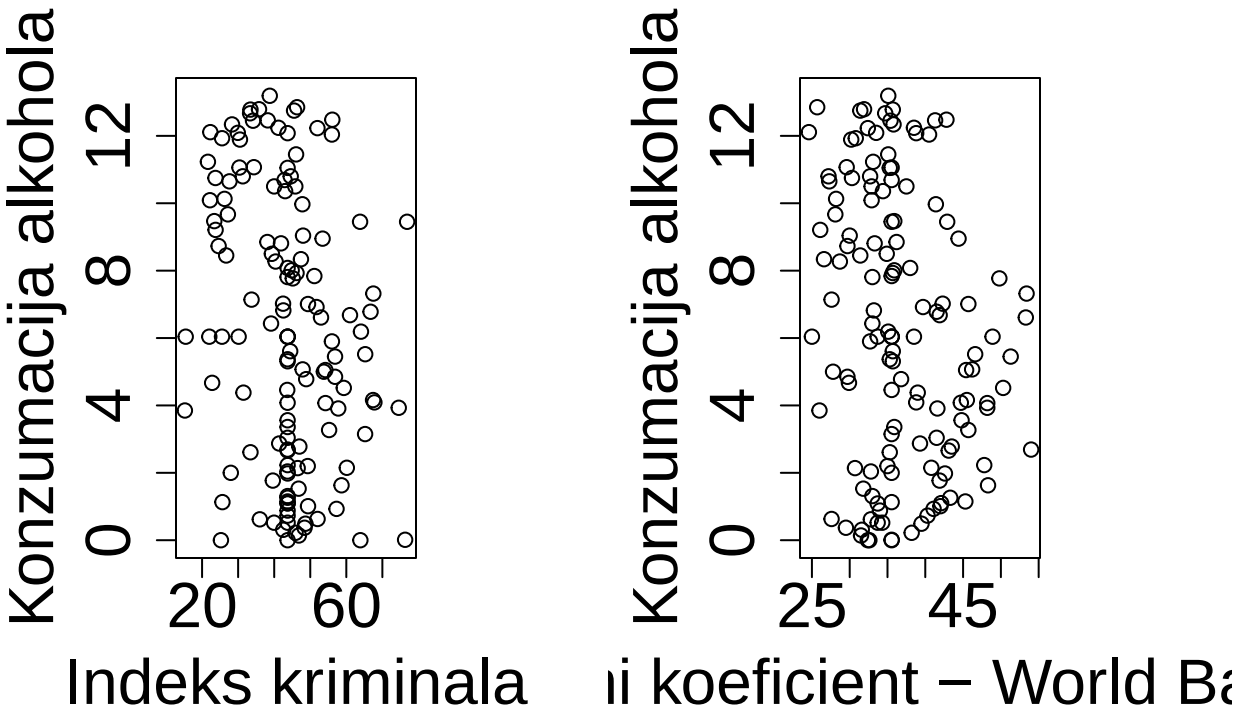

```
plot(data_23$"Generosity",data_23$"Alcohol consumption Both Sexes (L/year)", xlab="Darežljivost",ylab="
```

```
## Warning in title(...): conversion failure on 'Darežljivost' in 'mbcsToSbcs':
## dot substituted for <c5>

## Warning in title(...): conversion failure on 'Darežljivost' in 'mbcsToSbcs':
## dot substituted for <be>
```

```
plot(data_23$"Perceptions of corruption",data_23$"Alcohol consumption Both Sexes (L/year)", xlab="Percep
```




```
plot(data_23$"Crime rate Crime Index",data_23$"Alcohol consumption Both Sexes (L/year)", xlab="Indeks k
plot(data_23$"Gini Coefficient - World Bank",data_23$"Alcohol consumption Both Sexes (L/year)", xlab="G
```



Rezultati upućuju na to da bi koeficijent sreće, gdp po glavi stanovnika, socijalna skrb, očekivanje zdravog života i percepcija korupcije mogli imati utjecaj na konzumaciju alkohola u državi, dok darežljivost, indeks stope kriminala i Gini koeficijent od Svjeske banke nemaju utjecaja na konzumaciju alkohola po državi.
###Korelacijski koeficijent i veza s linearnim modelom

Korelacijski koeficijent opisuje smjer i prirodu veze dviju varijabli. Izvršit ćemo korelacijski test nad varijablama za koje se čini da imaju utjecaj na konzumaciju alkohola po državi.

```
cor.test(data_23$"Ladder score",data_23$"Alcohol consumption Both Sexes (L/year)")
```

```
##
##  Pearson's product-moment correlation
##
## data:  data_23$"Ladder score" and data_23$"Alcohol consumption Both Sexes (L/year)"
## t = 7.1832, df = 135, p-value = 4.154e-11
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3927769 0.6373568
## sample estimates:
##      cor
## 0.525852
```

```
cor.test(data_23$"GDP per capita",data_23$"Alcohol consumption Both Sexes (L/year)")
```

```
##
##  Pearson's product-moment correlation
##
## data:  data_23$"GDP per capita" and data_23$"Alcohol consumption Both Sexes (L/year)"
## t = 6.9354, df = 135, p-value = 1.527e-10
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3772514 0.6264073
## sample estimates:
##       cor
## 0.5125381
```

```
cor.test(data_23$"Social support",data_23$"Alcohol consumption Both Sexes (L/year)")
```

```
##
##  Pearson's product-moment correlation
##
## data:  data_23$"Social support" and data_23$"Alcohol consumption Both Sexes (L/year)"
## t = 7.8571, df = 135, p-value = 1.101e-12
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4331508 0.6653758
## sample estimates:
##      cor
## 0.560172
```

```
cor.test(data_23$"Healthy life expectancy",data_23$"Alcohol consumption Both Sexes (L/year)")
```

```
##
##  Pearson's product-moment correlation
##
## data:  data_23$"Healthy life expectancy" and data_23$"Alcohol consumption Both Sexes (L/year)"
```

```
## t = 6.3254, df = 135, p-value = 3.444e-09
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3374875 0.5979078
## sample estimates:
##       cor
## 0.4781394
```

```
cor.test(data_23$"Perceptions of corruption",data_23$"Alcohol consumption Both Sexes (L/year)")
```

```
##
##  Pearson's product-moment correlation
##
## data:  data_23$"Perceptions of corruption" and data_23$"Alcohol consumption Both Sexes (L/year)"
## t = -0.81844, df = 135, p-value = 0.4146
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.23520931  0.09861191
## sample estimates:
##         cor
## -0.07026568
```

