World Happiness Report

X Æ A-Xii

6.12.2021.

Motivacija i opis problema

World Happiness Report je publikacija Mreže rješenja za održivi razvoj Ujedinjenih naroda koja sadrži podatke o osjećaju sreće pojedinih nacija. Podatci su dobiveni kroz ankete koje provode Gallup i Lloyd's Register Foundation. Prvi je izvještaj objavljen 2012. godine, a od 2016. se objavljuje na Međunarodni dan sreće 20. ožujka.

Učitavanje podataka o svjetskom bogatstvu 2021. godine

```
wealth <- read_excel("C:/Users/Sara/Documents/My Fax/SAP/whr-sap/files/credit_suisse_global_wealth_datal
```

```
## New names:
## * ‘‘ -> ...6
## * ‘‘ -> ...7
## * ‘‘ -> ...8
## * ‘‘ -> ...9
```

```
dim(wealth)
```

```
## [1] 169  10
```

```
head(wealth)
```

```
## # A tibble: 6 x 10
##   ‘Country name‘ ‘Adults (thousands)‘ ‘Mean wealth per adu~ ‘Median wealth per ~
##   <chr>                         <dbl>                 <dbl>                <dbl>
## 1 <NA>                             NA                    NA                   NA
## 2 Afghanistan                   18356                  1744                  734
## 3 Albania                        2187                 30524                15363
## 4 Algeria                       27620                  8871                 2302
## 5 Angola                        14339                  3529                 1131
## 6 Argentina                     30799                  7224                 2157
## # ... with 6 more variables:
## #   Distribution of adults (%) by wealth range (USD) <chr>, ...6 <chr>,
## #   ...7 <chr>, ...8 <chr>, ...9 <chr>, Gini (%) <dbl>
```

Učitavanje podataka o globalnoj sreći 2020. godine

You can also embed plots, for example:

```
whr2020 <- read_excel("C:/Users/Sara/Documents/My Fax/SAP/whr-sap/files/WHR_2020.xlsx")

dim(whr2020)
```

```
## [1] 153   9
```

```
head(whr2020)
```

```
## # A tibble: 6 x 9
##   ‘Country name‘ ‘Regional indicator‘ ‘Ladder score‘ ‘Logged GDP per capita‘
##   <chr>          <chr>                         <dbl>                   <dbl>
## 1 Finland        Western Europe                 7.81                    10.6
## 2 Denmark        Western Europe                 7.65                    10.8
## 3 Switzerland    Western Europe                 7.56                    11.0
## 4 Iceland        Western Europe                 7.50                    10.8
## 5 Norway         Western Europe                 7.49                    11.1
## 6 Netherlands    Western Europe                 7.45                    10.8
## # ... with 5 more variables: Social support <dbl>,
## #   Healthy life expectancy <dbl>, Freedom to make life choices <dbl>,
## #   Generosity <dbl>, Perceptions of corruption <dbl>
```

Učitavanje podataka o globalnoj sreći 2021. godine

You can also embed plots, for example:

```
whr2021 <- read_excel("C:/Users/Sara/Documents/My Fax/SAP/whr-sap/files/WHR_2021.xlsx")

dim(whr2021)
```

```
## [1] 149  11
```

```
head(whr2021)
```

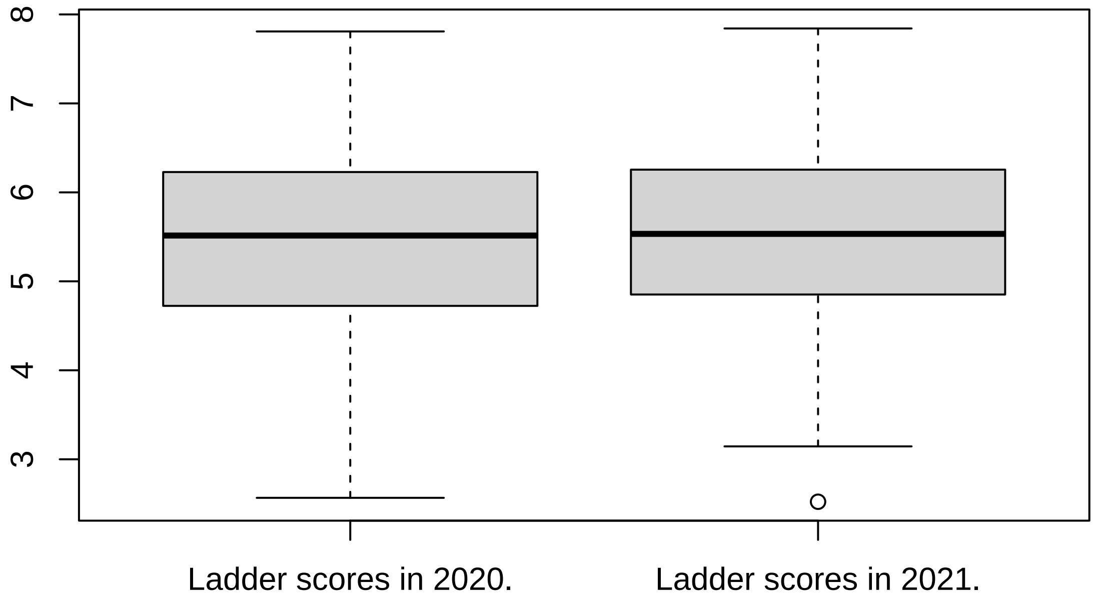
```
## # A tibble: 6 x 11
##   ‘Country name‘ ‘Regional indicator‘ ‘Ladder score‘ ‘Logged GDP per capita‘
##   <chr>          <chr>                         <dbl>                   <dbl>
## 1 Finland        Western Europe                 7.84                    10.8
## 2 Denmark        Western Europe                 7.62                    10.9
## 3 Switzerland    Western Europe                 7.57                    11.1
## 4 Iceland        Western Europe                 7.55                    10.9
## 5 Netherlands    Western Europe                 7.46                    10.9
## 6 Norway         Western Europe                 7.39                    11.1
## # ... with 7 more variables: Social support <dbl>,
## #   Healthy life expectancy <dbl>, Freedom to make life choices <dbl>,
## #   Generosity <dbl>, Perceptions of corruption <dbl>, Income Gini <dbl>,
## #   Wealth Gini <dbl>
```

Je li razina sreće veća u 2020. ili 2021. godini?

Ovo pitanje ćemo provjeravati uparenim t-testom. Podaci koje koristimo su razlike rezultata WHR-a u 2021. i 2020. godini za iste države.

Prvo ćemo napraviti dva boxplota kako bi vizualizirali podatke za pojedinu godinu.

```
boxplot(whr2020$`Ladder score`, whr2021$`Ladder score`, names = c("Ladder scores in 2020.", "Ladder sco
```

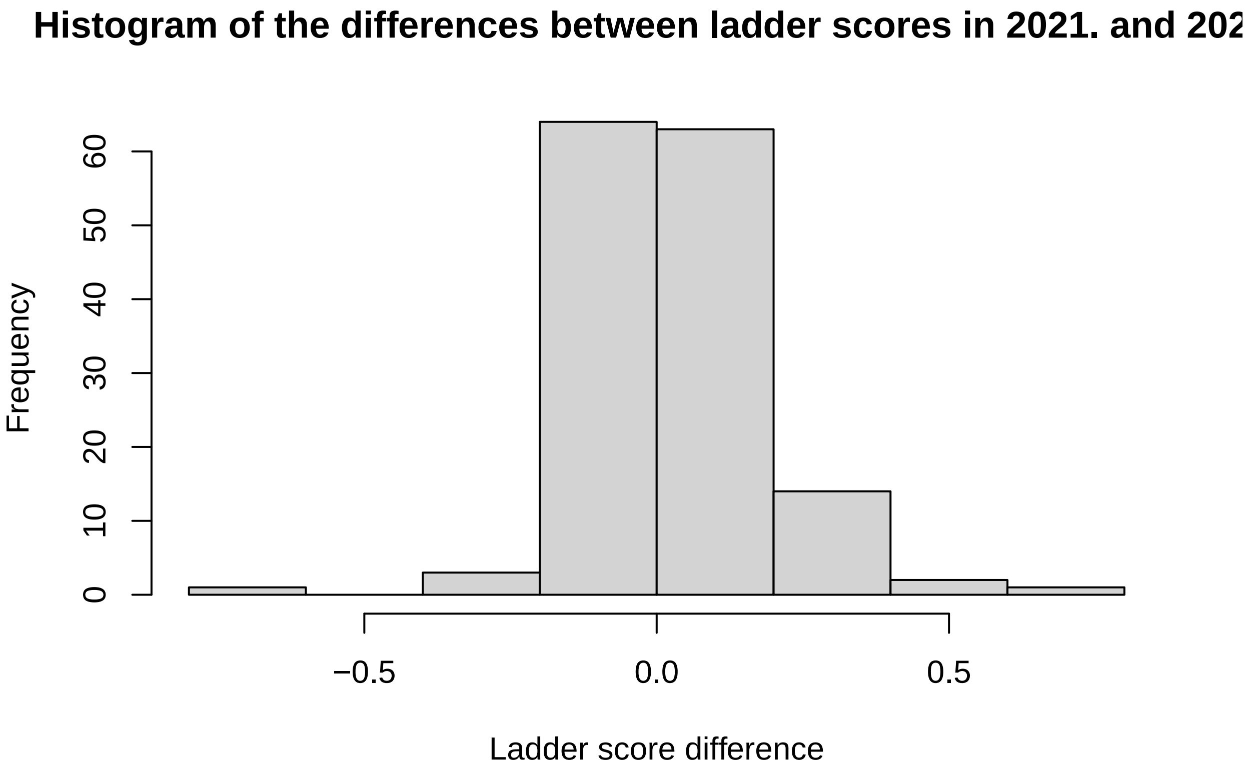


Vidimo da su srednje vrijednosti rezultata za obje godine otprilike jednake.

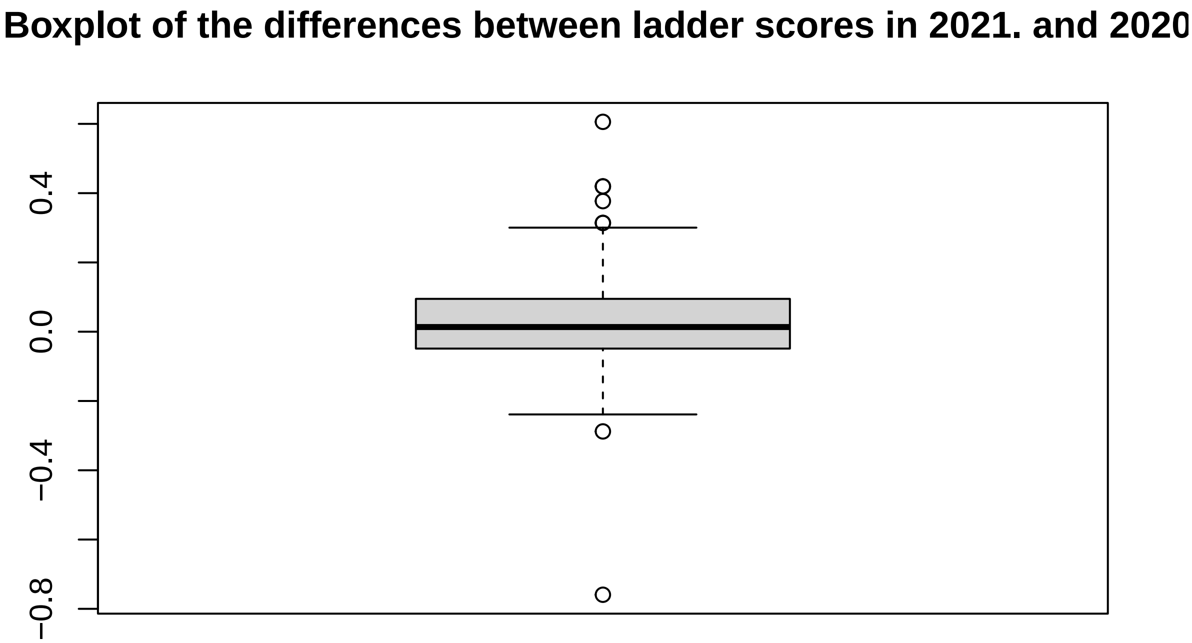
Sada ćemo prikazati razlike rezultata pomoću histograma kako bismo se uvjerili u normalnost podataka, budući da je to uvjet za provođenje uparenog t-testa.

Također ćemo ih prikazati pomoću boxplota, da lakše uočimo stršeće vrijednosti.

```
whr_merged = merge(whr2021, whr2020, by="Country name")
ladderScore_differences = whr_merged$`Ladder score.x`- whr_merged$`Ladder score.y`

hist(ladderScore_differences, xlab="Ladder score difference", main="Histogram of the differences betwee
```



```
boxplot(ladderScore_differences, main="Boxplot of the differences between ladder scores in 2021. and 20
```



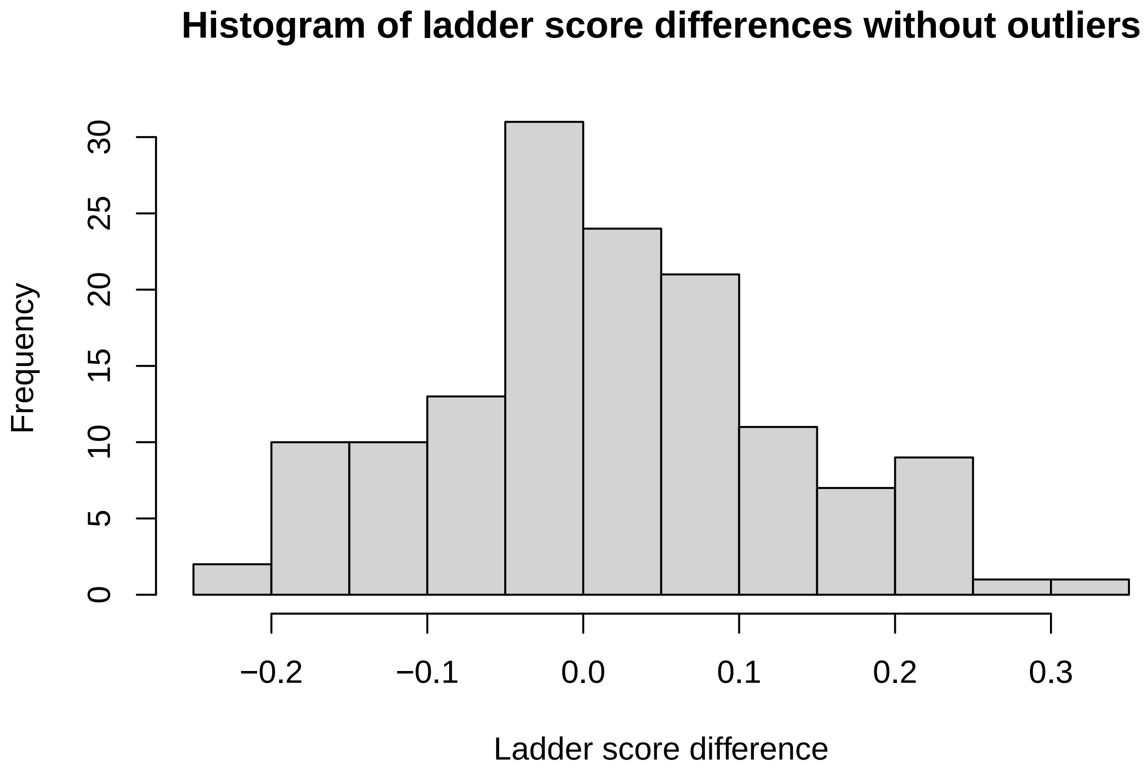
Mogli bismo otprilike reći da podaci jesu normalno distribuirani, no da su ipak više zbijeni oko sredine.

S ciljem povećavanja normalnosti, izbacit ćemo stršeće vrijednosti te ponovno prikazati histogram i boxplot dobivenih podataka.

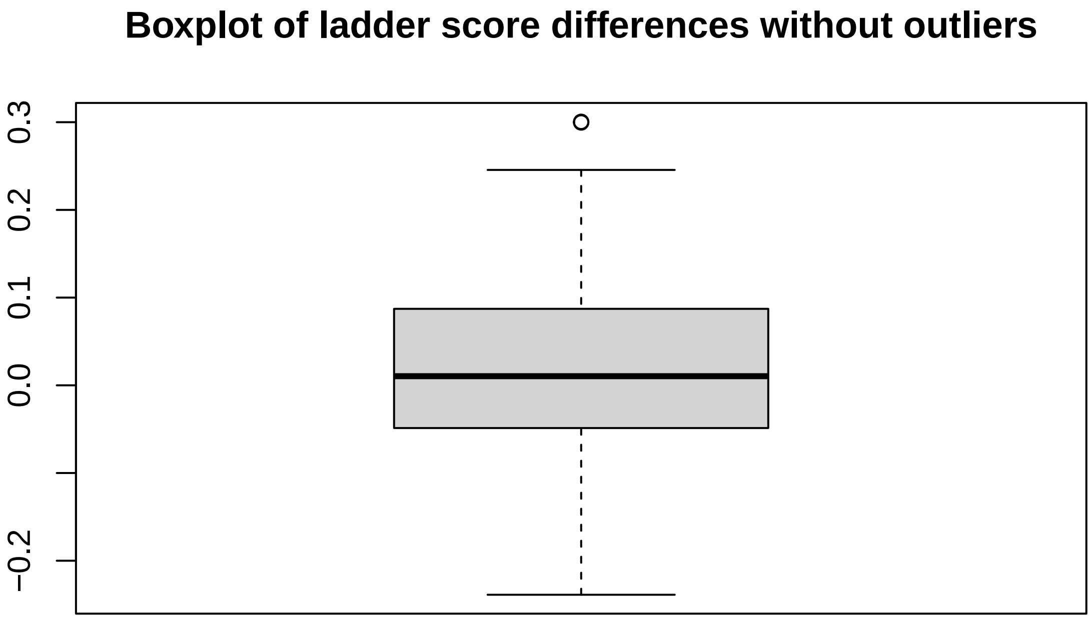
```
ladderScore_differences_no_outliers = ladderScore_differences[!ladderScore_differences %in% boxplot.stat

length(ladderScore_differences) - length(ladderScore_differences_no_outliers)
```

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

```
hist(ladderScore_differences_no_outliers, xlab="Ladder score difference", main="Histogram of ladder scor
```



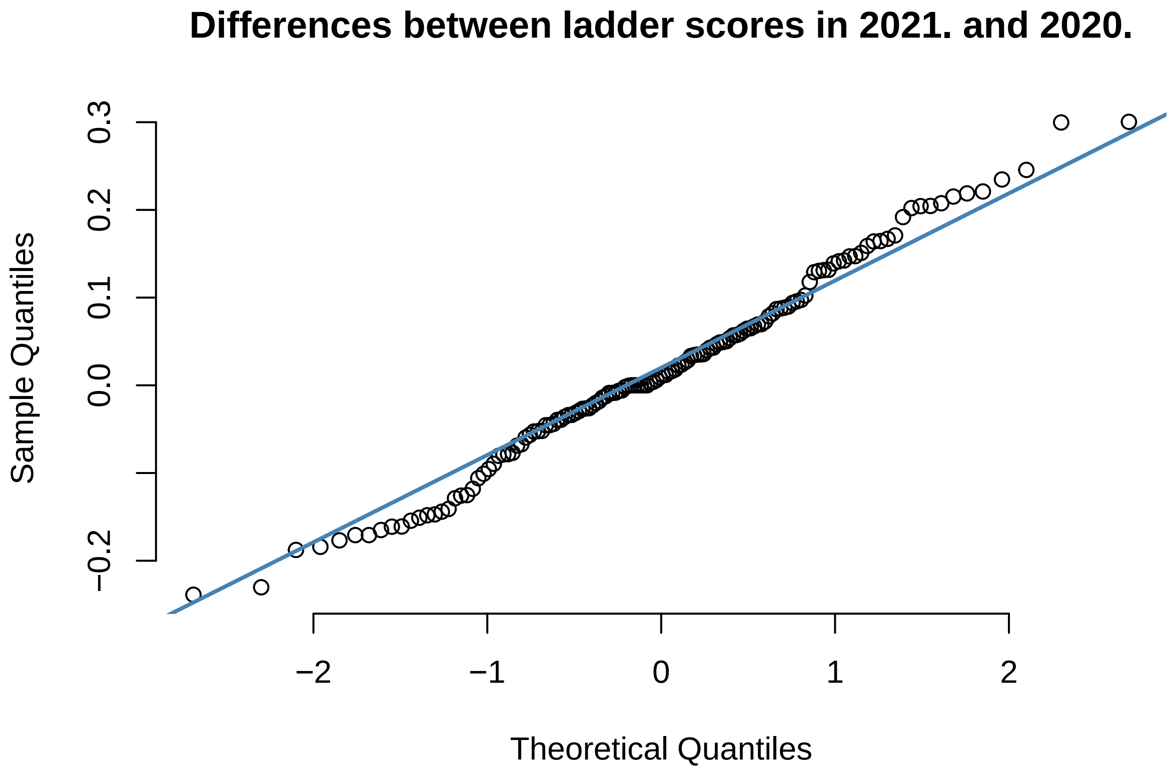
```
boxplot(ladderScore_differences_no_outliers, main="Boxplot of ladder score differences without outliers
```



Uočavamo da su nam sadašnji podaci distribuirani puno više normalno, nego što su prije bili, a izbacili smo samo 8 vrijednosti.

Da bismo se uvjerili u normalnost podataka, možemo je provjeriti i pomoću qq-plota.

```
qqnorm(ladderScore_differences_no_outliers, pch = 1, frame = FALSE, main="Differences between ladder sc
qqline(ladderScore_differences_no_outliers, col = "steelblue", lwd = 2)
```



Kao i histogram, qq-plot nam upućuje na normalnost podataka. Jeini podaci koji se ne ravnaju savršeno po normalnoj distribuciji su oni rubni.

Još da budemo sasvim sigurni u normalnost naših podataka, provest ćemo Kolmogorov-Smirnovljev test. Hipoteze su nam sljedeće:

$$H_0 : \text{podaci su normalno distribuirani}$$
$$H_1 : \text{podaci nisu normalno distribuirani}$$

```
ks.test(ladderScore_differences_no_outliers, "pnorm", mean(ladderScore_differences_no_outliers), sd(lad
```

```
## Warning in ks.test(ladderScore_differences_no_outliers, "pnorm",
## mean(ladderScore_differences_no_outliers), : ties should not be present for the
## Kolmogorov-Smirnov test
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  ladderScore_differences_no_outliers
## D = 0.04168, p-value = 0.9681
## alternative hypothesis: two-sided
```

Budući da je p-vrijednost znatno veća od 0.05, ne odbijamo nul hipotezu o normalnosti podataka te možemo krenuti s obostranim t-testom.

Hipoteze nam glase ovako:

$$H_0 : \mu_{2021} = \mu_{2020}$$
$$H_1 : \mu_{2021} \neq \mu_{2020}$$

```
t.test(whr_merged$`Ladder score.x`, whr_merged$`Ladder score.y`, paired=TRUE, alternative="two.sided",c
```

```
##
##  Paired t-test
##
## data:  whr_merged$‘Ladder score.x‘ and whr_merged$‘Ladder score.y‘
## t = 2.0746, df = 147, p-value = 0.03977
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.001230481 0.050664122
## sample estimates:
## mean of the differences
##               0.0259473
```

Budući da smo dobili p-vrijednost manju od 0.05, odbijamo hipotezu o jednakosti rezultata WHR-a u 2020. i 2021. godini u korist alternativne hipoteze.

Provest ćemo još jedan t-test, no ovaj put jednostrani sa sljedećim hipotezama:

$$H_0 : \mu_{2021} <= \mu_{2020}$$
$$H_1 : \mu_{2021} > \mu_{2020}$$

```
t.test(whr_merged$`Ladder score.x`, whr_merged$`Ladder score.y`, paired=TRUE, alternative="greater",con
```

```
##
##  Paired t-test
##
## data:  whr_merged$‘Ladder score.x‘ and whr_merged$‘Ladder score.y‘
## t = 2.0746, df = 147, p-value = 0.01988
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.005244588         Inf
## sample estimates:
## mean of the differences
##               0.0259473
```

Zbog p-vrijednosti manje od 0.05, odbijamo nul hipotezu i prihvaćemo alternativnu, odnosno da je razina sreće veća u 2021. nego što je bila u 2020. godini.

Možemo li temeljem drugih dostupnih variabli predvidjeti sreću neke nacije?

Koja je od njih najbolji prediktor sreće?

Jedna nacija u našem data setu jednaka jednom retku, te nam je dana varijabla Ladder score, tj. razina sreće, pa imamo podatak o razini sreće svake države. Stoga naslućujemo da je linearna regresija način kako odgovoriti na ova pitanja. Prvo ćemo nadopuniti nedostajuće vrijednosti, a onda se pozabaviti linearnom regresijom.

Nadopunjavanje nedostajućih vrijednosti

S obzirom da u stupcima Income Gini i Wealth Gini nedostaju neke vrijednosti, te ćemo vrijednosti nadopuniti prosječnim vrijednostima tih stupaca pomoću paketa imputeTS.

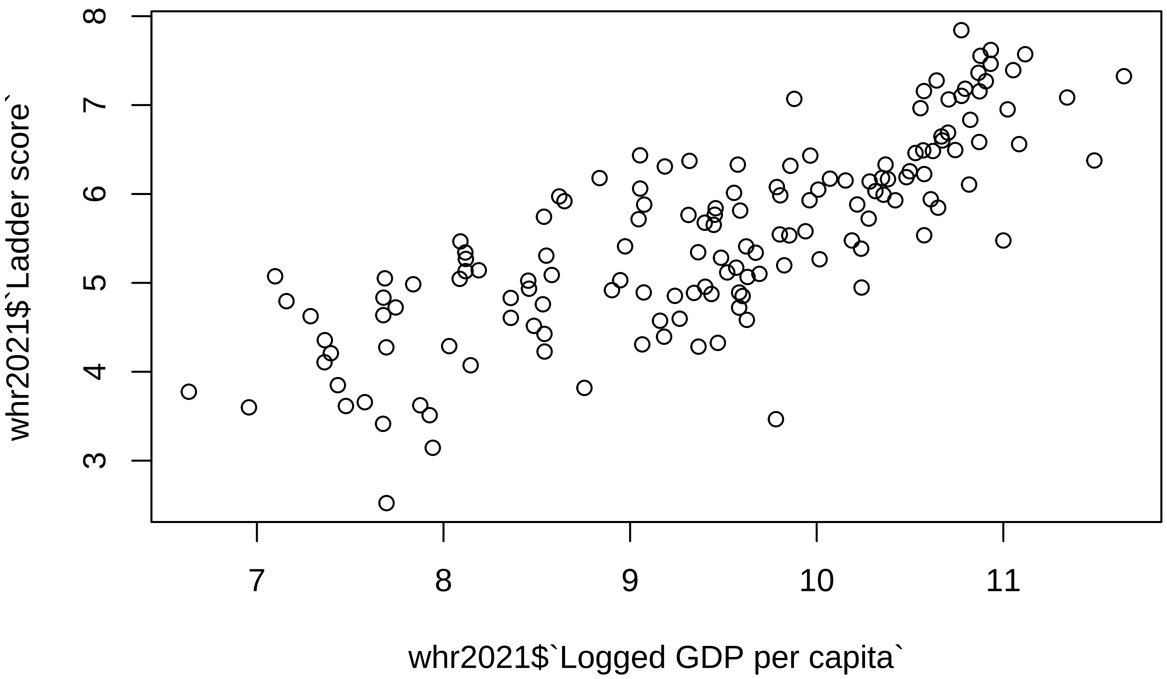
```
whr2021$`Wealth Gini` <- na_mean(whr2021$`Wealth Gini`)
whr2021$`Income Gini` <- na_mean(whr2021$`Income Gini`)
```

Kako bi znali predvidjeti razinu sreće, možemo ispitati različite varijable koje bi mogle utjecati na sreću:

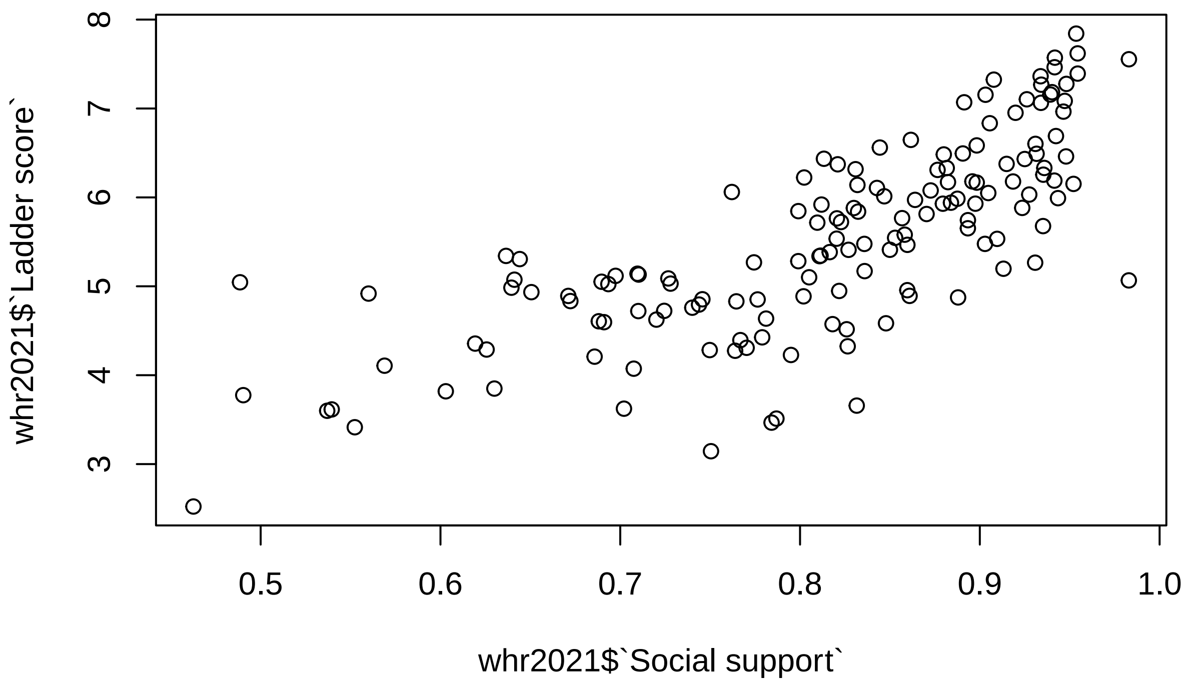
- GDP
- socijalna podrška
- očekivanje trajanja života
- sloboda donošenja odluka
- darežljivost
- percepcija korupcije
- dohodak
- bogatstvo

Kad promatramo utjecaj samo jedne nezavisne varijable X na neku zavisnu varijablu Y, grafički je moguće dobiti jako dobar dojam o njihovom odnosu - tu je najčešće od pomoći scatter plot. Pogledajmo kako izgledaju scatter plot-ovi naših varijabli:

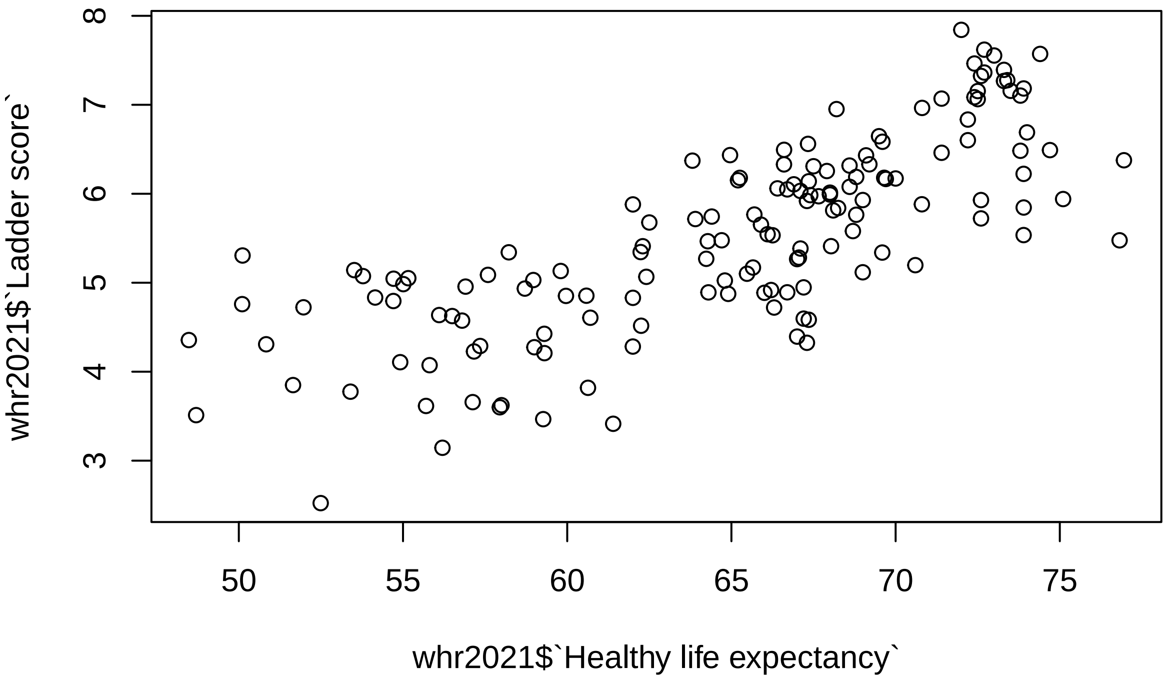
```
plot(whr2021$`Logged GDP per capita`, whr2021$`Ladder score`) #graficki prikaz podataka
```



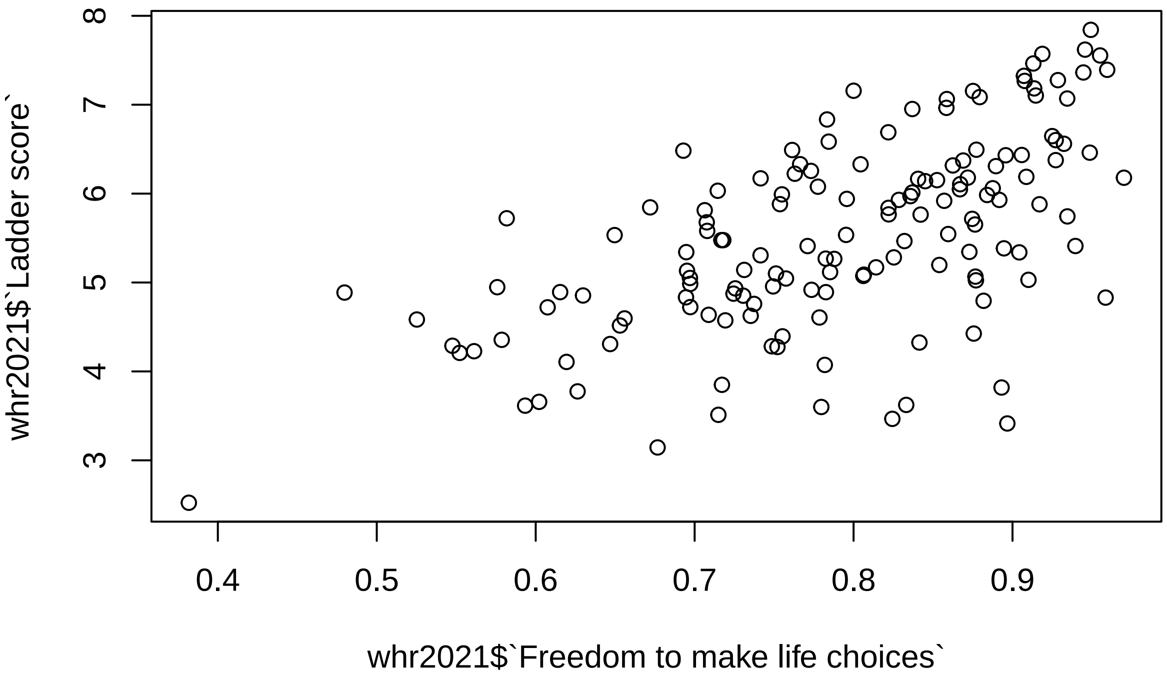
```
plot(whr2021$`Social support`, whr2021$`Ladder score`) #graficki prikaz podataka
```



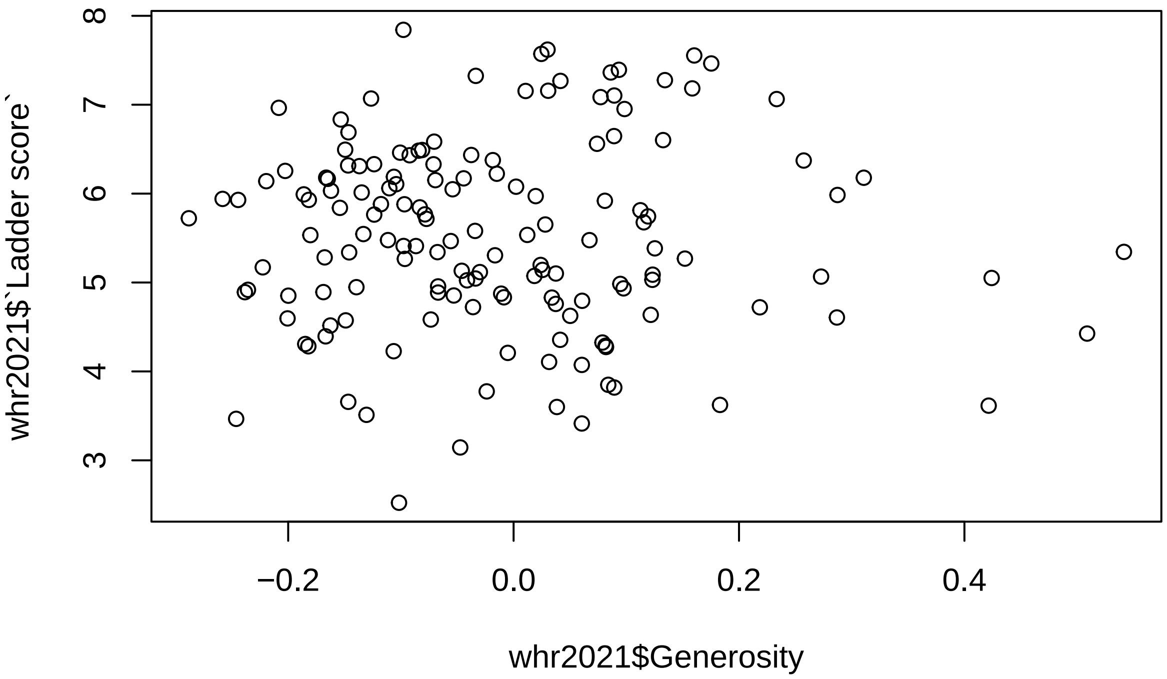
```
plot(whr2021$`Healthy life expectancy`, whr2021$`Ladder score`) #graficki prikaz podataka
```



