SAP - projekt

Utjecaj preventivne zdravstvene zaštite na zdravlje

Opis skupa podataka:

Skup podataka kojeg ćemo analizirati, ukupno 16000 podataka, sadrži podatke o metodama preventivne zdravstvene zaštite i ispitivanim bolestima među stanovništvom 500 gradova SAD-a.

Za svaki grad dani su podatci o udjelu stanovništva toga grada koje provodi određene metode zdravstvene zaštite i podatci o udjelu stanovištva toga grada koji pati od neke promatranih bolesti/zdravstvenih stanja. U istraživanju su promatrane četiri metode zdravstvene zaštite i 12 bolesti/zdravstvenih stanja.

Podatci su dani u 10 stupaca koji redom predstavljaju: redni broj podatka, ime savezne države u kojoj se nalazi grad, ime grada, kategoriju ("Prevention" ili "Health Outcomes"), opis mjere, jedinica u kojoj su dani podatci (% za sve podatke), tip podatka ("AgeAdjPrv" ili "CrdPrv"), rezutat mjerenja (udio stanovništva koje provodi neku preventivnu mjeru ili ima neku bolest/zdravstveno stanje), broj stanovnika pojednig grada i kratki opis mjerenja.

```
data = read.csv("dataset.csv")

dim(data)
```

```
## [1] 16000    10
```

Primijetili smo da u skupu podataka 'data' za svaki grad imamo dva zapisa koji se razlikuju samo u podatku koji je zapisan u stupcu 'DataValueTypeID':"AgeAdjPrv" ili "CrdPrv". Ako je podatak tipa "AgeAdjPrv" to znači da rezultat, tj. podatak zapisan u stupcu 'Data_Value', predstavlja udio stanovništva koje provodi neku preventivnu mjeru ili imaja bolest/zdravstveno stanje u odnosu na stanovništvo promatranog grada koje odgovara dobnoj skupini nad kojom je provedeno ispitivanje. Ako je podatak tipa "CrdPrv" to znači da rezultat predstavlja udio stanovništva koje provodi neku preventivnu mjeru ili ima bolest/zdravstveno stanje u odnosu na cjelokupno stanovništvo toga grada.

Odlučili smo za početak koristiti samo podatke čiji je tip "AgeAdjPrv", jer mislimo da ćemo time izbjeći pristranost. Naime, dobna struktura svih gradova nije ista i neki gradovi mogu imati pretežno starije stanovništvo koje je podložnije bolestima u odnosu na druge gradove.

U skup podataka 'ageAdjData' izlučit ćemo samo podatke čiji je tip "AgeAdjPrv".

```
ageAdjData = data[data["DataValueTypeID"] == "AgeAdjPrv",]

dim(ageAdjData)
```

```
## [1] 8000   10
```

Sada u skupu podataka 'ageAdjData' za svaki grad i za svaku mjeru imamo točno jedan podatak, ukupno 8000 podataka.

Postoje dvije kategorije mjerenja (stupac 'Categories'): "Health outcomes" i "Prevention". Kategorija označuje sadrži li podatak informaciju o udjelu stanovništva nekog grada koji poduzimaju neku vrstu preventivne zaštite ili informaciju o udjelu stanovništva koji boluju od neke bolest/zdravstvenog stanja. Stupac 'Short_Question_Text' predstavlja kratki opis mjere i upravo ćemo zbog jasnoće i sažetosti koristiti taj skraćeni opis, umjesto cijelog naziva mjere.

```
catPrevention = ageAdjData[ageAdjData$Category == "Prevention",]

catHealtOutcomes = ageAdjData[ageAdjData$Category == "Health Outcomes",]
```

Tablica ispod sadrži sve mjere i njihove kraće opise. Vidimo da za svaku mjeru postoji točno 500 podataka, tj. po jedan podatak za svaki grad.

```
library(dplyr)
```

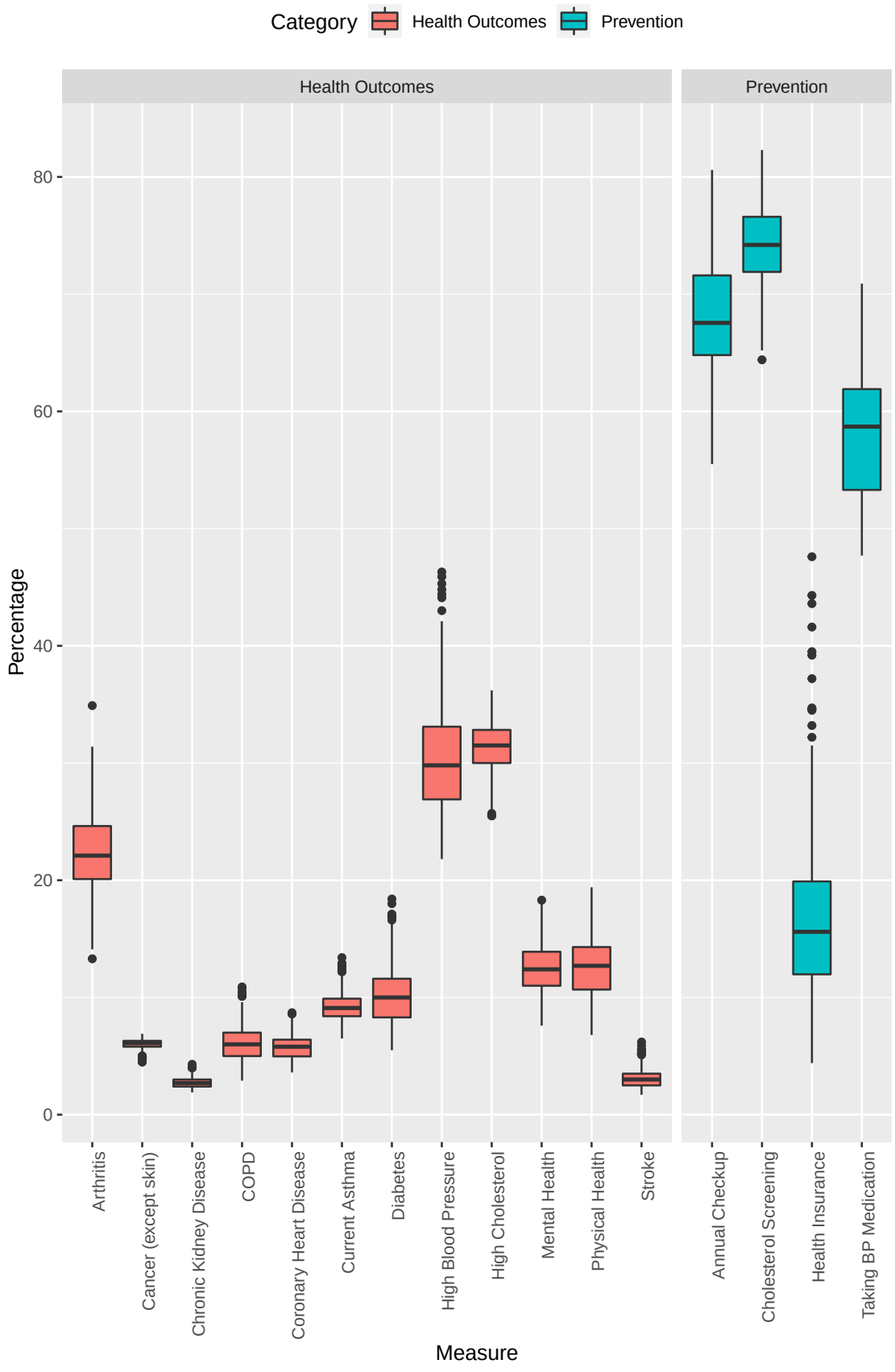
```
##
## Attaching package: 'dplyr'

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

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

```
ageAdjData %>% group_by(Category, Measure, Short_Question_Text) %>% summarise(count = n())
```

```
## `summarise()` regrouping output by 'Category', 'Measure' (override with `.groups` argument)

## # A tibble: 16 x 4
## # Groups:   Category, Measure [16]
##    Category     Measure                                  Short_Question_Te~ count
##    <chr>        <chr>                                    <chr>              <int>
##  1 Health Outc~ Arthritis among adults aged >=18 Years   Arthritis            500
##  2 Health Outc~ Cancer (excluding skin cancer) among a~ Cancer (except sk~   500
##  3 Health Outc~ Chronic kidney disease among adults ag~ Chronic Kidney Di~   500
##  4 Health Outc~ Chronic obstructive pulmonary disease ~ COPD                500
##  5 Health Outc~ Coronary heart disease among adults ag~ Coronary Heart Di~   500
##  6 Health Outc~ Current asthma among adults aged >=18 ~ Current Asthma       500
##  7 Health Outc~ Diagnosed diabetes among adults aged >~ Diabetes             500
##  8 Health Outc~ High blood pressure among adults aged ~ High Blood Pressu~   500
##  9 Health Outc~ High cholesterol among adults aged >=1~ High Cholesterol     500
## 10 Health Outc~ Mental health not good for >=14 days a~ Mental Health        500
## 11 Health Outc~ Physical health not good for >=14 days~ Physical Health      500
## 12 Health Outc~ Stroke among adults aged >=18 Years     Stroke               500
## 13 Prevention   Cholesterol screening among adults age~ Cholesterol Scree~   500
## 14 Prevention   Current lack of health insurance among~ Health Insurance     500
## 15 Prevention   Taking medicine for high blood pressur~ Taking BP Medicat~   500
## 16 Prevention   Visits to doctor for routine checkup w~ Annual Checkup       500
```

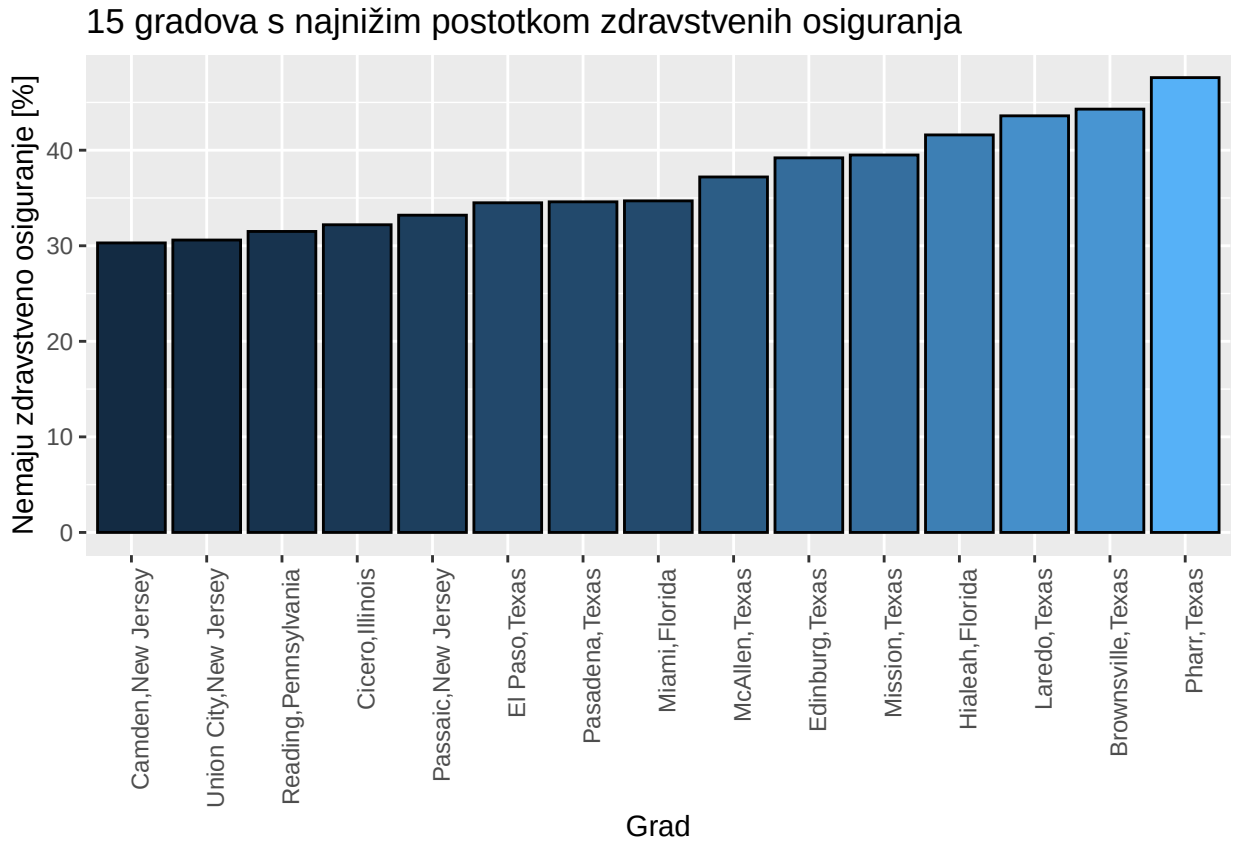
U nastavku su prikazani pravokutni dijagrami (box plot) za svaku mjeru zaštite i bolest te su grupirani prema kategoriji kojoj podatci pripadaju. Svaki pravokutni dijagram nastao je iz ukupno 500 podataka, jer je istraživanje provedeno u 500 gradova. Pravokutni dijagram kombinira prikaz medijana, kvartila podataka, te najmanje i najveće vrijednosti. Prikaz pravokutnog dijagrama bitan je jer iz njega možemo iščitati stršeće vrijednosti. Primjećujemo da za većinu kategorija postoje stršeće vrijednosti, a u nastavku ćemo prokomentirati one kategorije kod kojih su one najizraženije.