Najveću korelaciju sa konzumacijom alkohola po državi pokazuje socijalna skrb. P-vrijednosti svih varijabli osim percepcije korupcije ukazuju na to da postoji veza između njih i konzumacije alkohola po državama.
###Primjena modela linearne regresije

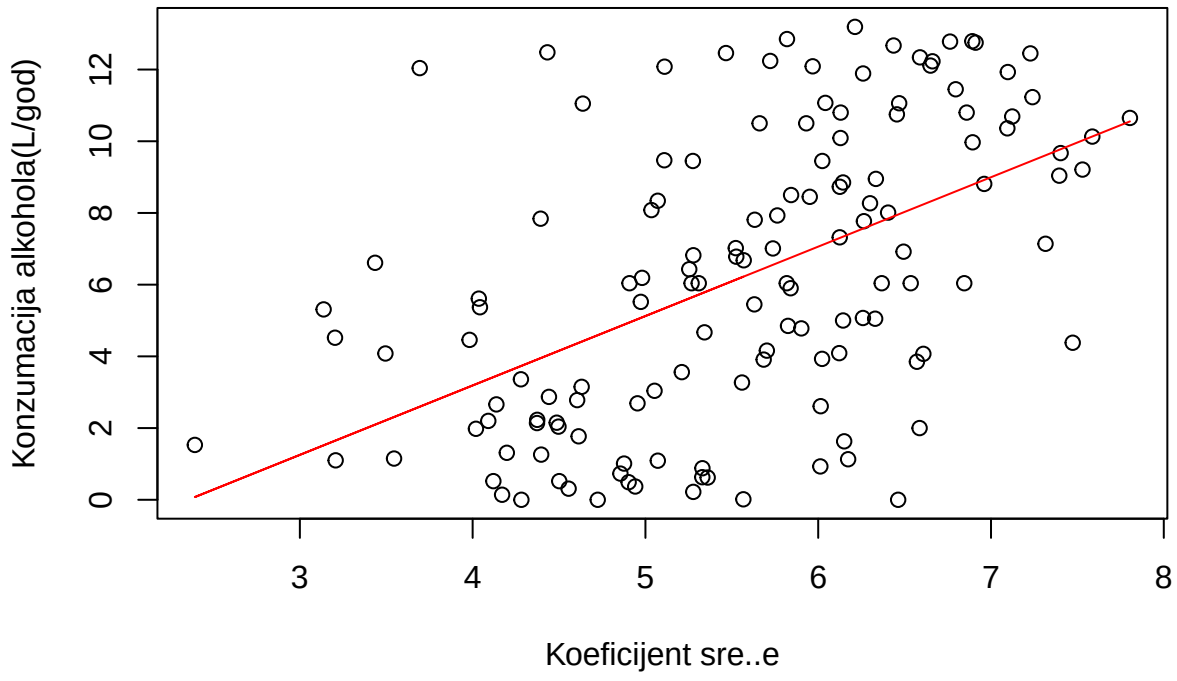
U nastavku je isproban model jednostavne linearne regresije - procjenjen je odnos jedne nezavisne varijable (regresora) i jedne zavisne varijable.

Svi modeli prikazani su grafički zajedno sa nezavisnom varijablom kojom predviđaju konzumaciju alkohola po državi.

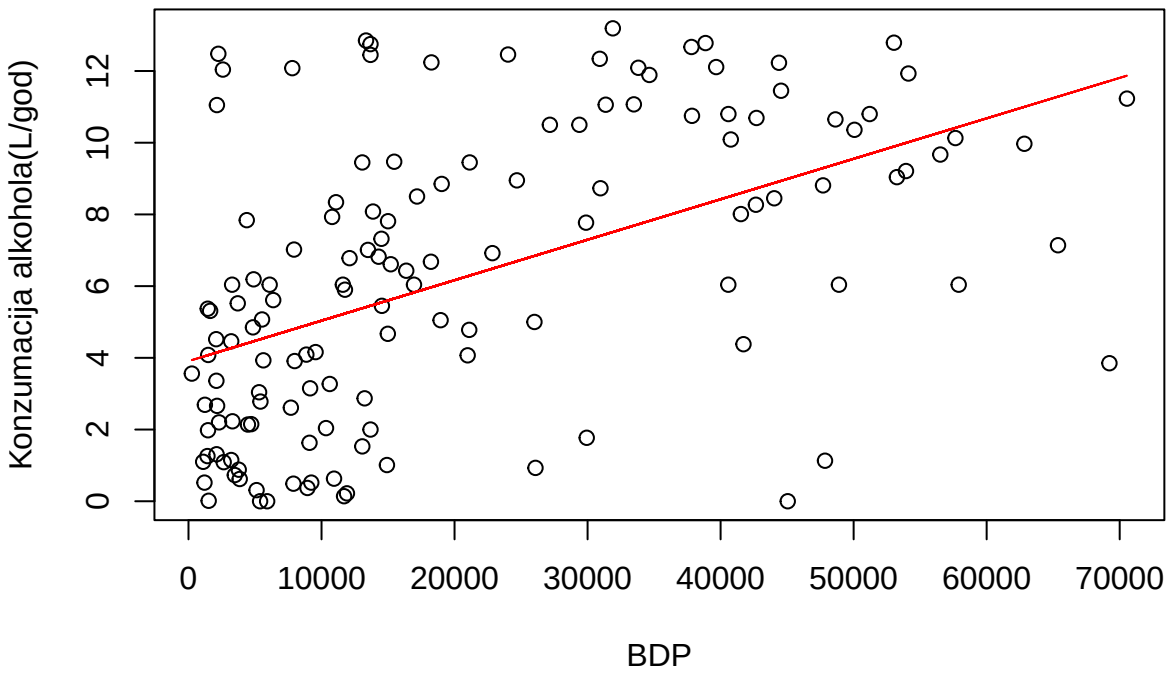
```
fit.ladder <- lm(`Alcohol consumption Both Sexes (L/year)` ~ `Ladder score`, data = data_23)

# Grafički prikaz podataka
plot(data_23$`Ladder score`, data_23$`Alcohol consumption Both Sexes (L/year)`,
     xlab = "Koeficijent sreće", ylab = "Konzumacija alkohola(L/god)", cex.main = 1, cex.lab = 1, cex.a
```

```
## Warning in title(...): conversion failure on 'Koeficijent sreće' in
## 'mbcsToSbcs': dot substituted for <c4>

## Warning in title(...): conversion failure on 'Koeficijent sreće' in
## 'mbcsToSbcs': dot substituted for <87>
```

