SAP - Projektni zadatak, grupa Sapsapsaptastično SAP!

Analiza kriminala i socioekonomskih faktora

Leon Novački, Iva Pezo, Lucija Vozab, Dorian Smoljan

7.12.2020.

Opis zadatka

Osim što kriminal i kriminalna djela kao društvo želimo adekvatno kazniti, želimo ići i korak dalje te ispitati kako ih možemo prevenirati. Koji su motivi, pokretači kriminala i kriminalnih djela teško je i kompleksno pitanje.

Cilj ovog projektnog zadatka je bolje razumijeti te istražiti same pojave kriminala te ispitati u kakvoj su oni vezi sa socio-ekonomskim uvjetima područja u kojem nastaju.

Učitavanje podataka - prosjek po četvrtima

```
dataByBorough = read.csv("data/Chicago_poverty_and_crime.csv")
dim(dataByBorough)
```

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

```
head(dataByBorough)
```

```
##   Community.Area Community.Area.Name Assault..Homicide. Firearm.related
## 1              1         Rogers Park                7.7             5.2
## 2              2          West Ridge                5.8             3.7
## 3              3              Uptown                5.4             4.6
## 4              4      Lincoln Square                5.0             6.1
## 5              5        North Center                1.0             1.0
## 6              6           Lake View                1.4             1.8
##   Below.Poverty.Level Crowded.Housing Dependency No.High.School.Diploma
## 1                22.7             7.9       28.8                   18.1
## 2                15.1             7.0       38.3                   19.6
## 3                22.7             4.6       22.2                   13.6
## 4                 9.5             3.1       25.6                   12.5
## 5                 7.1             0.2       25.5                    5.4
## 6                10.5             1.2       16.5                    2.9
##   Per.Capita.Income Unemployment
## 1             23714          7.5
## 2             21375          7.9
## 3             32355          7.7
## 4             35503          6.8
## 5             51615          4.5
## 6             58227          4.7
```

Učitavanje podataka - ukupni podaci svih zločina u proteklih godinu dana

```
lastYearCrimes = read.csv("data/Crimes_-_One_year_prior_to_present.csv")
dim(lastYearCrimes)
```

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

```
head(lastYearCrimes)
```

```
##      CASE.    DATE..OF.OCCURRENCE                BLOCK IUCR PRIMARY.DESCRIPTION
## 1 JD388829 10/04/2020 08:31:00 PM 086XX S CARPENTER ST 0560             ASSAULT
## 2 JD346990 08/26/2020 01:33:00 PM  011XX N DEARBORN ST 0890               THEFT
## 3 JD403530 10/18/2020 03:50:00 PM     049XX W ADAMS ST 0460             BATTERY
## 4 JD141525 02/05/2020 02:54:00 PM   030XX N HALSTED ST 0860               THEFT
## 5 JD366829 08/26/2020 02:19:00 AM 021XX W CULLERTON ST 0890               THEFT
## 6 JD205528 04/09/2020 02:00:00 PM   029XX S ARCHER AVE 1320     CRIMINAL DAMAGE
##   SECONDARY.DESCRIPTION LOCATION.DESCRIPTION ARREST DOMESTIC BEAT WARD FBI.CD
## 1                SIMPLE            APARTMENT      N        N  613   21    08A
## 2         FROM BUILDING            APARTMENT      N        N 1824    2     06
## 3                SIMPLE               STREET      N        N 1533   28    08B
## 4          RETAIL THEFT           DRUG STORE      N        N 1933   44     06
## 5         FROM BUILDING            APARTMENT      N        N 1234   25     06
## 6            TO VEHICLE               STREET      N        N  913   11     14
##   X.COORDINATE Y.COORDINATE LATITUDE LONGITUDE                    LOCATION
## 1      1170827      1847522 41.73707 -87.64972 (41.737074199, -87.64972468)
## 2           NA           NA       NA        NA
## 3           NA           NA       NA        NA
## 4           NA           NA       NA        NA
## 5           NA           NA       NA        NA
## 6      1168260      1885596 41.84161 -87.65803 (41.841609341, -87.65803375)
```

```
# Dimenzije dataseta dataByBorough:
names(dataByBorough)
```

```
##  [1] "Community.Area"         "Community.Area.Name"    "Assault..Homicide."
##  [4] "Firearm.related"        "Below.Poverty.Level"    "Crowded.Housing"
##  [7] "Dependency"             "No.High.School.Diploma" "Per.Capita.Income"
## [10] "Unemployment"
```

Skup podatak dataByBorough sastoji se od 77 zapisa od kojih svaki označava određeni kvart garda Chicaga te je opisan 10 varijabli.

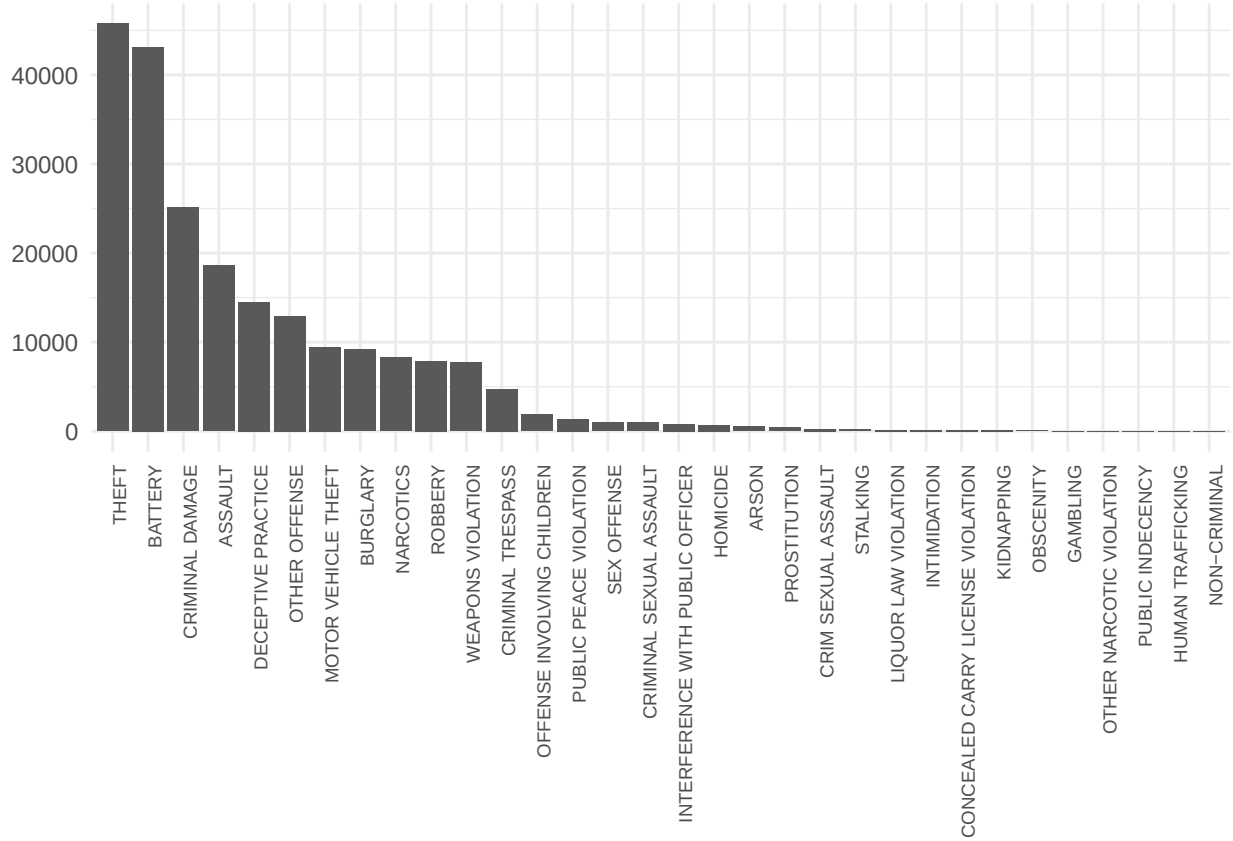
```
# Dimenzije dataseta lastYearCrimes:
names(lastYearCrimes) # imena stupaca
```

```
##  [1] "CASE."                "DATE..OF.OCCURRENCE"  "BLOCK"
##  [4] "IUCR"                 "PRIMARY.DESCRIPTION"  "SECONDARY.DESCRIPTION"
##  [7] "LOCATION.DESCRIPTION" "ARREST"               "DOMESTIC"
## [10] "BEAT"                 "WARD"                 "FBI.CD"
## [13] "X.COORDINATE"         "Y.COORDINATE"         "LATITUDE"
## [16] "LONGITUDE"            "LOCATION"
```

Skup podataka lastYearCrimes sastoji se od 216032 slučajeva opisanih pomoću 17 varijabli. Od kojih će nas najviše interesirati vrijeme događaja te primarni opis zločina.

Bavit ćemo se usporedbama različitih vrsta zločina s obzirom na različite faktore. Pogledajmo histogram frekvencija zločina različitih vrsta.

```
library(dplyr)
library(magrittr)
library(ggplot2)
library(knitr)

lastYearCrimes$svizlociniCount <- 1
sviZlociniPrimarniOpis <- aggregate(svizlociniCount ~ PRIMARY.DESCRIPTION,
                        data = lastYearCrimes, FUN=sum)


lastYearCrimes %>% group_by(PRIMARY.DESCRIPTION) %>%summarise(count=sum(svizlociniCount)) %>%ggplot(aes
              geom_bar(stat = "identity") +
              xlab("Primarni opis") + theme_minimal()+ theme(axis.title.x=element_blank()) +
  theme(axis.title.y=element_blank()) +
  theme(axis.text.y = element_text(size= 9),
        axis.text.x = element_text(size = 7, angle = 90, hjust=1))
```

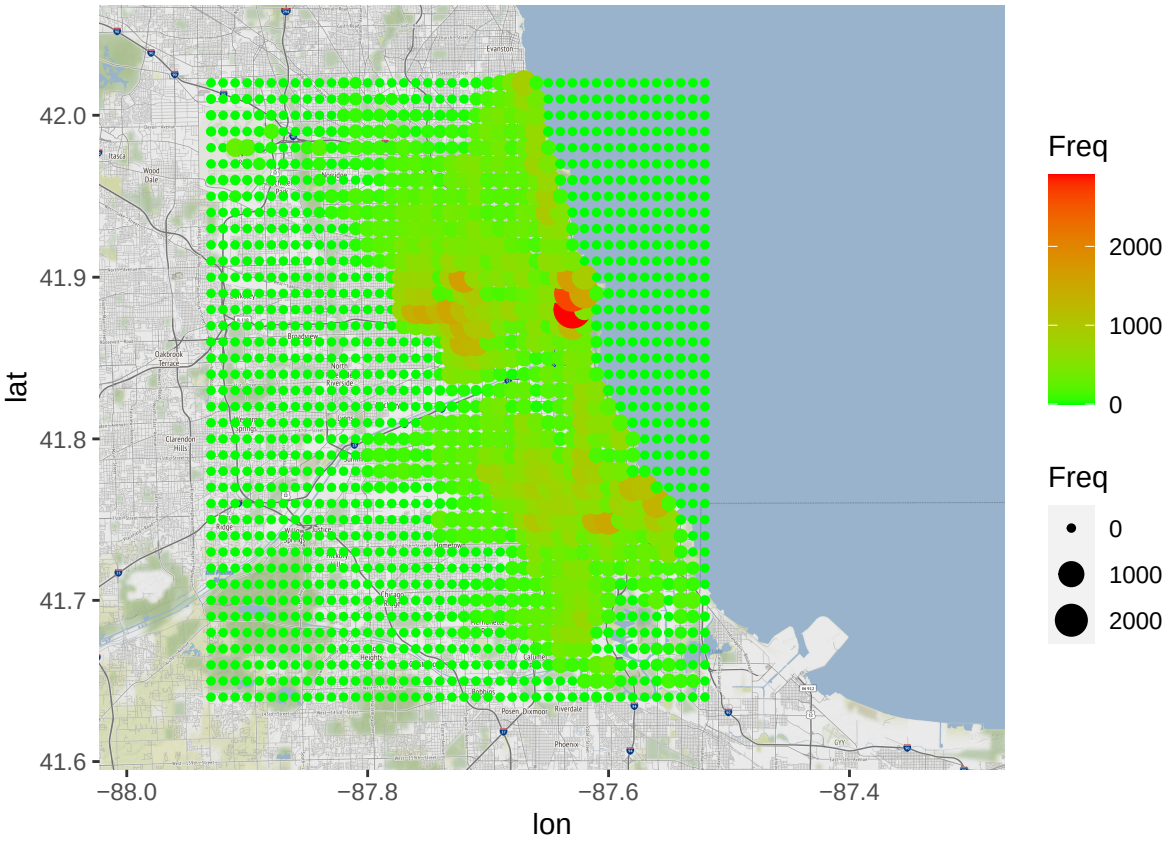


Najzastupljeniji zločini su krađe te fizički napadi dok su ne kriminalne radnje, trgovina ljudima i javna nepristojnost najmanje prisutni.

Frekvencija zločina nije jednaka u svakom dijelu grada. Poznato je da su određeni dijelovi grada sigurniji od drugih. Prikažimo to na karti.

```r
#ucitavanje karte Chicaga
library(ggmap)
 chicago <- get_stamenmap(bbox = c(left = -88.0225, bottom = 41.5949, right = -87.2713, top = 42.0677),

 #karta grada
 #ggmap(chicago)

 LatLonCounts<-as.data.frame(table(round(lastYearCrimes$LONGITUDE,2), round(lastYearCrimes$LATITUDE,2)))

 LatLonCounts$Long <- as.numeric(as.character(LatLonCounts$Var1))
 LatLonCounts$Lat <- as.numeric(as.character(LatLonCounts$Var2))

 #karta grada + intenzitet zločina na tom mjestu
 ggmap(chicago) + geom_point(data = LatLonCounts, aes(x = Long, y = Lat, color = Freq, size = Freq)) + s
```

```
## Warning: Removed 82 rows containing missing values (geom_point).
```



Iz prikaza možemo naslutiti da intenzitet zločina nije jednak u svim djelovima grada.

Case study: Učestalost zločina ovisno o dobu dana

Jedno od pitanja kojima ćemo se baviti jest veza doba dana i učestalosti zločina. Podatke ćemo prikazati podijelom na 12 razdoblja po 2 sata sa prvim razdobljem od ponoći.

Nakon učitavanja podataka potrebno je pretvoriti odgovarajući stupac "DATE.OF.OCCURENCE" u odgovarajući tip podataka nakon čega su podatci spremni za analizu.

Prvi korak u analizi podataka jest uočavanje anomalija u podatcima. Zbog toga ćemo prvo utvrditi ima li anomalija, te ako ima ukloniti ih.

Ima li nepostojećih zapisa?

```
anyNA(lastYearCrimes$DATE..OF.OCCURRENCE)
```

```
## [1] FALSE
```

```
anyNA(lastYearCrimes$PRIMARY.DESCRIPTION)
```

```
## [1] FALSE
```

Nepostojećih zapisa nema.

Ima li neunesenih zapisa?

```
any(is.null(lastYearCrimes$DATE..OF.OCCURRENCE))
```

```
## [1] FALSE
```

```
any(is.null(lastYearCrimes$PRIMARY.DESCRIPTION))
```

```
## [1] FALSE
```

Nepostojećih zapisa nema.

Svi zapisi imaju unesene podatke o vremenu i datumu te primarnom opisu. Sada krećemo s pretvaranjem podataka u odgovarajući oblik.

```
#pretvaramo datum u POSIXlt oblik za lakšu manipulaciju (POSIXlt je oblik u kojem su datumi "samo" zapi
lastYearCrimes$DATE..OF.OCCURRENCE <- strptime(lastYearCrimes$DATE..OF.OCCURRENCE, format="%m/%d/%Y %I:%

#pretvaramo datum u POSIXct oblik -> u njemu su datumi interno zapisani kao sekunde nakon 1.1.1970, pun
lastYearCrimes$TIMESTAMP <- strftime(lastYearCrimes$DATE..OF.OCCURRENCE, format="%m/%d/%Y %H:%M:%S %p",
```

Testiranje hipoteza

Prvo pitanje koje nas zanima je razlikuje li se učestalost zločina ovisno o tome koje je doba dana. Kako bi lakše došli do zaključka, prikažimo prvo podatke grafički.