```
library(ggplot2)

ggplot(ageAdjData, aes(x = Short_Question_Text, y = Data_Value, fill = Category)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1), legend.position = 'top') +
  xlab('Measure') + ylab('Percentage') +
  facet_grid(. ~ Category, scales = "free", space='free')
```


Category
Health Outcomes
Prevention
Health Outcomes
Prevention
Percentage
0
20
40
60
80
Arthritis
Cancer (except skin)
Chronic Kidney Disease
COPD
Coronary Heart Disease
Current Asthma
Diabetes
High Blood Pressure
High Cholesterol
Mental Health
Physical Health
Stroke
Annual Checkup
Cholesterol Screening
Health Insurance
Taking BP Medication
Measure

Primijetili smo da pravokutni dijagram za mjeru prevencije 'Health Insurance' ima najviše stršećih vrijednosti. U sljedećem dijagramu prikazano je 15 gradova koji imaju najveći postotak stanovništva koji nemaju zdravstveno osiguranje. Upravo su te vrijednosti stršeće vrijednosti pravokutnog dijagrama.

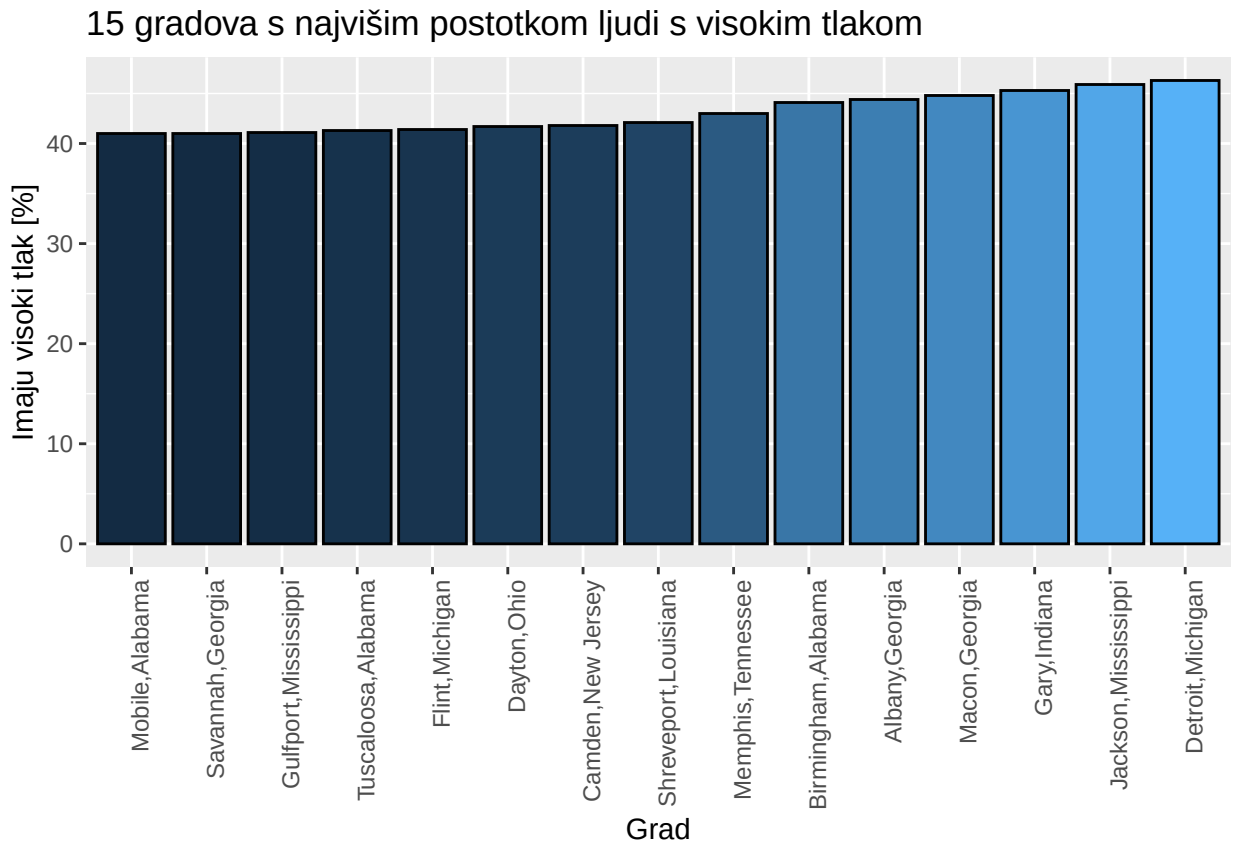
```
library(ggplot2)
data.health.insurance <- subset(ageAdjData, Short_Question_Text == 'Health Insurance')
data.health.insurance <- data.health.insurance[order(data.health.insurance$Data_Value,
                                                     decreasing = TRUE),]
data.health.insurance <- data.health.insurance[1:15,]

data.health.insurance$CityName <- paste0(data.health.insurance$CityName, ',',
                                         data.health.insurance$StateDesc)
ggplot(data.health.insurance, aes(x = reorder(CityName, Data_Value),
                                  y = Data_Value, fill = Data_Value)) +
  geom_bar(stat = "identity",col = "black")+
  theme(axis.text.x = element_text(angle = 90, hjust = 1), legend.position = 'none') +
  ggtitle('15 gradova s najnižim postotkom zdravstvenih osiguranja') +
  xlab('Grad') + ylab('Nemaju zdravstveno osiguranje [%]')
```



Istu stvar smo napravili i za bolest 'High Blood Pressure'. Primjećujemo da je nagib dijagrama za mjeru 'Health Insurance' puno strmiji od onog za bolest 'High Blood Pressure', zato što se u prvom slučaju stršeće vrijednosti nalaze u puno većem intervalu nego u drugom.

```
library(ggplot2)
data.high.blood.pressure <- subset(ageAdjData, Short_Question_Text == 'High Blood Pressure')
data.high.blood.pressure <- data.high.blood.pressure[order(data.high.blood.pressure$Data_Value,
```

```
                                                    decreasing = TRUE),]
data.high.blood.pressure <- data.high.blood.pressure[1:15,]

data.high.blood.pressure$CityName <- paste0(data.high.blood.pressure$CityName, ',',
                                            data.high.blood.pressure$StateDesc)
ggplot(data.high.blood.pressure, aes(x = reorder(CityName, Data_Value), y = Data_Value,
                                     fill = Data_Value)) +
  geom_bar(stat = "identity",col = "black")+
  theme(axis.text.x = element_text(angle = 90, hjust = 1), legend.position = 'none') +
  ggtitle('15 gradova s najvišim postotkom ljudi s visokim tlakom') +
  xlab('Grad') + ylab('Imaju visoki tlak [%]')
```



1) Postoji li neka metoda preventivne zaštite koja je "popularnija" u saveznoj državi Ohio nego u saveznoj državi Florida (odnosno, za koju je udio stanovnika savezne države Ohio veći od udjela stanovnika savezne države Florida)?

Da bismo odgovorili na ovo pitanje, prvo ćemo provesti manipulacije nad skupom podataka kako bismo skupili sve informacije koje su nam potrebne i sistematski ih organizirati. Zatim ćemo podatke prikazati stupičastim dijagramom, usporediti popularnost svake metode preventivne zaštite u državama Ohio i Florida i naslutiti mogući odgovor na postavljeno pitanje. Za kraj ćemo provesti test o dvije proporcije za svaku metodu preventivne zaštite i donijeti zaključke.

Podatci u skopu podataka kojeg koristimo dani su za gradove, a za potrebe ovog zadatka trebamo ih poopćiti na savezne države. To ćemo učiniti tako da ćemo prvo izračunati ukupan broj ispitanika u nekoj saveznoj državi zbrajanjem broja ispitanika iz svih gradova u kojima je provedeno istraživanje za svaku državu. Ti su podatci prikazani u tablici populationData. Stupac PopulationCount je broj stanovnika toga grada, StatePopulationCount broj koji predstavlja zbroj stanovika svih gradova te države u kojima je provedeno istraživanje. Taj broj smatramo za naš skup podataka aproksimacijom broja stanovnika neke države. Primjećujemo da su s barem jednim gradom zastupljene sve savezne države (svih 50 i District of Columbia).

```
cityPopulationData <- ageAdjData %>% group_by(StateDesc, CityName, PopulationCount) %>%
  distinct(StateDesc, CityName, PopulationCount)


statePopulationData <- cityPopulationData %>% group_by(StateDesc) %>%
  mutate(CityCount = n()) %>% group_by(StateDesc, CityCount) %>%
  summarise_at(vars(PopulationCount), list(StatePopulationCount = sum))


populationData <- merge(cityPopulationData, statePopulationData, by="StateDesc")
```

U sljedećoj tablici, stateData, su po svim saveznim državama za svaku mjeru (bolesti ili mjeru prevencije) prikazani podatci o broju stanovnika u državi (StatePopulationCount), broj stanovnika u toj državi koji odgovara nekoj mjeri, tj. koji ima određenu bolest ili prakticira mjeru prevencije (Data_Value_Population_Count) i odgovarajući postotak (Data_Value) u odnosu na ukupan broj ispitanika u državi.

```
stateData = ageAdjData %>% mutate(Data_Value_Population_Count =
                                    round(Data_Value*0.01*PopulationCount, digits = 0))

stateData = merge(stateData, populationData)

stateData = stateData %>%
  group_by(StateDesc, Short_Question_Text, Category, StatePopulationCount, Data_Value_Unit) %>%
  summarise_at(vars(Data_Value_Population_Count), list(Data_Value_Population_Count = sum))

stateData = stateData %>% mutate(Data_Value =
                                 round((Data_Value_Population_Count/StatePopulationCount)*100,
                                         digits = 1))

stateData = stateData[c(1, 2, 3, 4, 6, 5, 7)]
stateData
```

```
## # A tibble: 816 x 7
## # Groups:   StateDesc, Short_Question_Text, Category, StatePopulationCount
## #   [816]
##    StateDesc Short_Question_~ Category StatePopulation~ Data_Value_Popu~
##    <chr>     <chr>            <chr>               <dbl>            <dbl>
##  1 Alabama   Annual Checkup   Prevent~           965304           716538
##  2 Alabama   Arthritis        Health ~           965304           280786
##  3 Alabama   Cancer (except ~ Health ~           965304            59222
##  4 Alabama   Cholesterol Scr~ Prevent~           965304           741743
##  5 Alabama   Chronic Kidney ~ Health ~           965304            29327
##  6 Alabama   COPD             Health ~           965304            77282
##  7 Alabama   Coronary Heart ~ Health ~           965304            65678
```

```
##  8 Alabama   Current Asthma   Health ~           965304           102166
##  9 Alabama   Diabetes         Health ~           965304           127891
## 10 Alabama   Health Insurance Prevent~           965304           162337
## # ... with 806 more rows, and 2 more variables: Data_Value_Unit <chr>,
## #   Data_Value <dbl>
```

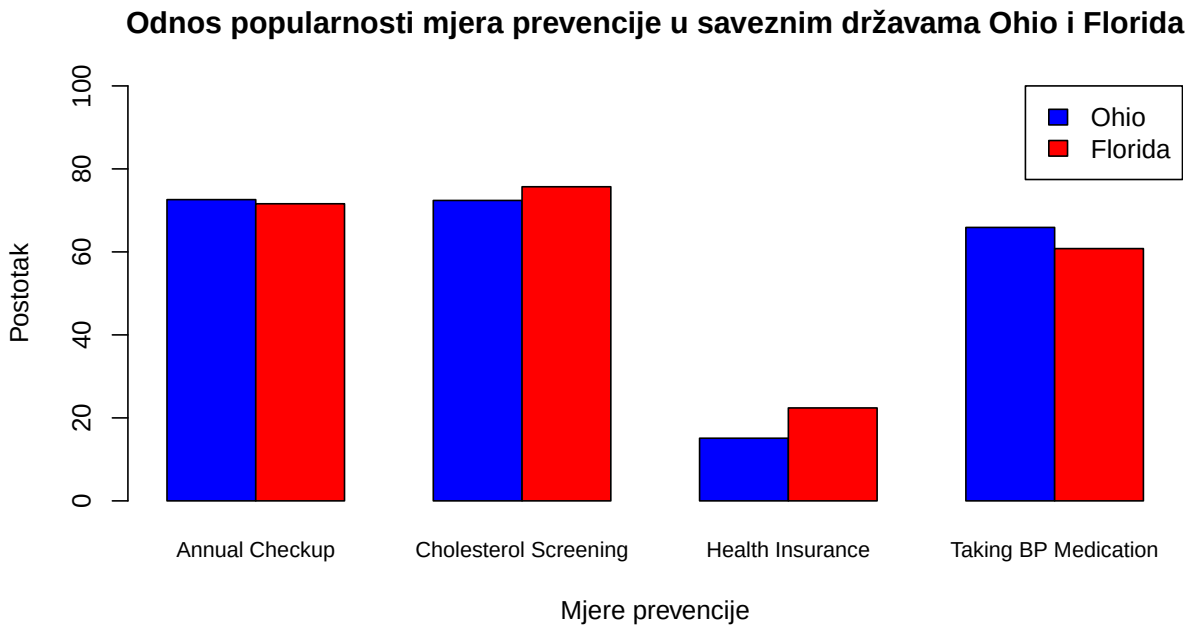
U sljedećoj tablici, o_f_prevention, su iz prethodne tablice filtrirani podatci za države Ohio i Florida i mjere prevencije.

```
o_f_prevention = stateData %>%
  filter(Category == "Prevention" & (StateDesc =="Ohio" | StateDesc =="Florida")) %>%
  arrange(desc(StateDesc))  %>% arrange(Short_Question_Text)
o_f_prevention
```

```
## # A tibble: 8 x 7
## # Groups:   StateDesc, Short_Question_Text, Category, StatePopulationCount [8]
##   StateDesc Short_Question_~ Category StatePopulation~ Data_Value_Popu~
##   <chr>     <chr>            <chr>               <dbl>            <dbl>
## 1 Ohio      Annual Checkup   Prevent~          2330226          1691026
## 2 Florida   Annual Checkup   Prevent~          5166487          3697308
## 3 Ohio      Cholesterol Scr~ Prevent~          2330226          1687122
## 4 Florida   Cholesterol Scr~ Prevent~          5166487          3913034
## 5 Ohio      Health Insurance Prevent~          2330226           351939
## 6 Florida   Health Insurance Prevent~          5166487          1157292
## 7 Ohio      Taking BP Medic~ Prevent~          2330226          1535235
## 8 Florida   Taking BP Medic~ Prevent~          5166487          3143467
## # ... with 2 more variables: Data_Value_Unit <chr>, Data_Value <dbl>
```

Na sljedećem dijagramu vidimo odnos popularnosti mjera prevencije za države Ohio i Florida. Iz dijagrama naslućujemo da su mjere prevencije "Anual Checkup" i "Taking BP Medication" popularnije u državi Ohio nego u Floridi. Također, treba obratiti pozornost na značenje mjere "Health Insurance", a pogledamo li njezin puni naziv vidjet ćemo da ona označava ispitanike koji nemaju zdravstveno osiguranje. Prema tome, za mjeru "Health Insurance" možemo naslutiti da u saveznoj državi Ohio preventivna mjera zdravstvenog osiguranja (u smislu ljudi koji imaju zdravstveno osiguranje) popularnija nego u Floridi.

```
ohio <- (o_f_prevention %>% filter(StateDesc == "Ohio"))$Data_Value
florida <- (o_f_prevention %>% filter(StateDesc == "Florida"))$Data_Value


states <- t(cbind(ohio, florida))


barplot(states, beside=TRUE, col=c("blue", "red"),xlab = "Mjere prevencije",
        ylab = "Postotak",cex.names=0.8,
        names.arg = unique(o_f_prevention$Short_Question_Text), ylim=c(0,100),
        main="Odnos popularnosti mjera prevencije u saveznim državama Ohio i Florida")


legend("topright",c("Ohio","Florida"),fill = c("blue", "red"))
```



Provedimo sada test o dvije proporcije za svaku mjeru prevencije.

Test o dvije proporcije u programskom paketu R implementiran je u funkciji `prop.test()`. Pretpostavka z-testa je da je n dovoljno velik, što u našem slučaju je.

Testovi za svaku mjeru prevencije će biti postavljeni kao: H0: p(ohio) = p(florida) H1: p(ohio) > p(florida) s 95%-tnim intervalom pouzdanosti. Osim u slučaju mjere prevencije "Health Insurance", tada će alternativna hipoteza biti H1: p(ohio) < p(florida), jer ta mjera obilježava ispitanike koji nemaju zdravstveno osiguranje.

Anual Checkup:

U donjem ispisu vidimo da je p-vrijednost manja od razine signifikantnosti 0.05 pa odbacujemo hipotezu H0 i zaključujemo da je mjera prevencije "Anual Checkup" popularnija u saveznoj državi Ohio nego u saveznoj državi Florida.

