Analiza kriminala i socio-ekonomskih faktora

Jan Grgić Bernard Spiegl Korina Šimičević Dunja Šmigovec

Uvod

Cilj ovog projekta je analizirati i vizualno prikazati podatke o kriminalu na području Chicaga bilježene u periodu od godinu dana.

Učitavanje paketa

```
library(dplyr)
library(corrplot)
library(car)
library(MASS)
library(lmtest)
library(het.test)
```

Učitavanje podataka

```
chicago_year = read.csv("data/Crimes_-_One_year_prior_to_present.csv")
chicago_pc = read.csv("data/Chicago_poverty_and_crime.csv")
dim(chicago_pc)
```

```
## [1] 77 10
```

```
dim(chicago_year)
```

```
## [1] 216032     17
```

Kategorije kriminala u datasetu

```
categories <- unique(chicago_year$PRIMARY.DESCRIPTION)
categories
```

```
##  [1] "ASSAULT"                           "THEFT"
##  [3] "BATTERY"                           "CRIMINAL DAMAGE"
##  [5] "CRIMINAL TRESPASS"                 "ROBBERY"
##  [7] "DECEPTIVE PRACTICE"                "MOTOR VEHICLE THEFT"
##  [9] "NARCOTICS"                         "OFFENSE INVOLVING CHILDREN"
## [11] "WEAPONS VIOLATION"                 "CONCEALED CARRY LICENSE VIOLATION"
## [13] "SEX OFFENSE"                       "BURGLARY"
## [15] "OTHER OFFENSE"                     "INTERFERENCE WITH PUBLIC OFFICER"
## [17] "HOMICIDE"                          "PUBLIC PEACE VIOLATION"
## [19] "ARSON"                             "CRIMINAL SEXUAL ASSAULT"
## [21] "CRIM SEXUAL ASSAULT"               "INTIMIDATION"
## [23] "KIDNAPPING"                        "STALKING"
## [25] "PROSTITUTION"                      "GAMBLING"
```

```
## [27] "LIQUOR LAW VIOLATION"              "OBSCENITY"
## [29] "PUBLIC INDECENCY"                  "OTHER NARCOTIC VIOLATION"
## [31] "HUMAN TRAFFICKING"                 "NON-CRIMINAL"
```

```
length(categories)
```

```
## [1] 32
```

Prikaz frekvencija različitih kategorija zločina unazad godinu dana

U nastavku (Figure 1) možemo vidjeti da je su najčešći lakši oblici zločina što je bilo i za očekivati.

```
table <- table(chicago_year$PRIMARY.DESCRIPTION)
data <- as.data.frame.table(table)


df <- data[order(data$Freq, decreasing = TRUE),]
op <- par(mar = c(12, 4, 1, 2))
barplot(
  df$Freq,
  names.arg = df$Var1,
  las = 2,
  cex.names = 0.6,
  col = "skyblue",
  ylim = range(pretty(c(0, df$Freq))),
  main = "Frequency of different crime activities"
)
```

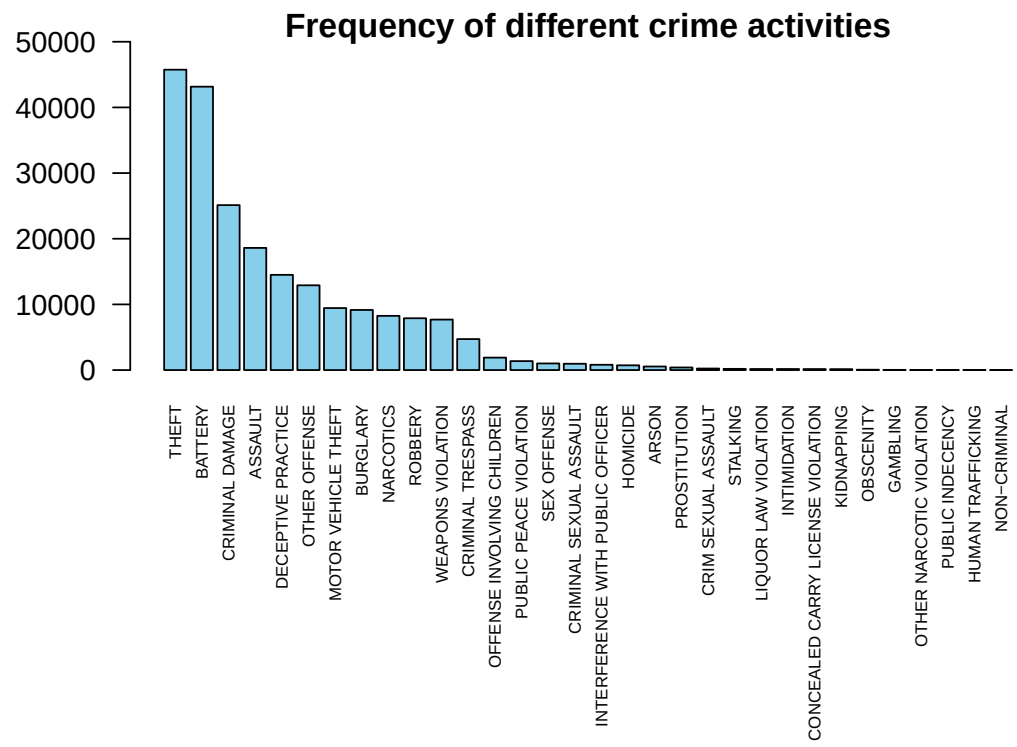

Figure 1: Frekventnost kategorija unazad godinu dana

Prikaz kategorija zločina koji su se pojavili manje od 2000 puta unazad godinu dana

Promotrimo li kategorije zločina koje s frekvencijama manjim od 2000 (Figure 2) uočavamo možemo vidjeti međusoban odnos frekventnosti težih kriminalnih radnji.

```
dfs <- df[df$Freq < 2000, ]
op <- par(mar = c(13, 4, 1, 2))
barplot(
  dfs$Freq,
  names.arg = dfs$Var1,
  las = 2,
  cex.names = 0.7,
  col = "skyblue",
  ylim = range(pretty(c(0, 2000))),
  main = "Crimes with a frequency <2000"
)
```

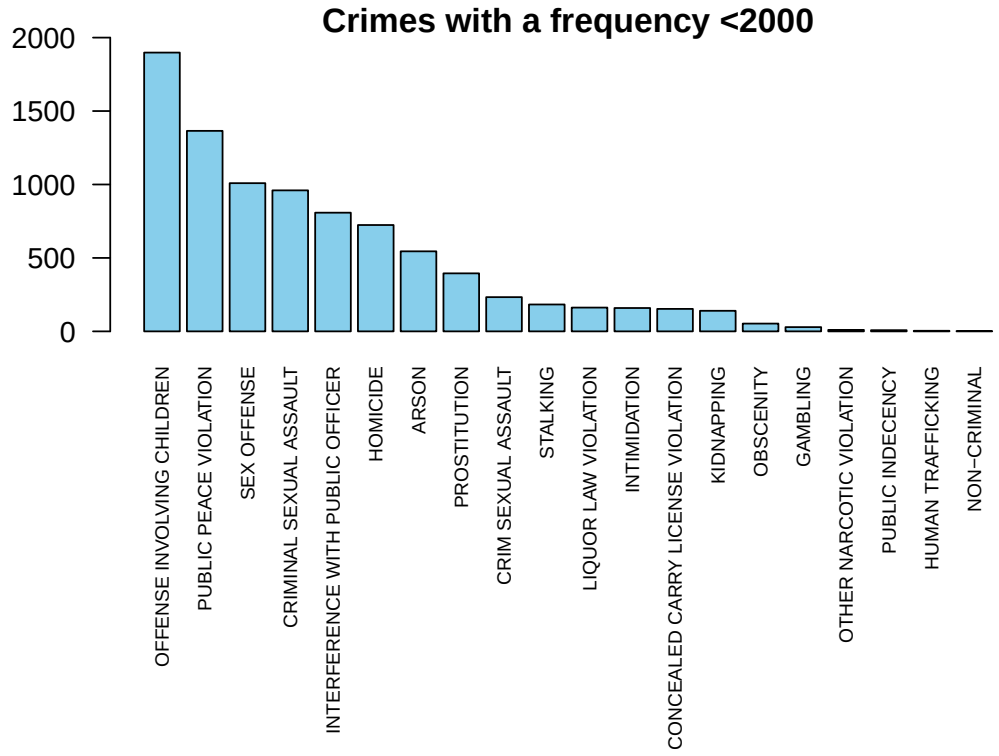

Figure 2: Kategorije kriminala s frekvencijom <2000

Učitavanje podataka o vremenu

```
#dohvaća listu datuma
dates <-
  strptime(c(chicago_year$DATE..OF.OCCURRENCE),
           format = "%m/%d/%Y %H:%M:%S %p",
           tz = "America/Chicago")

#dohvaća listu sati
```

```
hours <-
  as.numeric(format(
    strptime(
      chicago_year$DATE..OF.OCCURRENCE,
      format = "%m/%d/%Y %I:%M:%S %p",
      tz = "America/Chicago"
    ),
    format = "%H"
  ))
table(hours)
```

```
## hours
##     0     1     2     3     4     5     6     7     8     9    10    11    12
## 11314  7040  5989  5195  4152  3531  3579  4790  6861  9311  9738  9835 13028
##    13    14    15    16    17    18    19    20    21    22    23
## 10625 11086 11783 11801 11802 12101 11373 11459 10429 10350  8860
```

Podaci o učestalosti kriminala ovisno o dobu dana

Promotrimo najprije ovisnost frekventnosti kriminalnih aktivnosti o trenutku u danu (Figure 3). Inicijalna pretpostavka većine je vjerojatno da su kriminalne aktivnosti najfrekventnije u noćnim satima, ali histogram nam pokazuje drukčije.

```
hist(
  hours + 1,
  breaks = 0:24,
  main = "Frequency of criminal activities in 24 hours",
  xlab = "Hour",
  ylab = "Frequency",
  col = "light blue",
  xaxt = "n"
)
axis(side = 1,
     at = seq(0, 24, 2),
     tick = T)
rug(seq(1, 23, 2), ticksize = -0.03, side = 1)
```