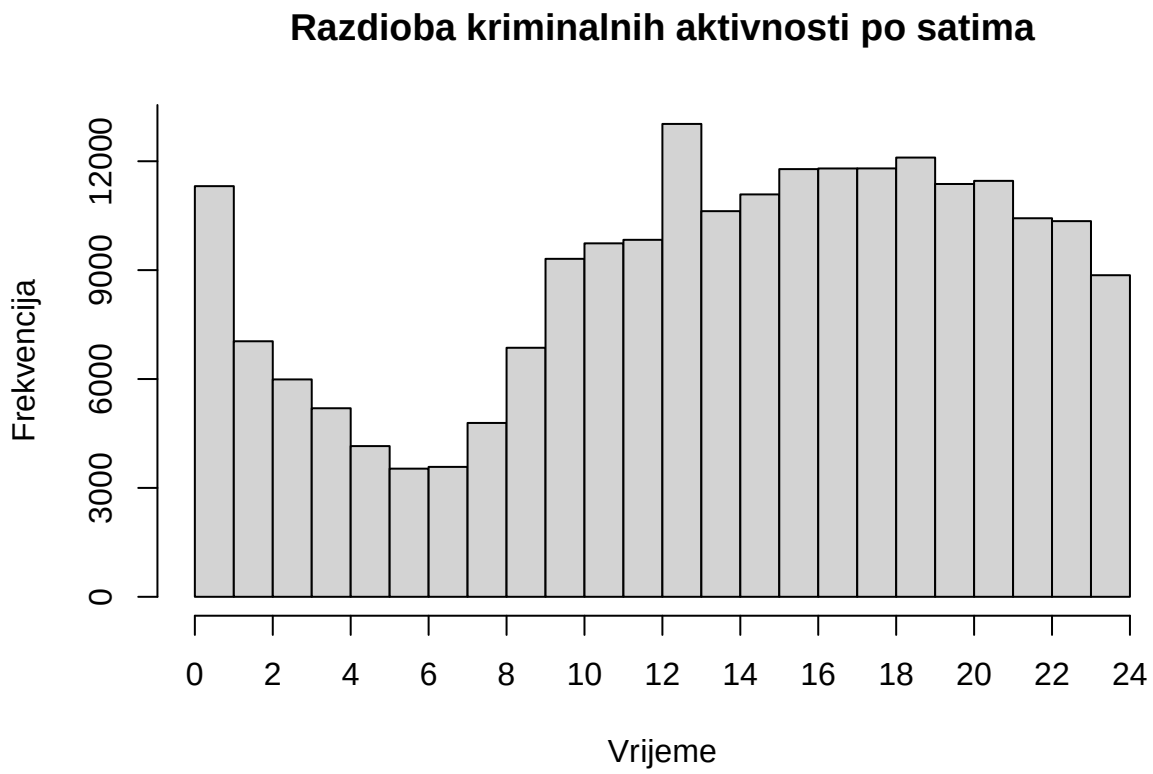
Prikazujemo frekvencije zločina u periodima od jednog i dva sata u danu počeviši od ponoći. .

```
breaksby2 <- seq(0,24, by=2)

lastYearCrimes$TIME.INTERVAL <- cut(lastYearCrimes$DATE..OF.OCCURRENCE$hour,
                                    breaks=breaksby2,
                                    right=FALSE, include.lowest = TRUE)

#frekvencijePoIntervalu <- cbind(table(lastYearCrimes$TIME.INTERVAL))
#colnames(frekvencijePoIntervalu) <- c("Frekvencija zločina")
```

```r
breaksEveryHour <- seq(0,24, by=1)
lastYearCrimes$TIME.EVERYHOUR <- cut(lastYearCrimes$DATE..OF.OCCURRENCE$hour,
                                     breaks=breaksEveryHour, right=FALSE,
                                     include.lowest = TRUE)
frekvencijePoSatu <- cbind(table(lastYearCrimes$TIME.EVERYHOUR))

#colnames(frekvencijePoSatu) <- c("Frekvencija zločina")
#frekvencijePoSatu
```

Grafički prikaz

Prikaz frekvencija po svakom satu prikazujemo pomoću histograma.



Podijelom dana na periode od 2 sata počevši od ponoći dobiven je prikazani histogram. Pogledom na histogram, čini nam se da zaista postoji određena zavisnost kriminala i doba dana - vidimo veću frekvenciju sredinom i u drugoj polovici dana u odnosu na npr. razdoblje ranog jutra (2-6h). No, ovo je samo pretpostavka i moramo je provjeriti odgovarajućim testovima.

Mjere centralne tendencije

```r
# Aritmetička sredina frekvencije zločina po satu
mean(frekvencijePoSatu)
```

```
## [1] 9001.333
```

```r
# Podrezana aritmetička sredina s uklanjanjem po 10% najmanjih i najvećih podataka
mean(frekvencijePoSatu, trim=0.1)
```

```
## [1] 9189.65
```

```r
# 1., 2. i 3. kvartil
quantile(frekvencijePoSatu, probs = c(0.25,0.5,0.75))
```

```
##     25%     50%     75%
##  6643.0 10092.5 11394.5
```

Mjere rasipanja

```r
# Rang - razlika između najvećeg i najmanjeg iznosa u podatcima
max(frekvencijePoSatu)-min(frekvencijePoSatu)
```

```
## [1] 9497
```

```r
# Interkvartilni rang - razlika trećeg i prvog kvartila podataka
IQR(frekvencijePoSatu)
```

```
## [1] 4751.5
```

```r
# Varijanca i standardna devijacija
var(frekvencijePoSatu)
```

```
##          [,1]
## [1,] 9174586
```

```r
sd(frekvencijePoSatu)
```

```
## [1] 3028.958
```

```r
#sqrt(var(frekvencijePoSatu))
#help(var)
```

```r
# Koeficijent varijacije -  relativna mjera rasipanja koja opisuje rasipanje podataka u odnosu na njiho
suppressWarnings(require(raster,quietly = TRUE))
```

```
##
## Attaching package: 'raster'

## The following object is masked from 'package:magrittr':
##
##     extract

## The following object is masked from 'package:dplyr':
##
##     select
```

```r
cv(frekvencijePoSatu)
```

```
## [1] 33.6501
```

```r
sd(frekvencijePoSatu)/mean(frekvencijePoSatu)
```

```
## [1] 0.336501
```

Mjere rasipanja u skladu su s očekivanim, nema prevelikih iznanađenja.

Kako bi ustanovili postoji li zaista veća vjerojatnost pojave zločina u određeno doba dana, krećemo od osnovne provjere - ravnaju li se frekvencije zločina po uniformnoj distribuciji - jer ukoliko se ravnaju, to bi značilo da je jednaka vjerojatnost pojave zločina kroz cijeli dan.

Za ovaj, a i za daljne testove dijelimo dan na 4 intervala po 6h: 2-8,8-14,14-20,20-02. Navedena razdoblja ugrubo odgovaraju jutru (2-8), prijepodnevu (8-14), poslijepodnevu(14-20), te večeri (20-02). Dodajemo novi stupac (PERIOD..OF.DAY) koji označava kojem u kojem od ta 4 dijela dana se zločin dogodio.

```r
jutroDonja = 7200 #02:00:00
jutroGornja = 28800 #08:00:00

prijepodneDonja = 28800 #08:00:00
prijepodneGornja = 50400 #14:00:00

poslijepodneDonja = 50400
poslijepodneGornja = 72000

calculatePeriod <- function (timestamp){
  secsAfterMidnight = timestamp$hour*3600 + timestamp$min*60 + timestamp$sec
    if ((secsAfterMidnight >= jutroDonja) & (secsAfterMidnight < jutroGornja)){
    return("jutro")
  }else if ((secsAfterMidnight >= prijepodneDonja) & (secsAfterMidnight < prijepodneGornja)){
    return("prijepodne")
  } else if ((secsAfterMidnight >= poslijepodneDonja) & secsAfterMidnight < poslijepodneGornja){
    return("poslijepodne")
  } else{
    return("vecer")
  }
}

periods <- sapply(lastYearCrimes$DATE..OF.OCCURRENCE, calculatePeriod) #primjena funkcije na vektor

lastYearCrimes$PERIOD..OF.DAY <- periods
```

Testiranje uniformnost frekvencije zločina kroz dan

Sada kada imamo odgovarajuće podatke, možemo preći na prvi test - test uniformnosti distribucije. Zanima nas slijede li vjerojatnosti pojave zločina uniformnu razdiobu u odnosu na ta 4 razdoblja. Ukoliko slijede, očekivali bi (ugrubo) jednak broj zločina u svakoj od 4 kategorije.

H0 - frekvencija pojavljivanja zločina ima uniformnu distribuciju, odnosno jednaka je vjerojatnost pojave zločina u bilo koje doba dana

H1 - u određena razdoblja u danu veća je vjerojatnost pojave zločina, odnosno distribucija nije uniformna.

Koristimo hi-kvadrat test za ocjenjivanje priladodbe modela podacima.

```
brojZločinaJutro = sum(lastYearCrimes$PERIOD..OF.DAY == "jutro")
brojZločinaPrijepodne = sum(lastYearCrimes$PERIOD..OF.DAY == "prijepodne")
brojZločinaPoslijepodne = sum(lastYearCrimes$PERIOD..OF.DAY == "poslijepodne")
brojZločinaVečer = sum(lastYearCrimes$PERIOD..OF.DAY == "vecer")

# brojZločinaJutro
# brojZločinaPrijepodne
# brojZločinaPoslijepodne
# brojZlocinaVecer

brojZlocinaJutro + brojZlocinaPrijepodne + brojZlocinaPoslijepodne + brojZlocinaVecer
```

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

```
# lastYearCrimes$DATE..OF.OCCURRENCE[2] < jutroDonja
# print(lastYearCrimes$DATE..OF.OCCURRENCE[2])


obs <- c(brojZločinaJutro, brojZlocinaPrijepodne,brojZlocinaPoslijepodne,brojZlocinaVecer)
exp <- c(.25, .25, .25, .25)
chisq.test(obs, p=exp)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  obs
## X-squared = 19061, df = 3, p-value < 2.2e-16
```

Kao što vidimo, p-vrijednost je jako malena, stoga možemo odbaciti nultu hipotezu, te zaključiti da frekvencija pojavljivanja zločina u danu nema uniformnu razdiobu, odnosno zaista postoji veća vjerojatnost pojavljivanja zločina u određeno doba dana, što je i u skladu s našim pretpostavkama.

Sada kada znamo da je za neke dijelove dana vjerojatnost pojave zločina veća, zanima nas koje je to doba dana, odnosno u kojem dijelu dana se događa najveći broj zločina. U tu svrhu stvaramo novi dataset, sa 365 unosa - po jedan za svaki dan. Svaki unos sadrži broj zločina koji se dogodio ujutro, prijepodne, poslijepodne i navečer. Pogledajmo primjer tih podataka.

##365 redaka, u svakom prosječan broj zločina za svaki tip zločina, te prosječan broj zločina koji se dogodio ujutro, prijepodne, poslijepodne i navečer

```
#transformiramo stupac DATE OF OCCURENCE kako bi izvukli samo datum
lastYearCrimes$DATE..OF.OCCURRENCE <- as.Date(lastYearCrimes$DATE..OF.OCCURRENCE)
DATE <- unique(lastYearCrimes$DATE..OF.OCCURRENCE)
crimesByDay <- data.frame(DATE)
head(crimesByDay[order(crimesByDay$DATE),])
```

```
## [1] "2019-10-24" "2019-10-25" "2019-10-26" "2019-10-27" "2019-10-28"
## [6] "2019-10-29"
```

```r
dim(crimesByDay)
```

```
## [1] 365   1
```

```r
for (i in 1:length(crimesByDay$DATE)){
  datum <- crimesByDay$DATE[i]
  crimesByDay$noOfCrimesMorning[i] <- sum((lastYearCrimes$PERIOD..OF.DAY == "jutro") & (lastYearCrimes$
  crimesByDay$noOfCrimesForenoon[i] <- sum((lastYearCrimes$PERIOD..OF.DAY == "prijepodne") & (lastYearC
  crimesByDay$noOfCrimesAfternoon[i] <- sum((lastYearCrimes$PERIOD..OF.DAY == "poslijepodne") & (lastYea
  crimesByDay$noOfCrimesEvening[i] <- sum((lastYearCrimes$PERIOD..OF.DAY == "vecer") & (lastYearCrimes$
}

head(crimesByDay)
```

```
##         DATE noOfCrimesMorning noOfCrimesForenoon noOfCrimesAfternoon
## 1 2020-10-04                79                 150                  157
## 2 2020-08-26                80                 164                  212
## 3 2020-10-18                96                 117                  136
## 4 2020-02-05                75                 205                  228
## 5 2020-04-09                52                 107                  155
## 6 2020-03-08                93                 142                  221
##   noOfCrimesEvening
## 1               183
## 2               179
## 3               145
## 4               123
## 5               108
## 6               180
```

Ovaj novi dataset nam otvara nove mogućnosti testiranja. Ukoliko ponovno pogledamo histogram razdiobe kriminalnih aktivnosti po satima, primjećuje se veća učestalost pojavljivanja zločina u poslijepodnevnom razdoblju dana. Također, ukoliko pogledamo prosječni broj zločina po razdoblju dana tijekom jedne godine, vidimo da je se prosječno najveći broj zločina (191.63) događa u poslijepodnevnom razdoblju.

```
## Prosjecna frekvencija zlocina ujutro po danu iznosi 74.61918
```

```
## Prosjecna frekvencija zlocina prijepodne po danu iznosi 162.7342
```

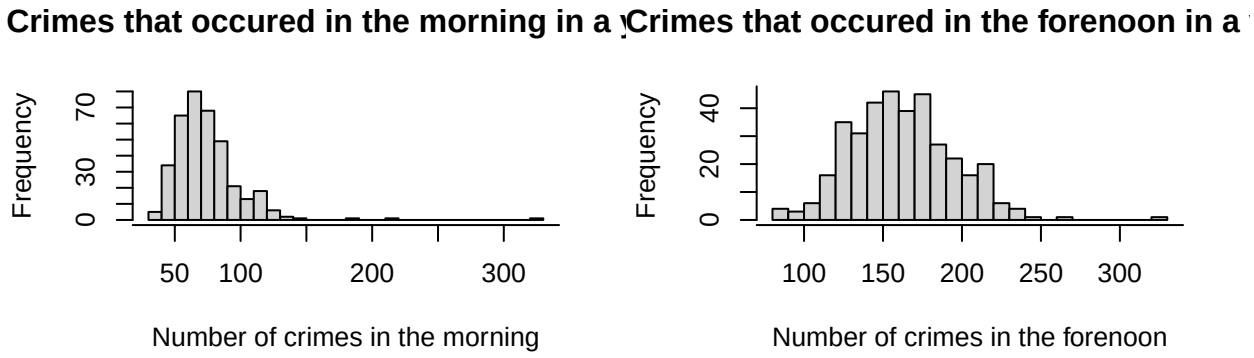
```
## Prosjecna frekvencija zlocina poslijepodne po danu iznosi 191.6329
```

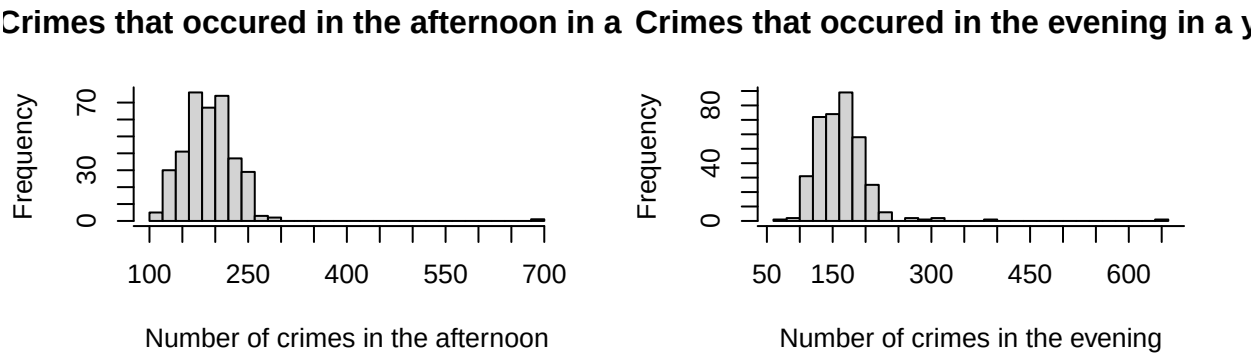
```
## Prosjecna frekvencija zlocina navecer po danu iznosi 162.8822
```

Dakle, imamo razloga vjerovati da je učestalost zločina poslijepodne veća nego učestalost u ostalim dijelovima dana. Međutim, kako bi se uvjerili u tu pretpostavku, moramo provesti odgovarajuće testove.

Pošto želimo testirati hipotezu da je srednja vrijednost frekvencije zločina po danu u poslijepodnevnom razdoblju veća od srednjih vrijednosti frekvencija zločina u ostalim razdobljima dana, koristimo test o jednakosti srednjih vrijednosti populacije, odnosno t-test. Preduvjet t-testa je normalnost i nezavisnost uzoraka. Kako se svaki zločin zabilježava samo jednom, odnosno nemamo situaciju da bi jedan zločin zabilježili 2 ili više puta, razumno je pretpostaviti nezavisnost zločina koji se događaju u različitim vremenskim razdobljima. Normalnost ćemo provjeriti prvo grafički, a zatim i odgovarajućim testom.

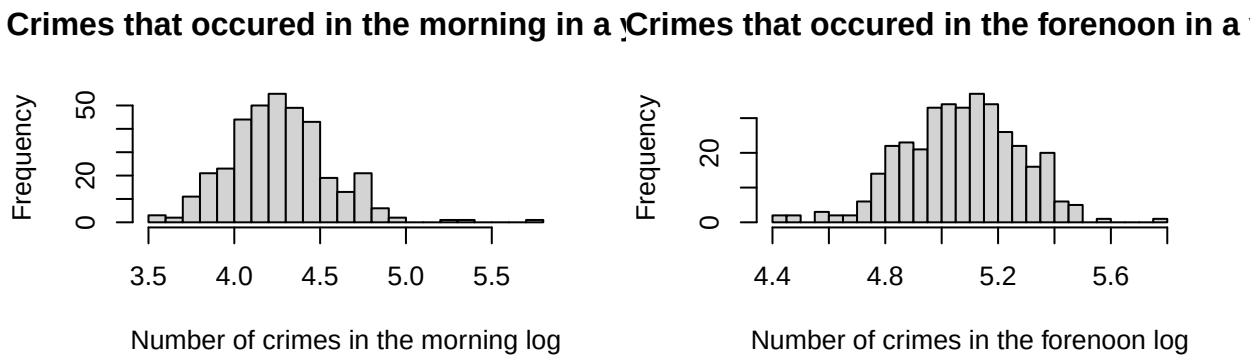
```
## Prikažimo prvo navedene podatke histogramom
```

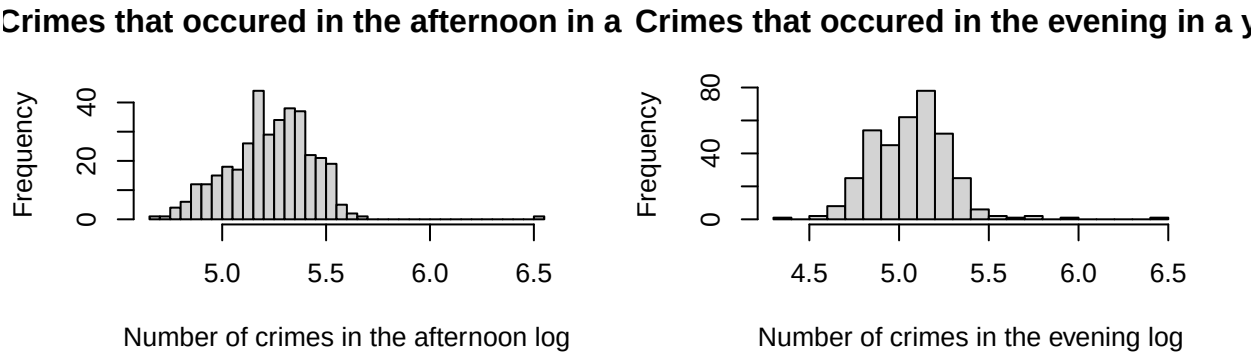




Vidimo da nam distribucija donekle nalikuje normalnoj, posebice za zločine prijepodne, no bolju sličnost sa normalnom možemo postići preko logaritamske transformacije.