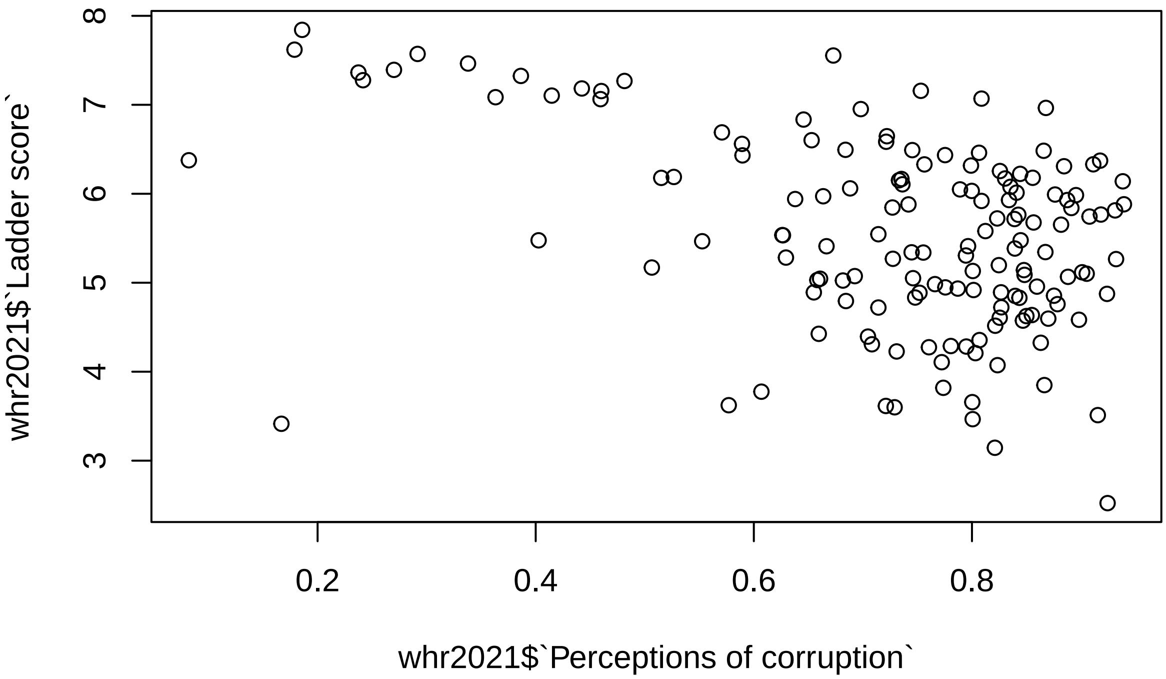
```
plot(whr2021$`Freedom to make life choices`, whr2021$`Ladder score`) #graficki prikaz podataka
```



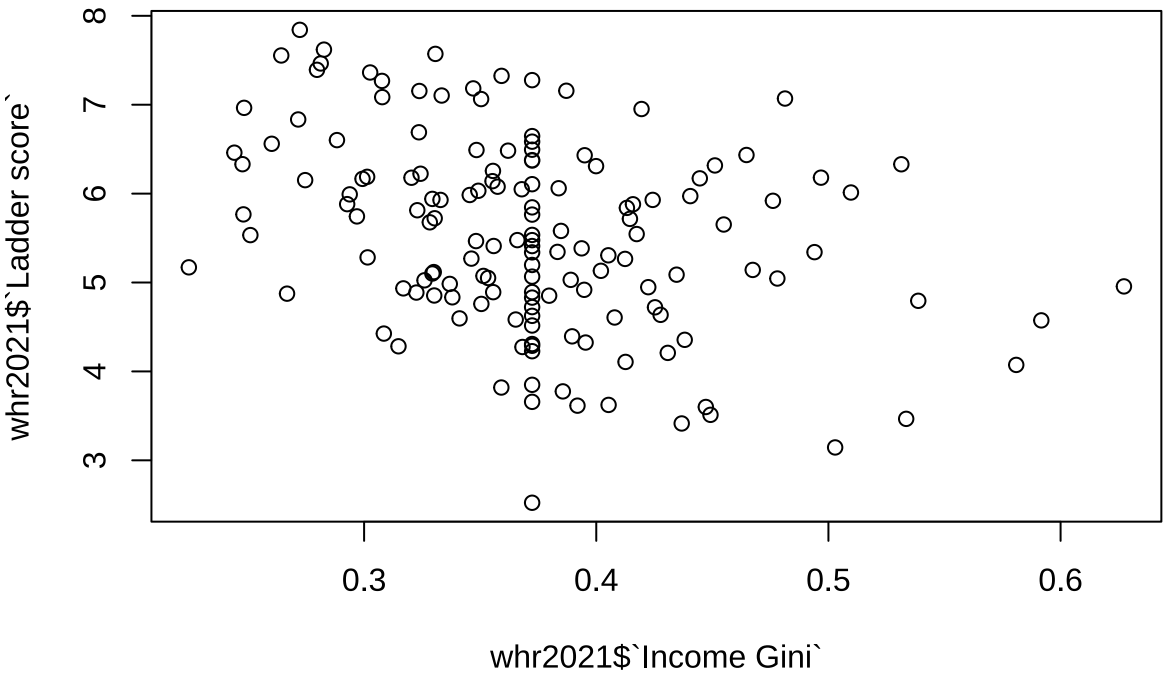
```
plot(whr2021$`Generosity`, whr2021$`Ladder score`) #graficki prikaz podataka
```



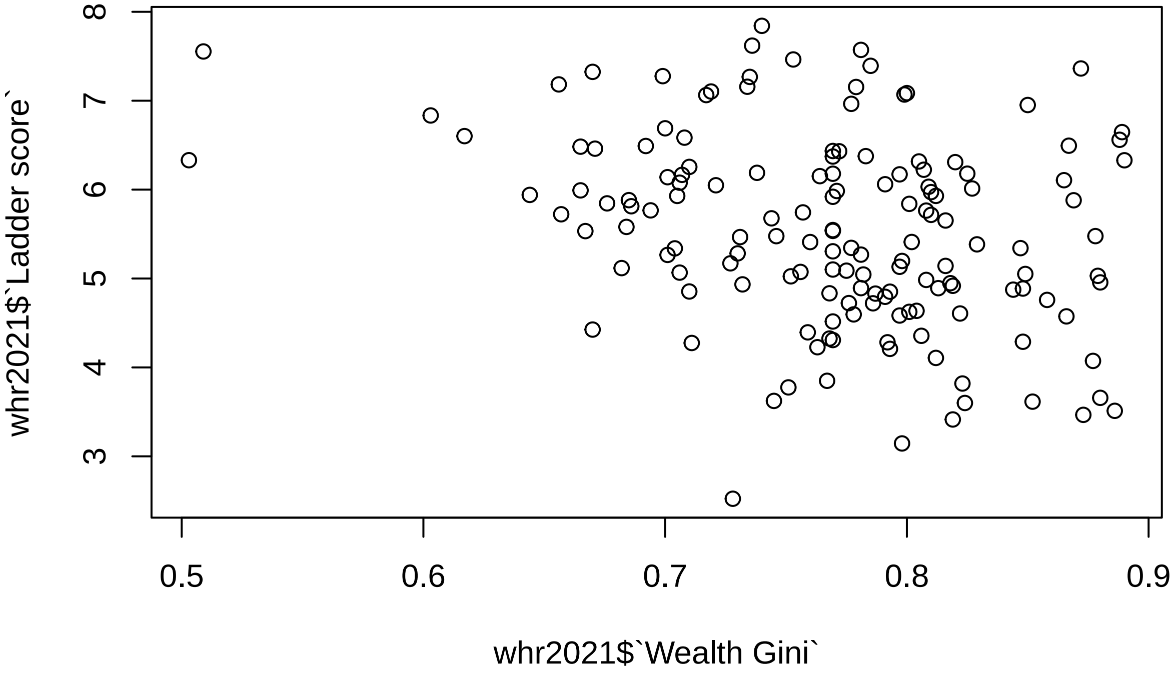
```
plot(whr2021$`Perceptions of corruption`, whr2021$`Ladder score`) #graficki prikaz podataka
```



```
plot(whr2021$`Income Gini`, whr2021$`Ladder score`) #graficki prikaz podataka
```



```
plot(whr2021$`Wealth Gini`, whr2021$`Ladder score`) #graficki prikaz podataka
```



Kao što smo već ranije spomenuli, modelom linearne regresije pokušati ćemo predvidīti sreću neke nacije pa napišimo stoga što smo dosad naučili o linearnoj regresiji.

Linearna regresija

Linearna regresija korisna je u raznim istraživačkim i praktičnim situacijama, a daje odgovore na nekoliko bitnih pitanja, od kojih nas zanimjaju sljedeća:

- Postoji li veza između ulazne varijable (ili više ulaznih varijabli) - regresora, i izlazne varijable (reakcije), unašem slučaju razine sreće?
- Koliko je jaka ta veza?
- Koje ulazne varijable najviše utječu na izlaznu varijablu i koliko je jak taj efekt?

Model linearne regresije i estimacija parametara

Model linearne regresije pretpostavlja linearnu vezu između ulaznih i izlaznih varijabli:

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \epsilon$$

Pretpostavke modela:

- linearnost veze X i Y
- pogreške nezavisne, homogene i normalno distribuirane s $\epsilon \sim \mathcal{N}(0, \sigma^2)$

Iz podataka je moguće dobiti procjenu modela:

$$\hat{Y} = b_0 + \sum_{j=1}^{p} b_j x_j + e,$$

odnosno:

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

u matričnom zapisu.

Procjena je zasnovana na metodi najmanjih kvadrata, tj. minimizaciji tzv. "sum of squared errors":

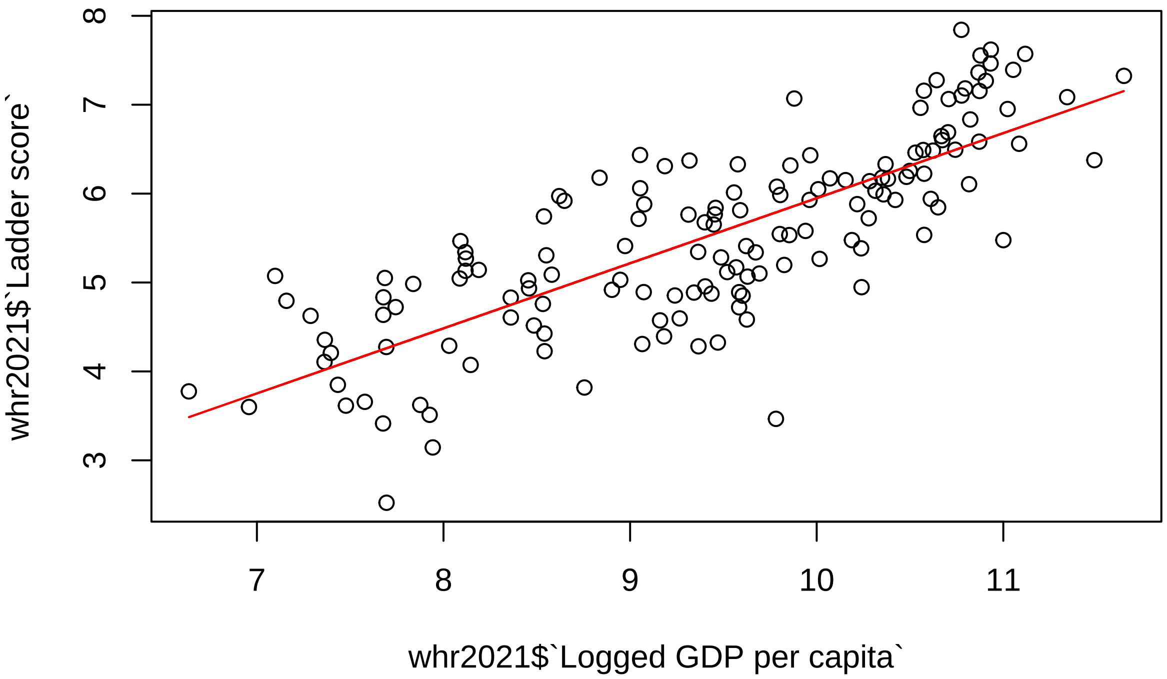
$$SSE = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = (\mathbf{y} - \mathbf{X}\mathbf{b})^T (\mathbf{y} - \mathbf{X}\mathbf{b})$$

Pogledajmo onda sada za koje od ponuđenih varijabli očekujemo da bi mogle biti dobar prediktor razine sreće.

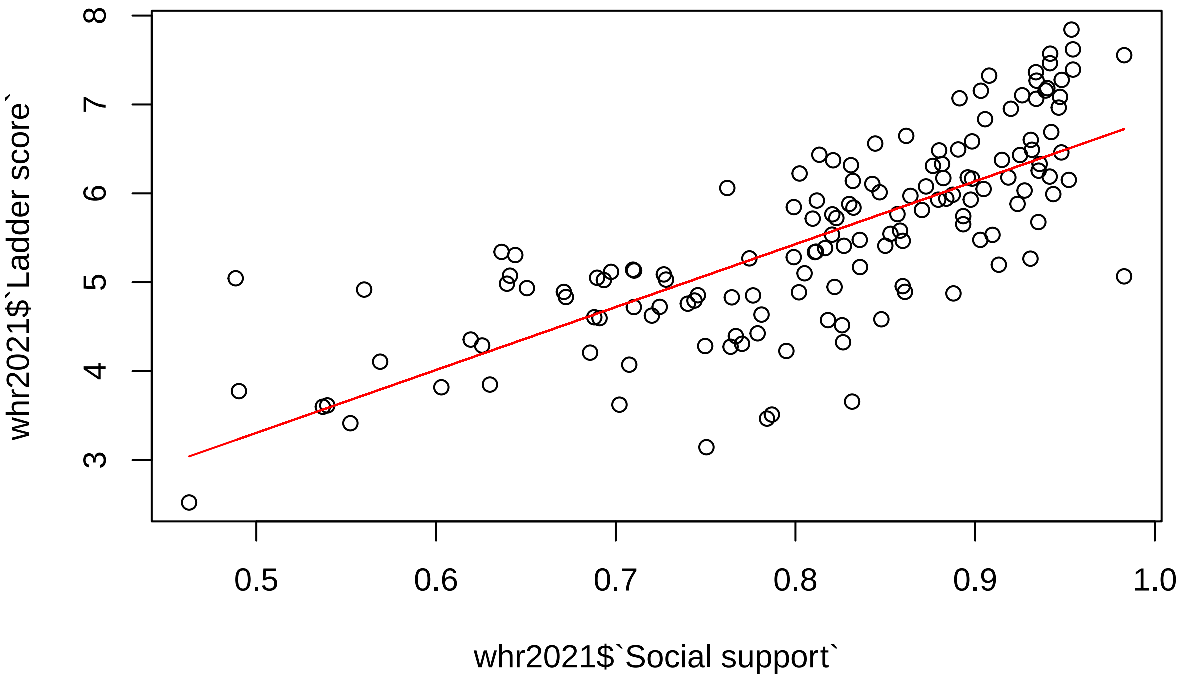
Iz grafova se vidi da GDP, socijalna podrška, očekivanje trajanja života i sloboda donošenja odluka imaju utjecaja na razinu sreće. Darežljivost, percepcija korupcije, dohodak i bogatstvo su slabiji kandidati za modeliranje razine sreće.

Pomoću modela jednostavne regresije, ispitati ćemo pojedinačni utjecaj svake varijable.

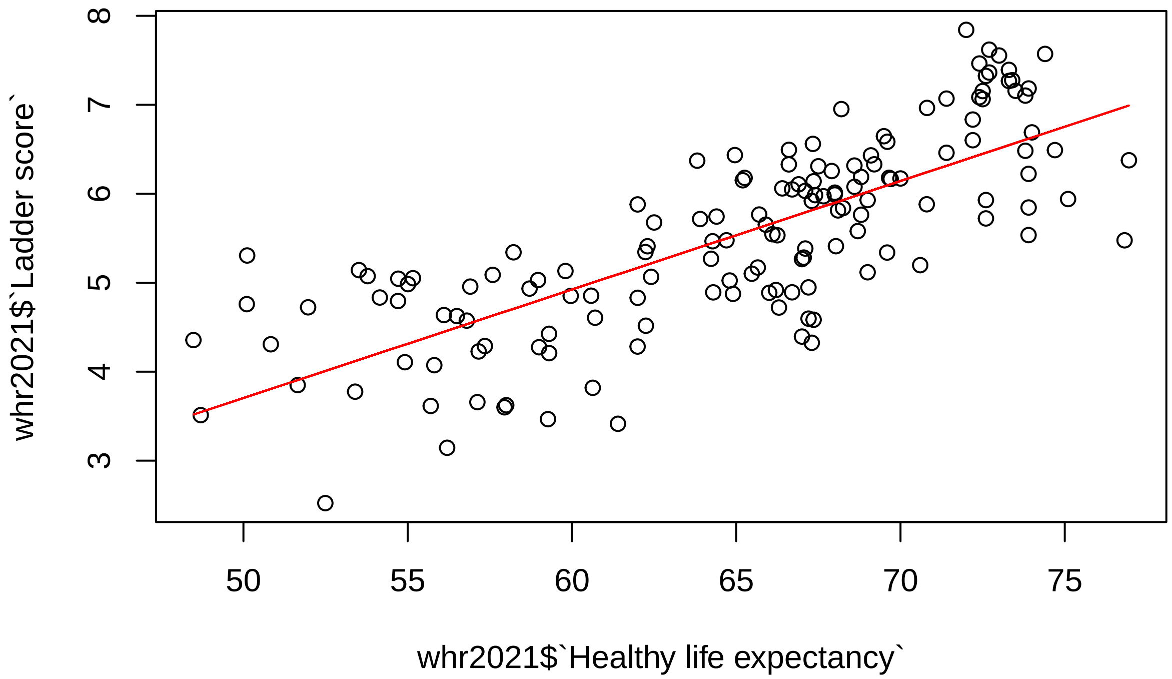
```
fit.gdp = lm(formula = whr2021$`Ladder score`~whr2021$`Logged GDP per capita`,data=whr2021) #linearni m
fit.ssupp = lm(formula = whr2021$`Ladder score`~whr2021$`Social support`,data=whr2021) #linearni model
fit.hle = lm(formula = whr2021$`Ladder score`~whr2021$`Healthy life expectancy`,data=whr2021) #linearni
fit.frd = lm(formula = whr2021$`Ladder score`~whr2021$`Freedom to make life choices`,data=whr2021) #lin
fit.gen = lm(formula = whr2021$`Ladder score`~whr2021$`Generosity`,data=whr2021) #linearni model ocjene
fit.corr = lm(formula = whr2021$`Ladder score`~whr2021$`Perceptions of corruption`,data=whr2021) #linea
fit.inc = lm(formula = whr2021$`Ladder score`~whr2021$`Income Gini`,data=whr2021) #linearni model ocjen
fit.wlth = lm(formula = whr2021$`Ladder score`~whr2021$`Wealth Gini`,data=whr2021) #linearni model ocje

plot(whr2021$`Logged GDP per capita`, whr2021$`Ladder score`) #graficki prikaz podataka
lines(whr2021$`Logged GDP per capita`,fit.gdp$fitted.values,col='red') #graficki prikaz procijenjenih v
```



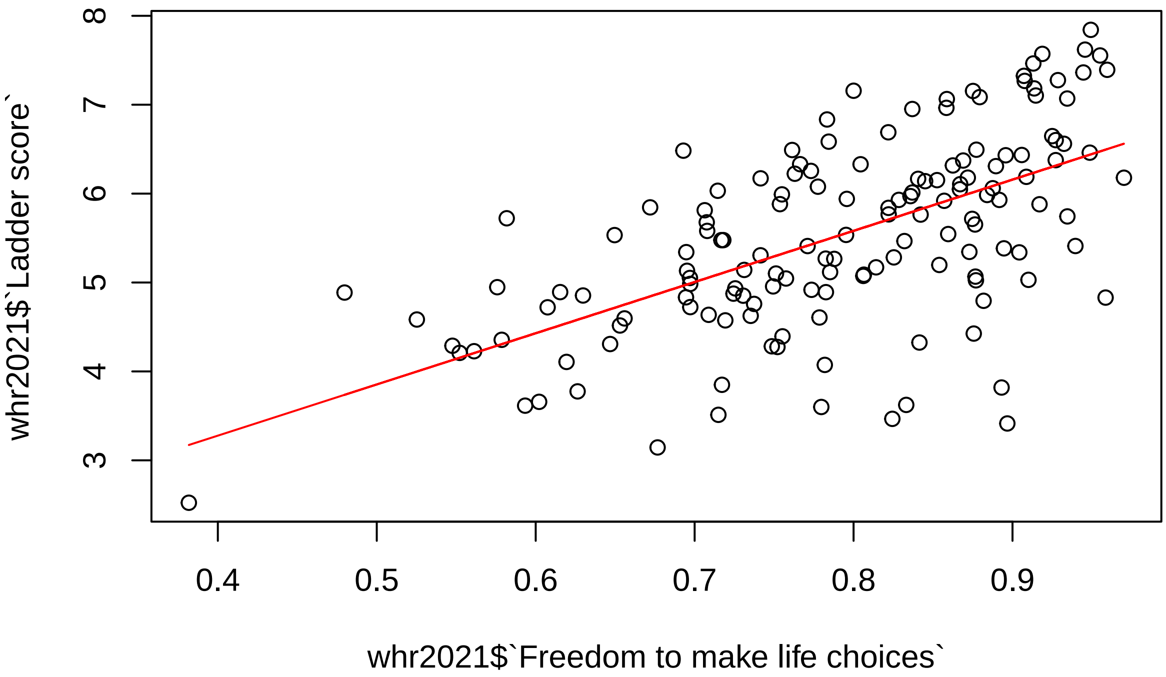
```
plot(whr2021$`Social support`, whr2021$`Ladder score`) #graficki prikaz podataka
lines(whr2021$`Social support`,fit.ssupp$fitted.values,col='red') #graficki prikaz procijenjenih vrijed
```



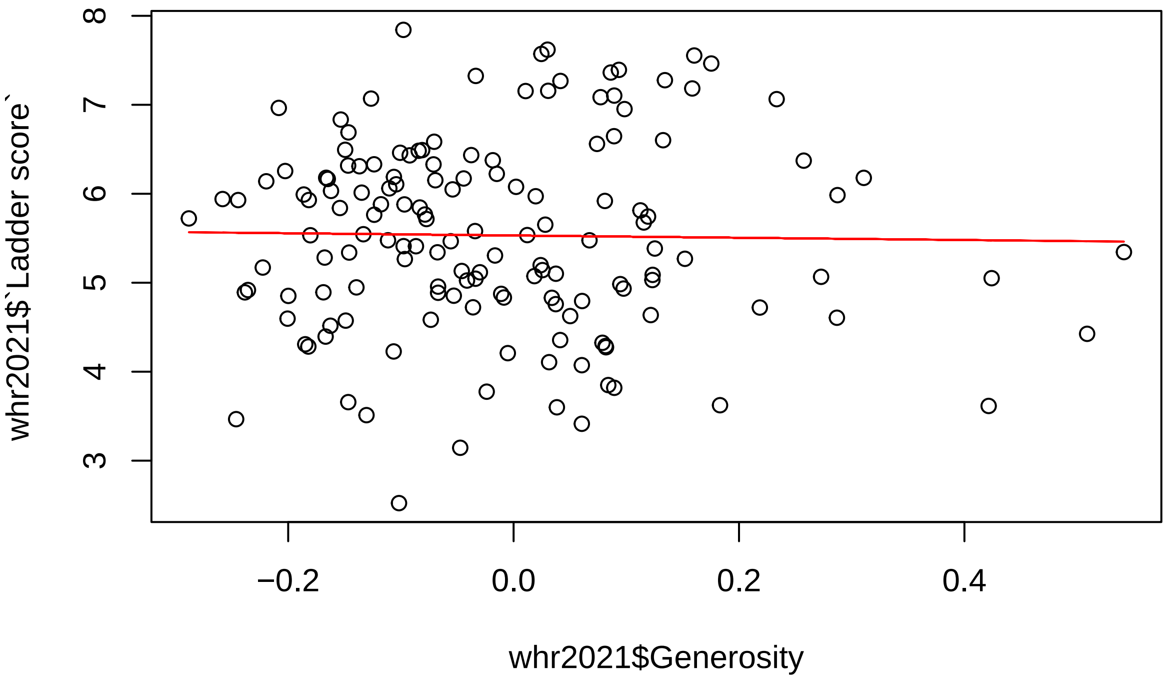
```
plot(whr2021$`Healthy life expectancy`, whr2021$`Ladder score`) #graficki prikaz podataka
lines(whr2021$`Healthy life expectancy`,fit.hle$fitted.values,col='red') #graficki prikaz procijenjenih
```



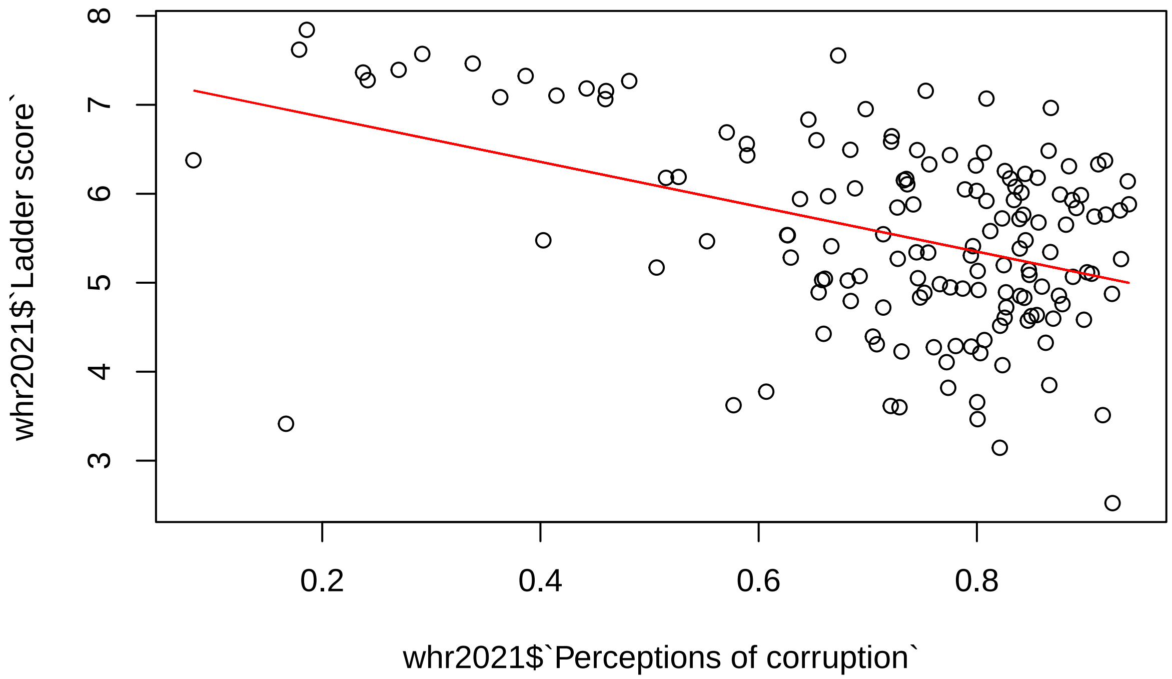
```
plot(whr2021$`Freedom to make life choices`, whr2021$`Ladder score`) #graficki prikaz podataka
lines(whr2021$`Freedom to make life choices`,fit.frd$fitted.values,col='red') #graficki prikaz procijen
```



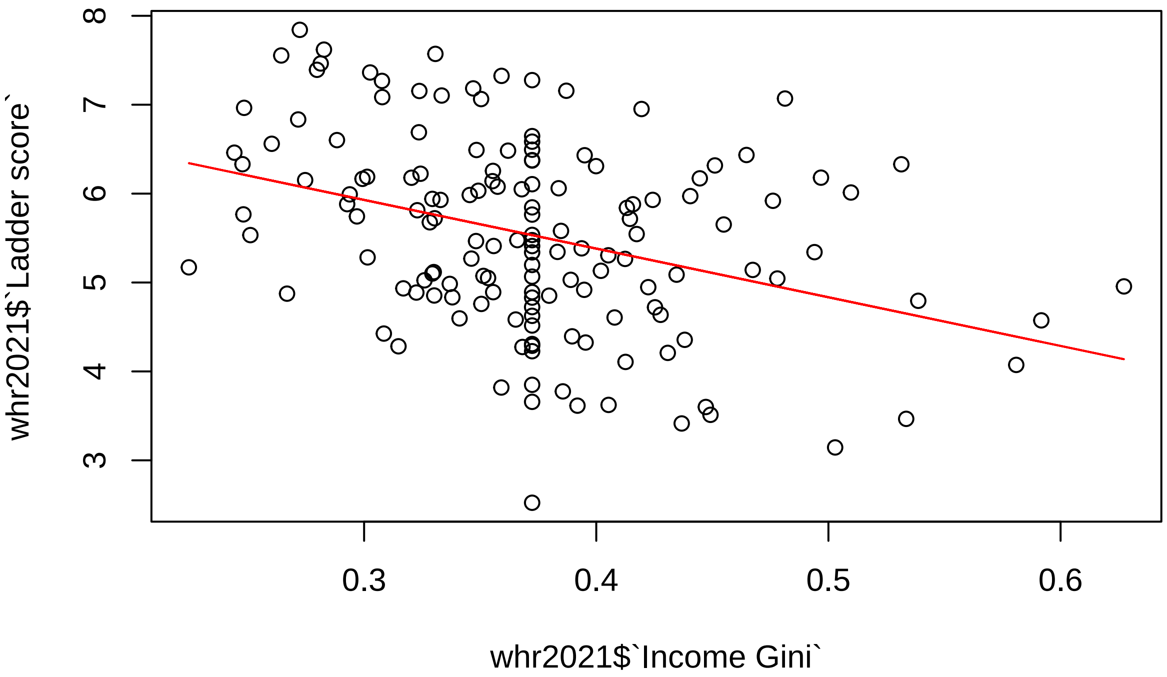
```
plot(whr2021$`Generosity`, whr2021$`Ladder score`) #graficki prikaz podataka
lines(whr2021$`Generosity`,fit.gen$fitted.values,col='red') #graficki prikaz procijenjenih vrijednosti
```



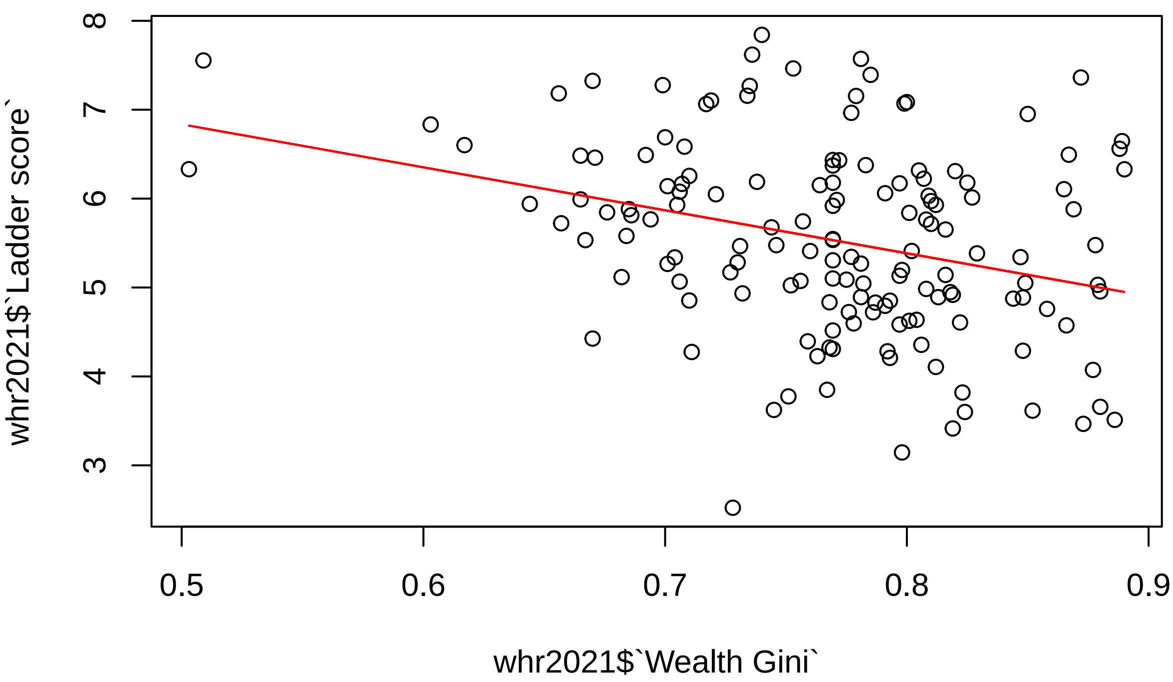
```
plot(whr2021$`Perceptions of corruption`, whr2021$`Ladder score`) #graficki prikaz podataka
lines(whr2021$`Perceptions of corruption`,fit.corr$fitted.values,col='red') #graficki prikaz procijenje
```



```
plot(whr2021$`Income Gini`, whr2021$`Ladder score`) #graficki prikaz podataka
lines(whr2021$`Income Gini`,fit.inc$fitted.values,col='red') #graficki prikaz procijenjenih vrijednosti
```



```
plot(whr2021$`Wealth Gini`, whr2021$`Ladder score`) #graficki prikaz podataka
lines(whr2021$`Wealth Gini`,fit.wlth$fitted.values,col='red') #graficki prikaz procijenjenih vrijednost
```

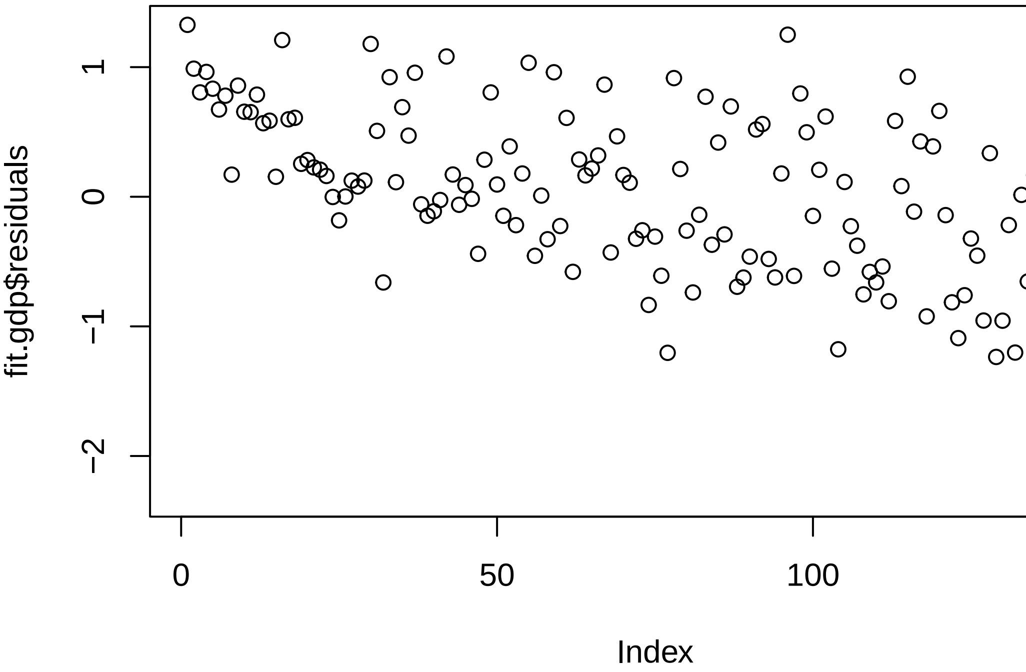


Nagibi pravaca linearne regresije nam pokazuju utjecaj pojedinih varijabli na izlaznu varijablu. Kako bismo usporedili i analizirali dobivene modele, provjeriti ćemo da pretpostavke modela nisu narušene. Pritom su najbitnije pretpostavke o regresorima (u multivarijatnoj regresiji regresori ne smiju biti međusobno jako korelirani) i o rezidualima (normalnost reziduala i homogenost varijance).