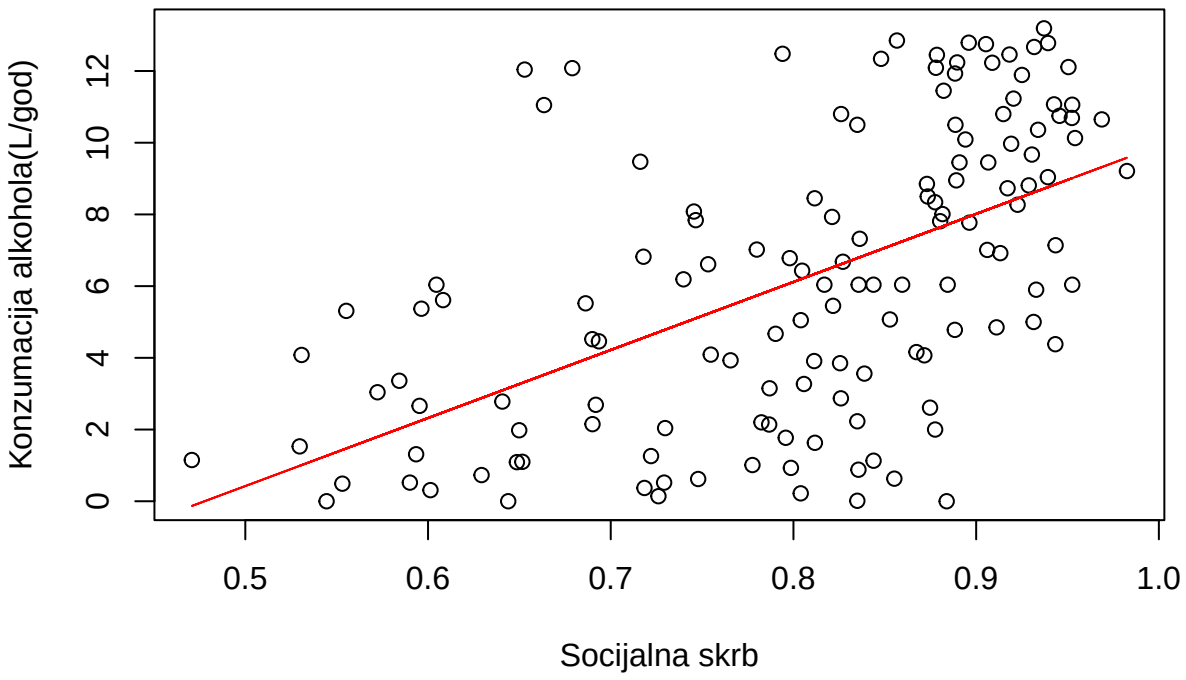
```
lines(data_23$`Ladder score`, fit.ladder$fitted.values, col = 'red')
```




```
fit.gdp <- lm(`Alcohol consumption Both Sexes (L/year)` ~ `GDP per capita`, data = data_23)
# Grafički prikaz podataka
plot(data_23$`GDP per capita`, data_23$`Alcohol consumption Both Sexes (L/year)`,
     xlab = "BDP", ylab = "Konzumacija alkohola(L/god)", cex.main = 1, cex.lab = 1, cex.axis = 1)

lines(data_23$`GDP per capita`, fit.gdp$fitted.values, col = 'red')
```



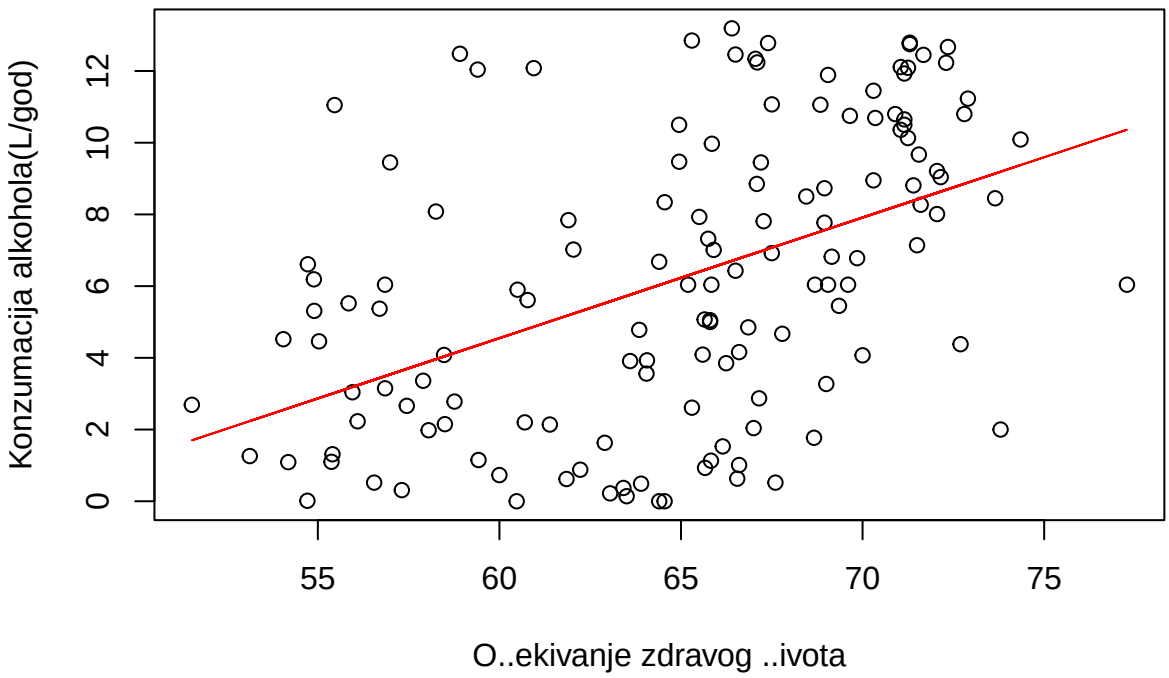

```
fit.ss <- lm(`Alcohol consumption Both Sexes (L/year)` ~ `Social support`, data = data_23)
# Grafički prikaz podataka
plot(data_23$`Social support`, data_23$`Alcohol consumption Both Sexes (L/year)`,
     xlab = "Socijalna skrb", ylab = "Konzumacija alkohola(L/god)", cex.main = 1, cex.lab = 1, cex.axis

lines(data_23$`Social support`, fit.ss$fitted.values, col = 'red')
```




```
fit.hle <- lm(`Alcohol consumption Both Sexes (L/year)` ~ `Healthy life expectancy`, data = data_23)

# Grafički prikaz podataka
plot(data_23$`Healthy life expectancy`, data_23$`Alcohol consumption Both Sexes (L/year)`,
     xlab = "Očekivanje zdravog života", ylab = "Konzumacija alkohola(L/god)", cex.main = 1, cex.lab =
```

```
## Warning in title(...): conversion failure on 'Očekivanje zdravog života' in
## 'mbcsToSbcs': dot substituted for <c4>

## Warning in title(...): conversion failure on 'Očekivanje zdravog života' in
## 'mbcsToSbcs': dot substituted for <8d>

## Warning in title(...): conversion failure on 'Očekivanje zdravog života' in
## 'mbcsToSbcs': dot substituted for <c5>

## Warning in title(...): conversion failure on 'Očekivanje zdravog života' in
## 'mbcsToSbcs': dot substituted for <be>
```

```
lines(data_23$`Healthy life expectancy`, fit.hle$fitted.values, col = 'red')
```