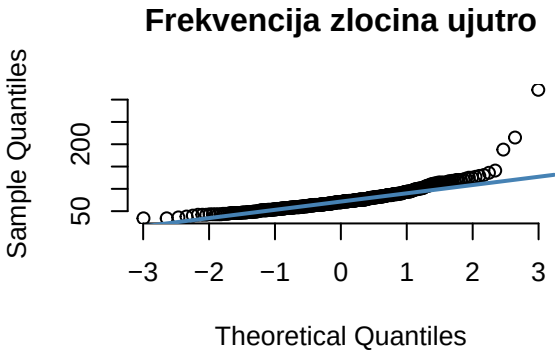
```
par(mfrow=c(2,2))

hist(log(crimesByDay$noOfCrimesMorning),
     main='Crimes that occured in the morning in a year',
     breaks=30,
     xlab='Number of crimes in the morning log')

hist(log(crimesByDay$noOfCrimesForenoon),
     main='Crimes that occured in the forenoon in a year',
     breaks = 30,
     xlab='Number of crimes in the forenoon log')

hist(log(crimesByDay$noOfCrimesAfternoon),
     main='Crimes that occured in the afternoon in a year',
     breaks = 30,
     xlab='Number of crimes in the afternoon log')

hist(log(crimesByDay$noOfCrimesEvening),
     main='Crimes that occured in the evening in a year',
     breaks = 30,
     xlab='Number of crimes in the evening log')
```

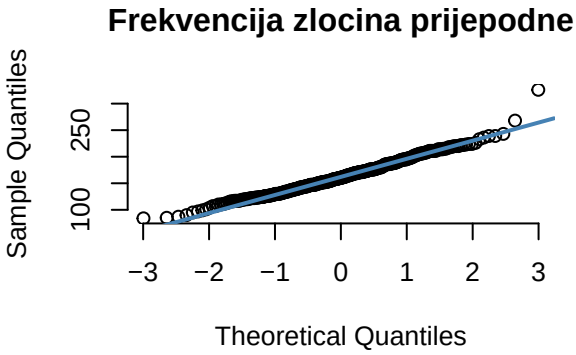


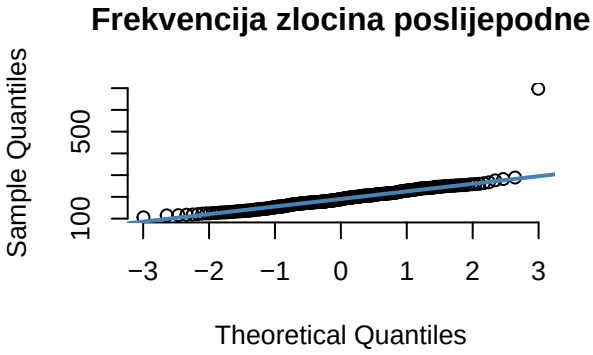


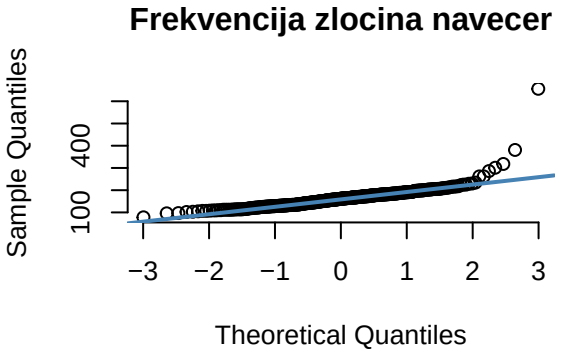
Sada već možemo primijetiti kako distribucija frekvencija zločina unutar svakog razdoblja više-manje nalikujue normalnoj. Kako bi se dodatno uvjerili, prikazati ćemo podatke i qq grafom. Testove za ispitivanje normalnosti nećemo koristiti; zbog velikog broja mjerenja, detektirala bi se i najmanja devijacija od normalne distribucije.

```
par(mfrow=c(2,2))

qqnorm(crimesByDay$noOfCrimesMorning, pch = 1, frame = FALSE,main='Frekvencija zločina ujutro')
qqline(crimesByDay$noOfCrimesMorning, col = "steelblue", lwd = 2)

qqnorm(crimesByDay$noOfCrimesForenoon, pch = 1, frame = FALSE,main='Frekvencija zločina prijepodne')
qqline(crimesByDay$noOfCrimesForenoon, col = "steelblue", lwd = 2)

qqnorm(crimesByDay$noOfCrimesAfternoon, pch = 1, frame = FALSE,main='Frekvencija zločina poslijepodne')
qqline(crimesByDay$noOfCrimesAfternoon, col = "steelblue", lwd = 2)

qqnorm(crimesByDay$noOfCrimesEvening, pch = 1, frame = FALSE,main='Frekvencija zločina navečer')
qqline(crimesByDay$noOfCrimesEvening, col = "steelblue", lwd = 2)
```









Primijećujemo da nam razdiobe poprilično nalikuju normalnoj razdiobi, iznimka je frekvencija zločina ujutro, te u manjoj mjeri frekvencija navečer koji se zbog par outliera donekle odmiču od očekivane vrijednosti za normalnu razdiobu. No, osloniti ćemo se na veliki broj podataka i centralni granični teorem, te nastaviti s testom uz pretpostavku normalnosti.

Posljedna stvar koju moramo ustvrditi prije nego li pređemo na testiranje je jednakost varijanci.

Prije svakog testa, provjeriti ćemo jesu li varijance jednake ili različite, i u skladu s tim koristiti odgovarajući test. U R-u test o jednakosti varijanci implementiran je kroz funkciju var.test(), koja prima uzorke iz populacija čije varijance želimo usporediti.

Prvo testiramo jednakost srednjih vrijednosti frekvencije zločina poslijepodne i zločina ujutro.

Kao što smo rekli, moramo prvo testirati varijance. Pogledajmo prvo varijance (transformiranih) podataka.

```
## Varijanca transformirane frekvencije zlocina ujutro 668.9233
```

```
## Varijacna transformirane frekvencije zlocina poslijepodne 1970.458
```

Vidimo da postoji određena razlika među varijancama. Provedimo sad test jednakosti varijanci.

```
## Test jednakosti varijanci frekvencija zlocina ujutro i frekvencija zlocina poslijepodne
```

```
##
##  F test to compare two variances
##
## data:  crimesByDay$noOfCrimesAfternoon and crimesByDay$noOfCrimesMorning
```

```
## F = 2.9457, num df = 364, denom df = 364, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  2.397844 3.618770
## sample estimates:
## ratio of variances
##           2.945716
```

Kako je p vrijednost izuzetno mala(<2.2e-16), možemo odbaciti H0 da su naše varijance jednake. Stoga t-test jednakosti varijanci provodimo uz uvjet različitosti varijanci.

```
t.test(crimesByDay$noOfCrimesAfternoon, crimesByDay$noOfCrimesMorning, alt = "greater", var.equal = FALS
```

```
##
##  Welch Two Sample t-test
##
## data:  crimesByDay$noOfCrimesAfternoon and crimesByDay$noOfCrimesMorning
## t = 43.514, df = 585.6, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  112.5835      Inf
## sample estimates:
## mean of x mean of y
## 191.63288  74.61918
```

H0 = srednje vrijednosti su jednake H1 = srednja vrijednost frekvencija zločina poslijepodne > srednja vrijednost frekvencija zločina ujutro

Kako je p-vrijednost izuzetno mala, odbacujemo H0 u korist H1, te zaključujemo da je prosječna vrijednost frekvencije pojavljivanja zločina poslijepodne veća od prosječne vrijednosti frekvencije pojavljivanja zločina ujutro.

Na potpuno isti način provodimo testove za ostala 2 razdoblja.

Test jednakosti varijanci frekvencija zločina prijepodne i frekvencija zločina poslijepodne.

```
## Varijanca frekvencije zlocina prijepodne 1085.003
```

```
## Varijacna frekvencije zlocina poslijepodne 1970.458
```

Vidimo da postoji razlika između varijanci, stoga pretpostavljamo da varijance nisu jednake. Provodimo test jednakosti varijanci.

```
## Test jednakosti varijanci frekvencija zlocina prijepodne i frekvencija zlocina poslijepodne
```

```
##
##  F test to compare two variances
##
## data:  crimesByDay$noOfCrimesAfternoon and crimesByDay$noOfCrimesForenoon
## F = 1.8161, num df = 364, denom df = 364, p-value = 1.635e-08
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.478312 2.231034
## sample estimates:
## ratio of variances
##           1.816085
```

P-vrijednost je vrlo mala, te nastavljamo s testom uz uvjet različitosti varijanci.

```
t.test(crimesByDay$noOfCrimesAfternoon, crimesByDay$noOfCrimesForenoon, alt = "greater", var.equal = FA
```

```
##
##  Welch Two Sample t-test
##
## data:  crimesByDay$noOfCrimesAfternoon and crimesByDay$noOfCrimesForenoon
## t = 9.9882, df = 671.6, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  24.13302      Inf
## sample estimates:
## mean of x mean of y
##  191.6329  162.7342
```

H0 = srednje vrijednosti su jednake H1 = srednja vrijednost frekvencija zločina poslijepodne > srednja vrijednost frekvencija zločina prijepodne

P-vrijednost je vrlo mala, odbacujemo H0 u korist H1, te zaključujemo da je srednja vrijednost frekvencije zločina poslijepodne veća od srednje vrijednosti frekvencije zločina prijepodne.

Preostaje nam samo testirati jednakost srednjih vrijednosti frekvencija zločina poslijepodne i frekvencija zločina navečer.

Test jednakosti varijanci frekvencija zločina navečer i frekvencija zločina poslijepodne.

```
## Varijanca frekvencije zlocina navecer 1861.5
```

```
## Varijacna frekvencije zlocina poslijepodne 1970.458
```

Čini nam se da bi varijance mogle biti jednake. Provjeravamo testom jednakosti varijanci.

Test jednakosti varijanci frekvencija zločina navečer i frekvencija zločina poslijepodne.

```
var.test(crimesByDay$noOfCrimesAfternoon, crimesByDay$noOfCrimesEvening)
```

```
##
##  F test to compare two variances
##
## data:  crimesByDay$noOfCrimesAfternoon and crimesByDay$noOfCrimesEvening
## F = 1.0585, num df = 364, denom df = 364, p-value = 0.5877
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8616565 1.3003921
## sample estimates:
## ratio of variances
##           1.058533
```

P-vrijednost je 0.5877, te stoga ne možemo odbaciti hipotezu jednakosti varijanci. Nastavljamo test uz uvjet jednakosti varijanci.

```r
t.test(crimesByDay$noOfCrimesAfternoon, crimesByDay$noOfCrimesEvening, alt = "greater", var.equal = TRU
```

```
##
##  Two Sample t-test
##
## data:  crimesByDay$noOfCrimesAfternoon and crimesByDay$noOfCrimesEvening
## t = 8.8733, df = 728, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  23.41434      Inf
## sample estimates:
## mean of x mean of y
##  191.6329  162.8822
```

H0 = srednje vrijednosti su jednake H1 = srednja vrijednost frekvencija zločina poslijepodne > srednja vrijednost frekvencija zločina navečer

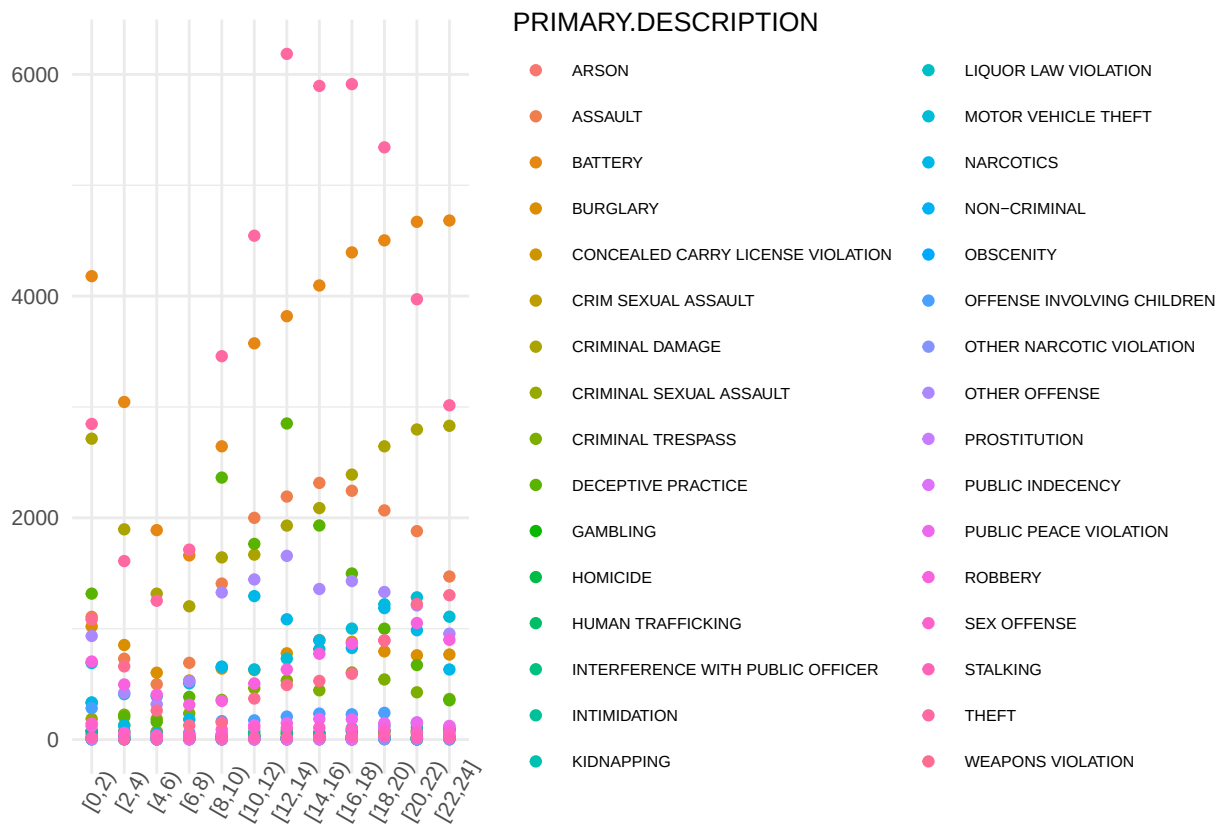
P-vrijednost je izrazito mala, te stoga možemo odbaciti H0 u korist H1, te zaključiti da je prosječna vrijednost frekvencije zločina poslijepodne veća od prosječne frekvencije zločina navečer.

Nakon provedenih testova, zaključujemo da se u (prosječnom) danu najveći broj zločina dogodi poslijepodne, odnosno u razdoblju od 14-20h.

Usporedba prisutnosti različitih zločina tijekom dana

Još nas zanima postoji li možda razlika u distribuciji različitih vrsta zločina tijekom dana, odnosno je li u nekom dijelu dana veća vjerojatnost pojave nekog zločina.

Kako bi stekli malo bolju predodžbu o podacima kojima baratamo i potencijalno uočili neku moguću povezanost, prikazati ćemo ove dvije varijable linijskim grafom.



Opažanja i zaključci

- u razdoblju od 10 do 20 sati krađe su zastupljenije od ostalih vrsta zločina te su češće nego u ostalim djelovima dana
- u razdoblju od 10 do 22 sata provale imaju linearan porast učestalosti
- ostale vrste zločina jednoliko su raspoređene tijekom

Dakle, iz grafičkih prikaza (histogram i linijski graf) zaključujemo da bi mogla postojati zavisnost između vremena u danu i prevalencije određenih vrsta zločina. Kako bi se uvjerili u točnost ove naše tvrdnje, odlično nam može poslužiti test nezavisnosti za kategorijske podatke -> zanima nas nosi li poznavanje vrijednosti jedne varijable (recimo doba dana, podijeljeno na 4 razdoblja) ikakvu informaciju o vrijednosti druge varijable (tip zločina).

Za provođenje ovog testa, prvo je potrebno kreirati kontigencijsku tablicu ove dvije varijable, te tablici dodati margine, odnosno sume redaka i stupaca

```
contTable = table( lastYearCrimes$PRIMARY.DESCRIPTION, lastYearCrimes$PERIOD..OF.DAY)
contTableAddedMargins = addmargins(contTable)
print(contTableAddedMargins)
```

Test nezavisnosti χ^2 test u programskom paketu R implementiran je u funkciji `chisq.test()` koja kao ulaz prima kontingencijsku tablicu podataka koje testiramo na nezavisnost.

