Projekt iz predmeta "Statistička analiza podataka"

Utjecaj preventivne zdravstvene zaštite na zdravlje

Ana Bagić, Tonio Ercegović, Nika Medić, Matej Škrabić

18.01.2021

Uvod

U okviru projektnog zadatka istražit ćemo zdravstvene indikatore koji spadaju u prevenciju i same zdravstvene tegobe i bolesti. Istraživanje ćemo provesti nad skupom podataka iz jedne godine za 500 američkih gradova. Za svaki grad su mjerene 4 vrste metoda preventivne zdravstvene zaštite i 12 zdravstvenih stanja ili bolesti. Kroz ovaj projekt analizirat ćemo veze između pojedinih metoda preventivne zdravstvene zaštite i zdravlja stanovništva, te usporediti rezultate u različitim gradovima.

Za početak, učitajmo podatke i potrebne pakete.

```
health = read.csv("./data_health_and_prevention.csv")
library(dplyr)
library(tidyverse)
library(ggplot2)
```

Počistimo dataframe za lakše rukovanje podatcima, tj. riješimo se nepotrebnih varijabli: X, Data_Value_Unit (uvijek %) i opredjelimo se za samo jedan postotak: AgeAdjPrv.

```
health <- health[health$DataValueTypeID == "AgeAdjPrv", c(2,3,4,5,8,9,10)]
head(health[, c(-4)])
```

```
##    StateDesc   CityName        Category Data_Value PopulationCount
## 1    Alabama Birmingham      Prevention       19.8          212237
## 3    Alabama Birmingham Health Outcomes       31.0          212237
## 5    Alabama Birmingham Health Outcomes       44.1          212237
## 7    Alabama Birmingham      Prevention       70.1          212237
## 9    Alabama Birmingham Health Outcomes        5.7          212237
## 11   Alabama Birmingham Health Outcomes       11.5          212237
##     Short_Question_Text
## 1      Health Insurance
## 3             Arthritis
## 5   High Blood Pressure
## 7  Taking BP Medication
## 9  Cancer (except skin)
## 11       Current Asthma
```

```
prevention = health[health$Category == "Prevention",]
outcomes = health[health$Category == "Health Outcomes",]
```

```
head(health$Measure, 3)
```

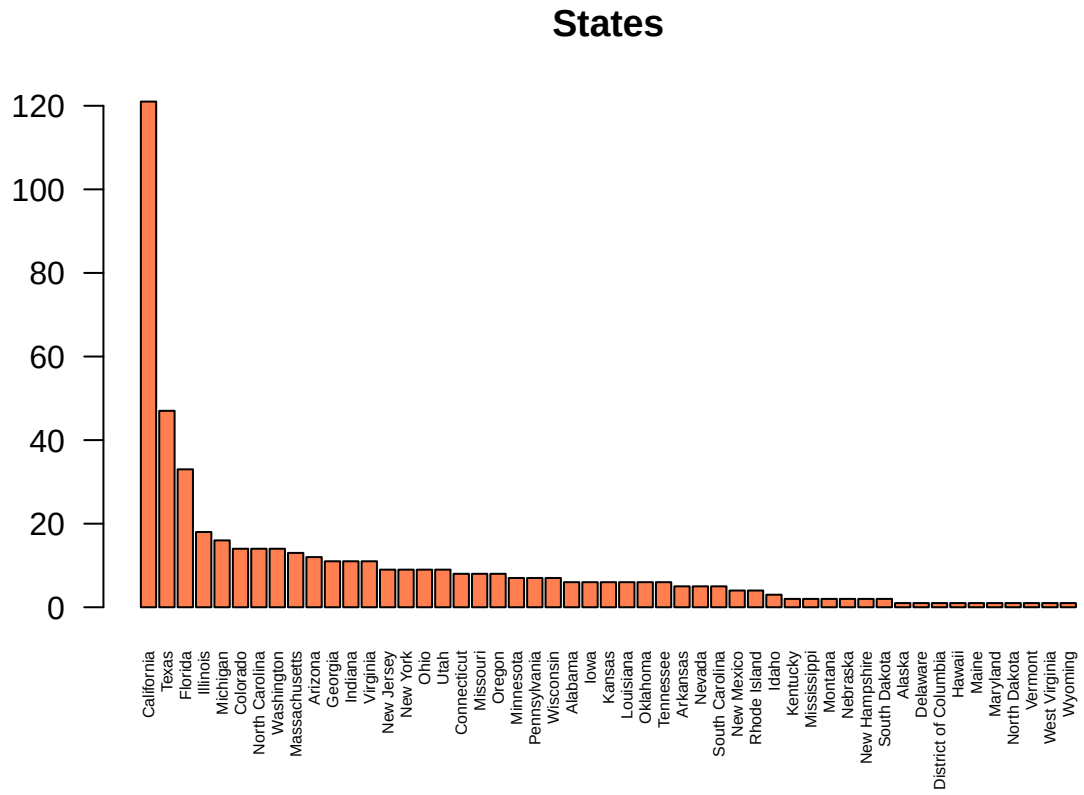
```
## [1] "Current lack of health insurance among adults aged 18â\200"64 Years"
## [2] "Arthritis among adults aged >=18 Years"
## [3] "High blood pressure among adults aged >=18 Years"
```

Uočimo da raspolažemo samo s ukupnim brojem stanovnika pojedinog grada, a udjeli stanovnika koji pate od neke bolesti ili primjenjuju neku od metoda preventivne zdravstvene zaštite dani su s starosnim ograničenjima (npr. samo stariji od 18 godina). Kako se proteže za sve gradove, metode i bolesti, prihvaćamo ovaj bias. Pretpostavljamo da nema ekstremnijih slučajeva (npr. jako velik udio djece u stanovništvu nekog grada) te smatramo da ova činjenica ne narušava previše rezultate naše analize.