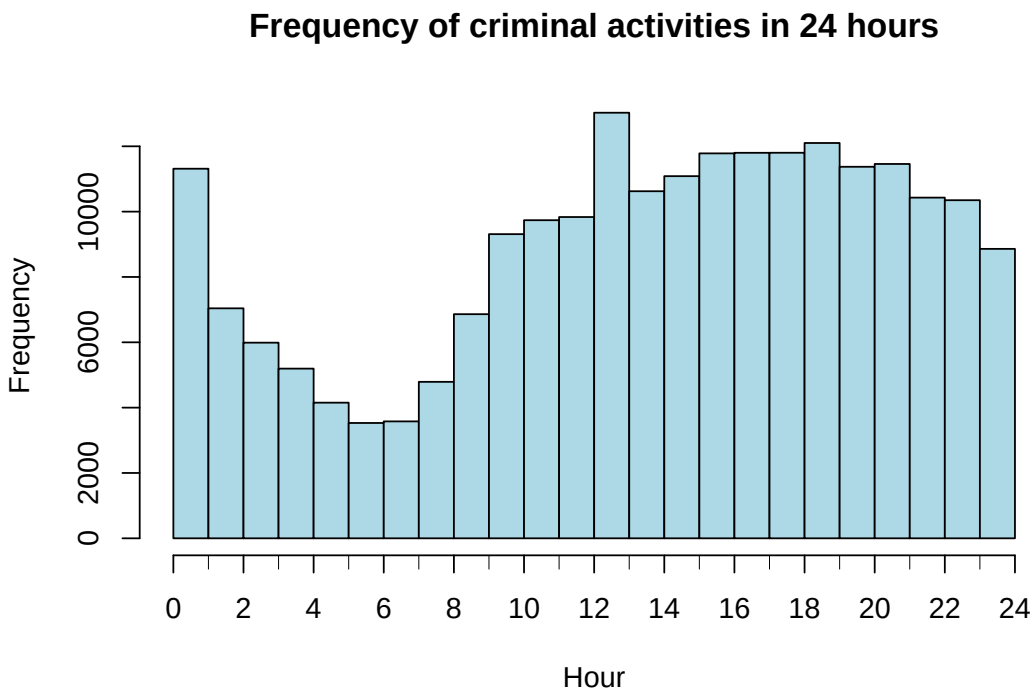

Figure 3: Frekventnost kriminalnih radnji u periodu od 24 sata

Izdvojimo sad broj kriminalnih djela u 3 različita doba dana: noć koja traje od 22h do 6h, jutro koje traje od 6h do 14h i popodne koje smo uzeli da traje od 14h do 22h.

```
day_parts<-ifelse(hours>=22|hours<6,rr2<-3,
                      ifelse(hours>=6&hours<14,rr2<-1,
                             rr2<-2))
mytable<-table(day_parts)
print(mytable)
```

```
## day_parts
##     1     2     3
## 67767 91834 56431
```

Nakon što smo izdvojili podatke, možemo nacrtati histogram po dijelovima dana.

```
hist(
  day_parts,
  breaks = 0:3,
  main = "Frequency of criminal activities for different parts of the day",
  xlab = "Day part",
  ylab = "Frequency",
  col = "light blue",
  xaxt = "n"
)
axis(side = 1,
     at = seq(0.5, 3, 1),
     labels = c("morning", "afternoon", "night")
     )
```

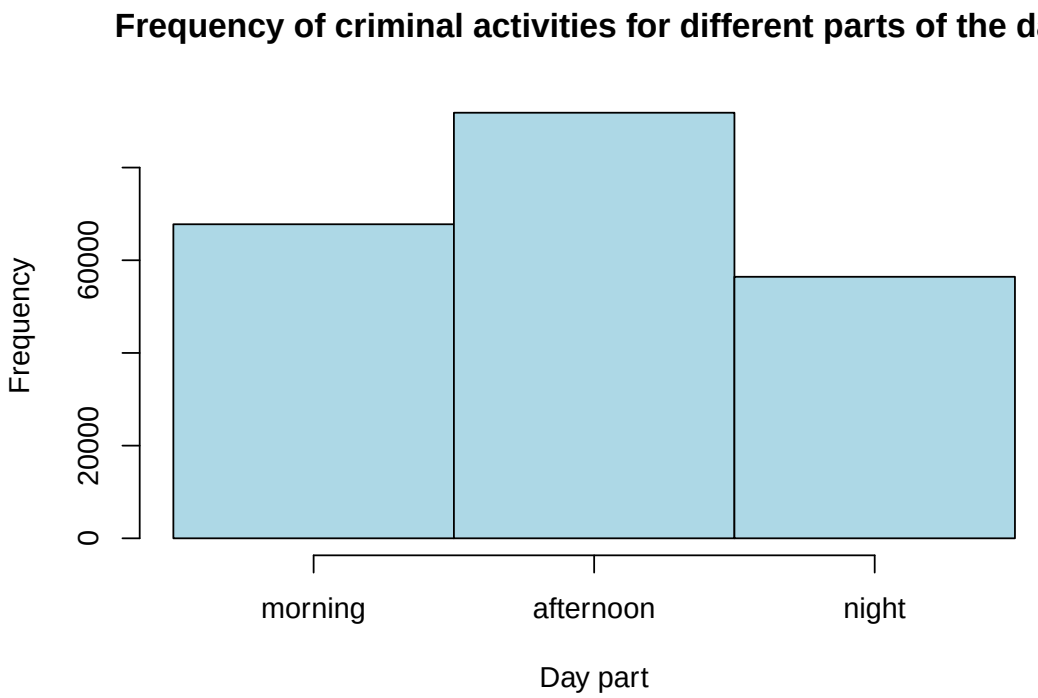

Figure 4: Frekventnost kriminalnih radnji po dijelovima dana

Iz histograma je očito da se najviše kriminalnih djela događa između 14 i 22h. Sada HI kvadrat testom možemo provjeriti uniformnost ove distribucije, odnosno vidjeti je li ta razlika statistički značajna ili nije. Pretpostavit ćemo da je distribucija uniformna. Testirajmo hipotezu H0 jesu li kriminalna djela jednako distribuirana po svim djelovima dana, uz hipotezu H1 da nisu jednako distribuirana uz razinu značajnosti alpha = 0.05.

```
res<-chisq.test(as.integer(table(day_parts)), p = c(1/3, 1/3, 1/3))
res
```

```
##
##  Chi-squared test for given probabilities
##
## data:  as.integer(table(day_parts))
## X-squared = 9077.8, df = 2, p-value < 2.2e-16
```

S obzirom da je p-value < 2.2e-16 izrazito manja od alpha = 0.05, odbacujemo H0 u korist H1 te bismo mogli zaključiti da ova distribucija nije uniformna, odnosno da broj kriminalnih djela nije isti u svakom dobu dana. Naravno, treba pripaziti i na veličinu podataka. S obzirom na velik broj podataka, može se dogoditi da se odbaci H0 u korist H1 kad to nije trebalo napraviti, ali s obzirom na izgled dijagrama, slažemo se s odbacivanjem H0 u korist H1.

Podaci o učestalosti kriminala ovisno o mjesecu u godini

Razmišljajući o prethodnom problemu, zaključujemo da bi bilo zanimljivo proučiti i ovisnost učestalosti kriminala o mjesecu u godini. U sljedećem dijelu, provjerit ćemo radi li se u ovom slučaju o uniformnoj razdiobi. Izdvojimo za početak broj kriminala vezan uz svaki mjesec.

```r
#dohvaća listu mjeseci
months <-
  as.numeric(format(
    strptime(
      chicago_year$DATE..OF.OCCURRENCE,
      format = "%m/%d/%Y %I:%M:%S %p",
      tz = "America/Chicago"
    ),
    format = "%m"
  ))
table(months)
```

```
## months
##     1     2     3     4     5     6     7     8     9    10    11    12
## 19682 18034 16522 12741 17381 17418 19289 19535 17403 17430 19872 20725
```

Podaci izgledaju kao da dolaze iz uniformne razdiobe, prikažimo ih histogramom bolje vizualizacije radi.

```r
hist(
  months,
  breaks = 0:12,
  main = "Frequency of criminal activities every month in a year",
  xlab = "Month",
  ylab = "Frequency",
  col = "light blue",
  xaxt = "n"
)
axis(side = 1,
     at = seq(0.5, 12, 1),
     labels = c(1:12)
     )
```

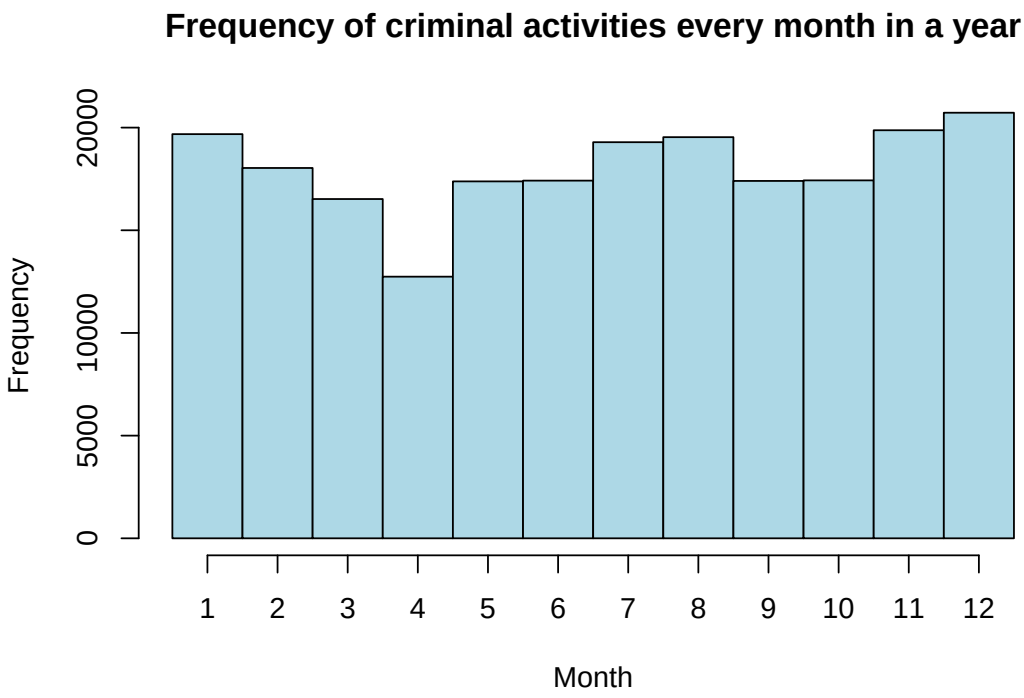

Figure 5: Frekventnost kriminalnih radnji svaki mjesec u godini

Iz histograma distribucija izgleda donekle uniformno. Vidimo ponešto manje brojke u ožujku i travnju. Opet ćemo iskoristiti HI kvadrat test da provjerimo uniformnost ove distribucije. Testirajmo hipotezu H0 jesu li kriminalna djela jednako distribuirana po svim mjesecima, uz hipotezu H1 da nisu jednako distribuirana uz razinu značajnosti alpha = 0.05.

```
res<-chisq.test(as.integer(table(months)), p = c(rep(1/12, each=12)))
res
```

```
##
##  Chi-squared test for given probabilities
##
## data:  as.integer(table(months))
## X-squared = 2723.1, df = 11, p-value < 2.2e-16
```

U ovom slučaju dobivamo nešto više od 3 puta manju vrijednost X-squared uz 11 stupnjeva slobode. p-value je poprilično mala te bismo mogli odbaciti H0 u korist H1. S obzirom da 3. i 4. mjesec odstupaju zbog pandemije koronavirusa, pokušajmo ih izbaciti iz podataka, te ponoviti test.

```
months_filtered<-ifelse(months==1,rr2<-1,
                        ifelse(months==2,rr2<-2,
                               ifelse(months==5,rr2<-5,
                                      ifelse(months==6,rr2<-6,
                                             ifelse(months==7,rr2<-7,
                                                    ifelse(months==8,rr2<-8,
                                                           ifelse(months==9,rr2<-9,
                                                                  ifelse(months==10,rr2<-10,
                                                                         ifelse(months==11,rr2<-11,
                                                                                ifelse(months==12,rr2<-12,
```

```
                            rr2<-NA))))))))))
table_months <- table(months_filtered)
table_months
```

```
## months_filtered
##     1     2     5     6     7     8     9    10    11    12
## 19682 18034 17381 17418 19289 19535 17403 17430 19872 20725
```

```
res<-chisq.test(as.integer(table_months), p = c(rep(1/10, each=10)))
res
```

```
##
##  Chi-squared test for given probabilities
##
## data:  as.integer(table_months)
## X-squared = 781.68, df = 9, p-value < 2.2e-16
```

X-squared je sad pao za još 4 puta u odnosu na prošli test. Iako je p vrijednost i dalje mala, odučili smo ne odbaciti H0 zbog velikog skupa podataka. Smatramo da je distribucija po mjesecima uniformna te da je u 2020. godini bilo drugačije zbog lockdowna u ožujku i travnju, ali ostali mjeseci imaju približno uniformnu distribuciju.

