SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Maja Benkus

Kristijan Maoduš

Matija Žinić

Očekivani životni vijek, Life expectancy (WHO)

SEMINARSKI RAD

Varaždin, 2020.

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Maja Benkus, JMBAG: 1191231766

Kristijan Maoduš, JMBAG: 0016116726

Matija Žinić, JMBAG: 0016116773

Studij: Baze podataka i baze znanja

Očekivani životni vijek, Life expectancy (WHO)

SEMINARSKI RAD

Mentorica:

Prof. dr. sc. Jasminka Dobša

Varaždin, lipanj 2020.

Sažetak

Ovaj rad je izrađen kao seminarski rad iz kolegija *Statističke metode za informatičare* na Fakultetu organizacije i informatike. Kroz nekoliko poglavlja u radu će biti opisani korišteni skup podataka i manipulacije nad istim te rezultati provedenih parametarskih i neparametarskih testova. Nad skupom podataka provedena je regresijska analiza.

Ključne riječi: statistika; WHO; hi-kvadrat; regresija; ANOVA; parametarski test; neparametarski test

Sadržaj

1. Uvod

U ovom seminarskom radu prikazat ćemo statističku obradu faktora koji utječu na očekivani životni vijek kroz nekoliko poglavlja. Podatkovni skup [1] sadrži 22 varijable koje su detaljno opisane u *zadatku a* u sljedećem poglavlju. Podaci su prikupljeni za ukupno 193 države i to u rasponu od 2000. do 2015. godine. Svjetska zdravstvena organizacija (WHO) javno objavljuje zdravstvene podatke, a u zadnjih 15 godina primijećen je značajan napredak u zdravstvenom sektoru što je rezultiralo smanjenjem stope smrtnosti, pogotovo država u razvoju u usporedbi s podacima zadnjih 30-ak godina. Kroz ove zadatke prikazat ćemo obradu podataka i rad u alatu RStudio, koji je integrirano razvojno okruženje (IDE) za programski jezik R. Prvo ćemo opisati same varijable statističkog skupa i grafički ih prikazati, zatim vidjeti postoji li korelacija među istima, nakon toga ćemo definirati nove kvalitativne varijable nad kojima ćemo provesti hi-kvadrat test, provest ćemo nekoliko parametarskih i neparametarskih testova, napraviti regresijsku analizu, te na kraju izvesti zaključak.

2. Deskriptivna statistika

2.1. Zadatak a

U zadatku a opisat ćemo varijable statističkog skupa i grafički ćemo ih prikazati. Naš skup podataka sadrži ukupno 22 varijable unutar preuzete .csv datoteke. Varijable možemo podijeliti na kvalitativne i kvantitativne. Kvalitativne varijable su:

- Država (eng. *Country*) – imamo podatke za ukupno 193 različite države

```
> #Grafici prikazi - kvalitativne
> distinct(.data = data, Country, .keep_all = FALSE)
                                          Country
1                                     Afghanistan
2                                         Albania
3                                         Algeria
4                                          Angola
5                                Antigua and Barbuda
6                                       Argentina
7                                         Armenia
8                                       Australia
9                                         Austria
10                                     Azerbaijan
11                                        Bahamas
12                                        Bahrain
13                                     Bangladesh
14                                       Barbados
15                                        Belarus
16                                        Belgium
17                                         Belize
18                                          Benin
19                                         Bhutan
20                  Bolivia (Plurinational State of)
21                      Bosnia and Herzegovina
22                                       Botswana
23                                         Brazil
24                           Brunei Darussalam
25                                       Bulgaria
26                                   Burkina Faso
27                                        Burundi
28                                  CÃ´te d'Ivoire
29                                      Cabo Verde
30                                       Cambodia
31                                       Cameroon
32                                         Canada
33                         Central African Republic
34                                           Chad
35                                          Chile
36                                          China
```

Slika 1. Popis država (Snimka zaslona, 2020.)

Na slici 1. vidimo neke od ukupno 193 različite države koje u RStudio možemo ispisati pomoću `distinct()` funkcije [2].

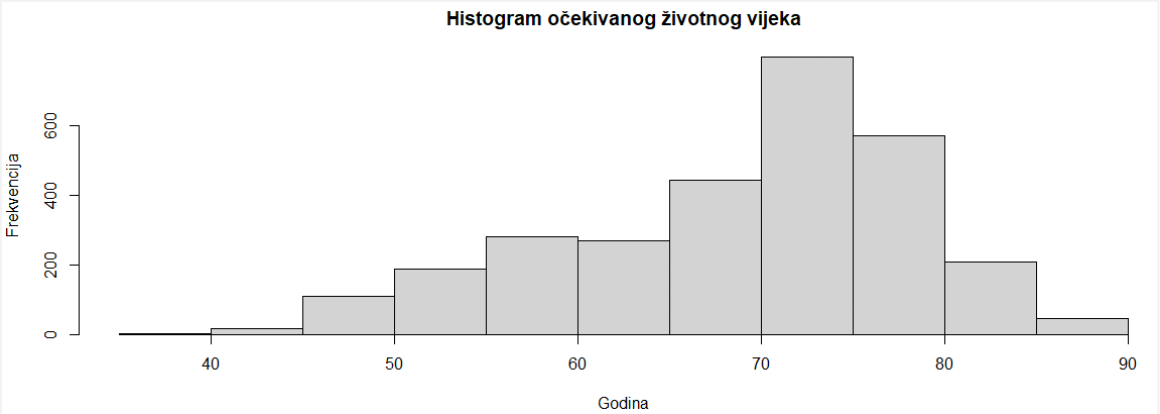
- Status – varijabla koja ima dvije vrijednosti, *Developing* i *Developed,* te predstavlja status razvijenosti određene države.

```
> count<-distinct(.data = data, Status, .keep_all = FALSE)
> count
      Status
1 Developing
2  Developed
```

Slika 2. Prikaz statusa pomoću funkcije distinct() (Snimka zaslona, 2020.)

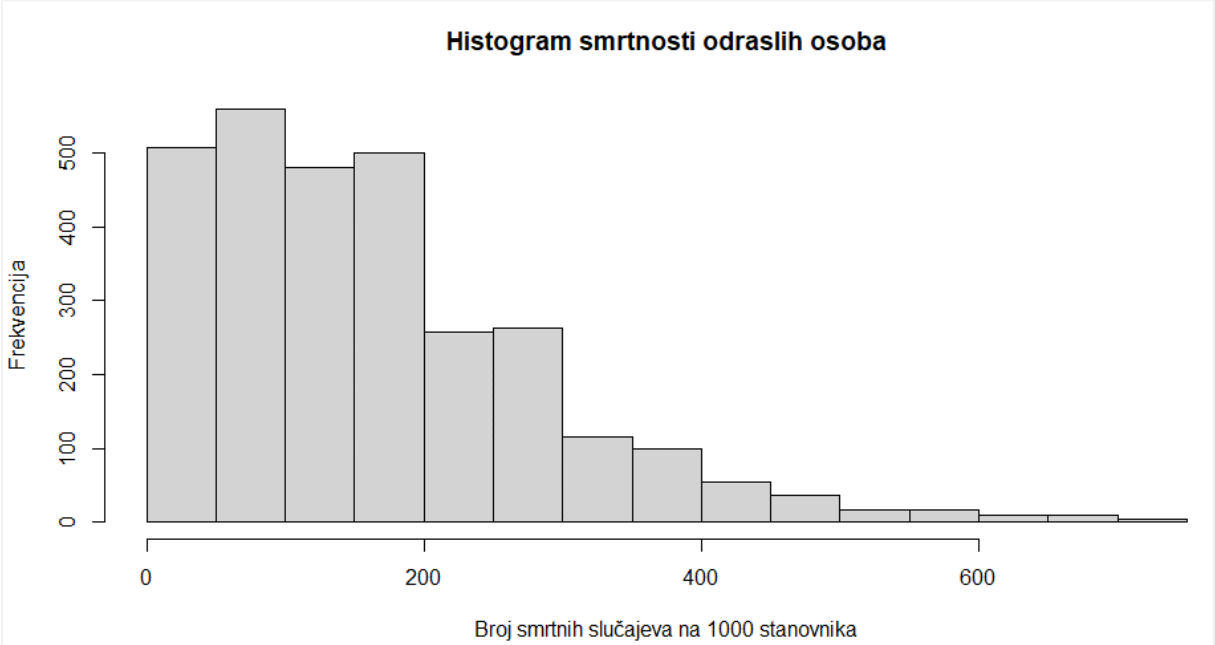
Kvantitativne varijable su:

- Godina (eng. *year*) – podaci su u rasponu od 2000. - 2015. godine
- Očekivani životni vijek (eng. *Life expectancy*) – izražen u godinama



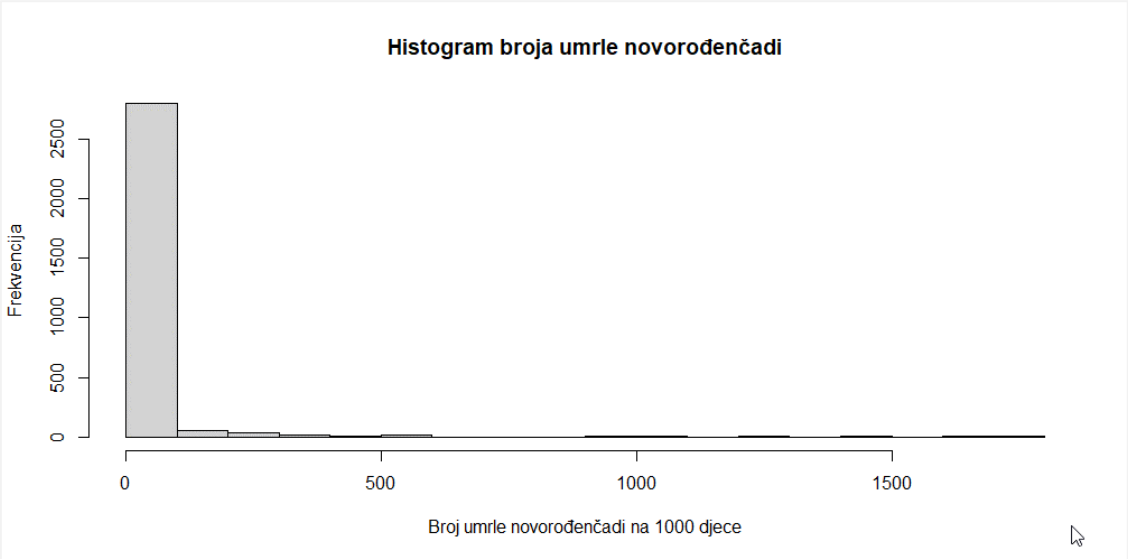
Slika 3. Histogram očekivanog životnog vijeka (Snimka zaslona, 2020.)

- Stopa smrtnosti odraslih (eng. *Adult Mortality*) - stopa smrtnosti oba spola (vjerojatnost umiranja u rasponu od 15 do 60 godina na 1000 stanovnika)



Slika 4. Histogram smrtnosti odraslih osoba (Snimka zaslona, 2020.)

- Broj umrlih novorođenčadi (eng. *Infant Deaths*) - broj umrlih novorođenčadi na 1000 djece

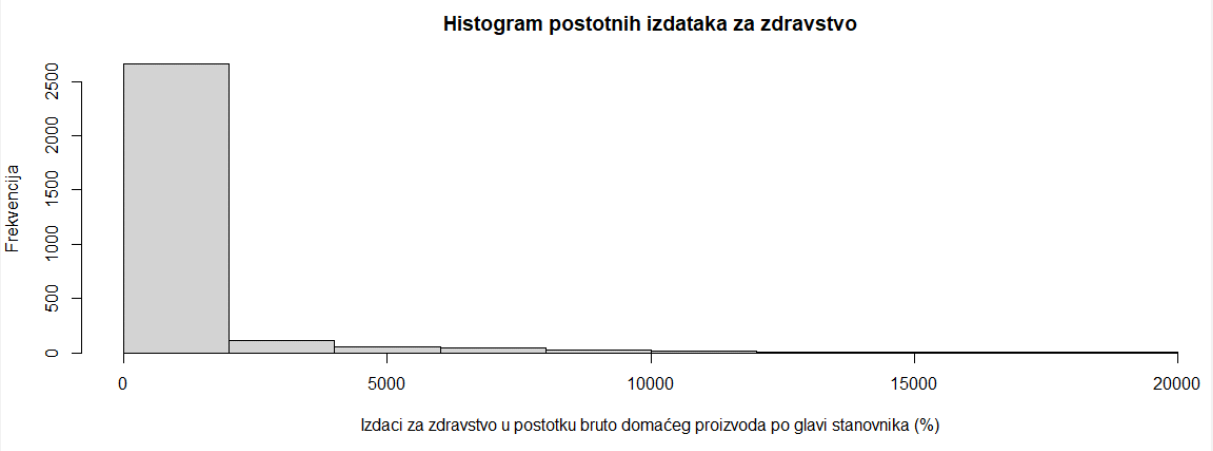


Slika 5. Histogram broja umrle novorođenčadi (Snimka zaslona, 2020.)

- Alkohol - zabilježen po glavi stanovnika (15+), izražen u litrama čistog alkohola

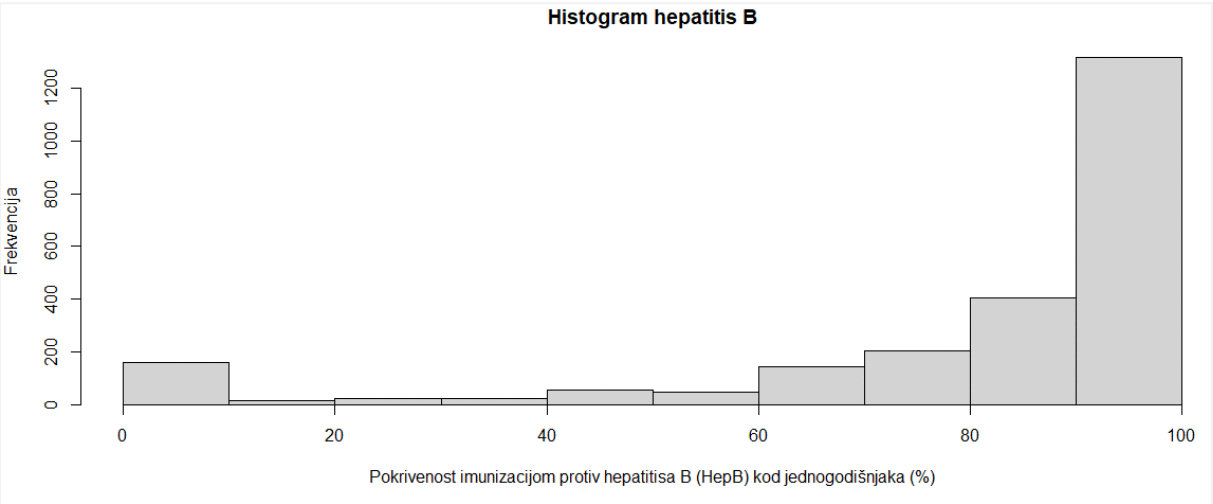
Slika 6. Histogram alkohola (Snimka zaslona, 2020.)

- Postotni izdaci (eng. *percentage expenditure*) - izdaci za zdravstvo u postotku bruto domaćeg proizvoda po glavi stanovnika (%)



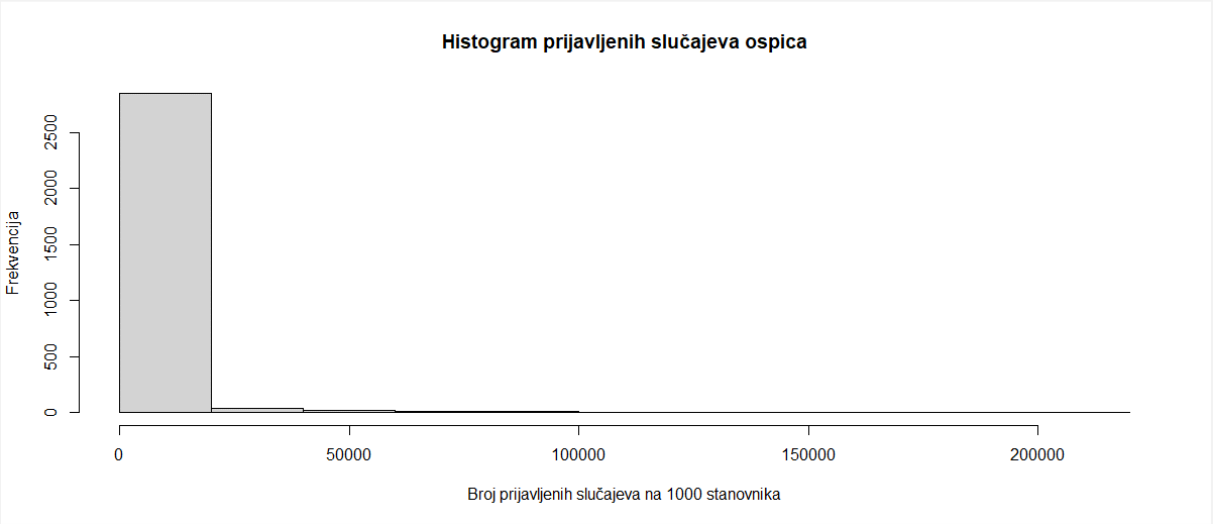
Slika 7. Histogram postotnih izdataka za zdravstvo (Snimka zaslona, 2020.)