Provjera normalnosti

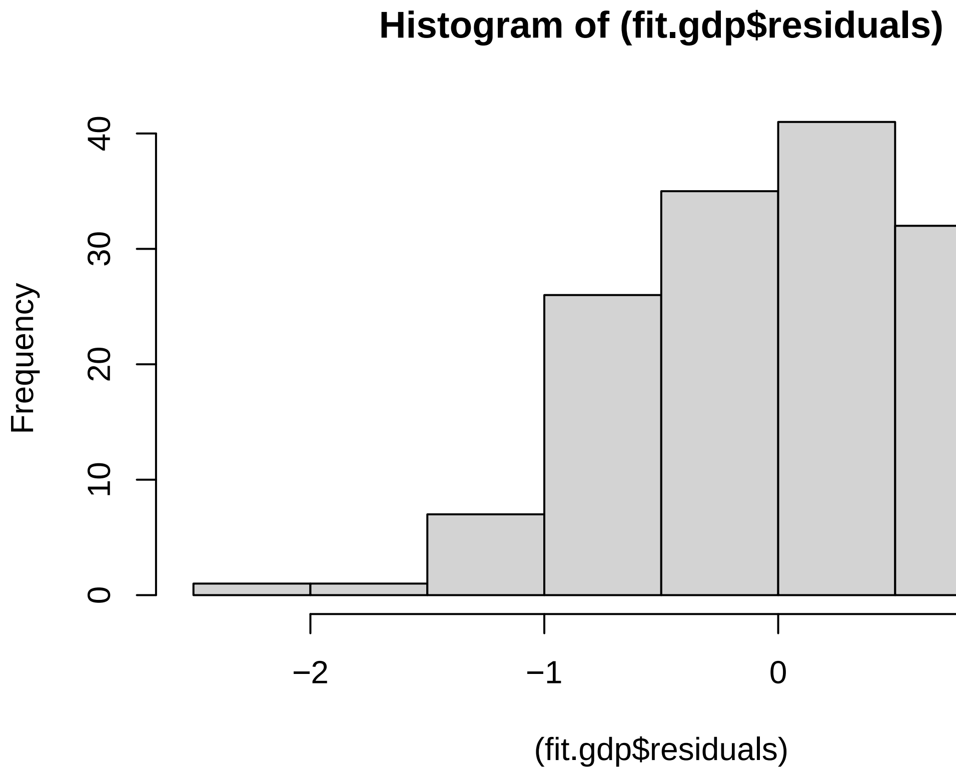
Normalnost reziduala provjeriti ćemo grafički pomoću histograma i kvantil-kvantil plota te statistički s Kolmogorov-Smirovljevim testom i Lillieforsovom korekcijom.

```
plot(fit.gdp$residuals) #gledajuci reziduale na ovaj nacin tesko je suditi o normalnosti
```

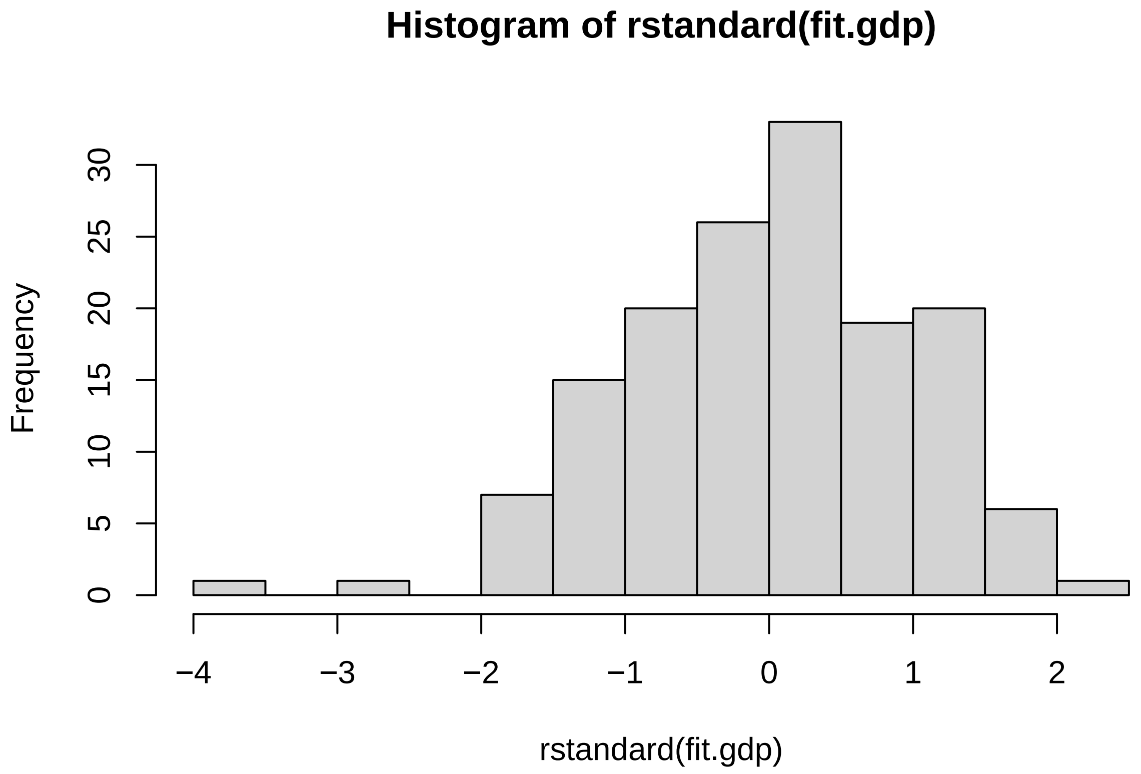


Provjera normalnosti GDP-a

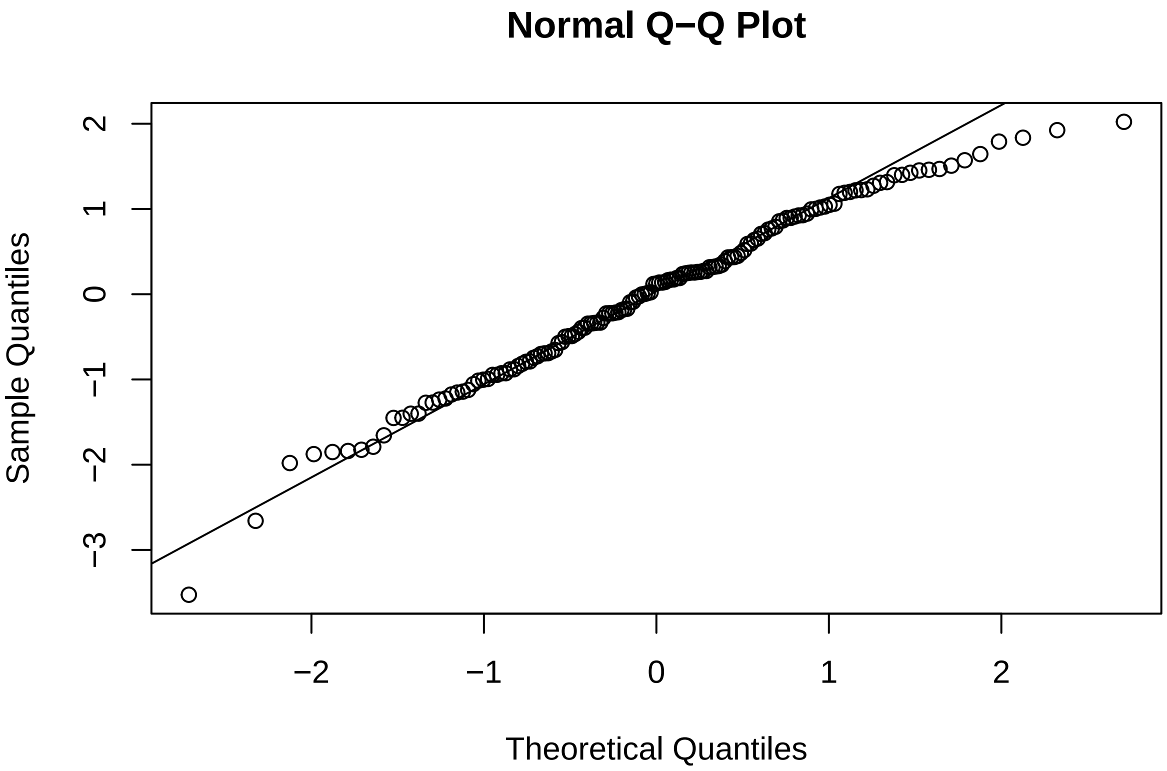
```
hist((fit.gdp$residuals))
```



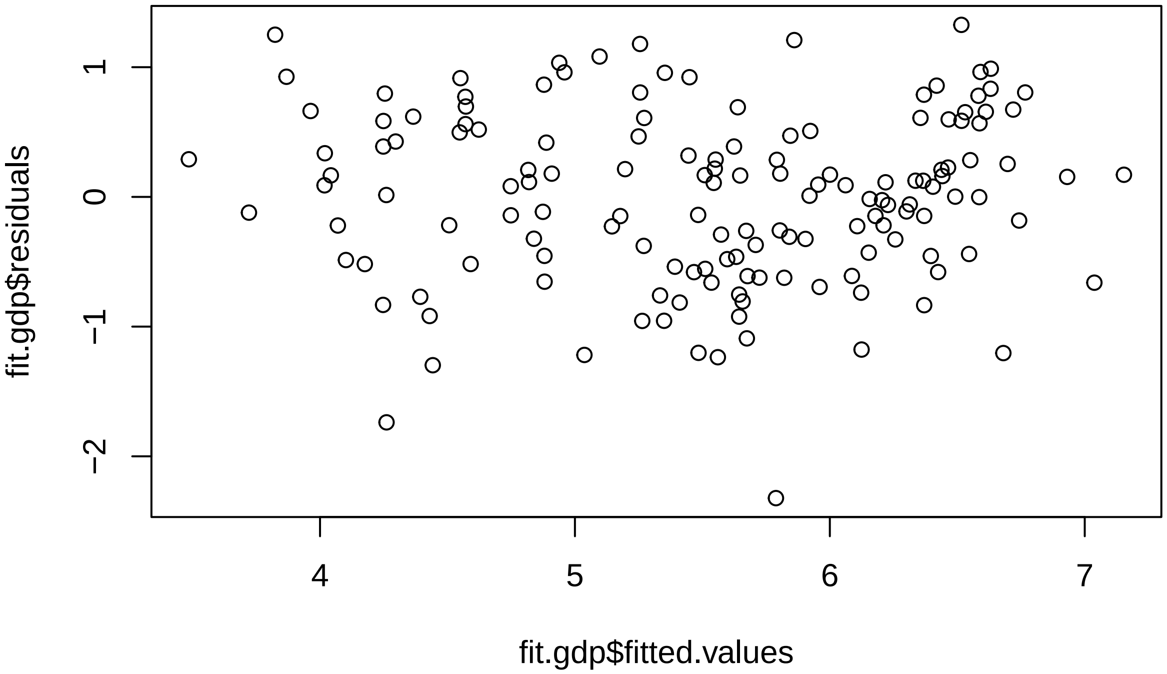
```
hist(rstandard(fit.gdp))
```



```
qqnorm(rstandard(fit.gdp))
qqline(rstandard(fit.gdp))
```



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



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

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  rstandard(fit.gdp)
## D = 0.057746, p-value = 0.7031
## alternative hypothesis: two-sided
```

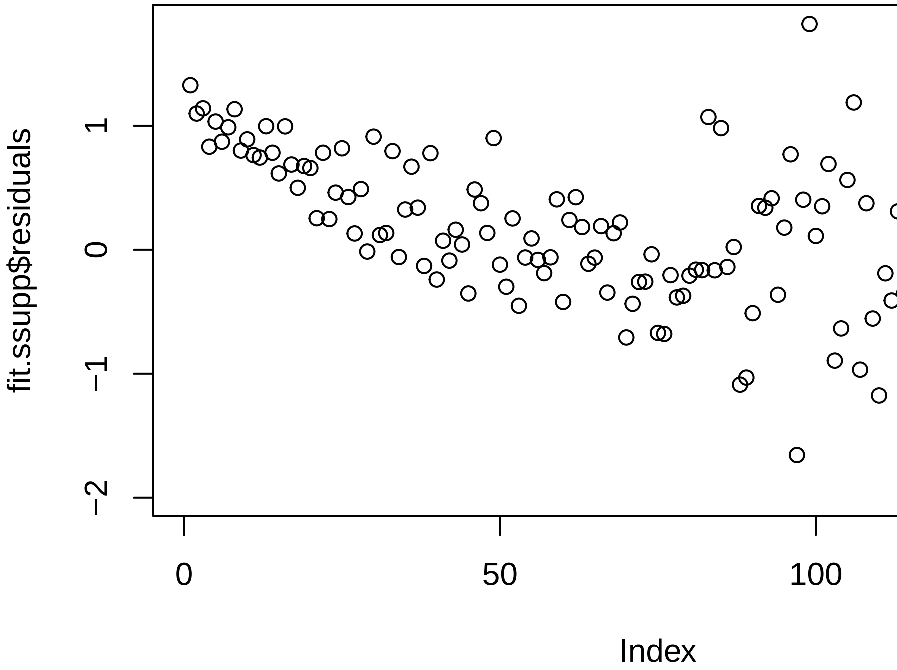
```
require(nortest)
```

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

```
lillie.test(rstandard(fit.gdp))
```

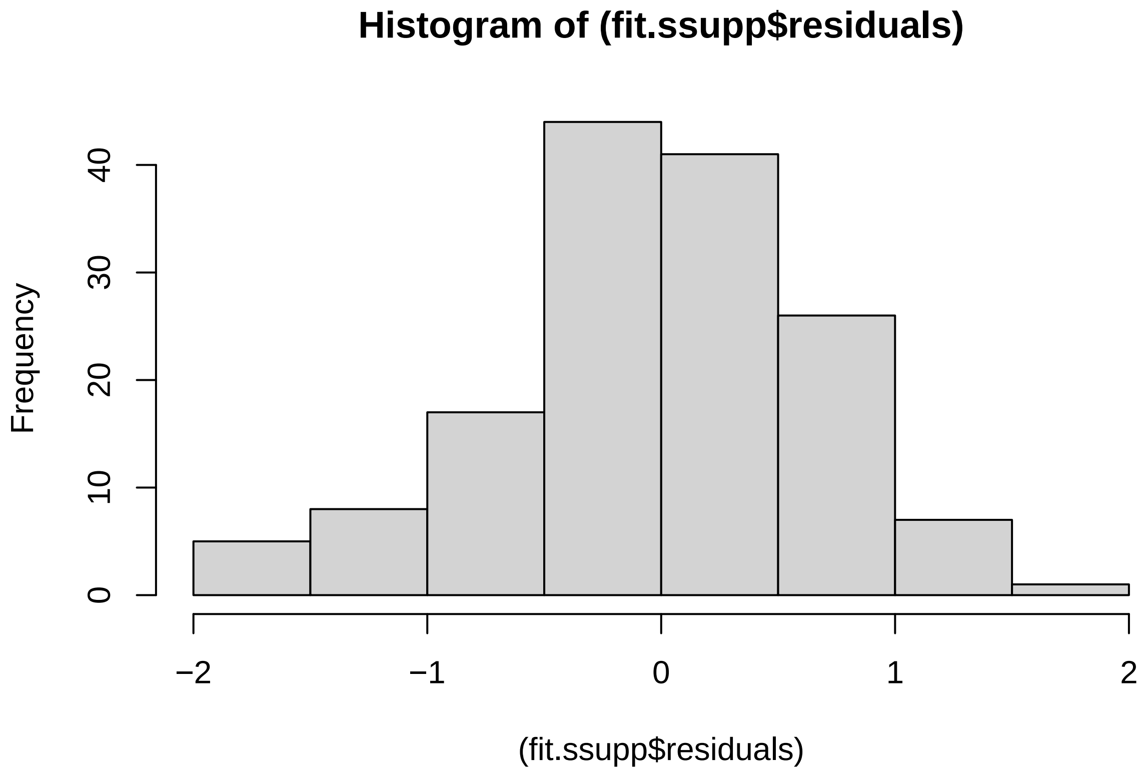
```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rstandard(fit.gdp)
## D = 0.057295, p-value = 0.2706
```

```
plot(fit.ssupp$residuals) #gledajuci reziduale na ovaj nacin tesko je suditi o normalnosti
```

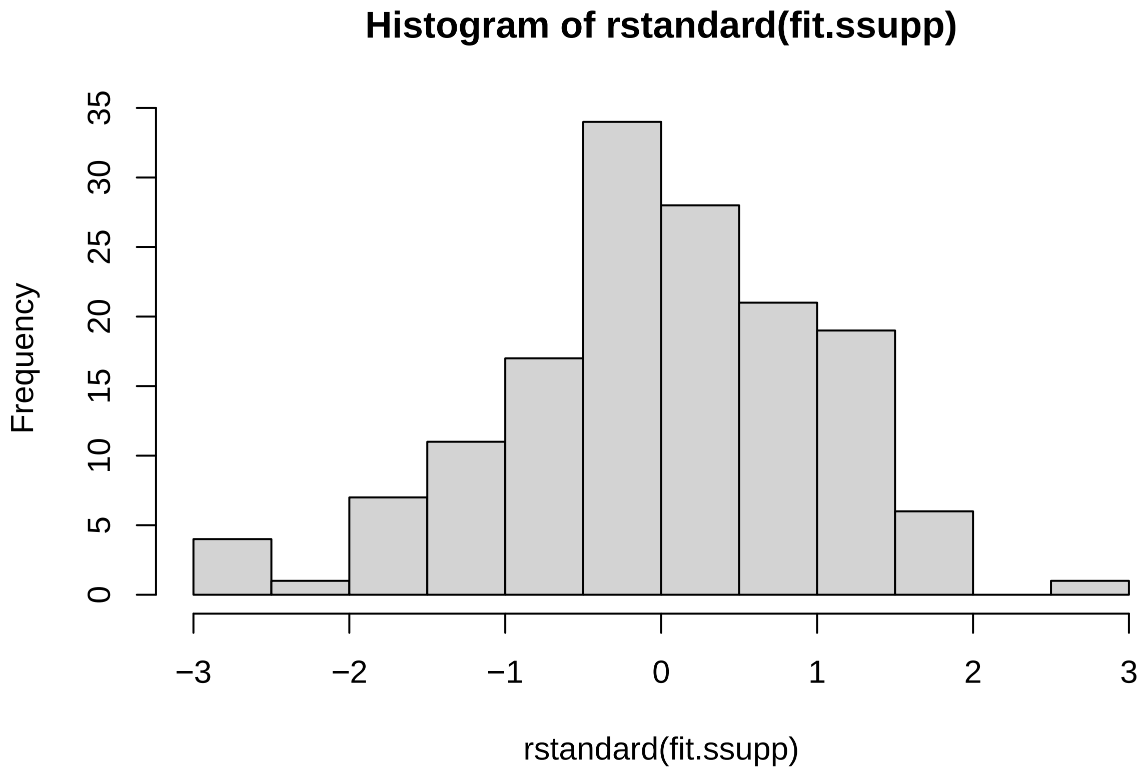


Provjera normalnosti socijalne podrške

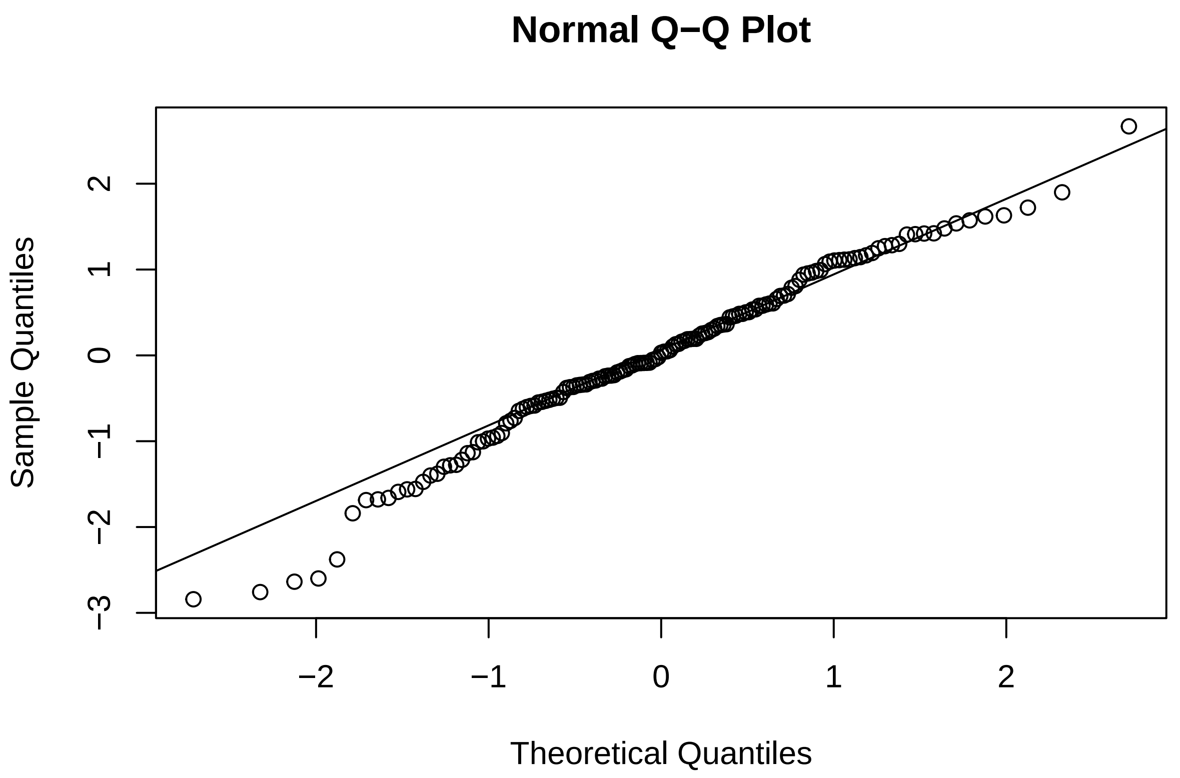
```
hist((fit.ssupp$residuals))
```



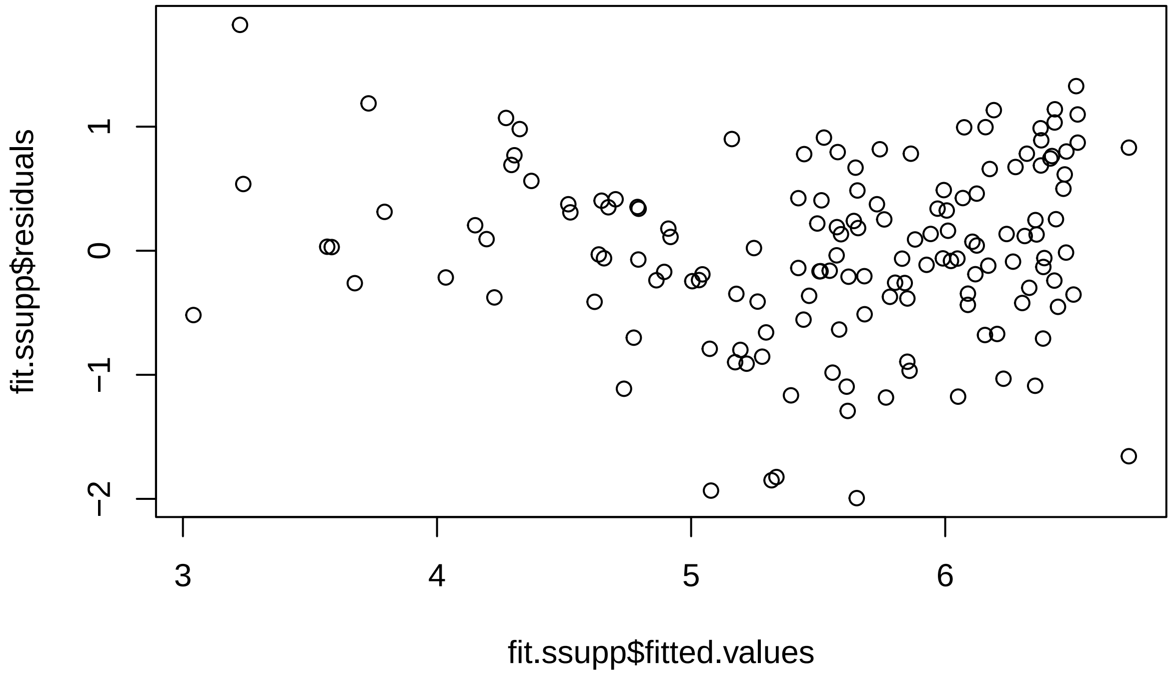
```
hist(rstandard(fit.ssupp))
```



```
qqnorm(rstandard(fit.ssupp))
qqline(rstandard(fit.ssupp))
```



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



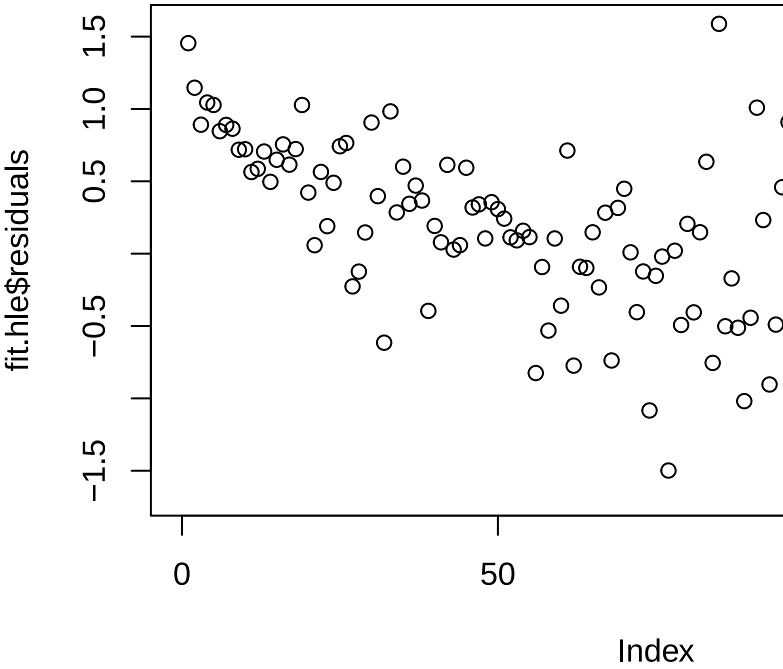
```
ks.test(rstandard(fit.ssupp),'pnorm')
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  rstandard(fit.ssupp)
## D = 0.063546, p-value = 0.5842
## alternative hypothesis: two-sided
```

```
lillie.test(rstandard(fit.ssupp))
```

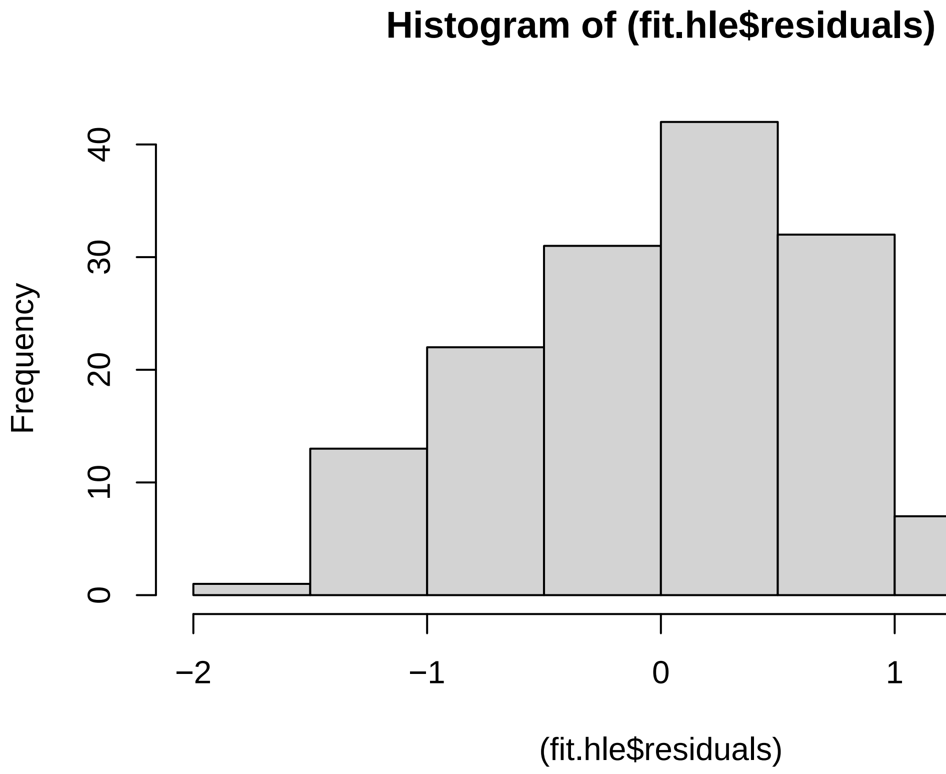
```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rstandard(fit.ssupp)
## D = 0.063637, p-value = 0.1475
```

```
plot(fit.hle$residuals) #gledajuci reziduale na ovaj nacin tesko je suditi o normalnosti
```

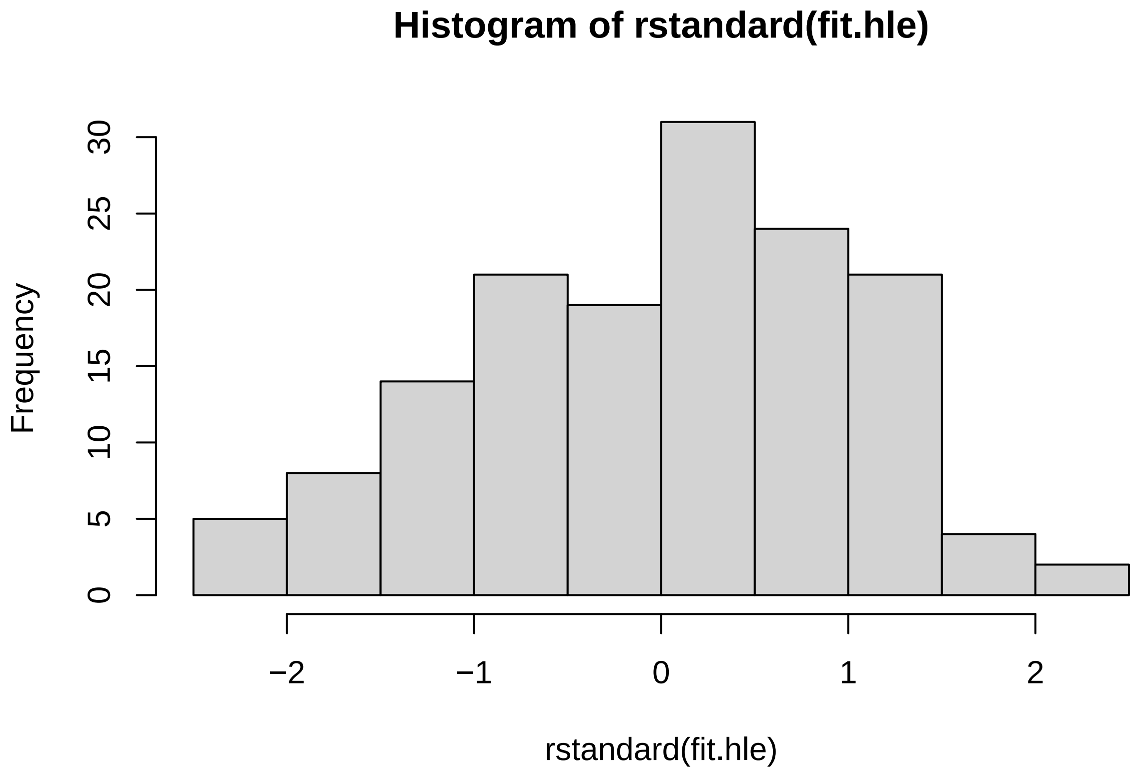


Provjera normalnosti očekivanja trajanja života

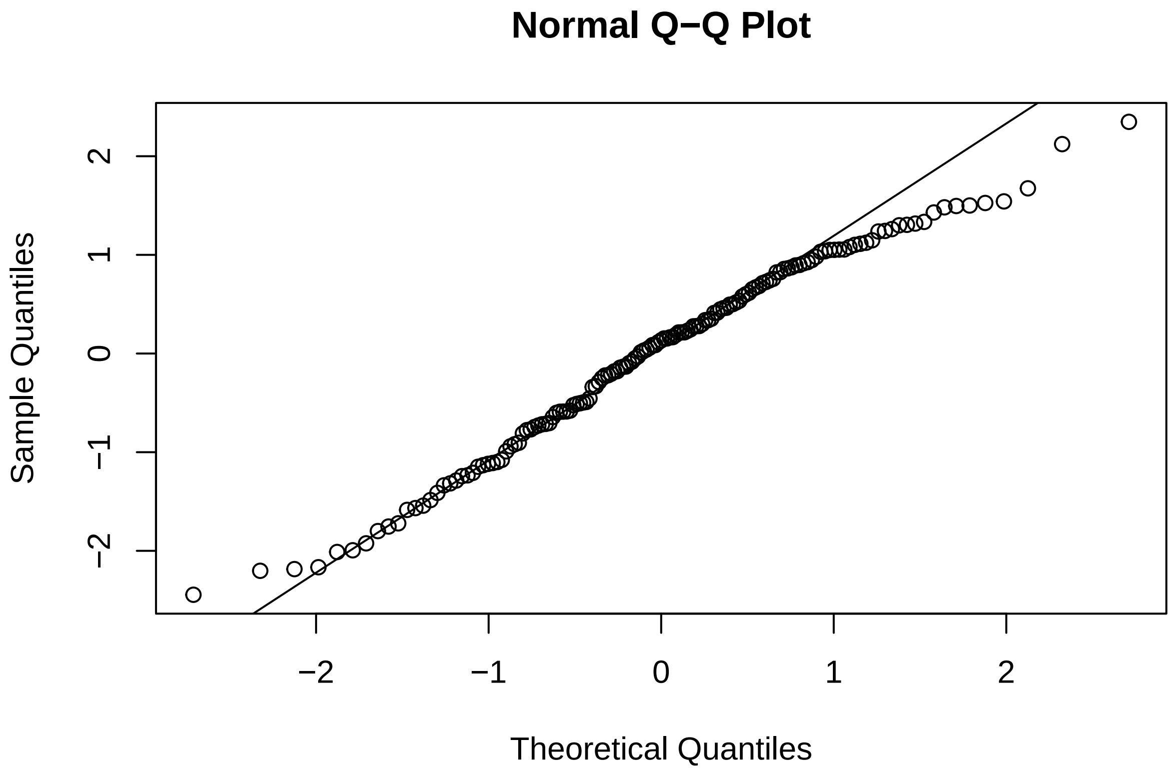
```
hist((fit.hle$residuals))
```



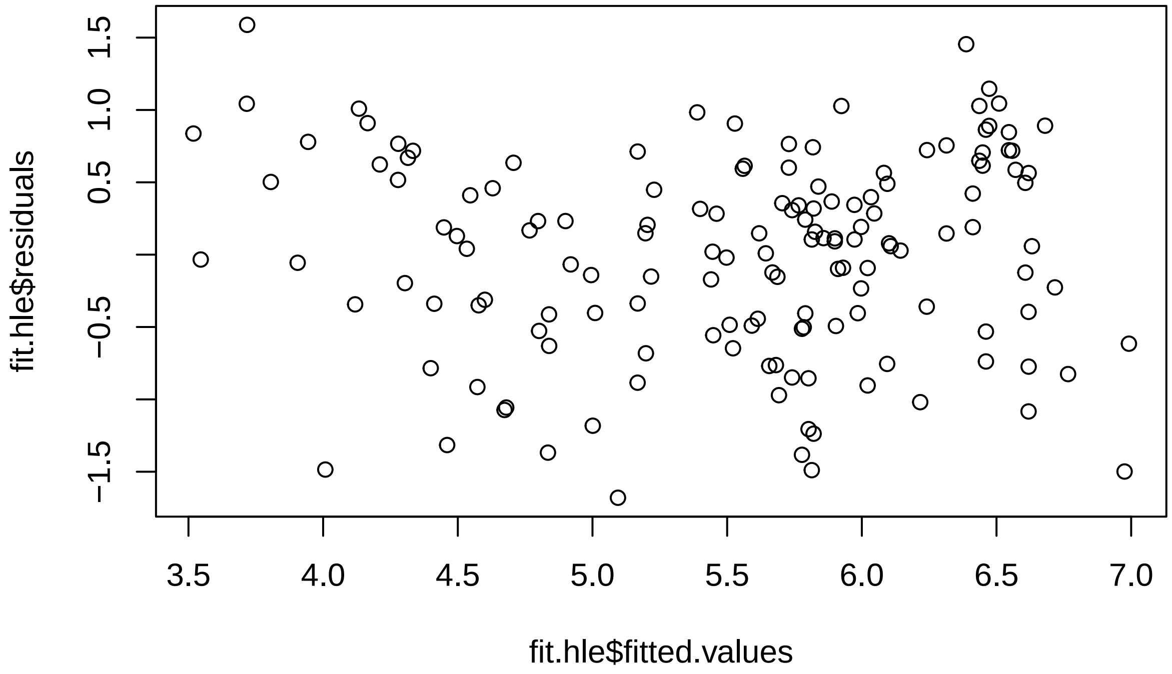
```
hist(rstandard(fit.hle))
```