Pretpostavka ovog testa je da očekivana frekvencija pojedinog razreda mora biti veća ili jednaka 5 (`chisq.test()` pretpostavlja da je ovaj uvjet zadovoljen stoga je prije provođenja testa potrebno to provjeriti).

```
## Ocekivane frekvencije za razred manje od 5  jutro - GAMBLING :  3.656144
## Ocekivane frekvencije za razred manje od 5  jutro - HUMAN TRAFFICKING :  0.5042957
## Ocekivane frekvencije za razred manje od 5  jutro - NON-CRIMINAL :  0.2521478
## Ocekivane frekvencije za razred manje od 5  jutro - OTHER NARCOTIC VIOLATION :  1.260739
## Ocekivane frekvencije za razred manje od 5  jutro - PUBLIC INDECENCY :  1.008591
## Ocekivane frekvencije za razred manje od 5  poslijepodne - HUMAN TRAFFICKING :  1.295104
## Ocekivane frekvencije za razred manje od 5  poslijepodne - NON-CRIMINAL :  0.6475522
## Ocekivane frekvencije za razred manje od 5  poslijepodne - OTHER NARCOTIC VIOLATION :  3.237761
## Ocekivane frekvencije za razred manje od 5  poslijepodne - PUBLIC INDECENCY :  2.590209
## Ocekivane frekvencije za razred manje od 5  prijepodne - HUMAN TRAFFICKING :  1.0998
## Ocekivane frekvencije za razred manje od 5  prijepodne - NON-CRIMINAL :  0.5499
## Ocekivane frekvencije za razred manje od 5  prijepodne - OTHER NARCOTIC VIOLATION :  2.7495
## Ocekivane frekvencije za razred manje od 5  prijepodne - PUBLIC INDECENCY :  2.1996
## Ocekivane frekvencije za razred manje od 5  vecer - HUMAN TRAFFICKING :  1.1008
## Ocekivane frekvencije za razred manje od 5  vecer - NON-CRIMINAL :  0.5503999
## Ocekivane frekvencije za razred manje od 5  vecer - OTHER NARCOTIC VIOLATION :  2.752
## Ocekivane frekvencije za razred manje od 5  vecer - PUBLIC INDECENCY :  2.2016
```

Vidimo da frekvencije određenih razreda ne zadovoljavaju uvjet da im očekivana frekvencija mora biti veća ili jednaka 5. To najvjerojatnije proizlazi iz činjenice da se određene kategrije zločina pojavljuju puno rijeđe od drugih (što je i logično, npr. naravno da je puno više zločina koji pripadaju kategoriji provala nego zločina koji pripadaju kategoriji krijumčarenja ljudima.). Pošto takvih razreda nema puno, te su svi rezultat navedene činjenice rijeđeg pojavljivanja određenih tipova zločina, najjednostavnije rješenje je iz kontigencijske tablice maknuti sve tipove zločina čiji razredi, u kombinaciji sa vremenima u dani, ne zadovoljavaju uvjet frekvencije. To su tipovi zločina GAMBLING, HUMAN TRAFFICKING, NON-CRIMINAL, OTHER NARCOTIC VIOLATION, PUBLIC INDECENCY.

Ponavljamo isti postupak kreiranja kontigencijske tablice, samo ne uključujući zločine navedenog tipa.

Svjesni smo da smo mogli koristiti i Fisherov egzaktni test, kojemu ne smeta činjenica da postoje očekivane frekvencije < 5. Međutim, naš dataframe je izuzetno velik, te nam je R javljao grešku "FEXACT error 40.Out of workspace.". Daljnim istraživanjem došli smo do zaključka da je Fisherov test komputacijski zahtjevniji, te da se također većinom koristi u situacijama kada veći broj razreda ima očekivanu frekvenciju < 5, te smo se stoga odlučili za ovu metodu izbacivanja konfliktnih razreda i hi-kvadrat test.

```
lastYearCrimesTmp <- subset(lastYearCrimes, !(lastYearCrimes$PRIMARY.DESCRIPTION %in% c("GAMBLING", "HU
contTable = table( lastYearCrimesTmp$PRIMARY.DESCRIPTION, lastYearCrimesTmp$PERIOD..OF.DAY)
contTableAddedMargins = addmargins(contTable)
```

Za novu kontigencijsku tablicu ponavljamo provjeru očekivane frekvencije svakog razreda.

```
noOfInadequateFrequencies = 0
for (col_names in colnames(contTableAddedMargins)){
  for (row_names in rownames(contTableAddedMargins)){
    if (!(row_names == 'Sum' | col_names == 'Sum') ){
      if (((contTableAddedMargins[row_names,'Sum'] * contTableAddedMargins['Sum',col_names]) / contTabl
        noOfInadequateFrequencies = noOfInadequateFrequencies + 1
      }

    }
  }
}

cat("Broj očekivani frekvencija razreda manjih od 5:", noOfInadequateFrequencies)
```

```
## Broj ocekivani frekvencija razreda manjih od 5: 0
```

Vidimo da je sad očekivana frekvencija svakog razreda > 5 te možemo nastaviti sa testom.

H0 - doba dana i kategorija kriminala su međusobno nezavisne varijable H1 - doba dana i kategorija kriminala su međusobno zavisne varijable

```
chisq.test(contTableAddedMargins,correct=F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  contTableAddedMargins
## X-squared = 12764, df = 108, p-value < 2.2e-16
```

```
#čišćenje nepotrebnih tablica iz memorije
rm(lastYearCrimesTmp)
rm(contTable)
rm(contTableAddedMargins)
```

Kao što vidimo, dobili smo p vrijednost $<$ 2.2e-16, odnosno možemo odbaciti H0 u korist H1, te zaključiti da zaista postoji zavisnost između doba dana i vrste zločina. Taj rezultat je u skladu s našim pretpostavkama.

Case study: Usporedba učestalosti krađa i učestalosti kriminala vezanih uz narkotike

Uspoređivati ćemo podatke o krađama te podatke o kriminalima vezanih uz narkotike.

Prvi korak je ispitivanje prisutnosti nepostojećih podataka o vrsti zločina.

```
anyNA(lastYearCrimes$PRIMARY.DESCRIPTION)
```

```
## [1] FALSE
```

Nepostojećih zapisa nema.

Ima li neunesenih zapisa?

```
any(is.null(lastYearCrimes$PRIMARY.DESCRIPTION))
```

```
## [1] FALSE
```

Svi zapisi imaju unesene podatke o primarnom opisu vrste zločina.

Pogledajmo podatke o navedenim vrstama zločina.

```
library(dplyr)
## dodajemo stupce koji označavaju je li se dogodio zločin krađe odnosno zločin s narkoticima
lastYearCrimes$krade <- ifelse(grepl("THEFT",  lastYearCrimes$PRIMARY.DESCRIPTION), 1, 0)
lastYearCrimes$narkotici <- ifelse(grepl("NARCOTICS",  lastYearCrimes$PRIMARY.DESCRIPTION), 1, 0)
```

Želimo prikazati distribuciju navedene dvije vrste zločina ovisno o dobu dana određenih prema predhodnom zadatku.

```r
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'

## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```r
indKrade = which(lastYearCrimes$krade == 1)
#lastYearCrimes[indKrade,]

dobaDanaZaKrade <- lastYearCrimes[indKrade,]$TIME.EVERYHOUR

frekvencijePoSatuKrade <- cbind(table(dobaDanaZaKrade))
colnames(frekvencijePoSatuKrade) <- c("Frekvencija krađa po satu")
#kbl(frekvencijePoSatuKrade, booktabs = T)
```

```r
indNarkotici = which(lastYearCrimes$narkotici == 1)
#lastYearCrimes[indNarkotici,]

dobaDanaZaNarkotike <- lastYearCrimes[indNarkotici,]$TIME.EVERYHOUR

frekvencijePoSatuNarkotici <- cbind(table(dobaDanaZaNarkotike))
colnames(frekvencijePoSatuNarkotici) <- c("Frekvencija narkotika po satu")
#kbl(frekvencijePoSatuNarkotici, booktabs = T)
```

```r
# Osnovna deskriptivna statistika:
summary(frekvencijePoSatuKrade)
```

```
##  Frekvencija krada po satu
##  Min.   : 778
##  1st Qu.:1298
##  Median :2396
##  Mean   :2300
##  3rd Qu.:3238
##  Max.   :3690
```

Iz sažetka mogu se isčitati osnovni podatci; najmanje zločina vezanih uz kađe dogodi se u razdoblju od 5 do 6 sati ujutro, dok se najviše dogodi u razdoblju od 12 do 13. Prosječan broj po satu jest 2300 zločina.

```r
# Osnovna deskriptivna statistika:
summary(frekvencijePoSatuNarkotici)
```

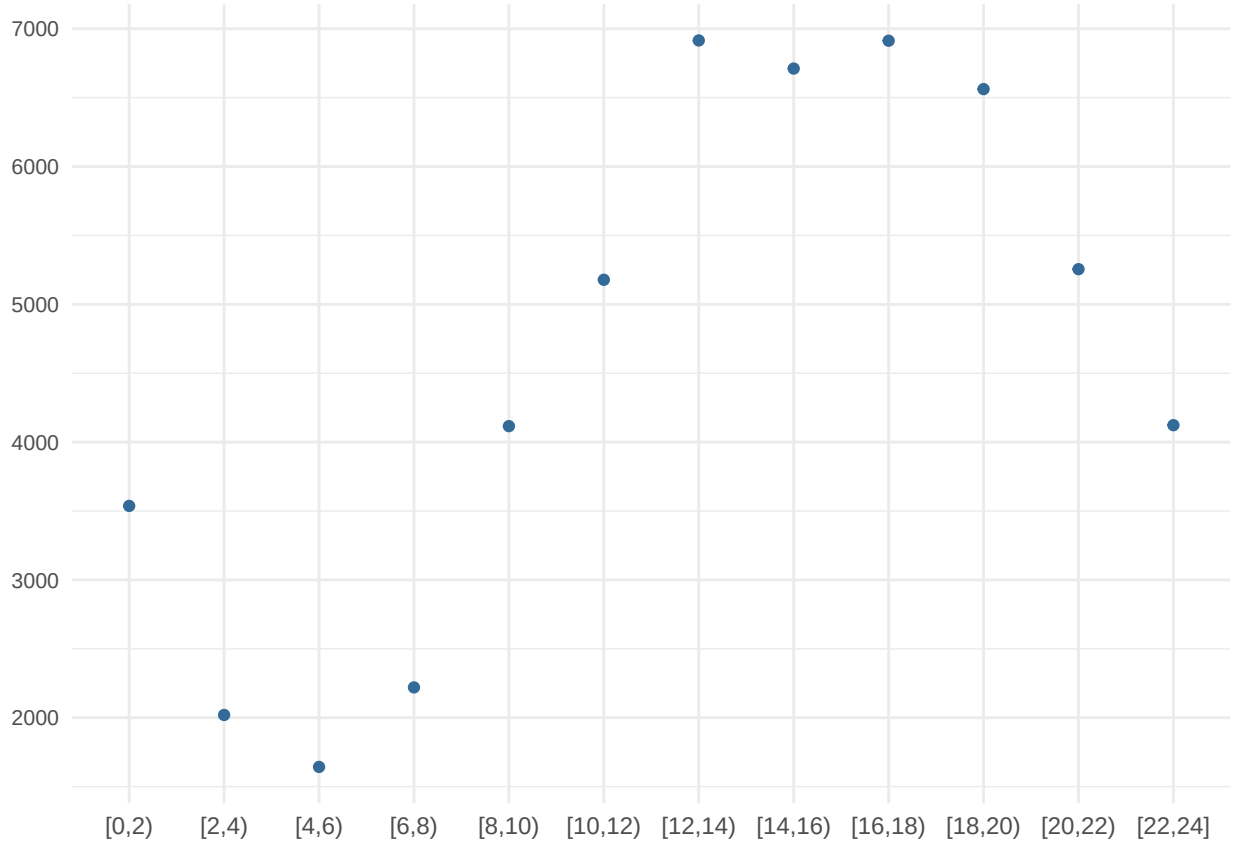
```
##  Frekvencija narkotika po satu
##  Min.   : 16.0
##  1st Qu.:127.2
##  Median :409.5
##  Mean   :344.1
##  3rd Qu.:509.8
##  Max.   :680.0
```

Iz sažetka mogu se isčitati osnovni podatci; najmanje zločina vezanih uz narkotike također se dogodi u razdoblju od 5 do 6 sati ujutro, dok se najviše dogodi u razdoblju od 11 do 12. Prosječan broj po satu jest 344.1 zločin.

Ako je frekvencija krađa daleko veća od učestalosti zločina vezanih uz narkotike, na prvi pogled vidi se da oba dvije raspodijele imaju minimume tijekom razdoblja između 3 i 7 sati ujutro. Prikažimo to grafički.

Grafički prikazujemo frekvencije krađa ovisno o dobu dana.

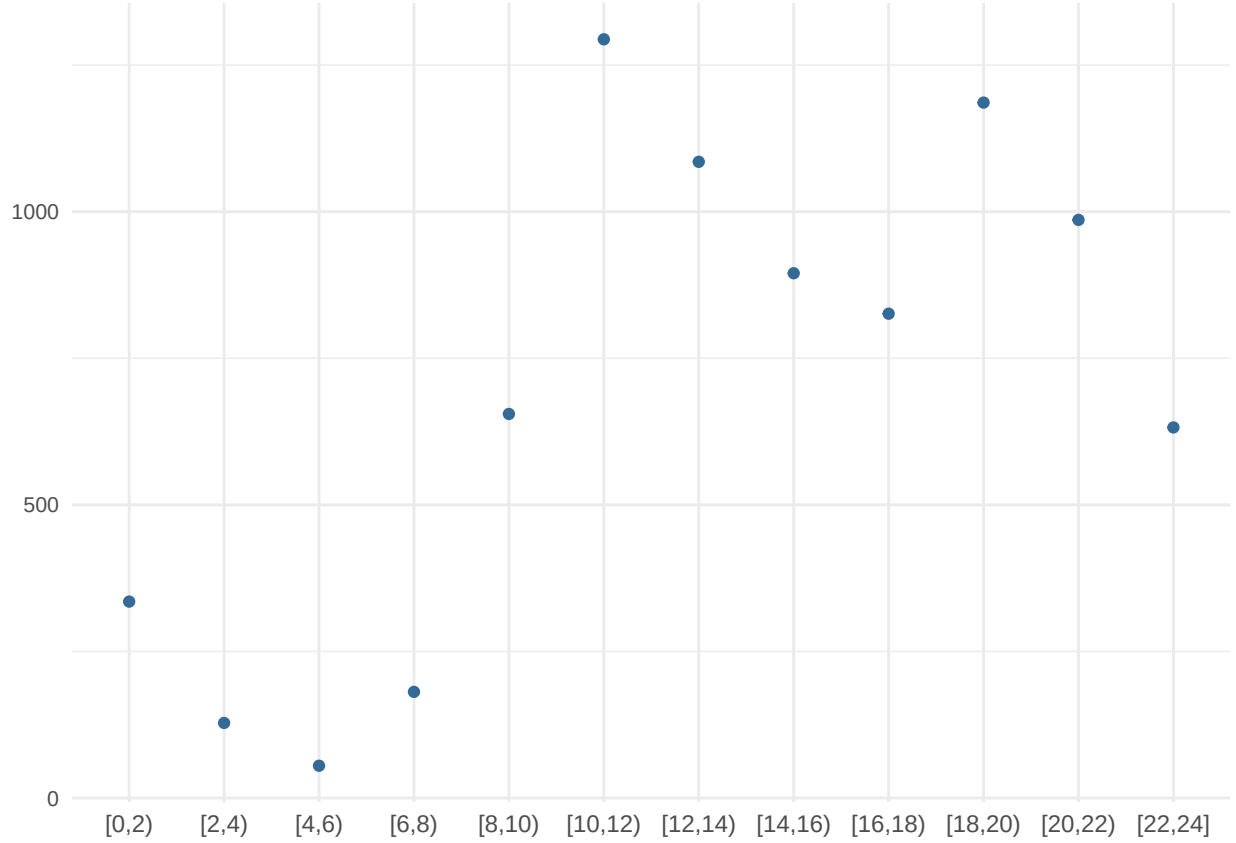
```
lastYearCrimes$kradeTrue <- as.integer(as.logical(lastYearCrimes$krade))

lastYearCrimes$KradeBr <- 1
sveKrade <- aggregate(kradeTrue ~ krade
                        + TIME.INTERVAL, data = lastYearCrimes, FUN=sum)
ind = which(sveKrade$krade == TRUE)

graf <- ggplot(
  data=sveKrade[ind,], aes(x=TIME.INTERVAL,
                       y=kradeTrue, color= krade),show.legend = FALSE) + geom_point()

graf + theme_minimal()+ theme(axis.title.x=element_blank()) +
  theme(axis.title.y=element_blank()) +
  theme(legend.position = "none",
        axis.text.y = element_text(size= 8))
```



Broj prisutnih zločina vezanih uz krađe najprisutniji je između podneva i 18 sati. nakon čega broj pada sve do 4 sata ujutro gdje ima najmanju vrijednost. Krivulja je naizgled pravilna.

Isti postupak provodimo za narkotike

```
lastYearCrimes$narkoticiTrue <- as.integer(as.logical(lastYearCrimes$narkotici))

lastYearCrimes$NarkoticiBr <- 1
sviNarkotici <- aggregate(narkoticiTrue ~ narkotici
                          + TIME.INTERVAL, data = lastYearCrimes, FUN=sum)
## treba maknuti false vrijednosti iz tablice
ind = which(sviNarkotici$narkotici == TRUE)

graf <- ggplot(
  data=sviNarkotici[ind,], aes(x=TIME.INTERVAL,
                        y=narkoticiTrue, color= narkotici),show.legend = FALSE) + geom_point()

graf + theme_minimal()+ theme(axis.title.x=element_blank()) +
  theme(axis.title.y=element_blank()) +
  theme(legend.position = "none",
        axis.text.y = element_text(size= 8))
```



Možemo primjetiti najveći broj zločina povezanih sa narkoticima u razdoblju od 10 do 12 sati nakon čega broj linearno pada sve do 18 sati. Daleko najmanji broj prisutaj je u razdoblju od 4 do 6 sati ujutro. Iz grafa ništa nemožemo zaključiti o prisutnosti zločina kroz doba dana.

Zanima nas odnos frekvencija kriminala vezanih uz narkotike u odnosu na kriminale vezane uz krađu. Najjednostavniji i najpraktičniji način da dobijemo uvid u odnos tih frekvencija je da izračunamo prosječan broj zločina u jednom danu za jednu kategoriju zločina, te za drugu, te provedemo t-test za jednakost srednjih vrijednosti.

```
## Prosjecan dnevni broj zlocina vezanih uz narkotike 22.62466
```

```
## Prosjecan dnevni broj zlocina vezanih uz krade 125.3342
```

Kao što vidimo, u Chicagu se dnevno, u prosjeku, događa gotovo 5.5 puta zločina vezanih uz krađu više nego zločina vezanih uz narkotike. Ovo nam daje snažnu indikaciju da je frekvencija krađa veća od frekvencije narkotika, no to moramo potvrditi odgovarajućim testom.