Učestalost krađa i kriminala vezanih uz narkotike

Dijagram (Figure 1) prikazuje učestalost različitih kategorija zločina unazad godinu dana. Najveću učestalost prema grafu ima kategorija krađe "THEFT", dok je kategorija "NARCOTICS" tek na 9. mjestu. Uz kategoriju krađe osim "THEFT" dodali smo "BURGLARY" i "ROBBERY" jer im je cilj krađa. Možemo reći da su specijalizacija krađe. Testom o jednoj proporciji provjeravamo je li razlika učestalosti krađa i kriminala vezanih uz narkotike statistički signifikantna, odnosno je li učestalost krađa doista veća od učestalosti kriminala vezanih za narkotike. Pretpostavit ćemo da se radi o normalnoj distribuciji.

Testiramo hipotezu H0: udio krađa je 0.5, nasuprot alternativi H1: udio krađa je veći od 0.5. Za razinu značajnosti alpha uzeli smo 0.05.

```
theft_crimes <- c("THEFT", "ROBBERY", "BURGLARY")
theft <- chicago_year[chicago_year$PRIMARY.DESCRIPTION %in% theft_crimes,]
narcotic_crimes <- c("NARCOTICS", "OTHER NARCOTIC VIOLATION")
all_narocitcs <- chicago_year[chicago_year$PRIMARY.DESCRIPTION %in% narcotic_crimes,]
x <- nrow(theft) # broj krađa
n <- nrow(theft) + nrow(all_narocitcs) # veličina uzorka

res <- prop.test(x, n, p = 0.5, correct = TRUE,
              alternative = "greater")

res
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  x out of n, null probability 0.5
## X-squared = 41834, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is greater than 0.5
## 95 percent confidence interval:
##  0.8816474 1.0000000
## sample estimates:
##         p
## 0.8836476
```

Dobili smo p-vrijednost 2.2e-16, koja je puno manja od alpha = 0.05, iz tog razloga odbacujemo H0 u korist H1 te zaključujemo da su krađe doista učestalije. U prilog ovom zaključku ide i procjena vjerojatnosti krađe koja je 0.8836476 i nalazi se u 95%-tnom intervalu povjerenja $0.8816474 < 0.8836476 < 1.0000000$.

Podaci o ukradenoj vrijednosti

Nakon što smo vidjeli da je krađa najčešći oblik kriminala, zanimala nas je vrijednost ukradenih stvari. Dijagram (Figure 6) pokazuje da je učestalost većih krađa (preko $500) manja. Ponovo ćemo testom o jednoj proporciji provjeriti je li to stvarno istina.

```
secondary_theft_types <- c("$500 AND UNDER", "OVER $500")
theft_secondary <- chicago_year[chicago_year$SECONDARY.DESCRIPTION %in% secondary_theft_types,]

table <- table(theft_secondary$SECONDARY.DESCRIPTION)
data <- as.data.frame.table(table)

barplot(
data$Freq,
names.arg = data$Var1,
cex.names = 0.6,
col = "skyblue",
ylim = range(pretty(c(0, data$Freq))),
main = "Theft value"
)
```

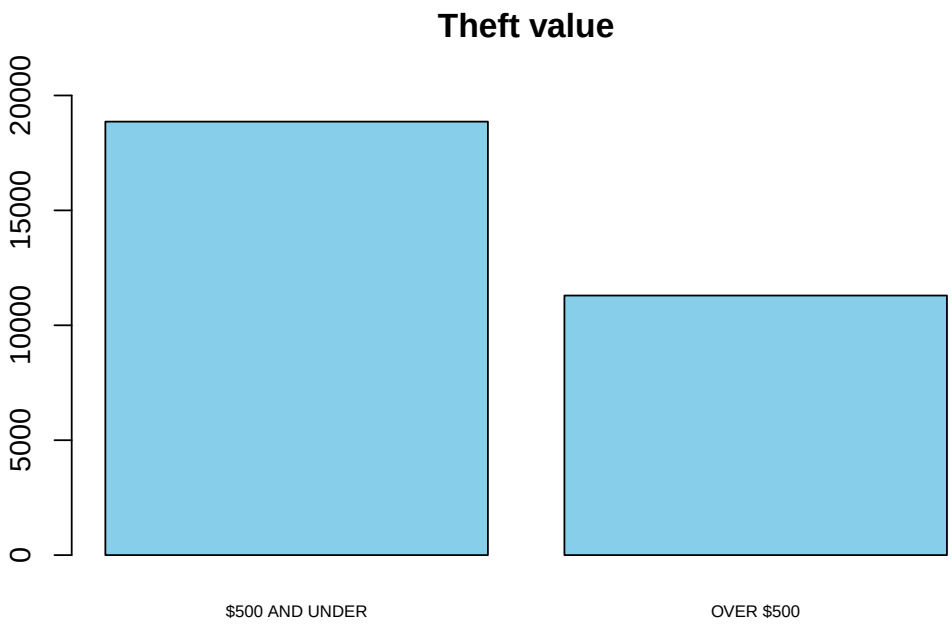

Figure 6: Podaci o vrijednosti krade

Testiramo hipotezu H0: udio krađa iznad $500 je 0.5, nasuprot alternativi H1: udio krađa iznad $500 je manji od 0.5. Za razinu značajnosti alpha uzeli smo 0.05.

```r
over500 <- c("OVER $500")

theft_more <- chicago_year[chicago_year$SECONDARY.DESCRIPTION %in% over500,]

x <- nrow(theft_more) # broj krađa većih od $500
n <- nrow(theft_secondary) # veličina uzorka

res <- prop.test(x, n, p = 0.5, correct = TRUE,
              alternative = "less")

res
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  x out of n, null probability 0.5
## X-squared = 1898.2, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is less than 0.5
## 95 percent confidence interval:
##  0.000000 0.379152
## sample estimates:
##         p
## 0.3745399
```

Dobili smo p-vrijednost 2.2e-16, koja je znatno manja od alpha = 0.05, iz tog razloga odbacujemo H0 u korist H1 te zaključujemo da su krađe iznad $500 manje učestale od krađa ispod $500. Vjerojatnost krađe iznad $500 je 0.3745399.

Pad kriminala vezan uz koronavirus

Sljedeći graf prikazuje fascinantu činjenicu o smanjenju broja kriminalnih djela nakon početka pandemije koronavirusa i lockdowna. Prikaz je informativan te nećemo ulaziti u detaljniju analizu.

```r
theft <- chicago_year[chicago_year$PRIMARY.DESCRIPTION == "THEFT",]
theft_dates <-
  table(as.Date(
    strptime(theft$DATE..OF.OCCURRENCE, format = "%m/%d/%Y %I:%M:%S %p", tz =
               "America/Chicago"),
  ))

narcotics <-
  chicago_year[chicago_year$PRIMARY.DESCRIPTION == "NARCOTICS",]
narcotics_dates <-
  table(as.Date(
    strptime(narcotics$DATE..OF.OCCURRENCE, format = "%m/%d/%Y %I:%M:%S %p", tz =
               "America/Chicago"),
  ))

all_crimes <-
  table(as.Date(
    strptime(chicago_year$DATE..OF.OCCURRENCE, format = "%m/%d/%Y %I:%M:%S %p", tz =
               "America/Chicago"),
  ))

plot(
```

```
  theft_dates,
  type = "l",
  main = "Criminal acts decrease during lockdown",
  xlab = "Date",
  ylab = "Number of crimes",
  ylim=c(0,750)
)
lines(all_crimes, col = "green")
lines(theft_dates, col = "red")
lines(narcotics_dates, col = "blue")
legend("topright", c("All crimes", "Theft", "Narcotics"), fill = c(rgb(0, 1, 0, 0.5), rgb(1, 0, 0, 0.5)
```

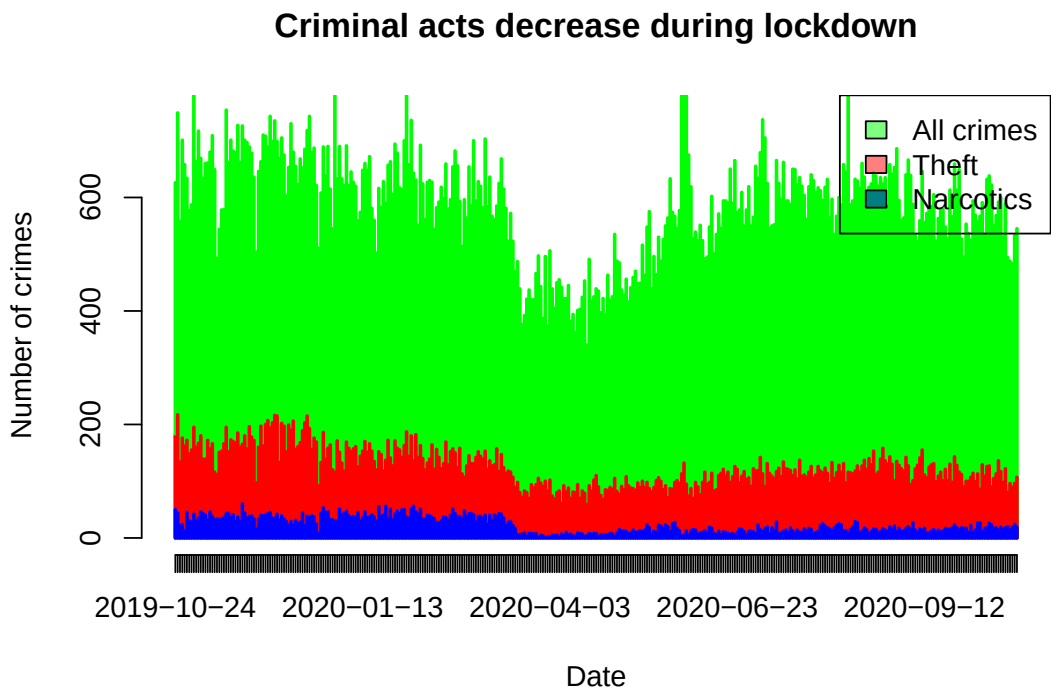

Figure 7: Pad kriminala nakon pandemije koronavirusa - informativno

Zavisnosti različitih socio-ekonomskih faktora

Prije nego što krenemo na veze između socio-ekonomskih faktora i kategorije kriminala, prvo nas zanima kakve su zavisnosti između različitih socio-ekonomskih faktora. U tu svrhu, napravili smo korelacijsku matricu.

```
matrixData <- chicago_pc[, c(-1, -2, -3, -4)]
matrix <- cor(matrixData)
matrix
```

```
##                       Below.Poverty.Level Crowded.Housing Dependency
## Below.Poverty.Level             1.0000000       0.3232420  0.4013540
## Crowded.Housing                 0.3232420       1.0000000  0.2444501
## Dependency                      0.4013540       0.2444501  1.0000000
## No.High.School.Diploma          0.4223819       0.9052740  0.4243563
## Per.Capita.Income              -0.5265178      -0.5452040 -0.7565786
```

```
## Unemployment                    0.7638170       0.1443044  0.6049994
##                       No.High.School.Diploma Per.Capita.Income Unemployment
## Below.Poverty.Level                0.4223819        -0.5265178    0.7638170
## Crowded.Housing                    0.9052740        -0.5452040    0.1443044
## Dependency                         0.4243563        -0.7565786    0.6049994
## No.High.School.Diploma             1.0000000        -0.7073543    0.3229021
## Per.Capita.Income                 -0.7073543         1.0000000   -0.6105529
## Unemployment                       0.3229021        -0.6105529    1.0000000
```

Uzmimo na primjer varijablu Below.Poverty.Level i pogledajmo kako se ona ponaša. Najveću pozitivnu korelaciju smo dobili za par varijabli Below.Poverty.Level i Unemployment (0.76). Dosta dobru negativnu korelaciju dobili smo za par varijabli Below.Poverty.Level i Per.Capita.Income (-0.53). Možemo zaključiti da je nezaposlenost veliki razlog zašto netko živi ispod granice siromaštva. Iz negativne korelacije između Below.Poverty.Level i Per.Capita.Income mogli bismo pretpostaviti da su kvartovi eventualno grupirani po socijalnom statusu te bogatiji kvartovi imaju manje siromašnih ljudi.