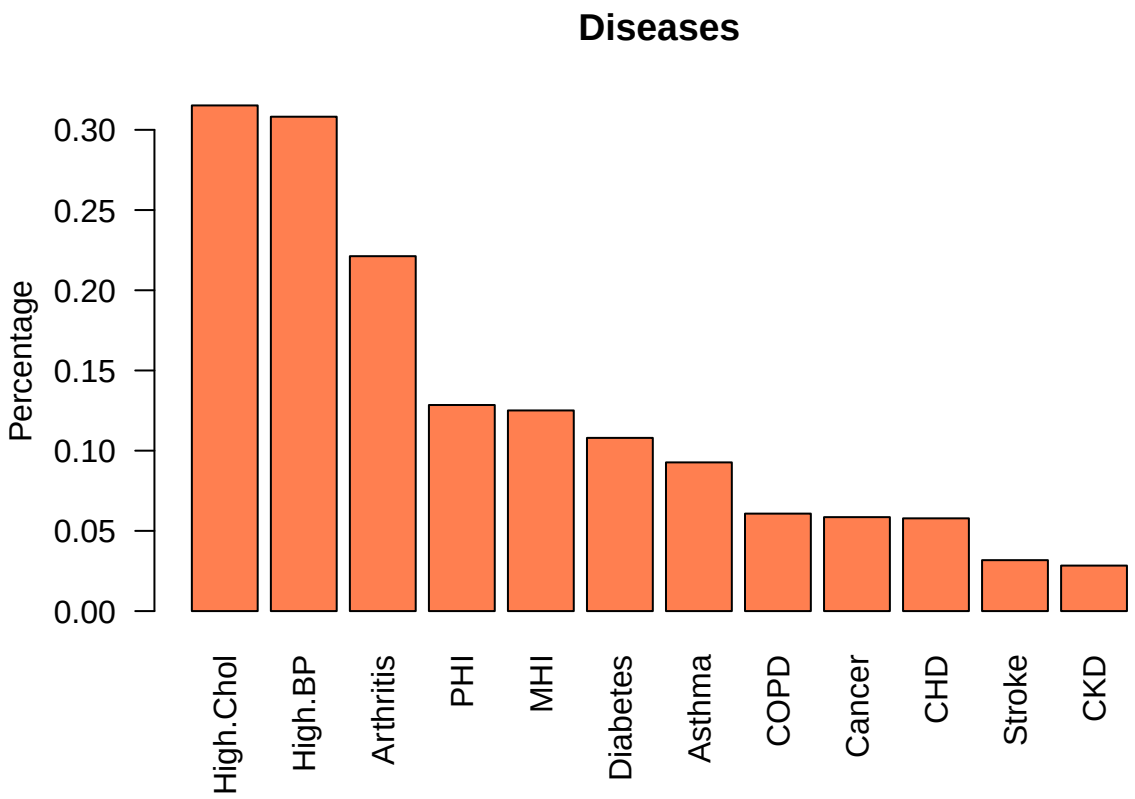
Opis podataka

Prije nego krenemo odgovarati na konkretna pitanja, upoznajmo se najprije s danim podatcima.

```
barplot(sort(table(health$StateDesc)/16, decreasing = TRUE),
        las=2,
        cex.names=.5,
        main='States',
        col = "coral")
```



Imamo najviše podataka za savezne države California, Texas i Florida, a najmanje za Wyoming, West Virginiu i Vermont.



Najviše zastupljene bolesti/zdravstvena stanja su povišeni kolesterol i krvni tlak, a najmanje je slučajeva kronične bolesti bubrega i moždanog udara.

Ohio i Florida

U ovome odjeljku ćemo pokušati odgovoriti na pitanje postoji li neka metoda preventivne zdravstvene zaštite koja je popularnija u saveznoj državi Ohio nego u saveznoj državi Florida.

Izdvojimo podatke za pojedinu saveznu državu u zasebni data frame.

```
ohio = health[health$StateDesc == "Ohio" & health$Category == "Prevention",]
florida = health[health$StateDesc == "Florida"  & health$Category == "Prevention",]
```

Veličinu uzorka za savezne države Ohio i Florida računamo na sljedeći način:

```
ohioCities = distinct(ohio[,c(2,6)], .keep_all = FALSE)
ohioPopulation = sum(ohioCities$PopulationCount)
floridaCities = distinct(florida[, c(2, 6)], .keep_all = FALSE)
floridaPopulation = sum(floridaCities$PopulationCount)
```

```
##     CityName PopulationCount
## 1      Akron          199110
## 2     Canton           73007
## 3 Cincinnati          296943
## 4  Cleveland          396815
## 5   Columbus          787033
## 6     Dayton          141527
## 7      Parma           81601
```

```
## 8     Toledo          287208
## 9 Youngstown           66982
```

```
## [1] "Ukupno stanovnika:  2330226"
```

```
##           CityName PopulationCount
## 1       Boca Raton           84392
## 2    Boynton Beach           68217
## 3       Cape Coral          154305
## 4       Clearwater          107685
## 5    Coral Springs          121096
## 6            Davie           91992
## 7  Deerfield Beach           75018
## 8          Deltona           85182
## 9  Fort Lauderdale          165521
## 10     Gainesville          124354
## 11         Hialeah          224669
## 12       Hollywood          140768
## 13    Jacksonville          821784
## 14        Lakeland           97422
## 15           Largo           77648
## 16      Lauderhill           66887
## 17       Melbourne           76068
## 18           Miami          399457
## 19     Miami Beach           87779
## 20   Miami Gardens          107167
## 21         Miramar          122041
## 22         Orlando          238300
## 23        Palm Bay          103190
## 24      Palm Coast           75180
## 25  Pembroke Pines          154750
## 26      Plantation           84955
## 27   Pompano Beach           99845
## 28  Port St. Lucie          164603
## 29  St. Petersburg          244769
## 30         Sunrise           84439
## 31     Tallahassee          181376
## 32           Tampa          335709
## 33 West Palm Beach           99919
```

```
## [1] "Ukupno stanovnika:  5166487"
```

Imamo podatke za 9 gradova iz savezne države **Ohio** koji ukupno imaju 2 330 226 stanovnika, te 33 grada iz savezne države **Florida** koji imaju 5 166 487 stanovnika.

Provest ćemo test o dvije proporcije za svaku od metoda preventivne zdravstvene zaštite. Hipoteze su sljedeće:

$$H_0 ... p_{Ohio} = p_{Florida}$$

$$H_1 ... p_{Ohio} > p_{Florida}$$

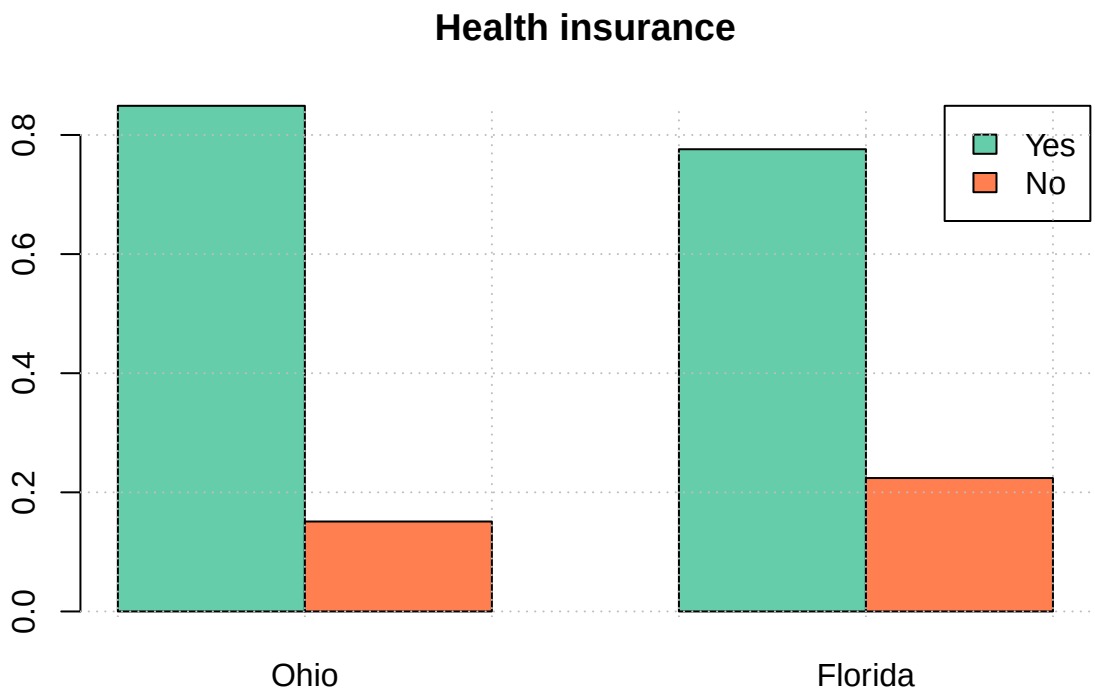
Prva metoda je zdravstveno osiguranje, ali kako su nam podatci dani kao postotak stanovništva koji *nema* zdravstveno osiguranje, najprije ćemo ga pretvoriti u postotak osiguranog stanovništva te izračunati ukupne proporcije.

```
ohio_health_insurance <- ohio[ohio$Short_Question_Text == "Health Insurance",]
ohio_health_insurance$Data_Value = 100- ohio_health_insurance$Data_Value
```

```r
k_ohio <- sum(ohio_health_insurance$Data_Value/100 * ohio_health_insurance$PopulationCount)
p_ohio <- k_ohio/ohioPopulation

florida_health_insurance <- florida[florida$Short_Question_Text == "Health Insurance",]
florida_health_insurance$Data_Value = 100 - florida_health_insurance$Data_Value
k_florida <- sum(florida_health_insurance$Data_Value/100 * florida_health_insurance$PopulationCount)
p_florida <- k_florida/floridaPopulation
```

```
## [1] "Ohio:  0.848967432772615 , Florida:  0.776000091358016"
```

Dakle, u saveznoj državi Ohio je osigurano 84.89674% stanovništva, a u saveznoj državi Florida 77.6%. Prije nego provedemo test da bismo usporedili ove uzorke, pogledajmo stupčasti dijagram kako bismo bolje vizualizirali podatke.