```
qqnorm(rstandard(fit.hle))
qqline(rstandard(fit.hle))
```



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



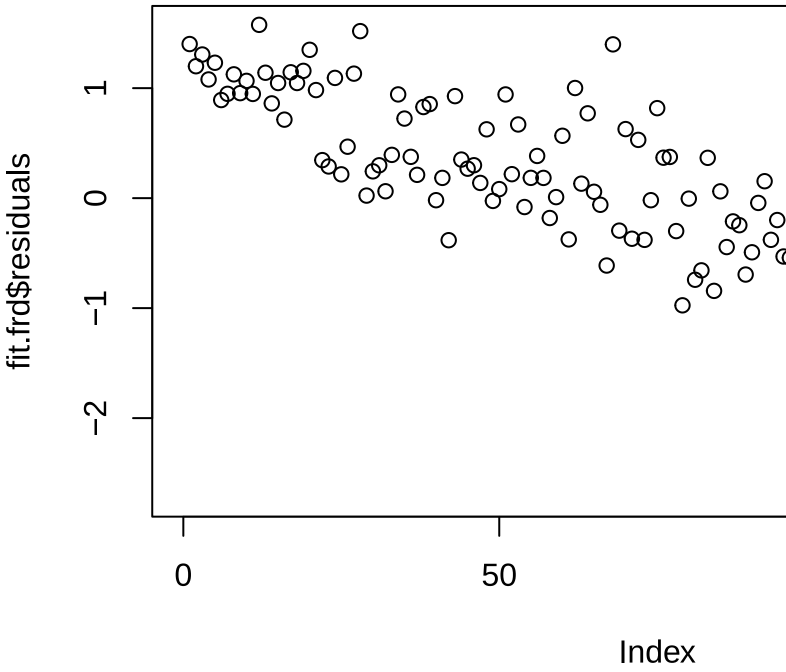
```
ks.test(rstandard(fit.hle),'pnorm')
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  rstandard(fit.hle)
## D = 0.057528, p-value = 0.7075
## alternative hypothesis: two-sided
```

```
lillie.test(rstandard(fit.hle))
```

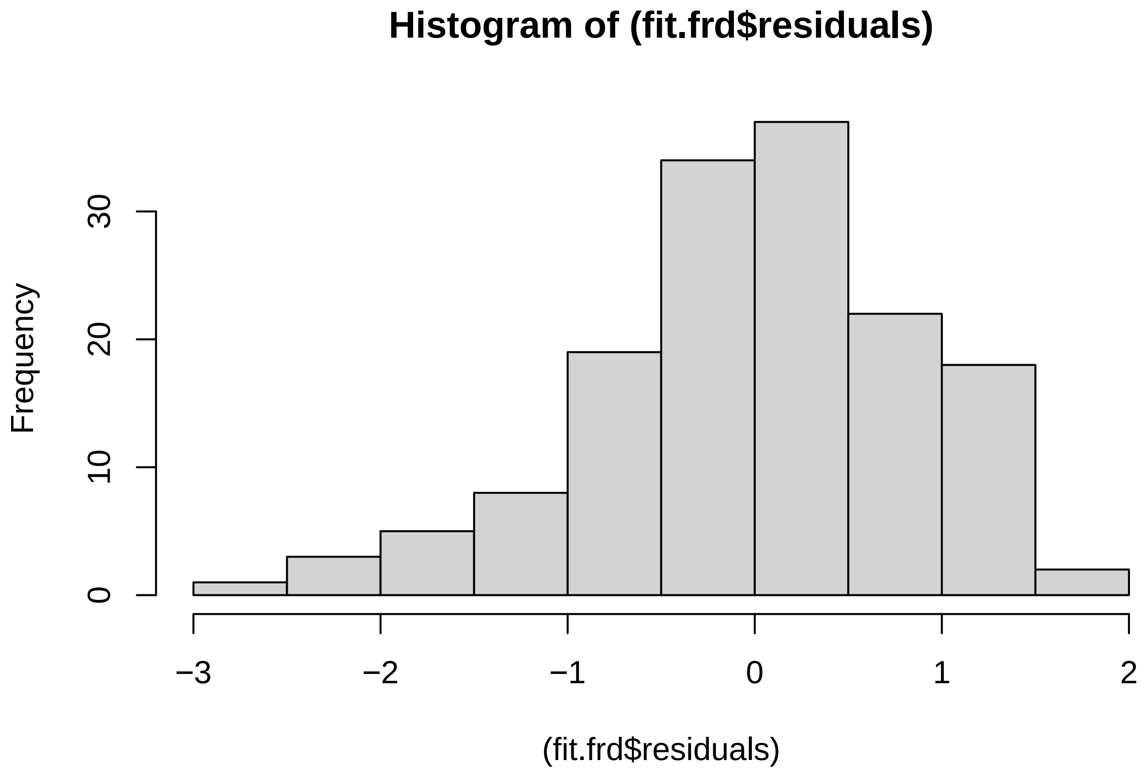
```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rstandard(fit.hle)
## D = 0.057051, p-value = 0.2765
```

```


plot(fit.frd$residuals) #gledajuci reziduale na ovaj nacin tesko je suditi o normalnosti
```

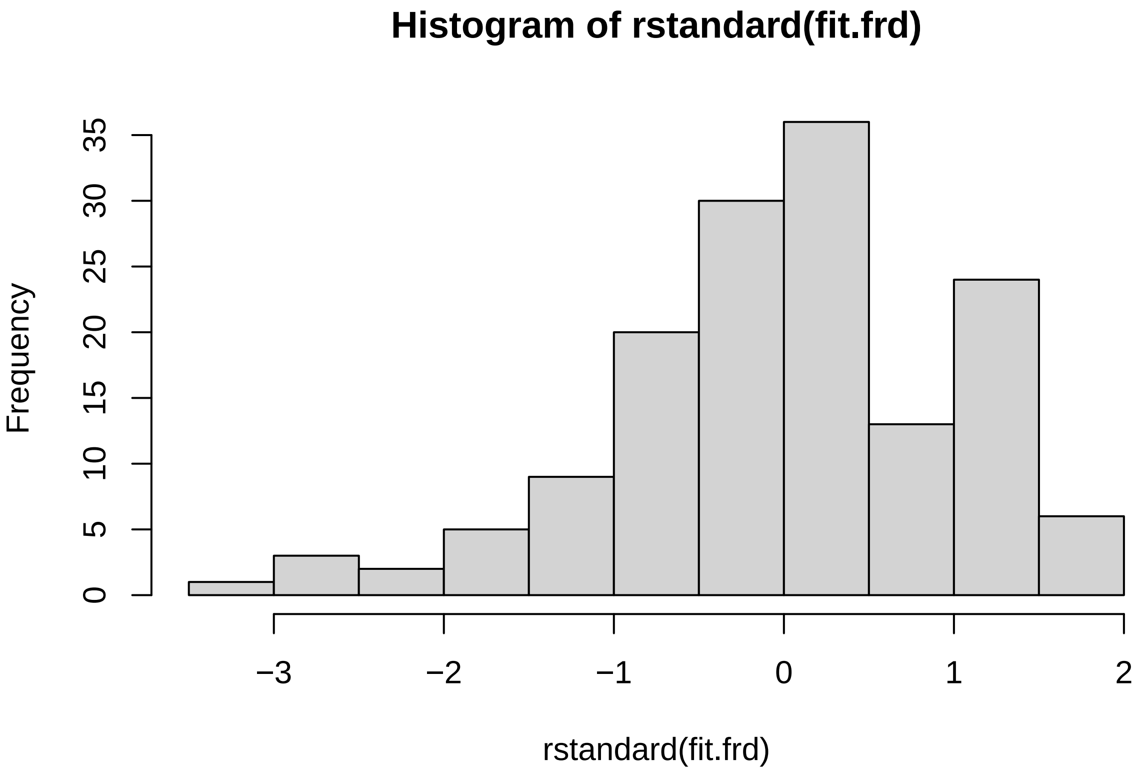


Provjera normalnosti slobode donošenja odluka

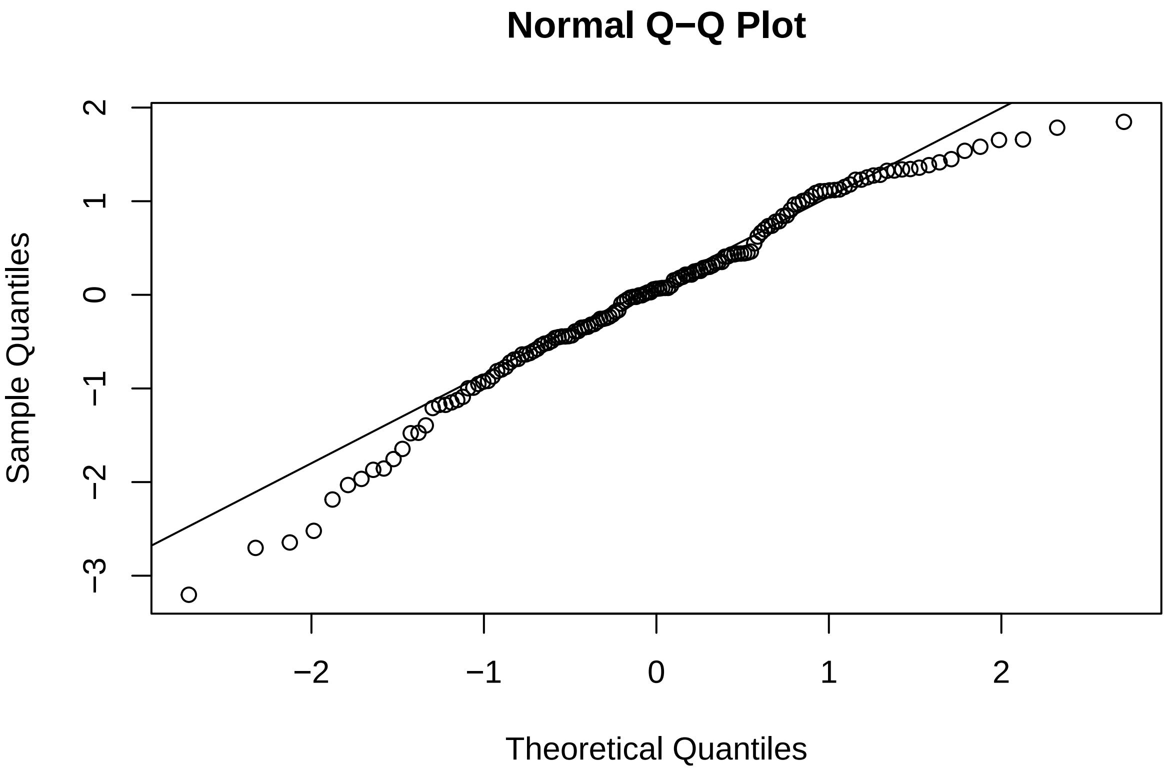
```
hist((fit.frd$residuals))
```



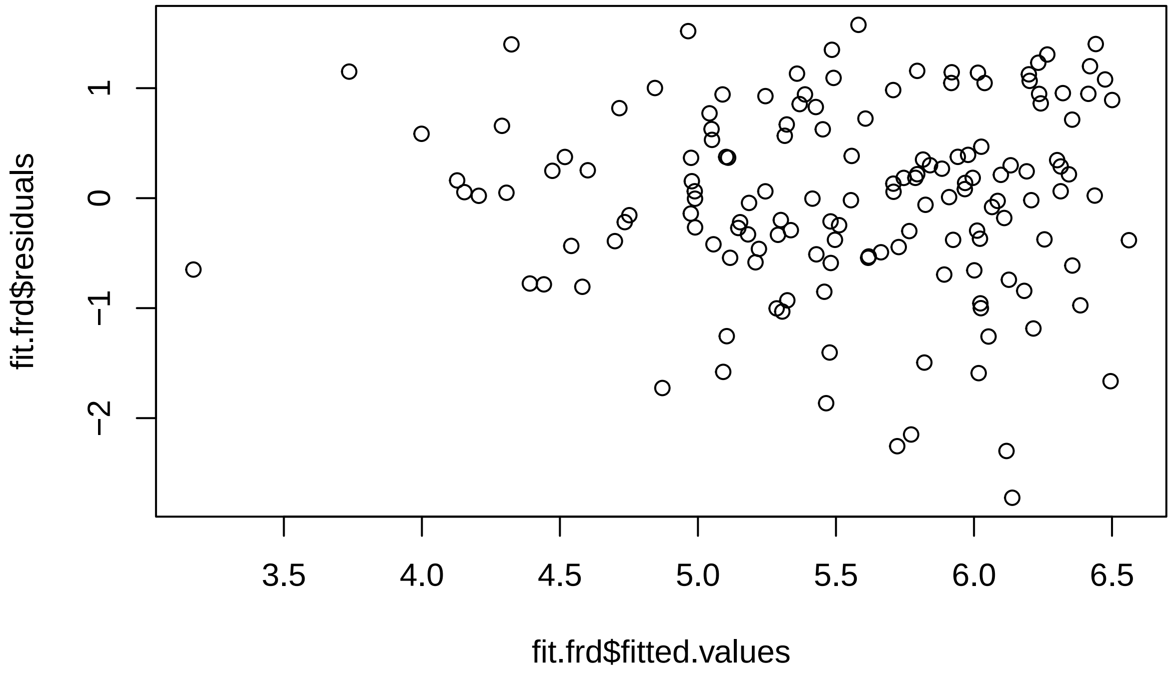
```
hist(rstandard(fit.frd))
```



```
qqnorm(rstandard(fit.frd))
qqline(rstandard(fit.frd))
```



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



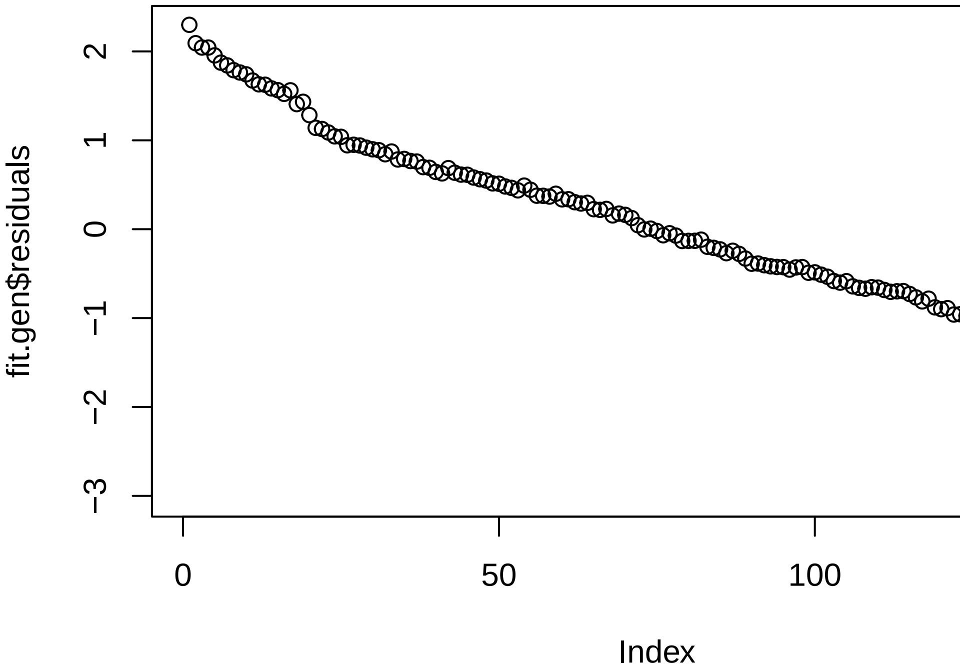
```
ks.test(rstandard(fit.frd),'pnorm')
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  rstandard(fit.frd)
## D = 0.051932, p-value = 0.8165
## alternative hypothesis: two-sided
```

```
lillie.test(rstandard(fit.frd))
```

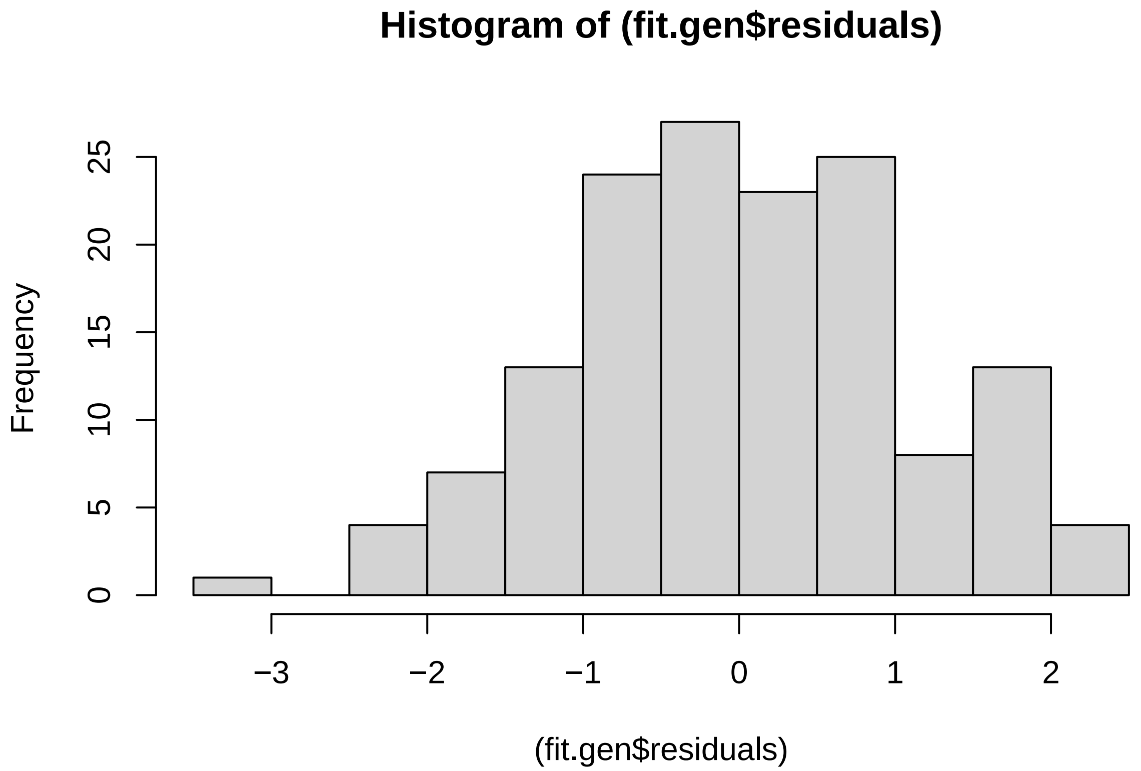
```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rstandard(fit.frd)
## D = 0.051823, p-value = 0.4228
```

```
plot(fit.gen$residuals) #gledajuci reziduale na ovaj nacin tesko je suditi o normalnosti
```

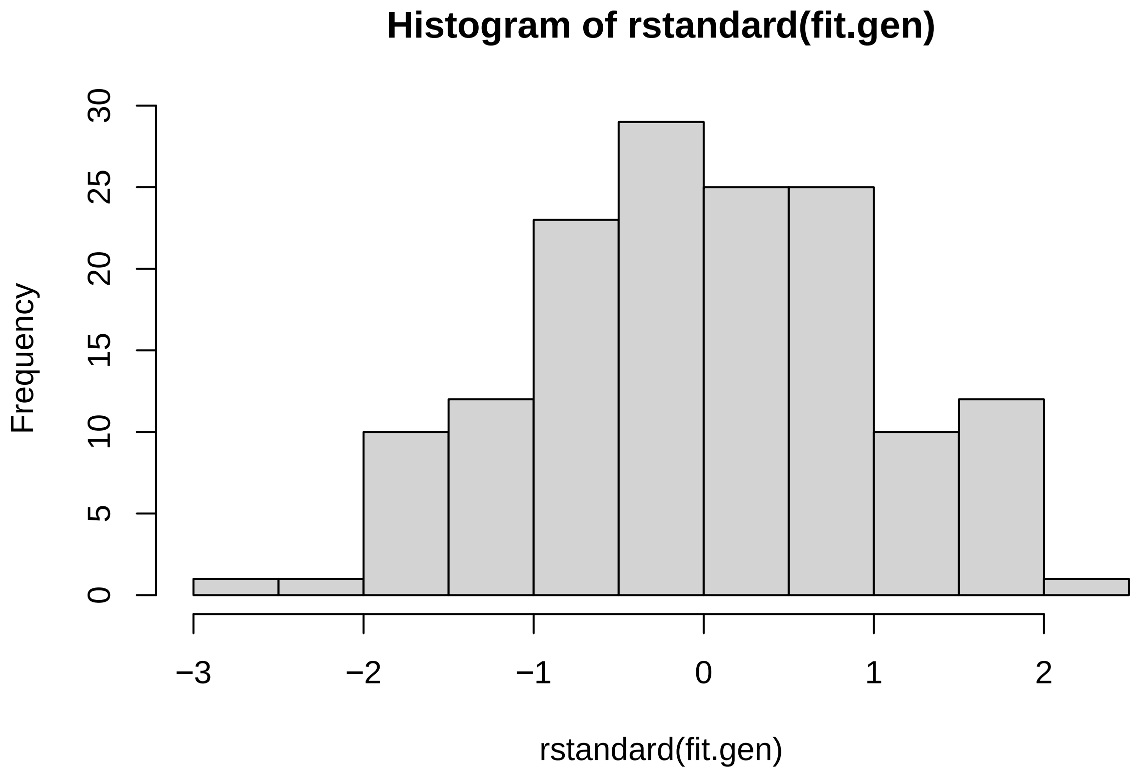


Provjera normalnosti darežljivosti

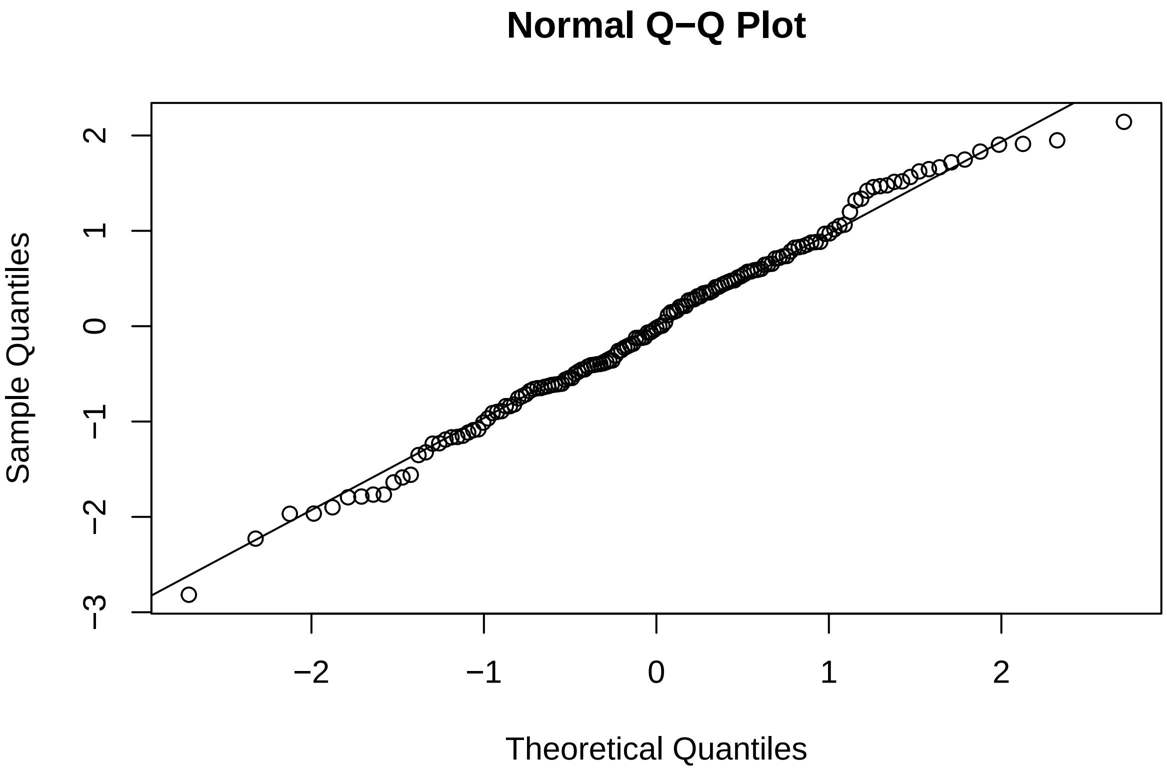
```
hist((fit.gen$residuals))
```



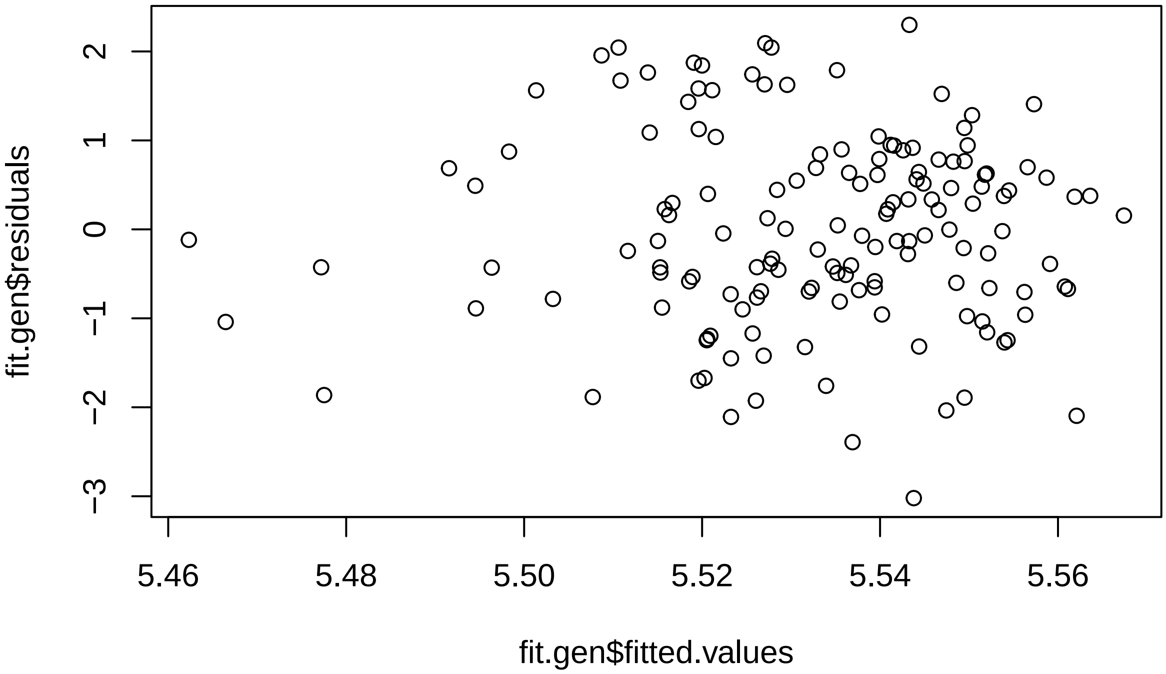
```
hist(rstandard(fit.gen))
```



```
qqnorm(rstandard(fit.gen))
qqline(rstandard(fit.gen))
```



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



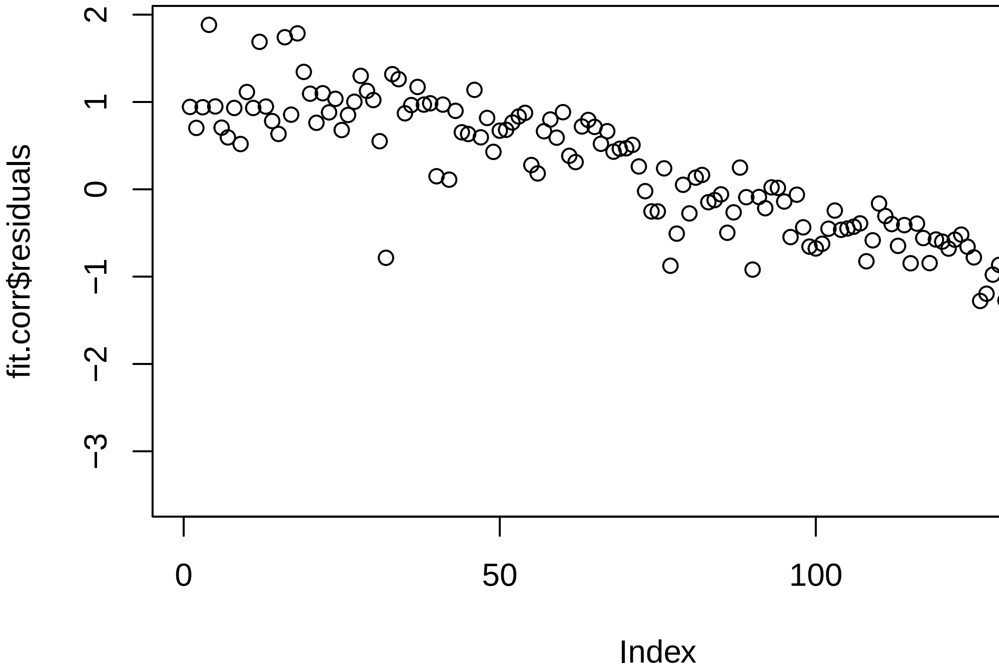
```
ks.test(rstandard(fit.gen),'pnorm')
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  rstandard(fit.gen)
## D = 0.043052, p-value = 0.9452
## alternative hypothesis: two-sided
```

```
lillie.test(rstandard(fit.gen))
```

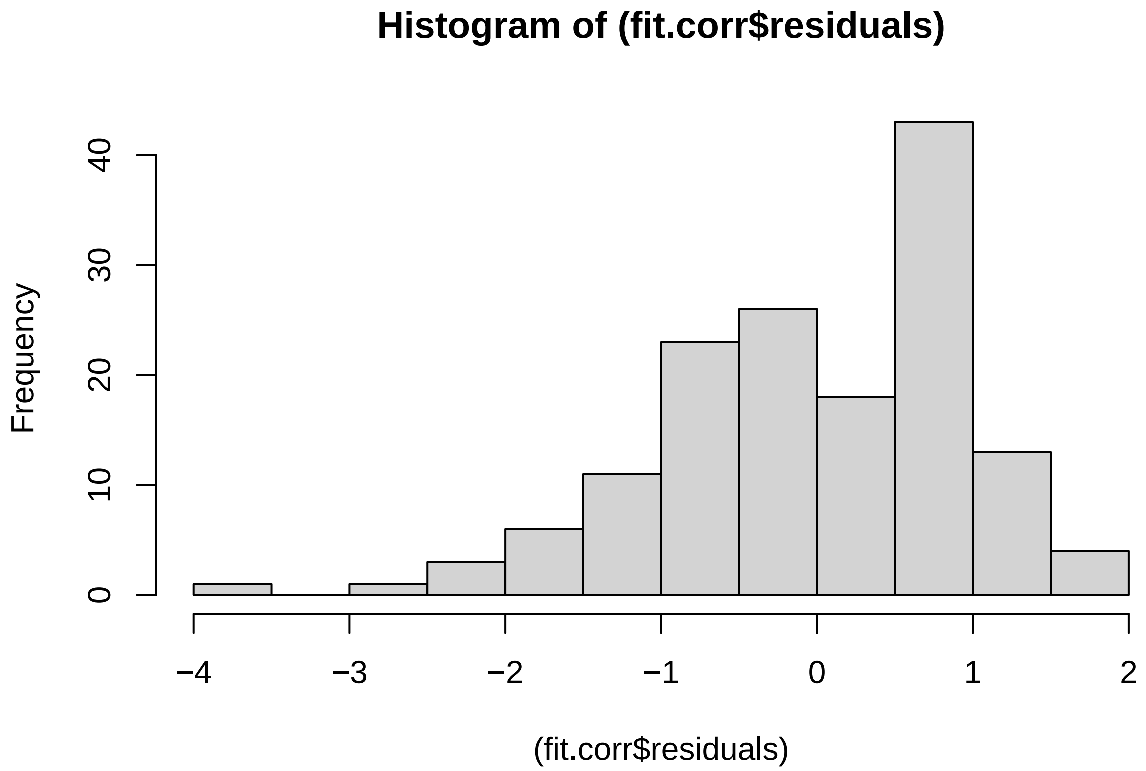
```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rstandard(fit.gen)
## D = 0.042407, p-value = 0.7372
```

```
plot(fit.corr$residuals) #gledajuci reziduale na ovaj nacin tesko je suditi o normalnosti
```

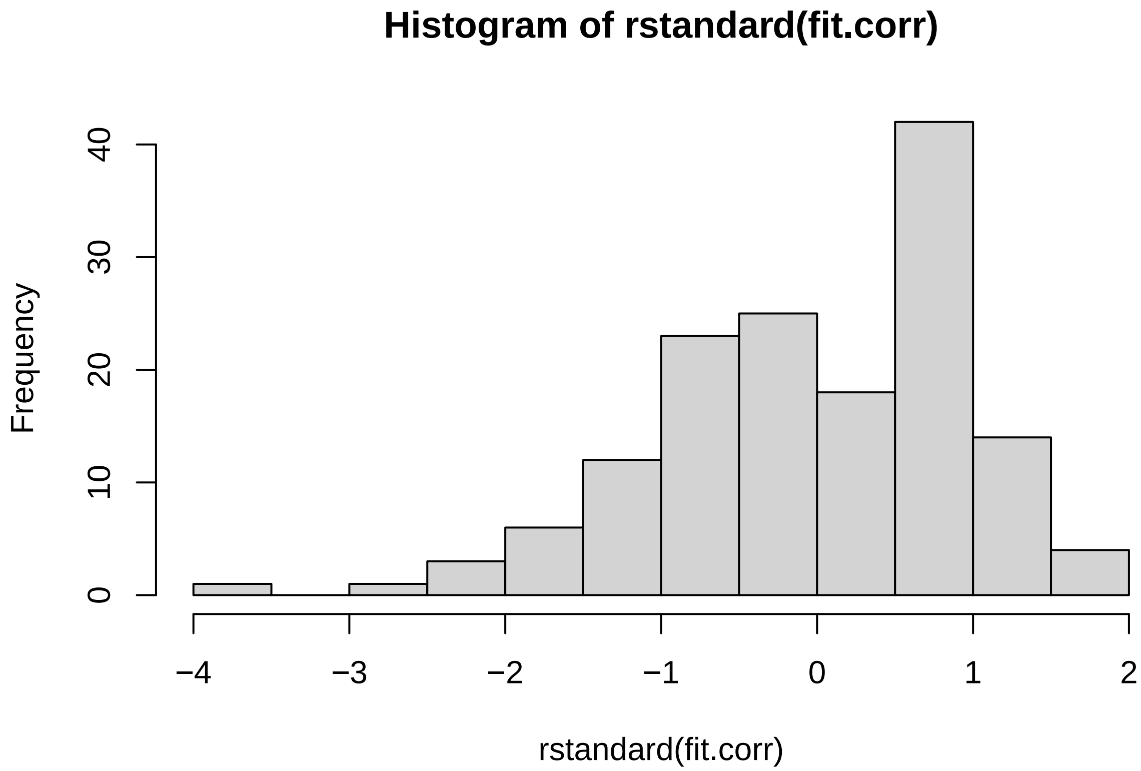


Provjera normalnosti korupcije

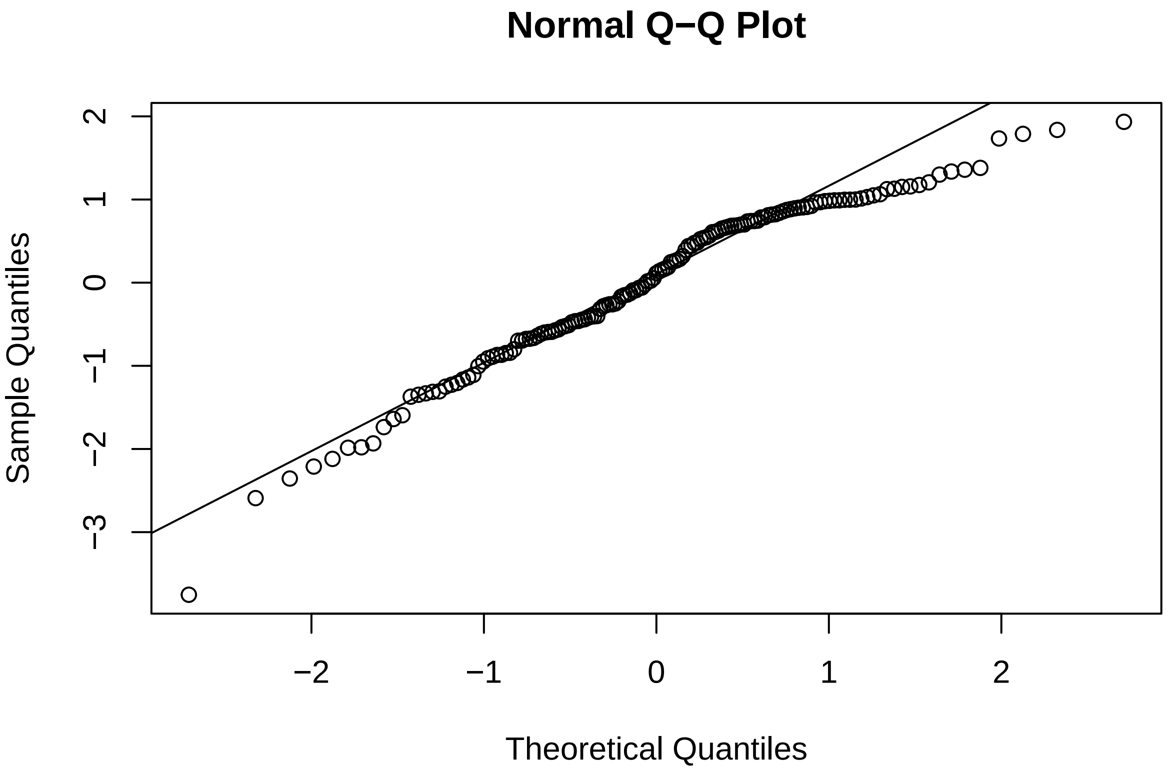
```
hist((fit.corr$residuals))
```



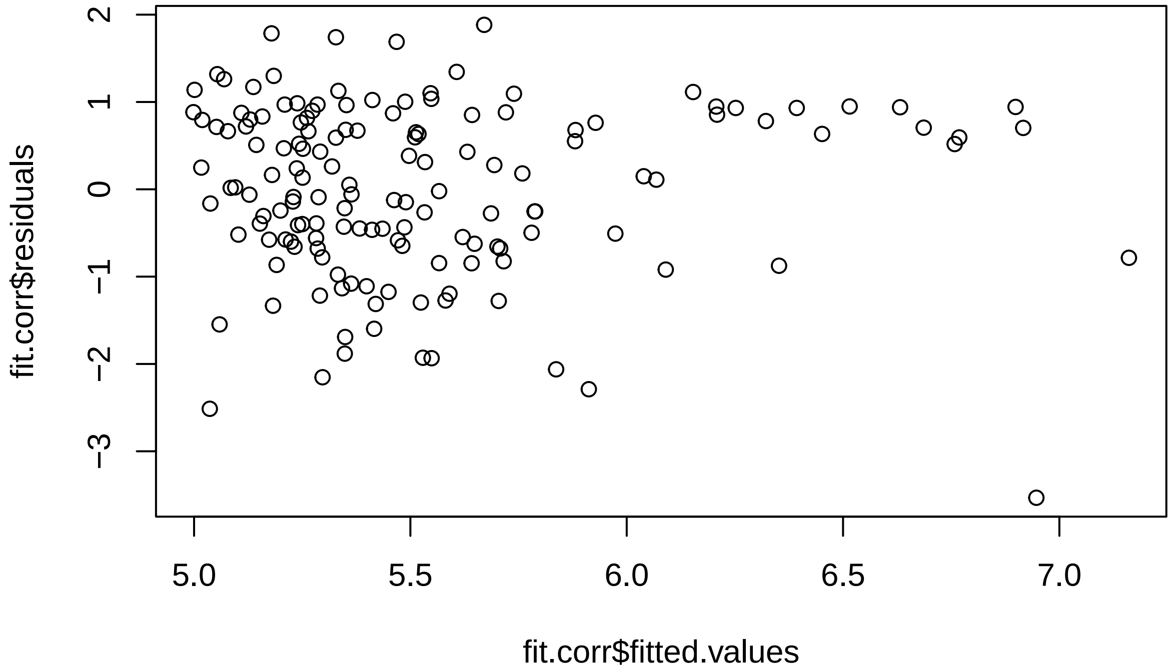
```
hist(rstandard(fit.corr))
```