```
o_f_anual_checkup = o_f_prevention %>% filter(Short_Question_Text == "Annual Checkup")

prop.test(x = o_f_anual_checkup$Data_Value_Population_Count,
          n = o_f_anual_checkup$StatePopulationCount, alternative = "greater",
          conf.level = 0.95, correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  o_f_anual_checkup$Data_Value_Population_Count out of o_f_anual_checkup$StatePopulationCount
## X-squared = 803.84, df = 1, p-value < 2.2e-16
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.009477843 1.000000000
## sample estimates:
```

```
##    prop 1    prop 2
## 0.7256918 0.7156329
```

Cholesterol Screening:

U donjem ispisu vidimo da je p-vrijednost veća od razine signifikantnosti 0.05 pa ne možemo odbaciti hipotezu H0.

```
o_f_cholesterol_screening = o_f_prevention %>%
  filter(Short_Question_Text == "Cholesterol Screening")
prop.test(x = o_f_cholesterol_screening$Data_Value_Population_Count,
          n = o_f_cholesterol_screening$StatePopulationCount, alternative = "greater",
          conf.level = 0.95, correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  o_f_cholesterol_screening$Data_Value_Population_Count out of o_f_cholesterol_screening$StateP
## X-squared = 9463.3, df = 1, p-value = 1
## alternative hypothesis: greater
## 95 percent confidence interval:
##  -0.0339442  1.0000000
## sample estimates:
##    prop 1    prop 2
## 0.7240165 0.7573878
```

Health Insurance:

U donjem ispisu vidimo da je p-vrijednost manja od razine signifikantnosti 0.05 pa odbacujemo hipotezu H0 i zaključujemo da je mjera prevencije "Health Insurance" popularnija (u smislu popularnije je imati zdravstveno osiguranje) u saveznoj državi Ohio nego u saveznoj državi Florida.

```
o_f_health_insurence = o_f_prevention %>%
  filter(Short_Question_Text == "Health Insurance")
prop.test(x = o_f_health_insurence$Data_Value_Population_Count,
          n = o_f_health_insurence$StatePopulationCount, alternative = "less",
          conf.level = 0.95, correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  o_f_health_insurence$Data_Value_Population_Count out of o_f_health_insurence$StatePopulationC
## X-squared = 53177, df = 1, p-value < 2.2e-16
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.00000000 -0.07247786
## sample estimates:
##    prop 1    prop 2
## 0.1510321 0.2239998
```

Taking BP Medication

U donjem ispisu vidimo da je p-vrijednost manja od razine signifikantnosti 0.05 pa odbacujemo hipotezu H0 i zaključujemo da je mjera prevencije "Taking BP Medication" popularnija u saveznoj državi Ohio nego u saveznoj državi Florida.

```
o_f_taking_bpmedication_ = o_f_prevention %>%
  filter(Short_Question_Text == "Taking BP Medication")
prop.test(x = o_f_taking_bpmedication_$Data_Value_Population_Count,
          n = o_f_taking_bpmedication_$StatePopulationCount, alternative = "greater",
          conf.level = 0.95, correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  o_f_taking_bpmedication_$Data_Value_Population_Count out of o_f_taking_bpmedication_$StatePopu
## X-squared = 17389, df = 1, p-value < 2.2e-16
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.04978004 1.00000000
## sample estimates:
##    prop 1    prop 2
## 0.6588352 0.6084341
```

Zaključujemo da postoji više metoda preventivne zaštite koja su "popularnije" u saveznoj državi Ohio nego u saveznoj državi Florida, i to su: Anual Checkup, Cholesterol Screening, Health Insurance i Taking BP Medication.

2) Sami proizvoljno izaberite neke tri savezne države koje su vam možda najzanimljivije. Je li postotak stanovništa koji boluje od kroničnih plućnih bolesti jednak za sve tri države?

Početni cilj nam je odrediti postotak stanovništva koji boluje od kroničnih plućnih bolesti u saveznim državama.

Radimo novu tablicu plućne u kojoj se nalaze samo plućne bolesti i u novu varijablu Apsbrojoboljelih izračunavamo broj oboljelih od određene plućne bolesti u pojedinom gradu.

```
plucne = catHealtOutcomes[catHealtOutcomes$Short_Question_Text == c("COPD"),]
plucne$ApsbrojOboljelih = round(plucne$Data_Value * plucne$PopulationCount * 0.01,
                                digits = 0)
```

Nakon što smo izračunali broj oboljelih i ukupan broj stanovništva saveznih država ubacujemo te podatke u novu tablicu brploboljelih i dijeljenjem ta dva podatka dobivamo postotak kronično plućno oboljelih u saveznim državama. Postotak smo poredali od većeg prema manjem.

```
brploboljelih = plucne %>% group_by(StateDesc) %>%
  summarise(sumbrploboljelih = sum(ApsbrojOboljelih))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
brploboljelih$PopulationCount = statePopulationData$StatePopulationCount
brploboljelih$postotak = round(brploboljelih$sumbrploboljelih /
                                 brploboljelih$PopulationCount * 100, digits = 2)
brploboljelih = brploboljelih %>% arrange(desc(postotak))
brploboljelih
```

```
## # A tibble: 51 x 4
##    StateDesc     sumbrploboljelih PopulationCount postotak
##    <chr>                    <dbl>           <dbl>    <dbl>
##  1 Kentucky                 79512          893140     8.9
##  2 Ohio                    206695         2330226     8.87
##  3 West Virginia             4369           51400     8.5
##  4 Michigan                181180         2225267     8.14
##  5 Tennessee               149066         1836343     8.12
##  6 Alabama                  77282          965304     8.01
##  7 Maryland                 47814          620961     7.7
##  8 Delaware                  5314           70851     7.5
##  9 Pennsylvania            171225         2290681     7.47
## 10 Indiana                 134627         1827472     7.37
## # ... with 41 more rows
```

Jedan od razloga plućnog oboljenja ljudi je zagađenost zraka do koje najčešće dolazi zbog prevelike napučenosti određenog područja. Stoga smo zaključili da bi bilo zanimljivo proučavati savezne države koje se razlikuju po napučenosti, odnosno države koje se razlikuju po broju stanovnika, a površinom su približno jednake.

Prvo smo u tablicu poredanoPopulationCount poredali države po broju stanovnika. Zatim smo iz tablice izdvojili saveznu državu Maine koja je jedna od najmanje napučenih, saveznu državu New York koja je jedna od najnapučenijih i saveznu državu Louisiana koja je srednje napučena. Za te tri savezne države ćemo utvrditi je li jednak postotak stanovništa koji boluje od kroničnih plućnih bolesti.

```
poredanoPopulationCount = brploboljelih %>% arrange(desc(PopulationCount))
brploboljelih$brZdravih = brploboljelih$PopulationCount - brploboljelih$sumbrploboljelih
```

Izdvajamo podatke za saveznu državu Maine u novu tablicu.

```
plMaine = brploboljelih %>% filter(StateDesc == "Maine")
plMaine
```

```
## # A tibble: 1 x 5
##   StateDesc sumbrploboljelih PopulationCount postotak brZdravih
##   <chr>                <dbl>           <dbl>    <dbl>     <dbl>
## 1 Maine                 4303           66194      6.5     61891
```

Izdvajamo podatke za saveznu državu New York u novu tablicu.

```
plNewYork = brploboljelih %>% filter(StateDesc == "New York")
plNewYork
```

```
## # A tibble: 1 x 5
##   StateDesc sumbrploboljelih PopulationCount postotak brZdravih
##   <chr>                <dbl>           <dbl>    <dbl>     <dbl>
## 1 New York            548516         9296499      5.9   8747983
```

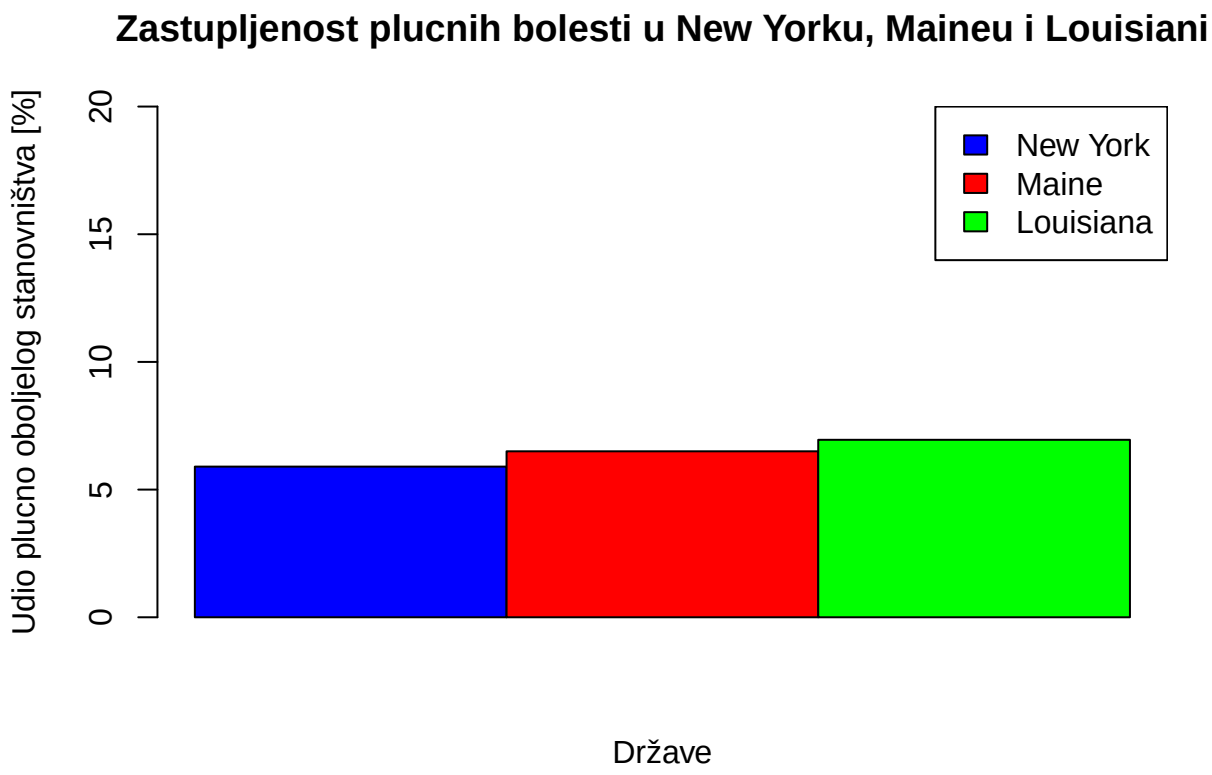
Izdvajamo podatke za saveznu državu Louisiana u novu tablicu.

```
plLouisiana = brploboljelih %>% filter(StateDesc == "Louisiana")
plLouisiana
```

```
## # A tibble: 1 x 5
##   StateDesc sumbrploboljelih PopulationCount postotak brZdravih
##   <chr>                <dbl>           <dbl>    <dbl>     <dbl>
## 1 Louisiana            71737         1031951     6.95    960214
```

Razliku u zastupljenosti kroničnih plućnih bolesti u ove tri savezne države predočavamo stupičasti dijagram u nastavku. Zbog veće uočljivosti razlika među državama y-os smo ograničili na 20%. Iz histograma možemo uočiti da napučenije države ne moraju imati veći postotak plućno oboljelih od manje napučenih država. Iz čega zaključujemo da napučenost nije jedini uzrok povečanog broja kroničnih plućnih bolesti kod ljudi. Također, možemo naslutiti da postotak plućno oboljelih u ove tri savezne države nije jednak što ćemo provjeriti testom u nastavku.

```
plStates <- t(cbind(plNewYork$postotak, plMaine$postotak, plLouisiana$postotak))

barplot(plStates,beside=TRUE, col=c("blue", "red", "green"),
        ylab = "Udio plućno oboljelog stanovništva [%]", cex.names=0.2, xlab = "Države",
        ylim=c(0,20), main="Zastupljenost plućnih bolesti u New Yorku, Maineu i Louisiani")


legend("topright",c("New York", "Maine", "Louisiana"),fill = c("blue", "red", "green"))
```



Sada konačno možemo testirati jednakost postotka kronično plućno oboljelog stanovnišva u ovim saveznim

državama. S obzirom da treba ispitati jednakost više proporcija koristit ćemo χ^2 test, odnosno test homogenosti. Za početak ćemo napravi kontigencijsku tablicu varijabli saveznih država i broja zdravih i oboljelih.

```
sumZarazeni = plNewYork$sumbrploboljelih + plMaine$sumbrploboljelih +
  plLouisiana$sumbrploboljelih
sumZdravi = plNewYork$brZdravih + plMaine$brZdravih + plLouisiana$brZdravih

tbl <- data.frame("NewYork" = c(plNewYork$sumbrploboljelih, plNewYork$brZdravih,
                                plNewYork$PopulationCount),
                  "Maine" = c(plMaine$sumbrploboljelih, plMaine$brZdravih,
                              plMaine$PopulationCount),
                  "Louisiana" = c(plLouisiana$sumbrploboljelih, plLouisiana$brZdravih,
                                  plLouisiana$PopulationCount),
                  "sum" = c(sumZarazeni, sumZdravi, sumZarazeni + sumZdravi))
rownames(tbl) <- c("oboljeli", "zdravi", "sum")
tbl
```

```
##          NewYork Maine Louisiana      sum
## oboljeli  548516  4303     71737   624556
## zdravi   8747983 61891    960214  9770088
## sum      9296499 66194   1031951 10394644
```

Test je postavljen kao: H0: p(NewYork) = p(Maine) = p(Louisiana) H1: p(NewYork), p(Maine), p(Louisiana) nisu jednake s 95%-tnim intervalom pouzdanosti.

Pretpostavka 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).

```
for (col_names in colnames(tbl)){
  for (row_names in rownames(tbl)){
    if (!(row_names == 'sum' | col_names == 'sum') ){
      cat('Očekivane frekvencije za razred ',col_names,'-',row_names,': ',
          (tbl[row_names,'sum'] * tbl['sum',col_names]) / tbl['sum','sum'],'\n')
    }
  }
}
```

```
## Očekivane frekvencije za razred  NewYork - oboljeli :  558574.6
## Očekivane frekvencije za razred  NewYork - zdravi :  8737924
## Očekivane frekvencije za razred  Maine - oboljeli :  3977.227
## Očekivane frekvencije za razred  Maine - zdravi :  62216.77
## Očekivane frekvencije za razred  Louisiana - oboljeli :  62004.16
## Očekivane frekvencije za razred  Louisiana - zdravi :  969946.8
```

Sve očekivane frekvencije su veće od 5. Možemo nastaviti sa χ^2 testom.