Za varijable Unemployment i Crowded.Housing dobivamo slabu pozitivinu korelaciju (0.14). Za te smo varijable očekivali pozitivnu korelaciju. Naime, ako netko nema posao onda vjerojatno nema ni dovoljno za stanarinu koja bi odgovarala veličini prostora koji je potreban za zdrav život. Međutim broj koji smo dobili je puno manji od očekivanog.

Kako smo dobili broj koji je blizu 0, ovaj primjer je dobar kandidant da napravimo korelacijski test.

```
cor.test(chicago_pc$Unemployment, chicago_pc$Crowded.Housing)
```

```
##
##  Pearson's product-moment correlation
##
## data:  chicago_pc$Unemployment and chicago_pc$Crowded.Housing
## t = 1.2629, df = 75, p-value = 0.2105
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.08233557  0.35675262
## sample estimates:
##       cor
## 0.1443044
```

P-vrijednost testa je 0.2105. To znači da u slučaju da varijable Unemployment i Crowded.Housing nisu u nikakvoj korelaciji (tj. korelacija je 0), vjerojatnost da dobijemo pozitivnu korelaciju od 0.144 i više je 21%. Možemo posumnjati u pozitivnu korelaciju ove dvije varijable.

Od drugih očekivanih vrijednosti imamo Per.Capita.Income i No.High.School.Diploma (-0.71). Ima smisla da su kvartovi siromašniji što je manji stupanj edukacije ljudi koji žive tamo. Također ako manji Dependecy znači zdraviju i bogatiju ekonomiju, onda također očekujemo da će varijable Dependecy i Per.Capita.Income biti jako negativno korelirane što i jesu (-0.76).

Najveća vrijednost koja nije korelacija varijable sa samom sobom vidimo u slučaju No.High.School.Diploma i Crowded.Housing (0.91). Ovo je dosta zanimljiv i na prvi pogled jako neočekivani rezultat koji smo dobili. Sljedće što smo napravili je test.

```
cor.test(chicago_pc$No.High.School.Diploma, chicago_pc$Crowded.Housing)
```

```
##
##  Pearson's product-moment correlation
##
## data:  chicago_pc$No.High.School.Diploma and chicago_pc$Crowded.Housing
## t = 18.454, df = 75, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
```

```
## 95 percent confidence interval:
##  0.8545696 0.9388829
## sample estimates:
##      cor
## 0.905274
```

Vidimo da je 95%-tni interval između 0.85 i 0.93. Zaključujemo da su ove dvije varijable zaista jako zavisne. Iz tog razloga, možemo zaključiti da manjak obrazovanja najviše pridonosi prenatrpanim domovima. Međutim još uvijek ne znamo objasniti zašto je to tako.

Ovo je dobar primjer da probamo napraviti linearnu regresiju i pogledamo kako to sve izgleda (Figure 10).

```
no.high.school.diploma <-
  as.matrix(chicago_pc[8]) #No.High.School.Diploma
crowded.housing <- as.matrix(chicago_pc[6]) #Crowded.Housing

reg1 <- lm(crowded.housing  ~ no.high.school.diploma, chicago_pc)

summary(reg1)
```

```
##
## Call:
## lm(formula = crowded.housing ~ no.high.school.diploma, data = chicago_pc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6566 -0.8833 -0.2186  0.9591  3.9239
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -0.87434    0.36072  -2.424   0.0178 *
## no.high.school.diploma  0.26798    0.01452  18.454   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.564 on 75 degrees of freedom
## Multiple R-squared:  0.8195, Adjusted R-squared:  0.8171
## F-statistic: 340.6 on 1 and 75 DF,  p-value: < 2.2e-16
```

```
plot(
  no.high.school.diploma,
  crowded.housing ,
  pch = 16,
  cex = 1.3,
  col = "lightgreen",
  xlab = "Percentage of people without high school diploma",
  ylab = "Percentage of people living in crowded houses",
)
abline(lm(crowded.housing  ~ no.high.school.diploma), col = "pink", lwd = 2)
```

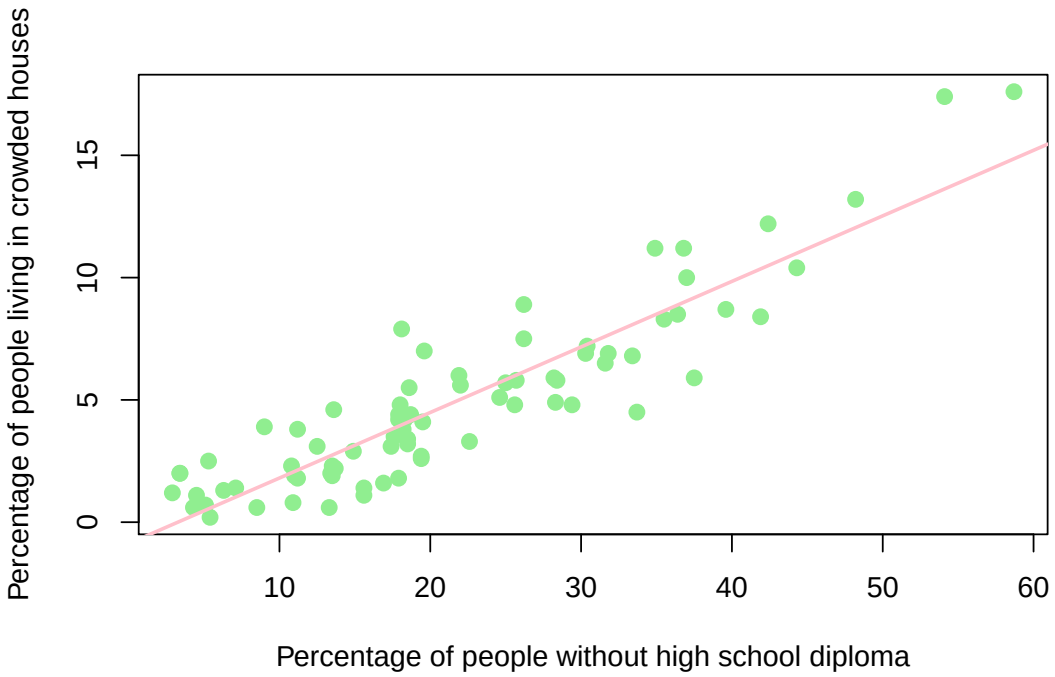

Figure 8: Prikaz linearne regresije (Crowded Housing)

Veze između socio-ekonomskih faktora i kategorije kriminala

Kategorija Firearm related

Kategorija kriminala koju najprije promatramo je Firearm.related. Prvo ćemo napraviti korelacijsku matricu da dobijemo dojam kako se brojke kreću i od kojih varijabli bismo mogli krenuti raditi regresijski model.

```
matrixData2 <- chicago_pc[, c(-1, -2, -3)]
matrix2 <- cor(matrixData2)
matrix2[,1]
```

```
##        Firearm.related    Below.Poverty.Level        Crowded.Housing
##             1.00000000             0.56575966             0.03445091
##             Dependency No.High.School.Diploma      Per.Capita.Income
##             0.59079639             0.13125365            -0.49685919
##           Unemployment
##             0.72257661
```

Firearm.related i Unemployment imaju najveći korelacijski koeficijent pa ćemo započeti graditi regresijski model s tom varijablom.

```
unemployment <- as.matrix(chicago_pc[10]) #Unemployment
firearm.related <- as.matrix(chicago_pc[4]) #Firearm.related
reg <- lm(firearm.related ~ unemployment, chicago_pc)

summary(reg)
```

```
##
## Call:
```

```
## lm(formula = firearm.related ~ unemployment, data = chicago_pc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.804  -5.035  -1.348   2.173  38.565
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -3.0395     2.4669  -1.232    0.222
## unemployment   1.4861     0.1642   9.052 1.18e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.06 on 75 degrees of freedom
## Multiple R-squared:  0.5221, Adjusted R-squared:  0.5157
## F-statistic: 81.94 on 1 and 75 DF,  p-value: 1.185e-13
```

```r
plot(
  unemployment,
  firearm.related,
  pch = 16,
  cex = 1.3,
  col = "lightgreen",
  xlab = "Percentage of unemployed people",
  ylab = "Firearm related crimes per 100 000",
)
abline(lm(firearm.related ~ unemployment), col = "pink", lwd = 2)
```

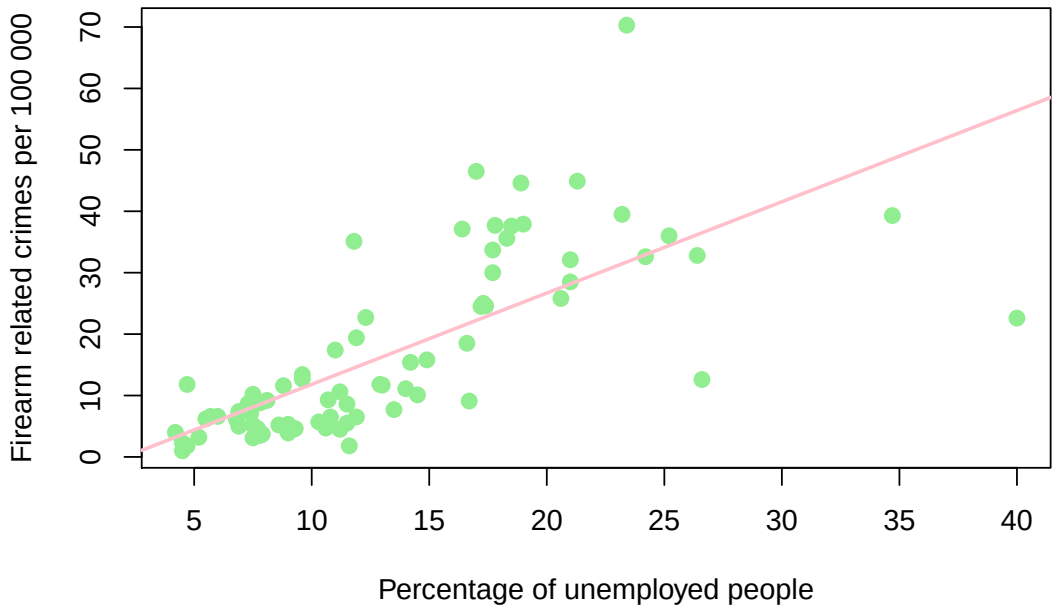

Figure 9: Prikaz linearne regresije (Firearm)