- Hepatitis B - pokrivenost imunizacijom protiv hepatitisa B (HepB) kod jednogodišnjaka (%)



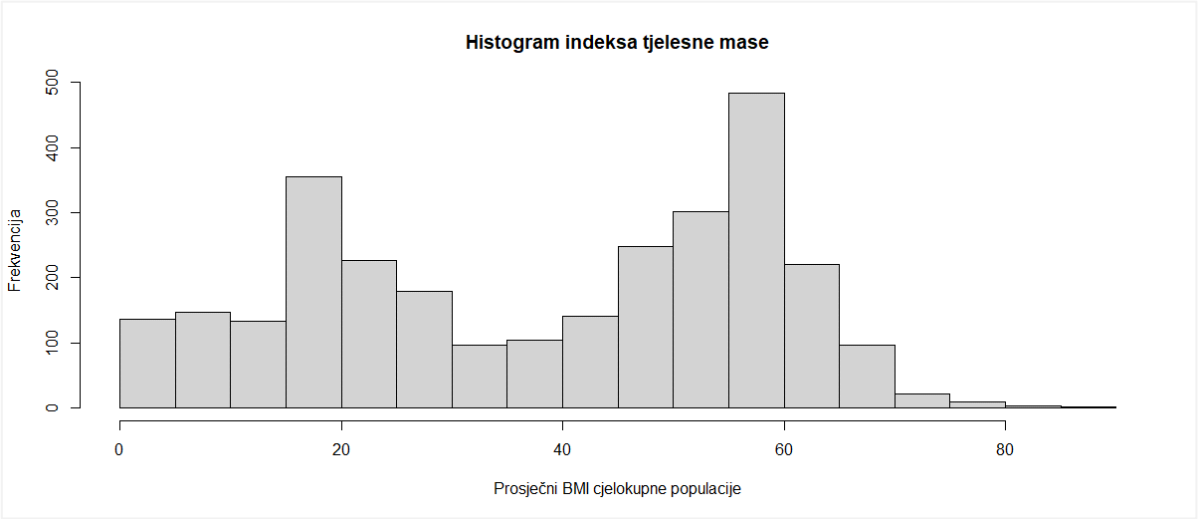
Slika 8. Histogram imunizacije hepatitisa B (Snimka zaslona, 2020.)

- Ospice (eng. *Measles*) - broj prijavljenih slučajeva na 1000 stanovnika



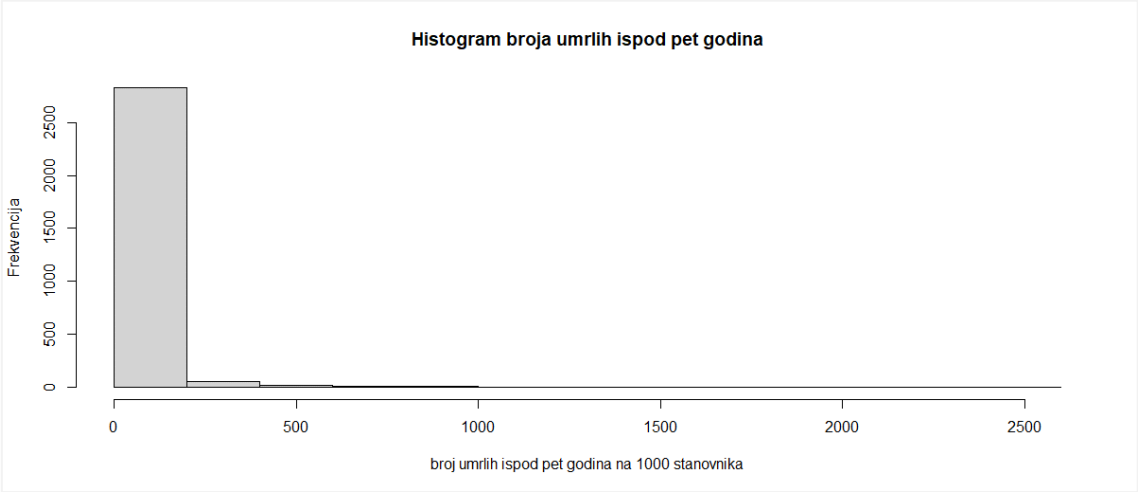
Slika 9. Histogram prijavljenih slučajeva ospica (Snimka zaslona, 2020.)

- Indeks tjelesne mase (eng. *Body Mass Indeks [BMI]*) - prosječni indeks tjelesne mase cjelokupne populacije



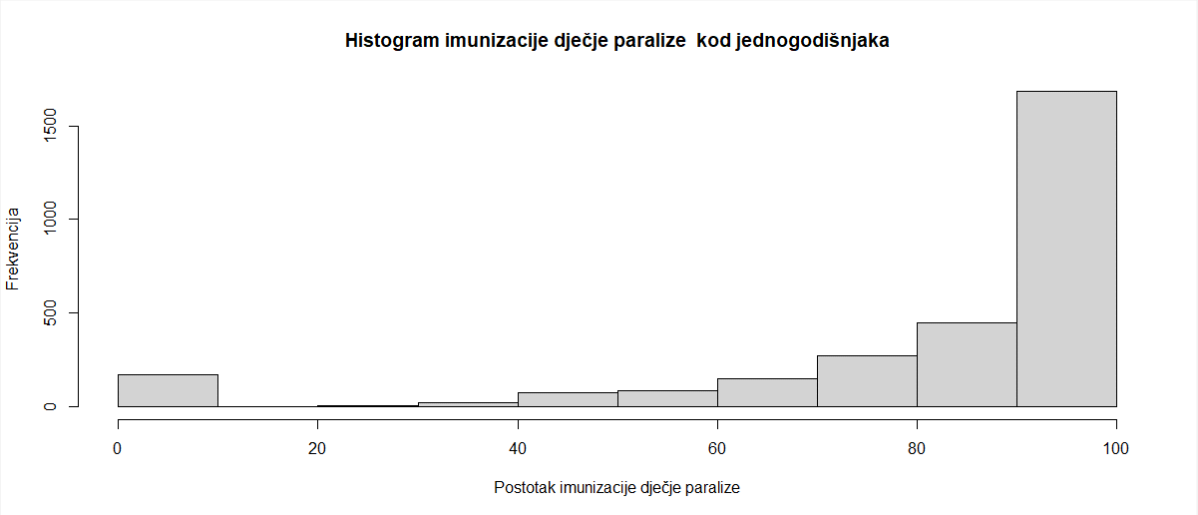
Slika 10. Histogram indeksa tjelesne mase (Snimka zaslona, 2020.)

- Broj umrlih ispod pet godina - broj umrlih ispod pet godina na 1000 stanovnika



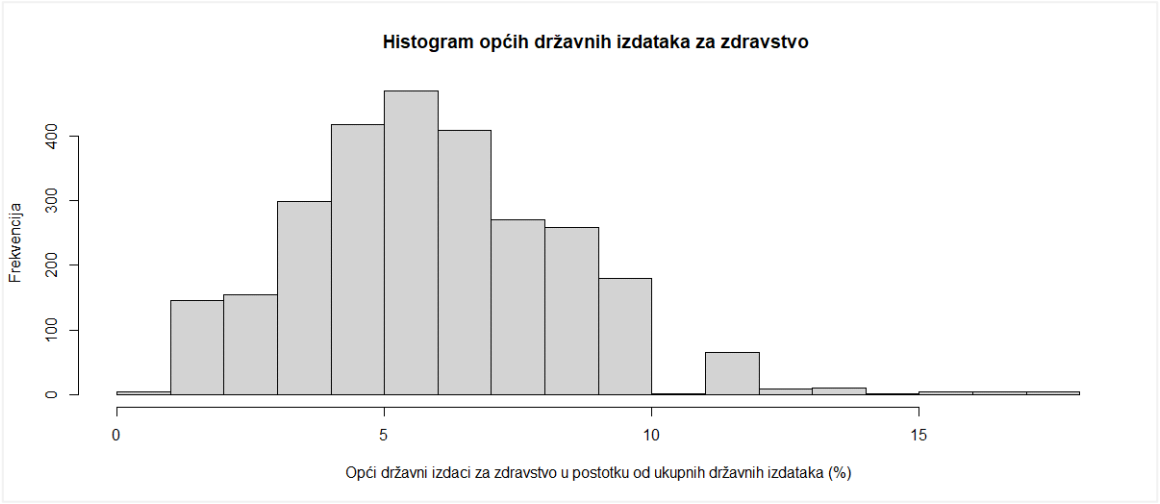
Slika 11. Histogram broja umrlih ispod pet godina (Snimka zaslona, 2020.)

- Imunizacija dječje paralize (eng. *Polio)* - pokrivenost imunizacijom protiv dječje paralize (Pol3) kod jednogodišnjaka (%)



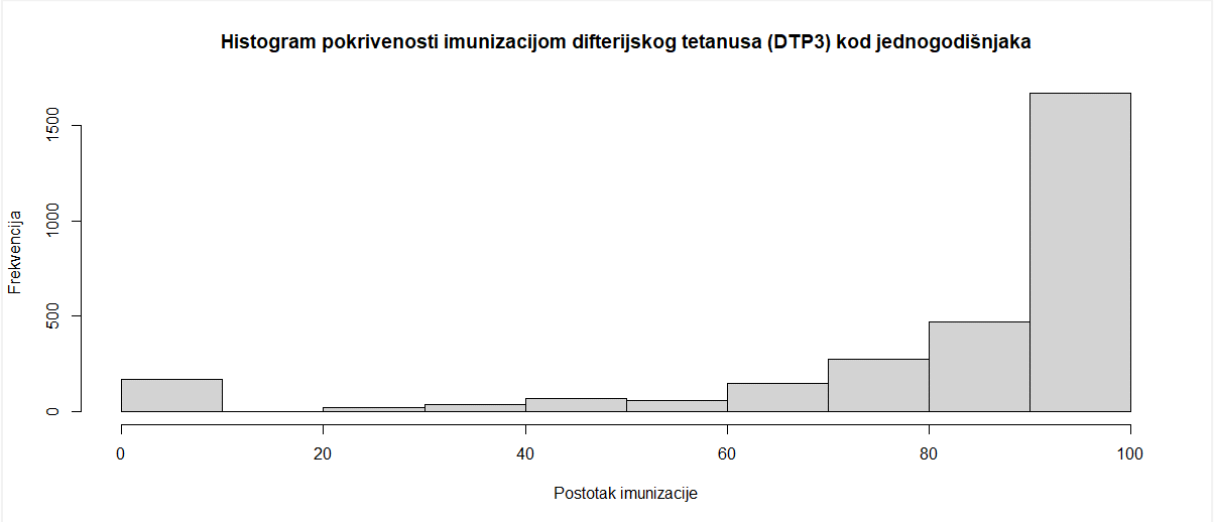
Slika 12. Histogram imunizacije dječje paralize kod jednogodišnjaka (Snimka zaslona, 2020.)

- Opći državni izdaci za zdravstvo (eng. *Total expenditure*) - opći državni izdaci za zdravstvo u postotku od ukupnih državnih izdataka (%)



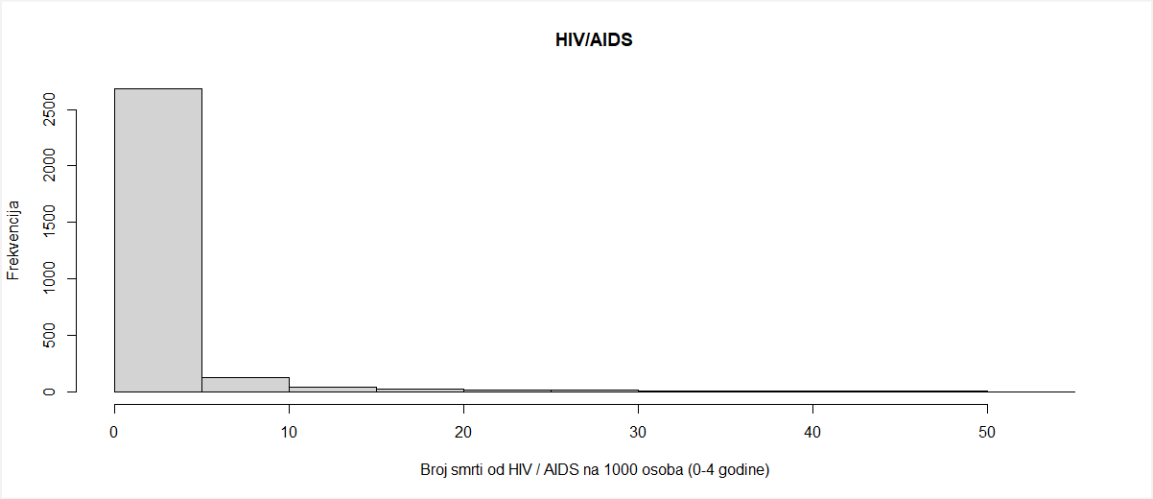
Slika 13. Histogram općih državnih izdataka za zdravstvo (Snimka zaslona, 2020.)

- Difterija - pokrivenost imunizacijom difterijskog tetanusa (DTP3) kod jednogodišnjaka (%)



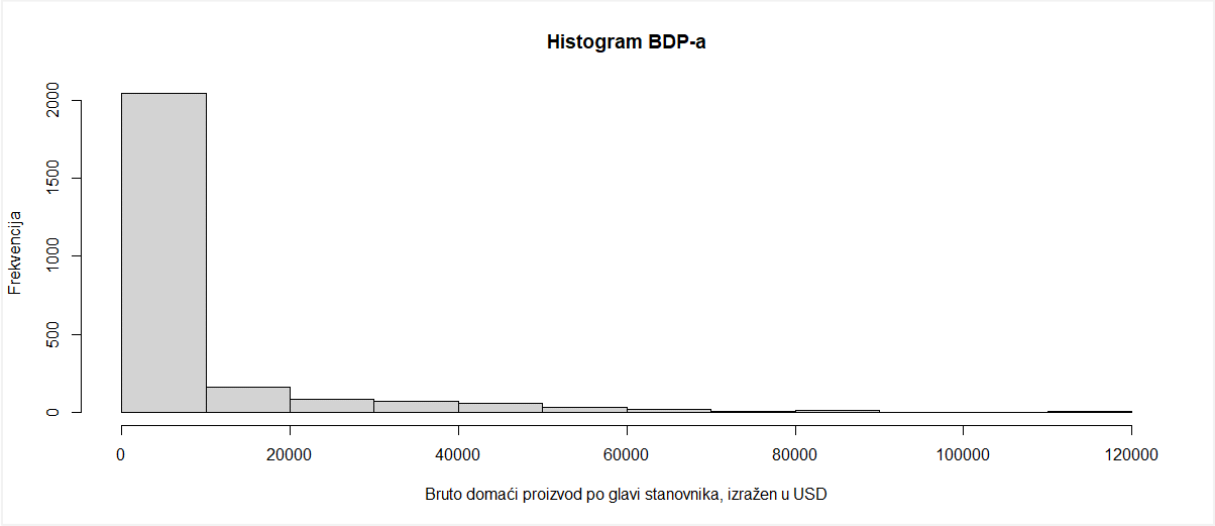
Slika 14. Histogram pokrivenosti imunizacijom difterijskog tetanusa (DTP3) kod jednogodišnjaka (Snimka zaslona, 2020.)

- HIV/AIDS - smrt na 1000 novorođenih od HIV / AIDS (0-4 godine)



Slika 15. Histogram HIV/AIDS (Snimka zaslona, 2020.)