```
chisq.test(tbl)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tbl
## X-squared = 1846.5, df = 6, p-value < 2.2e-16
```

Provedbom χ^2 testa zbog p-vrijednosti koja je manja od razine signifikantnosti 0.05 zaključujemo da postotak stanovništa koji boluje od kroničnih plućnih bolesti nije jednak u saveznim državama New York, Louisiana i Maine.

#Usporedba saveznih država s najmanjim i najvećim postotkom ljudi koji imaju zdravstveno osiguranje

```
#Savezna država s najvišim postotkom stanovništva koji imaju zdravstveno osiguranje
data.health.insurance <- subset(stateData, Short_Question_Text == 'Health Insurance')
data.health.insurance <- data.health.insurance[order(data.health.insurance$Data_Value,
                                                     decreasing = FALSE),]
data.health.insurance <- data.health.insurance[1:1,]
data.health.insurance
```

```
## # A tibble: 1 x 7
## # Groups:   StateDesc, Short_Question_Text, Category, StatePopulationCount [1]
##   StateDesc Short_Question_~ Category StatePopulation~ Data_Value_Popu~
##   <chr>     <chr>            <chr>               <dbl>            <dbl>
## 1 Vermont   Health Insurance Prevent~            42417             4114
## # ... with 2 more variables: Data_Value_Unit <chr>, Data_Value <dbl>
```

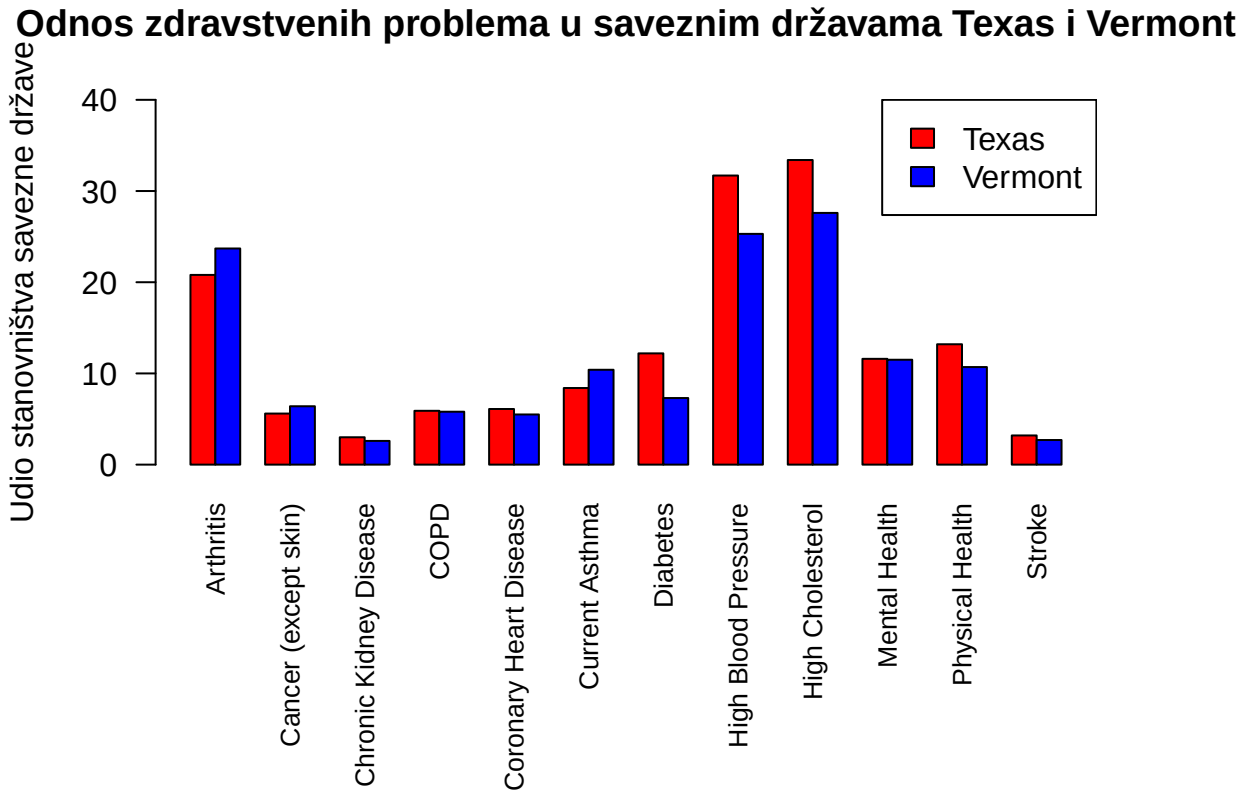
```
#savezna država s najnižim postotkom stanovništva koji imaju zdravstveno osiguranje
data.health.insurance <- subset(stateData, Short_Question_Text == 'Health Insurance')
data.health.insurance <- data.health.insurance[order(data.health.insurance$Data_Value,
                                                     decreasing = TRUE),]
data.health.insurance <- data.health.insurance[1:1,]
data.health.insurance
```

```
## # A tibble: 1 x 7
## # Groups:   StateDesc, Short_Question_Text, Category, StatePopulationCount [1]
##   StateDesc Short_Question_~ Category StatePopulation~ Data_Value_Popu~
##   <chr>     <chr>            <chr>               <dbl>            <dbl>
## 1 Texas     Health Insurance Prevent~         12325490          3362946
## # ... with 2 more variables: Data_Value_Unit <chr>, Data_Value <dbl>
```

Vermont je savezna država s najvišim postotkom stanovništva koji imaju zdravstveno osiguranje, a Texas je savezna država s najnižim postotkom stanovništva sa zdravstvenim osiguranjem. Odlučili smo na bar plotu vidjeti odnos postotka ljudi s određenim zdravstvenim problemima u saveznim državama Texas i Vermont te bi iz danog grafa mogli naslutiti na koje bolesti utječe mali postotak ljudi s zdravstvenim osiguranjem.

```
texasData <- subset(stateData, StateDesc == "Texas")
texasData <- subset(texasData, Category == "Health Outcomes")

vermontData <- subset(stateData, StateDesc == "Vermont")
vermontData <- subset(vermontData, Category == "Health Outcomes")

healthStates <- t(cbind(texasData$Data_Value, vermontData$Data_Value))

par(mar=c(9,4,4,4))

barplot(healthStates,beside=TRUE, col=c("red", "blue"),
        ylab = "Udio stanovništva savezne države",
         cex.names=0.8,
        names.arg = c(vermontData$Short_Question_Text), ylim=c(0,40),
```

```
        main="Odnos zdravstvenih problema u saveznim državama Texas i Vermont",

      las=2)

legend("topright",c("Texas","Vermont"),fill = c("red", "blue"))
```



Iz danog grafa naslućuje se da zdravstveno osiguranje najviše utječe na postotak stanovništva koji pate od dijabetesa, visokog krvnog tlaka i povišenog kolesterola. Također, naizgled se čini da zdravstveno osiguranje nema utjecaja neke druge bolesti, kao npr. plućne bolesti ili mentalne bolesti. Odlučili smo testirati jednakost postotka ljudi koji pate od visokog krvnog tlaka u Texasu i Vermontu.

U sljedećoj tablici v_t_healthOutcomes su iz tablice stateData filtrirani podatci za države Texas i Vermont i mjere bolesti.

```
v_t_healthOutcomes = stateData %>%
  filter(Category == "Health Outcomes" & (StateDesc =="Texas" | StateDesc =="Vermont")) %>%
  arrange(desc(StateDesc))  %>% arrange(Short_Question_Text)
v_t_healthOutcomes
```

```
## # A tibble: 24 x 7
## # Groups:   StateDesc, Short_Question_Text, Category, StatePopulationCount [24]
##    StateDesc Short_Question_~ Category StatePopulation~ Data_Value_Popu~
##    <chr>     <chr>            <chr>               <dbl>            <dbl>
##  1 Vermont   Arthritis        Health ~            42417            10053
##  2 Texas     Arthritis        Health ~         12325490          2562081
##  3 Vermont   Cancer (except ~ Health ~            42417             2715
```

```
##  4 Texas     Cancer (except ~ Health ~         12325490           688775
##  5 Vermont   Chronic Kidney ~ Health ~            42417             1103
##  6 Texas     Chronic Kidney ~ Health ~         12325490           368814
##  7 Vermont   COPD             Health ~            42417             2460
##  8 Texas     COPD             Health ~         12325490           732866
##  9 Vermont   Coronary Heart ~ Health ~            42417             2333
## 10 Texas     Coronary Heart ~ Health ~         12325490           754465
## # ... with 14 more rows, and 2 more variables: Data_Value_Unit <chr>,
## #   Data_Value <dbl>
```

Provodimo test o dvije proporcije za mjeru bolesti "High Blood Pressure" i za mjeru bolesti "Mental Health".

Testovi za svaku mjeru prevencije će biti postavljeni kao: H0: p(vermont) = p(texas) H1: p(vermont) < p(texas) s 95%-tnim intervalom pouzdanosti.

High Blood Pressure:

U donjem ispisu vidimo da je p-vrijednost manja od razine signifikantnosti 0.05 pa odbacujemo hipotezu H0 i zaključujemo da je mjera bolesti "High Blood Pressure" popularnija u saveznoj državi Texas nego u saveznoj državi Vermont.

```
v_t_highBloodPressure = v_t_healthOutcomes %>%
  filter(Short_Question_Text == "High Blood Pressure")

prop.test(x = v_t_highBloodPressure$Data_Value_Population_Count,
          n = v_t_highBloodPressure$StatePopulationCount, alternative = "less",
          conf.level = 0.95, correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  v_t_highBloodPressure$Data_Value_Population_Count out of v_t_highBloodPressure$StatePopulatio
## X-squared = 807.4, df = 1, p-value < 2.2e-16
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.00000000 -0.06083427
## sample estimates:
##    prop 1    prop 2
## 0.2530118 0.3173249
```

Mental Health:

U donjem ispisu vidimo da je p-vrijednost veća od razine signifikantnosti 0.05 te ne možemo odbaciti hipotezu H0. P-vrijednost je blizu vrijednosti 0.5 te možemo zaključiti da su postotci ljudi približno jednaki.

```
v_t_mentalHealth = v_t_healthOutcomes %>% filter(Short_Question_Text == "Mental Health")

prop.test(x = v_t_mentalHealth$Data_Value_Population_Count,
          n = v_t_mentalHealth$StatePopulationCount, alternative = "less",
          conf.level = 0.95, correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  v_t_mentalHealth$Data_Value_Population_Count out of v_t_mentalHealth$StatePopulationCount
## X-squared = 0.14846, df = 1, p-value = 0.35
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.000000000  0.001953072
## sample estimates:
##    prop 1    prop 2
## 0.1150011 0.1156003
```

Ova dva testa dala su nam zanimljive rezultate. Daje se naslutiti da zdravstveno osiguranje nema utjecaja na sve bolesti u pojedinoj državi. Definitvno je zanimljivo razmisliti o razlogu zašto bi zdravstveno osiguranje više utjecalo na količinu ljudi koji pate od povišenog krvnog tlaka nego na količinu ljudi koji pate od mentalnih bolesti. Vjerojatno postoje drugi faktori koji bolje opisuju zašto zdravstveno osiguranje manje utječe na neke bolesti, no takve zaključke ne možemo naslutiti iz provedenog istraživanja.

3) Ispitajte veze izmedu ove 4 metode preventivne zaštite i pojedine bolesti. Koje metode imaju najveći utjecaj? Na koje bolesti? Što ste od zavisnosti očekivali, a što vas je iznenadilo?

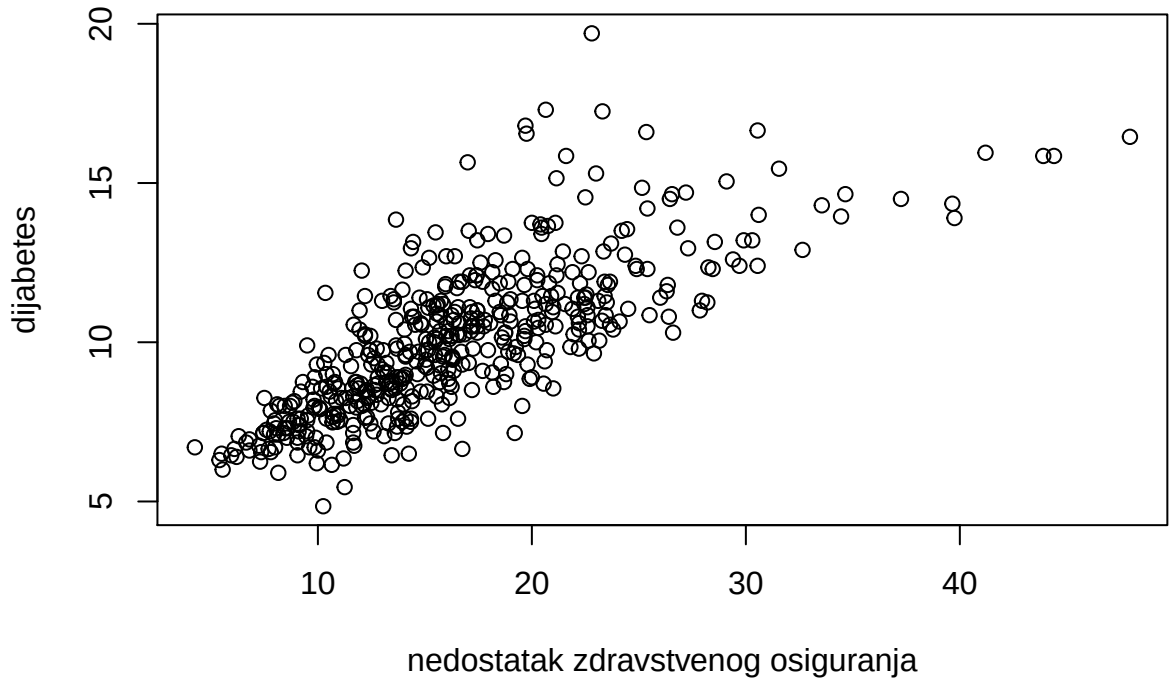
Ispitujemo vezu između nedostatka zdrastvenog osiguranja (u tablici "Current lack of health insurance among adults aged 18–64 Years") i dijabetesa (u tablici "Diagnosed diabetes among adults aged >=18 Years").

Prvo ćemo za svaki grad filtrirati dva podatka: broj građana koji nemaju zdravstveno osiguranje te broj građana koji imaju dijabetes. Budući da promatramo utjecaj jedne nezavisne varijable (nedostatak zdravstvenog osiguranja) na zavisnu (dijabetes), prigodni prikaz njihovog odnosa će nam pokazati scatter plot. Dakle svaki grad će biti predstavljen jednom točkom u dijagramu.