Dodajemo 2 nova stupa u dataframe crimesByDay - stupac noOfNarcoticsCrimes, te stupac noOfTheftCrimes Čni nam za svaki dan (tj. redak) u dataframeu govore koliko je taj dan bilo zločina povezanih sa narkoticima, odnosno krađama

```
#crimesByDay

for (i in 1:length(crimesByDay$DATE)){
  datum <- crimesByDay$DATE[i]
  crimesByDay$noOfNarcoticsCrimes[i] <- sum((lastYearCrimes$PRIMARY.DESCRIPTION == "NARCOTICS") & (last
  crimesByDay$noOfTheftCrimes[i] <- sum((lastYearCrimes$PRIMARY.DESCRIPTION == "THEFT") & (lastYearCrim
}

head(crimesByDay)
```

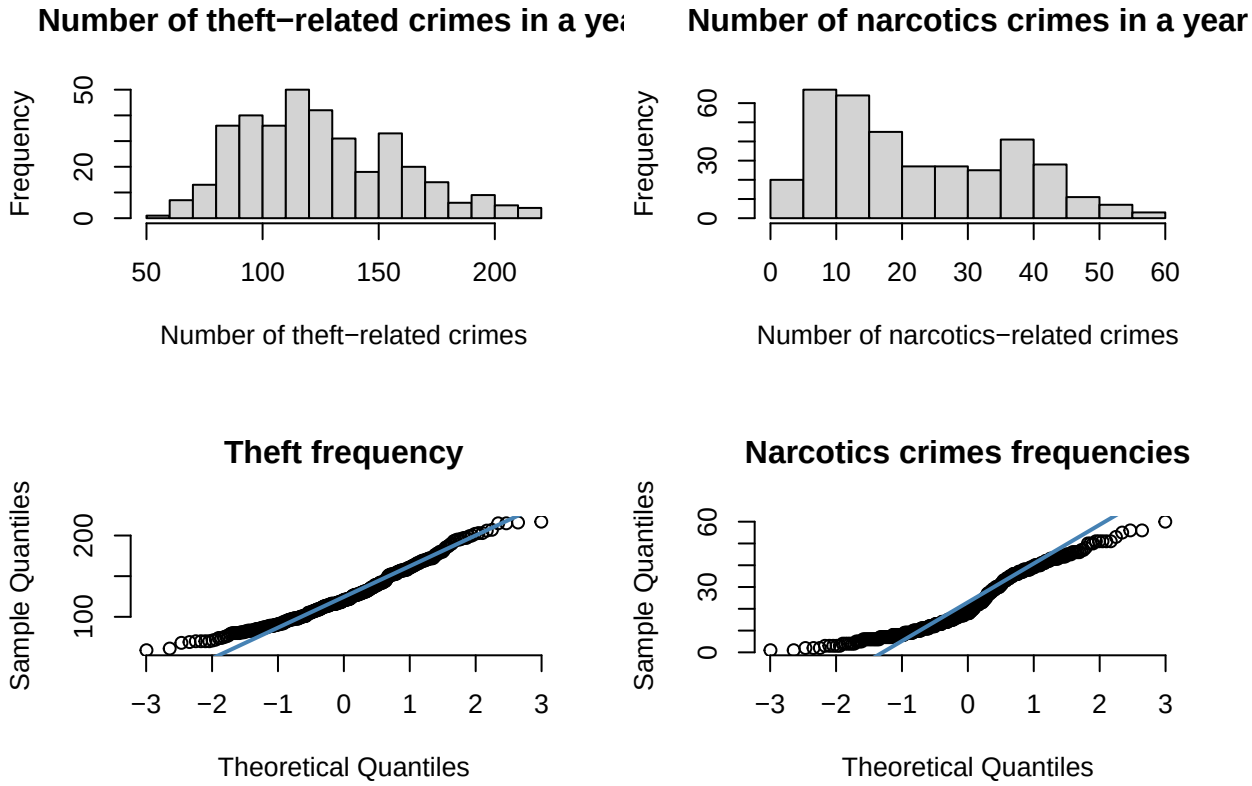
```
##         DATE noOfCrimesMorning noOfCrimesForenoon noOfCrimesAfternoon
## 1 2020-10-04                79                150                  157
## 2 2020-08-26                80                164                  212
## 3 2020-10-18                96                117                  136
## 4 2020-02-05                75                205                  228
## 5 2020-04-09                52                107                  155
## 6 2020-03-08                93                142                  221
##   noOfCrimesEvening noOfNarcoticsCrimes noOfTheftCrimes
## 1               183                  14              99
## 2               179                  14             136
## 3               145                  20              80
## 4               123                  46             182
## 5               108                   4              92
## 6               180                  40             136
```

Ono što želimo testirati je je li prosječni broj zločina vezanih uz krađe po danu veći od prosječnog broja zločina vezanih uz narkotike. Za tu svrhu služi nam t-test, detaljno opisan u 1. zadatku.

Kako smo već jednom imali posla s t-testom, znamo da bi ga uspješno proveli, moramo zadovoljiti neke preduvjete. Specifično, podaci koje testiramo moraju, barem ugrubo imati normalnu distribuciju, te biti međusobno nezavisni. Kako je svaki zločin označen samo jednim opisnikom, ne postoji mogućnost da je npr. isti zločin zabilježen i kao krađa i kao zločin vezan uz narkotike, stoga je razumno pretpostaviti nezavisnost. Da bi utvrdili zadovoljavaju li uzorci i uvjet normalnosti, moramo ih malo bolje pročiti.

```
par(mfrow=c(2,2))

hist(crimesByDay$noOfTheftCrimes,
     main='Number of theft-related crimes in a year',
     breaks=15,
     xlab='Number of theft-related crimes')

hist(crimesByDay$noOfNarcoticsCrimes,
```

```
     main='Number of narcotics crimes in a year',
     breaks=15,
     xlab='Number of narcotics-related crimes')

qqnorm(crimesByDay$noOfTheftCrimes, pch = 1, frame = FALSE,main='Theft frequency')
qqline(crimesByDay$noOfTheftCrimes, col = "steelblue", lwd = 2)

qqnorm(crimesByDay$noOfNarcoticsCrimes, pch = 1, frame = FALSE,main='Narcotics crimes frequencies')
qqline(crimesByDay$noOfNarcoticsCrimes, col = "steelblue", lwd = 2)
```

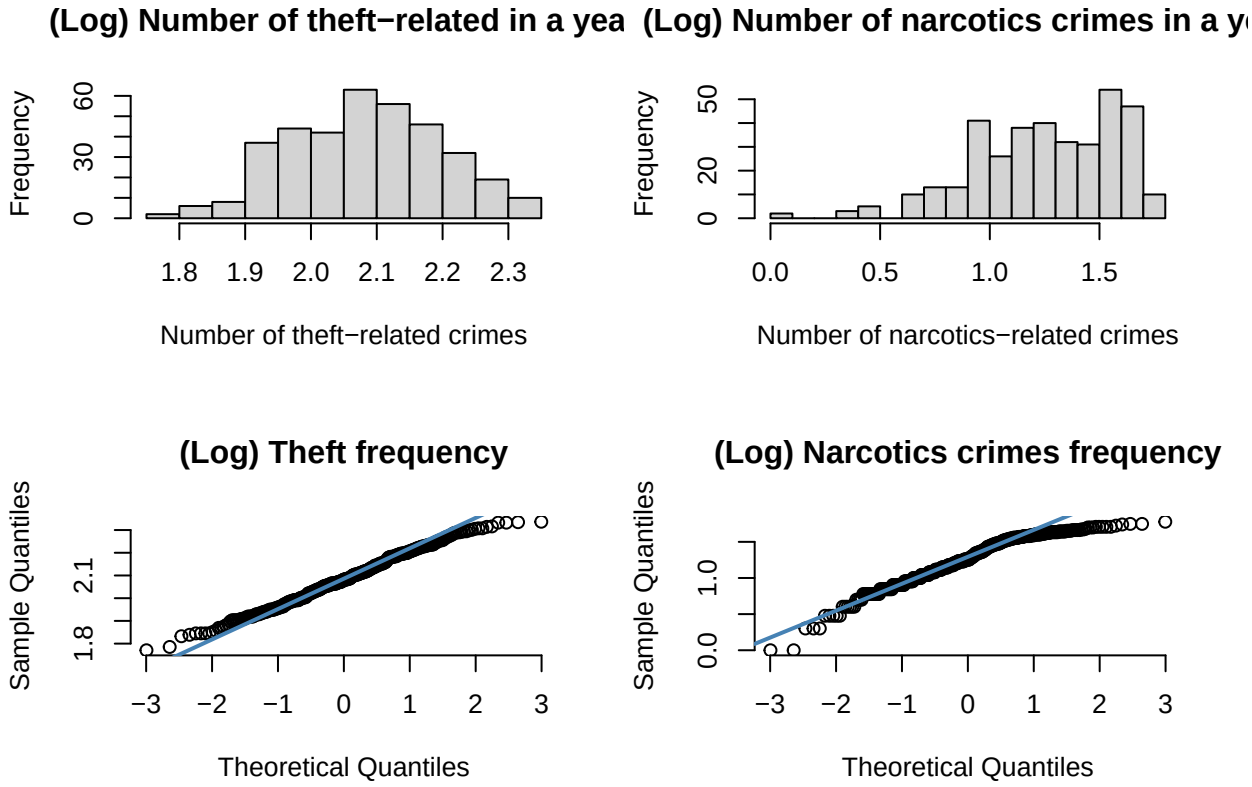


Iz histograma primjećujemo da se razdioba frekvencije krađa više-manje ravna po normalnoj distribuciji, međutim čini se da razdioba frekvencije kriminala vezanih uz narkotike dosta odstupa od normalne distribucije.

Iz qq dijagrama potvrđujemo navedenu tvrdnju - krađe poprilično dobro slijede normalnu distribuciju, dok zločini vezani uz narkotike ne. Znamo da je jedan od načina na koji možemo podatke svesti na normalnu razdiobu logaritamska transformacija, stoga idemo pogledati kako izgledaju transformirani podaci.

```
par(mfrow=c(2,2))

hist(log10(crimesByDay$noOfTheftCrimes),
     main='(Log) Number of theft-related in a year',
     breaks=15,
     xlab='Number of theft-related crimes')

hist(log10(crimesByDay$noOfNarcoticsCrimes),
```

```
     main='(Log) Number of narcotics crimes in a year',
     breaks=15,
     xlab='Number of narcotics-related crimes')

qqnorm(log10(crimesByDay$noOfTheftCrimes), pch = 1, frame = FALSE,main='(Log) Theft frequency')
qqline(log10(crimesByDay$noOfTheftCrimes), col = "steelblue", lwd = 2)

qqnorm(log10(crimesByDay$noOfNarcoticsCrimes), pch = 1, frame = FALSE,main='(Log) Narcotics crimes frequ
qqline(log10(crimesByDay$noOfNarcoticsCrimes), col = "steelblue", lwd = 2)
```



Vidimo da transformirani podaci ipak nešto bolje nalikuju normalnoj razdiobu, pa, oslanjajući se na velik broj podataka i centralni granični teorem, nastavljamo s testiranjem koristeći transformirane podatke.

```
## Varijanca frekvencije krada 0.01331923
```

```
## Varijacna frekvencije narkotika 0.1069384
```

Vidimo da postoji razlika između varijanci, stoga pretpostavljamo da varijance nisu jednake. Provodimo test jednakosti varijanci.

```
## Test jednakosti varijanci frekvencija kradi i narkotika
```

```
##
##  F test to compare two variances
##
```

```
## data:  crimesByDay$noOfTheftCrimes and crimesByDay$noOfNarcoticsCrimes
## F = 5.7402, num df = 364, denom df = 364, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  4.672592 7.051768
## sample estimates:
## ratio of variances
##           5.740212
```

Dobili smo malu p vrijednost, te zaključujemo da varijance nisu jednake. Možemo nastaviti sa testom

```
t.test(log10(crimesByDay$noOfTheftCrimes), log10(crimesByDay$noOfNarcoticsCrimes), alternative = "great
```

```
##
##  Welch Two Sample t-test
##
## data:  log10(crimesByDay$noOfTheftCrimes) and log10(crimesByDay$noOfNarcoticsCrimes)
## t = 45.778, df = 453.29, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.8010143       Inf
## sample estimates:
## mean of x mean of y
##  2.082851  1.251919
```

Dobili smo malu p-vrijednost, te zaključujuemo da je (transformirana) srednja vrijednost frekvencije krađa veća od (transformirane) srednje vrijednosti narkotika. Zbog svojstava logaritamske transformacije, to znači da ista relacija vrijedi i za srednje vrijednosti originalnih podataka.

Case study: Veza između socio-ekonomskih faktora i pojedine kategorije kriminala

Do sada smo razmatrali utjecaj vanjskih čimbenika na kriminalna dijela, no zanima nas i potencijalna korelacija između socioekonomskog statusa pojedinaca i pojedine kategorije kriminala. Socioekonomski status stanovnika pojedinih općina Chicaga praćen je pomoću prosjećne razine neimaštine, pretrpanog kućanstva, nezavršene srednje škole,dohotka po stanovniku te nezaposlenosti.

```
poverty_crime=read.csv("data/Chicago_poverty_and_crime.csv")
summary(poverty_crime)
```

```
##  Community.Area Community.Area.Name Assault..Homicide. Firearm.related
##  Min.   : 1     Length:77           Min.   : 0.00      Min.   : 1.00
##  1st Qu.:20     Class :character    1st Qu.: 4.90      1st Qu.: 5.50
##  Median :39     Mode  :character    Median :10.80      Median :10.60
##  Mean   :39                         Mean   :18.07      Mean   :16.73
##  3rd Qu.:58                         3rd Qu.:32.20      3rd Qu.:25.80
##  Max.   :77                         Max.   :70.30      Max.   :70.30
##  Below.Poverty.Level Crowded.Housing    Dependency    No.High.School.Diploma
##  Min.   : 3.10       Min.   : 0.200   Min.   :15.50   Min.   : 2.9
##  1st Qu.:12.00       1st Qu.: 2.000   1st Qu.:32.30   1st Qu.:13.4
##  Median :18.20       Median : 4.200   Median :38.30   Median :18.5
```

```
##  Mean   :20.29       Mean   : 4.913   Mean   :35.83   Mean   :21.6
##  3rd Qu.:26.10       3rd Qu.: 6.800   3rd Qu.:40.90   3rd Qu.:29.4
##  Max.   :61.40       Max.   :17.600   Max.   :50.20   Max.   :58.7
##  Per.Capita.Income  Unemployment
##  Min.   : 8535     Min.   : 4.2
##  1st Qu.:15467     1st Qu.: 7.8
##  Median :20489     Median :11.5
##  Mean   :25107     Mean   :13.3
##  3rd Qu.:29026     3rd Qu.:17.4
##  Max.   :87163     Max.   :40.0
```

Assault..Homicide prikazuje broj napada i ubojstva na 100.000 stanovnika.

Firearm.related prikazuje broj kriminala povezanih sa vatrenim oružjem na 100.000 stanovnika.

Below.Poverty.Level označava postotak stanovništva u kvartu koja živi ispod granice siromaštva. Granica siromaštva se definira kao količina novaca potrebna za osnovne potrebštine.

Crowded.Housing je postotak stanovništva koje živi u prenapučenom domu. Dom je prenapučen ako u njemu živi više ljudi nego namijenjeno.

Dependency označava postotak stanovništva koji financijski potpuno ili značajno ovisi o drugoj osobi. No.High.School.Diploma je postotak stanovništva bez srednje stručne spreme.

Per.Capita.Income je prosječan dohodak po osobi u pojedinom kvartu.

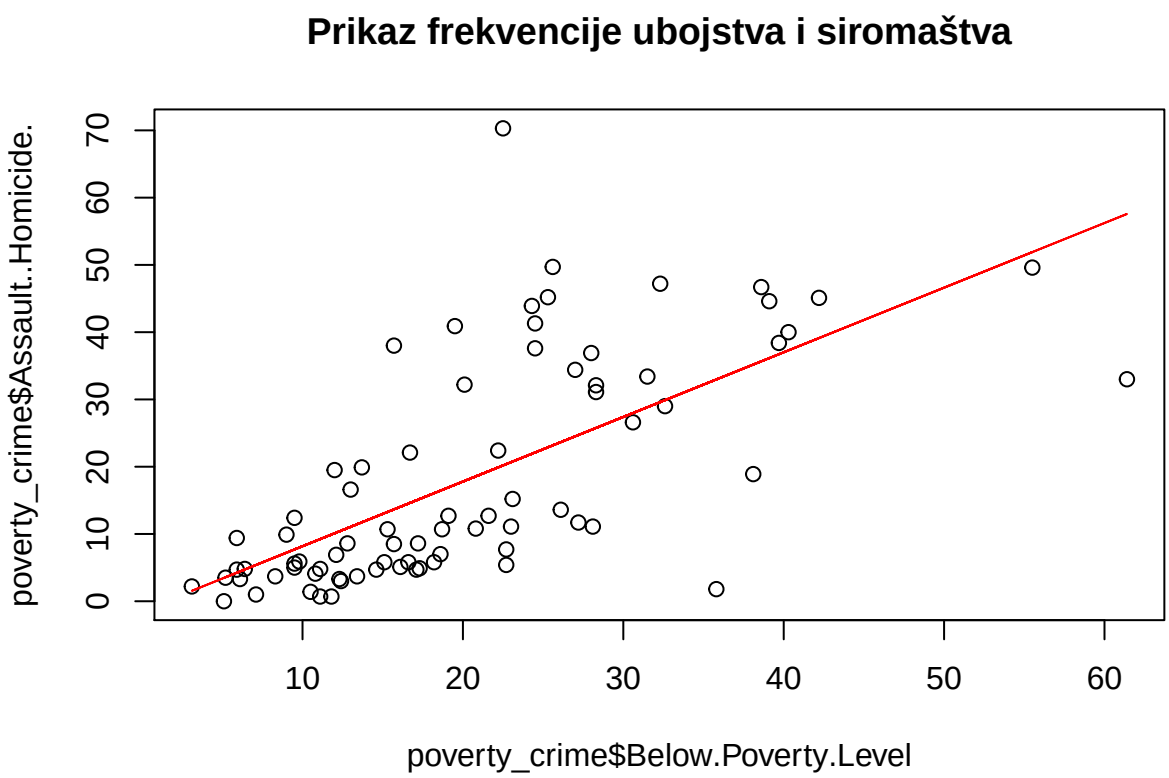
Unemployment označava postotak nezaposlenog stanovništva u kvartu.

Jednostavna linearna regresija

Radimo jednostavne regresijske modele o utjecaju pojedinih faktora na frekvenciju napada i ubojstva. Ne prikazujemo ih sve, ali neke ćemo istaknuti.