Slika i regresija koju smo dobili nije loša, međutim odmah primjećujemo da imamo stršećih vrijednosti. Također vidimo da nam koeficijent b0 nije statistički značajan. Primjenjujući transformaciju nad varijablom Unemployment, pokušat ćemo dobiti nešto bolju regresiju.

```
unemployment <- as.matrix(chicago_pc[10]) #Unemployment
firearm.related <- as.matrix(chicago_pc[4]) #Firearm.related
reg <- lm(firearm.related ~ log(unemployment), chicago_pc)

summary(reg)
```

```
##
## Call:
## lm(formula = firearm.related ~ log(unemployment), data = chicago_pc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.370  -6.010  -0.026   4.310  39.019
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -34.868      5.462  -6.384 1.30e-08 ***
## log(unemployment)   20.982      2.175   9.648 8.82e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.725 on 75 degrees of freedom
## Multiple R-squared:  0.5538, Adjusted R-squared:  0.5478
## F-statistic: 93.08 on 1 and 75 DF,  p-value: 8.816e-15
```

Ako primijenimo logaritamsku transformaciju, koeficijent b0 postaje jako značajan, a i koeficijent determinacije se povećao, te F-test daje puno bolje rezultate.

```
plot(
  log(unemployment),
  firearm.related,
  pch = 16,
  cex = 1.3,
  col = "lightgreen",
  xlab = "Percentage of unemployed people",
  ylab = "Firearm related crimes per 100 000",
)
abline(lm(firearm.related ~ log(unemployment)), col = "pink", lwd = 2)
```

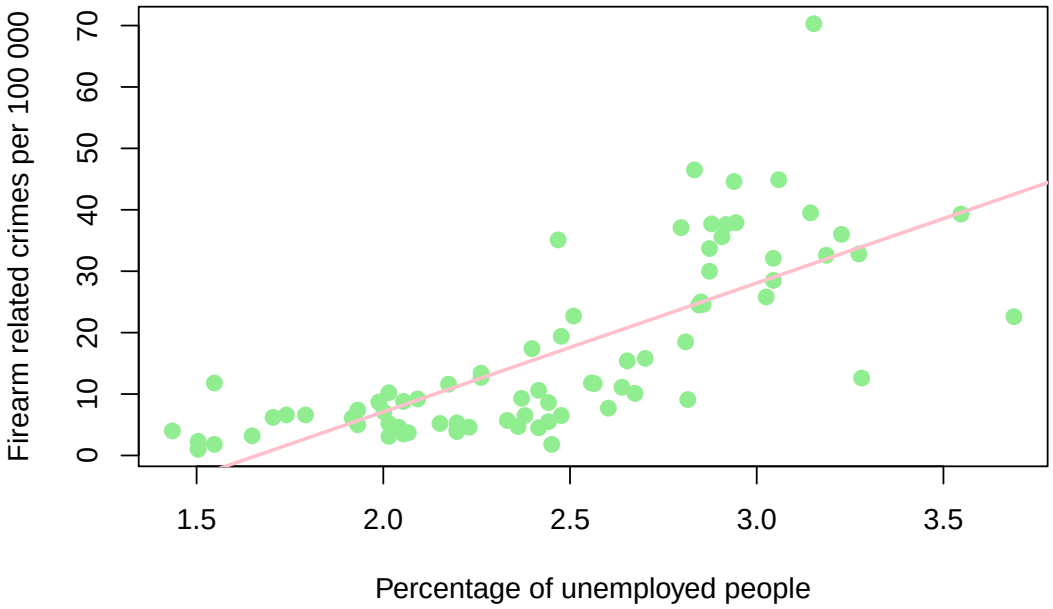

Figure 10: Prikaz linearne regresije (Firearm)

Na slici se još uvijek primjećju stršeće vrijednosti.

Sljedeće dvije varijable koje su u velikoj korelaciji su Below.Poverty.Level i Dependency. Obje varijable su visoko korelirane s varijablom Unemployment što smo promatrali prije.

```
unemployment <- as.matrix(chicago_pc[10]) #Unemployment
dependecy <- as.matrix(chicago_pc[7]) #Dependency
firearm.related <- as.matrix(chicago_pc[4]) #Firearm.related
reg <-
  lm(firearm.related ~ log(unemployment) + dependecy, chicago_pc)

summary(reg)
```

```
##
## Call:
## lm(formula = firearm.related ~ log(unemployment) + dependecy,
##     data = chicago_pc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.486  -6.364  -0.949   4.533  39.613
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -37.5400     5.9574  -6.301 1.91e-08 ***
## log(unemployment)  18.5212     3.0971   5.980 7.30e-08 ***
## dependecy           0.2434     0.2185   1.114    0.269
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.709 on 74 degrees of freedom
## Multiple R-squared:  0.5611, Adjusted R-squared:  0.5493
## F-statistic: 47.31 on 2 and 74 DF,  p-value: 5.84e-14
```

Ako dodamo varijablu Dependency u model, on se nije puno poboljšao, štoviše varijabla je statistički neznačajna.

```
unemployment <- as.matrix(chicago_pc[10]) #Unemployment
dependecy <- as.matrix(chicago_pc[7]) #Dependency
firearm.related <- as.matrix(chicago_pc[4]) #Firearm.related
below.poverty.level <- as.matrix(chicago_pc[5]) #Below.Poverty.Level
reg <-
  lm(firearm.related ~ log(unemployment) + dependecy + below.poverty.level,
     chicago_pc)

summary(reg)
```

```
##
## Call:
## lm(formula = firearm.related ~ log(unemployment) + dependecy +
##     below.poverty.level, data = chicago_pc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.777  -6.406  -0.251   4.247  40.642
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -35.9337     6.4219  -5.596 3.63e-07 ***
## log(unemployment)    16.4529     4.3323   3.798   0.0003 ***
## dependecy             0.2831     0.2268   1.248   0.2159
## below.poverty.level   0.1014     0.1479   0.685   0.4953
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.744 on 73 degrees of freedom
## Multiple R-squared:  0.5639, Adjusted R-squared:  0.546
## F-statistic: 31.47 on 3 and 73 DF,  p-value: 3.644e-13
```

Te ako dodamo i varijablu Below.Poverty.Level dobivamo sličnu stvar što je i očekivano. Sve tri varijable su dosta zavisne te ne dobivamo puno novih informacija.

Iz tog razloga ćemo izbaciti te varijable iz modela.

Tražimo neku drugu varijablu koja nije jako kolerirana s našom prvom odabranom varijablom i tražimo onu koja daje trenutačno najbolji model. Dolazimo do rezultata da je to varijabla No.High.School.Diploma.

```
unemployment <- as.matrix(chicago_pc[10]) #Unemployment
no.high.school.diploma <-
  as.matrix(chicago_pc[8]) #No.High.School.Diploma
firearm.related <- as.matrix(chicago_pc[4]) #Firearm.related
reg <-
  lm(firearm.related ~ log(unemployment) + no.high.school.diploma,
     chicago_pc)
```

```r
summary(reg)
```

```
##
## Call:
## lm(formula = firearm.related ~ log(unemployment) + no.high.school.diploma,
##     data = chicago_pc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.789  -5.188  -0.955   3.530  36.236
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -35.94655    5.23301  -6.869 1.72e-09 ***
## log(unemployment)       23.81559    2.30341  10.339 5.16e-16 ***
## no.high.school.diploma  -0.27280    0.09563  -2.853  0.00562 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.293 on 74 degrees of freedom
## Multiple R-squared:  0.598,  Adjusted R-squared:  0.5871
## F-statistic: 55.04 on 2 and 74 DF,  p-value: 2.276e-15
```

Tražimo postoji li još kakva varijabla koja bi mogla pridonjeti modelu.

Ako vizualiziramo parove Per.Capita.Income, Firearm.related možemo uočiti pravilnost. U siromašnijim kvartovima, kriminal koji uključuje vatreno oružje je puno češće nego u bogatijim kvartovima. Štoviše, uočavamo eksponencijalni pad pa ćemo primijeniti transformaciju podataka nad tom varijablom.

```r
plot(
  chicago_pc$Per.Capita.Income,
  firearm.related,
  pch = 16,
  cex = 1.3,
  col = "lightgreen",
  xlab = "Per capita income",
  ylab = "Firearm related crimes per 100 000",
)
```

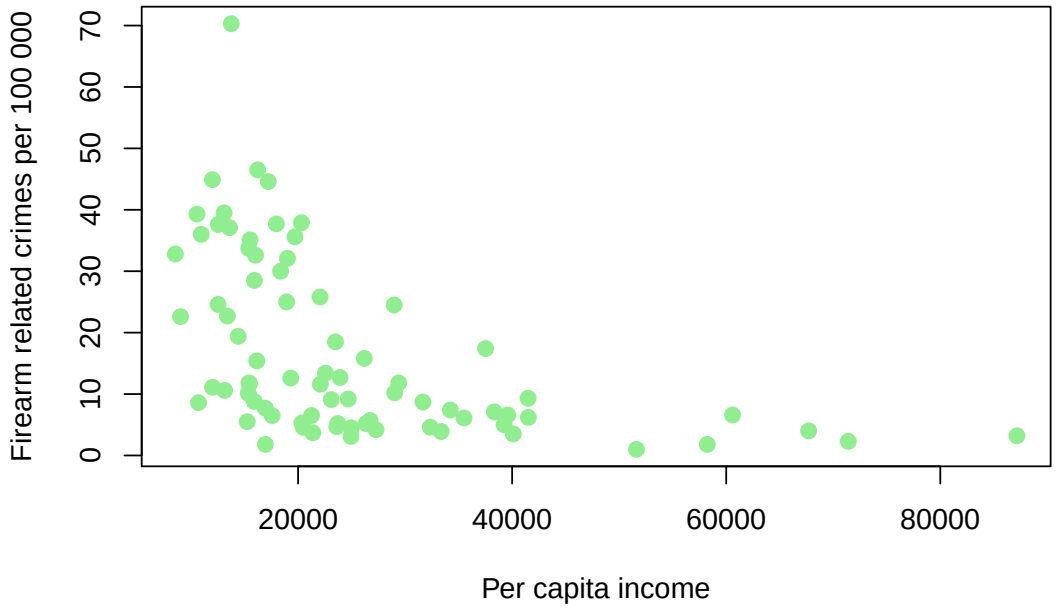

Figure 11: Odnos oružanih kriminala i GDP per capita

```
unemployment <- as.matrix(chicago_pc[10]) #Unemployment
no.high.school.diploma <-
  as.matrix(chicago_pc[8]) #No.High.School.Diploma
firearm.related <- as.matrix(chicago_pc[4]) #Firearm.related
per.capita.income <- as.matrix(chicago_pc[9]) #Per.Capita.Income
reg <-
  lm(
    log(firearm.related) ~ log(unemployment) + no.high.school.diploma + log(per.capita.income),
    chicago_pc
  )

summary(reg)
```

```
##
## Call:
## lm(formula = log(firearm.related) ~ log(unemployment) + no.high.school.diploma +
##     log(per.capita.income), data = chicago_pc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5864 -0.3415  0.1638  0.3963  0.8967
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           11.306360   4.134212   2.735  0.00783 **
## log(unemployment)      0.976573   0.230993   4.228 6.75e-05 ***
```

```
## no.high.school.diploma -0.032562   0.009433  -3.452  0.00093 ***
## log(per.capita.income) -1.058240   0.349350  -3.029  0.00339 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5351 on 73 degrees of freedom
## Multiple R-squared:  0.6858, Adjusted R-squared:  0.6729
## F-statistic: 53.11 on 3 and 73 DF,  p-value: < 2.2e-16
```

Dodavanjem ostalih dviju varijabli Crowded.Housing i Below.Poverty.Level nismo dobili bolji model pa s ovime završavamo gradnju modela.