```
data %>% filter(Short_Question_Text == "Health Insurance") %>%
  summarise(health_insurance_value = Data_Value, city = CityName) ->
  health_insurance_data
data %>% filter(Short_Question_Text == "Diabetes") %>%
  summarise(diabetes_value = Data_Value, city = CityName) -> diabetes_data
health_insurance_diabetes <- merge(diabetes_data,health_insurance_data,by="city")
health_insurance_diabetes %>% group_by(city) %>%
  summarise(diabetes_value = mean(diabetes_value),
  health_insurance_value = mean(health_insurance_value)) -> health_insurance_diabetes
```

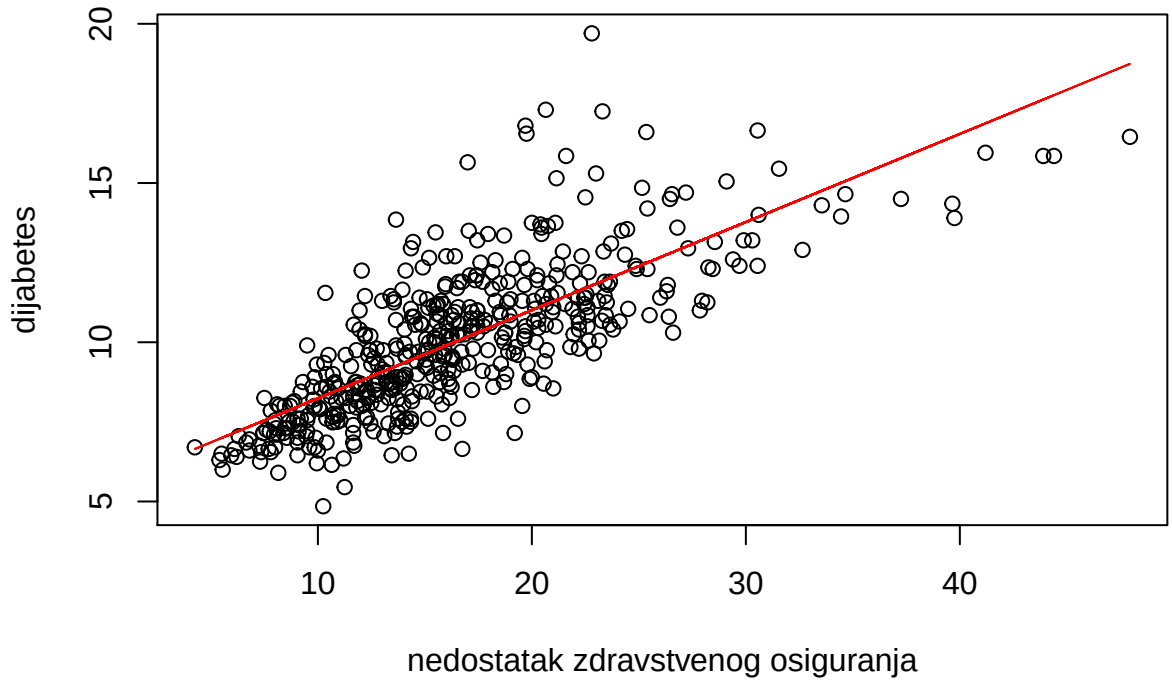
```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
plot(health_insurance_diabetes$health_insurance_value,
     health_insurance_diabetes$diabetes_value,
     xlab="nedostatak zdravstvenog osiguranja", ylab="dijabetes")
```



Vidimo iz scatter plota da nedostatak zdravstvenog osiguranja ima izražen utjecaj na broj dijabetičara. Možemo primjetiti da što više ljudi nema zdravsteno osiguranje to ima više ljudi koji imaju dijabetes, što ima smisla. Kako bismo ispitali utjecaj navedene prevencije na navedenu bolest, procijenit ćemo model jednostavne regresije sa prevencijom kao nezavisnom varijablom (regresorom) te bolešću kao zavisnom varijablom.

```
fit_ins_diabetes = lm(diabetes_value~health_insurance_value,
                      data=health_insurance_diabetes)
plot(health_insurance_diabetes$health_insurance_value,
     health_insurance_diabetes$diabetes_value,
     xlab="nedostatak zdravstvenog osiguranja",
     ylab="dijabetes") #graficki prikaz podataka
lines(health_insurance_diabetes$health_insurance_value,
      fit_ins_diabetes$fitted.values,col='red')
```

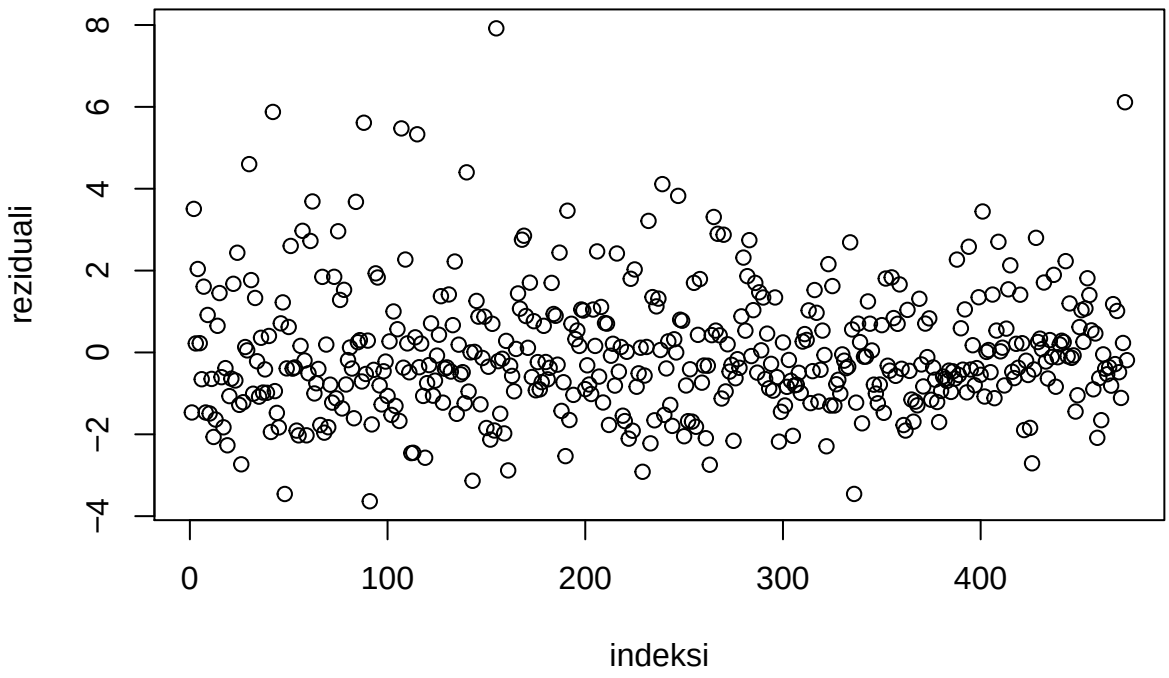



```
#grafici prikaz procijenjenih vrijednosti iz modela
```

Kako bismo dobiveni modeli mogli analizirati, prvo je potrebno provjeriti da pretpostavke modela nisu narušene. Pritom ćemo provjeravati dvije pretpostavke modela: normalnost reziduala i homogenost varijance.

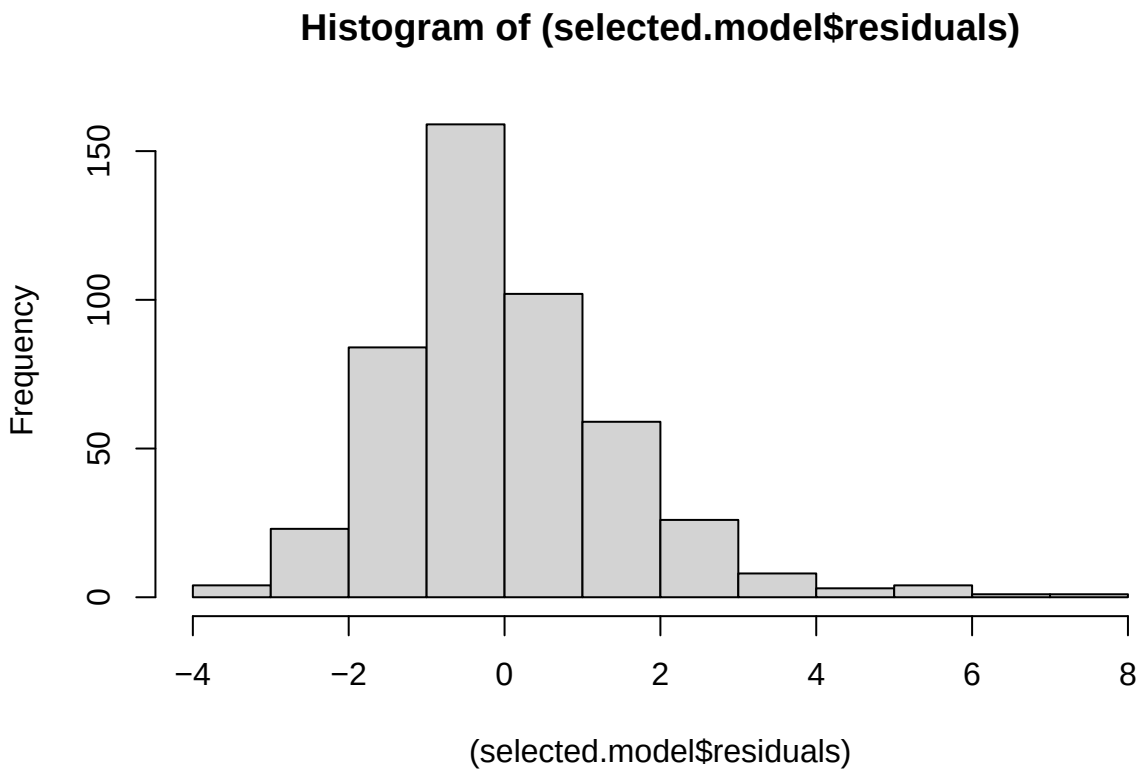
Normalnost reziduala ćemo provjeriti grafički, pomoću kvantil-kvantil plota (usporedbom s linijom normalne razdiobe), te statistički pomoću Kolmogorov-Smirnovljevog testa te Lillieforceve inačica Kolmogorov-Smirnovljevog testa (ne poznaju se očekivanje i varijanca populacije, a Lillieforceov test je upravo za takvu primjenu). Homogenost varijance ćemo provjeriti pomoću grafa reziduala u ovisnosti o izlazu modela.

```
selected.model = fit_ins_diabetes

#Grafički prikaz reziduala samo po indeksu - ne možemo zaključiti ništa o normalnosti
plot(selected.model$residuals,
     xlab="indeksi",
     ylab="reziduali")
```

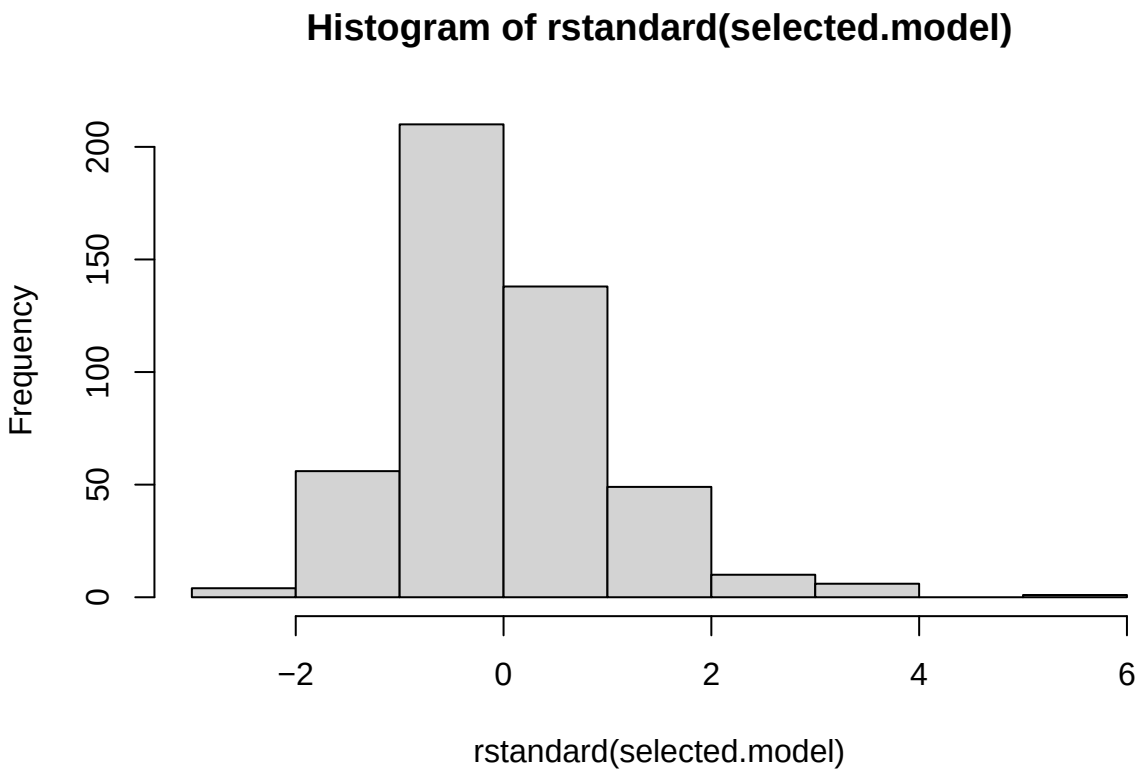


Na gornjem scatter plotu možemo vidjeti sve reziduale. Ovaj prikaz nam nije jako interpretativan pa ćemo preći na sljedeće prikaze.