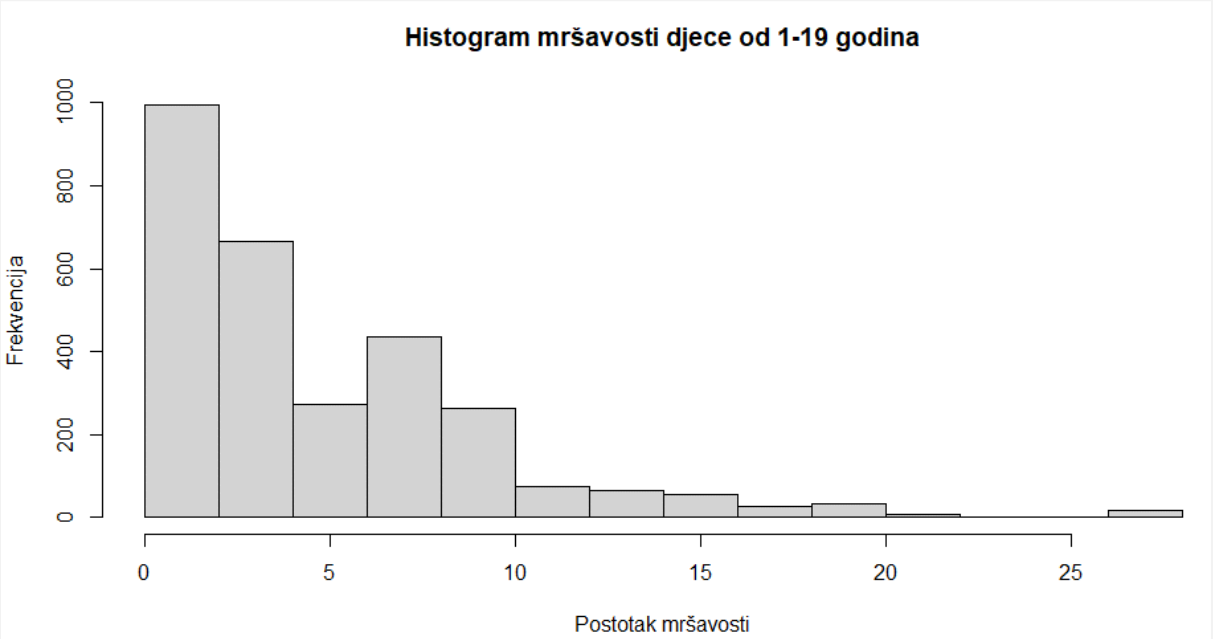
- BDP (eng. *GDP*) - bruto domaći proizvod po glavi stanovnika, izražen u USD



Slika 16. Histogram BDP-a (Snimka zaslona, 2020.)

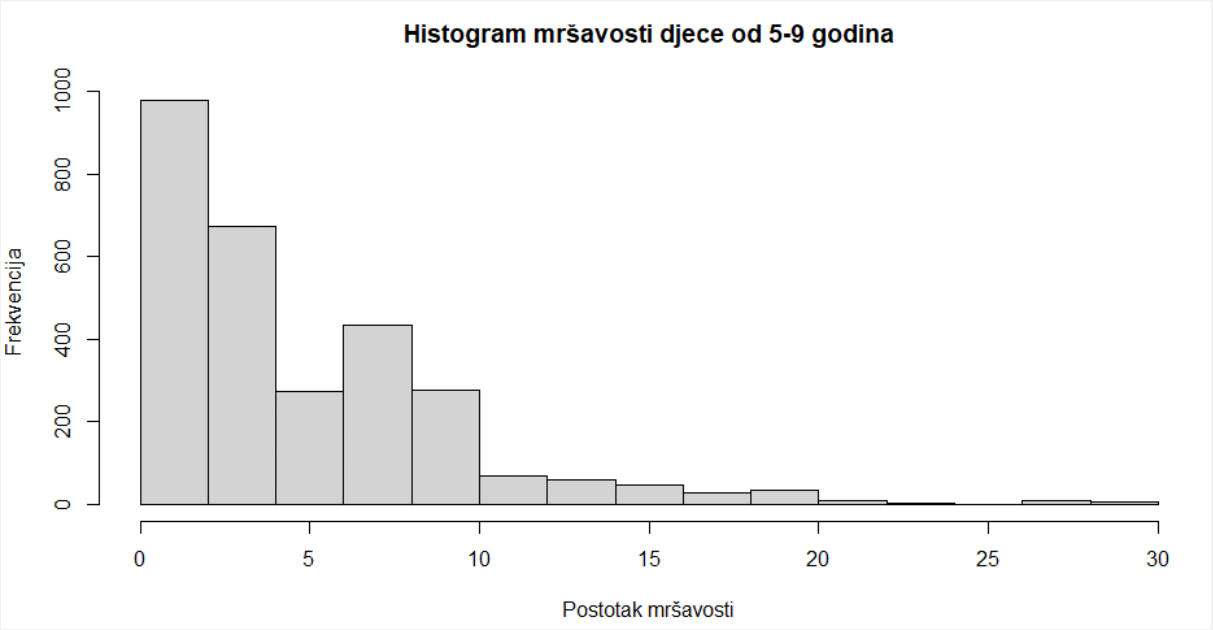
- Populacija (eng. *population*)

- Mršavost 1-19 godina - rasprostranjenost mršavosti kod djece i adolescenata u dobi od 10 do 19 godina (%)



Slika 17. Histogram postotka mršavosti djece od 1-19 godina (Snimka zaslona, 2020.)

- Mršavost 5-9 godina - rasprostranjenost mršavosti kod djece u dobi od 5 do 9 godina (%)

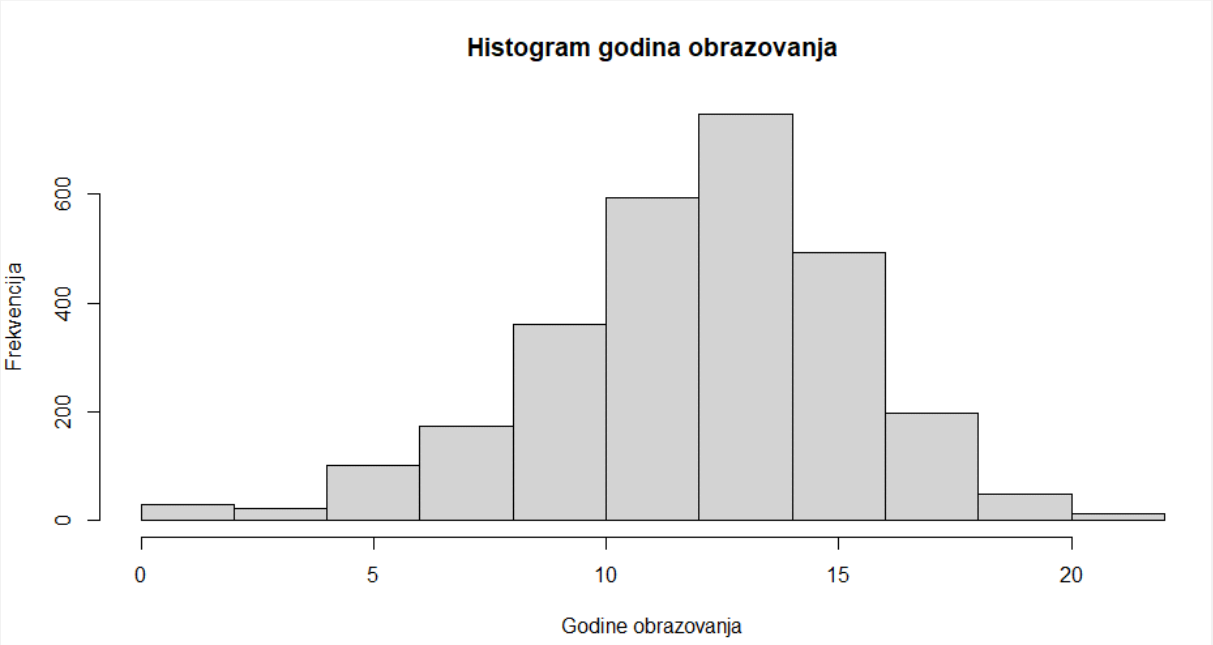


Slika 18. Histogram postotka mršavosti djece od 5-9 godina (Snimka zaslona, 2020.)

- Sastav dohotka (eng. *Income composition*) - indeks ljudskog razvoja s obzirom na sastav dohotka resursa (indeks se kreće od 0 do 1)

Slika 19. Histogram sastava dohotka (Snimka zaslona, 2020.)

- Školovanje (eng. *schooling*) - broj godina školovanja (godine)



Slika 20. Histogram godina obrazovanja (Snimka zaslona, 2020.)

2.1.1. Korelacije

Kako bi izračunali korelacijsku matricu prvo ćemo odabrati sve kvantitativne varijable iz našeg skupa podataka, te ih učitati u varijablu *dataKorelacija*.

```
# select variables
myvars <- c("Year", "Life.expectancy", "Adult.Mortality", "infant.deaths", "Alcohol", "percentage.expenditure",
            "Hepatitis.B", "Measles", "BMI", "under.five.deaths", "Polio", "Total.expenditure", "Diphtheria",
            "HIV.AIDS", "GDP", "Population", "thinness..1.19.years", "thinness.5.9.years",
            "Income.composition.of.resources", "Schooling")
dataKorelacija <- data[myvars]
```

Slika 21. Podskup skupa podataka (Snimka zaslona, 2020.)

Funkcijom [3]

```
res3 <- cor(dataKorelacija, use = "complete.obs")
round(res3, 2)
```

izračunat ćemo korelacijsku matricu koja je vidljiva na Slika 22.

```
                                 Year Life.expectancy Adult.Mortality infant.deaths Alcohol percentage.expenditure Hepatitis.B Measles   BMI under.five.deaths
Year                             1.00            0.05          -0.04         0.01   -0.11                  0.07        0.11   -0.05  0.01              0.01
Life.expectancy                  0.05            1.00          -0.70        -0.17    0.40                  0.41        0.20   -0.07  0.54             -0.19
Adult.Mortality                 -0.04           -0.70           1.00         0.04   -0.18                 -0.24       -0.11    0.00 -0.35              0.06
infant.deaths                    0.01           -0.17           0.04         1.00   -0.11                 -0.09       -0.23    0.53 -0.23              1.00
Alcohol                         -0.11            0.40          -0.18        -0.11    1.00                  0.42        0.11   -0.05  0.35             -0.10
percentage.expenditure           0.07            0.41          -0.24        -0.09    0.42                  1.00        0.02   -0.06  0.24             -0.09
Hepatitis.B                      0.11            0.20          -0.11        -0.23    0.11                  0.02        1.00   -0.12  0.14             -0.24
Measles                         -0.05           -0.07           0.00         0.53   -0.05                 -0.06       -0.12    1.00 -0.15              0.52
BMI                              0.01            0.54          -0.35        -0.23    0.35                  0.24        0.14   -0.15  1.00             -0.24
under.five.deaths                0.01           -0.19           0.06         1.00   -0.10                 -0.09       -0.24    0.52 -0.24              1.00
Polio                           -0.02            0.33          -0.20        -0.16    0.24                  0.13        0.46   -0.06  0.19             -0.17
Total.expenditure                0.06            0.17          -0.09        -0.15    0.21                  0.18        0.11   -0.11  0.19             -0.15
Diphtheria                       0.03            0.34          -0.19        -0.16    0.24                  0.13        0.59   -0.06  0.18             -0.18
HIV.AIDS                        -0.12           -0.59           0.55         0.01   -0.03                 -0.10       -0.09    0.00 -0.21              0.02
GDP                              0.10            0.44          -0.26        -0.10    0.44                  0.96        0.04   -0.06  0.27             -0.10
Population                       0.01           -0.02          -0.02         0.67   -0.03                 -0.02       -0.13    0.32 -0.08              0.66
thinness..1.19.years             0.02           -0.46           0.27         0.46   -0.40                 -0.26       -0.13    0.18 -0.55              0.46
thinness.5.9.years               0.01           -0.46           0.29         0.46   -0.39                 -0.26       -0.13    0.17 -0.55              0.46
Income.composition.of.resources  0.12            0.72          -0.44        -0.13    0.56                  0.40        0.18   -0.06  0.51             -0.15
Schooling                        0.09            0.73          -0.42        -0.21    0.62                  0.42        0.22   -0.12  0.55             -0.23
                                 Polio Total.expenditure Diphtheria HIV.AIDS   GDP Population thinness..1.19.years thinness.5.9.years
Year                             -0.02              0.06       0.03    -0.12  0.10       0.01                 0.02                 0.01
Life.expectancy                   0.33              0.17       0.34    -0.59  0.44      -0.02                -0.46                -0.46
Adult.Mortality                  -0.20             -0.09      -0.19     0.55 -0.26      -0.02                 0.27                 0.29
infant.deaths                    -0.16             -0.15      -0.16     0.01 -0.10       0.67                 0.46                 0.46
Alcohol                           0.24              0.21       0.24    -0.03  0.44      -0.03                -0.40                -0.39
percentage.expenditure            0.13              0.18       0.13    -0.10  0.96      -0.02                -0.26                -0.26
Hepatitis.B                       0.46              0.11       0.59    -0.09  0.04      -0.13                -0.13                -0.13
Measles                          -0.06             -0.11      -0.06     0.00 -0.06       0.32                 0.18                 0.17
BMI                               0.19              0.19       0.18    -0.21  0.27      -0.08                -0.55                -0.55
under.five.deaths                -0.17             -0.15      -0.18     0.02 -0.10       0.66                 0.46                 0.46
Polio                             1.00              0.12       0.61    -0.11  0.16      -0.05                -0.16                -0.17
Total.expenditure                 0.12              1.00       0.13     0.04  0.18      -0.08                -0.21                -0.22
Diphtheria                        0.61              0.13       1.00    -0.12  0.16      -0.04                -0.19                -0.18
HIV.AIDS                         -0.11              0.04      -0.12     1.00 -0.11      -0.03                 0.17                 0.18
GDP                               0.16              0.18       0.16    -0.11  1.00      -0.02                -0.28                -0.28
Population                       -0.05             -0.08      -0.04    -0.03 -0.02       1.00                 0.28                 0.28
thinness..1.19.years             -0.16             -0.21      -0.19     0.17 -0.28       0.28                 1.00                 0.93
thinness.5.9.years               -0.17             -0.22      -0.18     0.18 -0.28       0.28                 0.93                 1.00
Income.composition.of.resources   0.31              0.18       0.34    -0.25  0.45      -0.01                -0.45                -0.44
Schooling                         0.35              0.24       0.35    -0.21  0.47      -0.04                -0.49                -0.47
                                 Income.composition.of.resources Schooling
Year                                                        0.12      0.09
Life.expectancy                                             0.72      0.73
Adult.Mortality                                            -0.44     -0.42
infant.deaths                                              -0.13     -0.21
Alcohol                                                     0.56      0.62
percentage.expenditure                                      0.40      0.42
Hepatitis.B                                                 0.18      0.22
Measles                                                    -0.06     -0.12
BMI                                                         0.51      0.55
under.five.deaths                                          -0.15     -0.23
Polio                                                       0.31      0.35
Total.expenditure                                           0.18      0.24
Diphtheria                                                  0.34      0.35
HIV.AIDS                                                   -0.25     -0.21
GDP                                                         0.45      0.47
Population                                                 -0.01     -0.04
thinness..1.19.years                                       -0.45     -0.49
thinness.5.9.years                                         -0.44     -0.47
Income.composition.of.resources                             1.00      0.78
Schooling                                                   0.78      1.00
```

Slika 22. Korelacijska matrica (Snimka zaslona, 2020.)

Nažalost funkcija `cor()` vraća nam samo korelacijske koeficijente između varijabli. Zbog toga ćemo prikazati i izračun korelacijske matrice s p-vrijednostima. Potrebno je instalirati *Hmisc* paket i iskoristiti funkciju *`rcorr()`*. Matricu p-vrijednosti vidimo na Slika 24.

```
106 install.packages("Hmisc")
107 library("Hmisc")
108 res2 <- rcorr(as.matrix(dataKorelacija))
109 res2$P
110 res2$r
111 res2
```

Slika 23. Funkcija rcorr() (Snimka zaslona, 2020.)