Provedimo sada test o dvije proporcije.

```r
prop.test(c(k_ohio, k_florida),
          c(ohioPopulation, floridaPopulation),
          alternative = "greater")
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(k_ohio, k_florida) out of c(ohioPopulation, floridaPopulation)
## X-squared = 53176, df = 1, p-value < 2.2e-16
## alternative hypothesis: greater
## 95 percent confidence interval:
```

```
##  0.07247723 1.00000000
## sample estimates:
##    prop 1    prop 2
## 0.8489674 0.7760001
```

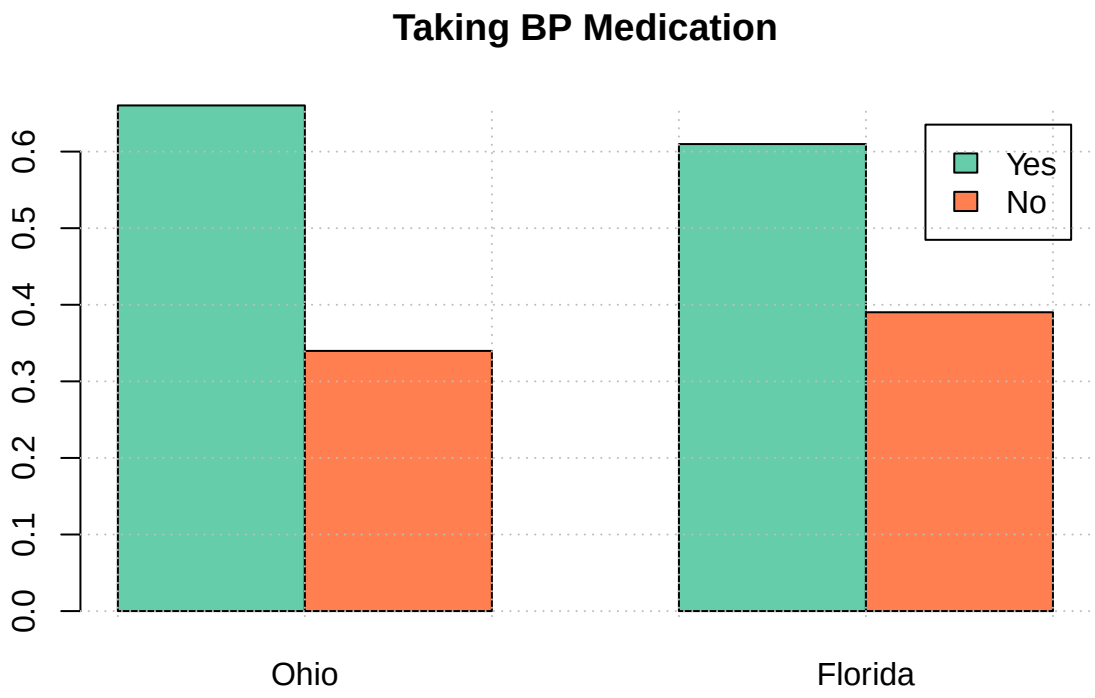
Na osnovu provedenog testa možemo odbaciti nultu hipotezu o jednakosti proporcija i zaključiti da je zdravstveno osiguranje popularnija metoda preventivne zdravstvene zaštite u saveznoj državi Ohio nego u saveznoj državi Florida.

Druga metoda je uzimanje lijekova za regulaciju krvnog tlaka. Izdvojimo potrebne podatke.

```
ohio_high_bp <- health[health$StateDesc == "Ohio"
                       & health$Short_Question_Text == "High Blood Pressure",]
ohio_high_bp$PopulationCount <- ohio_high_bp$Data_Value/100*ohio_high_bp$PopulationCount
ohio_high_bp <- ohio_high_bp[, c("CityName", "PopulationCount", "Short_Question_Text")]
ohio_total_high_bp <- sum(ohio_high_bp$PopulationCount)

ohio_bp_medication <- ohio[ohio$Short_Question_Text == "Taking BP Medication",]
ohio_bp_medication$PopulationCount <- ohio_high_bp$PopulationCount
k_ohio <- sum(ohio_bp_medication$Data_Value/100 * ohio_bp_medication$PopulationCount)

florida_high_bp <- health[health$StateDesc == "Florida"
                          & health$Short_Question_Text == "High Blood Pressure",]
florida_high_bp$PopulationCount <- florida_high_bp$Data_Value/100*florida_high_bp$PopulationCount
florida_high_bp <- florida_high_bp[, c("CityName", "PopulationCount", "Short_Question_Text")]
florida_total_high_bp <- sum(florida_high_bp$PopulationCount)

florida_bp_medication <- florida[florida$Short_Question_Text == "Taking BP Medication",]
florida_bp_medication$PopulationCount <- florida_high_bp$PopulationCount
k_florida <- sum(florida_bp_medication$Data_Value/100 * florida_bp_medication$PopulationCount)

p_ohio <- k_ohio/ohio_total_high_bp
p_florida <- k_florida/florida_total_high_bp
```

```
## [1] "Ohio:  0.660271781742056 , Florida:  0.609800961446317"
```

Sada možemo pogledati stupčasti dijagram.



U saveznoj državi Ohio 66.02718% ljudi starijih od 18 godina koji imaju povišen krvni tlak uzima lijekove za regulaciju krvnog tlaka, a u saveznoj državi Florida je taj udio 60.9801%.

Testirajmo jednakost proporcija.

```
prop.test(c(k_ohio, k_florida),
          c(ohio_total_high_bp, florida_total_high_bp),
          alternative = "greater")
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(k_ohio, k_florida) out of c(ohio_total_high_bp, florida_total_high_bp)
## X-squared = 6027.9, df = 1, p-value < 2.2e-16
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.04941161 1.00000000
## sample estimates:
##    prop 1    prop 2
## 0.6602718 0.6098010
```

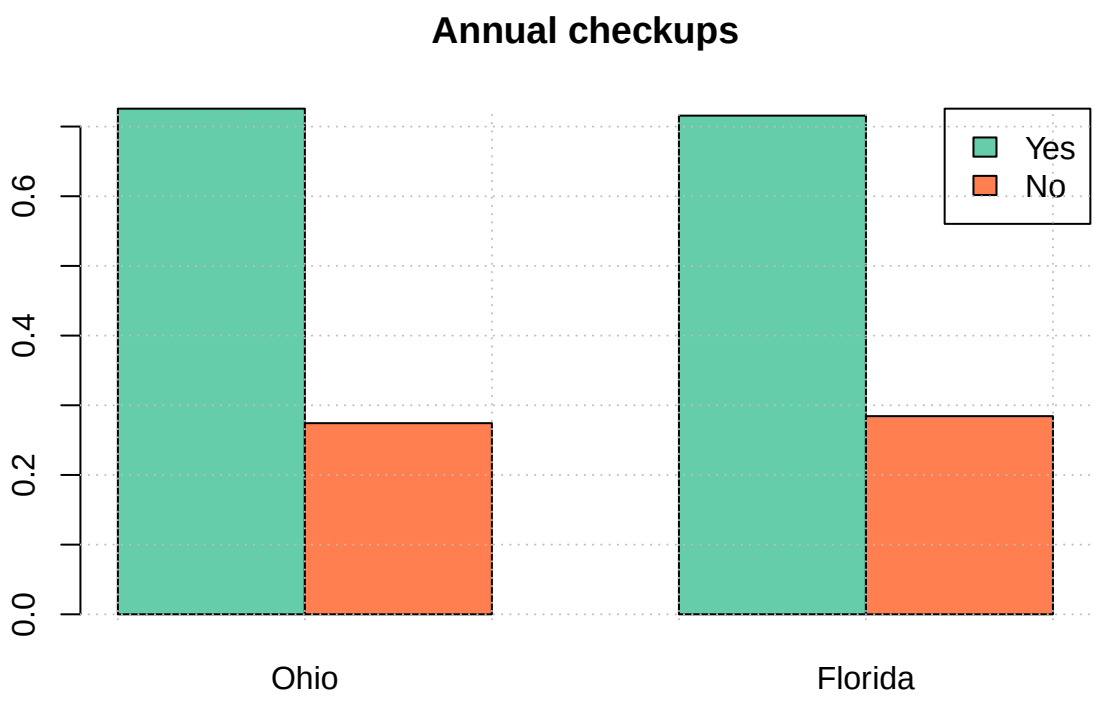
Na osnovu provedenog testa ponovno zaključujemo da je metoda popularnija u saveznoj državi Ohio nego u saveznoj državi Florida.