```
#Histogram reziduala
hist((selected.model$residuals))
```

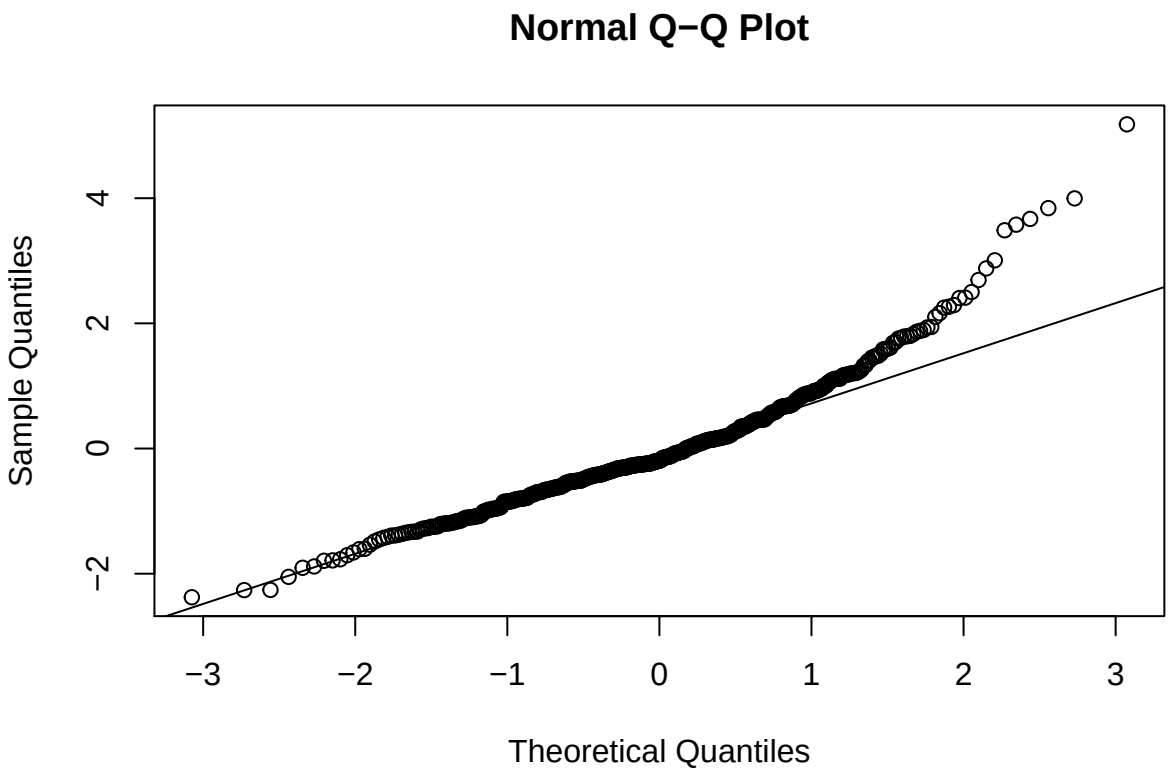



```
#Histogram standardiziranih reziduala
hist(rstandard(selected.model))
```



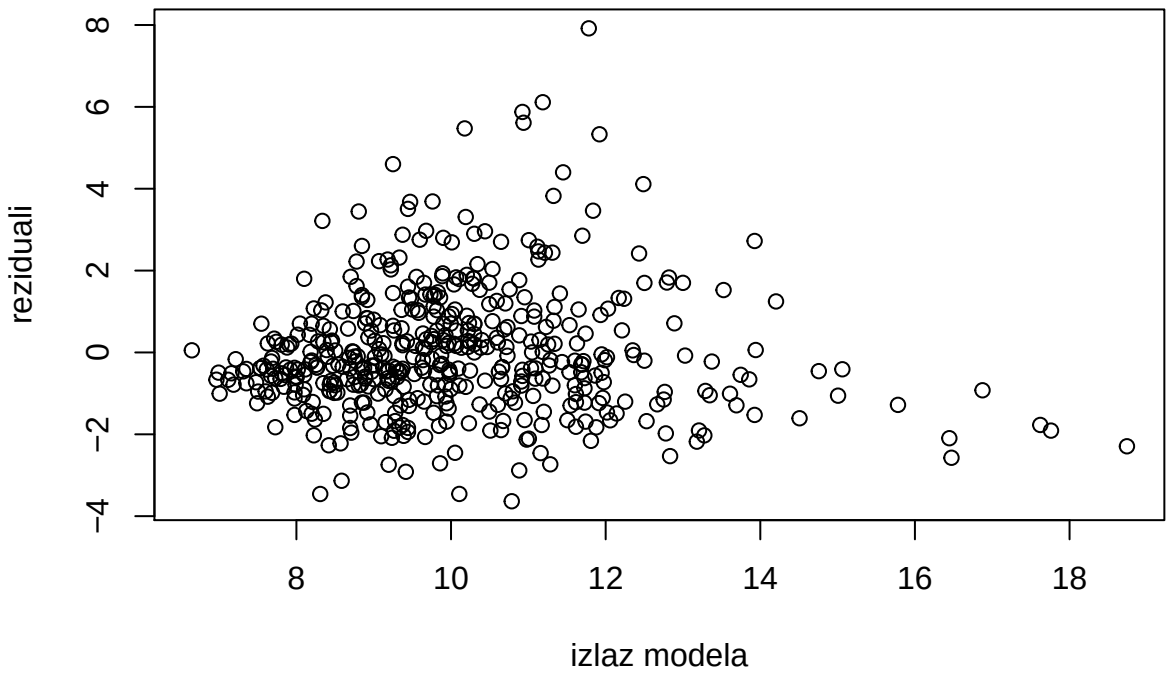
Gore su prikazana dva histograma. Prvi histogram prikazuje "sirove reziduale" dok drugi prikazuje standardizirane reziduale. Možemo primjetiti da su distribucije lagano zakrivljene u desno. Drugi histogram je interpretativniji jer očekujemo da se standardizirani reziduali ponašaju normalno. Također, standardizirane reziduale ćemo testirati na normalnost. To je ujedno i razlog zbog kojeg standardizirane reziduale koristimo i u q-q plotu.

```
#q-q plot reziduala s linijom normalne distribucije
qqnorm(rstandard(selected.model))
qqline(rstandard(selected.model))
```



Iznad je prikazan q-q plot reziduala s linijom normalne distribucije. Vidimo da q-q plot nije savršen zbog velikih odstupanja nekih točaka od linije normalnosti.

```
plot(selected.model$fitted.values,selected.model$residuals,
      xlab="izlaz modela",
     ylab="reziduali")
```




```
#reziduale je dobro prikazati u ovisnosti o procjenama modela
```

Iznad su prikazani reziduali u odnosu na izlazne vrijednosti modela. Postoji područje gdje su neki reziduali veći u odnosu na ostatak. Budući da reziduali nisu uvijek u istom intervalu, nije ispunjena pretpostavka homogenosti varijance. To znači da neki dio varijabilnosti nije objašnjen našim modelom.

```
#Kolmogorov-Smirnovljev test na normalnost
ks.test(rstandard(selected.model),'pnorm')
```

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

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  rstandard(selected.model)
## D = 0.0961, p-value = 0.0003153
## alternative hypothesis: two-sided
```

```
#Lillieforceov test na normalnost
require(nortest)
```

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

```
lillie.test(rstandard(selected.model))
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rstandard(selected.model)
## D = 0.09601, p-value = 2.842e-11
```

Gornji testovi ne mogu potvrditi našu pretpostavku da su reziduali normalni. Nul-hipoteza oba testa je da su reziduali normalni, a zbog male p-vrijednosti možemo odbaciti nul-hipotezu. Dakle, nećemo dalje analizirati ovaj model (osim summaryja) jer ne možemo pretpostaviti normalnost reziduala. Treba napomenuti da testove provodimo nad standardiziranim rezidualima.

Iz summaryja modela možemo očitati sljedeće nama značajne vrijednosti: vidimo da je p-vrijednost F-statistike jako mala pa možemo zaključiti da je model signifikantan.

Također možemo očitati koeficijent deteminacije R^2. On je bitan pokazatelj kvalitete prilagodbe modela koji nam govori koliki je postotak varijance u modelu obuhvaćen tj. objašnjen našim modelom.

```
summary(selected.model)
```

```
##
## Call:
## lm(formula = diabetes_value ~ health_insurance_value, data = health_insurance_diabetes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6353 -0.9480 -0.2960  0.7044  7.9183
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              5.47131    0.19266   28.40   <2e-16 ***
## health_insurance_value  0.27677    0.01088   25.44   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.532 on 472 degrees of freedom
## Multiple R-squared:  0.5783, Adjusted R-squared:  0.5774
## F-statistic: 647.2 on 1 and 472 DF,  p-value: < 2.2e-16
```

Provjerimo kako ostale prevencije utječu na broj dijabetičara. Provjerimo jesu li svi modeli signifikantni.

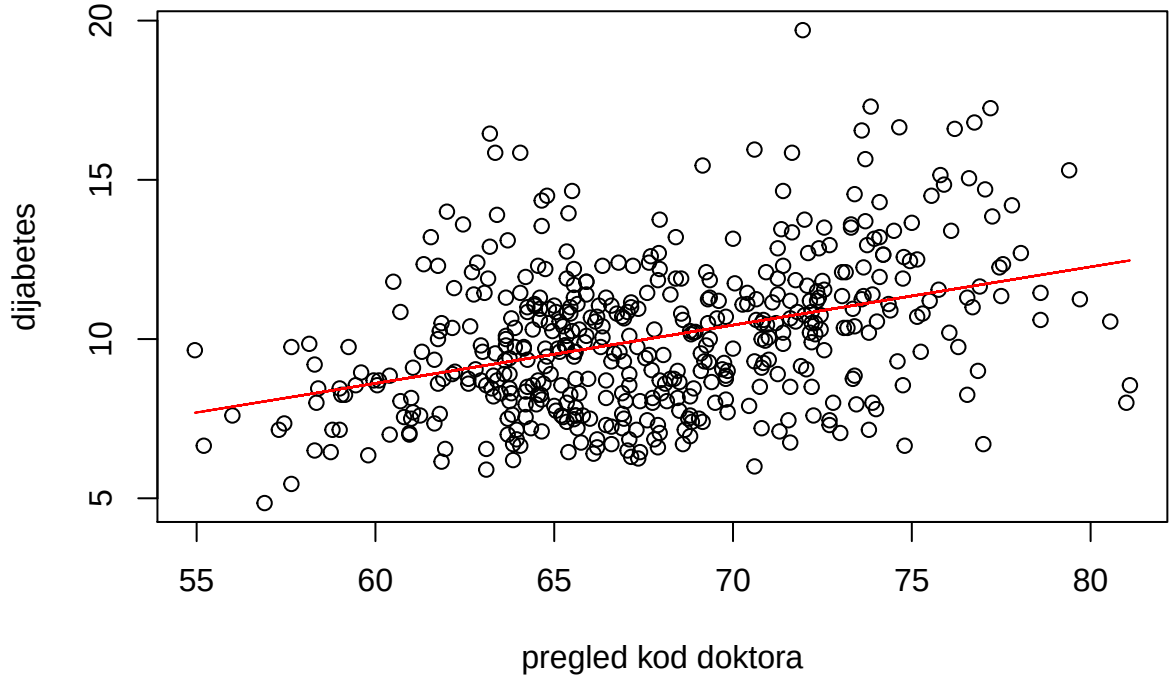
Prvo provjeravamo model kojem je nezavisna varijabla broj građana koji su obavili rutinski pregled kod u zadnjih godinu dana. (u tablici "Visits to doctor for routine checkup within the past Year among adults aged >=18 Years"). Zavisna varijabla ostaje u svakom modelu ista.

```
#regresije s ostalim metodama prevencije - jesu li signifikantne?

data %>% filter(Short_Question_Text == "Annual Checkup") %>%
  summarise(annual_checkup_value = Data_Value, city = CityName) -> annual_checkup_data
data %>% filter(Short_Question_Text == "Diabetes") %>%
  summarise(diabetes_value = Data_Value, city = CityName) -> diabetes_data
annual_checkup_diabetes <- merge(diabetes_data,annual_checkup_data,by="city")
```

```
annual_checkup_diabetes%>% group_by(city) %>%
  summarise(diabetes_value = mean(diabetes_value),
            annual_checkup_value = mean(annual_checkup_value)) -> annual_checkup_diabetes
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
fit_ann_diabetes = lm(diabetes_value~annual_checkup_value,data=annual_checkup_diabetes)
plot(annual_checkup_diabetes$annual_checkup_value, annual_checkup_diabetes$diabetes_value,
     xlab="pregled kod doktora", ylab="dijabetes")
lines(annual_checkup_diabetes$annual_checkup_value,fit_ann_diabetes$fitted.values,col='red')
```




```
summary(fit_ann_diabetes)
```

```
##
## Call:
## lm(formula = diabetes_value ~ annual_checkup_value, data = annual_checkup_diabetes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0194 -1.5427 -0.0833  1.2805  8.9045
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -2.3676     1.3800  -1.716   0.0869 .
```

```
## annual_checkup_value   0.1830     0.0203   9.010   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.179 on 472 degrees of freedom
## Multiple R-squared:  0.1468, Adjusted R-squared:  0.145
## F-statistic: 81.19 on 1 and 472 DF,  p-value: < 2.2e-16
```

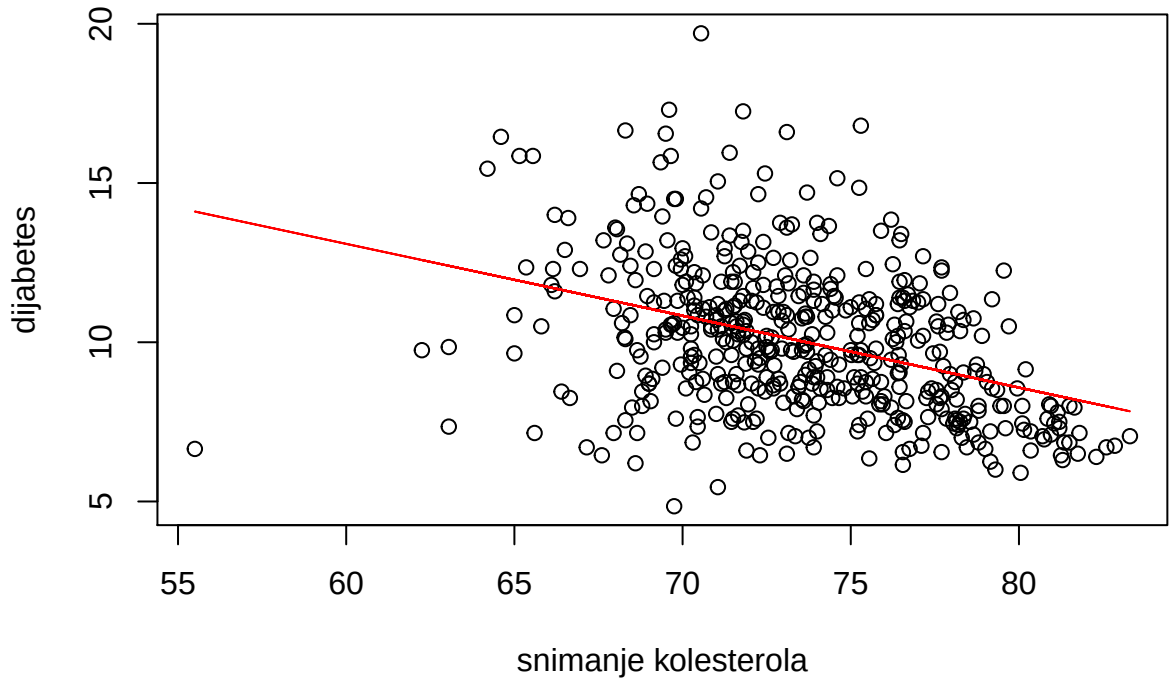
Model je signifikantan (jako mala p-vrijednost) i objašnjava otrpilike 14.5% varijance.