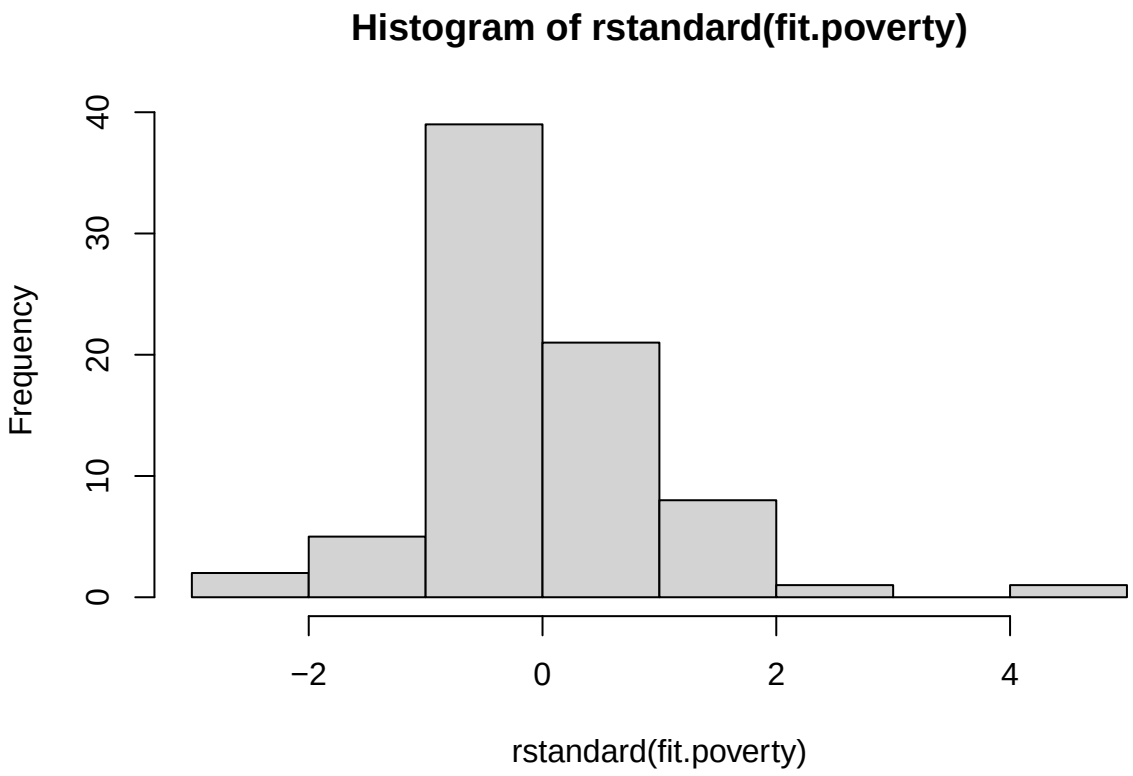
```
#nekoliko jednostavnih regresijskih modela
fit.poverty=lm(Assault..Homicide.~Below.Poverty.Level, data=poverty_crime)
fit.income=lm(Assault..Homicide.~Per.Capita.Income, data=poverty_crime)
fit.crowded=lm(Assault..Homicide.~Crowded.Housing, data=poverty_crime)
fit.dependency=lm(Assault..Homicide.~Dependency, data=poverty_crime)
fit.noHSdiploma=lm(Assault..Homicide.~No.High.School.Diploma, data=poverty_crime)
fit.unemployment=lm(Assault..Homicide.~Unemployment, data=poverty_crime)
```

```
#postotak siromaštva vs frekvencija ubojstva
plot(poverty_crime$Below.Poverty.Level, poverty_crime$Assault..Homicide., main = "Prikaz frekvencije ub
lines(poverty_crime$Below.Poverty.Level, fit.poverty$fitted.values, col="red")
```



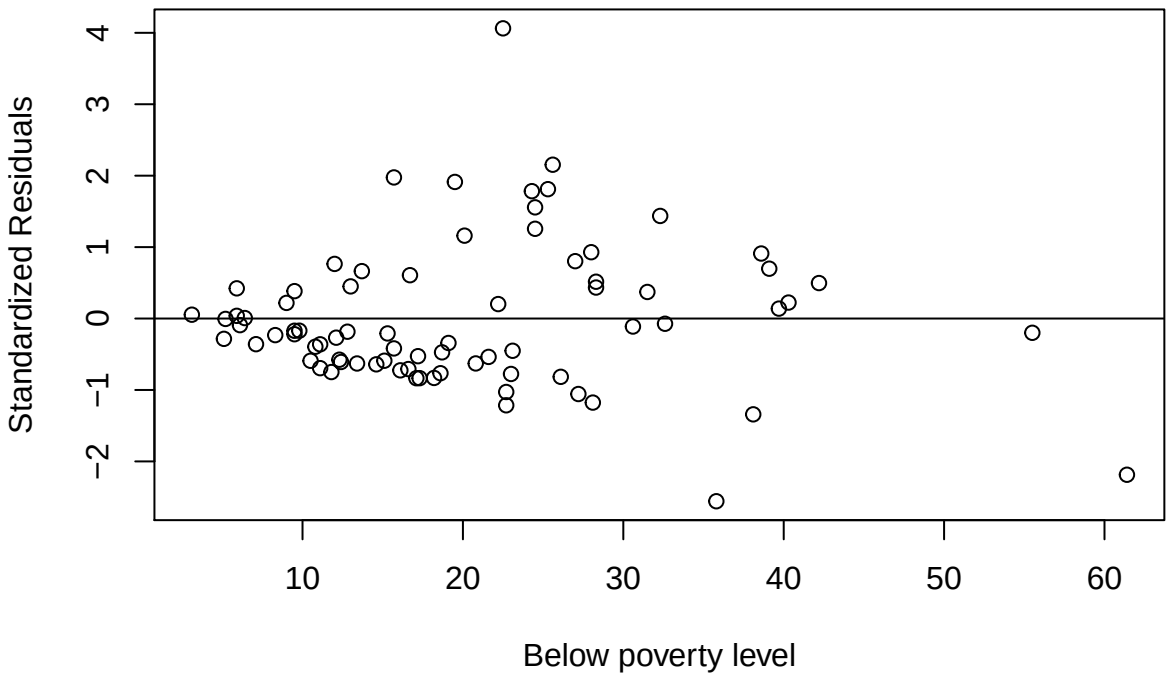
Kako bi mogli uspoređivati ovaj model sa drugima, moramo prvo provjeriti jesu li pretpostavke modela narušene. Normalnost reziduala ćemo provjeriti grafički pomoću histograma.

```
hist(rstandard(fit.poverty))
```



Reziduali ne pokazuju preveliko odstupanje od normalne distribucije, a znamo da je t-test robustan na normalnost pa i dalje možemo radili statističku analizu na temelju modela.

```
plot (poverty_crime$Below.Poverty.Level, rstandard(fit.poverty), ylab="Standardized Residuals", xlab="B
abline(0, 0)
```



Na ovom grafu vidimo da pretpostavka o homogenosti varijance nije točna, što je tipično za studije presjeka. Kod ovakvih istraživanja, treba uzeti u obzir da se greške modela mogu povećavati u ovisnosti o nekim varijablama, ali iz grafa nije očito koja bi varijabla to mogla biti.

Za daljnu analizu modela koristimo funkciju summary()

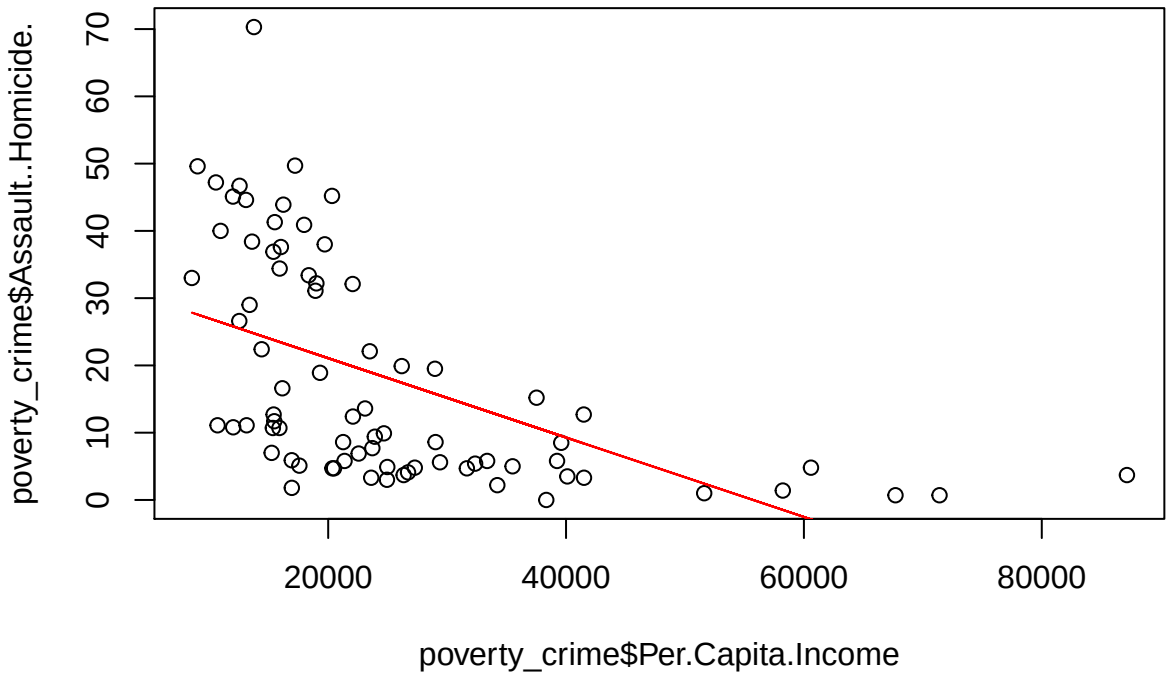
```
summary(fit.poverty)
```

```
##
## Call:
## lm(formula = Assault..Homicide. ~ Below.Poverty.Level, data = poverty_crime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.172  -7.745  -2.304   5.539  50.109
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -1.4320     2.8853  -0.496    0.621
## Below.Poverty.Level   0.9610     0.1239   7.756 3.47e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.42 on 75 degrees of freedom
## Multiple R-squared:  0.4451, Adjusted R-squared:  0.4377
## F-statistic: 60.15 on 1 and 75 DF,  p-value: 3.472e-11
```

Varijabla postotka siromaštva je statistički značajna te model dobro predviđa kriminal. Sama konstanta statistički nije značajna, ali nam to ne smeta jer ćemo kasnije raditi višestruku linearnu regresiju i tada će (nadamo se) konstanta biti točnija.

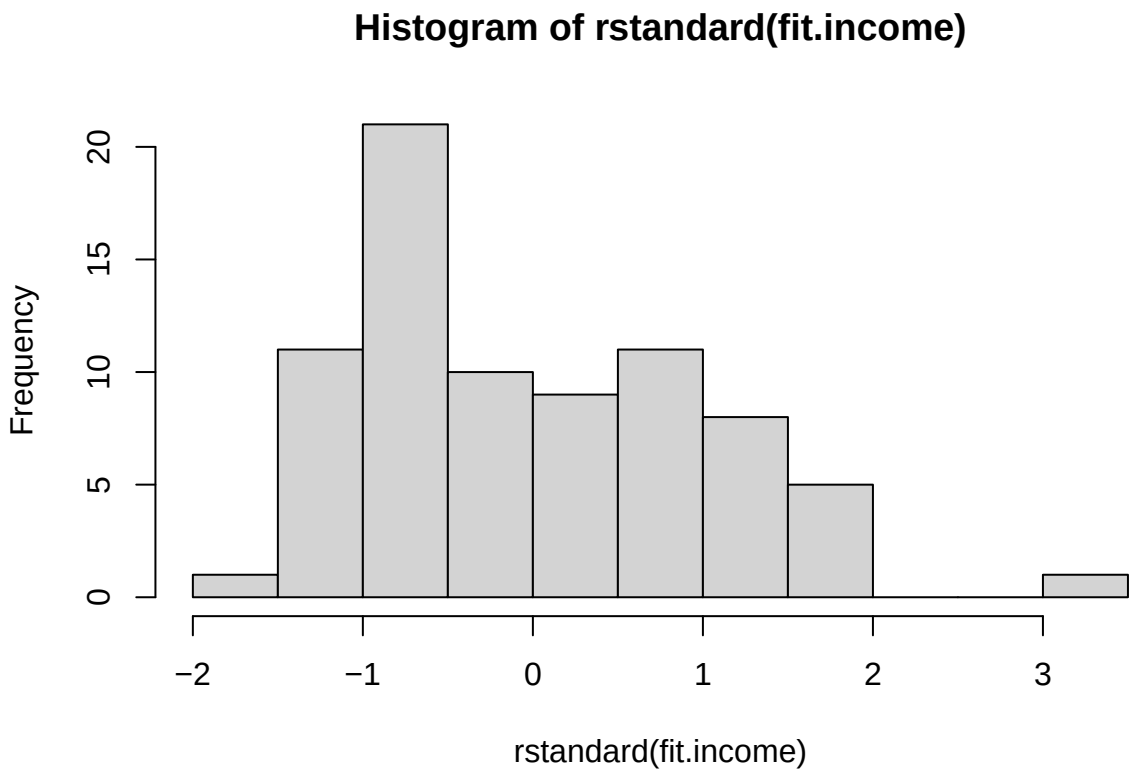
Sljedeće gledamo linearni model ovisnosti ubojstva o prosječnom dohotku

```
#prosječni dohodak vs kriminal
plot(poverty_crime$Per.Capita.Income, poverty_crime$Assault..Homicide.)
lines(poverty_crime$Per.Capita.Income, fit.income$fitted.values, col="red")
```



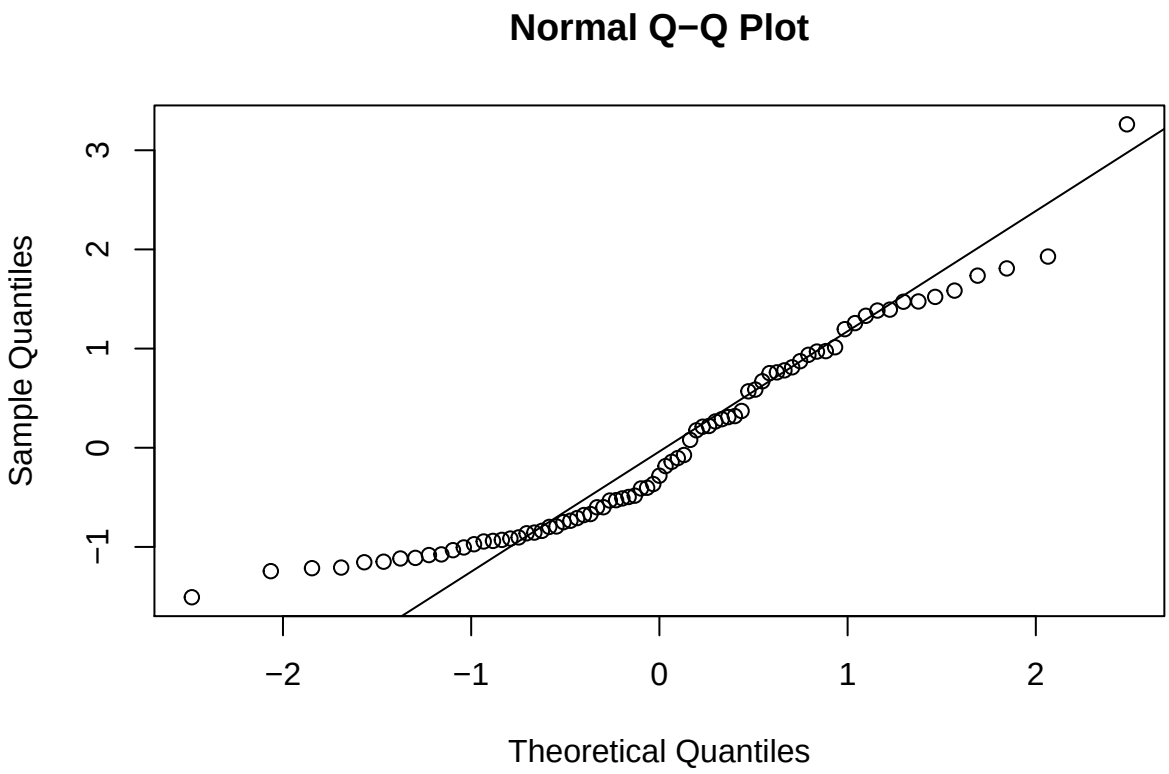
Provjeravamo normalnost reziduala.

```
hist(rstandard(fit.income))
```



Za provjeravanje normalnosti reziduala možemo koristiti i Q-Q plot.

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



Možemo dodatno statistički provjeriti i sa Lilliefors testom. H0 = standarnizirani reziduali slijede normalnu distribuciju.

H1 = standarnizirani reziduali ne slijede normalnu distribuciju.