```
qqnorm(rstandard(fit.corr))
qqline(rstandard(fit.corr))
```



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



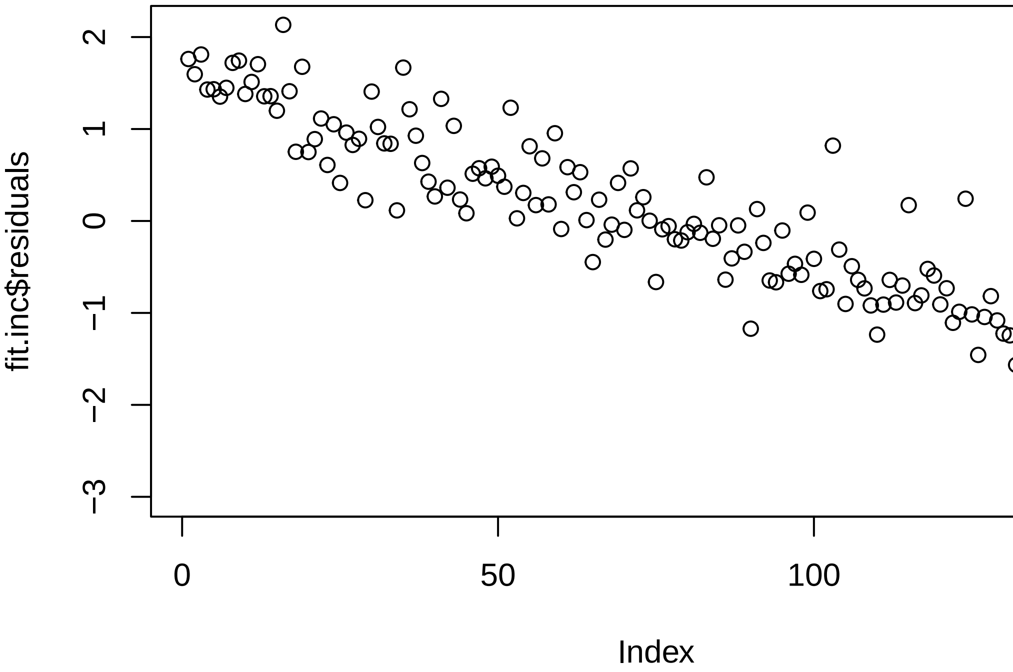
```
ks.test(rstandard(fit.corr),'pnorm')
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  rstandard(fit.corr)
## D = 0.10407, p-value = 0.07932
## alternative hypothesis: two-sided
```

```
lillie.test(rstandard(fit.corr))
```

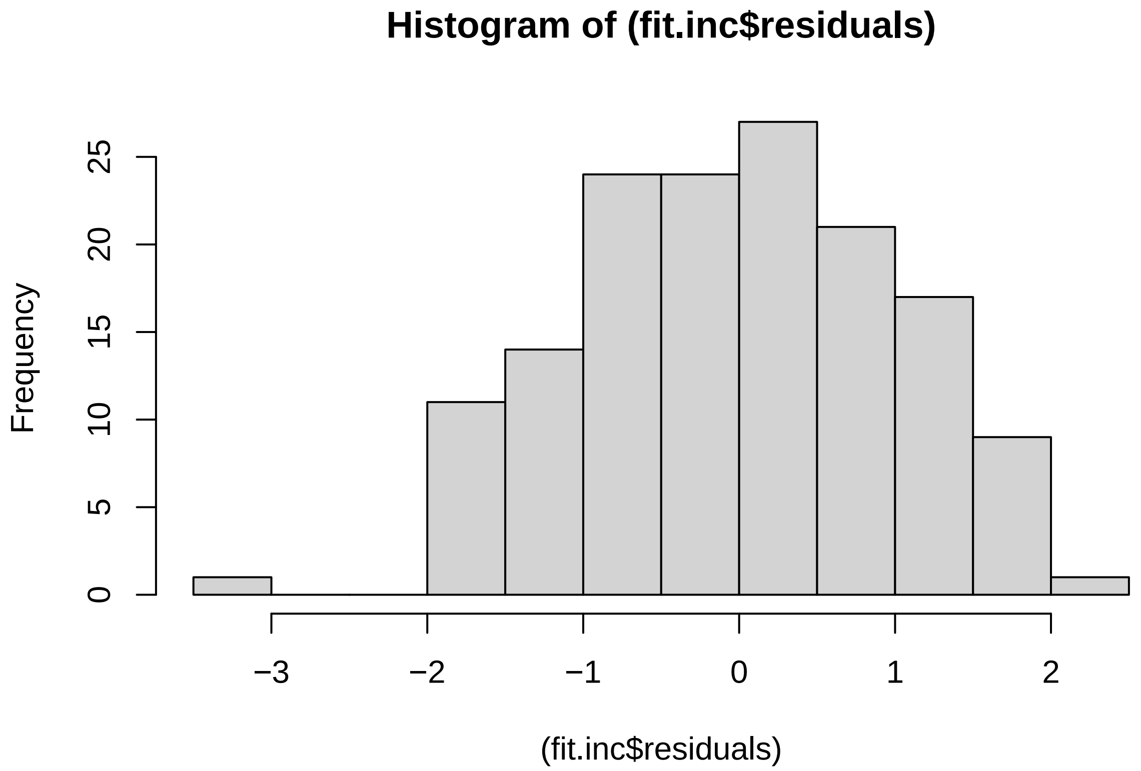
```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rstandard(fit.corr)
## D = 0.10286, p-value = 0.0005636
```

```
plot(fit.inc$residuals) #gledajuci reziduale na ovaj nacin tesko je suditi o normalnosti
```

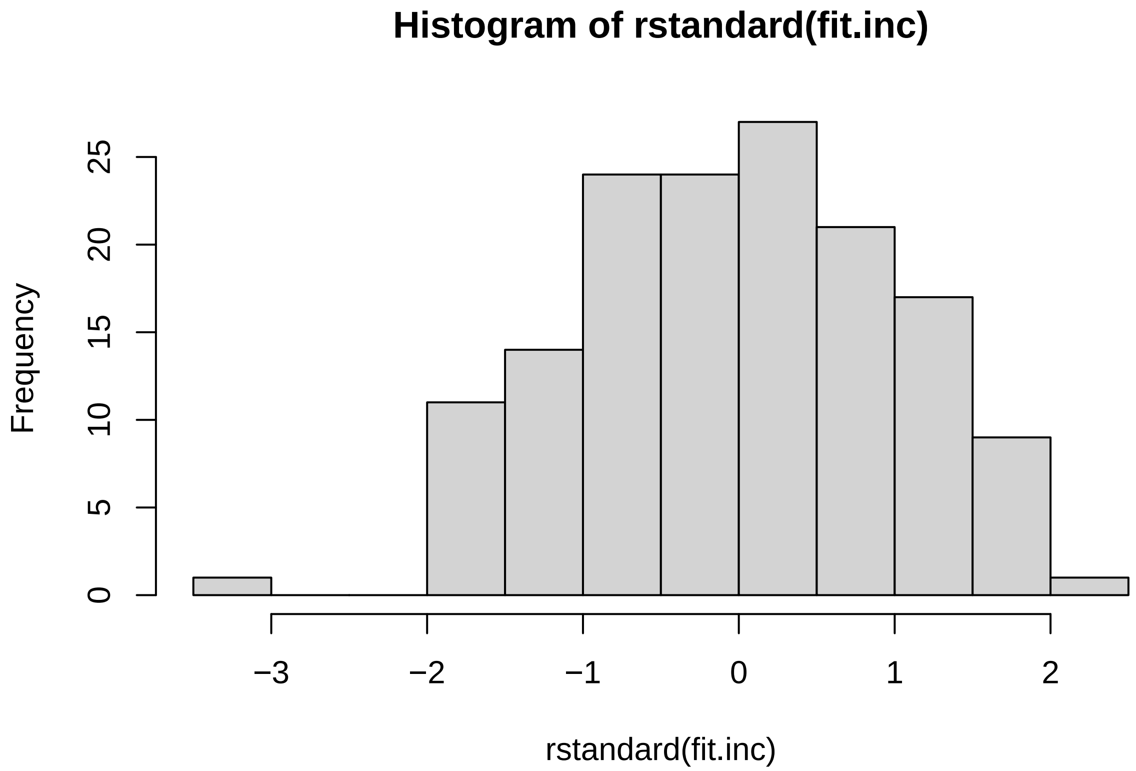


Provjera normalnosti dohotka

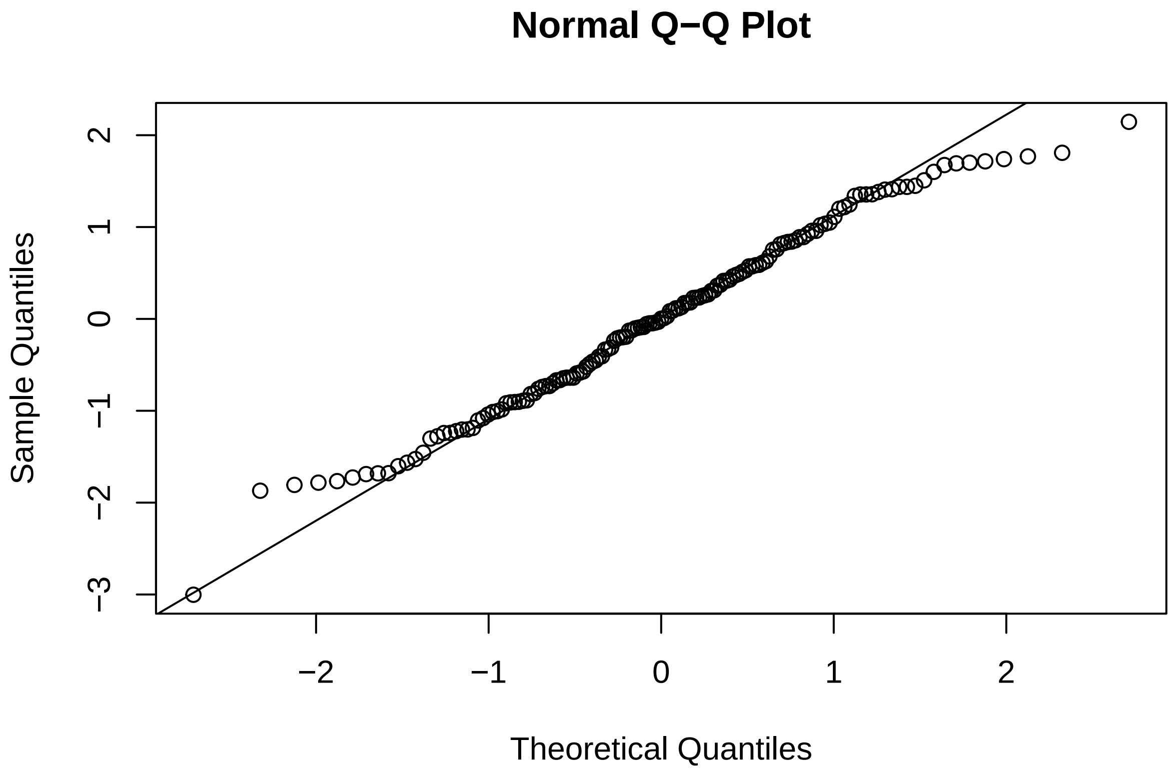
```
hist((fit.inc$residuals))
```



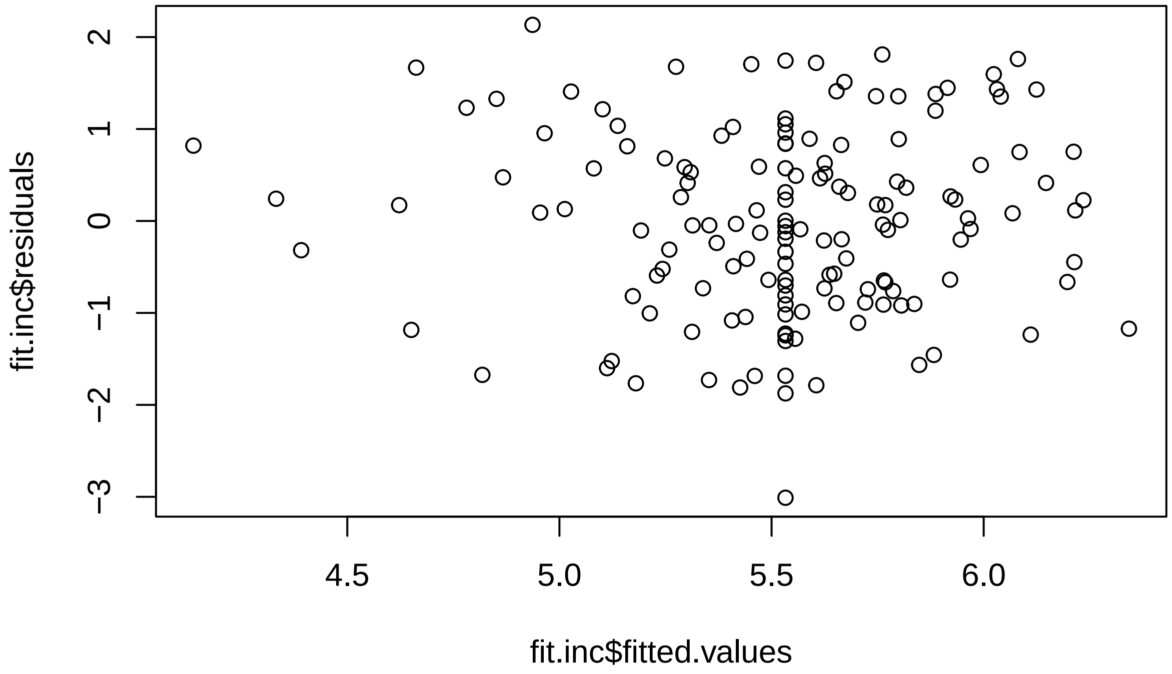
```
hist(rstandard(fit.inc))
```



```
qqnorm(rstandard(fit.inc))
qqline(rstandard(fit.inc))
```



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



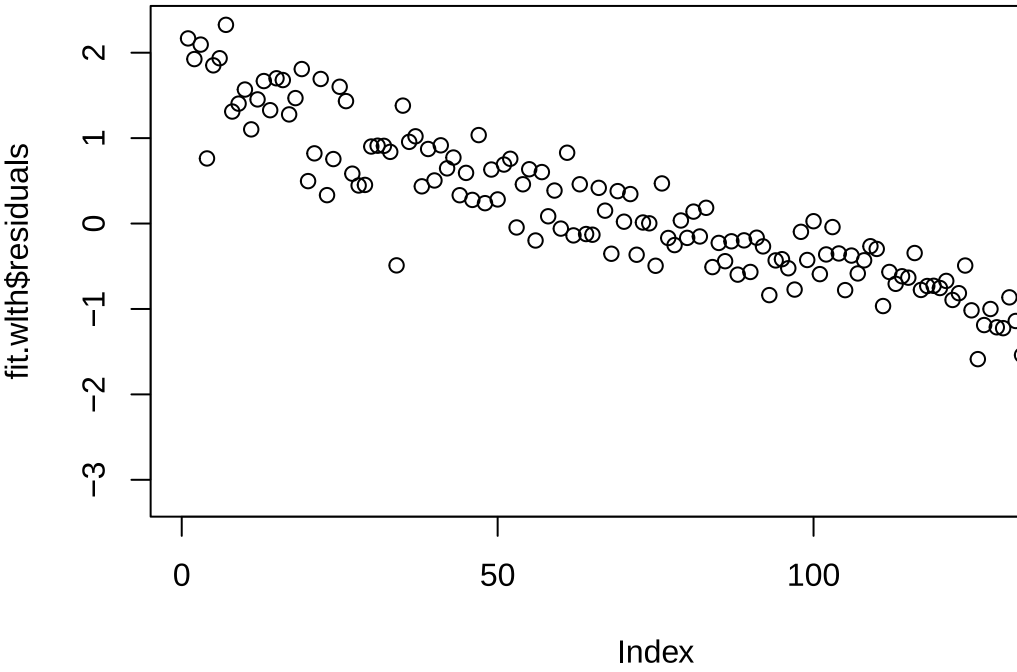
```
ks.test(rstandard(fit.inc),'pnorm')
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  rstandard(fit.inc)
## D = 0.047272, p-value = 0.8932
## alternative hypothesis: two-sided
```

```
lillie.test(rstandard(fit.inc))
```

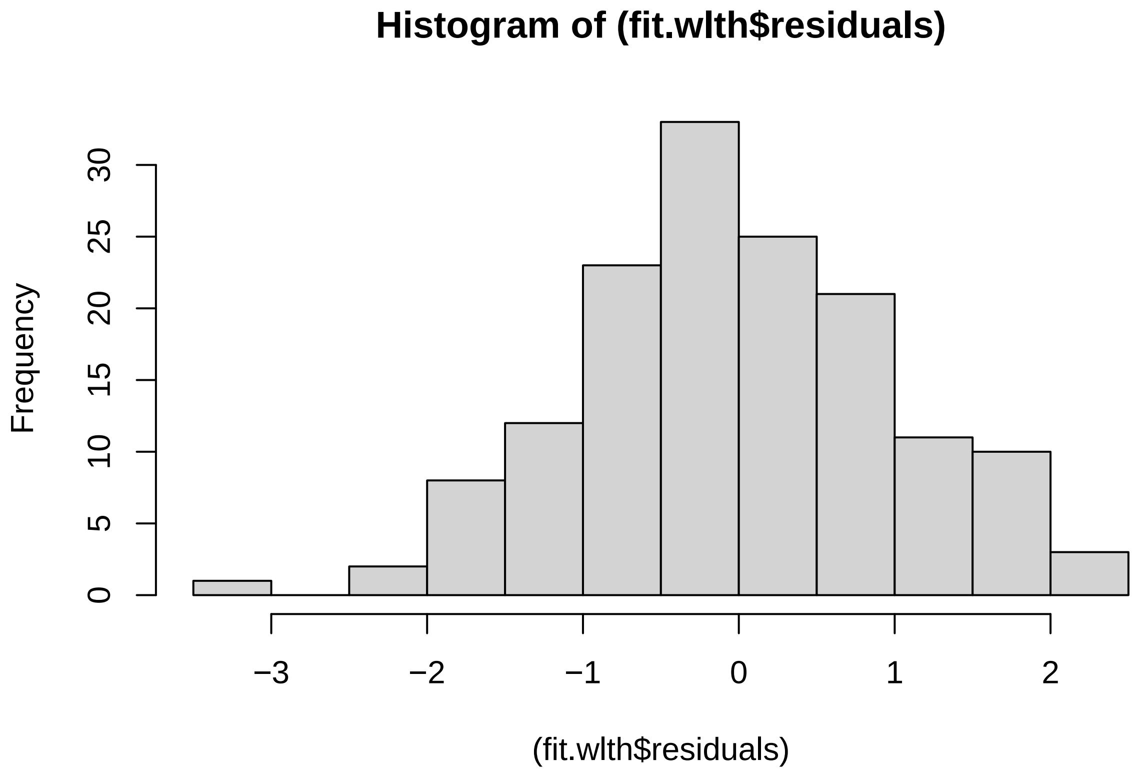
```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rstandard(fit.inc)
## D = 0.046873, p-value = 0.5864
```

```
plot(fit.wlth$residuals) #gledajuci reziduale na ovaj nacin tesko je suditi o normalnosti
```

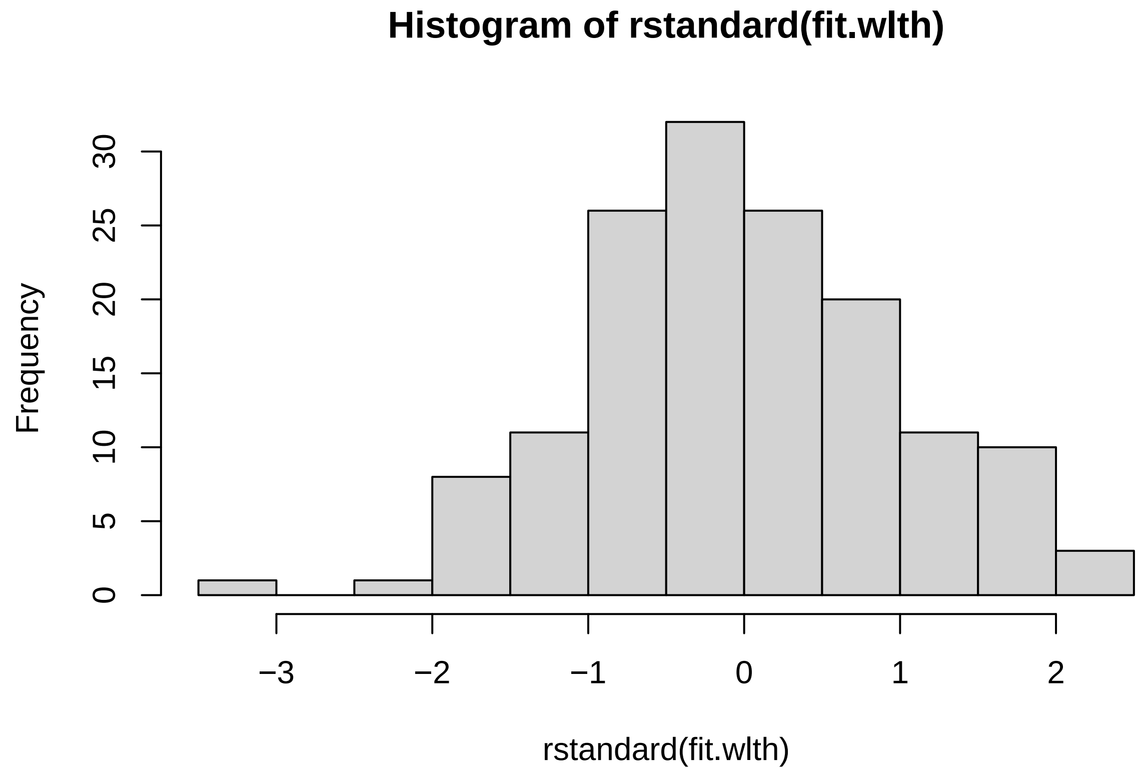


Provjera normalnosti prihoda

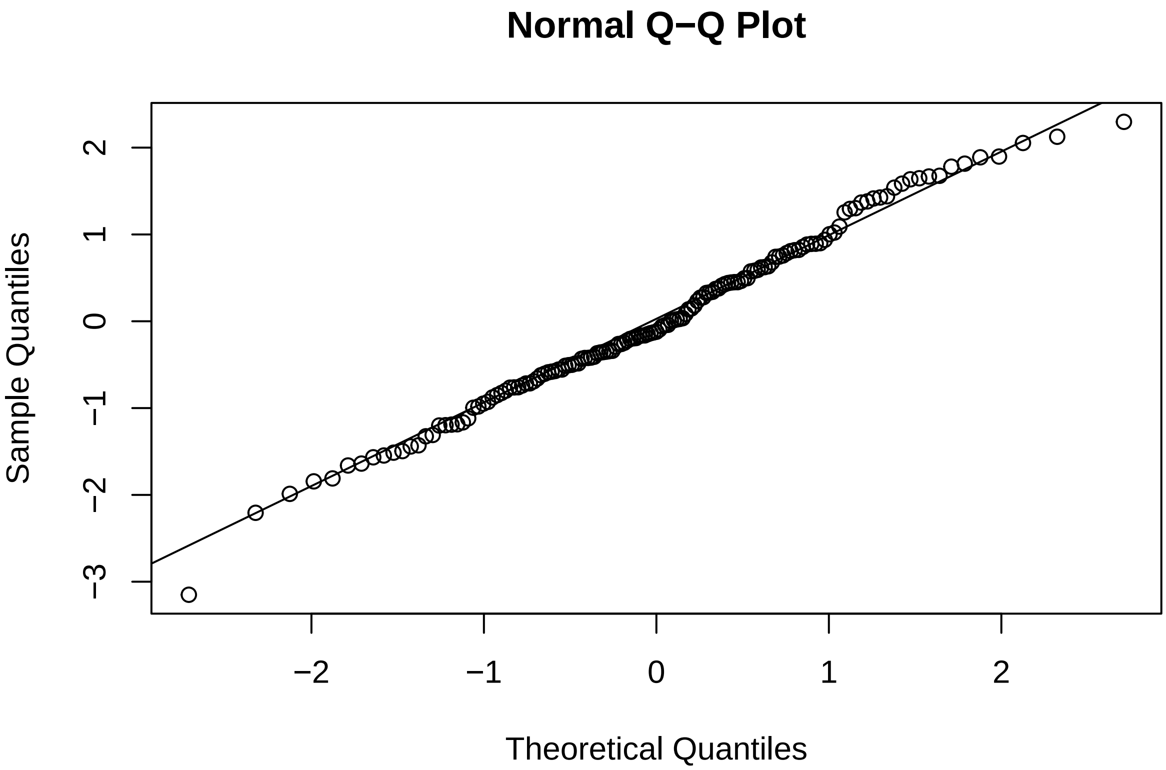
```
hist((fit.wlth$residuals))
```



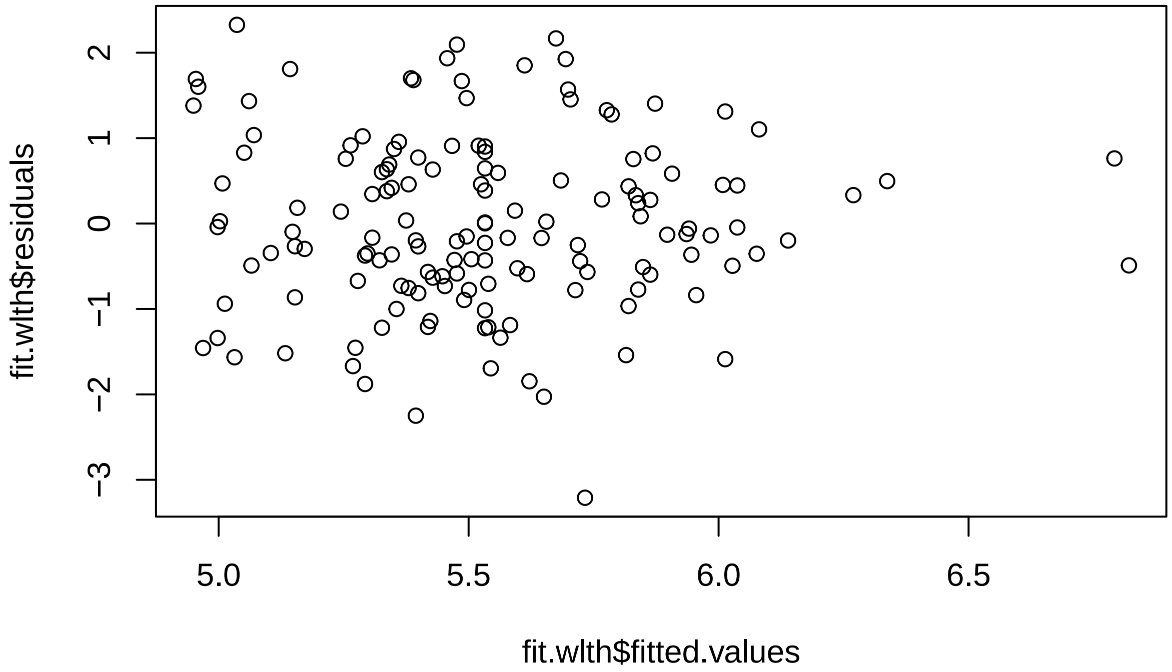
```
hist(rstandard(fit.wlth))
```



```
qqnorm(rstandard(fit.wlth))
qqline(rstandard(fit.wlth))
```



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



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

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  rstandard(fit.wlth)
## D = 0.051508, p-value = 0.8241
## alternative hypothesis: two-sided
```

```
lillie.test(rstandard(fit.wlth))
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rstandard(fit.wlth)
## D = 0.051536, p-value = 0.4318
```

Iz dobivenih histograma reziduala i QQ-plota, možemo zaključiti da svi osim varijable Perception of corruption približno zadovoljavaju zahtjev o normalnosti reziduala.

Koja je od dostupnih varijabli najbolji prediktor razine sreće?

Ocjena kvalitete modela i statističko zaključivanje o procijenjenom modelu

Ako pretpostavke modela nisu (neprihvatljivo) prekršene, moguće je primijeniti različite statističke testove o procijenjenim koeficijentima i modelu.

t-test koeficijenata modela Budući da vrijedi $B_i \sim N(\mu_{B_i}, \sigma_{B_i})$, $\mu_{B_i} = \beta_i$, statistika

$$T = \frac{B_i - \beta_i}{SE(B_i)}$$

ima t-distribuciju s $n-k-1$ stupnjeva slobode, gdje je k broj parametara. Većina programskih paketa, pa tako i R, pri estimiranju koeficijenata linearne regresije automatski testira $\beta_i = 0$. One koeficijente za koje možemo odbaciti $H_0 : \beta_i = 0$ u korist $H_1 : \beta_i \neq 0$ zovemo **značajni koeficijenti**.

Mjere kvalitete prilagodbe modela podatcima

SSE Mjera koju minimiziramo estimiranjem parametara modela ("fitanjem na podatke") je SSE:

$$SSE = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

$\mathbf{R^2}$ Vrlo česta mjera kvalitete prilagodbe modela je koeficijent deteminacije, definiran kao:

$$R^2 = 1 - \frac{SSE}{SST},$$

gdje je: $SST = \sum_{i=1}^{N} (y_i - \bar{y}_i)^2$ tzv. "total corrected sum of squares". Koeficijent determinacije R^2 je za linearne modele po definiciji $R^2 \in [0,1]$ i opisuje koji postotak varijance u izlaznoj varijabli Y je estimirani linearni model objasnio/opisao.

Adjusted $\mathbf{R^2}$ Prilagođeni koeficijent determinacije penalizira dodatne parametre u modelu:

$$R^2_{adj} = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}.$$

F-test

Za ispitivanje signifikantnosti čitavog modela koristi se F-statistika:

$$f = \frac{SSR/k}{SSE/(n-k-1)},$$

gdje je $SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$.

Sve navedene mjere vidjet ćemo pozivanjem summary() za modele jednostavne regresije, koje smo dosad napravili.

```
summary(fit.gdp)
```

```
##
## Call:
## lm(formula = whr2021$‘Ladder score‘ ~ whr2021$‘Logged GDP per capita‘,
##     data = whr2021)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.32205 -0.46197  0.08219  0.50811  1.32615
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      -1.3719     0.4456  -3.078  0.00248 **
## whr2021$‘Logged GDP per capita‘   0.7320     0.0469  15.609  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.661 on 147 degrees of freedom
## Multiple R-squared:  0.6237, Adjusted R-squared:  0.6211
## F-statistic: 243.6 on 1 and 147 DF,  p-value: < 2.2e-16
```

```
summary(fit.ssupp)
```

```
##
## Call:
## lm(formula = whr2021$‘Ladder score‘ ~ whr2021$‘Social support‘,
##     data = whr2021)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.99388 -0.37152  0.02153  0.46039  1.82052
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -0.2318     0.4145  -0.559    0.577
## whr2021$‘Social support‘   7.0756     0.5038  14.045   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.7041 on 147 degrees of freedom
## Multiple R-squared:  0.573,  Adjusted R-squared:  0.5701
## F-statistic: 197.3 on 1 and 147 DF,  p-value: < 2.2e-16
```

```
summary(fit.hle)
```

```
##
## Call:
## lm(formula = whr2021$‘Ladder score‘ ~ whr2021$‘Healthy life expectancy‘,
##     data = whr2021)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.67985 -0.49034  0.09193  0.56415  1.58813
##
```

```
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       -2.394987   0.548098   -4.37 2.34e-05 ***
## whr2021$‘Healthy life expectancy‘  0.121980   0.008388   14.54  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.6901 on 147 degrees of freedom
## Multiple R-squared:  0.5899, Adjusted R-squared:  0.5871
## F-statistic: 211.5 on 1 and 147 DF,  p-value: < 2.2e-16
```

```
summary(fit.frd)
```

```
##
## Call:
## lm(formula = whr2021$‘Ladder score‘ ~ whr2021$‘Freedom to make life choices‘,
##     data = whr2021)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.72397 -0.46161  0.05457  0.62788  1.57557
##
## Coefficients:
##                                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)                              0.9736     0.4964   1.962   0.0517 .
## whr2021$‘Freedom to make life choices‘   5.7597     0.6208   9.278   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.8557 on 147 degrees of freedom
## Multiple R-squared:  0.3693, Adjusted R-squared:  0.365
## F-statistic: 86.09 on 1 and 147 DF,  p-value: < 2.2e-16
```

```
summary(fit.gen)
```

```
##
## Call:
## lm(formula = whr2021$‘Ladder score‘ ~ whr2021$Generosity, data = whr2021)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.02089 -0.69621 -0.02025  0.69882  2.29880
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         5.53091    0.08871  62.347   <2e-16 ***
## whr2021$Generosity -0.12666    0.58785  -0.215     0.83
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.077 on 147 degrees of freedom
## Multiple R-squared:  0.0003157,  Adjusted R-squared:  -0.006485
## F-statistic: 0.04643 on 1 and 147 DF,  p-value: 0.8297
```

```
summary(fit.corr)
```

```
##
## Call:
## lm(formula = whr2021$‘Ladder score‘ ~ whr2021$‘Perceptions of corruption‘,
##     data = whr2021)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5321 -0.5989  0.1103  0.7941  1.8833
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          7.3675     0.3357  21.944   <2e-16 ***
## whr2021$‘Perceptions of corruption‘  -2.5219     0.4482  -5.627    9e-08 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.9774 on 147 degrees of freedom
## Multiple R-squared:  0.1772, Adjusted R-squared:  0.1716
## F-statistic: 31.66 on 1 and 147 DF,  p-value: 8.996e-08
```

```
summary(fit.inc)
```

```
##
## Call:
## lm(formula = whr2021$‘Ladder score‘ ~ whr2021$‘Income Gini‘,
##     data = whr2021)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.00993 -0.73327  0.00267  0.75265  2.13284
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)            7.5713     0.4451  17.011  < 2e-16 ***
## whr2021$‘Income Gini‘  -5.4742     1.1746  -4.661 7.01e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.006 on 147 degrees of freedom
## Multiple R-squared:  0.1287, Adjusted R-squared:  0.1228
## F-statistic: 21.72 on 1 and 147 DF,  p-value: 7.012e-06
```

```
summary(fit.wlth)
```

```
##
## Call:
## lm(formula = whr2021$‘Ladder score‘ ~ whr2021$‘Wealth Gini‘,
##     data = whr2021)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.2100 -0.6338 -0.1228  0.6908  2.3261 
## 
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)    
## (Intercept)             9.2527     0.9374   9.871  < 2e-16 ***
## whr2021$‘Wealth Gini‘  -4.8349     1.2135  -3.984 0.000106 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 1.024 on 147 degrees of freedom
## Multiple R-squared:  0.09747,    Adjusted R-squared:  0.09133 
## F-statistic: 15.87 on 1 and 147 DF,  p-value: 0.0001061
```

Zaključak 2. pitanja s pojavom novog pitanja

Dosad smo iz grafičkih prikaza mogli vidjeti kako najviše varijabla GDP, pa onda i varijable Social Support, Healthy life expectancy te Freedom to make life choices imaju vrlo jak efekt na Ladder score, tj. razinu sreće. Vidimo da su za točno te varijable i najveće vrijednosti R2). Dakle, te varijable su onda i najbolji prediktori razine sreće, a s obzirom da varijabla GDP ima najveći R2, zaključujemo da je ona najbolji prediktor sreće.