Konzumacija alkohola(L/god)
0
2
4
6
8
10
12
55
60
65
70
75
O..ekivanje zdravog ..ivota

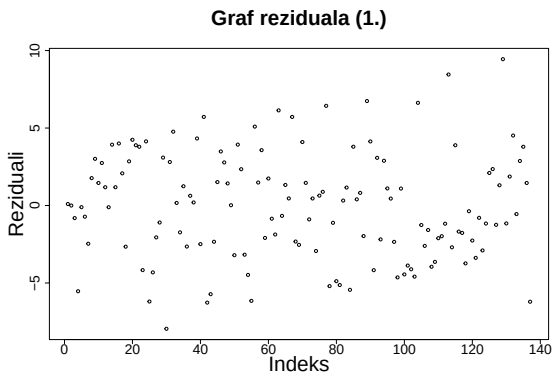
Da bi smo zadržali modele, potrebno ih je dalje analizirati i usporediti. Prvo moramo provjeriti da li su narušene pretpostavke modela jednostavne linearne regresije. Potrebno je provjeriti normalnost reziduala i homogenost varijance. Ukoliko je nešto od toga narušeno, model se odbacuje. Normalnost reziduala ćemo provjeriti pomoću kvantil-kvantil plota grafički, a statistički pomoću Kolmogorov-Smirnovljevog testa ili Lilifeforceovom inačicom. Razina signifikantnosti za svaki test će biti 5%. Ova funkcija koristi se za provjeru normalnosti reziduala i homogenosti varijance u modelima linearne regresije. U tu svrhu uvodimo sljedeću funkciju

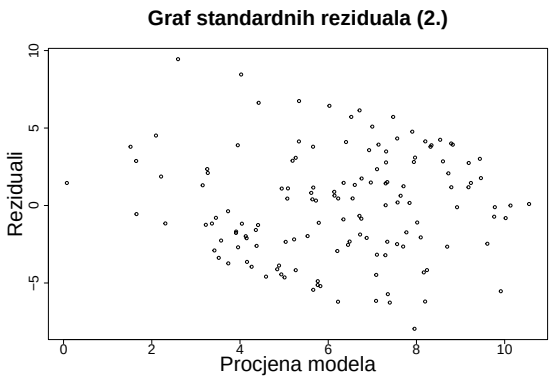
```r
normality_homogenity <- function(selected.model){

  par(mfrow=c(3,2), mai=c(1,1,1,1))

  #1 prikaz reziduala po indeksu danom u podacima
  p1 = plot(selected.model$residuals,
       main="Graf reziduala (1.)", xlab = "Indeks", ylab = "Reziduali",
       cex.main = 3, cex.lab = 3, cex.axis = 2)

  #2 prikaz reziduala u ovisnosti o procjenama modela
  p2 = plot(selected.model$fitted.values,selected.model$residuals,
       main="Graf standardnih reziduala (2.)", xlab = "Procjena modela", ylab = "Reziduali",
       cex.main = 3, cex.lab = 3, cex.axis = 2)

  #3 histogram reziduala
  hist((selected.model$residuals),
       xlab = "Reziduali", main = "Histogram reziduala (3.)",
       cex.main = 3, cex.lab = 3, cex.axis = 2)

  #4 histogram standardiziranih reziduala
  hist(rstandard(selected.model),
       xlab = "Standardni reziduali", main = "Histogram standardnih reziduala (4.)",
       cex.main = 3, cex.lab = 3, cex.axis = 2)

  #5 q-q plot standardiziranih reziduala s linijom normalne distribucije
  qqnorm(rstandard(selected.model),
         main="q-q plot standardiziranih reziduala (5.)", cex.main = 3, cex.lab = 3, cex.axis = 2)
  qqline(rstandard(selected.model))

  #6 Kolmogorov-Smirnovljev test - usporedba standardiziranih reziduala s normalnom razdiobom
  print(ks.test(rstandard(selected.model),'pnorm'))

  #7 Lillieforce inaciča KS-testa
  print(lillie.test(rstandard(selected.model)))

}
```

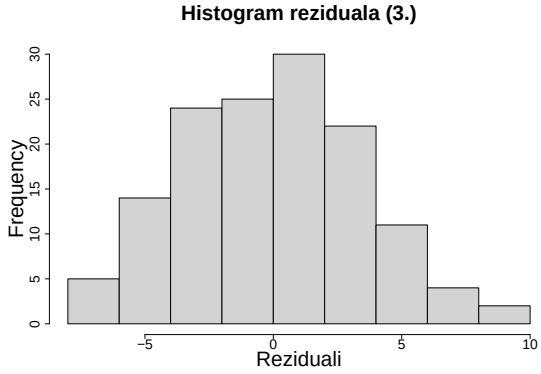
###Koeficijent sreće

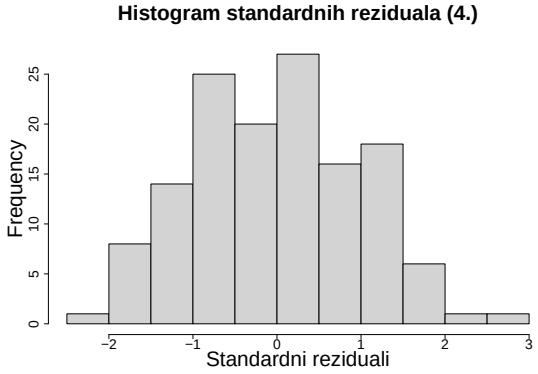
```
normality_homogenity(fit.ladder)
```