Sljedeća metoda je odlazak na rutinske sistematske preglede. Provedimo analizu na jednak način kao do sada.

```
ohio_checkup <- ohio[ohio$Short_Question_Text == "Annual Checkup",]
k_ohio <- sum(ohio_checkup$Data_Value/100 * ohio_checkup$PopulationCount)
p_ohio <- k_ohio/ohioPopulation
```

```
florida_checkup <- florida[florida$Short_Question_Text == "Annual Checkup",]
k_florida <- sum(florida_checkup$Data_Value/100 * florida_checkup$PopulationCount)
p_florida <- k_florida/floridaPopulation
```

```
## [1] "Ohio:  0.725691213212796 , Florida:  0.715632734389925"
```



72.56912% stanovništva savezne države Ohio i 71.56326% stanovništva savezne države Florida odlazi na rutinske sistematske preglede. Provedimo test o dvije proporcije.

```
prop.test(c(k_ohio, k_florida),
          c(ohioPopulation, floridaPopulation),
          alternative = "greater")
```

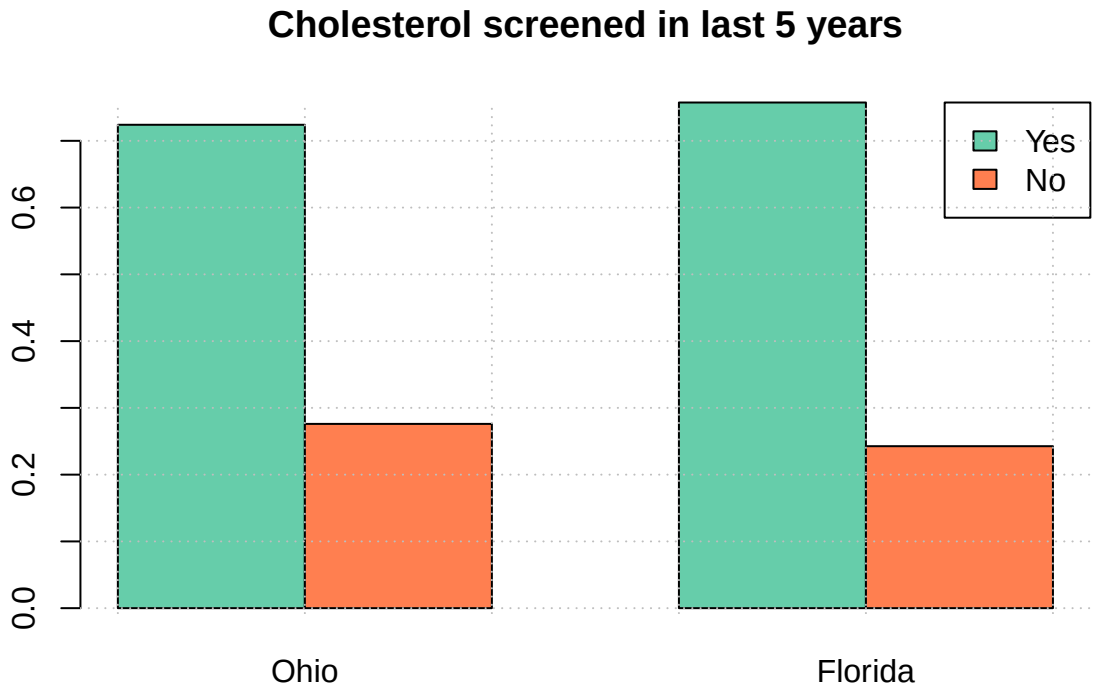
```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(k_ohio, k_florida) out of c(ohioPopulation, floridaPopulation)
## X-squared = 803.71, df = 1, p-value < 2.2e-16
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.009477053 1.000000000
## sample estimates:
##    prop 1    prop 2
## 0.7256912 0.7156327
```

Na osnovu provedenog testa zaključujemo da je metoda popularnija u saveznoj državi Ohio nego u saveznoj državi Florida.

Posljednja metoda preventivne zdravstvene zaštite je kontrola kolesterola. Vizualizirajmo podatke kao i do sada.

```
ohio_cholesterol = ohio[ohio$Short_Question_Text == "Cholesterol Screening",]
k_ohio <- sum(ohio_cholesterol$Data_Value/100 * ohio_cholesterol$PopulationCount)
p_ohio <- k_ohio/ohioPopulation

florida_cholesterol <- florida[florida$Short_Question_Text == "Cholesterol Screening",]
k_florida <- sum(florida_cholesterol$Data_Value/100 * florida_cholesterol$PopulationCount)
p_forida <- k_florida/floridaPopulation
```

```
## [1] "Ohio:  0.724016377381421 , Florida:  0.715632734389925"
```



U saveznoj državi Ohio 72.40164% stanovništva kontrolira svoj kolesterol, a u saveznoj državi Florida je taj udio 75.3788%. Testirajmo proporcije na jednak način kao i za prethodne metode.

```
prop.test(c(k_ohio, k_florida),
          c(ohioPopulation, floridaPopulation),
          alternative = "greater")
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(k_ohio, k_florida) out of c(ohioPopulation, floridaPopulation)
## X-squared = 9463.4, df = 1, p-value = 1
## alternative hypothesis: greater
## 95 percent confidence interval:
##  -0.03394484  1.00000000
```

```
## sample estimates:
##    prop 1    prop 2
## 0.7240164 0.7573880
```

Na osnovu provedenog testa ne možemo odbaciti nultu hipotezu o jednakosti proporcija. Štoviše, test snažno upućuje da je metoda zapravo popularnija u saveznoj državi Florida, što nam se i čini očitim iz ukupnih udjela, pa se možemo i uvjeriti u to.