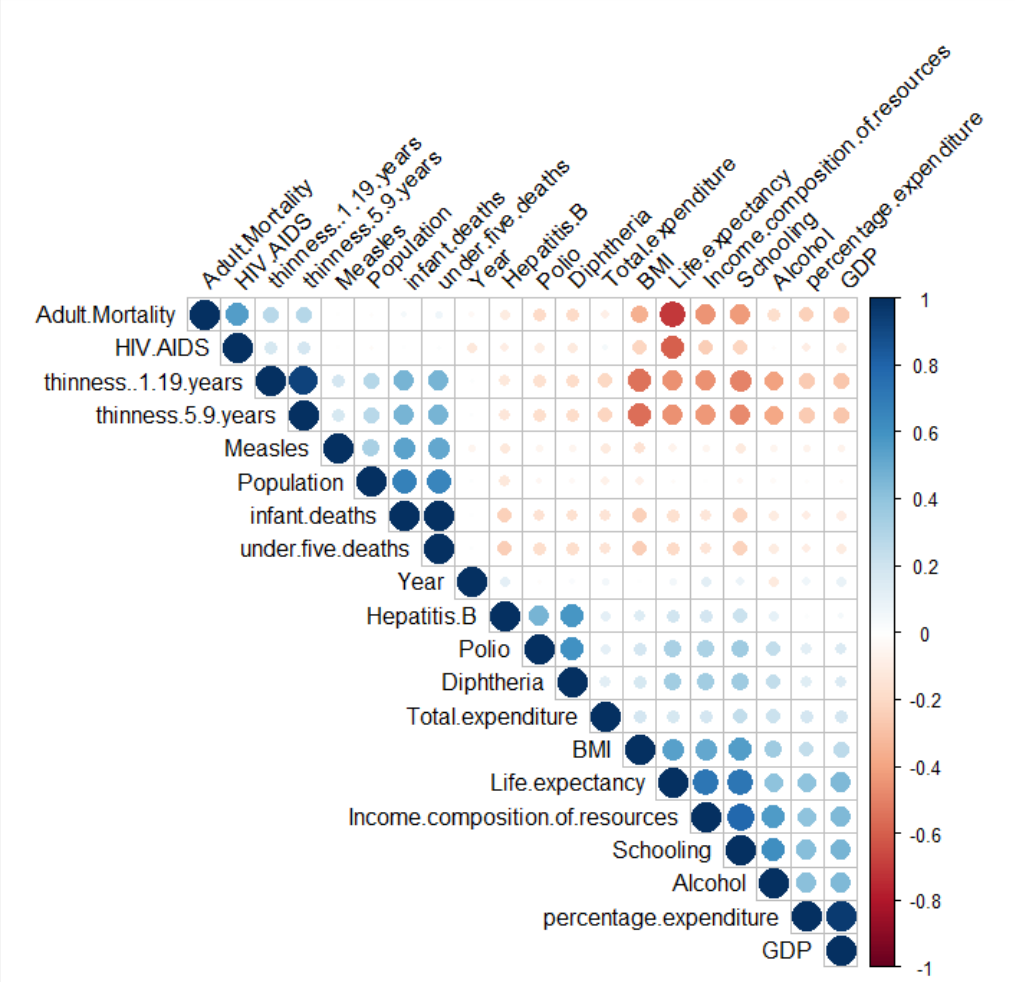
	Year	Life.expectancy	Adult.Mortality	infant.deaths	Alcohol	percentage.expenditure	Hepatitis.B	Measles
Year	NA	0.0000000	1.848578e-05	4.257452e-02	5.495436e-03	8.881567e-02	3.281079e-07	7.565579e-06
Life.expectancy	0.000000e+00	NA	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
Adult.Mortality	1.848578e-05	0.0000000	NA	1.986753e-05	0.000000e+00	0.000000e+00	1.776357e-15	9.166560e-02
infant.deaths	4.257452e-02	0.0000000	1.986753e-05	NA	1.239093e-09	3.367554e-06	0.000000e+00	0.000000e+00
Alcohol	5.495436e-03	0.0000000	0.000000e+00	1.239093e-09	NA	0.000000e+00	3.706709e-05	6.618874e-03
percentage.expenditure	8.881567e-02	0.0000000	0.000000e+00	3.367554e-06	0.000000e+00	NA	4.269710e-01	2.149003e-03
Hepatitis.B	3.281079e-07	0.0000000	1.776357e-15	0.000000e+00	3.706709e-05	4.269710e-01	NA	3.533331e-09
Measles	7.565579e-06	0.0000000	9.166560e-02	0.000000e+00	6.618874e-03	2.149003e-03	3.533331e-09	NA
BMI	3.921566e-09	0.0000000	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.834088e-13	0.000000e+00
under.five.deaths	1.994345e-02	0.0000000	3.334215e-07	0.000000e+00	3.583787e-09	1.850661e-06	0.000000e+00	0.000000e+00
Polio	3.463240e-07	0.0000000	0.000000e+00	0.000000e+00	0.000000e+00	1.332268e-15	0.000000e+00	1.489919e-13
Total.expenditure	2.211745e-06	0.0000000	1.859234e-09	1.782818e-11	0.000000e+00	0.000000e+00	6.297607e-03	2.925670e-08
Diphtheria	3.148592e-13	0.0000000	0.000000e+00	0.000000e+00	0.000000e+00	6.217249e-15	0.000000e+00	1.332268e-14
HIV.AIDS	2.775558e-14	0.0000000	0.000000e+00	1.715448e-01	1.049749e-02	1.068527e-07	3.441238e-08	9.403233e-02
GDP	3.745829e-07	0.0000000	0.000000e+00	5.834695e-08	0.000000e+00	0.000000e+00	1.634528e-04	1.338173e-04
Population	4.173942e-01	0.3035322	5.144795e-01	0.000000e+00	1.038404e-01	2.200179e-01	1.485790e-07	0.000000e+00
thinness..1.19.years	9.869528e-03	0.0000000	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	4.044164e-09	0.000000e+00
thinness.5.9.years	6.049171e-03	0.0000000	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.025443e-09	0.000000e+00
Income.composition.of.resources	0.000000e+00	0.0000000	0.000000e+00	1.620926e-14	0.000000e+00	0.000000e+00	0.000000e+00	7.578160e-12
Schooling	0.000000e+00	0.0000000	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	3.874678e-13

	BMI	under.five.deaths	Polio	Total.expenditure	Diphtheria	HIV.AIDS	GDP	Population
Year	3.921566e-09	1.994345e-02	3.463240e-07	2.211745e-06	3.148592e-13	2.775558e-14	3.745829e-07	4.173942e-01
Life.expectancy	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	3.035322e-01
Adult.Mortality	0.000000e+00	3.334215e-07	0.000000e+00	1.859234e-09	0.000000e+00	0.000000e+00	0.000000e+00	5.144795e-01
infant.deaths	0.000000e+00	0.000000e+00	0.000000e+00	1.782818e-11	0.000000e+00	1.715448e-01	5.834695e-08	0.000000e+00
Alcohol	0.000000e+00	3.583787e-09	0.000000e+00	0.000000e+00	0.000000e+00	1.049749e-02	0.000000e+00	1.038404e-01
percentage.expenditure	0.000000e+00	1.850661e-06	1.332268e-15	0.000000e+00	6.217249e-15	1.068527e-07	0.000000e+00	2.200179e-01
Hepatitis.B	1.834088e-13	0.000000e+00	0.000000e+00	6.297607e-03	0.000000e+00	3.441238e-08	1.634528e-04	1.485790e-07
Measles	0.000000e+00	0.000000e+00	1.489919e-13	2.925670e-08	1.332268e-14	9.403233e-02	1.338173e-04	0.000000e+00
BMI	NA	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	5.921647e-04
under.five.deaths	0.000000e+00	NA	0.000000e+00	1.020739e-11	0.000000e+00	3.911929e-02	2.050190e-08	0.000000e+00
Polio	0.000000e+00	0.000000e+00	NA	7.398526e-13	0.000000e+00	0.000000e+00	0.000000e+00	6.655204e-02
Total.expenditure	0.000000e+00	1.020739e-11	7.398526e-13	NA	1.332268e-15	9.423688e-01	1.940448e-11	2.322268e-04
Diphtheria	0.000000e+00	0.000000e+00	0.000000e+00	1.332268e-15	NA	0.000000e+00	0.000000e+00	1.757934e-01
HIV.AIDS	0.000000e+00	3.911929e-02	0.000000e+00	9.423688e-01	0.000000e+00	NA	7.947643e-12	1.830894e-01
GDP	0.000000e+00	2.050190e-08	0.000000e+00	1.940448e-11	0.000000e+00	7.947643e-12	NA	1.787417e-01
Population	5.921647e-04	0.000000e+00	6.655204e-02	2.322268e-04	1.757934e-01	1.830894e-01	1.787417e-01	NA
thinness..1.19.years	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
thinness.5.9.years	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
Income.composition.of.resources	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	6.764449e-01
Schooling	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.301149e-01

	thinness..1.19.years	thinness.5.9.years	Income.composition.of.resources	Schooling
Year	9.869528e-03	6.049171e-03	0.000000e+00	0.000000e+00
Life.expectancy	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
Adult.Mortality	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
infant.deaths	0.000000e+00	0.000000e+00	1.620926e-14	0.000000e+00
Alcohol	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
percentage.expenditure	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
Hepatitis.B	4.044164e-09	1.025443e-09	0.000000e+00	0.000000e+00
Measles	0.000000e+00	0.000000e+00	7.578160e-12	3.874678e-13
BMI	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
under.five.deaths	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
Polio	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
Total.expenditure	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
Diphtheria	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
HIV.AIDS	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
GDP	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
Population	0.000000e+00	0.000000e+00	6.764449e-01	1.301149e-01
thinness..1.19.years	NA	0.000000e+00	0.000000e+00	0.000000e+00
thinness.5.9.years	0.000000e+00	NA	0.000000e+00	0.000000e+00
Income.composition.of.resources	0.000000e+00	0.000000e+00	NA	0.000000e+00

Slika 24. Korelacijska matrica varijabli s p-vrijednostima (Snimka zaslona, 2020.)

Korelacijsku matricu možemo i vizualizirati s korelogramom (eng. *correlogram*). Potrebno je instalirati paket *corrplot*, a rezultat vizualizacije je na Slika 25.

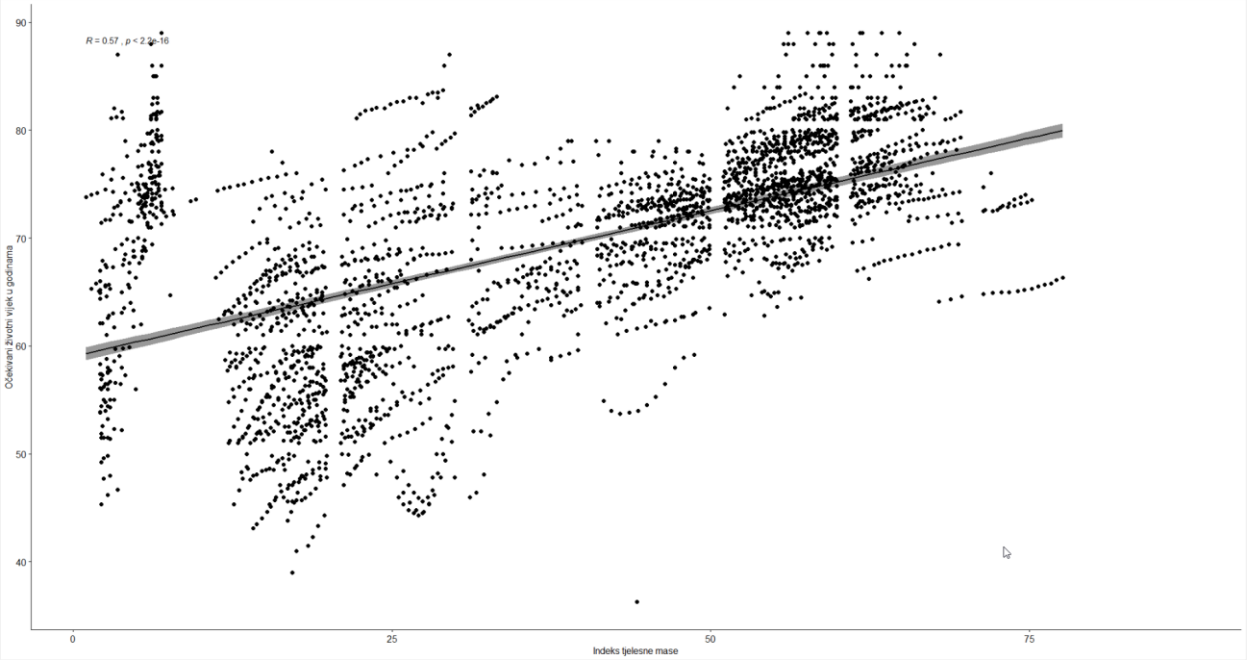


Slika 25. Korelogram (Snimka zaslona, 2020.)

U potpoglavlju 2.1.1.1 prikazali smo izračun korelacije između dvije varijable pomoću Pearsonovog korelacijskog testa [4]. Kako ne bi prikazivali izračun korelacije svake moguće kombinacije varijabli koristili smo matrice prikazane u ovom poglavlju, no postupak bi bio isti za sve kombinacije kvantitativnih varijabli.

2.1.1.1. Korelacija varijabli indeks tjelesne mase i očekivani životni vijek

Prvo ćemo vizualizirati podatke ove dvije varijable pomoću funkcije `ggscatter()` koja je dostupna instalacijom *ggpubr* paketa [5]. Na slici u nastavku vidimo raspršeni graf varijabli.



Slika 26. Raspršeni graf varijabli indeksa tjelesne mase i očekivanog životnog vijeka (Snimka zaslona, 2020.)

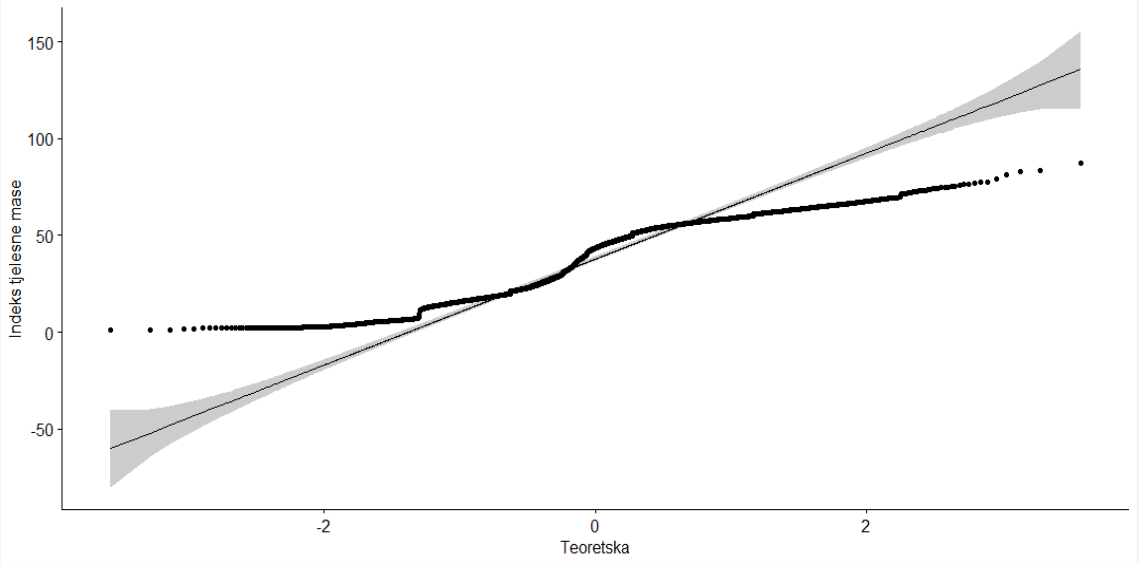
Iz grafa iznad vidimo da je odnos varijabli linearan. Sada trebamo provjeriti imaju li podaci obje varijable normalnu distribuciju. Za tu provjeru iskoristit ćemo Shapiro-Wilkov test. Nulta hipoteza nam je da su podaci normalno distribuirani, a alternativna hipoteza da podaci nisu normalno distribuirani.

```
> # Shapiro-Wilk normality test
> shapiro.test(data$BMI) #p-value < 2.2e-16

        Shapiro-Wilk normality test

data:  data$BMI
W = 0.93033, p-value < 2.2e-16

> # Shapiro-Wilk normality test
> shapiro.test(data$Life.expectancy) #p-value < 2.2e-16

        Shapiro-Wilk normality test

data:  data$Life.expectancy
W = 0.95605, p-value < 2.2e-16
```

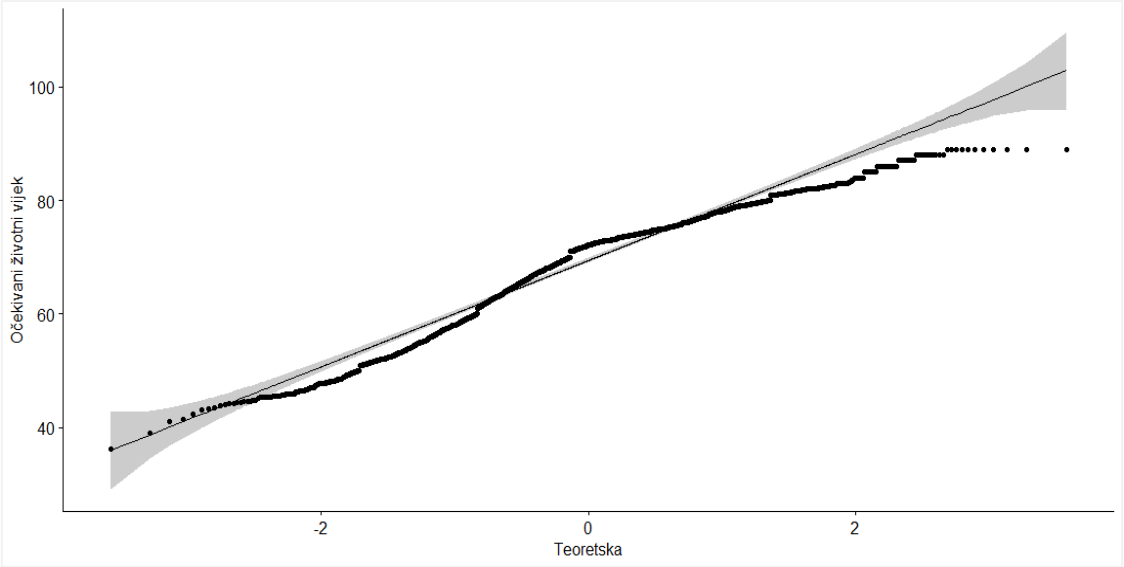
Slika 27. Shapiro-Wilkov test 1 (Snimka zaslona, 2020.)