Sada je vrijeme da vidimo zadovoljala li model sve pretpostavke.

Provjeravamo homogenost varijanci. Kako bi vrijedila ta pretpostavka, trebamo dobiti što bolju moguću horizontalnu liniju.

```
plot(reg, 3, col="lightgreen", pch=16)
```

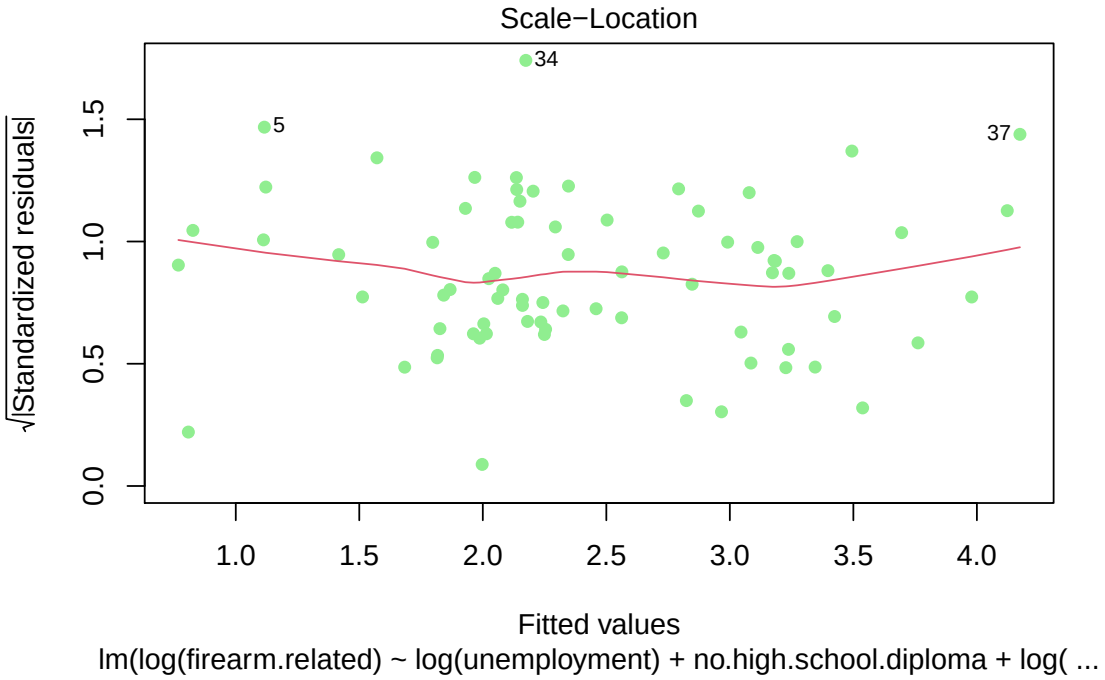

Figure 12: Prikaz standardiziranih reziduala

Horizontalna linija je dobar znak da homogenost varijance vrijedi.
Provjerimo sada to i testom.

```
bptest(reg)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  reg
## BP = 0.64481, df = 3, p-value = 0.8861
```

Test prolazi na razini značajnosti od 0.05.
Dalje nas zanima normalnost reziduala.

```
plot(reg, 2, col ="lightgreen", pch=16)
```

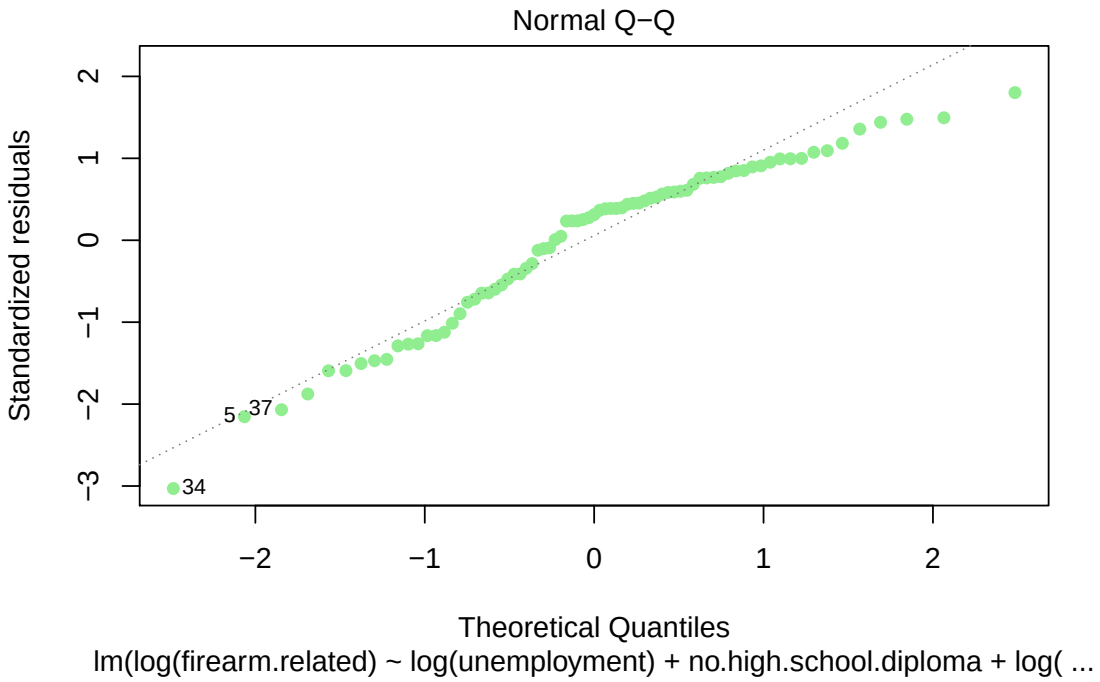

Figure 13: Q-Q Plot

Reziduali ne prate najbolje liniju koju trebaju pa možemo pretpostaviti da reziduali nisu normalno distribuirani. To znači da naš model ne bi davao pouzdane rezultate kada bismo ga koristili za predikciju vrijednosti. Našu pretpostavku možemo još provjeriti Kolmogorov-Smirnovljevim testom.

```
ks.test(rstandard(reg), 'pnorm')
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  rstandard(reg)
## D = 0.164, p-value = 0.0281
## alternative hypothesis: two-sided
```

Vidimo da model pada na KS testu te zaključujemo da reziduali nisu normalno distribuirani.

Zadnje što provjeravamo je nezavisnost pomoću Durbin-Watsonovog testa.

```
dwtest(reg)
```

```
##
##  Durbin-Watson test
##
## data:  reg
```

```
## DW = 1.7064, p-value = 0.07272
## alternative hypothesis: true autocorrelation is greater than 0
```

Vidimo da reziduali nisu autokorelirani što znači da su nezavisni.

Zaključak

Na kraju zaključujemo da na kriminal koji uključuje vatreno oružje utječe nezaposlenost, manjak više edukacije i GDP doprinos kvarta. Rezultati koji smo dobili nisu ništa neobičajni već konzistentni s onime što bi povezali s kriminalom. Loše socio-ekonomske prilike stvaraju atmosferu za kriminal.

Kategorija Assault Homicide

Promotrimo sada kategoriju Assault Homicide i njezinu korelaciju s drugim faktorima.

```
matrixHomicide <- chicago_pc[,c(-1, -2, -4)]
corMatrixHomicide <- cor(matrixHomicide)
corMatrixHomicide[,1]
```

```
##     Assault..Homicide.    Below.Poverty.Level        Crowded.Housing
##              1.0000000              0.6671429              0.0662508
##             Dependency No.High.School.Diploma      Per.Capita.Income
##              0.5748271              0.1822667             -0.5327565
##           Unemployment
##              0.8148348
```

Kao i kod kriminala vezanih za vatrena oružja vidimo jaku korelaciju s nezaposlenosti, što sugerira da regresijski model ponovo gradimo od te dvije varijable.

```
unemployment <- as.matrix(chicago_pc[10])
homicide <- as.matrix(chicago_pc[3])
regression <- lm(homicide ~ unemployment, chicago_pc)

summary(regression)
```

```
##
## Call:
## lm(formula = homicide ~ unemployment, data = chicago_pc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.684  -5.110  -0.898   2.974  32.856
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -7.4617     2.3689   -3.15  0.00235 **
## unemployment   1.9190     0.1576   12.17  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.664 on 75 degrees of freedom
## Multiple R-squared:  0.664,  Adjusted R-squared:  0.6595
## F-statistic: 148.2 on 1 and 75 DF,  p-value: < 2.2e-16
```

Kako bismo dobili bolje rezultate na naše podatke možemo primijeniti neku od transformacija.

```
regression_sqrt <- lm(homicide ~ sqrt(unemployment), chicago_pc)
regression_log <- lm(homicide ~ log(unemployment), chicago_pc)
```

```r
summary(regression_sqrt)
```

```
##
## Call:
## lm(formula = homicide ~ sqrt(unemployment), data = chicago_pc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.523  -5.776  -1.253   4.086  32.676
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -34.881      4.293  -8.126 6.88e-12 ***
## sqrt(unemployment)    14.989      1.177  12.736  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.374 on 75 degrees of freedom
## Multiple R-squared:  0.6838, Adjusted R-squared:  0.6796
## F-statistic: 162.2 on 1 and 75 DF,  p-value: < 2.2e-16
```

```r
summary(regression_log)
```

```
##
## Call:
## lm(formula = homicide ~ log(unemployment), data = chicago_pc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.712  -6.414  -1.009   6.886  34.049
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -46.413      5.464  -8.495 1.36e-12 ***
## log(unemployment)    26.220      2.175  12.053  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.728 on 75 degrees of freedom
## Multiple R-squared:  0.6595, Adjusted R-squared:  0.655
## F-statistic: 145.3 on 1 and 75 DF,  p-value: < 2.2e-16
```

Usporedbom rezultata uočavamo da je sqrt transformacija prikladnija u ovoj situaciji, također primijećujemo da smo dobili bolju signifikantnost varijable.
Vizualizirajmo dobivenu regresiju.

```r
plot(
  sqrt(unemployment),
  homicide,
  pch = 16,
  cex = 1.3,
  col = "cornflowerblue",
  xlab = "Unemployment percentage",
  ylab = "Homicide crimes per 100 000"
)
```

```
abline(regression_sqrt, col = "lightgreen", lwd = 2)
```

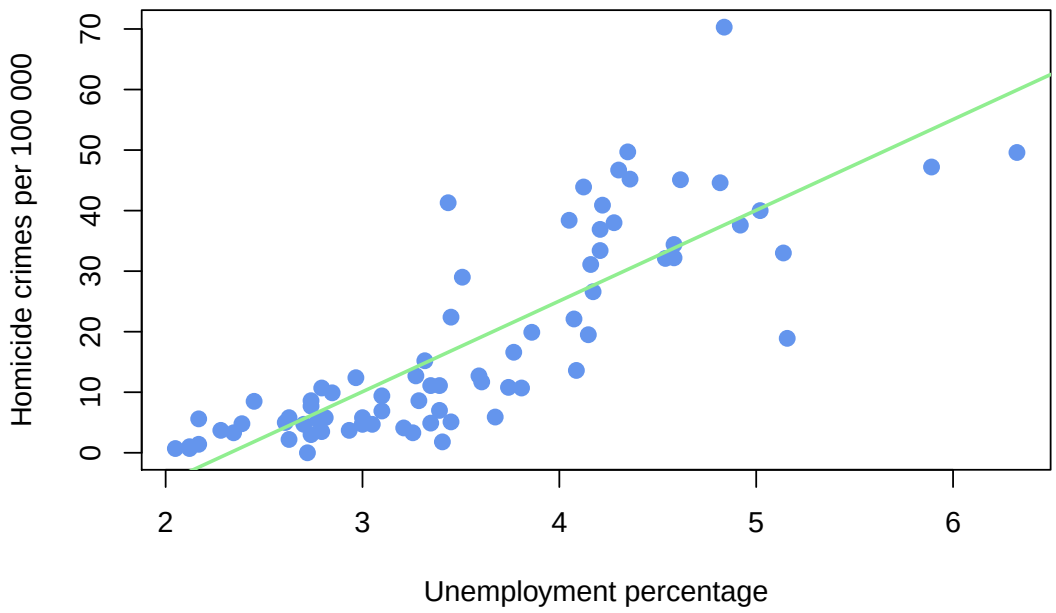

Figure 14: Prikaz linearne regresije (Homicide)