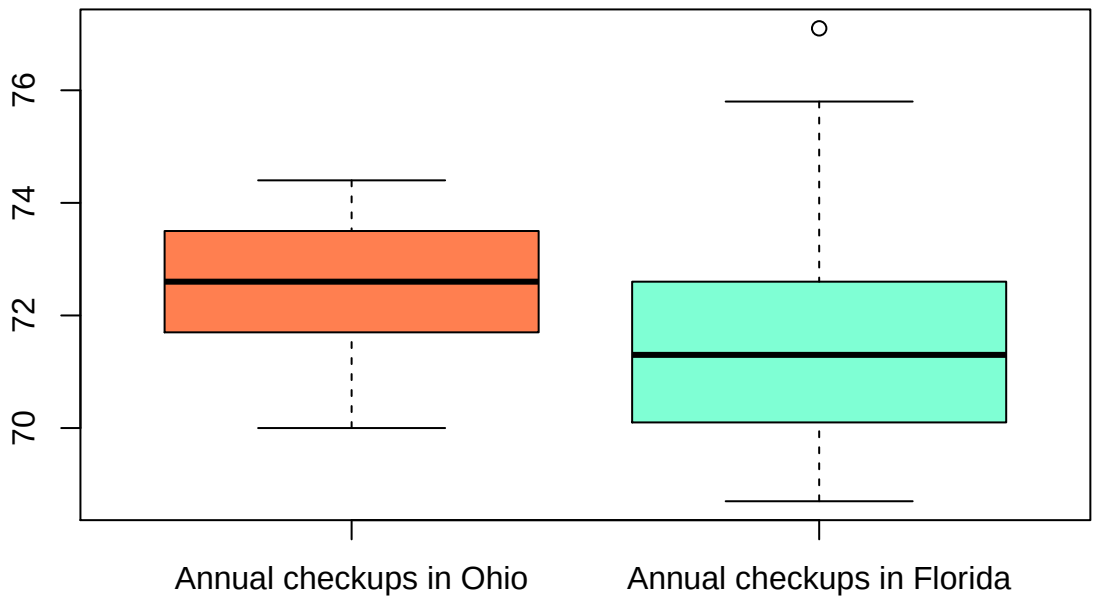
```
prop.test(c(k_ohio, k_florida),
          c(ohioPopulation, floridaPopulation),
          alternative = "less")
```

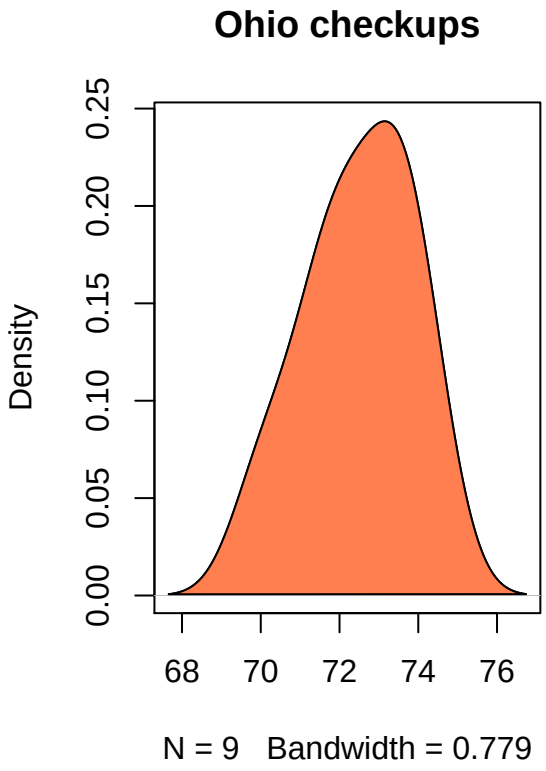
```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(k_ohio, k_florida) out of c(ohioPopulation, floridaPopulation)
## X-squared = 9463.4, df = 1, p-value < 2.2e-16
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.0000000 -0.0327984
## sample estimates:
##    prop 1    prop 2
## 0.7240164 0.7573880
```

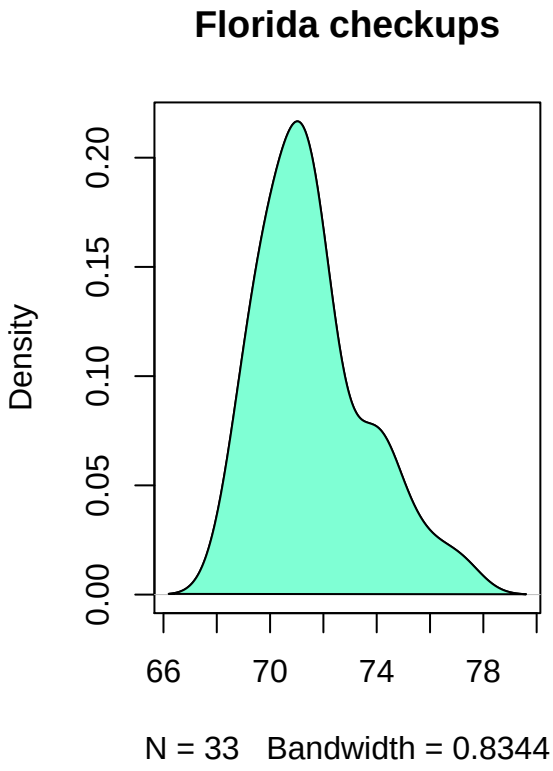
Možemo reći da su rezultati očekivani, s obzirom na veličinu uzorka čak su i relativno male razlike u proporciji(kao npr. kod sistematskih pregleda), koje nam možda kontekstualno nisu značajne, ipak statistički značajne.

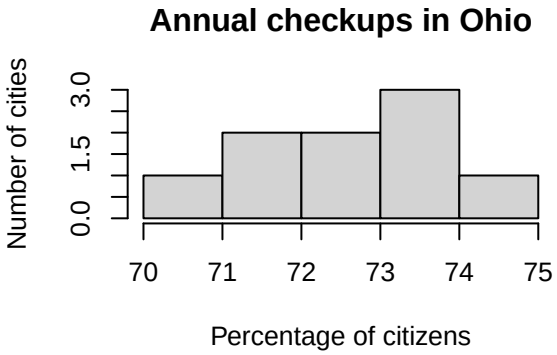
Analizi ovog problema mogli smo pristupiti i na malo drugačiji način - mogli smo provjeriti jednakost srednjih vrijednosti postotaka stanovništva koji primjenjuju neku metodu po gradovima saveznih država Ohio i Florida. Pritom valja napomenuti da bismo time odgovorili na nešto drugačije pitanje: u prethodnoj analizi uspoređivali smo ukupne proporcije u saveznim državama, dok bi sada uspoređivali je li prosječni udio stanovnika koji primjenjuje neku metodu jednak za gradove u saveznoj državi Ohio, odnosno Florida. Promotrimo za primjer sistematske preglede.

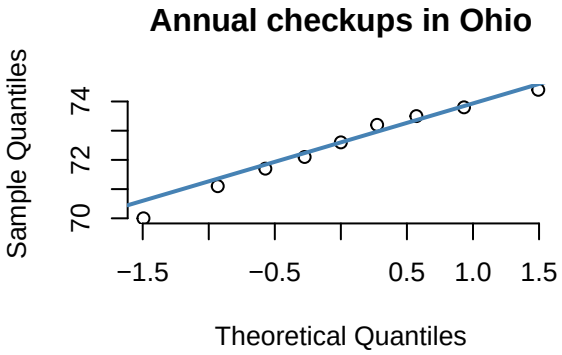
Sada podatke promatramo kao metričke, te želimo provesti t-test o jednakosti srednjih vrijednosti. Pretpostavke testa su *nezavisnost* i *normalnost* podataka. Nezavisnost možemo pretpostaviti s obzirom da se podatci odnose na različite savezne države, a normalnost ćemo provjeriti u nastavku. Pogledajmo najprije box-and-whiskers plot, histograme i QQ-plot.