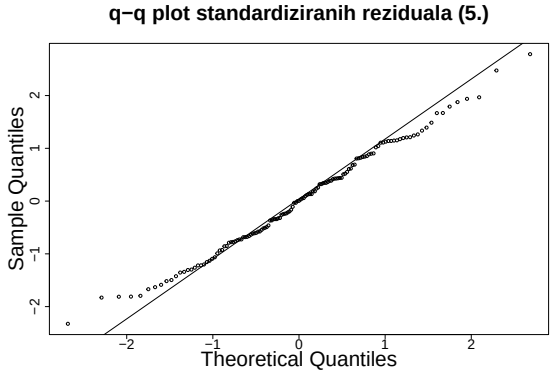
```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  rstandard(selected.model)
## D = 0.04653, p-value = 0.9281
## alternative hypothesis: two-sided
##
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rstandard(selected.model)
## D = 0.046196, p-value = 0.6737
```







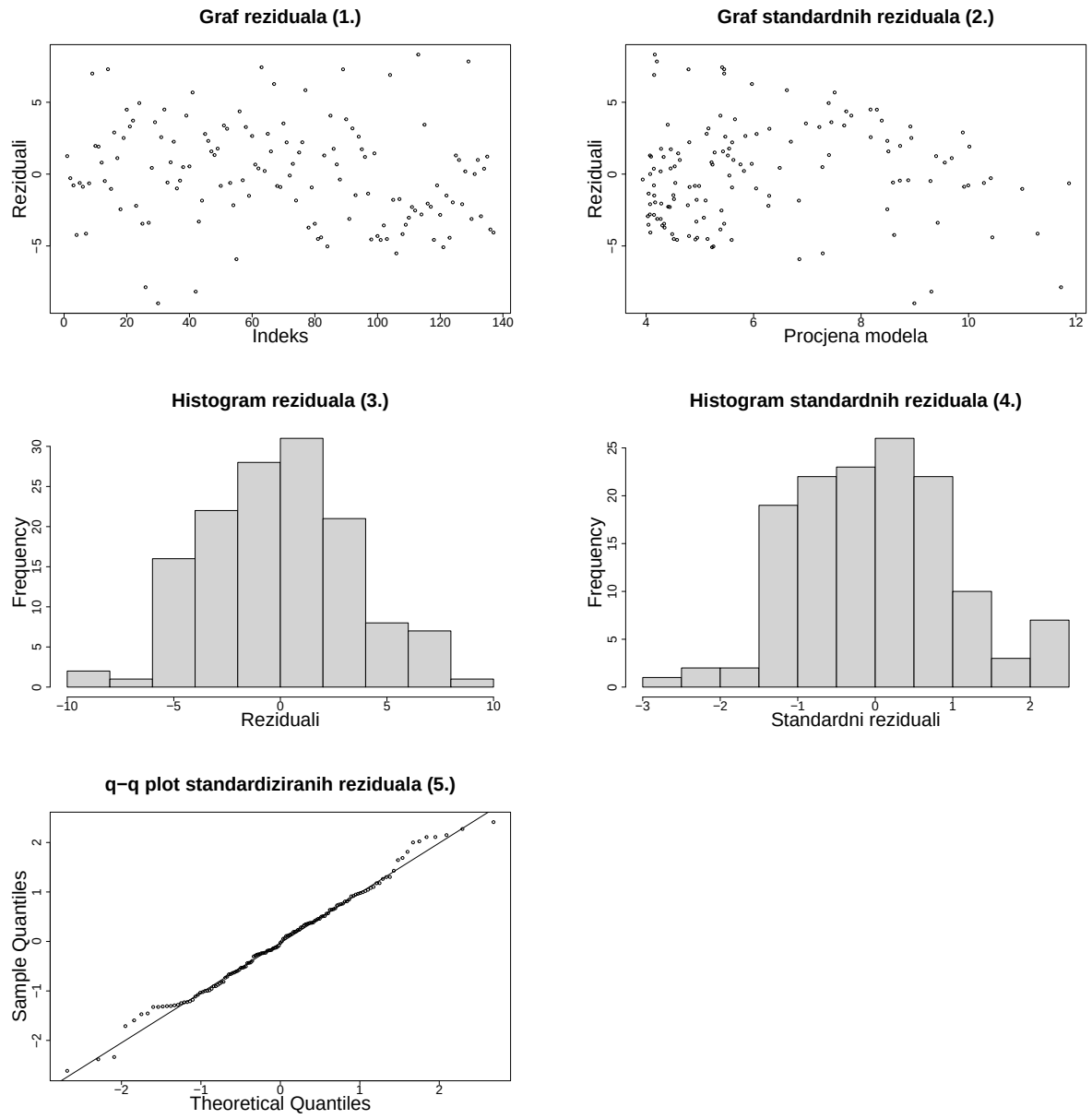




1.Reziduali se čine jednoliko raspršeno. 2. Reziduali se čine jednoliko raspršeni.3. Čini se da su reziduali raspodijeljeni simetričnom normalnom distribucijom. 4. Čini se da su reziduali podijeljeni dvostranom normalnom distribucijom. 5. Oblik qq-plota ukazuje na dvostranu normalnu distribuciju s debelim repovima. KM TEST i LILLIE TEST ukazuju da distribucija reziduala odgovara normalnoj distribuciji(p-value = 0.9281, p-value = 0.6737) ### Bruto društveni proizvod

```r
normality_homogenity(fit.gdp)
```