Sljedeća varijabla po razini korelacije Below poverty level.
Ako promotrimo ovisnot stope ubojstava sa stopom siromaštva uočavamo dosta jasnu poveznicu.

```
homicide_gdp <-
  data.frame(x = chicago_pc$Below.Poverty.Level,
             y = chicago_pc$Assault..Homicide.)
plot(homicide_gdp,
     main = "Homicide rate in relation to GDP per capita",
     xlab = "Below poverty level",
     ylab = "Homicide rate (n/100 000)",
     col = "cornflowerblue",
     pch = 16)
```

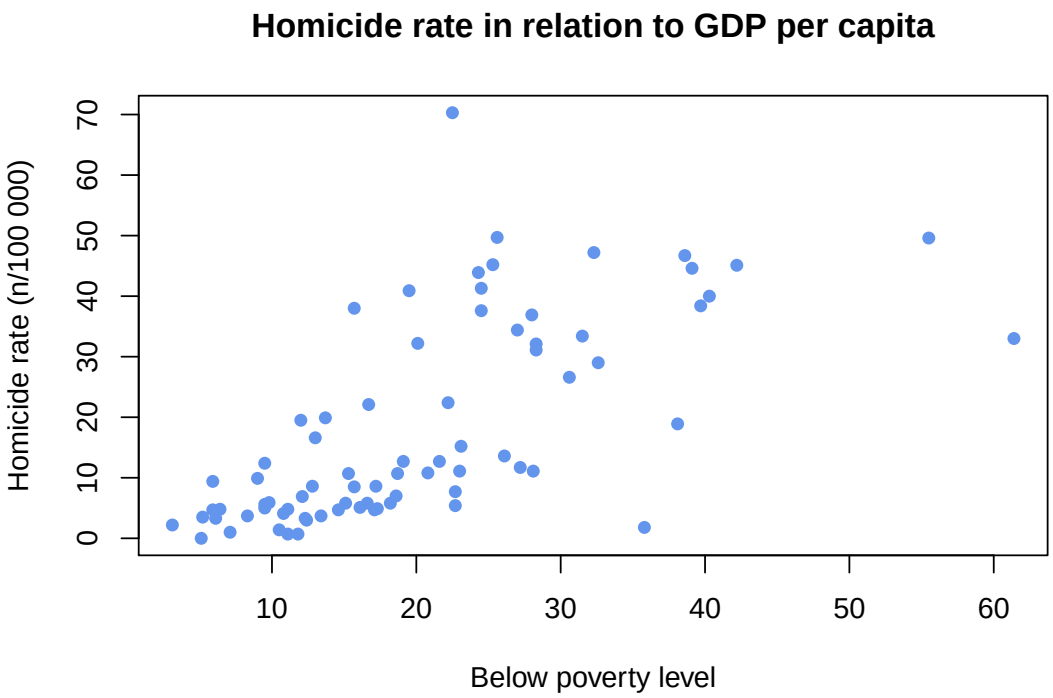

Figure 15: Odnos stope ubojstava i stope siromaštva

Pokušajmo sada dodati varijablu u regresijski model.

```
below_poverty <- as.matrix(chicago_pc[5])
regression_2 <-
  lm(homicide ~ sqrt(unemployment) + below_poverty, chicago_pc)
regression_2_sqrt <-
  lm(homicide ~ sqrt(unemployment) + sqrt(below_poverty), chicago_pc)
regression_2_log <-
  lm(homicide ~ sqrt(unemployment) + log(below_poverty), chicago_pc)

summary(regression_2)
```

```
##
## Call:
## lm(formula = homicide ~ sqrt(unemployment) + below_poverty, data = chicago_pc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.837  -5.075  -0.831   3.713  34.072
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -33.0532     4.7067  -7.023 8.92e-10 ***
## sqrt(unemployment)  13.6880     1.8058   7.580 8.04e-11 ***
## below_poverty        0.1363     0.1435   0.950    0.345
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 9.38 on 74 degrees of freedom
## Multiple R-squared:  0.6876, Adjusted R-squared:  0.6792
## F-statistic: 81.44 on 2 and 74 DF,  p-value: < 2.2e-16
```

```
summary(regression_2_sqrt)
```

```
##
## Call:
## lm(formula = homicide ~ sqrt(unemployment) + sqrt(below_poverty),
##     data = chicago_pc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.833  -5.149  -0.851   3.997  34.301
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -36.140      4.372  -8.266 4.06e-12 ***
## sqrt(unemployment)    13.190      1.780   7.409 1.69e-10 ***
## sqrt(below_poverty)    1.757      1.310   1.341    0.184
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.325 on 74 degrees of freedom
## Multiple R-squared:  0.6913, Adjusted R-squared:  0.683
## F-statistic: 82.86 on 2 and 74 DF,  p-value: < 2.2e-16
```

```
summary(regression_2_log)
```

```
##
## Call:
## lm(formula = homicide ~ sqrt(unemployment) + log(below_poverty),
##     data = chicago_pc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.635  -5.271  -0.938   4.090  34.014
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -39.456      5.231  -7.542 9.46e-11 ***
## sqrt(unemployment)    13.178      1.676   7.861 2.37e-11 ***
## log(below_poverty)     3.853      2.561   1.505    0.137
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.296 on 74 degrees of freedom
## Multiple R-squared:  0.6932, Adjusted R-squared:  0.6849
## F-statistic:  83.6 on 2 and 74 DF,  p-value: < 2.2e-16
```

Usporedbom 3 mogućnosti kombiniranja varijabli, primijećujemo da zadnja opcija (kombinacija sqrt i log transformacije) daje najbolje rezultate.
Pokušajmo sada dodati i treću varijablu po koreliranosti, Dependency.

```r
dependency <- as.matrix(chicago_pc[7])
regression_3 <-
  lm(homicide ~ sqrt(unemployment) + log(below_poverty) + dependency,
     chicago_pc)

summary(regression_3)
```

```
##
## Call:
## lm(formula = homicide ~ sqrt(unemployment) + log(below_poverty) +
##     dependency, data = chicago_pc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.849  -5.616  -0.810   4.263  34.775
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -43.9214     6.8940  -6.371 1.49e-08 ***
## sqrt(unemployment)  11.7114     2.2327   5.245 1.47e-06 ***
## log(below_poverty)   4.6583     2.6857   1.734   0.0871 .
## dependency           0.2052     0.2064   0.994   0.3233
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.297 on 73 degrees of freedom
## Multiple R-squared:  0.6973, Adjusted R-squared:  0.6849
## F-statistic: 56.05 on 3 and 73 DF,  p-value: < 2.2e-16
```

Kao i kod analize varijable Firearm related, za varijablu Homicide primijećujemo slično ponašanje - dodavanje varijable Dependency uzrokuje minimalna poboljšanja u modelu te ima vrlo malu signifikantnost.
Nakon ispitivanja nekoliko modela pokazalo se da varijable per_capita_income i crowded housing također imaju vrlo malu signifikanost te ne poboljšavaju razine signifikantnosti drugih varijabli. No dodamo li No high school diploma varijablu u model, ona ne da sama sebe ima veliku signifikantnost nego i značajno poboljšava signifikantnost varijabli dependency i below poverty level.
Sljedeći model pokazao je najbolje performanse:

```r
no_diploma <- as.matrix(chicago_pc[8])

regression_4 <-
  lm(
    homicide ~ sqrt(unemployment) + log(below_poverty) + dependency + no_diploma,
    chicago_pc
  )

summary(regression_4)
```

```
##
## Call:
## lm(formula = homicide ~ sqrt(unemployment) + log(below_poverty) +
##     dependency + no_diploma, data = chicago_pc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.356  -4.202  -0.743   4.218  33.452
```

```
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -53.04672    6.78760  -7.815 3.39e-11 ***
## sqrt(unemployment)    9.91023    2.10877   4.700 1.22e-05 ***
## log(below_poverty)    9.13968    2.74127   3.334 0.001355 **
## dependency            0.50491    0.20578   2.454 0.016562 *
## no_diploma           -0.37094    0.09827  -3.775 0.000327 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.553 on 72 degrees of freedom
## Multiple R-squared:  0.7473, Adjusted R-squared:  0.7333
## F-statistic: 53.23 on 4 and 72 DF,  p-value: < 2.2e-16
```

```
plot(regression_4, 3, col="cornflowerblue", pch=16)
```

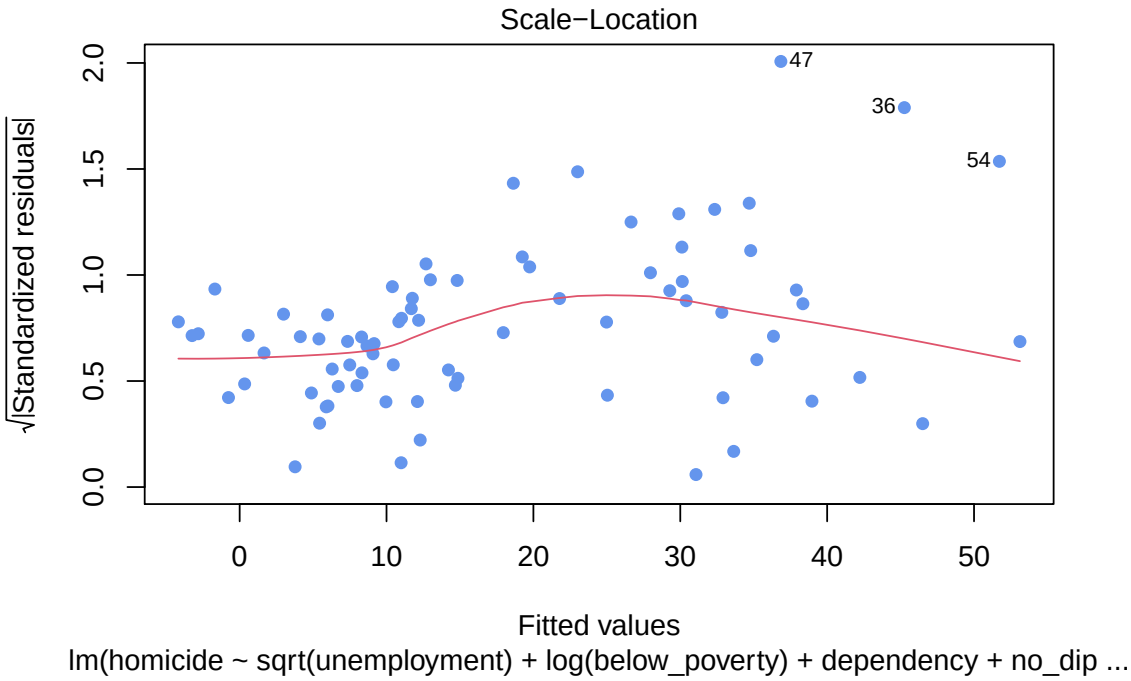

Figure 16: Prikaz standardiziranih reziduala

Breusch-Paganovim testom možemo provjeriti homoskedastičnost reziduala.

```
bptest(regression_4)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  regression_4
## BP = 12.048, df = 4, p-value = 0.017
```

Obzirom na činjenicu da linija u grafu iznad nije horizontalna i bilo je za očekivati da varijance nisu homogene.