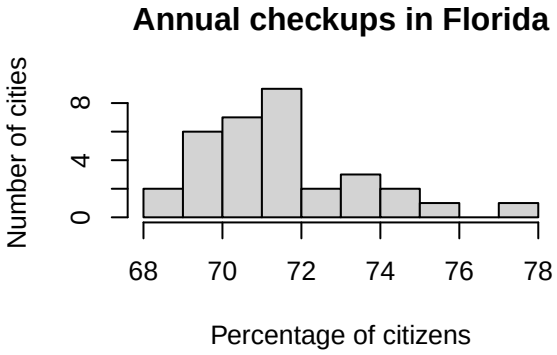

76
74
72
70
Annual checkups in Ohio
Annual checkups in Florida

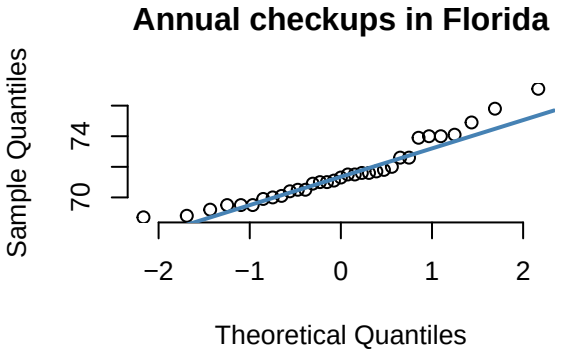

Ohio checkups
Density
0.00
0.05
0.10
0.15
0.20
0.25
68
70
72
74
76
N = 9 Bandwidth = 0.779


Florida checkups
Density
0.00
0.05
0.10
0.15
0.20
66
70
74
78
N = 33 Bandwidth = 0.8344









Razdiobe odstupaju od normalne, ali nisu previše zakrivljene ili nepravilne. Provedimo sada Lillieforsovu inačicu Kolmogorov-Smirnovljevevog testa.

```
lillie.test(ohio_checkup$Data_Value)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  ohio_checkup$Data_Value
## D = 0.13797, p-value = 0.8908
```

```
lillie.test(florida_checkup$Data_Value)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  florida_checkup$Data_Value
## D = 0.15663, p-value = 0.03877
```

Iako uz uobičajenu razinu sigurnosti od $\alpha = 0.05$ test odbacuje pretpostavku normalnosti za podatke iz Floride, p-vrijednost je skoro 4% te imajući na umu robusnost t-testa na normalnost, zaključujemo da možemo pretpostaviti normalnost podataka. Provjerimo sada jednakost varijanci.

```
var.test(ohio_checkup$Data_Value, florida_checkup$Data_Value)
```

```
##
##  F test to compare two variances
##
## data:  ohio_checkup$Data_Value and florida_checkup$Data_Value
```

```
## F = 0.47754, num df = 8, denom df = 32, p-value = 0.2743
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.1822572 1.8531292
## sample estimates:
## ratio of variances
##          0.4775422
```

Na osnovu p-vrijednosti od 0.2743 ne odbacujemo nultu hipotezu te konačno možemo provesti t-test s pretpostavkom o jednakosti varijanci. Uzmimo razinu značajnosti $\alpha = 0.05$.

```
t.test(ohio_checkup$Data_Value, florida_checkup$Data_Value, alternative = "greater", var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  ohio_checkup$Data_Value and florida_checkup$Data_Value
## t = 1.2363, df = 40, p-value = 0.1118
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.3239973        Inf
## sample estimates:
## mean of x mean of y
##  72.48889  71.59394
```

Na osnovu ovog testa ne možemo odbaciti nultu hipotezu o jednakosti srednjih vrijednosti. Uz razinu značajnosti od 5% zaključujemo da je prosječni udio stanovništva koji ide na redovne sistematske preglede jednak za gradove saveznih država Ohio i Florida. Kao neparametarsku alternativu ovome testu, mogli smo provesti Mann-Whitney-Wilcoxonov test koji je slabiji, ali ne zahtjeva normalnost podataka.(izračunata p-vrijednost će biti aproksimativna jer postoje "duplikati")

```
wilcox.test(ohio_checkup$Data_Value, florida_checkup$Data_Value, alternative='greater')
```

```
## Warning in wilcox.test.default(ohio_checkup$Data_Value,
## florida_checkup$Data_Value, : cannot compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  ohio_checkup$Data_Value and florida_checkup$Data_Value
## W = 201.5, p-value = 0.05366
## alternative hypothesis: true location shift is greater than 0
```

p-vrijednost je očekivano manja nego kod t-testa, ali i dalje veća od 5% pa opet ne bismo odbacili nultu hipotezu o jednakosti srednjih vrijednosti.

Kronične plućne bolesti - astma i COPD

U uvome dijelu izabrat ćemo 3 savezne države i usporediti njihove proporcije stanovništva koje boluje od kroničnih plućnih bolesti. Tri države koje smo izabrali su: **Arizona**, **Colorado** i **Utah**. Najprije ćemo se upoznati s promatranim podacima te promotriti kontigencijsku tablicu za ove savezne države.

Kako test homogenosti zahtjeva da zbrojevi redaka ili stupaca budu unaprijed zadani, uzmimo uzorak od 500 000 ljudi iz svake od tri savezne države.

Pogledajmo najprije podatke za kroničnu opstruktivnu bolest pluća.

```
arizona = outcomes[outcomes$StateDesc == "Arizona" & outcomes$Short_Question_Text == "COPD",]
colorado = outcomes[outcomes$StateDesc == "Colorado" & outcomes$Short_Question_Text == "COPD",]
```

```r
utah = outcomes[outcomes$StateDesc == "Utah" & outcomes$Short_Question_Text == "COPD",]

n <- 500000

arizona_population <- sum(arizona$PopulationCount)
colorado_population <- sum(colorado$PopulationCount)
utah_population <- sum(utah$PopulationCount)

arizona_COPD <- sum(arizona$PopulationCount * arizona$Data_Value/100)
colorado_COPD <- sum(colorado$PopulationCount * colorado$Data_Value/100)
utah_COPD <- sum(utah$PopulationCount * utah$Data_Value/100)

arizona_sample <- rbinom(n, 1, arizona_COPD/arizona_population)
colorado_sample <- rbinom(n, 1, colorado_COPD/colorado_population)
utah_sample <- rbinom(n, 1, utah_COPD/utah_population)

tmp <- c(sum(arizona_sample),
         n-sum(arizona_sample),
         sum(colorado_sample),
         n-sum(colorado_sample),
         sum(utah_sample),
         n-sum(utah_sample))
ctablica <- matrix(tmp,ncol=3)
colnames(ctablica) <- c("Arizona","Colorado","Utah")
rownames(ctablica) <- c("COPD","bez COPD")
ctablica = as.table(ctablica)
ctablica
```

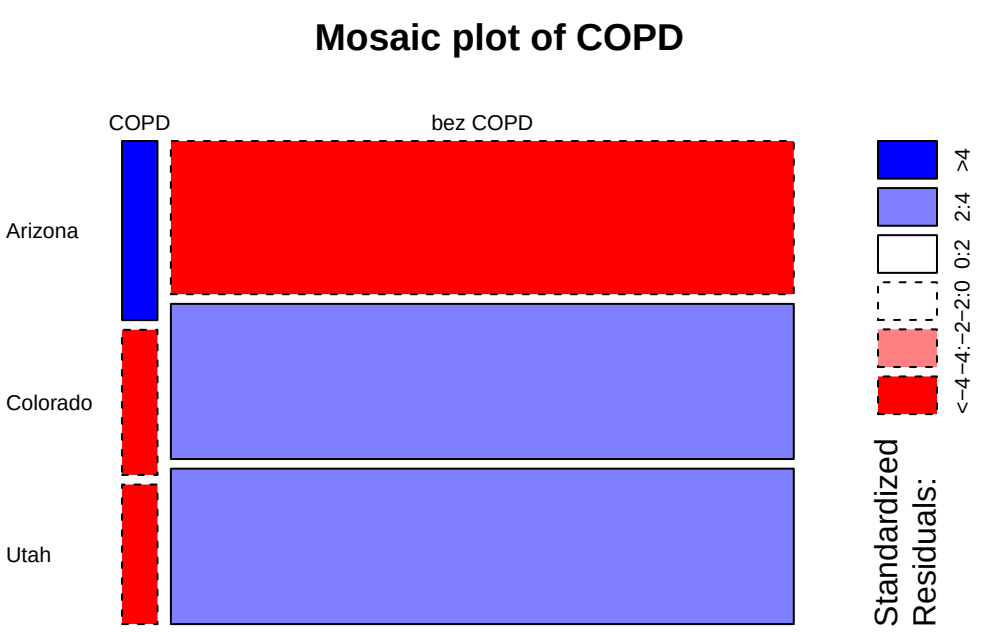
```
##          Arizona Colorado   Utah
## COPD       31179    25246  24287
## bez COPD  468821   474754 475713
```

Sljedeći graf prikazuje tu tablicu i odskakanje njenih vrijednosti od očekivanih.