Zbog daljnje analize, prisjetimo se da varijabla Perception of Corruption ne zadovoljava pretpostavku o normalnosti reziduala, te primijetimo da je za varijablu Generosity udio varijance blizu nuli te nam F-test pokazuje da taj model nije statistički značajan.

Također, iako nisu svi modeli jednako kvalitetni, u svim ostalim slučajevima, osim varijabli Percepction of Corruption i Generosity, koeficijenti uz zavisnu varijablu su značajni, te F-testovi upućuju na to i da su svi modeli značajni (objašnjavaju značajno više varijance od nul modela). Zaključujemo da čak i varijable Income gini i Wealth gini nisu suvišne u modeliranju.

No, postavlja nam se prirodno pitanje - mogu li sve varijable višestrukom regresijom bolje predvidjeti razinu sreće od varijable GDP i kojom kombinacijom tih varijabli? U nastavku se bavimo tim pitanjem.

NAPOMENA: pogledati valja li ovaj zaključak i na temelju gore danih t-testa

##itd proučiti to još malo prije slanja i prokomentirati

Dodatno pitanje: Može li višestruka regresija bolje predvidjeti razinu sreće od varijable GDP per Capital?

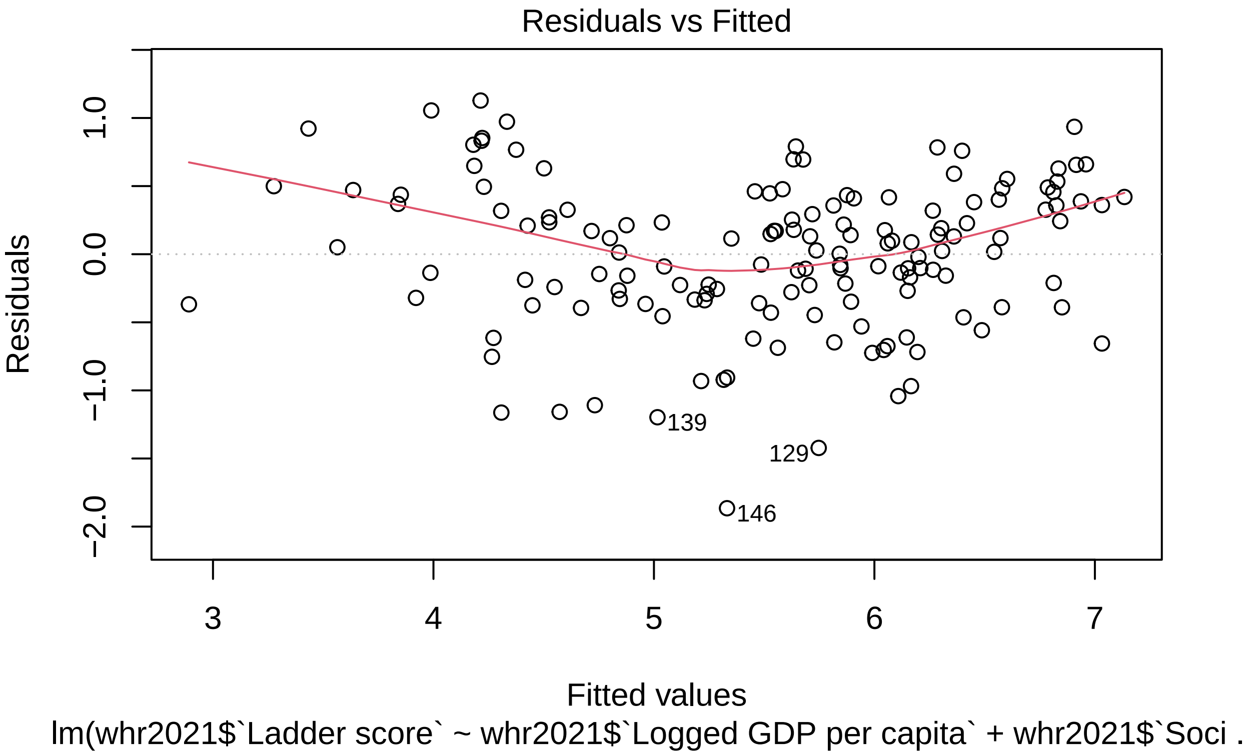
Koje će varijable biti uključene?

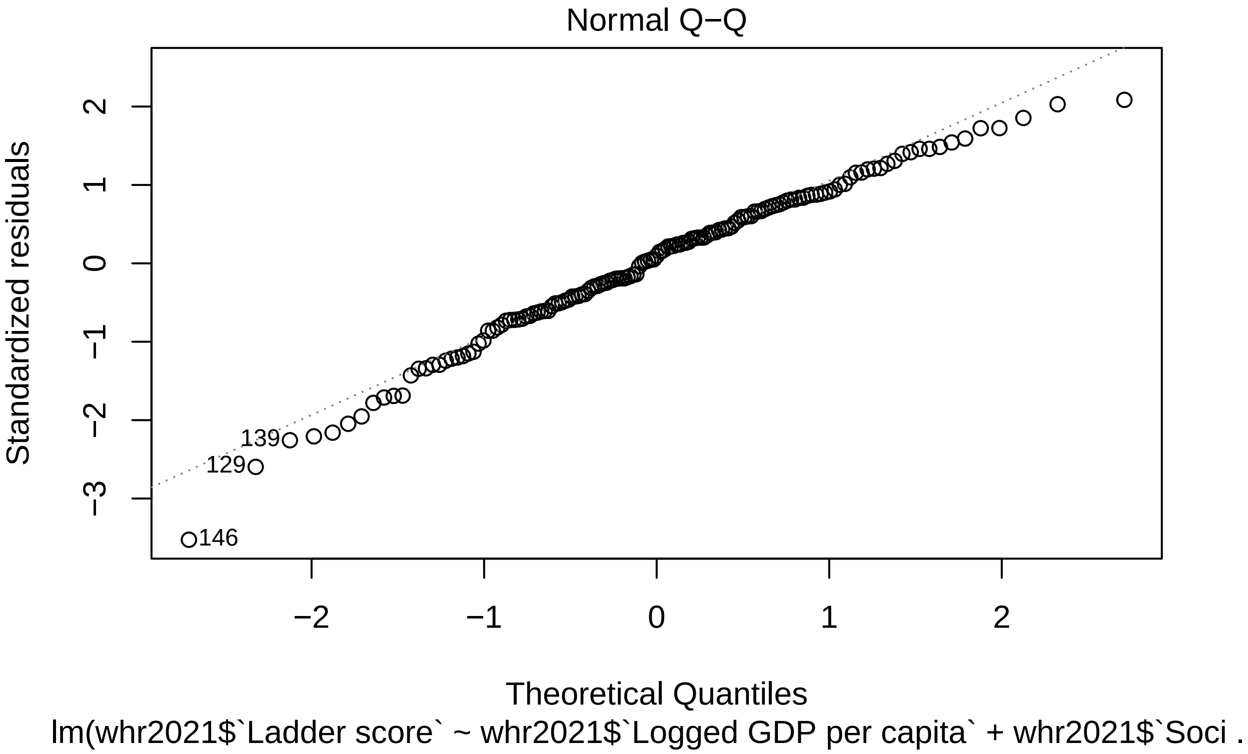
Višestruka regresija

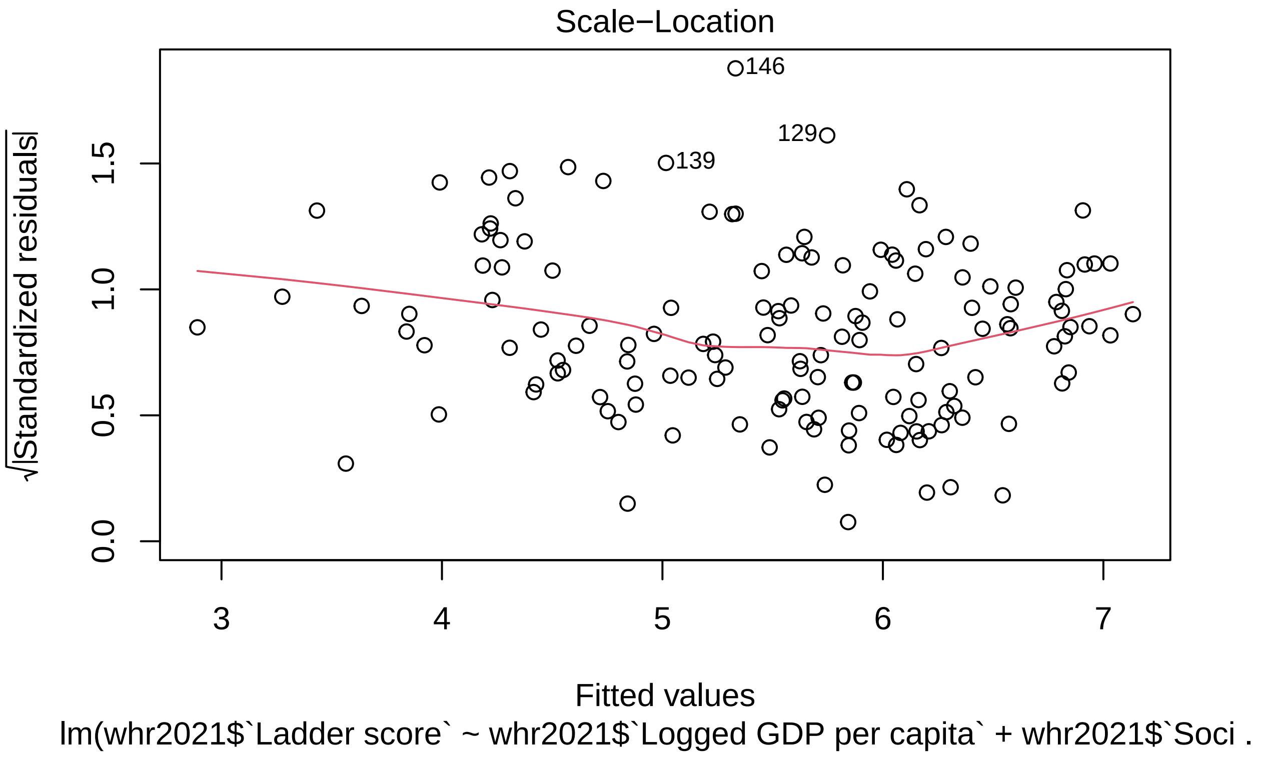
```
fit.multi = lm(formula=whr2021$`Ladder score`~whr2021$`Logged GDP per capita`+whr2021$`Social support`+
summary(fit.multi)
```

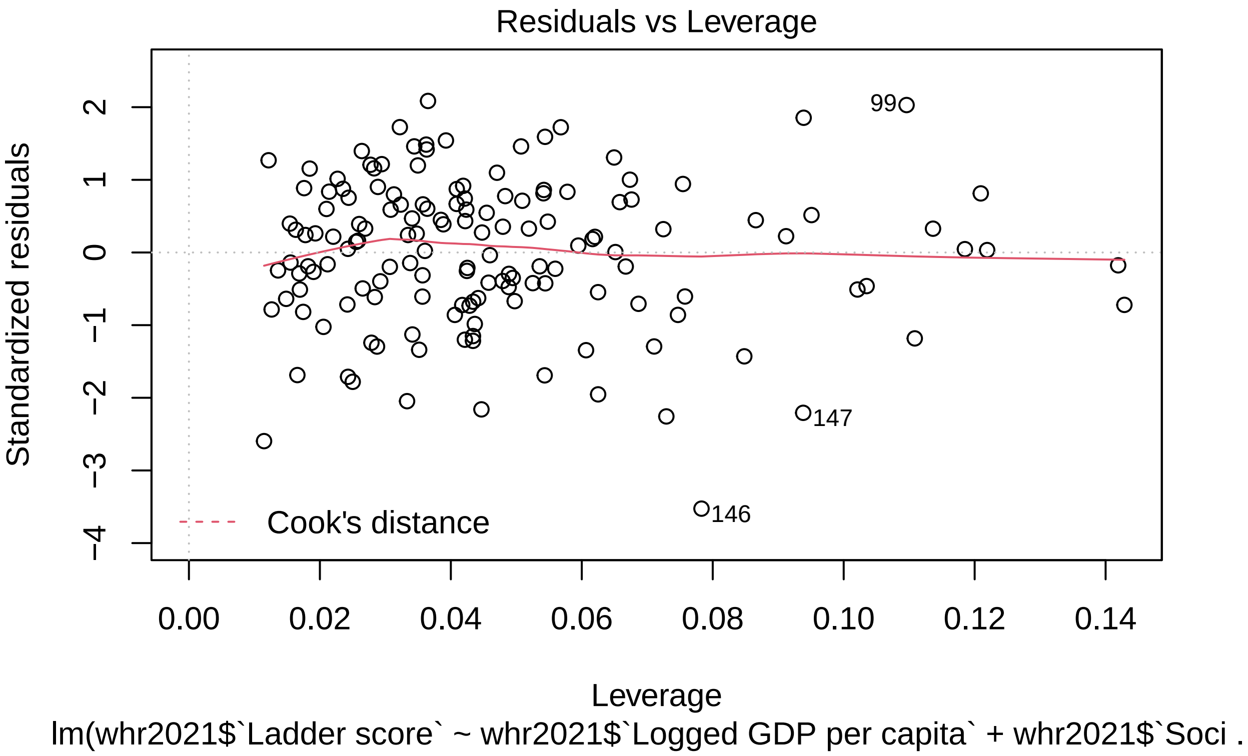
```
## 
## Call:
## lm(formula = whr2021$‘Ladder score‘ ~ whr2021$‘Logged GDP per capita‘ + 
##     whr2021$‘Social support‘ + whr2021$‘Healthy life expectancy‘ +
##     whr2021$‘Freedom to make life choices‘ + whr2021$‘Income Gini‘ +
##     whr2021$‘Wealth Gini‘, data = whr2021)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.86501 -0.33337  0.05079  0.38774  1.12796
##
## Coefficients:
##                                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)                            -2.28228    0.87070  -2.621  0.00972 **
## whr2021$‘Logged GDP per capita‘         0.29106    0.08721   3.338  0.00108 **
## whr2021$‘Social support‘                2.09146    0.66663   3.137  0.00207 **
## whr2021$‘Healthy life expectancy‘       0.02915    0.01395   2.089  0.03851 *
## whr2021$‘Freedom to make life choices‘  2.56419    0.46953   5.461 2.06e-07 ***
## whr2021$‘Income Gini‘                  -0.73233    0.77888  -0.940  0.34869
## whr2021$‘Wealth Gini‘                  -0.37103    0.80046  -0.464  0.64370
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.551 on 142 degrees of freedom
## Multiple R-squared:  0.7474, Adjusted R-squared:  0.7368
## F-statistic: 70.04 on 6 and 142 DF,  p-value: < 2.2e-16
```

```
plot(fit.multi)
```




Normal Q−Q
Standardized residuals
2
1
0
−1
−2
−3
−2
−1
0
1
2
139
129
146
Theoretical Quantiles
lm(whr2021$`Ladder score` ~ whr2021$`Logged GDP per capita` + whr2021$`Soci .


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
3
4
5
6
7
146
129
139
Fitted values
lm(whr2021$`Ladder score` ~ whr2021$`Logged GDP per capita` + whr2021$`Soci .



Kovarijacijska matrica

Regresija s jako koreliranim ulaznim varijablama će uglavnom dati neke rezultate, ali na temelju njih ne možemo donositi nikakve zaključke. U slučaju savršene linearne zavisnosti ili koreliranosti ulaznih varijabli, procjena regresijskog modela će biti nestabilna i barem jedan koeficijent će biti NA. Stoga je potrebno odabrati onaj podskup varijabli za koje smatramo da objašnjavaju različite efekte u podatcima i nisu međusobno (previše) korelirane.

Gledat ćemo matricu Pearsonovih koeficijenata korelacije, a ne kovarijacijsku matricu jer podatci u varijablama nisu na jednakim skalama. UMETNUTI OVDJE FORMULU PEARSONOVIH KOEFICIJENATA AKO JE OVAJ DIO UOPCE BITAN Ako je koeficijent korelacije = 1, vektori varijabli leže na istom pravcu i imaju istu orijentaciju, a ako koeficijent korelacije = -1, suprotnu orijentaciju. Ako je koeficijent korelacije = 0, vektori su neovisni.

NAPOMENA: Pogledati treba li ova kovarijacijska matrica uopće??

```
M <- cbind(whr2021$`Ladder score`,whr2021$`Logged GDP per capita`,whr2021$`Social support`,whr2021$`Hea
cor(M)
```

```
##             [,1]      [,2]      [,3]      [,4]       [,5]        [,6]
## [1,] 1.00000000 0.7897476 0.7569675 0.7680603 0.60773272 -0.01776866
## [2,] 0.78974762 1.0000000 0.7852911 0.8594858 0.43231011 -0.19938590
## [3,] 0.75696748 0.7852911 1.0000000 0.7232477 0.48307259 -0.11496782
## [4,]  0.76806029  0.8594858  0.7232477  1.0000000  0.46136792 -0.16181509
## [5,]  0.60773272  0.4323101  0.4830726  0.4613679  1.00000000  0.16945077
## [6,] -0.01776866 -0.1993859 -0.1149678 -0.1618151  0.16945077  1.00000000
## [7,] -0.42097416 -0.3422316 -0.2034292 -0.3642919 -0.40103339 -0.16389962
## [8,] -0.35879747 -0.3714221 -0.3277160 -0.3858341 -0.13904290 -0.02564794
## [9,] -0.31219754 -0.3133797 -0.3118741 -0.3959141 -0.08324258  0.04599525
##            [,7]        [,8]        [,9]
## [1,] -0.4209742 -0.35879747 -0.31219754
## [2,] -0.3422316 -0.37142211 -0.31337971
## [3,] -0.2034292 -0.32771597 -0.31187410
## [4,] -0.3642919 -0.38583413 -0.39591408
## [5,] -0.4010334 -0.13904290 -0.08324258
## [6,] -0.1638996 -0.02564794  0.04599525
## [7,]  1.0000000  0.23986922  0.07718160
## [8,]  0.2398692  1.00000000  0.51824277
## [9,]  0.0771816  0.51824277  1.00000000
```

Primijetimo da varijable Logged GDP per capita, Social support i Healthy life expectancy imaju koeficijent korelacije veći od 0.75 s varijablom Ladder score - razinom sreće. Pogledajmo to malo detaljnije, kao i njihove međusobne koeficijente korelacije:

```
cor.test(whr2021$`Ladder score`,whr2021$`Logged GDP per capita`)
```

```
##
##  Pearson’s product-moment correlation
##
## data:  whr2021$‘Ladder score‘ and whr2021$‘Logged GDP per capita‘
## t = 15.609, df = 147, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7204370 0.8434382
## sample estimates:
##       cor
## 0.7897476
```

```
cor.test(whr2021$`Ladder score`,whr2021$`Social support`)
```

```
##
##  Pearson’s product-moment correlation
##
## data:  whr2021$‘Ladder score‘ and whr2021$‘Social support‘
## t = 14.045, df = 147, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6787901 0.8181785
## sample estimates:
##       cor
## 0.7569675
```

```
cor.test(whr2021$`Ladder score`,whr2021$`Healthy life expectancy`)
```

```
##
```

```
##  Pearson’s product-moment correlation
##
## data:  whr2021$‘Ladder score‘ and whr2021$‘Healthy life expectancy‘
## t = 14.542, df = 147, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6928273 0.8267529
## sample estimates:
##       cor
## 0.7680603
```

```
#2.Logged GDP per capita vs. 3.Social Support
cor.test(whr2021$`Logged GDP per capita`,whr2021$`Social support`)
```

```
##
##  Pearson’s product-moment correlation
##
## data:  whr2021$‘Logged GDP per capita‘ and whr2021$‘Social support‘
## t = 15.378, df = 147, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7147454 0.8400180
## sample estimates:
##       cor
## 0.7852911
```

```
#2.Logged GDP per capita vs. 4.Healthy life expectancy
cor.test(whr2021$`Logged GDP per capita`,whr2021$`Healthy life expectancy`)
```

```
##
##  Pearson’s product-moment correlation
##
## data:  whr2021$‘Logged GDP per capita‘ and whr2021$‘Healthy life expectancy‘
## t = 20.386, df = 147, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8107335 0.8963987
## sample estimates:
##       cor
## 0.8594858
```

```
#3.Social support vs. 4.Healthy life expectancy
cor.test(whr2021$`Social support`,whr2021$`Healthy life expectancy`)
```

```
##
##  Pearson’s product-moment correlation
##
## data:  whr2021$‘Social support‘ and whr2021$‘Healthy life expectancy‘
## t = 12.698, df = 147, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6364678 0.7919458
## sample estimates:
##       cor
## 0.7232477
```

Vidimo da varijable Logged GDP per capita, Social support i Healthy life expectancy međusobno imaju visok koeficijent korelacije, dakle za pretpostaviti je da sigurno treba izbaciti dvije, a ostaviti jednu. Također, varijable Social support i Healthy life expectancy imaju jako velik koeficijent korelacije s varijablom GDP per capita,a manji međusobno, pa bi pretpostavka bila da možda izbacimo GDP, a gledamo samo te druge dvije varijable. No, s obzirom da nam je GDP najbolje modelirao razinu sreće, pretpostavljamo da će naš model ipak biti bolji s uključenom varijablom GDP.

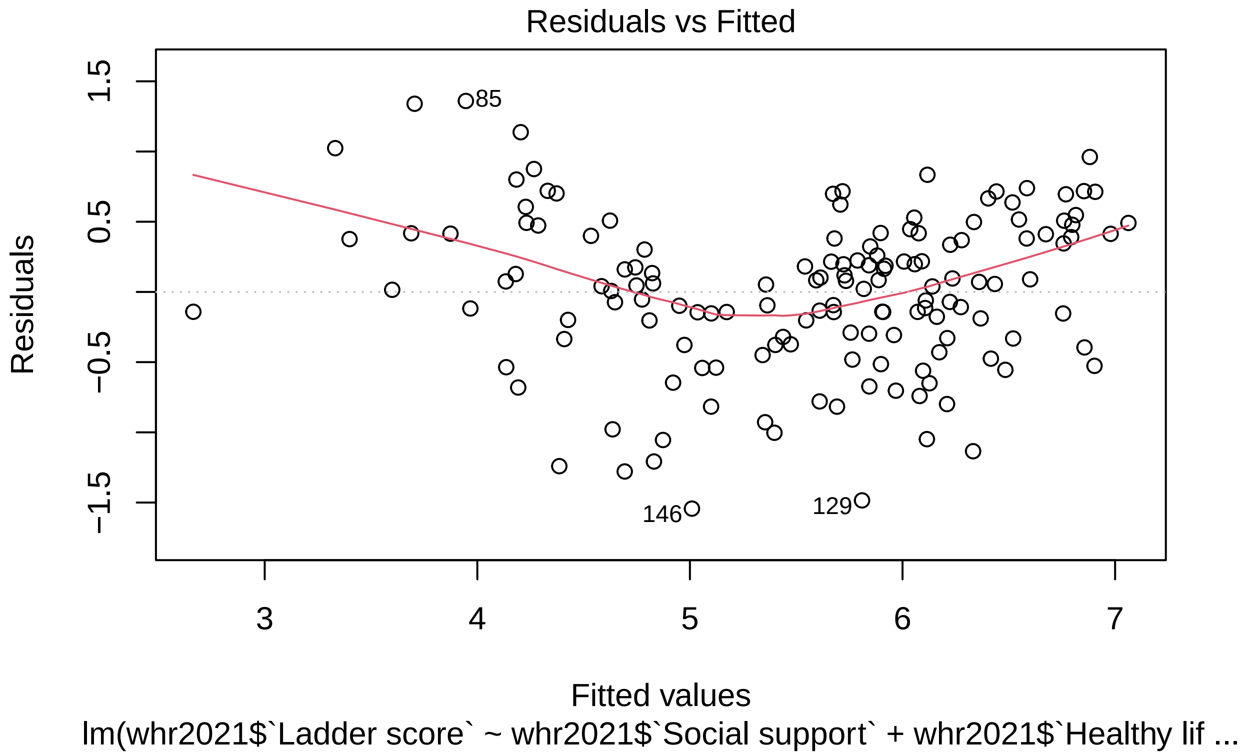
Pogledajmo kako izgleda regresija bez varijable GDP.

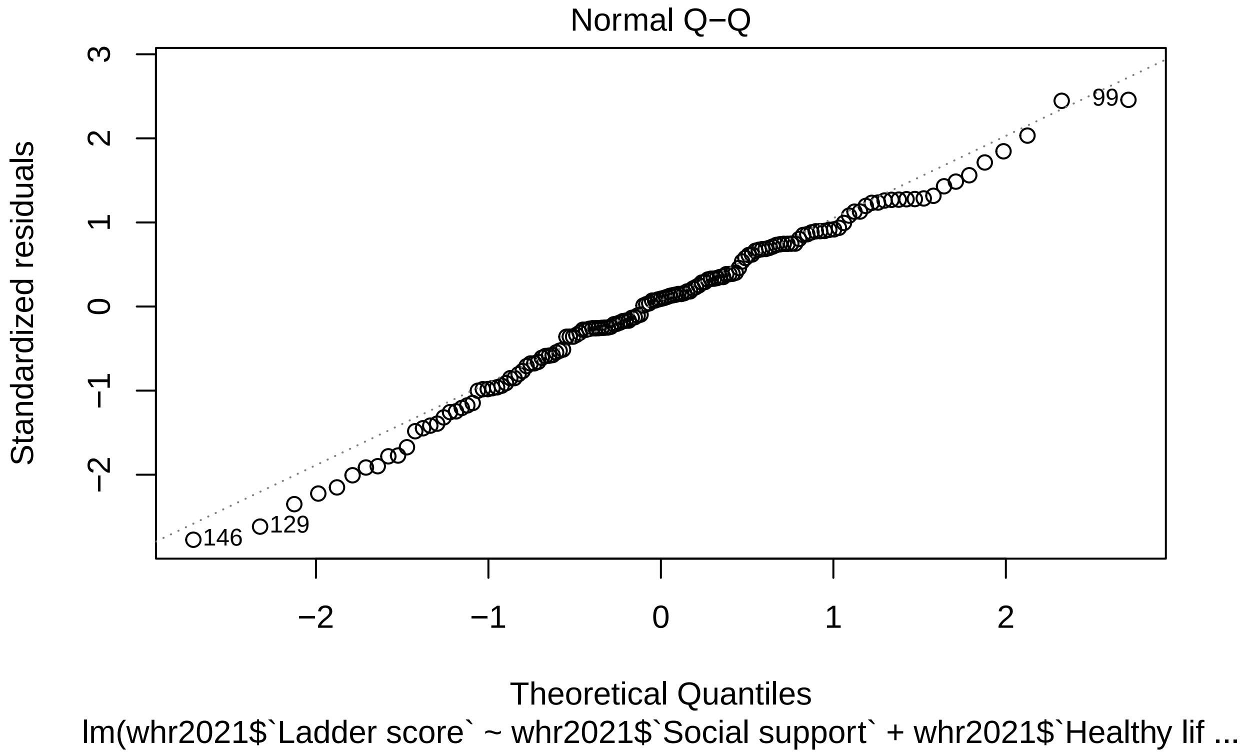
Višestruka regresija bez varijable GDP per capita

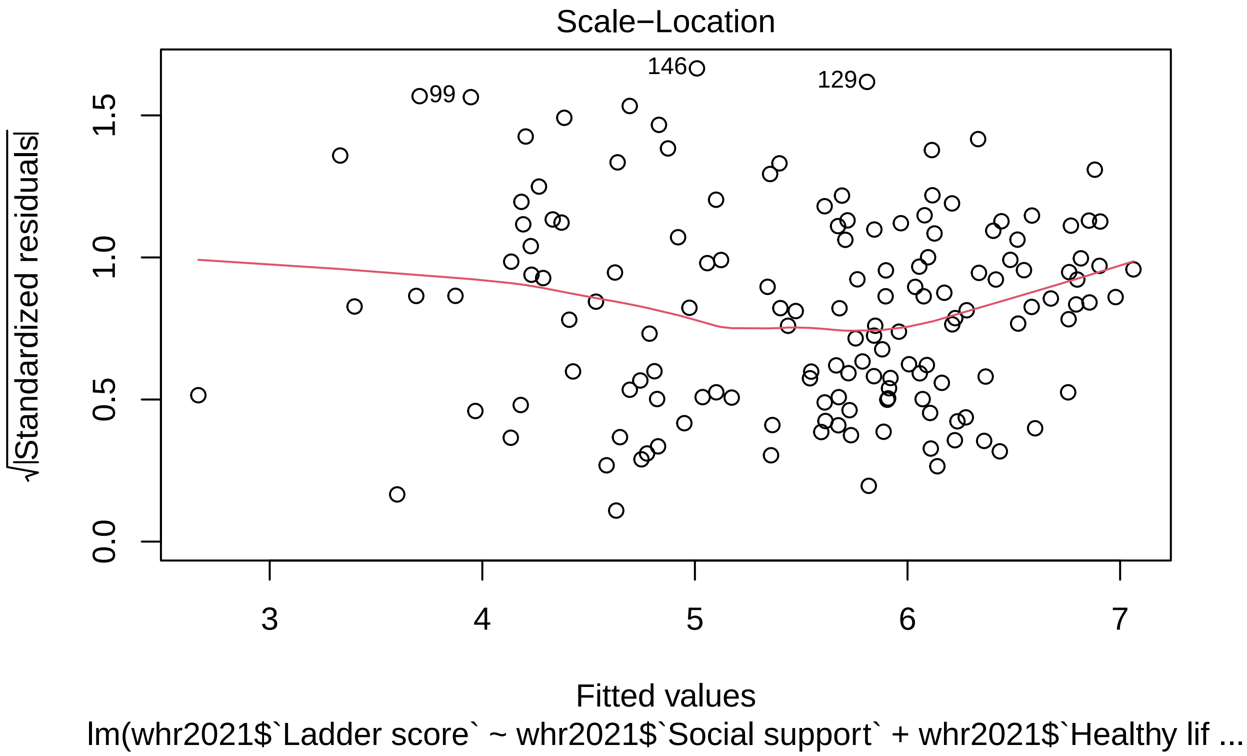
```
fit.nogdp = lm(formula=whr2021$`Ladder score`~whr2021$`Social support`+whr2021$`Healthy life expectancy
#without GDP
summary(fit.nogdp)
```

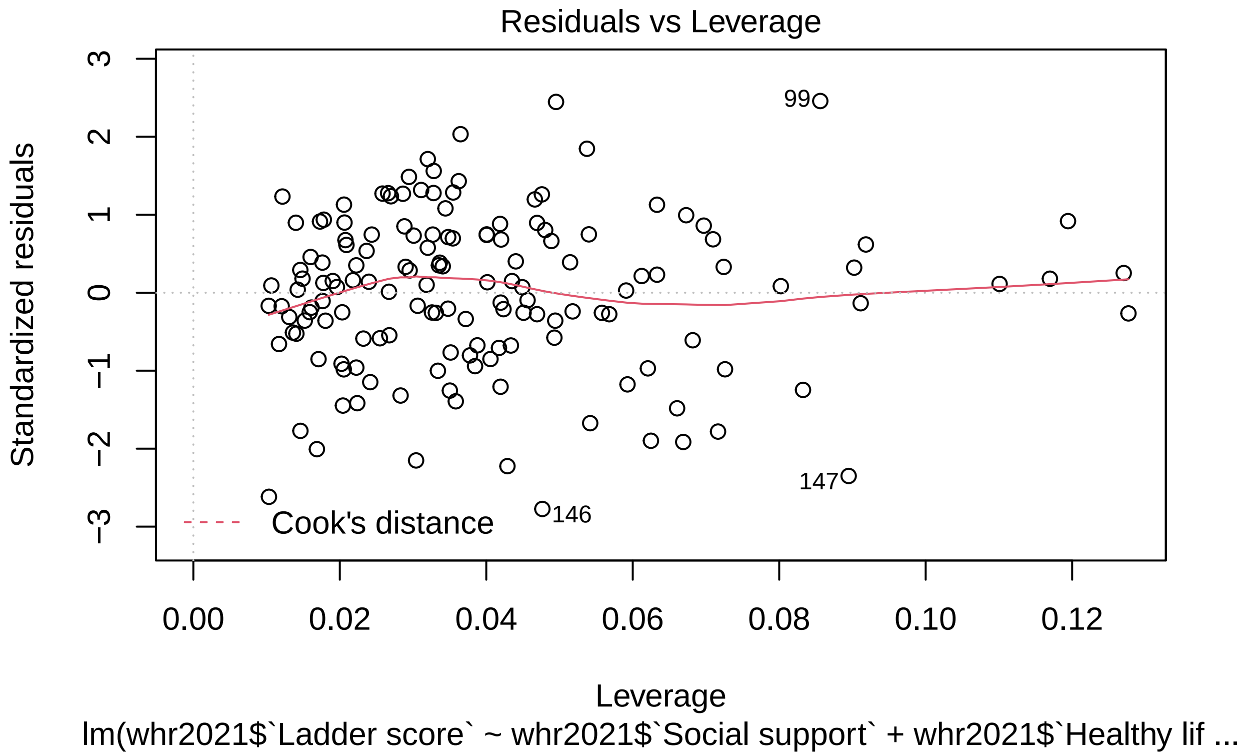
```
##
## Call:
## lm(formula = whr2021$‘Ladder score‘ ~ whr2021$‘Social support‘ +
##     whr2021$‘Healthy life expectancy‘ + whr2021$‘Freedom to make life choices‘ +
##     whr2021$‘Income Gini‘ + whr2021$‘Wealth Gini‘, data = whr2021)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.54319 -0.33172  0.05234  0.41099  1.35996
##
## Coefficients:
##                                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)                            -2.49788    0.89856  -2.780  0.00617 **
## whr2021$‘Social support‘                3.11790    0.61206   5.094 1.09e-06 ***
## whr2021$‘Healthy life expectancy‘       0.06032    0.01072   5.625 9.45e-08 ***
## whr2021$‘Freedom to make life choices‘  2.48388    0.48525   5.119 9.75e-07 ***
## whr2021$‘Income Gini‘                  -0.99809    0.80180  -1.245  0.21524
## whr2021$‘Wealth Gini‘                  -0.03201    0.82166  -0.039  0.96898
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.5702 on 143 degrees of freedom
## Multiple R-squared:  0.7276, Adjusted R-squared:  0.7181
## F-statistic:  76.4 on 5 and 143 DF,  p-value: < 2.2e-16
```

```
plot(fit.nogdp)
```


Residuals vs Fitted
Residuals
1.5
0.5
−0.5
−1.5
3
4
5
6
7
85
146
129
Fitted values
lm(whr2021$`Ladder score` ~ whr2021$`Social support` + whr2021$`Healthy lif ...


Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
−2
−1
0
1
2
99
129
146
Theoretical Quantiles
lm(whr2021$`Ladder score` ~ whr2021$`Social support` + whr2021$`Healthy lif ...


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
3
4
5
6
7
99
146
129
Fitted values
lm(whr2021$`Ladder score` ~ whr2021$`Social support` + whr2021$`Healthy lif ...



Možemo primijetiti da je R2 ipak veći kada je varijabla GDP uključena. U nastavku ćemo još vidjeti možemo li transformirati podatke da dobijemo bolji model.

Transformacije podataka, dodavanje interakcijskih članova

Promatranjem scatter plotova s početka, zaključujemo da bismo mogli probati modificirati varijable Income Gini i Social support, dodavanjem kvadrata tih ulaznih varijabli. Potencijalno, s obzirom na sličnost grafova GDP i HLE s grafom varijable Social Support, naknadno ćemo ispitati ima li njihova transformacija utjecaja na R2 modela.