Uzmemo li npr. razinu signifikantnosti od 5%, vidljivo je da su p-vrijednosti manje od 0,05 što implicira da su podaci signifikantno drugačije distribuirani od normalne distribucije, odnosno odbacujemo nultu hipotezu i prihvaćamo alternativnu.

Vizualnom inspekcijom, takozvanim Q-Q (kvantil-kvantil) grafom, prikazat ćemo korelaciju podataka odabranih varijabli i normalne distribucije.



Slika 28. Q-Q graf indeksa tjelesne mase (Snimka zaslona, 2020.)



Slika 29. Q-Q graf očekivanog životnog tijeka (Snimka zaslona, 2020.)

Iz prethodnih grafova zaključujemo da prva varijabla nije potpuno normalno distribuirana, ali ima linearan trend, dok druga varijabla može doći iz normalne distribucije, suprotno rezultatima Shapiro-Wilkovog testa. Pearsonov test korelacije provodimo naredbom:

```
cor.test(var1, var2, method = "pearson".
```

```
> res <- cor.test(data$BMI, data$Life.expectancy,
+                 method = "pearson")
> res

        Pearson's product-moment correlation

data:  data$BMI and data$Life.expectancy
t = 37.097, df = 2894, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5424873 0.5918785
sample estimates:
      cor
0.5676935
```

Slika 30. Rezultati Pearsonovog testa 1 (Snimka zaslona, 2020.)

Prema rezultatima sa slike iznad, p-vrijednost je puno manja od 0,05 ako promatramo razinu signifikantnosti od 5%, te zaključujemo da postoji signifikantna **pozitivna** korelacija između varijabli indeksa tjelesne mase i očekivanog životnog vijeka s koeficijentom korelacije 0,567.

2.2. Zadatak b

U zadatku b je bilo potrebno kreirati dvije nove varijable temeljem postojećih kvantitativnih varijabli. Prva varijabla koju je bilo potrebno kreirati je varijabla $Number\ of\ years\ schooling$ koja će biti kvalitativna varijabla i poprimat će sljedeće vrijednosti:

$$Number\ of\ years\ schooling\ kvalitativna = \begin{cases} 0 & 0 \leq Number\ of\ years\ schooling \leq 8 \\ 1 & 8 < Number\ of\ yers\ schooling \leq 12 \\ 2 & 12 < Number\ of\ years\ schooling \end{cases}$$

Varijablu smo kreirali pomoću stupca *Schooling* koji poprima vrijednosti od 0 do 22. Za kreiranje nove kvalitativne varijable koriste se sljedeće naredbe u R-u:

```
NumberOfYearsSchooling <- cut(Schooling, breaks=c(0, 8, 12, 22),
                              labels=c(0, 1, 2), as.factor.result=TRUE)

data1 = cbind(data, NumberOfYearsSchooling)
detach(data)
attach(data1)
```

Prilikom poziva naredbe `cut` odredili smo granice grupa nove kvalitativne varijable te njihove nazive. Parametas `as.factor.result` postavljen je na `TRUE` kako bi R znao da je nova varijabla `NumberOfYearsSchooling` kvalitativna varijabla. Nakon kreiranja nove varijable dodajemo je u skup podataka pomoću naredbe `cbind`.

Sljedeća varijabla koju je bilo potrebno kreirati je varijabla koja opisuje BMI (*Body Mass Index*) prema kategorijama:

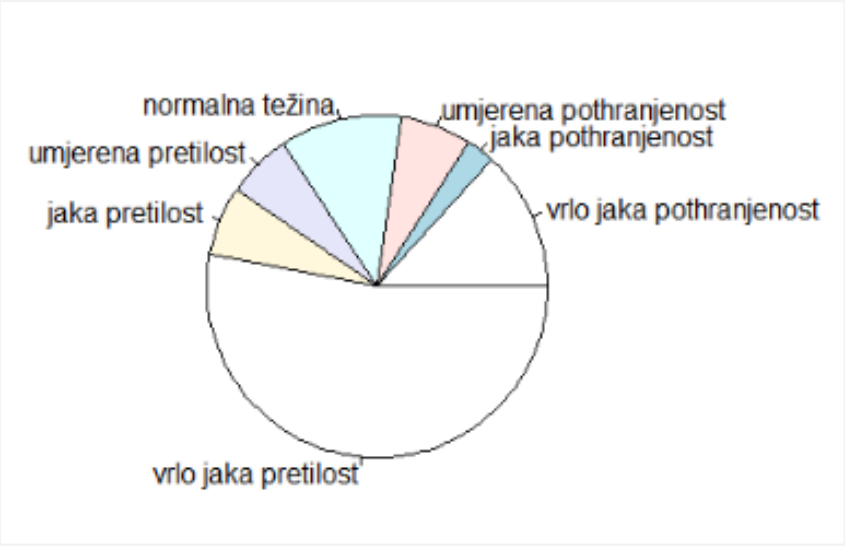
$$BMI\ kvalitativna = \begin{cases} vrlo\ jaka\ pothranjenost & 0 \leq BMI \leq 15 \\ jaka\ pothranjenost & 15 < BMI \leq 16 \\ umjeren\ pothranjenost & 16 < BMI \leq 18{,}5 \\ normalna\ težina & 18{,}5 < BMI \leq 25 \\ umjerena\ pretilost & 25 < BMI \leq 30 \\ jaka\ pretilost & 30 < BMI \leq 40 \\ vrlo\ jaka\ pretilost & BM > 40 \end{cases}$$

Varijablu smo kreirali na sličan način kao i prethodnu pomoću naredbe `cut` u R-u:

```
BMIqualitative <- cut(BMI, breaks=c(0,15,16,18.5,25,30,40,100),
                  labels = c("vrlo jaka pothranjenost", "jaka pothranjenost",
                  "umjerena     pothranjenost",     "normalna     težina",
                  "umjerena       pretilost",       "jaka       pretilost",
                  "vrlo jaka pretilost"))

data2 = cbind(data1, BMIqualitative)
detach(data1)
attach(data2)
```

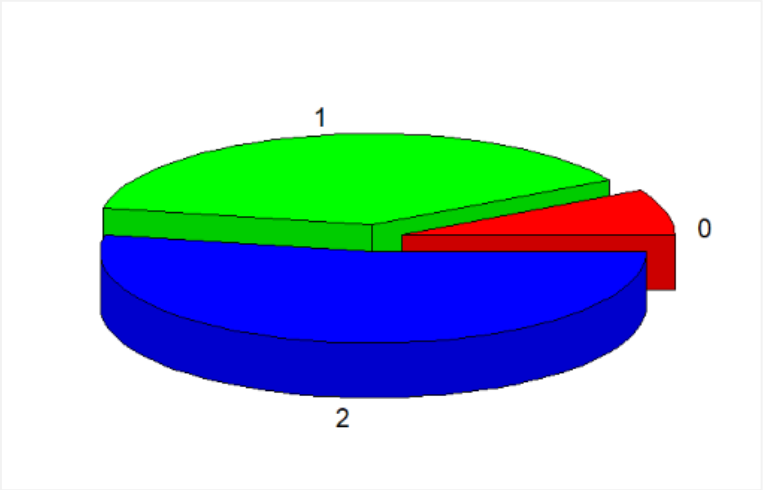
2.3. Zadatak c

U zadatku c je potrebno opisati dvije novokreirane varijable: *BMIqualitative* i *NumberOfYearsSchooling*. Varijabla *BMIqualitative* može poprimiti sedam različitih vrijednosti koje opisuju indeks tjelesne mase pojedinca. Granice za svaku od grupa navedene su u zadatku ranije, ali, ako je primjerice indeks tjelesne mase 32, tada će *BMIqualitative* varijabla poprimiti vrijednost „jaka pretilost“. Na Slika 31. prikazan je *pie-chart* prikaz novonastale kvalitativne varijable. Vidljivo je da je u ulaznom skupu podataka najčešća vrijednost „vrlo jaka pretilost“, a najrjeđa vrijednost „jaka pothranjenost.“



Slika 31. Kvalitativna varijabla *BMI kvalitativna* (Snimka zaslona, 2020.)

Sljedeća novokreirana varijabla je kvalitativna varijabla *NumberOfYearsSchooling*. Varijabla može poprimiti tri vrijednosti: 0, 1 ili 2 ovisno o duljini školovanja. Ukoliko je duljina školovanja kraća od 8, tada varijabla poprima vrijednost 0; ako je duljina između 8 i 12 godina varijabla poprima vrijednost 1; ako je duljina školovanja dulja od 12 godina, tada varijabla poprima vrijednost 2. Na *pie-chart* prikazu, Slika 32., vidljivo je da je najčešća duljina školovanja upravo ona iznad 12 godina.



Slika 32. Kvalitativna varijabla NumberOfYearsSchooling (Snimka zaslona, 2020.)

2.4. Izbacivanje nepoznatih vrijednosti

Prije nego što smo krenuli s provedbom hi-kvadrat testa, parametarskih i neparametarskih testova te regresijskom analizom, bilo je potrebno ukloniti sve primjere iz ulaznog skupa podataka koji ne poprimaju nikakvu vrijednost.

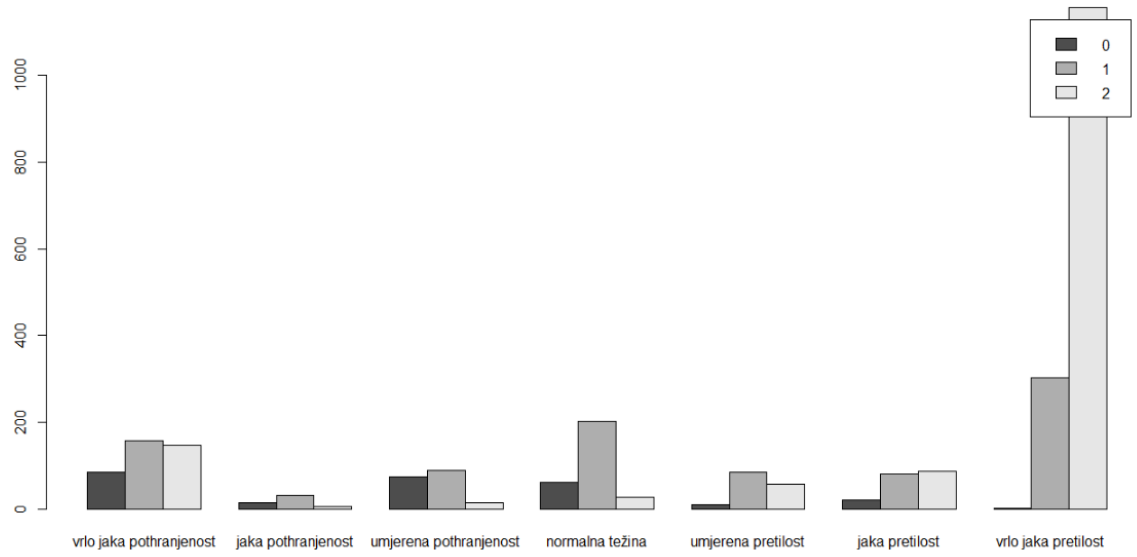
```
dataFinal <- data2[complete.cases(data2),]
detach(data2)
attach(dataFinal)
```

Primjere koji ne poprimaju nikakvu vrijednost uklonili smo pomoću R naredbe `complete.cases` koja vraća samo one primjere koji za svaku varijablu poprimaju neku vrijednost, odnosno, nemaju vrijednosti koje nedostaju.

3. Hi-kvadrat test

3.1. Zadatak d

Korištenjem hi-kvadrat testa ispitat ćemo povezanost kvalitativnih varijabli *Number of years schooling kvalitativna* i *BMI kvalitativna* iz prethodnog zadatka. Prvo smo tražene varijable spremili u tablicu, čiji *barplot* vidimo na slici ispod.



Slika 33. Barplot kvalitativnih varijabli *Number of years schooling kvalitativna* i *BMI kvalitativna*

Ovi grafovi bi trebali izgledati slično, ako nema ovisnosti o varijabli *BMI kvalitativna.* Vidimo na slici da baš i ne izgledaju slično, no idemo provjeriti samim testom. Nulta hipoteza nam je da ne postoji ovisnost *Number of years schooling kvalitativna* i *BMI kvalitativna*, a alternativna hipoteza nam je da postoji ovisnost kvalitativnih varijabli *Number of years schooling kvalitativna* i *BMI kvalitativna*.

```
209 tab <- table(NumberOfYearsSchooling, BMIqualitative)
210 tab
211 barplot(tab, beside = T, legend = T)
212
213 CTest <- chisq.test(tab, correct = TRUE)
214 CTest
```

Slika 34. Prvi hi-kvadrat test

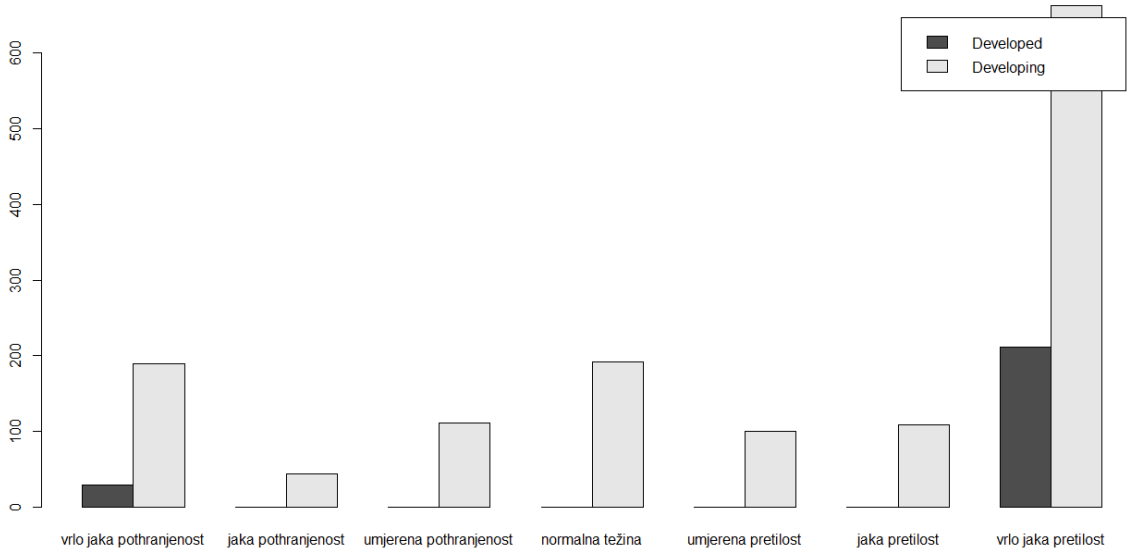
```
> CTest <- chisq.test(tab, correct = TRUE)
> CTest

        Pearson's Chi-squared test

data:  tab
X-squared = 1048, df = 12, p-value < 2.2e-16
```

Slika 35. Rezultati prvog hi-kvadrat testa

Prema rezultatima hi-kvadrat testa, uzmemo li npr. razinu signifikantnosti od 5%, vidimo da je p-vrijednost manja od 0,05, te stoga odbacujemo nultu hipotezu i prihvaćamo alternativnu hipotezu koja govori da postoji ovisnost kvalitativnih varijabli *Number of years schooling kvalitativna* i *BMI kvalitativna* na razini signifikantnosti od 5%.