Primijenimo sqrt transformaciju na homicide varijablu.

```
regression_4_sqrt <-
  lm(
    sqrt(homicide) ~ sqrt(unemployment) + log(below_poverty) + dependency + no_diploma,
    chicago_pc
  )

summary(regression_4_sqrt)
```

```
##
## Call:
## lm(formula = sqrt(homicide) ~ sqrt(unemployment) + log(below_poverty) +
##     dependency + no_diploma, data = chicago_pc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.80684 -0.37760 -0.03299  0.54123  2.49527
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -5.21795    0.73950  -7.056 8.71e-10 ***
## sqrt(unemployment)  1.02877    0.22975   4.478 2.77e-05 ***
## log(below_poverty)  1.24684    0.29866   4.175 8.24e-05 ***
## dependency          0.07191    0.02242   3.208  0.00200 **
## no_diploma         -0.03531    0.01071  -3.298  0.00151 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9319 on 72 degrees of freedom
## Multiple R-squared:  0.7848, Adjusted R-squared:  0.7728
## F-statistic: 65.64 on 4 and 72 DF,  p-value: < 2.2e-16
```

Promotrimo sada graf sa standardiziranim rezidualima:

```
plot(regression_4_sqrt, 3, col = "cornflowerblue", pch =16)
```

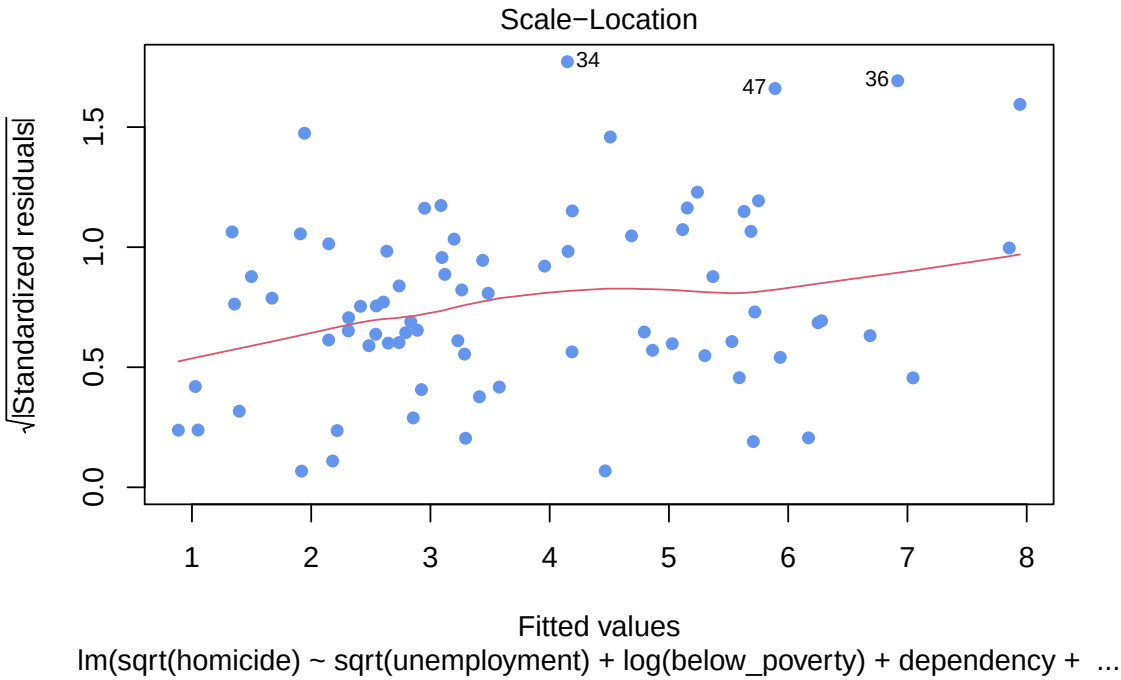

Figure 17: Prikaz standardiziranih reziduala

Vidimo da se pravac znatno izravnao u usporedbi s modelom.
Pokažimo to i testom.

```
bptest(regression_4_sqrt)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  regression_4_sqrt
## BP = 9.2326, df = 4, p-value = 0.05554
```

Test sada prolazi na razini značajnosti 0.05.

Durbin-Watsonovim testom možemo provjeriti postoji li koreliranost između reziduala.

```
dwtest(regression_4_sqrt)
```

```
##
##  Durbin-Watson test
##
## data:  regression_4_sqrt
## DW = 1.7605, p-value = 0.1038
## alternative hypothesis: true autocorrelation is greater than 0
```

Test nam potvrđuje da nema koreliranosti između reziduala što je poželjan ishod kod regresijskih modela.
Promotrimo sada QQ plot. Vidimo da reziduali poprilično dobro slijede pravac što bi sugeriralo njihovu normalnost.

```
plot(regression_4_sqrt, 2, col = "cornflowerblue", pch=16)
```

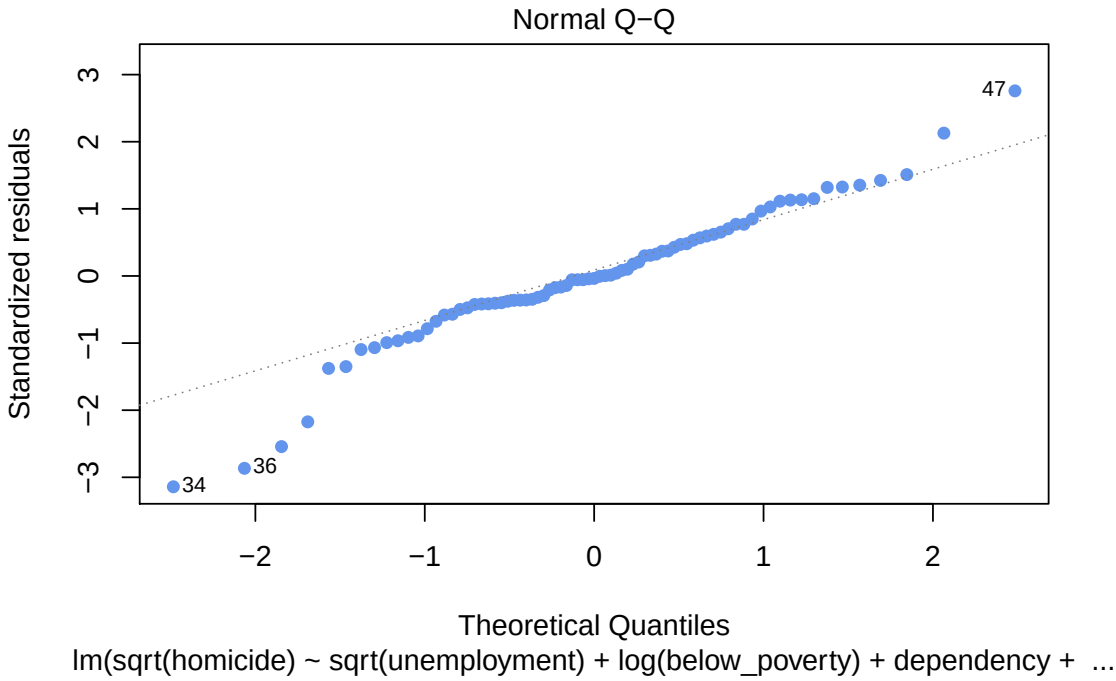

Figure 18: Q-Q Plot

Prethodnu pretpostavku možemo provjeriti Kolmogorov-Smirnovljevim testom.

```
ks.test(rstandard(regression_4_sqrt), 'pnorm')
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  rstandard(regression_4_sqrt)
## D = 0.10191, p-value = 0.3754
## alternative hypothesis: two-sided
```

Test nam govori da su reziduali normalni.
Kao dodatnu provjeru možemo provesti i Lillieforseovu inačicu testa.

```
require(nortest)
```

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

```
lillie.test(rstandard(regression_4_sqrt))
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rstandard(regression_4_sqrt)
## D = 0.10465, p-value = 0.03631
```

Vidimo da Lillieforseov test ne prolazi, to se događa jer je on nešto stroži i precizniji od KS testa.
Promotrimo li histogram reziduala vidimo da oni donekle oponašaju izgled normalne razdiobe.

```
hist((regression_4_sqrt$residuals),
     col = 'cornflowerblue')
```

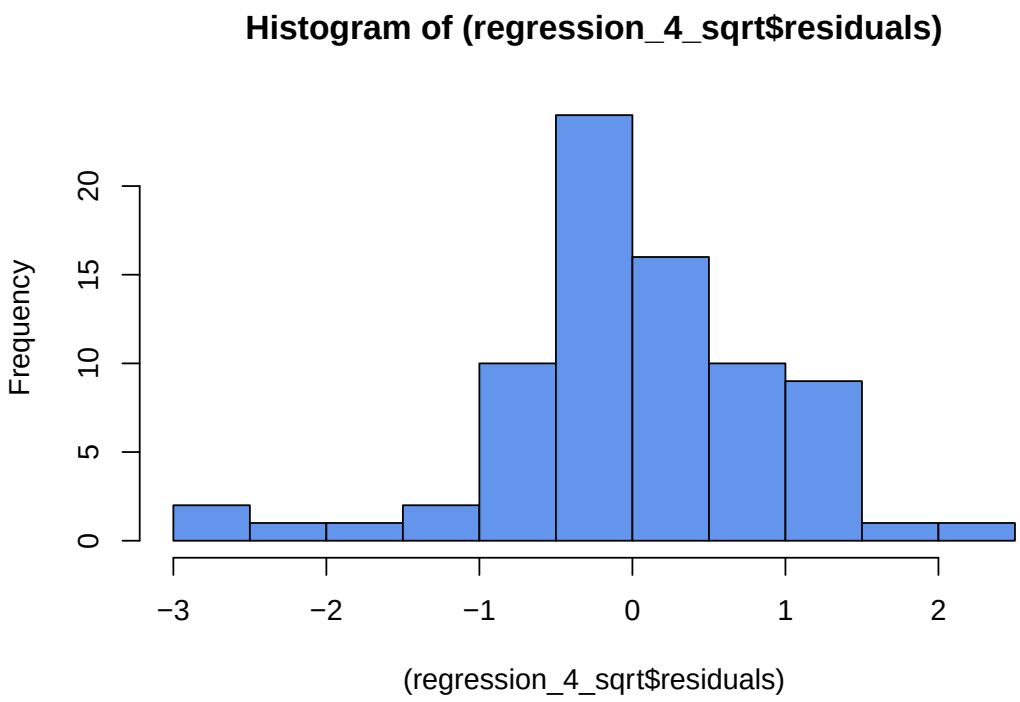

Figure 19: Histogram reziduala

Zaključak

Možemo zaključiti da na stopu ubojstava najviše utječu nezaposlenost i stopa ljudi koji žive ispod razine siromaštva, također znatan utjecaj pokazuju manjak više edukacije i radna sposobnost.
Rezultati su očekivani i stvaraju jasnu poveznicu između negativnih karakteristika kvarta i stope ubojstava.