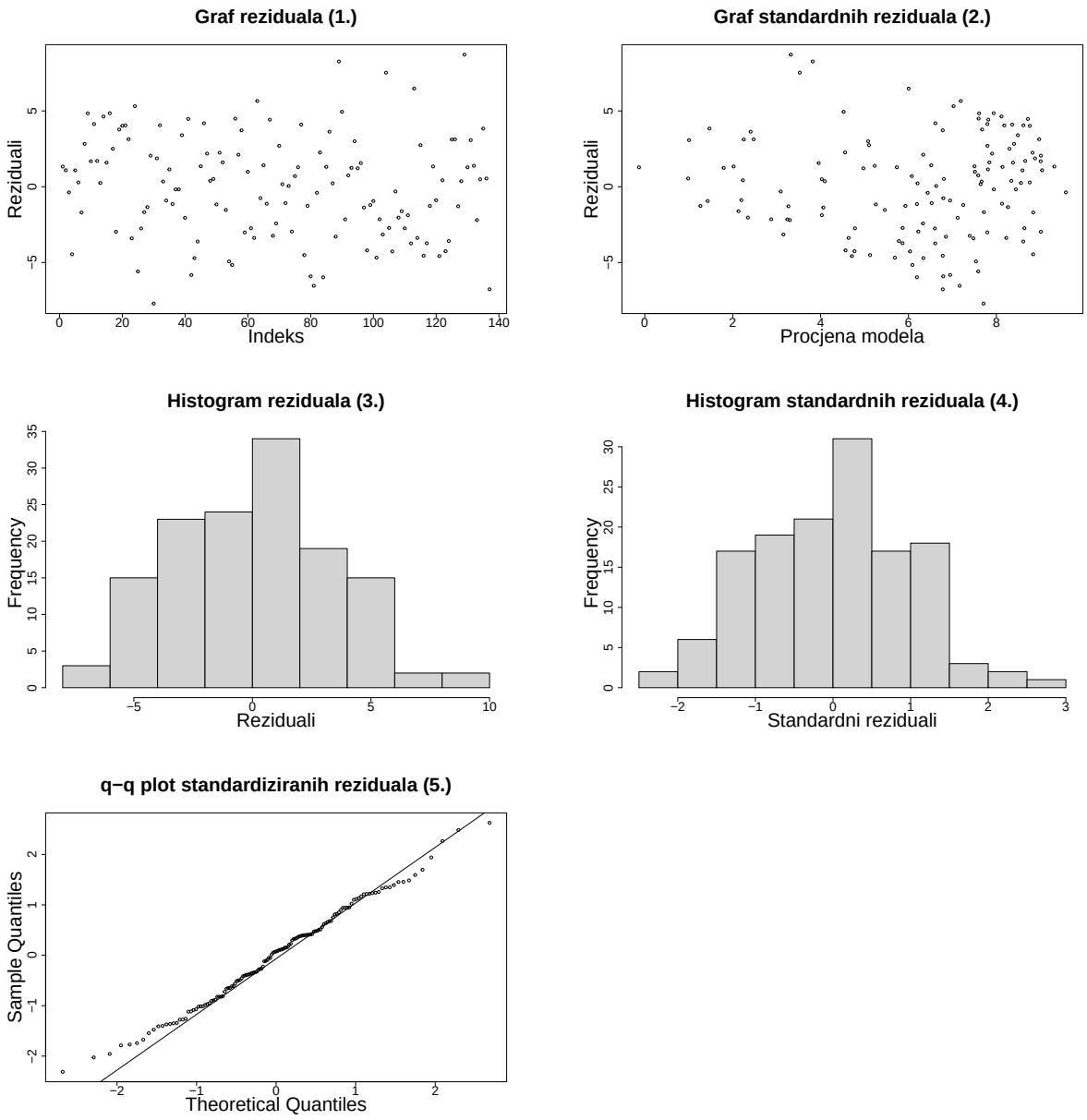
```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  rstandard(selected.model)
## D = 0.041525, p-value = 0.9722
## alternative hypothesis: two-sided
##
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rstandard(selected.model)
## D = 0.042569, p-value = 0.7868
```



1.Reziduali se čine jednoliko raspršeno. 2. Reziduali se čine jednoliko raspršeni.3. Čini se da su reziduali raspodijeljeni simetričnom normalnom distribucijom. 4. Čini se da su reziduali podijeljeni normalnom distribucijom. 5. Oblik qq-plota ukazuje na normalnu distribuciju s debelim repovima. KM TEST i LILLIE TEST ukazuju da distribucija reziduala odgovara normalnoj distribuciji(p-value = 0.9722, p-value = 0.7868)
###Socijalna skrb

```r
normality_homogenity(fit.ss)
```

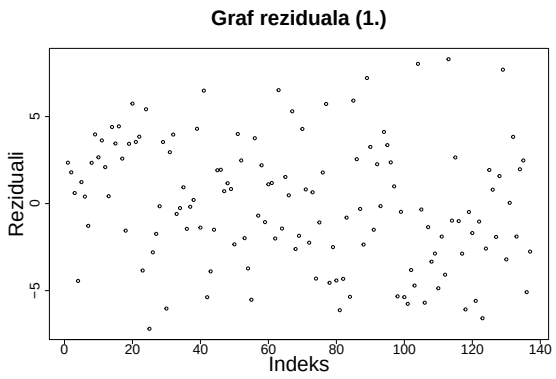
```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  rstandard(selected.model)
## D = 0.047662, p-value = 0.9147
## alternative hypothesis: two-sided
##
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rstandard(selected.model)
## D = 0.047298, p-value = 0.6377
```

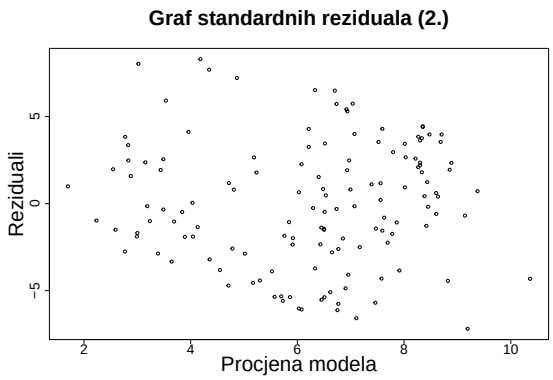


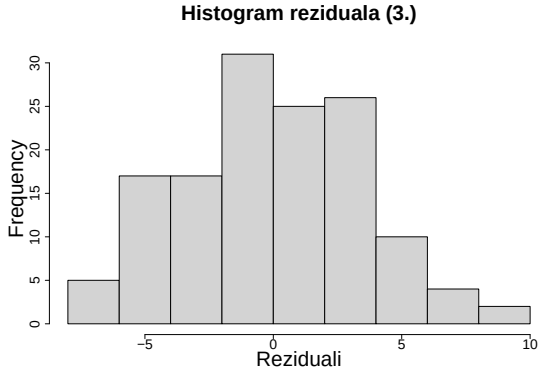
1.Reziduali se čine jednoliko raspršeno. 2. Reziduali se ne čine jednoliko raspršeni.3. Čini se da su reziduali raspodijeljeni simetričnom normalnom distribucijom. 4. Čini se da su reziduali podijeljeni normalnom distribucijom. 5. Oblik qq-plota ukazuje na normalnu distribuciju s debelim repovima. KM TEST i LILLIE TEST ukazuju da distribucija reziduala odgovara normalnoj distribuciji(p-value = 0.9147, p-value = 0.6377)
###Očekivanje zdravog života