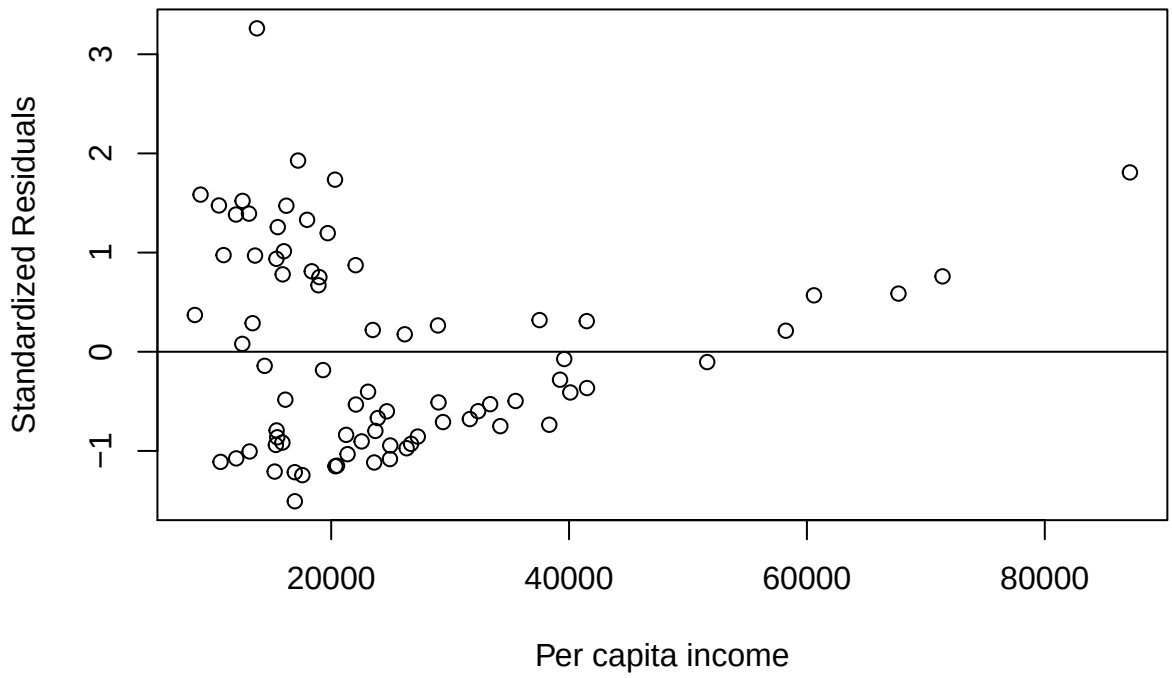
```
library(nortest)
lillie.test(rstandard(fit.income))
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rstandard(fit.income)
## D = 0.14034, p-value = 0.0007087
```

Po testu odbacujemo nultu hipotezu ali i dalje nastavljamo sa analizom jer distribucija reziduala mora barem donekle ličiti na normalnu distribuciju.

Potrebno je i provjeriti homoskedastičnost.

```
plot (poverty_crime$Per.Capita.Income, rstandard(fit.income), ylab="Standardized Residuals", xlab="Per
abline(0, 0)
```



Zbog vrste istraživanja uzorci nisu jednoliko raspoređeni po promatranoj varijabli, te ne možemo ispitati homoskedastičnost. Ima samo 72 uzorka.

Primijetimo da kvartova sa prosječnim dohotkom iznad 60,000 ima malo i zbog toga ne možemo odrediti varijancu reziduala. Takvi kvartovi su rubni slučajevi. Zbog istog razloga ne ispitujemo homoskedastičnost nad ostalim modelima.

```
summary(fit.income)
```

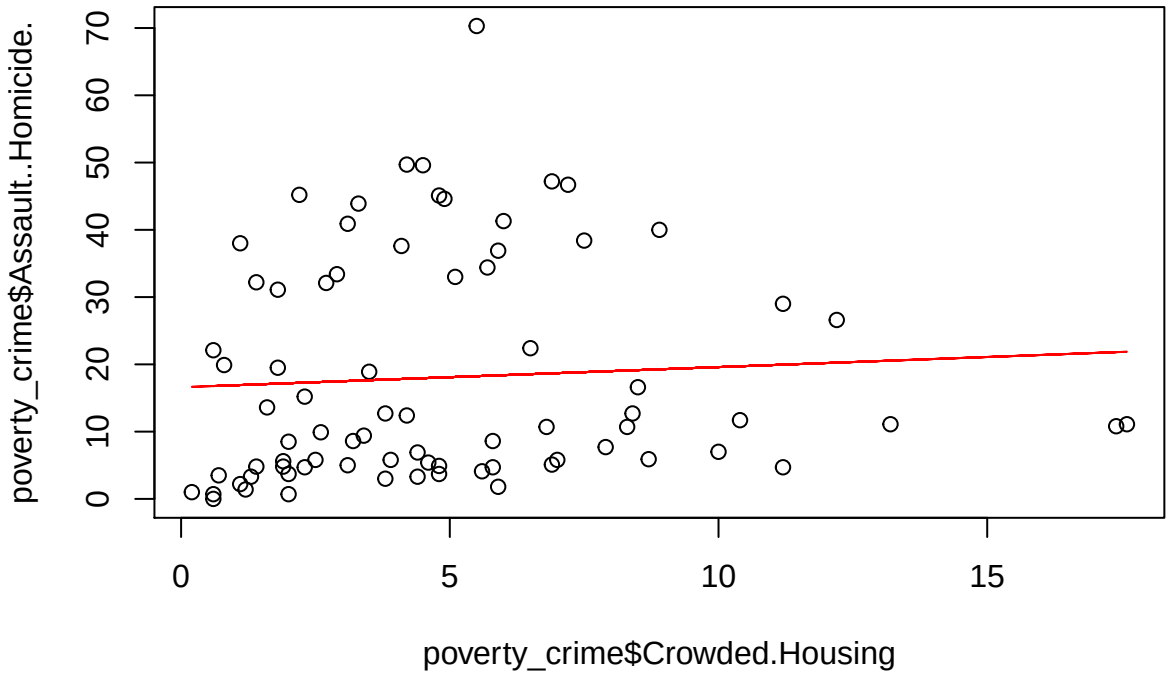
```
##
## Call:
## lm(formula = Assault..Homicide. ~ Per.Capita.Income, data = poverty_crime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.087 -11.986  -3.928  10.910  45.534
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       32.8833950  3.1573233  10.415 3.19e-16 ***
## Per.Capita.Income -0.0005901  0.0001082  -5.452 6.11e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.11 on 75 degrees of freedom
## Multiple R-squared:  0.2838, Adjusted R-squared:  0.2743
## F-statistic: 29.72 on 1 and 75 DF,  p-value: 6.114e-07
```

Koeficijent uz dohodak je vrlo mali, ali to je očekivano jer dohodak se izražava relativno velikim brojevima u odnosu na broj kriminala.

Također, p-vrijednost je vrlo mala pa je model statistički značajan.

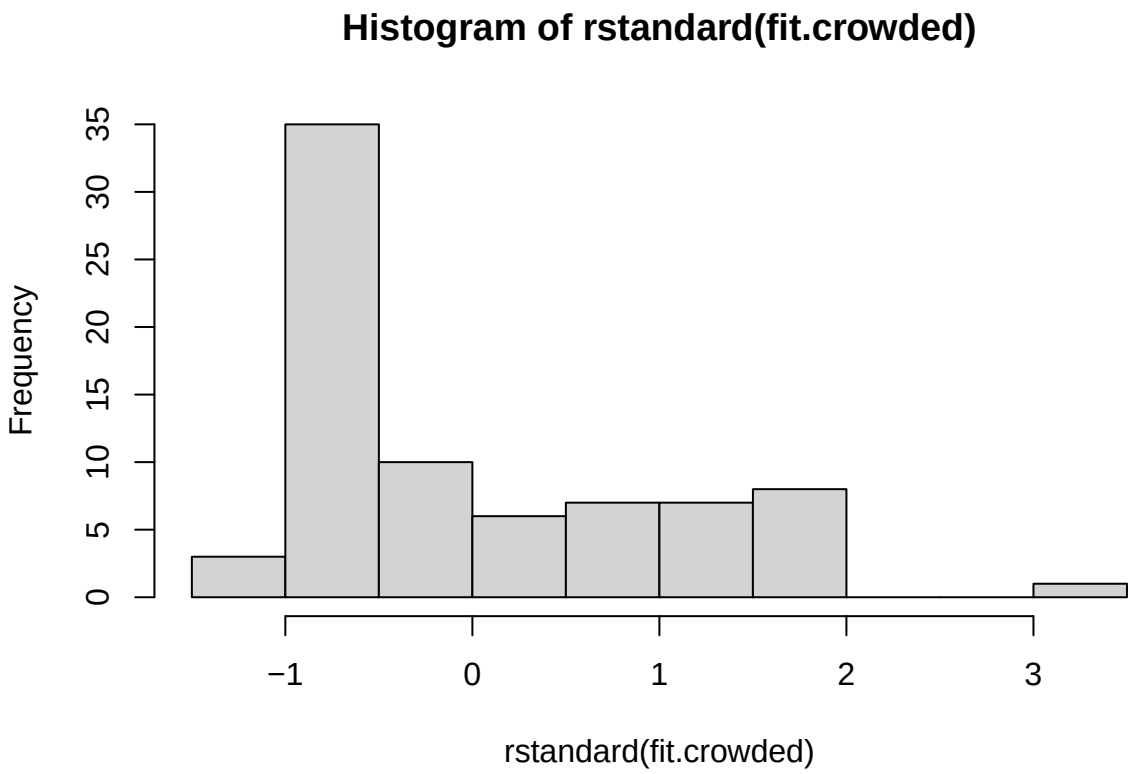
Dalje proučavamo ovisnost kriminala o prenapućenosti kućanstva

```
#prenapućenost vs kriminal
plot(poverty_crime$Crowded.Housing, poverty_crime$Assault..Homicide.)
lines(poverty_crime$Crowded.Housing, fit.crowded$fitted.values, col="red")
```



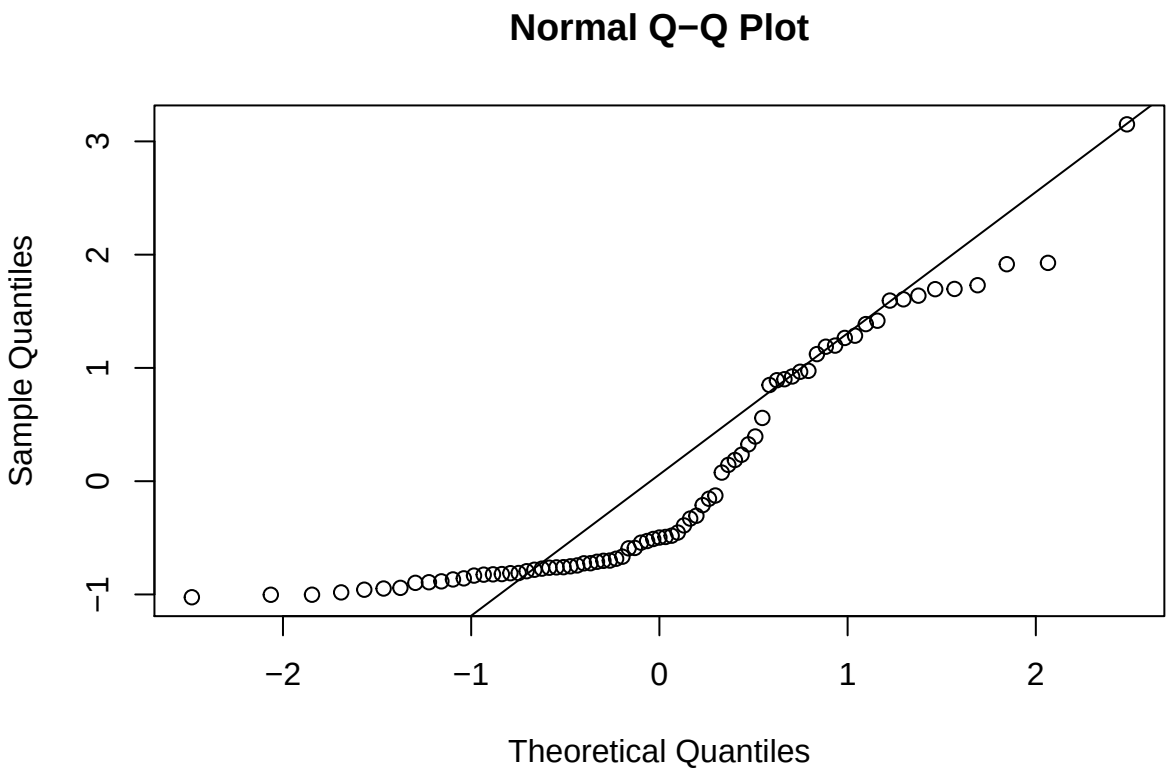
Nagib je blago rastuć, a pojedini uzorci su dosta udaljeni od pravca regresije.

```
hist(rstandard(fit.crowded))
```



Graf se dosta razlikuje od normalne razdiobe, pokazuje kršenje pretpostavke o normalnosti reziduala.

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



QQ graf također pokazuje narušavanje pretpostavke o normalnosti. Navodimo ga za usporedbu s drugim QQ grafom.

Prikazan oblik nam govori da su podaci preraspršeni za normalnu razdiobu.

Iako varijabla nije pogodna za jednostavnu linearnu regresiju, može biti značajna kod višestruke regresije u kombinaciji s drugim varijablama, pa ćemo je probati iskorititi kasnije.

Sljedeće prikazujemo ovisnost kriminala o nezaposlenosti

```
summary(fit.unemployment)
```

```
##
## Call:
## lm(formula = Assault..Homicide. ~ Unemployment, data = poverty_crime)
##
## 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
```

Koeficijent determinante je 0.664, što je vrlo uspješno za statistiku koja se bavi ljudima.

Izdvojili smo ju jer nezaposlenost vrlo direktno utječe na prosječan dohodak stanovništva u kvartu.

```
cor(poverty_crime$Unemployment, poverty_crime$Per.Capita.Income)
```

```
## [1] -0.6105529
```

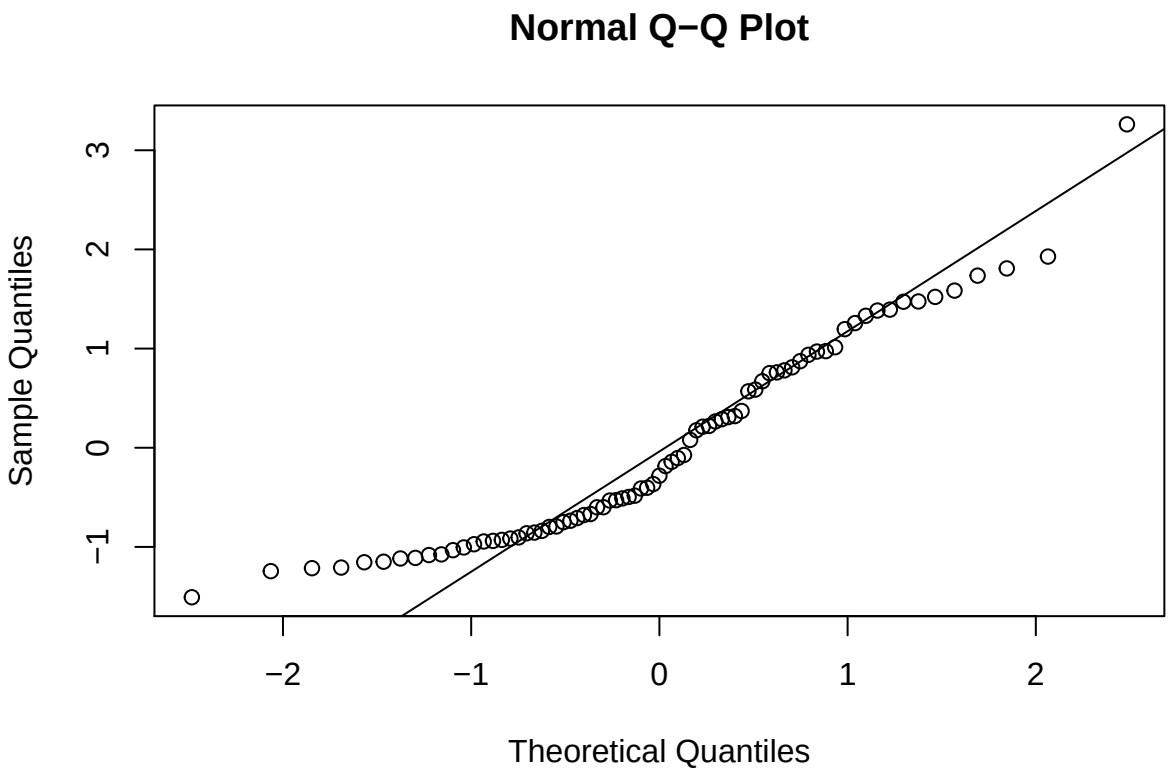
Njihova korelacije je relativno velika. No kada prikažemo ovisnost kriminala o prosječnom dohotku, dobijemo model s iznenađujuće malim koeficijentom determinacije

```
summary(fit.income)
```

```
##
## Call:
## lm(formula = Assault..Homicide. ~ Per.Capita.Income, data = poverty_crime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.087 -11.986  -3.928  10.910  45.534
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       32.8833950  3.1573233  10.415 3.19e-16 ***
## Per.Capita.Income -0.0005901  0.0001082  -5.452 6.11e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.11 on 75 degrees of freedom
## Multiple R-squared:  0.2838, Adjusted R-squared:  0.2743
## F-statistic: 29.72 on 1 and 75 DF,  p-value: 6.114e-07
```

Zbog ovoga nismo mogli napraviti dobre višestruke modele koje sadrže prosječni dohodak. Zbog velike korelacije s drugim varijablama gubili smo statističku značajnost varijabli, a koeficijent determinacije ne bi postao dovoljno veći da to opravda.

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



QQ graf ne pokazuje velika odstupanja od normalnosti,

Nećemo prikazivati ostale jednostavne linearne regresije, jer ćemo neke od njih koristiti u višestrukoj linearnoj regresiji.

##Višestruke linearne regresije

Prvo moramo testirati na korelaciju između faktora.

[1] - Siromaštvo

[2] - Prenapučenost kućanstva

[3] - Financijska ovisnost

[4] - Postotak ljudi bez SSS

[5] - Prosječni dohodak

[6] - Nezaposlenost

```
cor(cbind(poverty_crime$Below.Poverty.Level, poverty_crime$Crowded.Housing,poverty_crime$Dependency, pov
```

```
##            [,1]       [,2]       [,3]       [,4]       [,5]       [,6]
## [1,]  1.0000000  0.3232420  0.4013540  0.4223819 -0.5265178  0.7638170
## [2,]  0.3232420  1.0000000  0.2444501  0.9052740 -0.5452040  0.1443044
## [3,]  0.4013540  0.2444501  1.0000000  0.4243563 -0.7565786  0.6049994
## [4,]  0.4223819  0.9052740  0.4243563  1.0000000 -0.7073543  0.3229021
## [5,] -0.5265178 -0.5452040 -0.7565786 -0.7073543  1.0000000 -0.6105529
## [6,]  0.7638170  0.1443044  0.6049994  0.3229021 -0.6105529  1.0000000
```

U slučaju visoke koreliranosti ulaznih varijabli procjena regresijskog modela će biti nestabilna. Koristimo gornju tablicu kako bih lakše odabrali varijable za dobar model.