3.2. Zadatak e

Korištenjem hi-kvadrat testa ispitat ćemo postoji li povezanost kvalitativnih statističkih varijabli *Status (Developed/Developing)* i *BMI kvalitativna* iz prethodnog zadatka. Prvo smo tražene varijable spremili u tablicu, čiji barplot vidimo na slici ispod.



Slika 36. Barplot kvalitativnih varijabli *Status (Developed/Developing)* i *BMI kvalitativna*

Ovi grafovi bi trebali izgledati slično, ako nema ovisnosti o varijabli *BMI kvalitativna*. Vidimo na slici da baš i ne izgledaju slično, no idemo provjeriti samim testom. Nulta hipoteza nam je da ne postoji ovisnost *Status (Developed/Developing)* i *BMI kvalitativna*, a alternativna hipoteza nam je da postoji ovisnost kvalitativnih varijabli *Status (Developed/Developing)* i *BMI kvalitativna*.

```
        Pearson's Chi-squared test

data:  tab
X-squared = 159.87, df = 6, p-value < 2.2e-16
```

Slika 37. Rezultati prvog hi-kvadrat testa

Prema rezultatima hi-kvadrat testa, uzmemo li npr. razinu signifikantnosti od 5%, vidimo da je p-vrijednost manja od 0,05, te stoga odbacujemo nultu hipotezu i prihvaćamo alternativnu hipotezu koja govori da postoji ovisnost kvalitativnih varijabli *Status (Developed/Developing)* i *BMI kvalitativna* na razini signifikantnosti od 5%.

4. Provedba parametarskih i neparametarskih testova

U ovom poglavlju bit će riješena četiri zadatka: f, g, h i j koji se tiču provedbe odgovarajućih parametarskih i neparametarskih testova.

4.1. Zadatak f

Korištenjem odgovarajućeg parametarskog i neparametarskog testa ispitajte postoje li razlike varijable Life expectancy in years po razinama (modalitetima) varijable Number of years schooling kvalitativna. Za provedbu parametarskog testa ispitajte pretpostavke za provedbu. Ukoliko se korištenjem parametarskog testa utvrdi da postoje signifikantne razlike provedite i post hoc test.

Prvi korak u rješavanju ovog zadatka je odabir adekvatnog parametarskog i neparametarskog testa. Budući da varijabla *NumberOfYearsSchooling* posjeduje 3 razine, odnosno 3 modaliteta, tada je kao parametarski test potrebno koristiti jednofaktorsku ANOVU, a kao neparametarski test Kruskal-Wallisov test. Sve testove ćemo provoditi i interpretirati na razini signifikantnosti od 5%.

Prije samog provođenja parametarskog testa, jednofaktorske ANOVE, potrebno je ispitati pretpostavke za provedbu: varijabla za koju provodimo test raspoređena je po normalnoj distribuciji u svakom skupu, razdiobe osnovnih skupova imaju jednake varijance i uzorci izabrani iz osnovnih skupova su nezavisni **Error! Reference source not found.**. Najprije ćemo provjeriti je li varijabla *Life expectancy* normalno distribuirana prema modalitetima varijable *NumberOfYearsSchooling.* Normalnost razdiobe provjeravat ćemo Shapiro-Wilkovim testom, a prije toga ćemo podatke rasporediti u 3 podskupa prema modalitetima varijable *NumberOfYearsSchooling*.

```
NumberOfYearsSchooling0 <- subset(dataFinal,
                                  subset = NumberOfYearsSchooling==0)
NumberOfYearsSchooling1 <- subset(dataFinal,
                                  subset = NumberOfYearsSchooling==1)
NumberOfYearsSchooling2 <- subset(dataFinal,
                                  subset = NumberOfYearsSchooling==2)

shapiro.test(NumberOfYearsSchooling0$Life.expectancy)
shapiro.test(NumberOfYearsSchooling1$Life.expectancy)
shapiro.test(NumberOfYearsSchooling2$Life.expectancy)
```

Rezultati testova prikazani su na slici u nastavku (Slika 38).

```
> shapiro.test(NumberOfYearsSchooling0$Life.expectancy)

        Shapiro-Wilk normality test

data:  NumberOfYearsSchooling0$Life.expectancy
W = 0.98134, p-value = 0.06306

> shapiro.test(NumberOfYearsSchooling1$Life.expectancy)

        Shapiro-Wilk normality test

data:  NumberOfYearsSchooling1$Life.expectancy
W = 0.9535, p-value = 2.429e-13

> shapiro.test(NumberOfYearsSchooling2$Life.expectancy)

        Shapiro-Wilk normality test

data:  NumberOfYearsSchooling2$Life.expectancy
W = 0.96793, p-value = 6.097e-13
```

Slika 38. Rezultati Shapiro-Wilkovog testa za f zadatak (Snimka zaslona, 2020.)

Ukoliko promatramo normalnost razdiobe za varijablu *Life expectancy* kada je vrijednost varijable *NumberOfYearsSchooling* jednaka 0 dobiva se p-vrijednost jednaka 0,06306 što je veće od 0,05 (odabrane razine signifikantnosti) pa zaključujemo da normalnost razdiobe vrijedi. Za ostala dva podskupa, kada su vrijednosti varijable *NumberOfYearsSchooling* 1 ili 2 ne vrijedi normalnost razdiobe jer je p-vrijednost puno manja od 0,05.

Sljedeći korak je ispitivanje jednakosti varijanci. Moguće je koristiti dva testa: Bartlettov test i Leveneov test.

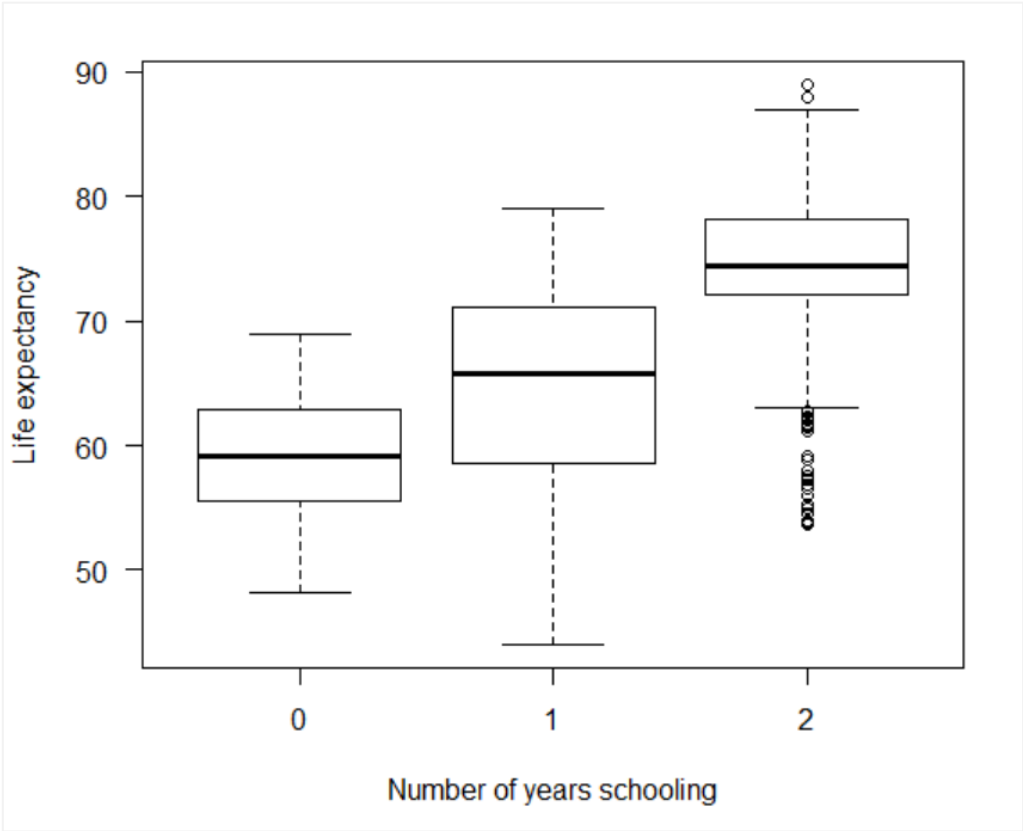
```
tapply(dataFinal$Life.expectancy, dataFinal$NumberOfYearsSchooling, var)
bartlett.test(Life.expectancy ~ NumberOfYearsSchooling, data=dataFinal)
library(car)
leveneTest(dataFinal$Life.expectancy,     dataFinal$NumberOfYearsSchooling,
center = mean)
```

Na slici u nastavku vidljivi su rezultati oba provedena testa (Slika 39). Najprije izračunavamo varijance redom po podskupovima i dobivamo vrijednosti: 25.5104, 66.21682 i 30.83572 te je već iz izračuna očito da varijance nisu jednake. Prilikom izvršavanja oba testa p-vrijednosti su puno manje od 0,05 što znači da jednakost po varijancama nije zadovoljena. Provest ćemo ANOVU.

```
> tapply(dataFinal$Life.expectancy, dataFinal$NumberOfYearsSchooling, var)
       0        1        2
25.55104 66.21682 30.83572
> bartlett.test(Life.expectancy ~ NumberOfYearsSchooling, data=dataFinal)

        Bartlett test of homogeneity of variances

data:  Life.expectancy by NumberOfYearsSchooling
Bartlett's K-squared = 127.04, df = 2, p-value < 2.2e-16

> library(car)
Loading required package: carData
> leveneTest(dataFinal$Life.expectancy, dataFinal$NumberOfYearsSchooling, center
 = mean)
Levene's Test for Homogeneity of Variance (center = mean)
        Df F value    Pr(>F)
group    2  74.711 < 2.2e-16 ***
      1646
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Slika 39. Rezultati Bartlettovog i Leveneovog testa za f zadatak (Snimka zaslona, 2020.)

ANOVU provodimo pomoću R naredba aov. Prije ANOVE ćemo pomoću boxplot-a prikazati koliko je očekivanje životnog vijeka prema modalitetima varijable *NumberOfYearsSchooling* (Slika 40). Na boxplot prikazu vidljiva je razlika medijana.



Slika 40. Box plot za f zadatak (snimka zaslona, 2020).

```
boxplot(Life.expectancy ~ NumberOfYearsSchooling, data=dataFinal,
        xlab = "Number of years schooling", ylab="Life expectancy",   las=1)
tapply(dataFinal$Life.expectancy, dataFinal$NumberOfYearsSchooling, mean)
tapply(dataFinal$Life.expectancy, dataFinal$NumberOfYearsSchooling, sd)


AnovaModel1 <- aov(Life.expectancy ~ NumberOfYearsSchooling, data=dataFinal)
summary(AnovaModel1)
```

Rezultati provedene ANOVE vidljivi su na slici u nastavku (Slika 41). Prije samog računanja ANOVE pomoću R naredbe `tapply` izračunata je srednja vrijednost i standardna devijacija. Vidljivo je da je srednja vrijednost očekivanja životnog vijeka u grupi koja se školovala od 0 do 8 godina 59 godina, u grupi koja se školovala od 8 do 12 godina 64 godine, a u grupi koja se školovala više od 12 godina skoro 75 godina. P-vrijednost je puno manja od 0,05 što znači da je analiza varijance signifikantna, odnosno, postoje signifikantne razlike među grupama.

```
> tapply(dataFinal$Life.expectancy, dataFinal$NumberOfYearsSchooling, mean)
       0        1        2
59.06418 64.24212 74.58318
> tapply(dataFinal$Life.expectancy, dataFinal$NumberOfYearsSchooling, sd)
       0        1        2
5.054803 8.137372 5.552992
> AnovaModel1 <- aov(Life.expectancy ~ NumberOfYearsSchooling, data=dataFinal)
> summary(AnovaModel1)
                      Df Sum Sq Mean Sq F value Pr(>F)
NumberOfYearsSchooling    2  54833   27416   620.8 <2e-16 ***
Residuals              1646  72697      44
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Slika 41. Rezultati ANOVE za f zadatak (Snimka zaslona, 2020.)

Budući da je parametarski test pokazao da postoje signifikantne razlike, potrebno je provesti post hoc testove da se utvrdi među kojim grupama razlike postoje. Provodi se Tukeyev test.

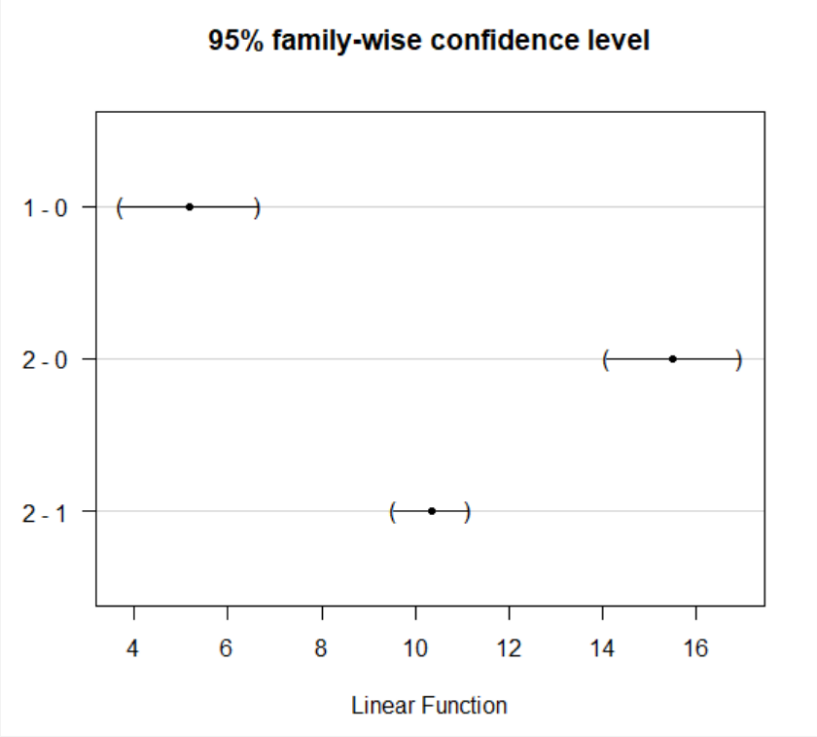
```
TukeyHSD(AnovaModel1)
library(multcomp)
.Pairs <- glht(AnovaModel1, linfct = mcp(NumberOfYearsSchooling = "Tukey"))
summary(.Pairs)
confint(.Pairs)
cld(.Pairs)
plot(confint(.Pairs))
```

Rezultati Tukeyevog testa vidljivi su na Slika *42*. Budući da su sve p vrijednosti manje od 0,05 zaključujemo da postoje signifikantne razlike među svim grupama, odnosno između grupe 1 i 0, grupe 2 i 0 i grupe 2 i 1. Dodatnim naredbama moguće je vidjeti još neke dodatne statistike.

```
> TukeyHSD(AnovaModel1)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Life.expectancy ~ NumberOfYearsSchooling, data = dataFinal)

$NumberOfYearsSchooling
         diff       lwr       upr p adj
1-0  5.177943  3.697104  6.658781     0
2-0 15.519002 14.072695 16.965308     0
2-1 10.341059  9.530360 11.151758     0
```

Slika *42*. Tukeyev post hoc test za f zadatak (Snimka zaslona, 2020.)

Moguće je iscrtati intervale pouzdanosti te pomoću naredbe cld() vidjeti dodijeljena slova grupi, tj. naredba će svakoj grupi dodijeliti slovo ovisno o tome postoje li ili ne signifikatne razlike. U ovom zadatku svaka od grupa će poprimiti svoje slovo. Intervali pouzdanosti vidljivi su na Slika 43.



Slika 43. Intervali pouzdanosti za f zadatak (Snimka zaslona, 2020.)

Sljedeći zadatak je provesti odgovarajući neparametarski test, a to je, kao što je već ranije navedeno, Kruskal-Wallisov test. Kruskal-Wallisov test provodimo naredbom:

```
kruskal.test(Life.expectancy ~ NumberOfYearsSchooling, data=dataFinal)
```

Vidljivo je da je p-vrijednost puno manja od 0,05 te ponovno zaključujemo da postoje signifikantne razlike među grupama (Slika 44.).

```
> kruskal.test(Life.expectancy ~ NumberOfYearsSchooling, data=dataFinal)

        Kruskal-Wallis rank sum test

data:  Life.expectancy by NumberOfYearsSchooling
Kruskal-Wallis chi-squared = 771.81, df = 2, p-value < 2.2e-16
```

Slika 44. Rezultat Kruskal-Wallisovog testa za f zadatak (Snimka zaslona, 2020.)

4.2. Zadatak g

Korištenjem odgovarajućeg parametarskog i neparametarskog testa ispitajte postoje li razlike varijable Life expectancy in years po razinama (modalitetima) varijable BMI kvalitativna. Za provedbu parametarskog testa ispitajte pretpostavke za provedbu. Ukoliko se korištenjem parametarskog testa utvrdi da postoje signifikantne razlike provedite i post hoc test.