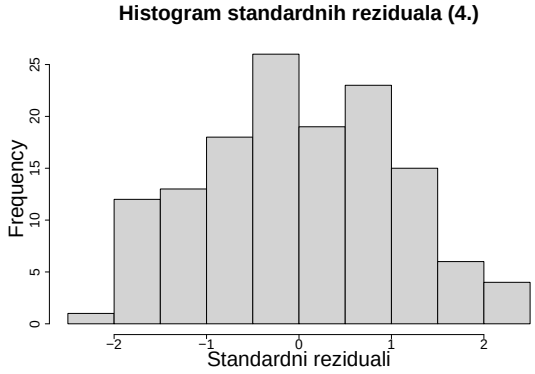
```r
normality_homogenity(fit.hle)
```

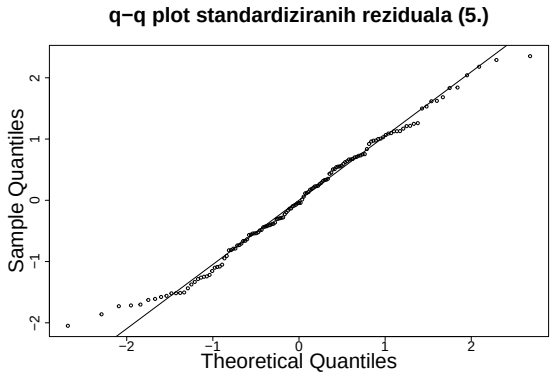
```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  rstandard(selected.model)
## D = 0.043532, p-value = 0.9575
## alternative hypothesis: two-sided
##
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rstandard(selected.model)
## D = 0.042808, p-value = 0.7798
```











1.Reziduali se čine jednoliko raspršeno. 2. Reziduali se ne čine jednoliko raspršeni.3. Čini se da su reziduali raspodijeljeni simetričnom normalnom distribucijom. 4. Čini se da su reziduali podijeljeni dvostranom normalnom distribucijom. 5. Oblik qq-plota ne ukazuje na normalnu distribuciju s debelim repovima. KM TEST i LILLIE TEST ukazuju da distribucija reziduala odgovara normalnoj distribuciji(p-value = 0.9575, p-value = 0.7798) ###Zaključak Svi modeli zadovoljavaju uvjete jednostavne linearne regresije, pa možemo reći da se svaki može uzeti kao model linearne regresije.

3. Postoje li razlike u kvaliteti zdravstvene skrbi medu različitim regijama?

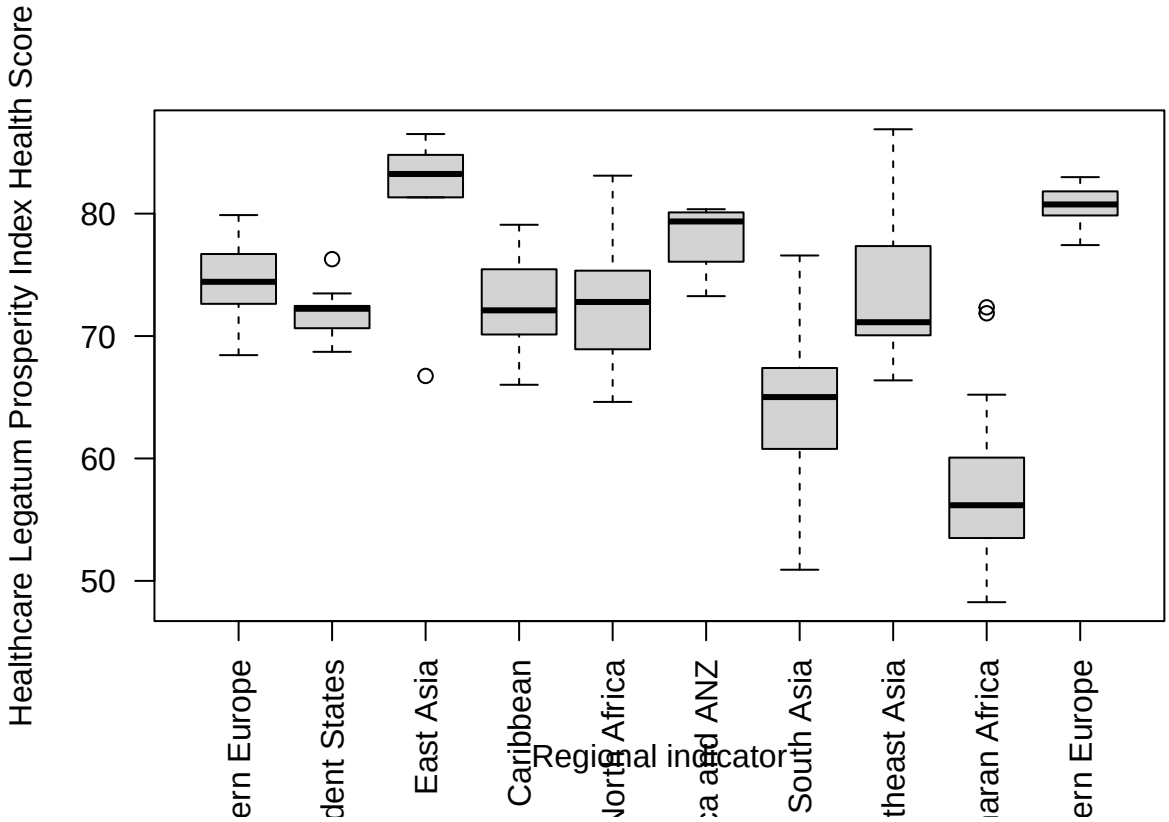
Prvo izoliramo stupce interesa:

```
zdravlje <- data_23[, c('Regional indicator', 'Healthcare Legatum Prosperity Index Health Score')]
```

Želimo utvrditi postoji li statistički značajna razlika u srednim vrijednostima indeksa zdrastvenih skrbi između barem jedne regije, i svih ostalih regija.

Prvo vizualiziramo podatke boxplot grafom:

```
boxplot(`Healthcare Legatum Prosperity Index Health Score` ~ `Regional indicator`, data = zdravlje, las
```



Evidentno kako su razlike u kvaliteti indeksa zdravstva regija prisutne, no potrebno ih je matematički dokazati.

ANOVA testira hipotezu jednakosti više grupa (n>2). Ukoliko je razlika varijance barem jedne grupe i varijance svih grupa stastistički značajna, odbacujemo hipotezu da razlike nema, i zaključujemo da se srednje vrijednosti grupa razlikuju.

Uvijeti za ANOVA analizu normalnost su distribucije unutar zasebnih grupa, i homogenost varijanci među grupama.