Pogledajmo sada kako izgleda model predikcije razine sreće transformiranom varijablom Social Support.

```
# moguce je provjeriti gore navedenu tvrdnju prvo na primjeru samo Social Support
fit.ssupp.sq = lm(formula=whr2021$`Ladder score`~ whr2021$`Social support` + I(whr2021$`Social support`
summary(fit.ssupp.sq)
```

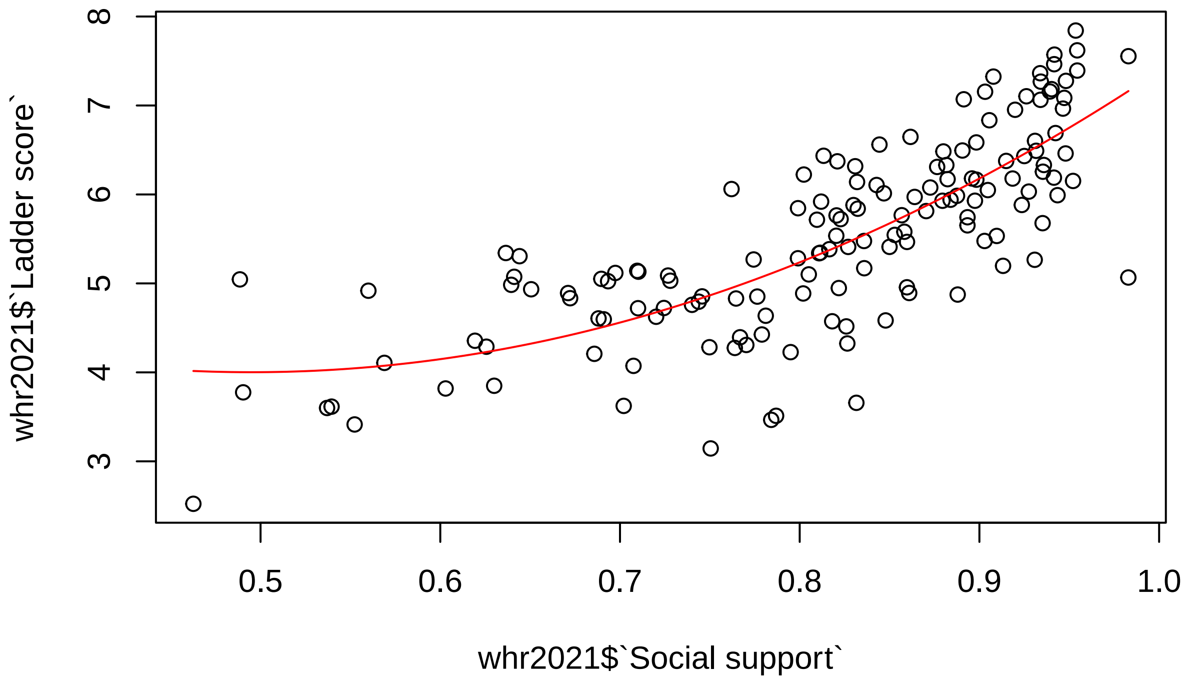
```
##
## Call:
## lm(formula = whr2021$‘Ladder score‘ ~ whr2021$‘Social support‘ +
##     I(whr2021$‘Social support‘^2), data = whr2021)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.09453 -0.39457  0.02724  0.51420  1.11166
##
## Coefficients:
```

```
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      7.252      1.989   3.647 0.000369 ***
## whr2021$`Social support`       -13.134      5.285  -2.485 0.014074 *
## I(whr2021$`Social support`^2)   13.268      3.455   3.840 0.000183 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6734 on 146 degrees of freedom
## Multiple R-squared:  0.6122, Adjusted R-squared:  0.6069
## F-statistic: 115.2 on 2 and 146 DF,  p-value: < 2.2e-16
```

Vidimo da je R2 ovakvog modela veći od R2 linearnog modela na netransformiranim podatcima.

Prikažimo grafički tu nelinearnu krivulju:

```
f = function(x, coeffs)
return(coeffs[[1]] + coeffs[[2]] * x + coeffs[[3]] * x^2)
plot(whr2021$`Social support`,whr2021$`Ladder score`)
curve(f(x, fit.ssupp.sq$coefficients), add = TRUE, col = "red")
```

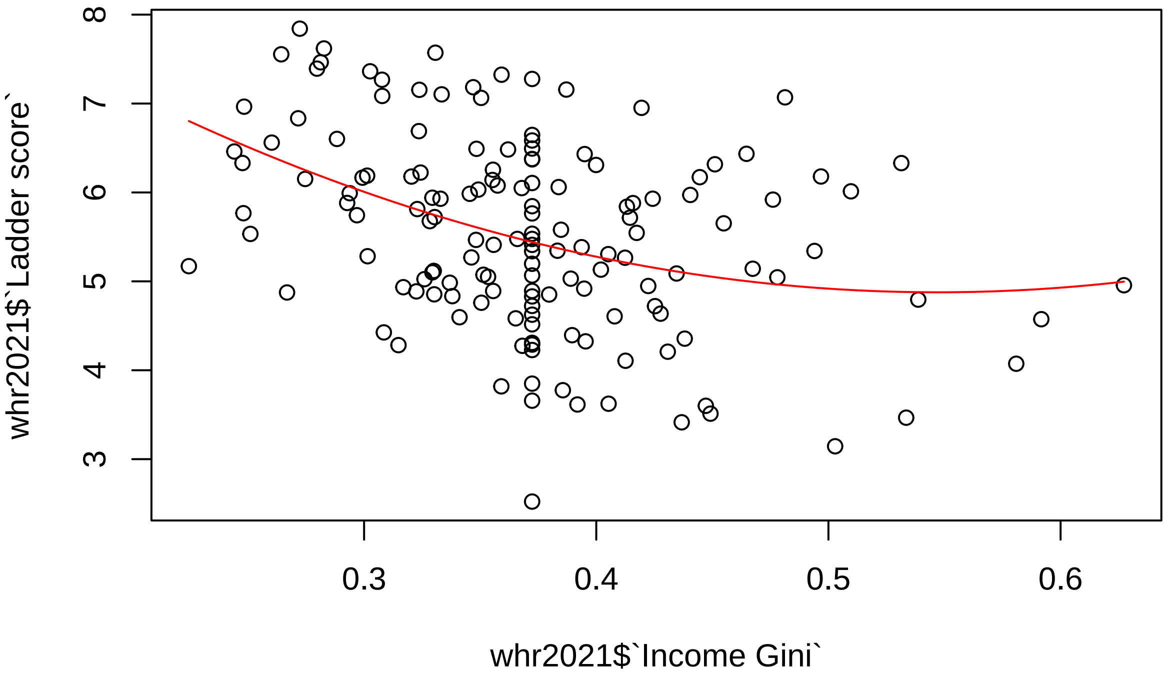


Pogledajmo sada za Income Gini:

```
# moguce je provjeriti gore navedenu tvrdnju prvo na primjeru samo Social Support
fit.inc.sq = lm(formula=whr2021$`Ladder score`~ whr2021$`Income Gini` + I(whr2021$`Income Gini`^2), data
summary(fit.inc.sq)
```

```
##
## Call:
## lm(formula = whr2021$‘Ladder score‘ ~ whr2021$‘Income Gini‘ +
##     I(whr2021$‘Income Gini‘^2), data = whr2021)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.91878 -0.74117 -0.02322  0.69929  2.11252
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    10.419      1.614   6.456 1.49e-09 ***
## whr2021$‘Income Gini‘         -20.263      8.145  -2.488   0.0140 *
## I(whr2021$‘Income Gini‘^2)     18.522     10.097   1.834   0.0686 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.9978 on 146 degrees of freedom
## Multiple R-squared:  0.1484, Adjusted R-squared:  0.1367
## F-statistic: 12.72 on 2 and 146 DF,  p-value: 8.1e-06
```

```
plot(whr2021$`Income Gini`,whr2021$`Ladder score`)
curve(f(x, fit.inc.sq$coefficients), add = TRUE, col = "red")
```

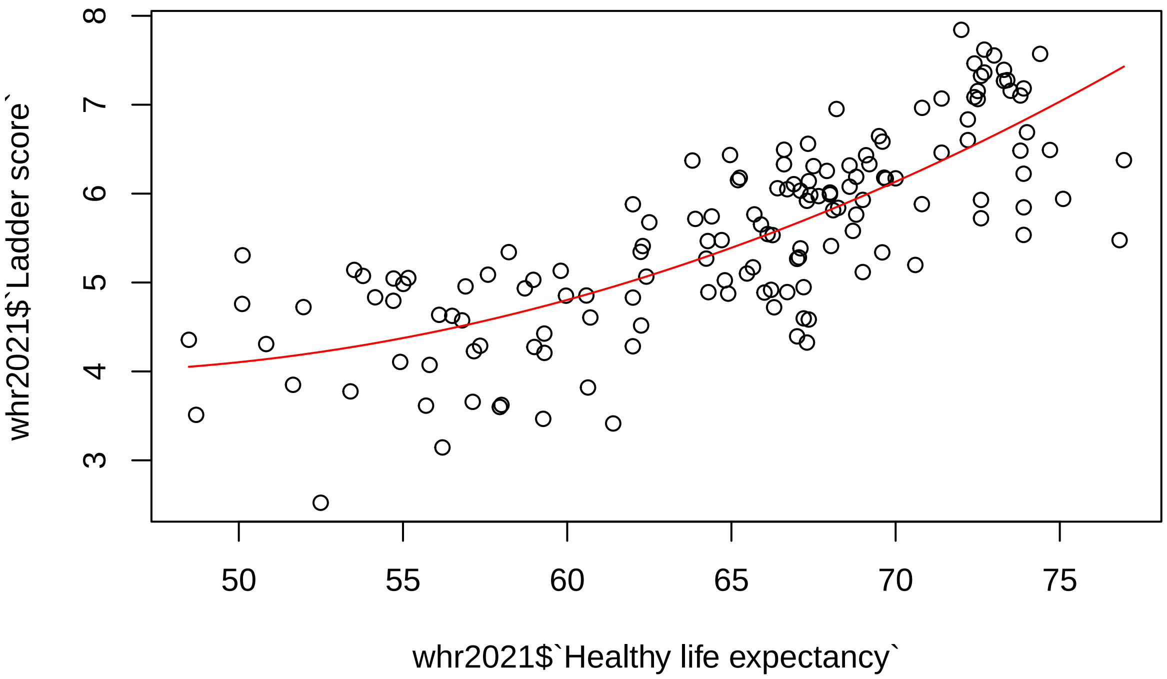


Model predikcije sreće s transformiranom varijablom Healthy life expectancy:

```r
# moguce je provjeriti gore navedenu tvrdnju prvo na primjeru samo Social Support
fit.hle.sq = lm(formula=whr2021$`Ladder score`~ whr2021$`Healthy life expectancy` + I(whr2021$`Healthy l
summary(fit.hle.sq)
```

```
##
## Call:
## lm(formula = whr2021$‘Ladder score‘ ~ whr2021$‘Healthy life expectancy‘ +
##     I(whr2021$‘Healthy life expectancy‘^2), data = whr2021)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9255 -0.4126  0.1375  0.5104  1.3662
##
## Coefficients:
##                                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)                            10.036170   4.523444   2.219  0.02805 *
## whr2021$‘Healthy life expectancy‘      -0.275909   0.143997  -1.916  0.05731 .
## I(whr2021$‘Healthy life expectancy‘^2)  0.003145   0.001136   2.768  0.00638 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.6749 on 146 degrees of freedom
## Multiple R-squared:  0.6104, Adjusted R-squared:  0.605
## F-statistic: 114.4 on 2 and 146 DF,  p-value: < 2.2e-16
```

```r
plot(whr2021$`Healthy life expectancy`,whr2021$`Ladder score`)
curve(f(x, fit.hle.sq$coefficients), add = TRUE, col = "red")
```

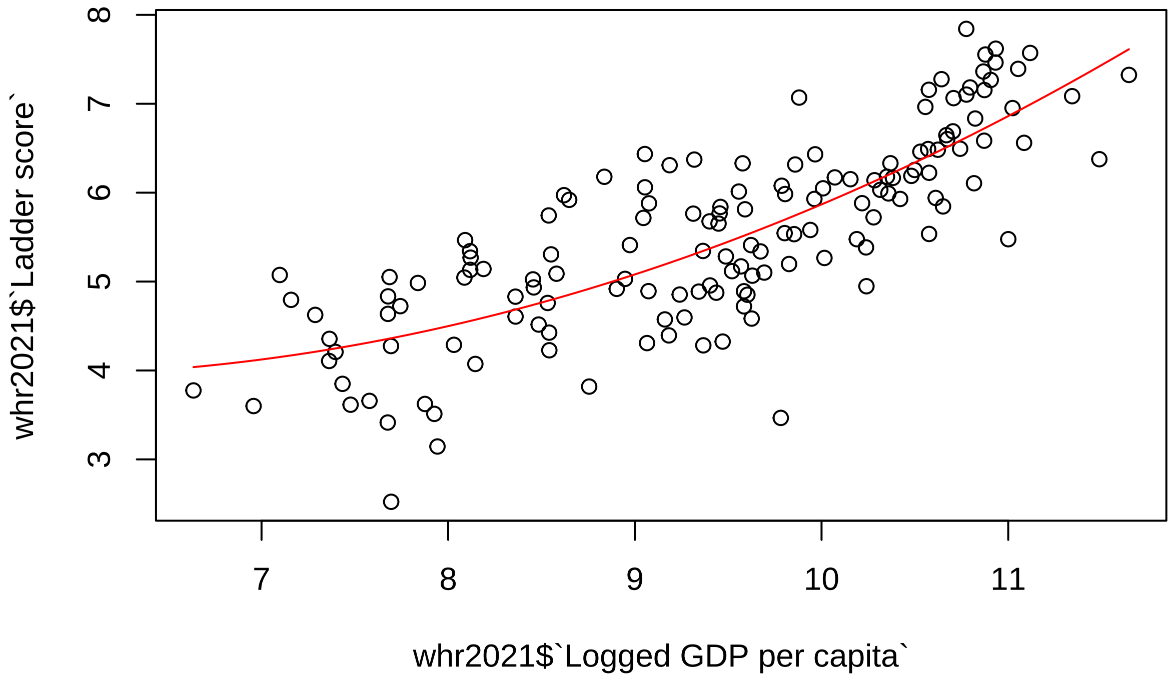


Model predikcije sreće s transformiranom varijablom Logged GDP per capita:

```
# moguce je provjeriti gore navedenu tvrdnju prvo na primjeru samo Social Support
fit.gdp.sq = lm(formula=whr2021$`Ladder score`~ whr2021$`Logged GDP per capita` + I(whr2021$`Logged GDP
summary(fit.gdp.sq)
```

```
##
## Call:
## lm(formula = whr2021$‘Ladder score‘ ~ whr2021$‘Logged GDP per capita‘ +
##     I(whr2021$‘Logged GDP per capita‘^2), data = whr2021)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.21176 -0.41625 -0.01271  0.50384  1.31724
##
## Coefficients:
##                                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)                             7.26442    3.28557   2.211  0.02859 *
## whr2021$‘Logged GDP per capita‘       -1.16978    0.71857  -1.628  0.10570
## I(whr2021$‘Logged GDP per capita‘^2)  0.10301    0.03884   2.652  0.00888 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6479 on 146 degrees of freedom
## Multiple R-squared:  0.641,  Adjusted R-squared:  0.6361
## F-statistic: 130.3 on 2 and 146 DF,  p-value: < 2.2e-16
```

```
plot(whr2021$`Logged GDP per capita`,whr2021$`Ladder score`)
curve(f(x, fit.gdp.sq$coefficients), add = TRUE, col = "red")
```



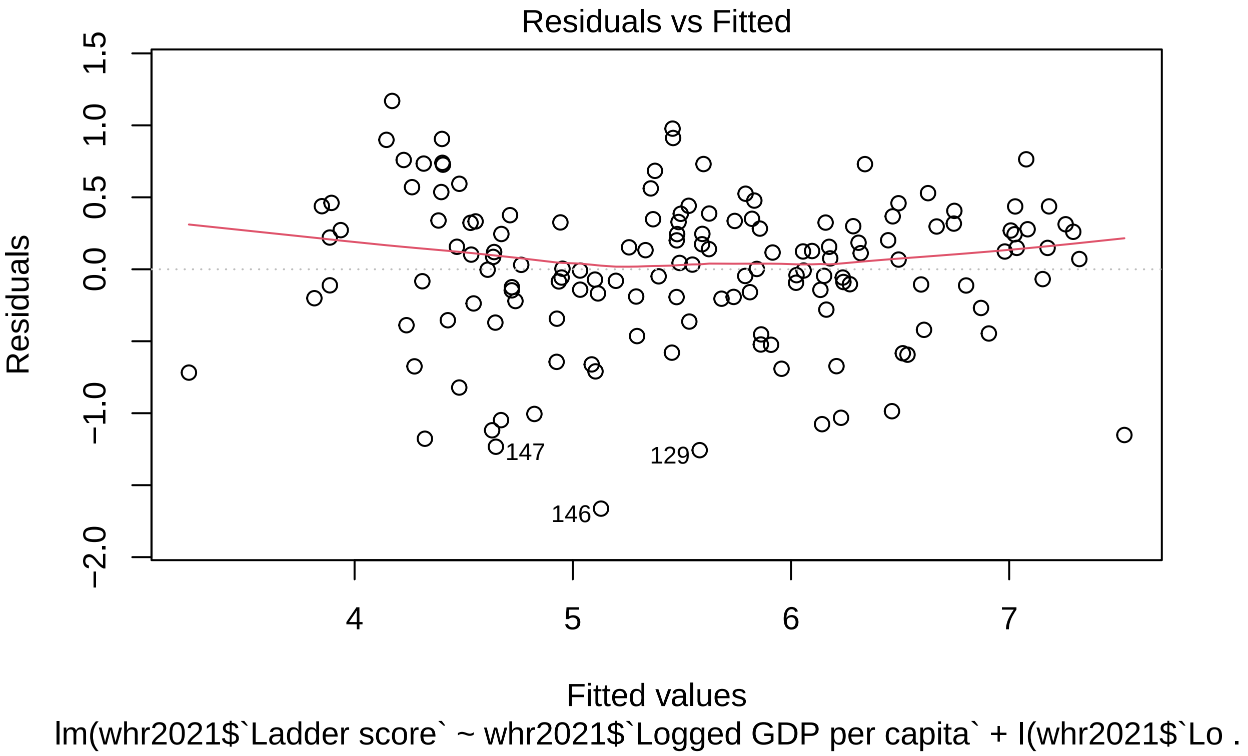
Uključivanjem ovako transformiranih varijabli moguće je dodatno poboljšati ukupni model višestruke regresije. Budući da vidimo da je, nakon kvadratne transformacije varijabli GDP, Social Support, Healthy life expectancy i Income Gini, postignuto povećanje R2 u njihovim modelima razine sreće, uključit ćemo ih u višestruku regresiju.

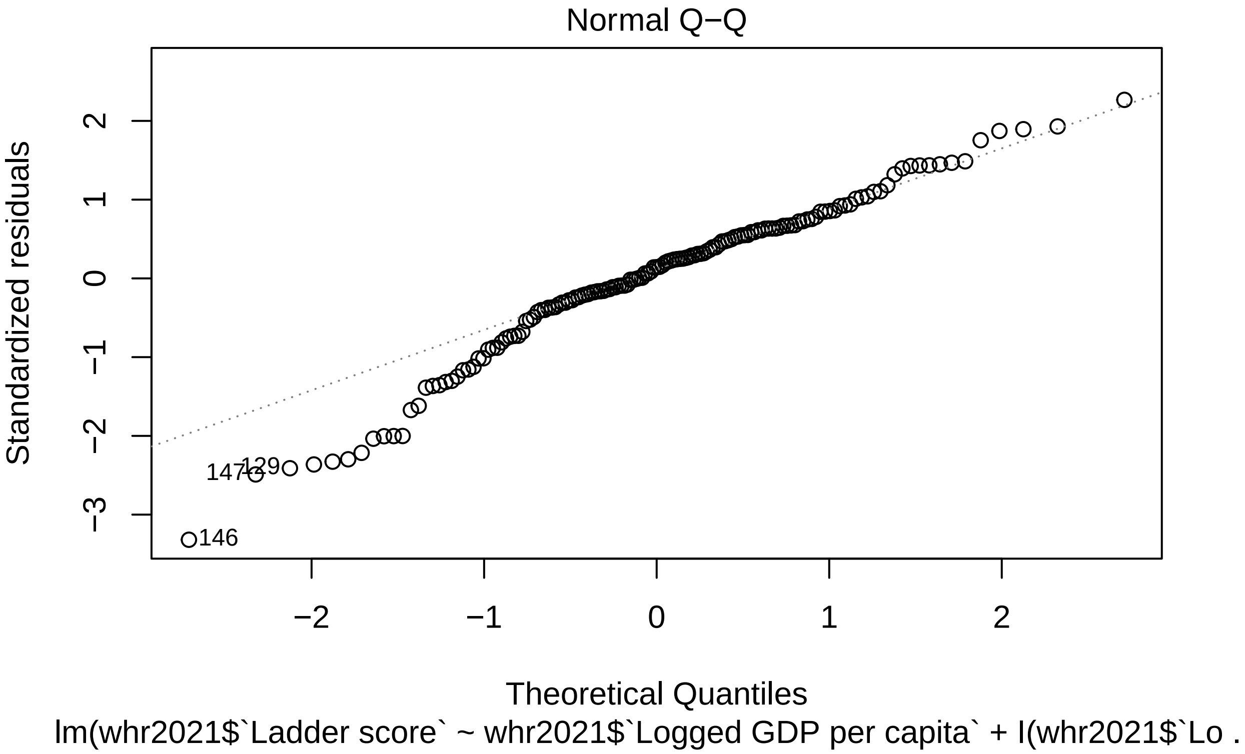
```
fit.multi.sq = lm(formula=whr2021$`Ladder score`~whr2021$`Logged GDP per capita`+I(whr2021$`Logged GDP
summary(fit.multi.sq)
```

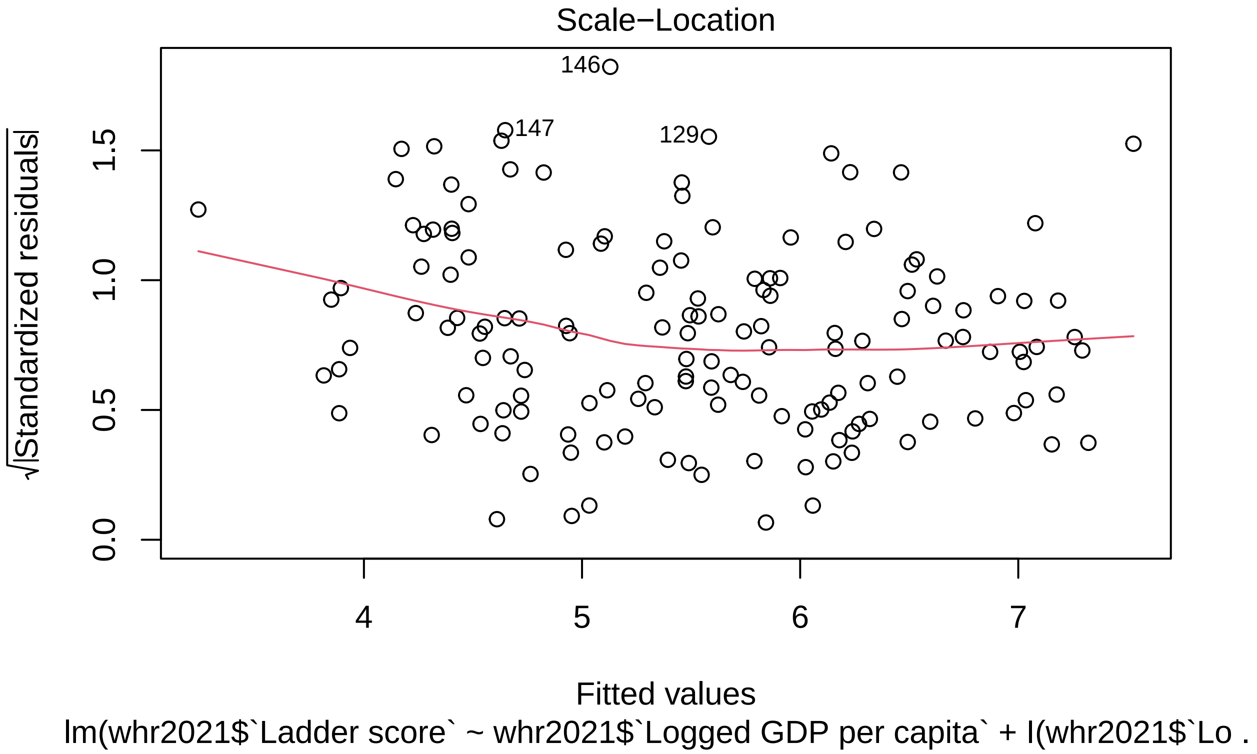
```
##
## Call:
## lm(formula = whr2021$‘Ladder score‘ ~ whr2021$‘Logged GDP per capita‘ +
##     I(whr2021$‘Logged GDP per capita‘^2) + whr2021$‘Social support‘ +
##     I(whr2021$‘Social support‘^2) + whr2021$‘Healthy life expectancy‘ +
##     I(whr2021$‘Healthy life expectancy‘^2) + whr2021$‘Freedom to make life choices‘ +
##     whr2021$‘Income Gini‘ + I(whr2021$‘Income Gini‘^2) + whr2021$‘Wealth Gini‘,
##     data = whr2021)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.66299 -0.20454  0.07102  0.32546  1.16930
##
## Coefficients:
```

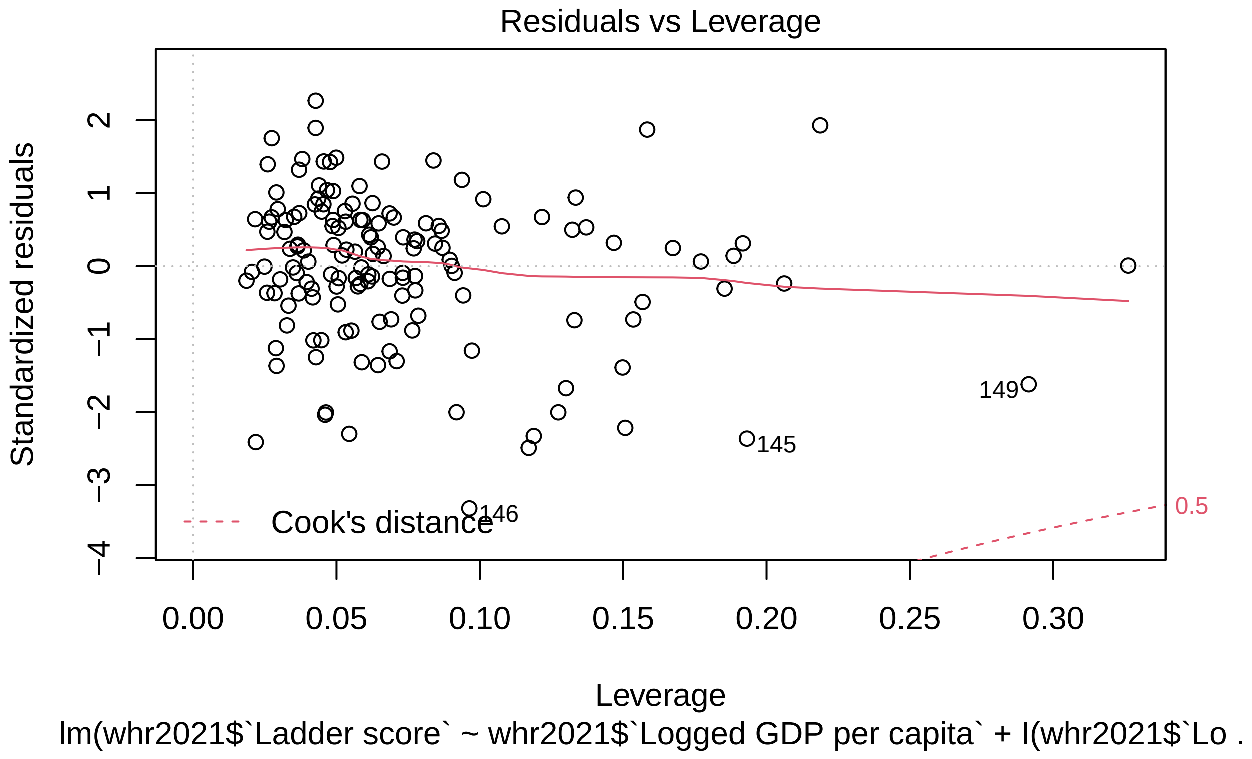
```
##                                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)                            11.798625   3.969721   2.972  0.00349
## whr2021$‘Logged GDP per capita‘        -0.892542   0.762502  -1.171  0.24380
## I(whr2021$‘Logged GDP per capita‘^2)    0.060083   0.041739   1.439  0.15227
## whr2021$‘Social support‘               -3.878645   4.727423  -0.820  0.41337
## I(whr2021$‘Social support‘^2)           4.361640   3.145884   1.386  0.16784
## whr2021$‘Healthy life expectancy‘      -0.203877   0.140733  -1.449  0.14970
## I(whr2021$‘Healthy life expectancy‘^2)  0.001879   0.001137   1.653  0.10060
## whr2021$‘Freedom to make life choices‘  2.499563   0.455178   5.491 1.86e-07
## whr2021$‘Income Gini‘                   0.700427   4.720163   0.148  0.88225
## I(whr2021$‘Income Gini‘^2)             -1.580214   5.740703  -0.275  0.78352
## whr2021$‘Wealth Gini‘                   0.193933   0.800238   0.242  0.80887
##
## (Intercept)                            **
## whr2021$‘Logged GDP per capita‘
## I(whr2021$‘Logged GDP per capita‘^2)
## whr2021$‘Social support‘
## I(whr2021$‘Social support‘^2)
## whr2021$‘Healthy life expectancy‘
## I(whr2021$‘Healthy life expectancy‘^2)
## whr2021$‘Freedom to make life choices‘ ***
## whr2021$‘Income Gini‘
## I(whr2021$‘Income Gini‘^2)
## whr2021$‘Wealth Gini‘
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.5271 on 138 degrees of freedom
## Multiple R-squared:  0.7754, Adjusted R-squared:  0.7591
## F-statistic: 47.64 on 10 and 138 DF,  p-value: < 2.2e-16
```

```
plot(fit.multi.sq)
```


Residuals vs Fitted
Residuals
1.5
1.0
0.5
0.0
−1.0
−2.0
4
5
6
7
Fitted values
lm(whr2021$`Ladder score` ~ whr2021$`Logged GDP per capita` + I(whr2021$`Lo .
147
129
146


Normal Q−Q
Standardized residuals
2
1
0
−1
−2
−3
147
129
146
−2
−1
0
1
2
Theoretical Quantiles
lm(whr2021$`Ladder score` ~ whr2021$`Logged GDP per capita` + I(whr2021$`Lo .


Scale−Location
146
147
129
√|Standardized residuals|
0.0
0.5
1.0
1.5
4
5
6
7
Fitted values
lm(whr2021$`Ladder score` ~ whr2021$`Logged GDP per capita` + I(whr2021$`Lo .



Pogledajmo kako izgleda model s varijablama Social Support i HLE, ali bez varijable GDP.

```
fit.sh.sq = lm(formula=whr2021$`Ladder score`~whr2021$`Social support`+I(whr2021$`Social support`^2)+wh
#bez GDP
summary(fit.sh.sq)
```

```
##
## Call:
## lm(formula = whr2021$`Ladder score` ~ whr2021$`Social support` +
##     I(whr2021$`Social support`^2) + whr2021$`Healthy life expectancy` +
##     I(whr2021$`Healthy life expectancy`^2) + whr2021$`Freedom to make life choices` +
##     whr2021$`Income Gini` + I(whr2021$`Income Gini`^2) + whr2021$`Wealth Gini`,
##     data = whr2021)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.43637 -0.25181  0.07338  0.32870  1.17997
##
## Coefficients:
##                                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           13.7127850  3.8138031   3.596 0.000448
## whr2021$`Social support`              -6.3339448  4.4603972  -1.420 0.157819
## I(whr2021$`Social support`^2)          6.3610698  2.9824759   2.133 0.034684
## whr2021$`Healthy life expectancy`     -0.3660137  0.1191494  -3.072 0.002557
## I(whr2021$`Healthy life expectancy`^2)  0.0033555  0.0009421   3.562 0.000504
## whr2021$`Freedom to make life choices`  2.5169170  0.4602100   5.469 2.02e-07
## whr2021$‘Income Gini‘                  -1.1982546  4.7366039  -0.253 0.800657
## I(whr2021$‘Income Gini‘^2)              0.3745125  5.7758806   0.065 0.948393
## whr2021$‘Wealth Gini‘                   0.6978076  0.7913722   0.882 0.379413
##
## (Intercept)                            ***
## whr2021$‘Social support‘
## I(whr2021$‘Social support‘^2)          *
## whr2021$‘Healthy life expectancy‘      **
## I(whr2021$‘Healthy life expectancy‘^2) ***
## whr2021$‘Freedom to make life choices‘ ***
## whr2021$‘Income Gini‘
## I(whr2021$‘Income Gini‘^2)
## whr2021$‘Wealth Gini‘
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.5365 on 140 degrees of freedom
## Multiple R-squared:  0.7639, Adjusted R-squared:  0.7504
## F-statistic: 56.62 on 8 and 140 DF,  p-value: < 2.2e-16
```

Kao što smo naslutili, rezultati upućuju na to da varijabla Logged GDP per capita daje dosta korisnu informaciju u modelu, čak i kad koristimo R^2_{adj}. Pogledajmo sada kako izgleda model sa svim varijablama osim Wealth Gini i Income Gini s obzirom da one imaju najmanji utjecaj na model.

```
fit.s.sq = lm(formula=whr2021$`Ladder score`~whr2021$`Logged GDP per capita`+I(whr2021$`Logged GDP per 
#bez GDP
summary(fit.s.sq)
```

```
##
## Call:
## lm(formula = whr2021$‘Ladder score‘ ~ whr2021$‘Logged GDP per capita‘ +
##     I(whr2021$‘Logged GDP per capita‘^2) + whr2021$‘Social support‘ +
##     I(whr2021$‘Social support‘^2) + whr2021$‘Healthy life expectancy‘ +
##     I(whr2021$‘Healthy life expectancy‘^2) + whr2021$‘Freedom to make life choices‘,
##     data = whr2021)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.74445 -0.20810  0.06919  0.33133  1.11759
##
## Coefficients:
##                                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)                            11.967401   3.781248   3.165   0.0019
## whr2021$‘Logged GDP per capita‘        -0.958425   0.749349  -1.279   0.2030
## I(whr2021$‘Logged GDP per capita‘^2)    0.063856   0.040899   1.561   0.1207
## whr2021$‘Social support‘               -4.019484   4.575976  -0.878   0.3812
## I(whr2021$‘Social support‘^2)           4.445226   3.024569   1.470   0.1439
## whr2021$‘Healthy life expectancy‘      -0.193209   0.136449  -1.416   0.1590
## I(whr2021$‘Healthy life expectancy‘^2)  0.001804   0.001096   1.646   0.1019
## whr2021$‘Freedom to make life choices‘  2.476156   0.446162   5.550 1.37e-07
##
## (Intercept)                            **
## whr2021$‘Logged GDP per capita‘
## I(whr2021$‘Logged GDP per capita‘^2)
## whr2021$‘Social support‘
## I(whr2021$‘Social support‘^2)
## whr2021$‘Healthy life expectancy‘
## I(whr2021$‘Healthy life expectancy‘^2)
## whr2021$‘Freedom to make life choices‘ ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.5227 on 141 degrees of freedom
## Multiple R-squared:  0.7743, Adjusted R-squared:  0.7631
## F-statistic:  69.1 on 7 and 141 DF,  p-value: < 2.2e-16
```

Rezultati upućuju na to da varijable Wealth Gini i Income Gini ipak sadrže korisne informacije za predviđanje sreće u modelu.

NAPOMENA: Razmisliti treba li ubaciti kategoriJsku varijablu regija ili ?? DOVRŠITI DO KRAJA ZAKLJUČAK

Zaključak 2. pitanja i dodatnog pitanja

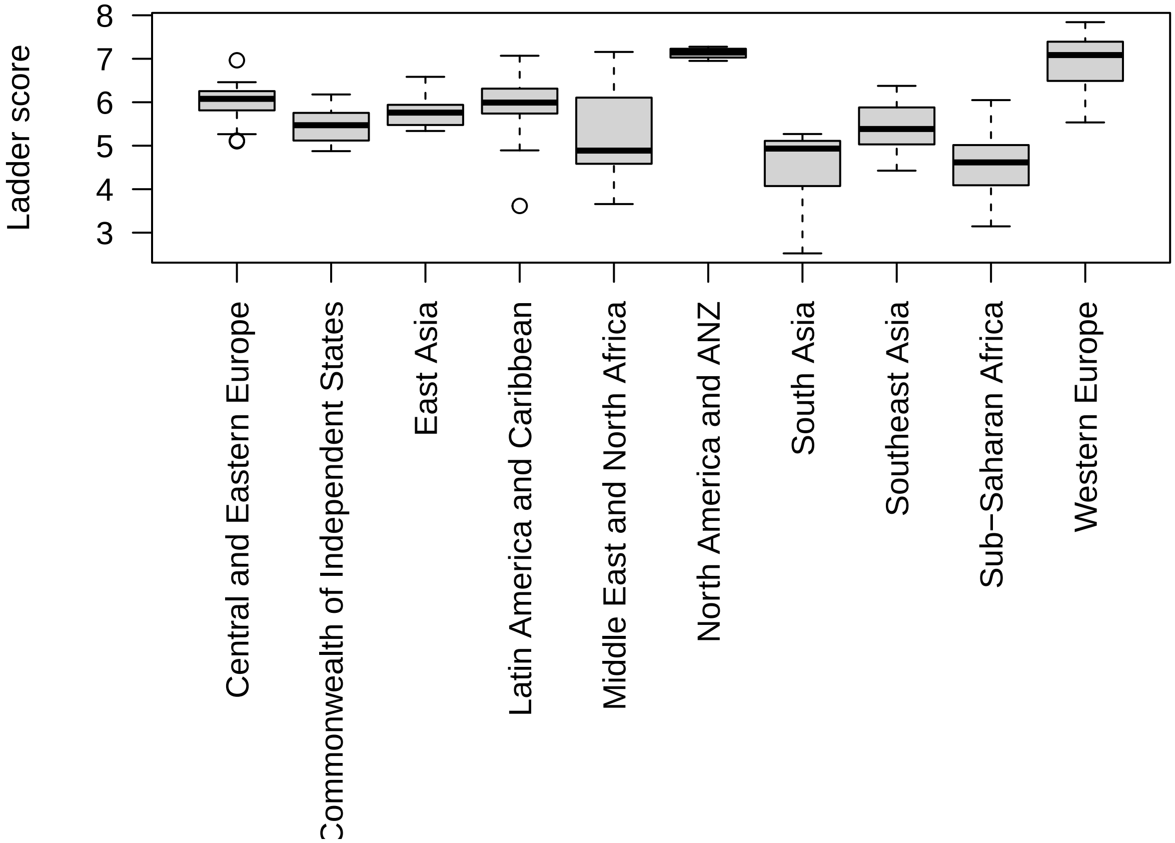
Ranije smo zaključili da je od ponuđenih varijabli, varijabla Logged GDP per capita najbolji prediktor sreće jednostavnom regresijom. No, što je s višestrukom regresijom? Rezultati višestruke regresije nam pak upućuju da je za bolje predviđanje razine sreće ipak bolje uključiti više varijabli.