Zadatak g će se riješiti na veoma sličan način kao i f zadatak budući da se radi o jednofaktorskoj analizi varijance kao parametarskom testu i Kruskal-Wallisovom testu kao neparametarskom testu. Testove ponovno provodimo na razini signifikantnosti od 5%. Budući da će se koristiti jednake R naredbe kao i u f zadatku bit će prikazane one najbitnije te njihovi rezultati. Cijeli zadatak riješen u R-u nalazi se u prilozima na kraju dokumenta.

Najprije je potrebno kreirati podskupove prema modalitetima koje poprima novokreirana varijabla *BMIqualitative*.

```
BMI0 <- subset(dataFinal, subset = BMIqualitative=="vrlo jaka pothranjenost")
BMI1 <- subset(dataFinal, subset = BMIqualitative=="jaka pothranjenost")
BMI2 <- subset(dataFinal, subset = BMIqualitative=="umjerena pothranjenost")
BMI3 <- subset(dataFinal, subset = BMIqualitative=="normalna težina")
BMI4 <- subset(dataFinal, subset = BMIqualitative=="umjerena pretilost")
BMI5 <- subset(dataFinal, subset = BMIqualitative=="jaka pretilost")
BMI6 <- subset(dataFinal, subset = BMIqualitative=="vrlo jaka pretilost")
```

Nakon kreiranja skupova potrebno je ispitati njihovu normalnost razdiobe pomoću Shapiro-Wilkovog testa. Rezultati Shapiro-Wilkovog testa vidljivi su na Slika 45. Podsjetimo se, ukoliko je p-vrijednost veća od 0,05 tada postoji normalnost razdiobe. Prema rezultatima vidljivo je da normalnost razdiobe postoji kada vrijednost varijable *BMIqualitative* poprima vrijednost „jaka pothranjenost“ i „umjerena pothranjenost“, a za ostale grupe uvjet normalnosti nije zadovoljen.

```
> shapiro.test(BMI0$Life.expectancy)

        Shapiro-Wilk normality test

data:  BMI0$Life.expectancy
W = 0.9791, p-value = 0.002468

> shapiro.test(BMI1$Life.expectancy)

        Shapiro-Wilk normality test

data:  BMI1$Life.expectancy
W = 0.96183, p-value = 0.1523

> shapiro.test(BMI2$Life.expectancy)

        Shapiro-Wilk normality test

data:  BMI2$Life.expectancy
W = 0.98109, p-value = 0.1172

> shapiro.test(BMI3$Life.expectancy)

        Shapiro-Wilk normality test

data:  BMI3$Life.expectancy
W = 0.98244, p-value = 0.01661

> shapiro.test(BMI4$Life.expectancy)

        Shapiro-Wilk normality test

data:  BMI4$Life.expectancy
W = 0.88439, p-value = 2.797e-07

> shapiro.test(BMI5$Life.expectancy)

        Shapiro-Wilk normality test

data:  BMI5$Life.expectancy
W = 0.94505, p-value = 0.0002067

> shapiro.test(BMI6$Life.expectancy)

        Shapiro-Wilk normality test

data:  BMI6$Life.expectancy
W = 0.9822, p-value = 7.867e-09
```

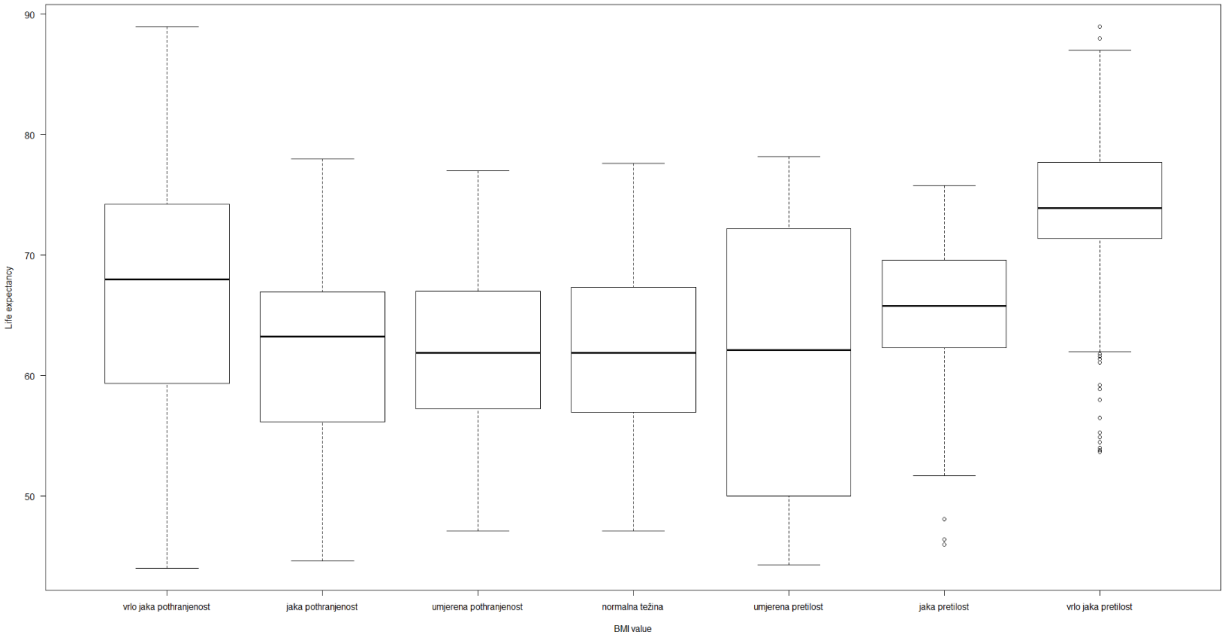
Slika 45. Rezultati Shapiro-Wilkovog testa za g zadatak (Snimka zaslona, 2020.)

Sljedeći uvjet koji mora biti ispunjen prije provedbe ANOVE je jednakost varijanci. Jednakost varijanci testirat ćemo pomoću Bartlettovog i Levenovog testa. U rezultatima oba testa, vidljivo je da je p-vrijednost manja od 0,05 pa zaključujemo da jednakost varijanci ne vrijedi (Slika 46.)

```
> tapply(dataFinal$Life.expectancy, dataFinal$BMIqualitative, var)
vrlo jaka pothranjenost      jaka pothranjenost  umjerena pothranjenost
               89.78632                57.10742                47.31307
        normalna težina      umjerena pretilost          jaka pretilost
               53.42741               121.71285                39.85458
    vrlo jaka pretilost
               29.84961
> bartlett.test(Life.expectancy ~ BMIqualitative, data=dataFinal)

        Bartlett test of homogeneity of variances

data:  Life.expectancy by BMIqualitative
Bartlett's K-squared = 200.53, df = 6, p-value < 2.2e-16

> library(car)
> leveneTest(dataFinal$Life.expectancy, dataFinal$BMIqualitative, center = mean)
Levene's Test for Homogeneity of Variance (center = mean)
        Df F value    Pr(>F)
group    6  51.684 < 2.2e-16 ***
      1642
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Slika 46. Rezultati Bartlettovog i Levenovog testa za g zadatak (Snimka zaslona, 2020).

Sada provodimo ANOVU te iscrtavamo boxplot prema modalitetima koje poprima varijabla *BMIqualitative* (Slika 47). Na boxplot prikazu možemo primijetiti da su razlike između medijana svake grupe različite, tj. postoji veća razlika u medijanima grupa koje za varijablu *BMIqualitative* poprimaju vrijednosti „vrlo jaka pothranjenost“, „jaka pretilost“ i „vrlo jaka pretilost“ naspram grupa koje za varijablu *BMIqualitative* poprimaju vrijednosti „jaka pothranjenost“, „umjerena pothranjenost“, „normalna težina“ i „umjerena pretilost“.



Slika 47. Boxplot prikaz za g zadatak (Snimka zaslona, 2020.)

Rezultati ANOVE pokazuju da postoji signifikantna razlika u grupama budući da je p-vrijednost značajno manja od 0,05.

```
> AnovaModel2 <- aov(Life.expectancy ~ BMIqualitative, data=dataFinal)
> summary(AnovaModel2)
                Df Sum Sq Mean Sq F value Pr(>F)
BMIqualitative    6  47679    7946   163.4 <2e-16 ***
Residuals      1642  79851      49
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Slika 48. Rezultati ANOVE za g zadatak (Snimka zaslona, 2020.)

Budući da postoji signifikantna razlika između grupa potrebno je provesti post hoc test, Tukeyev test kako bi se otkrilo među kojim grupama postoji signifikantna razlika. Na Slika 49. vidljive su p-vrijednosti prema grupama.

	upr	p adj
jaka pothranjenost-vrlo jaka pothranjenost	-2.1946640	0.0000269
umjerena pothranjenost-vrlo jaka pothranjenost	-3.2798216	0.0000000
normalna težina-vrlo jaka pothranjenost	-3.3032934	0.0000000
umjerena pretilost-vrlo jaka pothranjenost	-3.3191694	0.0000000
jaka pretilost-vrlo jaka pothranjenost	0.7760465	0.4132817
vrlo jaka pretilost-vrlo jaka pothranjenost	8.3212357	0.0000000
umjerena pothranjenost-jaka pothranjenost	3.5845538	1.0000000
normalna težina-jaka pothranjenost	3.6977539	0.9999904
umjerena pretilost-jaka pothranjenost	3.5160195	0.9999983
jaka pretilost-jaka pothranjenost	7.6354014	0.0253205
vrlo jaka pretilost-jaka pothranjenost	15.5418024	0.0000000
normalna težina-umjerena pothranjenost	2.7943725	0.9996417
umjerena pretilost-umjerena pothranjenost	2.7128773	0.9999996
jaka pretilost-umjerena pothranjenost	6.8172119	0.0003660
vrlo jaka pretilost-umjerena pothranjenost	14.5182431	0.0000000
umjerena pretilost-normalna težina	2.0735370	0.9982213
jaka pretilost-normalna težina	6.1703210	0.0002067
vrlo jaka pretilost-normalna težina	13.7449036	0.0000000
jaka pretilost-umjerena pretilost	7.0172031	0.0003392
vrlo jaka pretilost-umjerena pretilost	14.7424674	0.0000000
vrlo jaka pretilost-jaka pretilost	10.4937514	0.0000000

Slika 49. Rezultati Tukeyevog testa za g zadatak (Snimka zaslona, 2020.)

Ukoliko je p-vrijednost manja od 0,05 tada postoji signifikantna razlika između dvije grupe. Prema rezultatima je vidljivo da **ne postoji** signifikantna razlika među sljedećim grupama:

- Jaka pretilost – vrlo jaka pothranjenost
- Umjerena pothranjenost – jaka pothranjenost
- Normalna težina – jaka pothranjenost
- Umjerena pretilost – jaka pothranjenost
- Normalna težina – umjerena pothranjenost
- Umjerena pretilost – umjerena pothranjenost
- Umjerena pretilost – normalna težina

Među ostalim grupama postoji signifikantna razlika što možemo i vidjeti pomoću R naredbe `cld()` čiji rezultat je vidljiv na Slika 49. Jasno je da se rezultati naredbe slažu s rezultatima provedenog Tukeyevog testa.

```
> cld(.Pairs)
vrlo jaka pothranjenost      jaka pothranjenost
                    "b"                     "a"
 umjerena pothranjenost         normalna težina
                    "a"                     "a"
     umjerena pretilost          jaka pretilost
                    "a"                     "b"
    vrlo jaka pretilost
                    "c"
```

Slika 50. Rezultat naredbe cld za g zadatak (Snimka zaslona, 2020.)

Vidljivo je da su grupe između kojih nema signifikantnih razlika označene istim slovima:

- „a“: umjerena pothranjenost, umjerena pretilost, jaka pothranjenost, normalna težina
- „b“: vrlo jaka pothranjenost, jaka pretilost
- „c“: vrlo jaka pretilost

Posljednje što je potrebno je provesti neparametarski test, Kruskal-Wallisov test. Test provodimo naredbom:

```
kruskal.test(Life.expectancy ~ BMIqualitative, data=dataFinal)
```

te je rezultat testa vidljiv na **Error! Reference source not found.**. P-vrijednost je manja od 0,05 što znači da postoje signifikantne razlike među grupama što je i pokazano ANOVA testom.

```
> kruskal.test(Life.expectancy ~ BMIqualitative, data=dataFinal)

        Kruskal-Wallis rank sum test

data:  Life.expectancy by BMIqualitative
Kruskal-Wallis chi-squared = 620.42, df = 6, p-value <
2.2e-16
```

Slika 51. Rezultati Kruskal-Wallisovog testa za g zadatak (Snimka zaslona, 2020.)

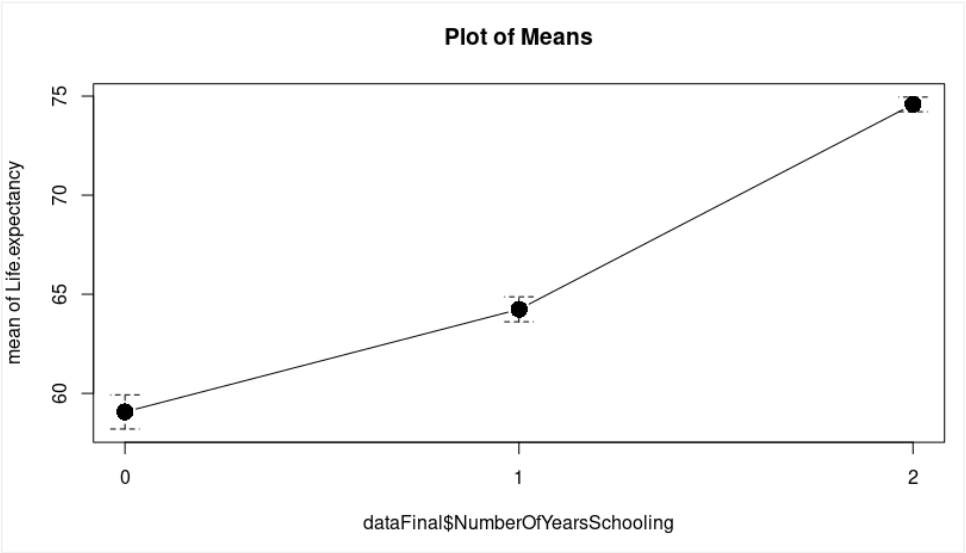
4.3. Zadatak h

Korištenjem dvofaktorske analize varijance ispitajte postoje li signifikantne razlike varijable Life expectancy in years po tretmanima varijabli Status i NumberOfYearsSchooling kvalitativna.

```
> AnovaModel <- (lm(Life.expectancy ~ Status * dataFinal$NumberOfYearsSchooling))
> Anova(AnovaModel)
Note: model has aliased coefficients
      sums of squares computed by model comparison
Anova Table (Type II tests)

Response: Life.expectancy
                                       Sum Sq   Df  F value  Pr(>F)
Status                                   5952    1 146.9544 < 2e-16 ***
dataFinal$NumberOfYearsSchooling        35780    2 441.7298 < 2e-16 ***
Status:dataFinal$NumberOfYearsSchooling   164    1   4.0493 0.04435 *
Residuals                               66581 1644
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Slika 52. Dvofaktorska analiza varijance (Snimka zaslona, 2020.)

Provođenjem dvofaktorske analize varijance, zaključuje se da postoje signifikantne razlike u ovisnosti o varijabli *Status* i varijabli *NumberOfYearsSchooling*. Također se vidi da na razini signifikantnosti od 5%, ali granično, postoji interakcija između varijabli *Status* i *NumberOfYearsSchooling*. To se sve zaključilo na temelju p-vrijednosti sa Slika 52.



Slika 53. Prikaz srednjih vrijednosti (Snimka zaslona, 2020)

Sa Slika 53 je vidljivo da postoje signifikantne razlike u ovisnosti varijable *Life expectancy* o tretmanima varijable *NumberOfYearsSchooling*. Intervali pouzdanosti se ne preklapaju.

4.4. Zadatak i

Korištenjem odgovarajućeg parametarskog i neparametarskog testa ispitajte postoje li razlike varijable Life expectancy in years po razinama (modalitetima) varijable Status. Za provedbu parametarskog testa ispitajte pretpostavke za provedbu.

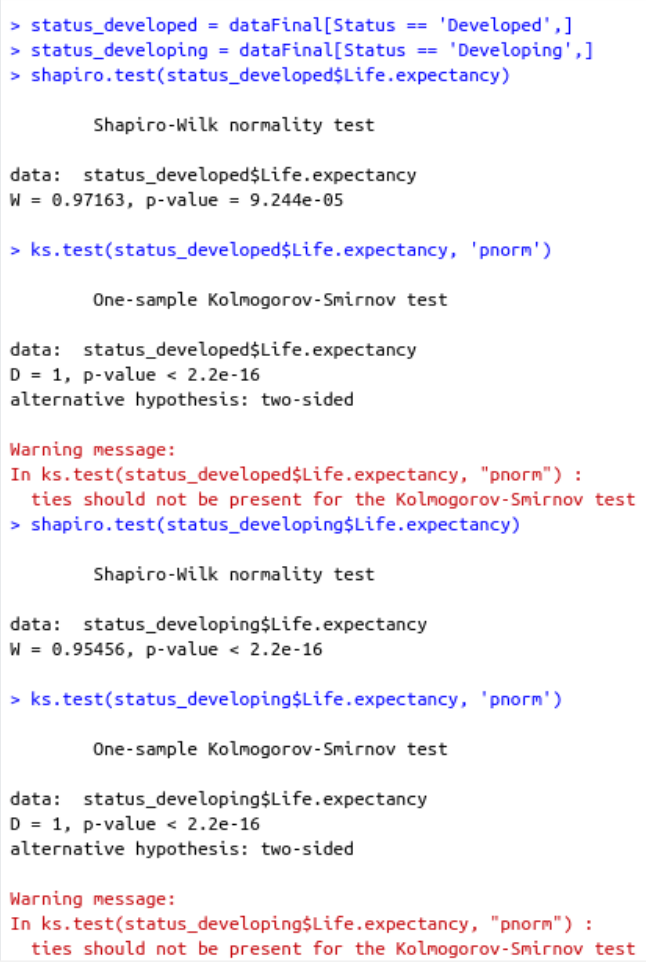
Kako su ove dvije varijable nezavisne, ovaj zadatak riješit će se pomoću parametarskog i neparametarskog testa za nezavisne uzorke.