Prvo ćemo testirati normalnost podataka unutar grupa:

```
# Brojimo koliko regija je prisutno u skupu podataka
broj_grupa <- zdravlje %>%
  group_by(`Regional indicator`) %>% # Grupiramo po regionalnom indikatoru
  summarize() %>% # Sumariziramo
  count() # Brojimo grupe

nenormalne_grupe <- zdravlje %>%
  group_by(`Regional indicator`) %>%
  summarize( # Vrtimo Shapirom test nad grupiranim vrijednostima svake regije
```

```
    shapiro_test = shapiro.test(`Healthcare Legatum Prosperity Index Health Score`)$p.value
  ) %>%
  filter(shapiro_test < 0.05) %>% # filtriramo p vrijednost testa
  pull(`Regional indicator`) # i izvlačimo ime stupca


print(paste("Ukupni broj regija u podatcima:", broj_grupa))
```

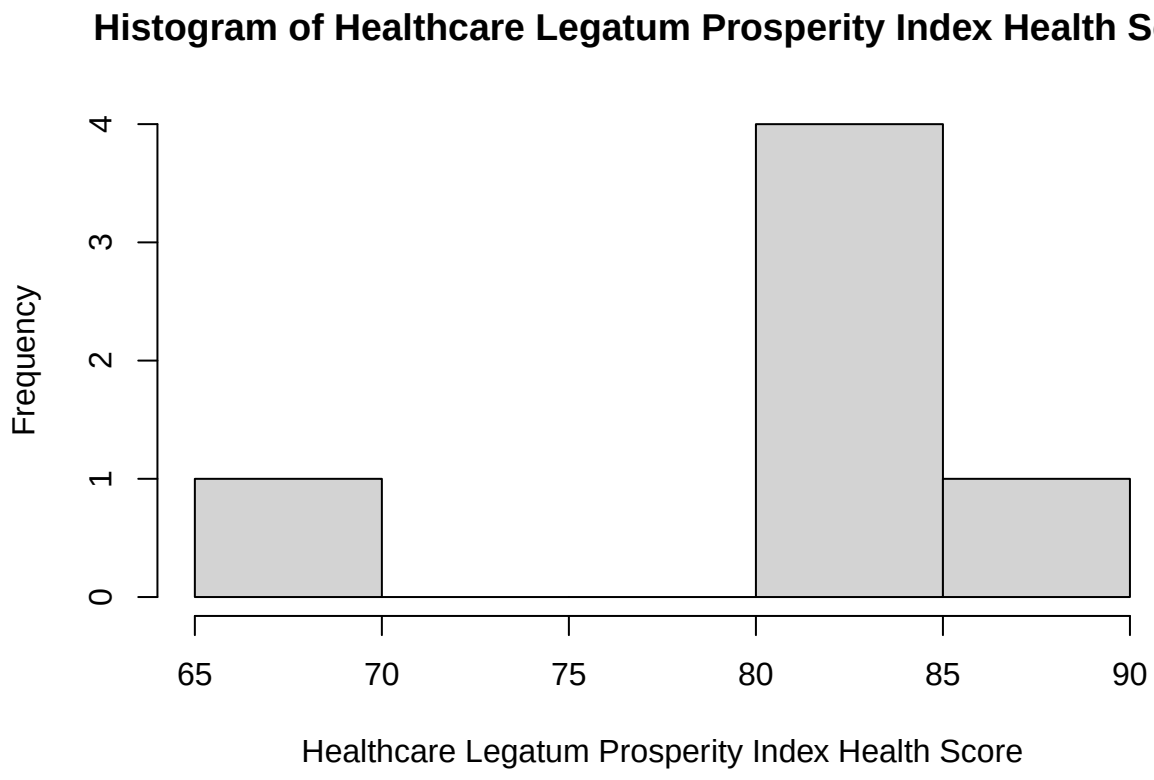
```
## [1] "Ukupni broj regija u podatcima: 10"
```

```
print(paste("Nenormalno distributirane regije:", nenormalne_grupe))
```

```
## [1] "Nenormalno distributirane regije: East Asia"
```

Jedna regija od 10 prisutnih nema normalno raspoređene podatke. ANOVA je otporna na nenormalnost podataka u malim količinama no pretpostavku da to ne utječe na rezultate provjeravamo ispisom histograma indeksa regije Istočne Azije.

```
zdravlje %>%
    filter(`Regional indicator` == 'East Asia') %>%
    with(hist(`Healthcare Legatum Prosperity Index Health Score`)) # With() omogućava hist() koristiti
```



Vidimo da nenormalnost podataka dolazi od jednog unosa. Mogli bi ga ukloniti, no lakše ga je ignorirati. ANOVA neće promijeniti rezultat na njegovoj osnovi.

Slijedeći test je homogenost varijanci među grupama, za koji koristimo Barlettov test.

```r
bartlett.test(`Healthcare Legatum Prosperity Index Health Score` ~ `Regional indicator`, data = zdravlj
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  Healthcare Legatum Prosperity Index Health Score by Regional indicator
## Bartlett's K-squared = 49.922, df = 9, p-value = 1.114e-07
```

Barlettova H0 hipoteza je jednakost varijanci. P vrijednost od 4.758e-08 manja je od 0.05, zbog čega odbacujemo H0 u korist alternativne hipoteze da se varijance razlikuju.

Zaključujemo da bi isti podatci pružali neispravne rezultate kada bi ih proveli kroz ANOVA metodu.

Kao alternativu koristimo Kruskal-Wallis-ov neparametarski test, koji drži iste početne hipoteze i zahtjeva najmanje 5 vrijednosti po grupi. Filtrirati ćemo te regije i provesti podatke kroz test.

Ispitujemo koje regije imaju manje od 5 vrijednosti:

```r
zdravlje_filtered <- zdravlje %>%
  group_by(`Regional indicator`) %>%
  filter(n() < 5) %>%
  summarize()

print(zdravlje_filtered)
```

```
## # A tibble: 1 x 1
##   `Regional indicator`
##   <chr>
## 1 North America and ANZ
```

Te ih izbacujemo iz podataka i vršimo Kruskal-ov test bez njih:

```r
zdravlje_kruskal = zdravlje[c(zdravlje$'Regional indicator' != 'North America and ANZ'), ]

kruskal.test(`Healthcare Legatum Prosperity Index Health Score` ~ `Regional indicator`, data = zdravlje
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Healthcare Legatum Prosperity Index Health Score by Regional indicator
## Kruskal-Wallis chi-squared = 95.976, df = 8, p-value < 2.2e-16
```

Kruskal rezultira p vrijednošću od 2.2e-16, ~ 0, pa odbacujemo hipotezu i zaključujemo da ima razlika između grupa.

Iz pokazanih testova zaključujemo da **se kvaliteta zdrastvene skrbi razlikuje između regija.**