Iduće provjeravamo model kojem je nezavisna varijabla broj građana koji obavljaju snimanje kolesterola u krvi (u tablici "Cholesterol screening among adults aged >=18 Years").

```
data %>% filter(Short_Question_Text == "Cholesterol Screening") %>%
  summarise(cholesterol_screening_value = Data_Value,
            city = CityName) -> cholesterol_screening_data
data %>% filter(Short_Question_Text == "Diabetes") %>%
  summarise(diabetes_value = Data_Value, city = CityName) -> diabetes_data
cholesterol_screening_diabetes <- merge(diabetes_data,cholesterol_screening_data,by="city")
cholesterol_screening_diabetes %>% group_by(city) %>%
  summarise(diabetes_value = mean(diabetes_value),
            cholesterol_screening_value = mean(cholesterol_screening_value)) ->
  cholesterol_screening_diabetes
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
fit_cs_diabetes = lm(diabetes_value~cholesterol_screening_value,
                     data=cholesterol_screening_diabetes)
plot(cholesterol_screening_diabetes$cholesterol_screening_value,
     cholesterol_screening_diabetes$diabetes_value, xlab="snimanje kolesterola",
     ylab="dijabetes")
lines(cholesterol_screening_diabetes$cholesterol_screening_value,
      fit_cs_diabetes$fitted.values,col='red')
```




```
summary(fit_cs_diabetes)
```

```
##
## Call:
## lm(formula = diabetes_value ~ cholesterol_screening_value, data = cholesterol_screening_diabetes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.4545 -1.4249 -0.2429  1.3783  8.9916
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  26.62835    1.83997  14.472   <2e-16 ***
## cholesterol_screening_value -0.22565    0.02498  -9.032   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.178 on 472 degrees of freedom
## Multiple R-squared:  0.1474, Adjusted R-squared:  0.1456
## F-statistic: 81.58 on 1 and 472 DF,  p-value: < 2.2e-16
```

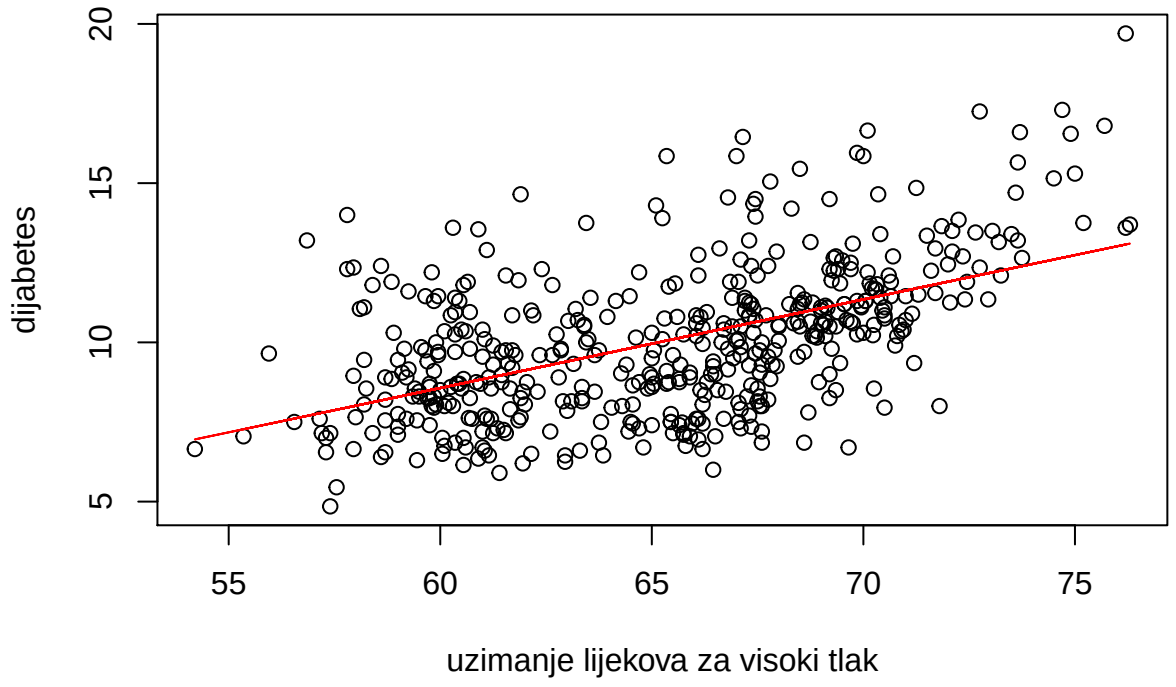
Model je signifikantan (jako mala p-vrijednost) i objašnjava otrpilike 14.56% varijance.

Konačno, provjeravamo model kojem je nezavisna varijabla broj građana koji imaju visok tlak te uzimaju lijekove za visoki tlak (u tablici "Taking medicine for high blood pressure control among adults aged >=18 Years with high blood pressure").

```
data %>% filter(Short_Question_Text == "Taking BP Medication") %>%
  summarise(taking_bp_med_value = Data_Value, city = CityName) -> taking_bp_med_data
data %>% filter(Short_Question_Text == "Diabetes") %>%
  summarise(diabetes_value = Data_Value, city = CityName) -> diabetes_data
taking_bp_med_diabetes <- merge(diabetes_data,taking_bp_med_data,by="city")
taking_bp_med_diabetes  %>% group_by(city) %>%
  summarise(diabetes_value = mean(diabetes_value),
            taking_bp_med_value = mean(taking_bp_med_value)) -> taking_bp_med_diabetes
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
fit_bp_diabetes = lm(diabetes_value~taking_bp_med_value,data=taking_bp_med_diabetes)
plot(taking_bp_med_diabetes $taking_bp_med_value, taking_bp_med_diabetes $diabetes_value,
     xlab="uzimanje lijekova za visoki tlak", ylab="dijabetes")
lines(taking_bp_med_diabetes$taking_bp_med_value,fit_bp_diabetes$fitted.values,col='red')
```




```
summary(fit_bp_diabetes)
```

```
##
## Call:
## lm(formula = diabetes_value ~ taking_bp_med_value, data = taking_bp_med_diabetes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -4.5515 -1.2860 -0.1543  1.0374  6.6259
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -8.12954    1.34201  -6.058 2.82e-09 ***
## taking_bp_med_value  0.27826    0.02051  13.567  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.001 on 472 degrees of freedom
## Multiple R-squared:  0.2805, Adjusted R-squared:  0.279
## F-statistic: 184.1 on 1 and 472 DF,  p-value: < 2.2e-16
```

Model je signifikantan (jako mala p-vrijednost) i objašnjava otrpilike 27.9% varijance.

Svi su modeli signifikantni odnosno sve prevencije utječu na broj dijabetičara. Dakle, možemo odabrati bilo koji skup prevencija od navedenih četiri za višestruku regresiju (uz poštivanje daljnih pretpostavki modela).

Za višestruku regresiju osim normalnosti reziduala i homogenosti varijance moramo provjeriti i da regresori nisu međusobno jako korelirani.

```
#višestruka regresija

#korelacija

cor(cbind(health_insurance_diabetes$health_insurance_value,
          annual_checkup_diabetes$annual_checkup_value,
          taking_bp_med_diabetes$taking_bp_med_value,
          cholesterol_screening_diabetes$cholesterol_screening_value))
```

```
##             [,1]       [,2]      [,3]       [,4]
## [1,]  1.00000000 0.03755962 0.2183712 -0.5530699
## [2,]  0.03755962 1.00000000 0.7958518  0.4254834
## [3,]  0.21837123 0.79585179 1.0000000  0.1442575
## [4,] -0.55306990 0.42548338 0.1442575  1.0000000
```

```
# korelacijski koeficijenti parova regresora
```

Iz tablice vidimo da su druga (rutinski pregled kod doktora) i treća (uzimanje lijekova za visoki tlak kod pacijenata s visokim tlakom) varijabla jako korelirane. Njihov korelacijski koeficijent iznosi približno 0.7959%. Zato ćemo jednu od tih dviju varijabli morati maknuti iz modela višestruke regresije.

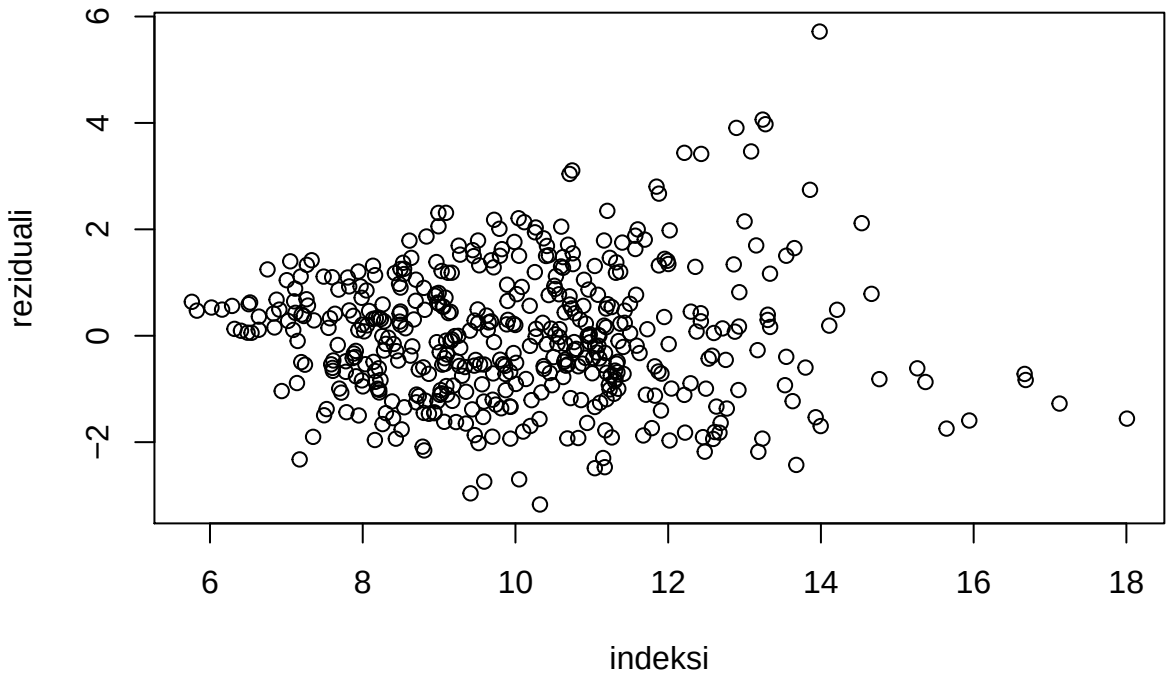
Odlučivat ćemo na temelju R^2 vrijednosti gornjih modela. Naime, model sa rutinskim pregledom kao zavisnom varijablom imao je vrijednost $R^2 = 14.5$ dok je model sa uzimanjem lijekova za visoki tlak kod pacijenata s visokim tlakom kao zavisnom varijablom imao $R^2 = 27.9$. Uzimajući te vrijednosti u obzir, izbacujemo varijablu "rutinski pregled kod doktora" iz modela.

```
#zajednicki dataset
total <- merge(health_insurance_diabetes, taking_bp_med_diabetes)
total <- merge(total, cholesterol_screening_diabetes)


#višetruka regresija
fit.multi = lm(diabetes_value ~ health_insurance_value + taking_bp_med_value +
                 cholesterol_screening_value, total)
```

Kao i u jednostavnoj regresiji, potrebno je provjeriti da pretpostavke modela nisu narušene. Pritom ćemo provjeravati dvije pretpostavke modela: normalnost reziduala i homogenost varijance.

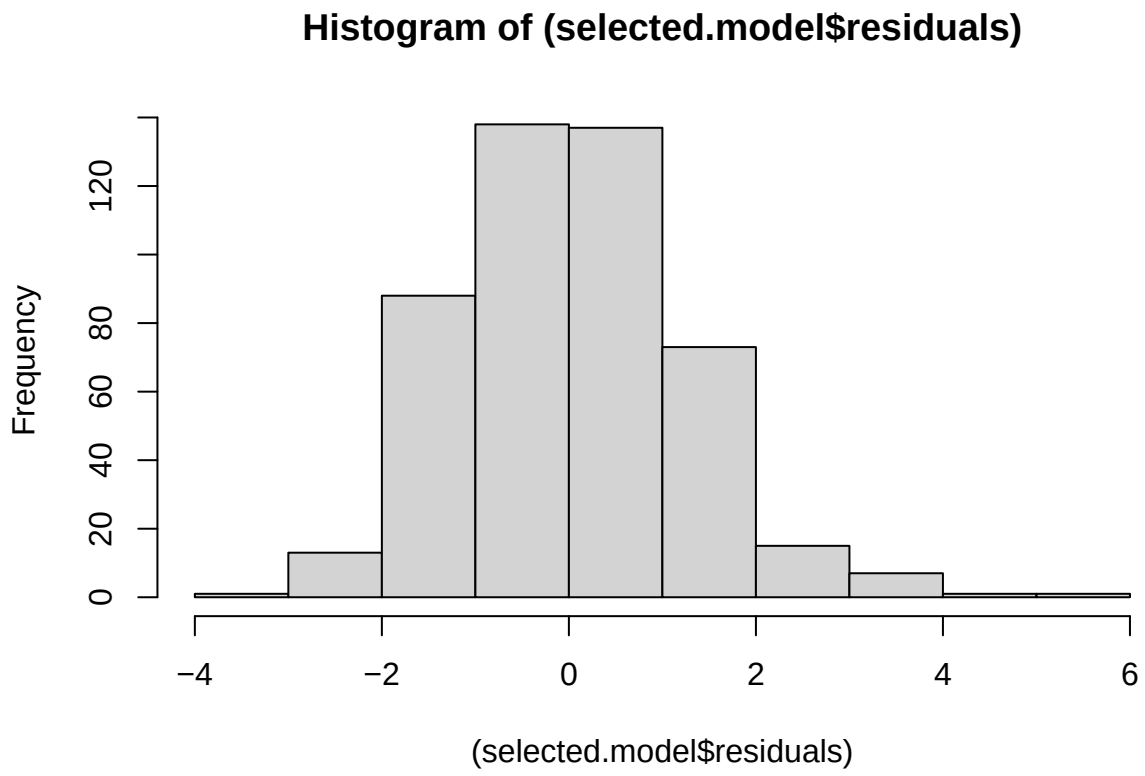
Normalnost reziduala ćemo provjeriti grafički, pomoću kvantil-kvantil plota (usporedbom s linijom normalne razdiobe), te statistički pomoću Kolmogorov-Smirnovljevog testa te Lillieforceve inačica Kolmogorov-Smirnovljevog testa (ne poznaju se očekivanje i varijanca populacije, a Lillieforceov test je upravo za takvu primjenu). Homogenost varijance ćemo provjeriti pomoću grafa reziduala u ovisnosti o izlazu modela.

```
selected.model = fit.multi

plot(selected.model$fitted.values,selected.model$residuals,
      xlab="indeksi",
     ylab="reziduali")
```