Postoje li razlike u iskazanoj sreći medu različitim regijama?

Na ovo pitanje ćemo odgovoriti korištenjem jednofaktorskom ANOVA metodom.

U sljedećem isječku ćemo prikazati box plot dijagrame sreće po pojedinim regijama.

```
par(mar=c(15,5,1,1))
boxplot(`Ladder score`~`Regional indicator`,data = whr2021, las = 2, xlab = "" )
```



Boxplot nas upućuje da postoje razlike u iskazanim srećama po regijama. To ćemo potvrditi ANOVA metodom.

Uvjeti za ANOVA-u su normalnost i nezavisnost podataka, te homogenost varijanci među regijama. Nezavisnost podataka možemo pretpostaviti. Normalnost podataka po regijama ćemo provjeriti s Kolmogorov-Smirnovim testom.Hipoteze su nam sljedeće:

$$H_0 : \text{podaci su normalno distribuirani}$$
$$H_1 : \text{podaci nisu normalno distribuirani}$$

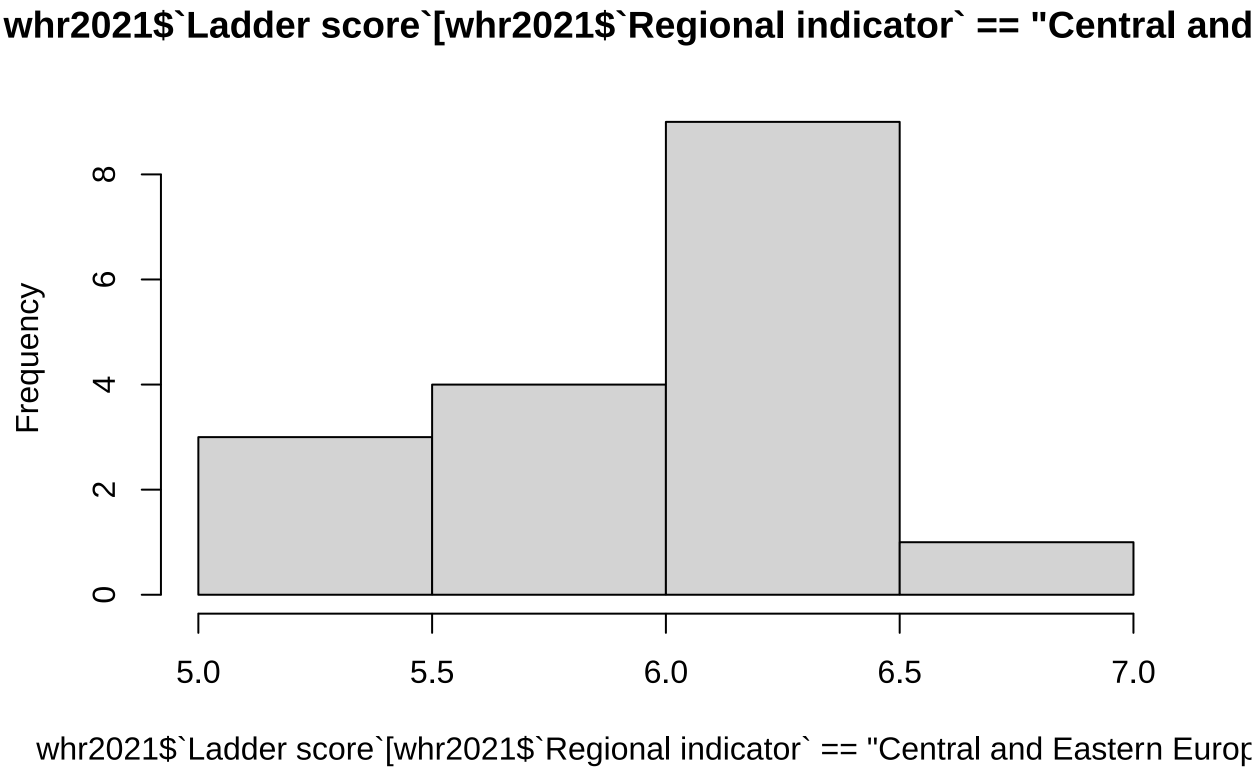
te

$$\alpha = 0.05$$

```
ks.test(whr2021$`Ladder score`[whr2021$`Regional indicator`=='Central and Eastern Europe'], "pnorm", mea
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  whr2021$‘Ladder score‘[whr2021$‘Regional indicator‘ == "Central and Eastern Europe"]
## D = 0.15266, p-value = 0.7689
## alternative hypothesis: two-sided
```

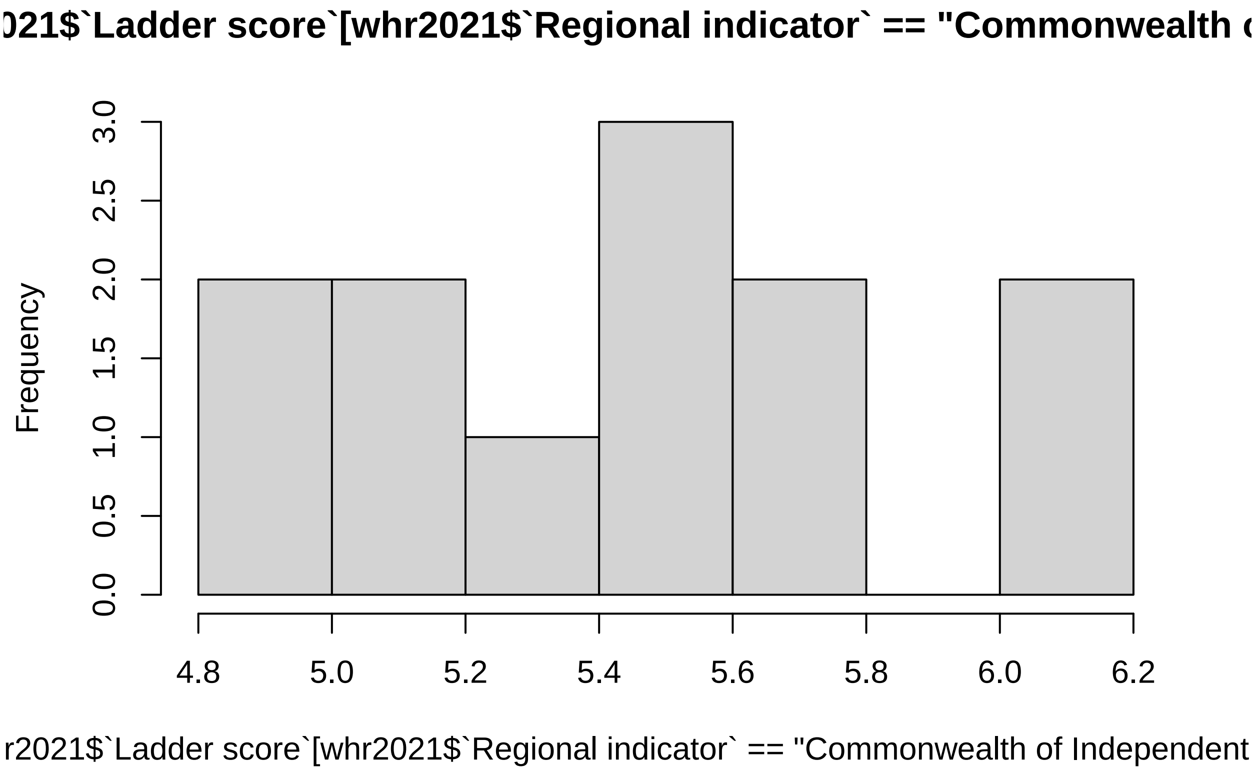
```
hist(whr2021$`Ladder score`[whr2021$`Regional indicator`=='Central and Eastern Europe'])
```



```
ks.test(whr2021$`Ladder score`[whr2021$`Regional indicator`=='Commonwealth of Independent States'], "pn
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  whr2021$‘Ladder score‘[whr2021$‘Regional indicator‘ == "Commonwealth of Independent States"]
## D = 0.1077, p-value = 0.9962
## alternative hypothesis: two-sided
```

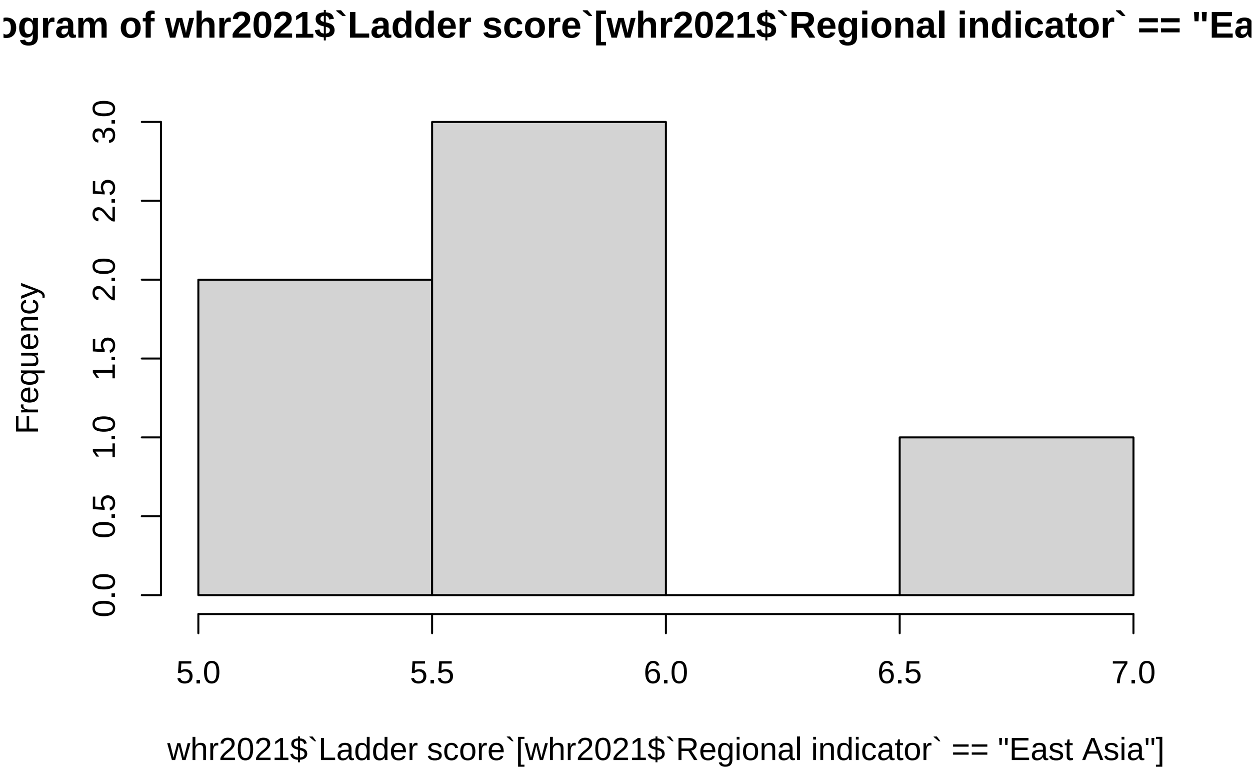
```
hist(whr2021$`Ladder score`[whr2021$`Regional indicator`=='Commonwealth of Independent States'])
```



```
ks.test(whr2021$`Ladder score`[whr2021$`Regional indicator`=='East Asia'], "pnorm", mean(whr2021$`Ladde
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  whr2021$‘Ladder score‘[whr2021$‘Regional indicator‘ == "East Asia"]
## D = 0.21724, p-value = 0.8868
## alternative hypothesis: two-sided
```

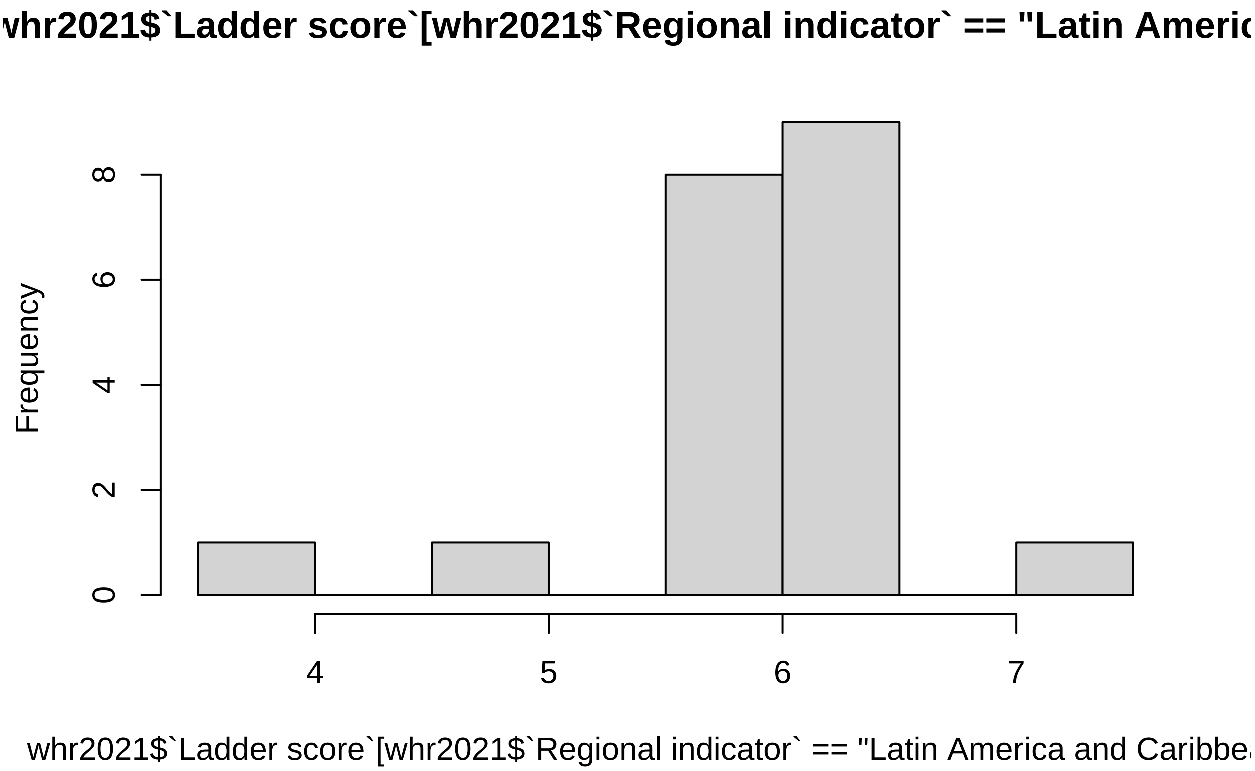
```
hist(whr2021$`Ladder score`[whr2021$`Regional indicator`=='East Asia'])
```



```
ks.test(whr2021$`Ladder score`[whr2021$`Regional indicator`=='Latin America and Caribbean'], "pnorm", m
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  whr2021$‘Ladder score‘[whr2021$‘Regional indicator‘ == "Latin America and Caribbean"]
## D = 0.20631, p-value = 0.3171
## alternative hypothesis: two-sided
```

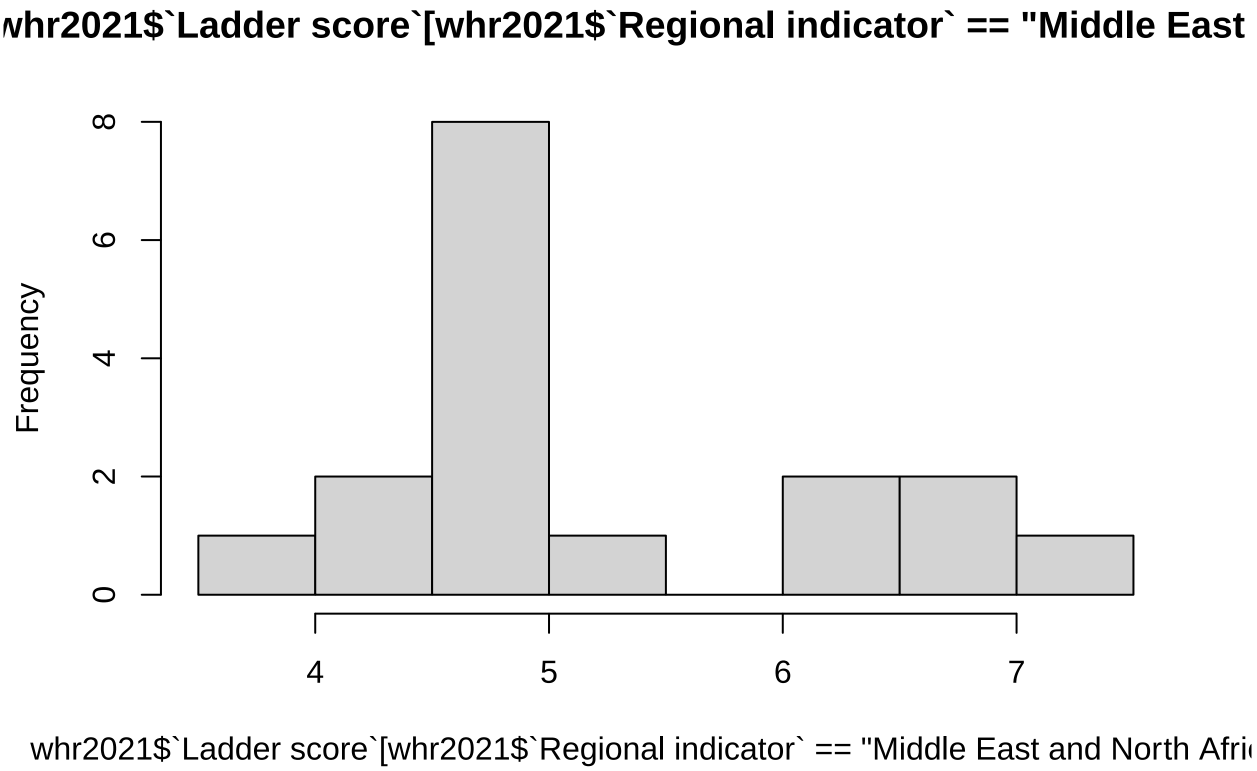
```
hist(whr2021$`Ladder score`[whr2021$`Regional indicator`=='Latin America and Caribbean'])
```



```
ks.test(whr2021$`Ladder score`[whr2021$`Regional indicator`=='Middle East and North Africa'], "pnorm", 
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  whr2021$‘Ladder score‘[whr2021$‘Regional indicator‘ == "Middle East and North Africa"]
## D = 0.25437, p-value = 0.186
## alternative hypothesis: two-sided
```

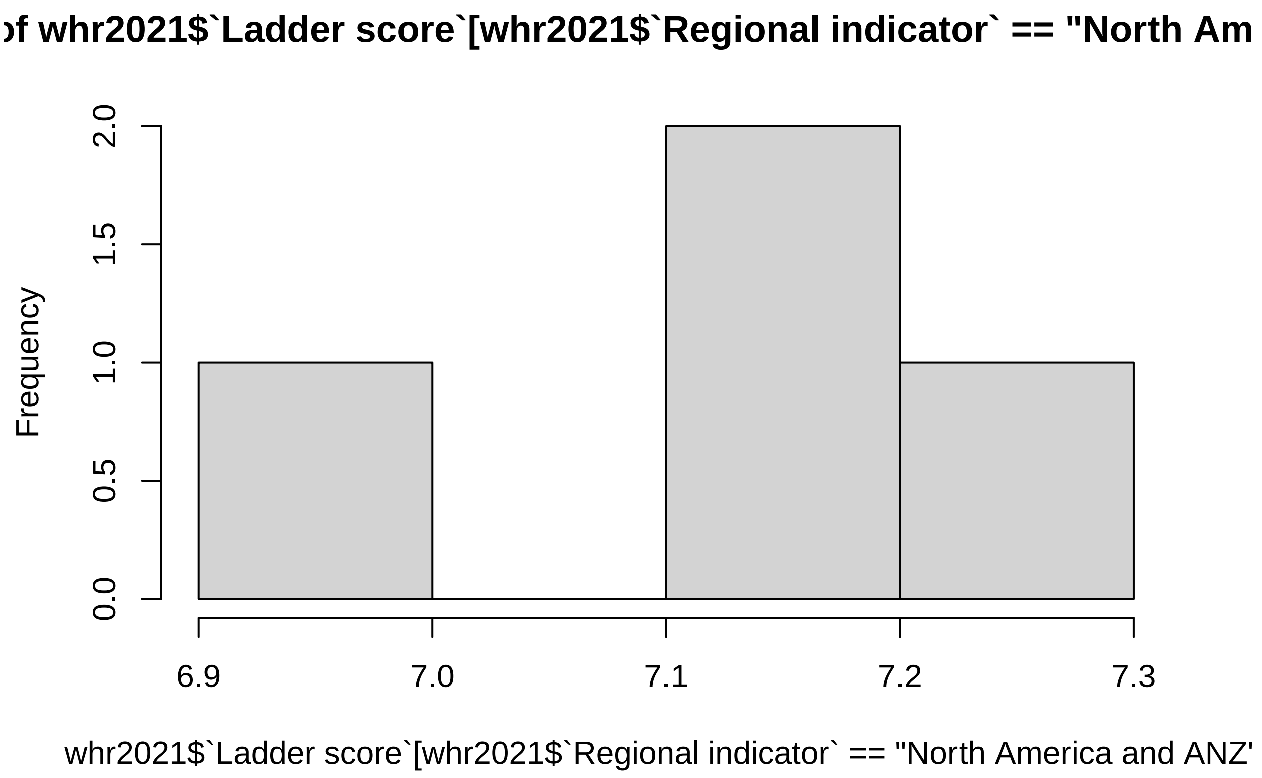
```
hist(whr2021$`Ladder score`[whr2021$`Regional indicator`=='Middle East and North Africa'])
```



```
ks.test(whr2021$`Ladder score`[whr2021$`Regional indicator`=='North America and ANZ'], "pnorm", mean(wh
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  whr2021$‘Ladder score‘[whr2021$‘Regional indicator‘ == "North America and ANZ"]
## D = 0.17678, p-value = 0.9972
## alternative hypothesis: two-sided
```

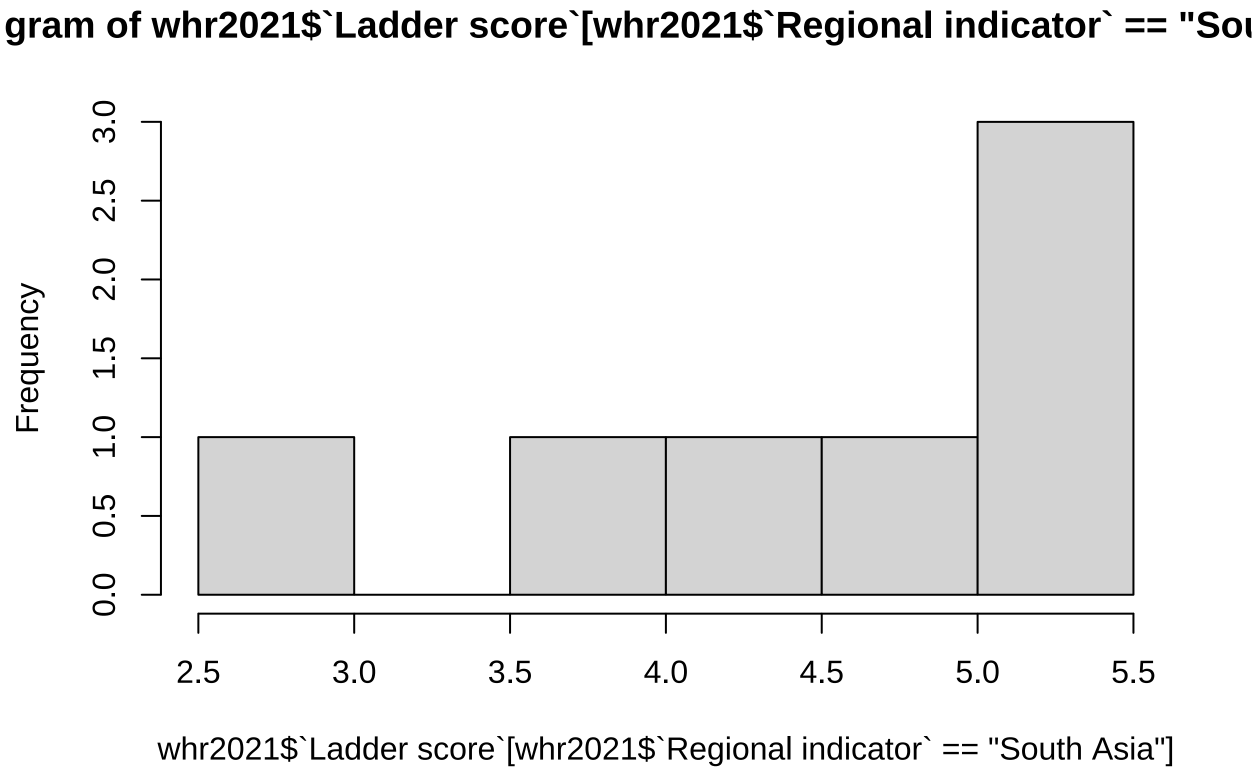
```
hist(whr2021$`Ladder score`[whr2021$`Regional indicator`=='North America and ANZ'])
```



```
ks.test(whr2021$`Ladder score`[whr2021$`Regional indicator`=='South Asia'], "pnorm", mean(whr2021$`Ladd
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  whr2021$‘Ladder score‘[whr2021$‘Regional indicator‘ == "South Asia"]
## D = 0.26133, p-value = 0.6354
## alternative hypothesis: two-sided
```

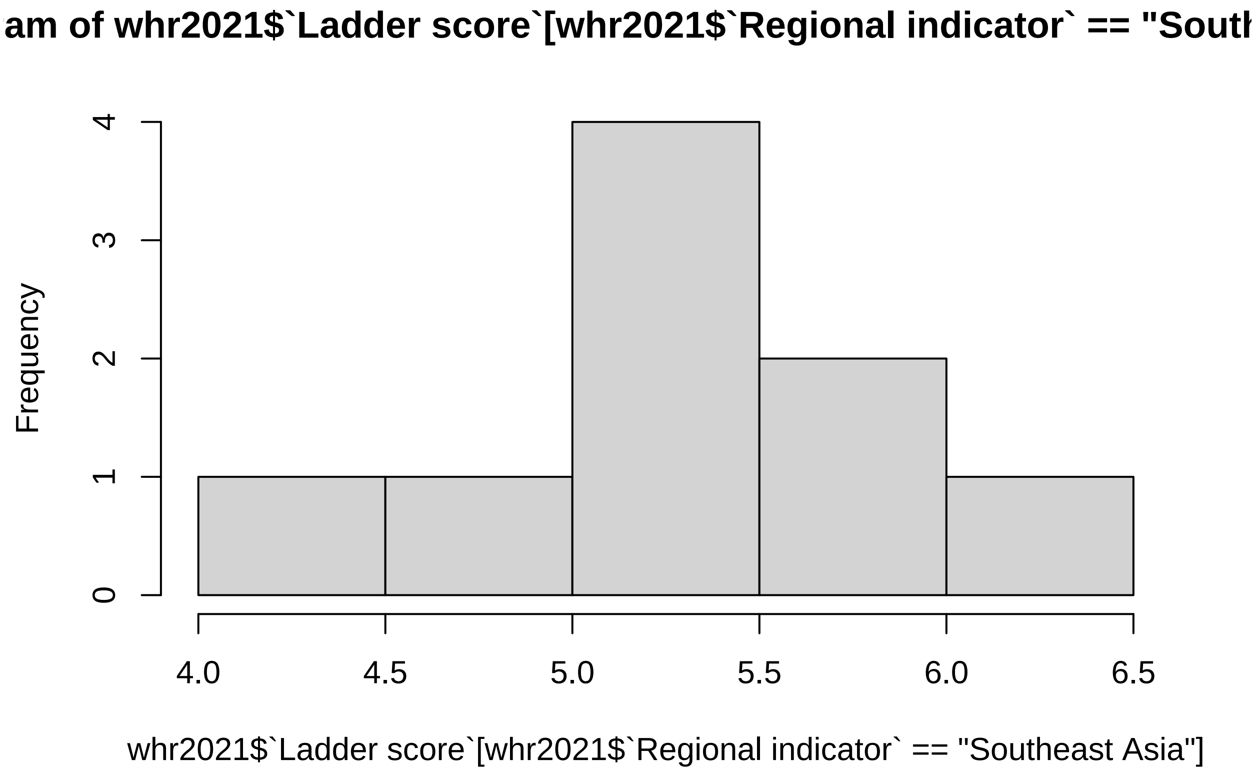
```
hist(whr2021$`Ladder score`[whr2021$`Regional indicator`=='South Asia'])
```



```
ks.test(whr2021$`Ladder score`[whr2021$`Regional indicator`=='Southeast Asia'], "pnorm", mean(whr2021$`l
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  whr2021$‘Ladder score‘[whr2021$‘Regional indicator‘ == "Southeast Asia"]
## D = 0.16447, p-value = 0.9367
## alternative hypothesis: two-sided
```

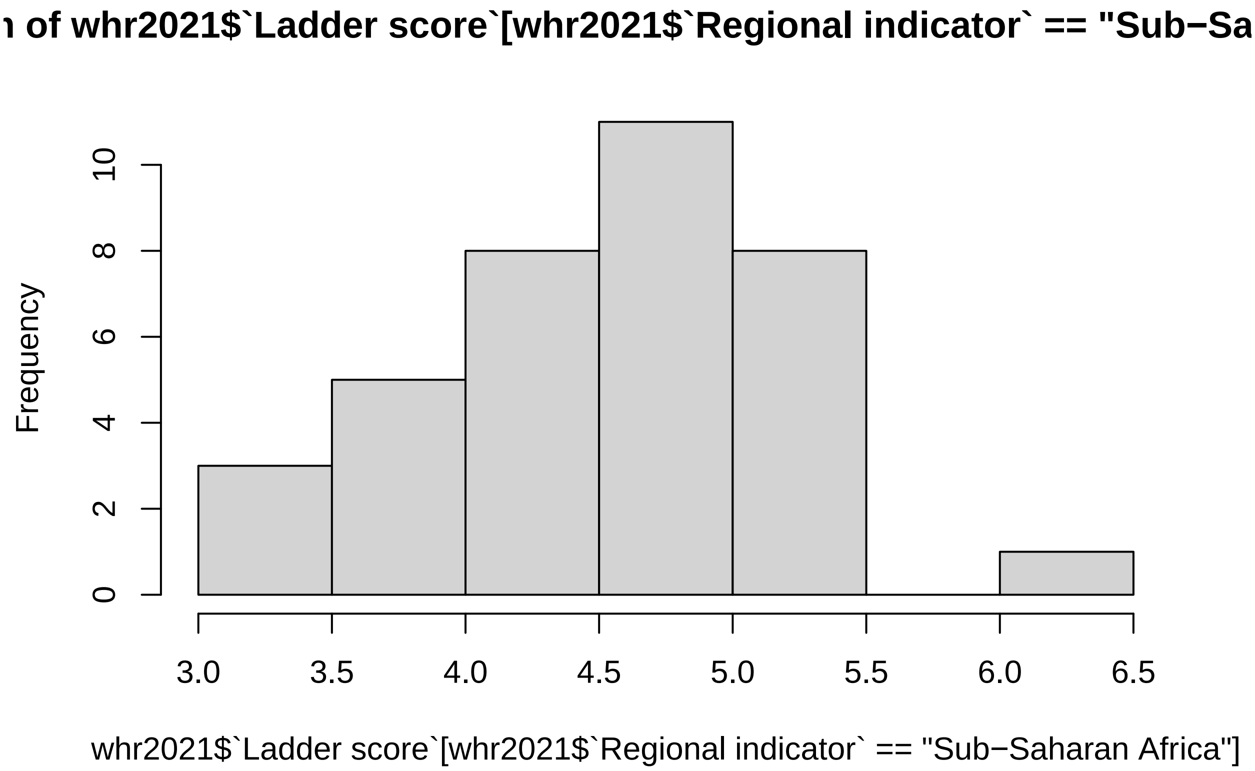
```
hist(whr2021$`Ladder score`[whr2021$`Regional indicator`=='Southeast Asia'])
```



```
ks.test(whr2021$`Ladder score`[whr2021$`Regional indicator`=='Sub-Saharan Africa'], "pnorm", mean(whr20
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  whr2021$‘Ladder score‘[whr2021$‘Regional indicator‘ == "Sub-Saharan Africa"]
## D = 0.1039, p-value = 0.7942
## alternative hypothesis: two-sided
```

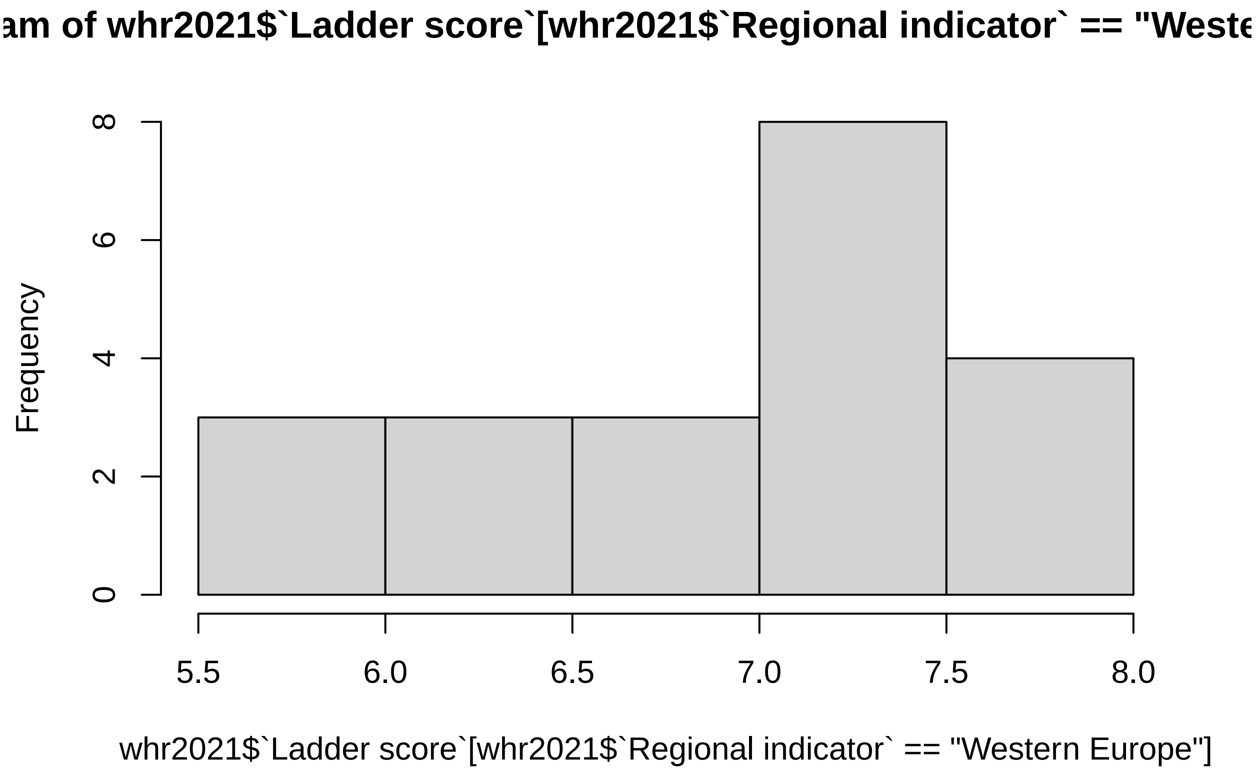
```
hist(whr2021$`Ladder score`[whr2021$`Regional indicator`=='Sub-Saharan Africa'])
```



```
ks.test(whr2021$`Ladder score`[whr2021$`Regional indicator`=='Western Europe'], "pnorm", mean(whr2021$`
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  whr2021$‘Ladder score‘[whr2021$‘Regional indicator‘ == "Western Europe"]
## D = 0.16103, p-value = 0.5918
## alternative hypothesis: two-sided
```

```
hist(whr2021$`Ladder score`[whr2021$`Regional indicator`=='Western Europe'])
```



P-vrijednosti na svim testovima su nam veće od kritične vrijednosti te ne odbijamo nul hipotezu.

Sada trebamo analizirati homogenost varijanci regija što ćemo napraviti s Bartlettovim testom. Hipoteze su nam sljedeće:

$$H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$$
$$H_1 : \neg H_0.$$

te

$$\alpha = 0.05$$

```
bartlett.test(whr2021$`Ladder score` ~ whr2021$`Regional indicator`)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  whr2021$‘Ladder score‘ by whr2021$‘Regional indicator‘
## Bartlett’s K-squared = 21.976, df = 9, p-value = 0.008955
```

P-vrijednost je manja od kritične vrijednosti tako da ne odbacujemo nul hipotezu.

Sada možemo napraviti jednofaktorsku ANOVA-u. Hipoteze su nam sljedeće:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$
$$H_1 : \neg H_0.$$

te

$$\alpha = 0.05k = 10$$

```
luck = aov(whr2021$`Ladder score` ~ whr2021$`Regional indicator`)
summary(luck)
```

```
##                                Df Sum Sq Mean Sq F value Pr(>F)
## whr2021$‘Regional indicator‘   9 106.05  11.783   25.34 <2e-16 ***
## Residuals                    139  64.64   0.465
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Kao što je sugerirano u grafu s početka, ANOVA potvrđuje da postoji razlika u iskazanoj sreći među regijama.