```
> tapply(Life.expectancy, Status, median)
 Developed Developing
     78.95      69.20
> wilcox.test(Life.expectancy ~ Status, alternative = 'greater')

        Wilcoxon rank sum test with continuity correction

data:  Life.expectancy by Status
W = 304438, p-value < 2.2e-16
alternative hypothesis: true location shift is greater than 0
```

Slika 54. Rezultati neparametarskog testa za zadatak i (Snimka zaslona, 2020.)

Za početak će se provjeriti medijan varijable *Life expectancy* s obzirom na modalitete varijable *Status*. Vidi se, na Slika 54, da postoji relativno velika razlika između medijana (78.95 i 69.20). Na prvu bi se odmah moglo zaključiti da postoji značajna razlika, ali to će se svejedno provjeriti korištenjem odgovarajućih testova. Prvi od njih je Mann-Whitney-Wilcoxon neparametarski test za nezavisne uzorke. Taj se test provodi pomoću funkcije `wilcox.test()`. Dobivena p-vrijednost iznosi 2.2e-16 što znači da se s velikom sigurnošću odbacuje nulta hipoteza i prihvaća alternativna koja navodi da je razlika između uzoraka veća od nule (test na gornju granicu).

Sada se prelazi na parametarske testove za nezavisne uzorke. Prije samo testiranja potrebno je napraviti dvije stvari. Prvo je potrebno podijeliti početni skup podataka na onoliko podskupova koliko modaliteta ima varijabla *Status*. To je u ovom slučaju dva. Prvi podskup će sadržavati samo one opservacije koje imaju modalitet "*Developed*", a drugi podskup one koje imaju modalitet "*Developing*". Kao drugo, zbog nezavisnosti uzoraka, potrebno je provjeriti normalnost razdiobe po grupama/podskupovima i ispitati jednakost varijanci.



Slika 55. Ispitivanje pretpostavki za zadatak i (Snimka zaslona, 2020.)

Vidi se, na Slika 55, da je na oba testa, Shapiro-Wilk i Kolmogorov-Smirnov, i za oba uzorka p-vrijednost jako mala. Dakle, i na razini signifikantnosti od 5% i od 1%, odbacuje se nulta hipoteza prema kojoj su podaci normalno distribuirani. To bi značilo da nema smisla nastavljati dalje s parametarskim testom, ali svejedno će se ovdje još provjeriti kakvi su rezultati kod testova za provjeru jednakosti varijanci.

```
> tapply(Life.expectancy, Status, var)
 Developed Developing
  18.26267   69.78903
> var.test(Life.expectancy ~ Status, alternative = 'two.sided', conf.level = .95)

        F test to compare two variances

data:  Life.expectancy by Status
F = 0.26168, num df = 241, denom df = 1406, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.2170738 0.3197240
sample estimates:
ratio of variances
          0.261684

> bartlett.test(Life.expectancy ~ Status)

        Bartlett test of homogeneity of variances

data:  Life.expectancy by Status
Bartlett's K-squared = 134.6, df = 1, p-value < 2.2e-16

> leveneTest(Life.expectancy, Status, center = mean)
Levene's Test for Homogeneity of Variance (center = mean)
        Df F value    Pr(>F)
group    1  109.83 < 2.2e-16 ***
      1647
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Warning message:
In leveneTest.default(Life.expectancy, Status, center = mean) :
  Status coerced to factor.
```

Slika 56. Provjera jednakosti varijanci za zadatak i (Snimka zaslona, 2020.)

Za početak se s funkcijom `tapply()` provjeri varijanca varijable *Life expectancy* po modalitetima varijable *Status*. Vidljivo je, sa Slika 56, da postoji značajna razlika između njih. Taj zaključak se može potvrditi i u nastavku, provođenjem F, Bartlettovog i Levenovog testa za provjeru jednakosti varijanci. Za sva tri testa p-vrijednost iznosi 2.2e-16 čime se odbacuje nulta hipoteza kojom se tvrdi da su varijance između uzoraka/podskupova jednake.

5. Regresijska analiza

5.1. Zadatak j

Provedite regresijsku analizu kod koje će zavisna varijabla biti Life expectancy in years, a nezavisne sve ostale kvantitativne varijable. Komentirajte parametre regresije: koeficijent determinacije i korigirani koeficijent determinacije. Interpretirajte skupni i pojedinačne testove signifikantnosti regresije za svaku od promatranih nezavisnih varijabli. Komentirajte rezultate regresije. Provedite izbor varijabli koristeći neku od metoda za izbor varijabli (Forward Selection Procedure, Backward Selection Procedure, Backward/Forward). Nacrtajte normalni prikaz rezidualnih vrijednosti.

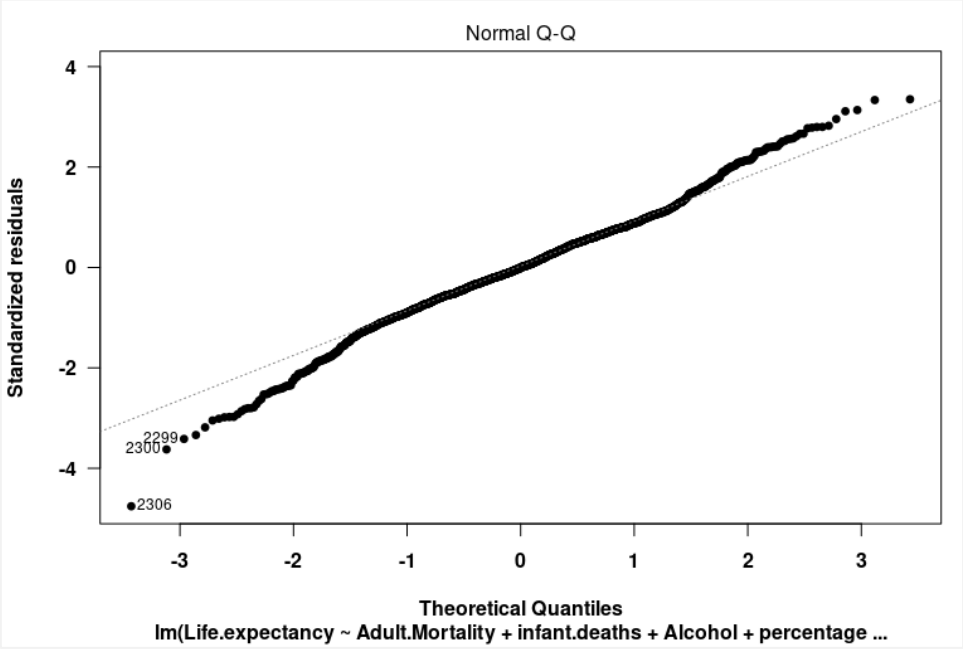
```
> RegModelLifeExpectancy <- lm(Life.expectancy ~ Adult.Mortality + infant.deaths + Alcohol + percentage.expenditure + Hepatitis.B + Measles + BMI + under.five.deaths + Polio
 + Total.expenditure + Diphtheria + HIV.AIDS + GDP + Population + thinness..1.19.years + thinness.5.9.years + Income.composition.of.resources + Schooling, data = dataFinal)
> summary(RegModelLifeExpectancy)

Call:
lm(formula = Life.expectancy ~ Adult.Mortality + infant.deaths +
    Alcohol + percentage.expenditure + Hepatitis.B + Measles +
    BMI + under.five.deaths + Polio + Total.expenditure + Diphtheria +
    HIV.AIDS + GDP + Population + thinness..1.19.years + thinness.5.9.years +
    Income.composition.of.resources + Schooling, data = dataFinal)

Residuals:
     Min       1Q   Median       3Q      Max
-17.0176  -2.0454  -0.0185   2.2260  11.9157

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     5.328e+01  7.358e-01  72.412  < 2e-16 ***
Adult.Mortality                -1.689e-02  9.473e-04 -17.828  < 2e-16 ***
infant.deaths                   9.369e-02  1.068e-02   8.776  < 2e-16 ***
Alcohol                        -5.435e-02  3.061e-02  -1.776   0.0760 .
percentage.expenditure          3.777e-04  1.805e-04   2.093   0.0365 *
Hepatitis.B                    -5.582e-03  4.446e-03  -1.256   0.2095
Measles                        -8.617e-06  1.081e-05  -0.797   0.4253
BMI                             3.350e-02  6.011e-03   5.573 2.92e-08 ***
under.five.deaths              -7.047e-02  7.728e-03  -9.119  < 2e-16 ***
Polio                           7.836e-03  5.163e-03   1.518   0.1293
Total.expenditure               7.975e-02  4.074e-02   1.958   0.0505 .
Diphtheria                      1.439e-02  5.938e-03   2.423   0.0155 *
HIV.AIDS                       -4.383e-01  1.788e-02 -24.519  < 2e-16 ***
GDP                             1.383e-05  2.838e-05   0.487   0.6260
Population                     -6.917e-10  1.753e-09  -0.395   0.6931
thinness..1.19.years           -8.670e-03  5.310e-02  -0.163   0.8703
thinness.5.9.years             -5.123e-02  5.242e-02  -0.977   0.3286
Income.composition.of.resources  9.824e+00  8.340e-01  11.780  < 2e-16 ***
Schooling                       8.783e-01  5.939e-02  14.789  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.596 on 1630 degrees of freedom
Multiple R-squared:  0.8347,    Adjusted R-squared:  0.8329
F-statistic: 457.4 on 18 and 1630 DF,  p-value: < 2.2e-16
```

Slika 57. Regresijska analiza (Snimka zaslona, 2020.)

Na Slika 57, vide se rezultati provedene regresijske analize u kojoj ima jedna zavisna varijabla (*Life expectancy*) i osamnaest nezavisnih. Koeficijent determinacije iznosi 83,47% što znači da je ovim modelom objašnjeno/protumačeno 83,47% varijacije podataka oko aritmetičke sredine. Korigirani koeficijent determinacije iznosi 83.29% i on je uvijek manji ili jednak "običnom" koeficijentu determinacije. Prema njegovoj vrijednosti može se zaključiti da ne postoji neka ovisnost između nezavisnih varijabli (skoro je isti kao i "obični" koeficijent determinacije). Sve u svemu, to je jedan zadovoljavajući postotak.

Kod pojedinačnih testova signifikantnosti regresije za svaku od promatranih nezavisnih varijabli vidi se da ih ima 9 koje su značajne/signifikantne, na razini signifikantnosti od 5%, dok ih je isto toliko koje na toj razini nisu signifikantne.

S druge strane, kod skupnog testa signifikantnosti (F-statistic), vidi se da je dobiveni regresijski model, kao cjelina, poprilično signifikantan, p-vrijednost mu je 2.2e-16.



Slika 58. Q-Q plot (Snimka zaslona, 2020.)

Na Slika 58, prikazan je Normal Q-Q plot, odnosno grafički prikaz reziduala u ovisnosti o kvantilima normalne distribucije. Vidi se da ne postoji normalnost reziduala. To se također može provjeriti i pomoću Shapiro-Wilkovog testa normalnosti za određeni regresijski model.

```
> shapiro.test(residuals(RegModelLifeExpectancy))

        Shapiro-Wilk normality test

data:  residuals(RegModelLifeExpectancy)
W = 0.98981, p-value = 2.464e-09
```

Slika 59. Shapiro-Wilkov test za zadatak j (Snimka zaslona, 2020.)

Na Slika 59, vidi se da je p-vrijednost vrlo mala (2.464e-09) što znači da se odbacuje nulta hipoteza koja navodi da postoji normalnost reziduala.

```
> step <- stepAIC(RegModelLifeExpectancy, direction = 'both')
Start:  AIC=4239.58
Life.expectancy ~ Adult.Mortality + infant.deaths + Alcohol +
    percentage.expenditure + Hepatitis.B + Measles + BMI + under.five.deaths +
    Polio + Total.expenditure + Diphtheria + HIV.AIDS + GDP +
    Population + thinness..1.19.years + thinness.5.9.years +
    Income.composition.of.resources + Schooling

                                 Df Sum of Sq   RSS    AIC
- thinness..1.19.years            1       0.3 21076 4237.6
- Population                      1       2.0 21078 4237.7
- GDP                             1       3.1 21079 4237.8
- Measles                         1       8.2 21084 4238.2
- thinness.5.9.years              1      12.3 21088 4238.5
- Hepatitis.B                     1      20.4 21096 4239.2
<none>                                       21076 4239.6
- Polio                           1      29.8 21106 4239.9
- Alcohol                         1      40.8 21117 4240.8
- Total.expenditure               1      49.5 21125 4241.5
- percentage.expenditure          1      56.6 21132 4242.0
- Diphtheria                      1      75.9 21152 4243.5
- BMI                             1     401.6 21477 4268.7
- infant.deaths                   1     995.8 22072 4313.7
- under.five.deaths               1    1075.1 22151 4319.6
- Income.composition.of.resources 1    1794.3 22870 4372.3
- Schooling                       1    2828.0 23904 4445.2
- Adult.Mortality                 1    4109.7 25186 4531.3
- HIV.AIDS                        1    7773.2 28849 4755.3
```

Slika 60. Backward/Forward Selection Procedure (Snimka zaslona, 2020.)

Pomoću Backward/Forward Selection Procedure, odnosno obostrane metode za izbor varijabli dobiveno je da postoji 6 varijabli koje bi trebalo izbaciti (varijable iznad oznake <*none*> na slici 25). To su one varijable koje ne doprinose puno kod varijacije zavisne varijable, tj. ne igraju veliku ulogu kod njezine procjene. Također se može primijetiti, iz jedne od prethodnih slika, da su to sve nesignifikantne varijable (na razini od 5%).

6. Zaključak

Ovdje treba sažeto rezimirati najvažnije rezultate razrade teme rada.

Popis literature

[1] „Life Expectancy (WHO) | Kaggle“. [Na internetu]. Dostupno na: https://www.kaggle.com/kumarajarshi/life-expectancy-who. [Pristupljeno: 12-lip-2020].

[2] „Subset distinct/unique rows — distinct • dplyr“. [Na internetu]. Dostupno na: https://dplyr.tidyverse.org/reference/distinct.html. [Pristupljeno: 16-lip-2020].

[3] „Correlation matrix : A quick start guide to analyze, format and visualize a correlation matrix using R software - Easy Guides - Wiki - STHDA“. [Na internetu]. Dostupno na: http://www.sthda.com/english/wiki/correlation-matrix-a-quick-start-guide-to-analyze-format-and-visualize-a-correlation-matrix-using-r-software. [Pristupljeno: 17-lip-2020].

[4] „Correlation Test Between Two Variables in R - Easy Guides - Wiki - STHDA“. [Na internetu]. Dostupno na: http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r. [Pristupljeno: 17-lip-2020].

[5] „ggpubr: Publication Ready Plots - Articles - STHDA“. [Na internetu]. Dostupno na: http://www.sthda.com/english/articles/24-ggpubr-publication-ready-plots/. [Pristupljeno: 17-lip-2020].

Popis slika