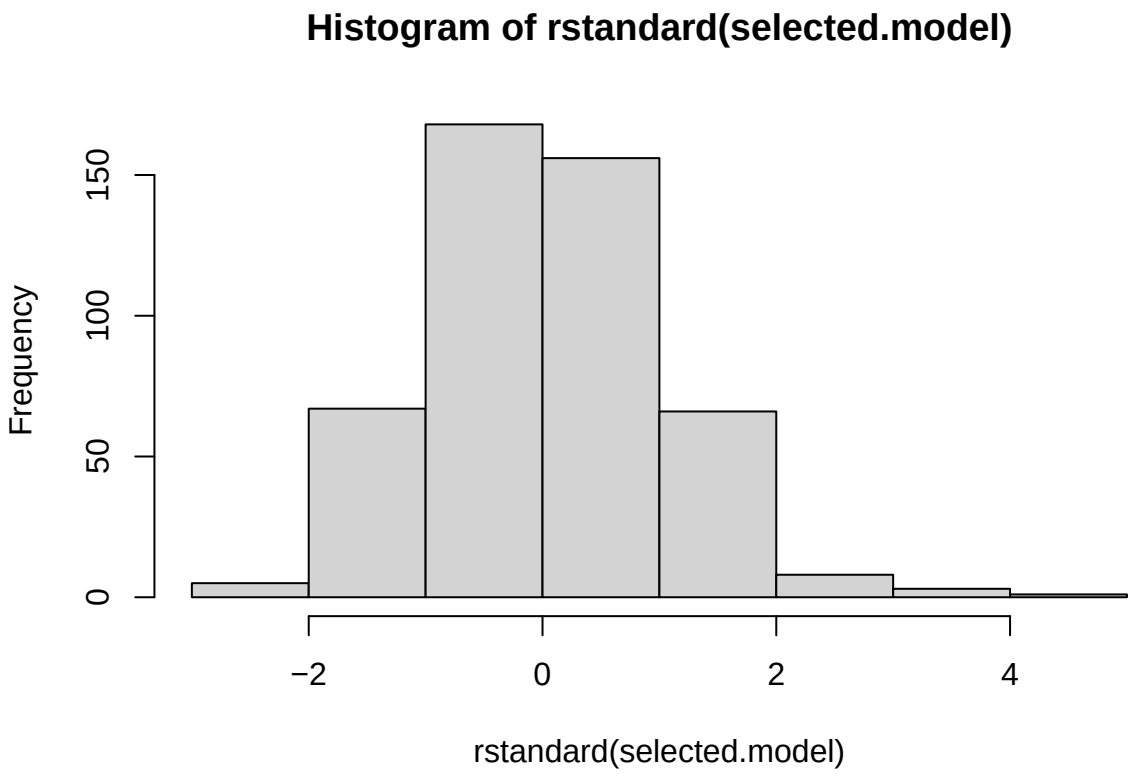


Na gornjem scatter plotu možemo vidjeti sve reziduale. Ovaj prikaz nam nije jako interpretativan pa ćemo preći na sljedeće prikaze.

```
hist((selected.model$residuals))
```

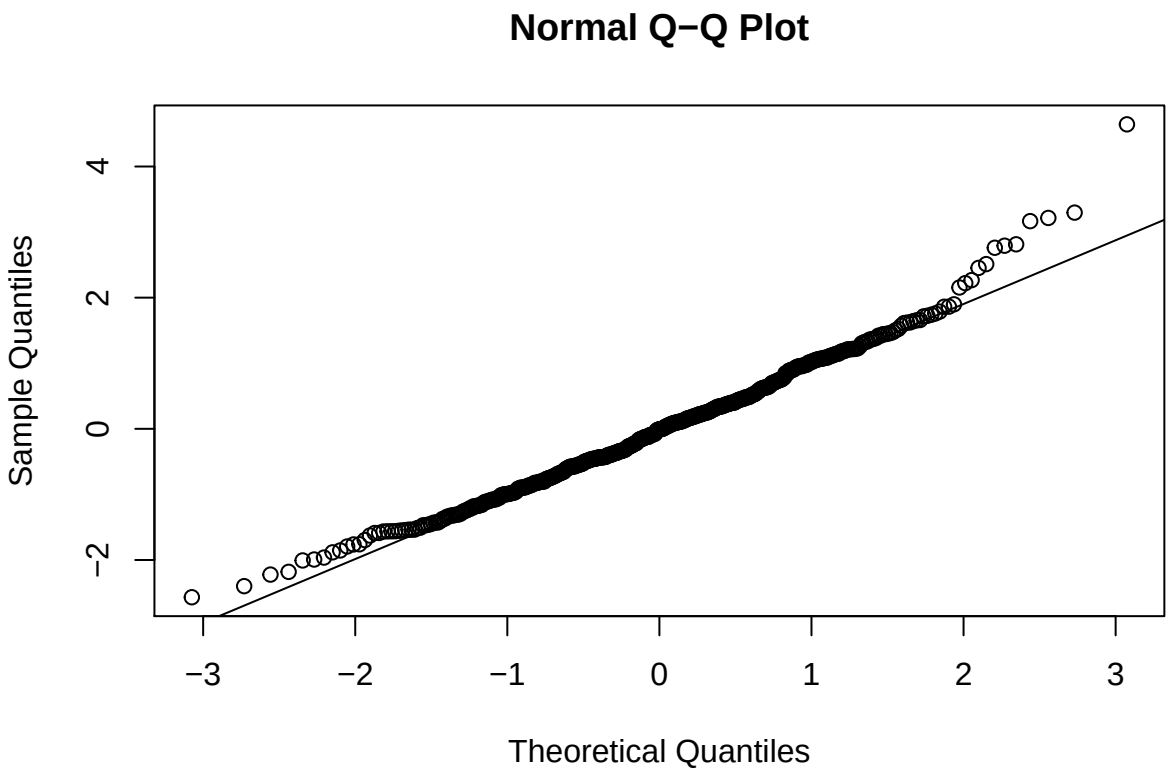



```
hist(rstandard(selected.model))
```



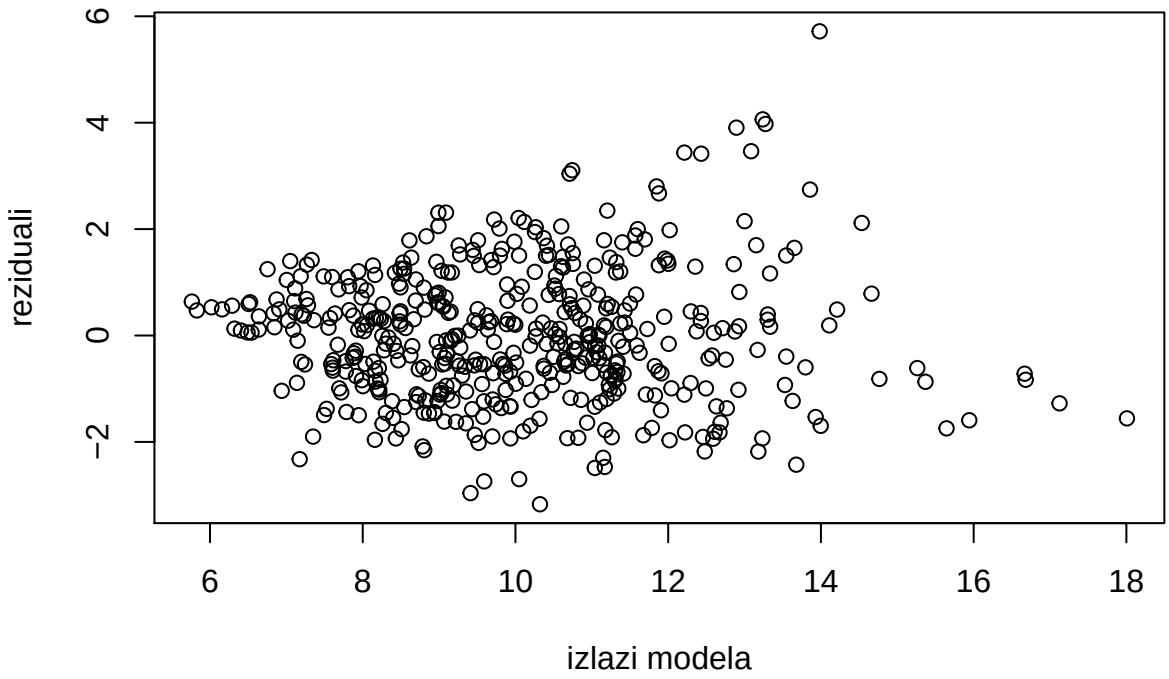
Gore su prikazana dva histograma. Prvi histogram prikazuje "sirove reziduale" dok drugi prikazuje standardizirane reziduale. Kao i u jednostavnoj regresiji, možemo vidjeti da su distribucije lagano zakrivljene u desno.

```
qqnorm(rstandard(selected.model))
qqline(rstandard(selected.model))
```



Iznad je prikazan q-q plot reziduala s linijom normalne distribucije. Vidimo da q-q plot i dalje nije savršen, no izgleda "bolje" nego u jednostanvoj regresiji.

```
plot(selected.model$fitted.values,selected.model$residuals,
      xlab="izlazi modela",
     ylab="reziduali")
```



Iznad su prikazani reziduali u odnosu na izlazne vrijednosti modela. Postoji područje gdje su neki reziduali veći u odnosu na ostatak. Budući da reziduali nisu uvijek u istom intervalu, ne možemo pretpostaviti homogenost varijance. To znači da neki dio varijabilnosti nije objašnjen našim modelom.

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

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  rstandard(fit.multi)
## D = 0.037184, p-value = 0.5287
## alternative hypothesis: two-sided
```

```
require(nortest)
lillie.test(rstandard(fit.multi))
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rstandard(fit.multi)
## D = 0.037407, p-value = 0.1103
```

Rezultati testova na normalnost su mnogo bolji nego kod jednostavne regresije. Uz razinu signifikantnosti 5% ne možemo odbaciti nul-hipotezu (normalnost reziduala).

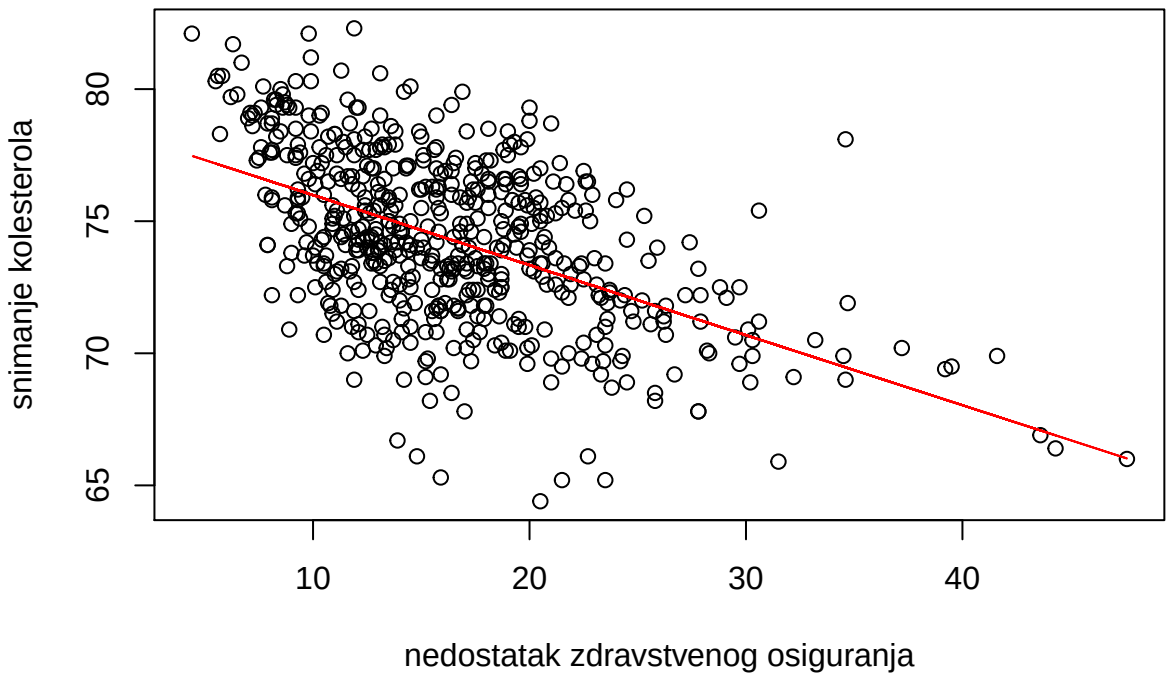
```
summary(fit.multi)
```

```
##
## Call:
## lm(formula = diabetes_value ~ health_insurance_value + taking_bp_med_value +
##     cholesterol_screening_value, data = total)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1710 -0.8622 -0.0110  0.7612  5.7167
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -3.23730    1.42067  -2.279 0.023131 *
## health_insurance_value       0.22320    0.01136  19.650  < 2e-16 ***
## taking_bp_med_value          0.21579    0.01380  15.632  < 2e-16 ***
## cholesterol_screening_value -0.06111    0.01809  -3.378 0.000791 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.242 on 470 degrees of freedom
## Multiple R-squared:  0.7238, Adjusted R-squared:  0.7221
## F-statistic: 410.6 on 3 and 470 DF,  p-value: < 2.2e-16
```

Iz summaryja modela višestruke regresije možemo isčitati da je vrijednost $R^2 = 0.7221$, odnosno da je 72.21% varijance objašnjeno našim modelom. To je značajno više nego u modelu jednostavne regresije. Također, p-vrijednost je očekivano mala (u svakoj jednostavnoj regresiji je bila mala) pa je model signifikantan.

Osim ovisnosti bolesti o mjerama prevencije, odlučili smo istražiti ovisi li ikoja mjera prevencije o nekoj drugoj. Odlučili smo se za model gdje je nezavisna varijabla nedostatak zdravstvenog osiguranja, a zavisna snimanje kolesterola. Drugim riječima, provjeravali smo hoće li ljudi koji imaju zdravstveno osiguranje češće kotrolirati kolesterol od onih koji nemaju. Naš je model nepotpun pa nismo ni provjeravali pretpostavke modela, no možemo vidjeti da ovisnost između varijabli postoji. Moguće je da zdravstveno osiguranje pokriva kontrolu kolesterola u nekim državama pa smo zato dobili takve rezultate.

```
health_insurance_data <- ageAdjData %>%
  filter(Short_Question_Text == "Health Insurance") %>%
  summarise(health_insurance_value = Data_Value, city = CityName)

cholesterol_screening_data <- ageAdjData %>%
  filter(Short_Question_Text == "Cholesterol Screening") %>%
  summarise(cholesterol_screening_value = Data_Value, city = CityName)

total <- merge(cholesterol_screening_data,health_insurance_data,by="city")
total <- total %>% group_by(city)

fit.cholesterol = lm(cholesterol_screening_value~health_insurance_value, data=total)
plot(total$health_insurance_value, total$cholesterol_screening_value,
      xlab="nedostatak zdravstvenog osiguranja",
     ylab="snimanje kolesterola")
lines(total$health_insurance_value,fit.cholesterol$fitted.values,col='red')
```




```
#graficki prikaz procijenjenih vrijednosti iz modela
```