Visina pravokutnika koji označavaju jednu vrijednost je proporcionalna udjelu broja stanovnika koji imaju/nemaju COPD, dok širina prikazuje udio broja stanovnika u nekoj saveznoj državi. Bijela boja znači da je vrijednost prilično jednaka očekivanoj, plava da je vrijednost veća te crvena da je vrijednost manja.

Odmah možemo primjetiti da na grafu postoje samo vrijednosti označene plavom ili crvenom bojom. To sugerira jako odstupanje promatranih vrijednosti od očekivanih te već sada možemo naslutiti rezultate testa o proporcijama tih saveznih država. Baratamo s velikim uzorcima i sigurni smo da ako postoje stvarne razlike u udjelima stanovnika koji boluju od COPD-a u te 3 savezne države da će test o homogenosti to i pokazati.

```r
dt <- as.table(as.matrix(ctablica))
library("graphics")
mosaicplot(dt,
           shade=TRUE,
           las=1,
           main="Mosaic plot of COPD")
```



Za nultu hipotezu uzimamo da su proporcije države jednake, a za alternativnu hipotezu da se bar jedna proporcija razlikuje. Sada ćemo provesti test za homogenost nad već ispisanom tablicom i ispisati promatrane i očekivane vrijednosti te komentirati njihove razlike.

$$H_0 ... p_{Arizona} = p_{Colorado} = p_{Utah}$$

$$H_1 ...\text{bar jedna proprocija nije jednaka}$$

```
ctest <- chisq.test(ctablica)
ctest
```

```
##
##  Pearson's Chi-squared test
##
## data:  ctablica
## X-squared = 1094.9, df = 2, p-value < 2.2e-16
```

```
##
## Promatrane vrijednosti:
```

```
##          Arizona Colorado   Utah
## COPD       31179    25246  24287
## bez COPD  468821   474754 475713
```

```
##
## Očekivane vrijednosti:
```

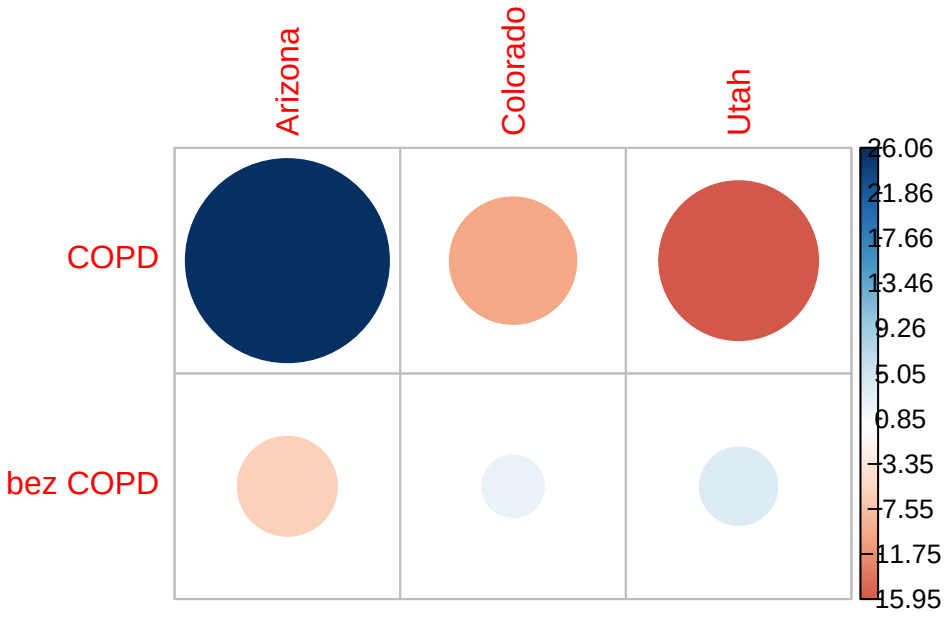
```
##          Arizona Colorado   Utah
## COPD       26904    26904  26904
```

```
## bez COPD  473096   473096 473096
```

Na osnovi testa odbacujemo nultu hipotezu i prihvaćamo da udjeli stanovnika koji boluju od COPD-a u saveznim državama Arizona, Colorado i Utah nisu isti. Usporedbom očekivanih vrijednosti vidimo da Arizona ima veći udio bolesnika od očekivanog dok Colorado i Utah imaju manji.

Pogledajmo još jedan grafički prikaz koji pruža bolje objašnjenje utjecaja odstupanja promatranih vrijednosti.

```
corrplot(ctest$residuals, is.cor = FALSE)
```



Polja matrice obojana plavom bojom označavaju veći udio od očekivanog, dok ona obojana crvenom označavaju manji udio od očekivanog. Iz grafa vidimo visoko odstupanje u razlici broja bolesnih kod Arizone i Utah te nešto manje u saveznoj državi Colorado. Ovime smo grafički pokazali rezultate koje je pokazao prethodno provedeni test.

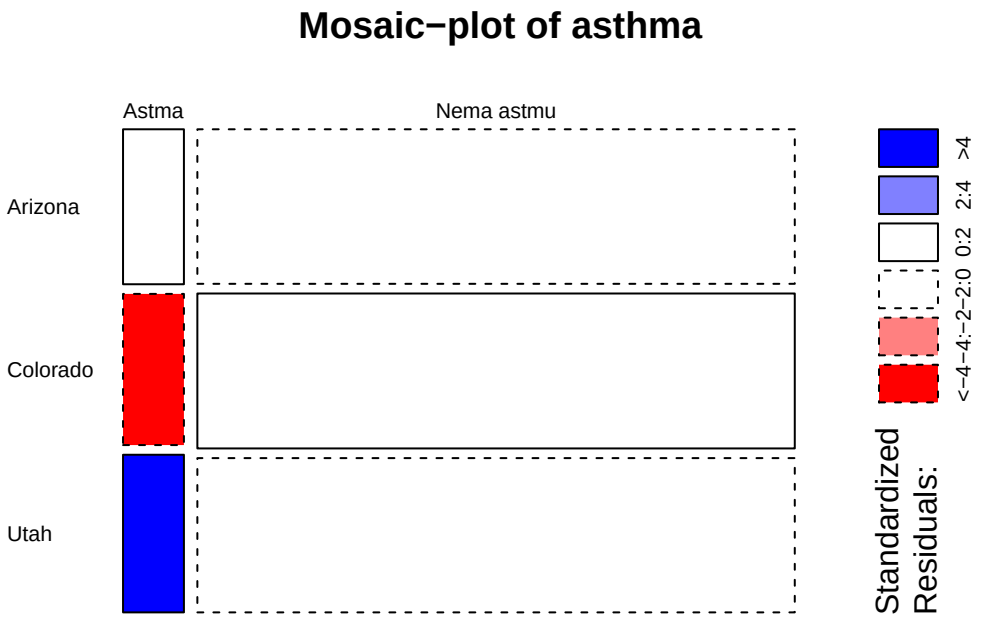
Provedimo istu analizu i za astmu.

```
arizona = outcomes[outcomes$StateDesc == "Arizona"
                   & outcomes$Short_Question_Text == "Current Asthma",]
colorado = outcomes[outcomes$StateDesc == "Colorado"
                    & outcomes$Short_Question_Text == "Current Asthma",]
utah = outcomes[outcomes$StateDesc == "Utah"
                & outcomes$Short_Question_Text == "Current Asthma",]

arizona_asthma <- sum(arizona$PopulationCount * arizona$Data_Value/100)
colorado_asthma <- sum(colorado$PopulationCount * colorado$Data_Value/100)
utah_asthma <- sum(utah$PopulationCount * utah$Data_Value/100)

arizona_sample <- rbinom(n, 1, arizona_asthma/arizona_population)
```

```
colorado_sample <- rbinom(n, 1, colorado_asthma/colorado_population)
utah_sample <- rbinom(n, 1, utah_asthma/utah_population)

tmp <- c(sum(arizona_sample),
         n-sum(arizona_sample),
         sum(colorado_sample),
         n-sum(colorado_sample),
         sum(utah_sample),
         n-sum(utah_sample))
ctablica <- matrix(tmp,ncol=3)
colnames(ctablica) <- c("Arizona","Colorado","Utah")
rownames(ctablica) <- c("Astma","Nema astmu")
ctablica = as.table(ctablica)
ctablica
```

```
##            Arizona Colorado   Utah
## Astma        46348    45211  47148
## Nema astmu  453652   454789 452852
```

```
dt <- as.table(as.matrix(ctablica))
library("graphics")
mosaicplot(dt,
           shade=TRUE,
           las=1,
           main="Mosaic-plot of asthma")
```