Radimo model s regresorima financijske neovisnosti i postotkom siromaštva.

```
fit.multi = lm(Assault..Homicide. ~ Below.Poverty.Level+Dependency, poverty_crime)
summary(fit.multi)
```

```
##
## Call:
## lm(formula = Assault..Homicide. ~ Below.Poverty.Level + Dependency,
##     data = poverty_crime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.616  -6.699  -0.352   4.654  46.766
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -27.0141     6.4458  -4.191 7.60e-05 ***
## Below.Poverty.Level   0.7494     0.1216   6.162 3.43e-08 ***
## Dependency            0.8338     0.1923   4.335 4.52e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.16 on 74 degrees of freedom
## Multiple R-squared:  0.5575, Adjusted R-squared:  0.5455
## F-statistic: 46.61 on 2 and 74 DF,  p-value: 7.938e-14
```

P-vrijednosti t-testova pojedinih varijabli nam pokazuju da su sve varijable statistički značajne.

Ako varijable među sobom imaju veliku korelaciju, to može narušiti statističku značajnost koje bi one pojedinačno imale.

```
cor(poverty_crime$Below.Poverty.Level, poverty_crime$Dependency)
```

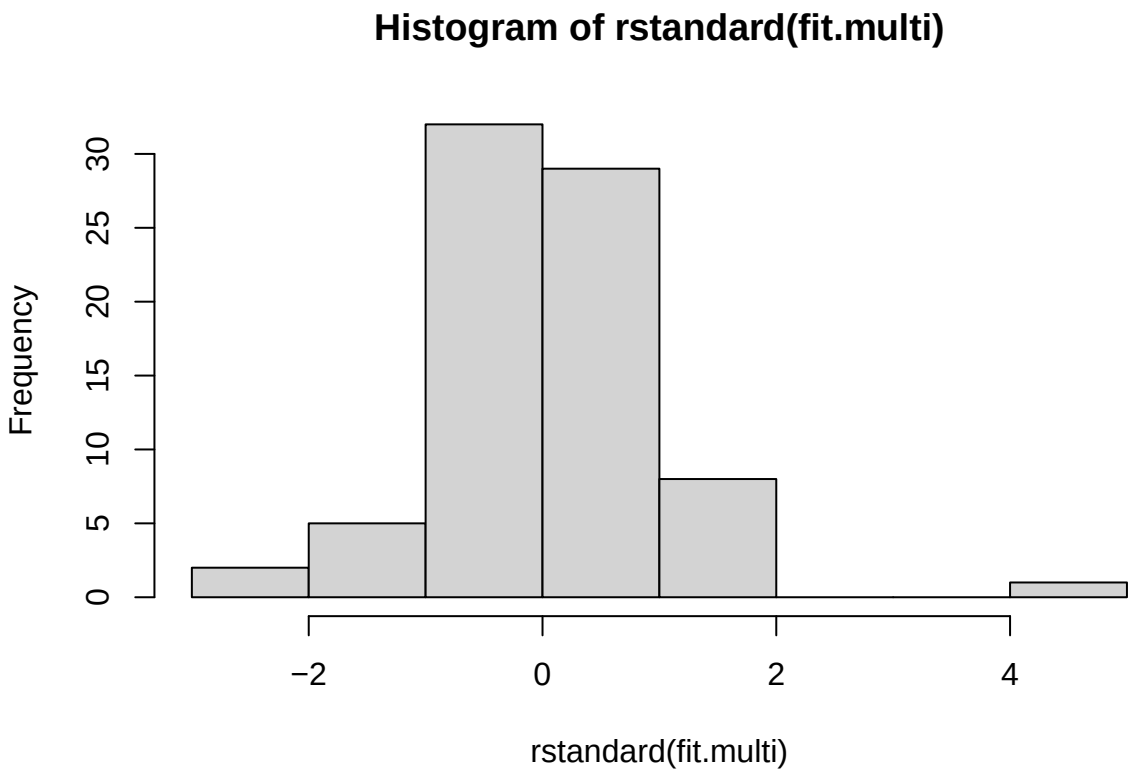
```
## [1] 0.401354
```

Razlika koeficijent determinacije i prilagođenog koeficijenta determinacije nije velika. Zaključujemo da su obje varijable pridonijele modelu te zato nisu loš izbor unatoč njihovoj korelaciji od 0.401354.

Dobiveni model višestruke regresije uspješno opisuje 55.75% varijacije u podacima.

Kako bi prihvatili model potrebno je provjeriti normalnost reziduala.

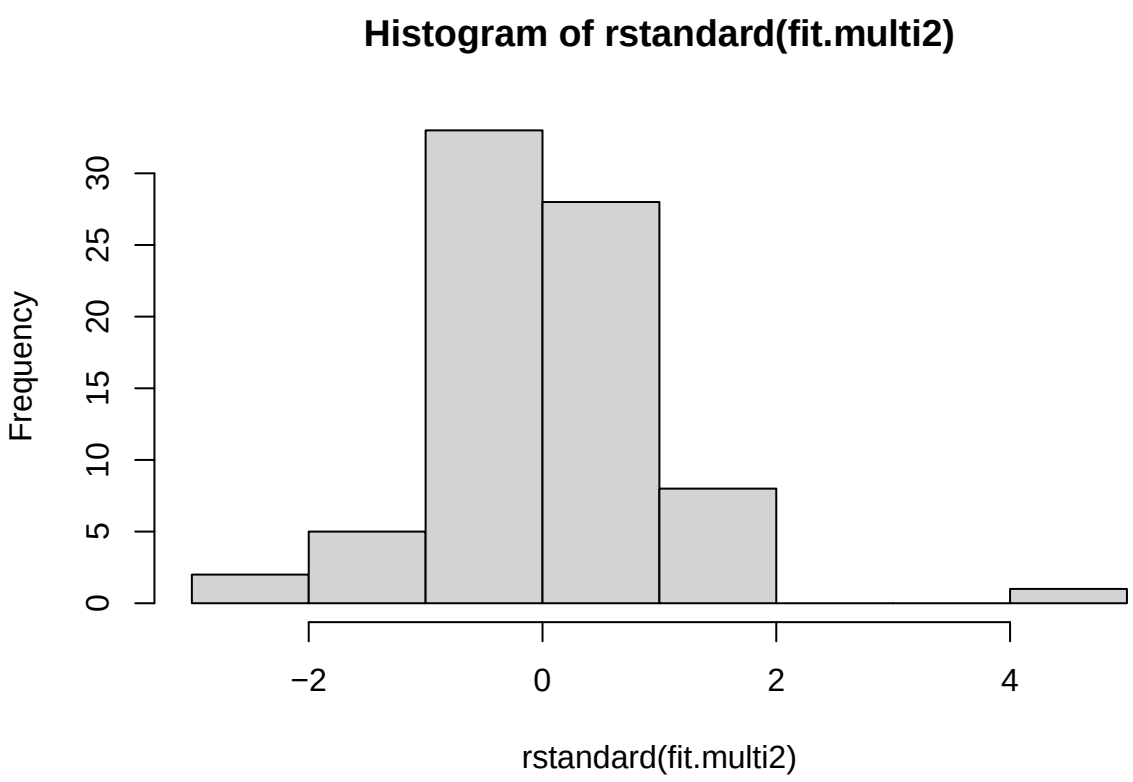
```
hist(rstandard(fit.multi))
```



Grafički prikaz pokazuje da normalnost reziduala je očuvana, uzimajući u obzir robustnost te pretpostavke.

Dodajemo prosječni dohodak u model.

```
fit.multi2 = lm(Assault..Homicide. ~ Below.Poverty.Level+Dependency+Per.Capita.Income, poverty_crime)
hist(rstandard(fit.multi2))
```



Graf razdiobe reziduala je dovoljno sličan normalnom.

```
summary(fit.multi2)
```

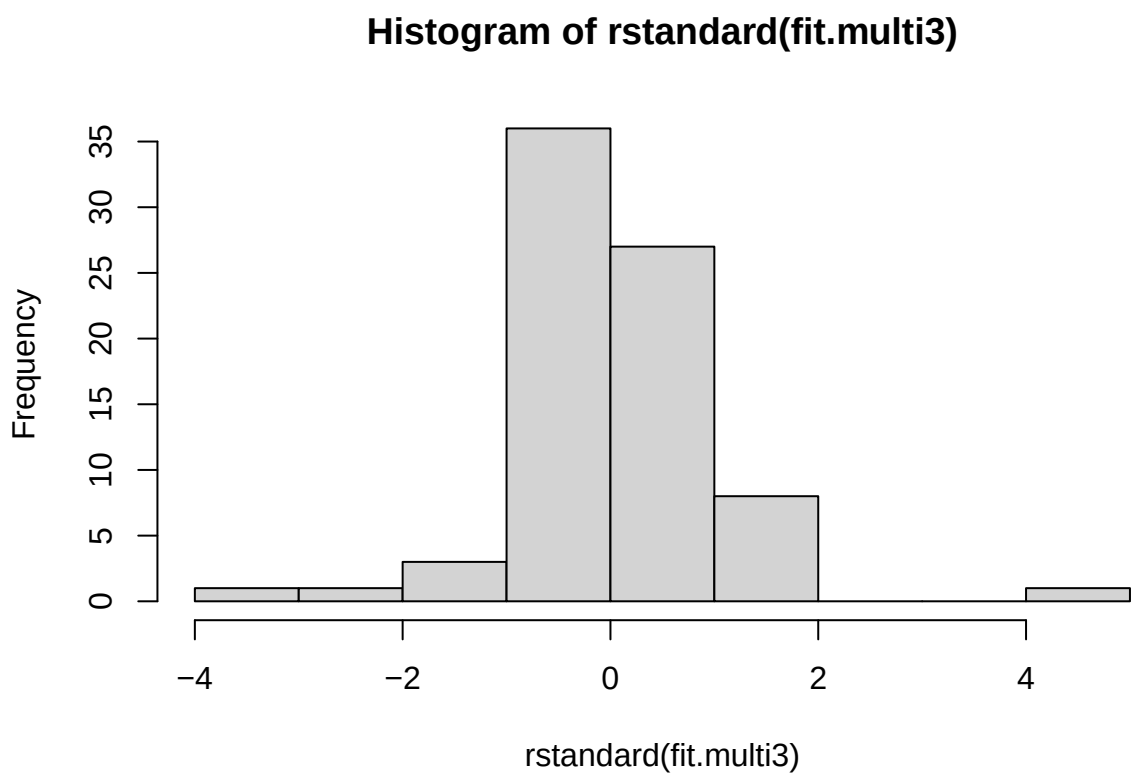
```
##
## Call:
## lm(formula = Assault..Homicide. ~ Below.Poverty.Level + Dependency +
##     Per.Capita.Income, data = poverty_crime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.614  -6.994  -0.333   4.892  47.009
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -3.137e+01  1.308e+01  -2.398   0.0190 *
## Below.Poverty.Level  7.682e-01  1.318e-01   5.829 1.41e-07 ***
## Dependency           9.065e-01  2.710e-01   3.345   0.0013 **
## Per.Capita.Income    5.438e-05  1.419e-04   0.383   0.7028
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.23 on 73 degrees of freedom
## Multiple R-squared:  0.5584, Adjusted R-squared:  0.5402
## F-statistic: 30.76 on 3 and 73 DF,  p-value: 5.766e-13
```

Model se nije puno poboljšao dodavanjem varijable dohotka. Sam dohodak je statistički neznačajan, te radi

jednostavnosti, bolje je odbaciti ga.

Nadograđujemo prethodni model sa regresorom prenapučenosti kućanstva

```
fit.multi3 = lm(Assault..Homicide. ~ Below.Poverty.Level + Dependency + Crowded.Housing, poverty_crime)
hist(rstandard(fit.multi3))
```



Graf razdiobe reziduala je dovoljno sličan normalnoj.

```
summary(fit.multi3)
```

```
##
## Call:
## lm(formula = Assault..Homicide. ~ Below.Poverty.Level + Dependency +
##     Crowded.Housing, data = poverty_crime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.103  -5.597  -0.347   3.932  46.849
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -26.3090     6.1823  -4.256 6.11e-05 ***
## Below.Poverty.Level   0.8335     0.1205   6.918 1.48e-09 ***
## Dependency            0.9016     0.1860   4.849 6.82e-06 ***
## Crowded.Housing      -0.9850     0.3577  -2.753  0.00744 **
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.7 on 73 degrees of freedom
## Multiple R-squared:  0.5991, Adjusted R-squared:  0.5826
## F-statistic: 36.36 on 3 and 73 DF,  p-value: 1.742e-14
```

Koeficijent determinacije je bolji do prethodnog za 0.04. Sve prethodne varijable su ostale vrlo značajne, a varijabla napučenosti kućanstva je isto statistički značajna, ali manje u odnosu na ostale varijable.

Ulazne varijable ovog modela su financijska ovisnost, postotak ljudi bez SSS i postotak ljudi ispod linije siromaštva. Prilikom odabira modela pazili smo na korelacije varijabli, jer visoke korelacije mogu uzrokovati probleme u interpretaciji regresijskih rezultata.

Prikaz korelacije varijabli

```
##           [,1]      [,2]      [,3]
## [1,] 1.0000000 0.4013540 0.4223819
## [2,] 0.4013540 1.0000000 0.4243563
## [3,] 0.4223819 0.4243563 1.0000000
```

Kvadriranje postotka ljudi ispod linije siromaštva je inspirirano začaranim krugovima koje siromaštvo stvara. Siromaštvo samo sebe pogoni pomoću kredita sa većim kamatama i lošijim financijskim odlukama koje su siromašni ljudi primorani napraviti.

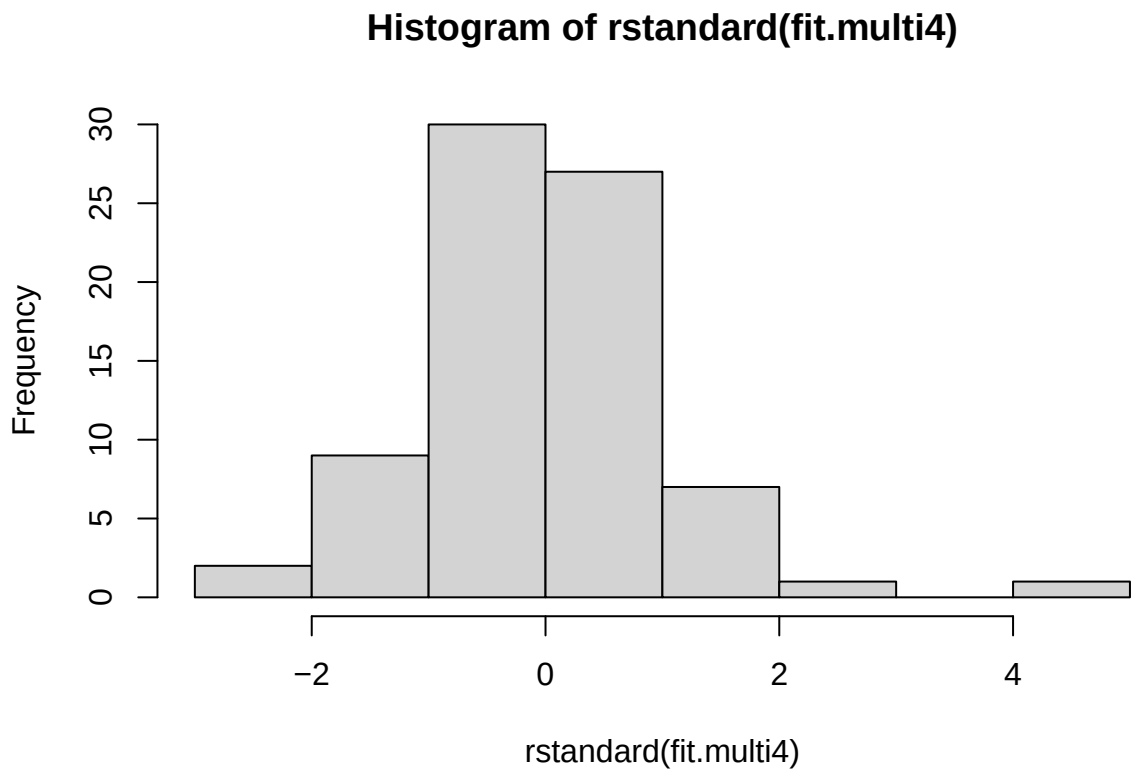
Sljedeći model prima ulaze: Below.Poverty.Level, (Below.Poverty.Level)^2, Dependency, No.High.School.Diploma

```
fit.multi4 = lm(Assault..Homicide. ~ Below.Poverty.Level + I(Below.Poverty.Level^2) + Dependency + No.H
summary(fit.multi4)
```

```
##
## Call:
## lm(formula = Assault..Homicide. ~ Below.Poverty.Level + I(Below.Poverty.Level^2) +
##     Dependency + No.High.School.Diploma, data = poverty_crime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.930  -5.032  -0.561   4.450  39.861
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -43.354125   6.553360  -6.616 5.61e-09 ***
## Below.Poverty.Level        2.206932   0.333533   6.617 5.58e-09 ***
## I(Below.Poverty.Level^2)  -0.024599   0.005764  -4.268 5.92e-05 ***
## Dependency                 1.128955   0.174518   6.469 1.04e-08 ***
## No.High.School.Diploma    -0.484924   0.107457  -4.513 2.44e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.524 on 72 degrees of freedom
## Multiple R-squared:  0.6867, Adjusted R-squared:  0.6693
## F-statistic: 39.45 on 4 and 72 DF,  p-value: < 2.2e-16
```

Sve ulazne varijable su statistički značajne. Koeficijent korelacije je 0.6867.

```
hist(rstandard(fit.multi4))
```



Pretpostavka o linearnosti reziduala je prihvatljiva.

Zaključak

Konačan model sadržava relevantne varijable koje objašnjavaju čak preko 68.67% varijance broja napada i ubojstva. Osim varijabli Dependency, Below.Poverty.Level i No.High.School.Diploma, uključen je i kvadrat Below.Poverty.Level-a (zbog nelinearnog efekta).

Sve navedene varijable su značajne na razini 0.01, kao i sam model, na što upućuju rezultati t-testova pojedinih koeficijenata i F-testa čitavog modela. Za usporedbu, jednostavni linearni regresijski model s regresorskom varijablom Per.Capita.Income ima koeficijent korelacije R=0.664 Treba procijeniti je li vrijedna zamjena jednostavnosti za tako sitno poboljšanje modela