Ovaj uzorak stanovnika, zaključujući prema grafu, bliži je homogenosti nego u prethodnom slučaju. Vidimo da

su opažene vrijednosti za saveznu državu Arizona vrlo bliske očekivanima. Provedimo sada test o homogenosti.

```
ctest <- chisq.test(ctablica)
ctest
```

```
##
##  Pearson's Chi-squared test
##
## data:  ctablica
## X-squared = 45.16, df = 2, p-value = 1.562e-10
```

```
##
## Promatrane vrijednosti:
```

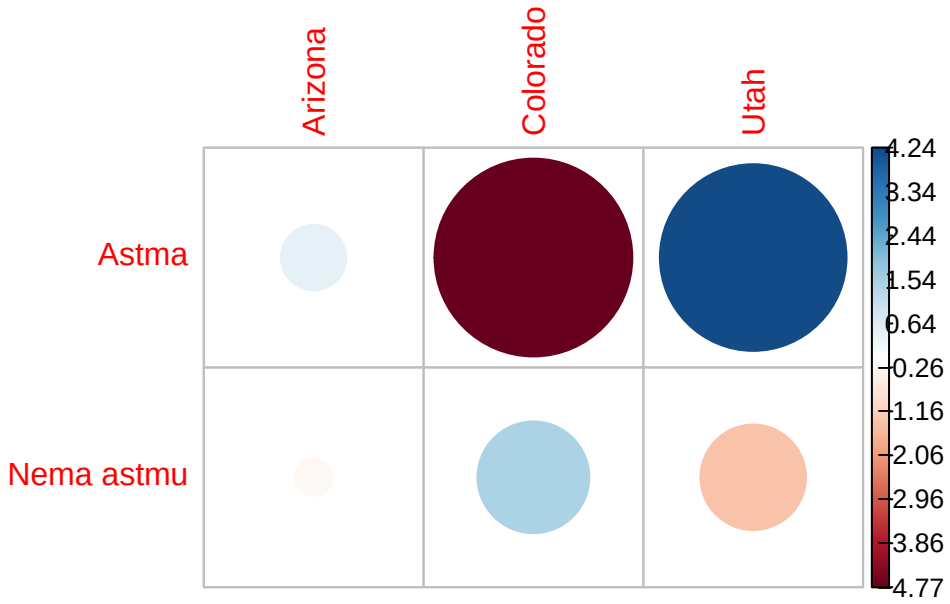
```
##            Arizona Colorado   Utah
## Astma        46348    45211  47148
## Nema astmu  453652   454789 452852
```

```
##
## Očekivane vrijednosti:
```

```
##              Arizona  Colorado      Utah
## Astma       46235.67  46235.67  46235.67
## Nema astmu 453764.33 453764.33 453764.33
```

p-vrijednost je i dalje izrazito mala i odbacujemo pretpostavku o homogenosti proporcija.

```
corrplot(ctest$residuals, is.cor = FALSE)
```



I ovaj graf nam potvrđuje što smo prethodno zaključili - proporcije u Arizoni su u skladu s očekivanima, ali

za Colorado i Utah postoje poprilična odstupanja.

Veze između metoda preventivne zaštite i bolesti

Podsjetimo se, raspolažemo podatcima za 4 metode preventivne zdravstvene zaštite: zdravstveno osiguranje, uzimanje lijekova za regulaciju krvnog tlaka, redovni sistematski pregledi i kontrola kolesterola te 12 bolesti odnosno zdravstvenih tegoba: artritis, povišeni krvni tlak, rak, astma, koronarna bolest srca, kronična opstruktivna bolest pluća, dijabetes, povišeni kolesterol, kronična bolest bubrega, produljeni problemi s mentalnim zdravljem, produljeni problemi s fizičkim zdravljem te moždani udar.

Prije provođenja ikakvih testova, od ponuđenih podataka, očekujemo najjaču zavisnost između uzimanja lijekova za regulaciju krvnog tlaka i udjela stanovništva koji imaju problema s krvnim tlakom te između udjela ljudi koji su pregledali kolesterol i udjela stanovništva s povišenim kolesterolom. Zdravstveno osiguranje i redovni sistematski pregledi su "općenitije" metode zdravstvene zaštite pa nam se unaprijed ne čini da će imati posebni utjecaj na neku određenu bolest već će doprinositi relativno manji, podjednak utjecaj na sve bolesti.

Za početak, formatirajmo podatke u prikladniji oblik za predstojeću analizu.

```
health_grouped <- health %>% group_by(StateDesc, CityName) %>% ungroup
health_grouped <- health_grouped[, c(1,2,5, 6, 7)]
health_overview <- data.frame(health_grouped)
health_overview <- reshape(health_overview,
                           idvar=c("StateDesc", "CityName", "PopulationCount"),
                           timevar = "Short_Question_Text",
                           direction="wide")

health_overview <- health_overview %>% rename(
  "Health.Insurance" = `Data_Value.Health Insurance`,
  "Arthritis" = Data_Value.Arthritis,
  "High.Blood.Pressure" = `Data_Value.High Blood Pressure`,
  "Taking.BP.Medication" = `Data_Value.Taking BP Medication`,
  "Cancer" = `Data_Value.Cancer (except skin)`,
  "Asthma" = `Data_Value.Current Asthma`,
  "Coronary.Heart.Disease" = `Data_Value.Coronary Heart Disease`,
  "Annual.Checkup" = `Data_Value.Annual Checkup`,
  "Cholesterol.Screening" = `Data_Value.Cholesterol Screening`,
  "COPD" = Data_Value.COPD,
  "Diabetes" = Data_Value.Diabetes,
  "High.Cholesterol" = `Data_Value.High Cholesterol`,
  "Chronic.Kidney.Disease" = `Data_Value.Chronic Kidney Disease`,
  "Mental.Health.Issues" = `Data_Value.Mental Health`,
  "Physical.Health.Issues" = `Data_Value.Physical Health`,
  "Stroke" = Data_Value.Stroke,
)
health_overview$`Health.Insurance` <- 100-health_overview$`Health.Insurance`
colorder <- c(1,2,3, 4, 7, 11, 12, 5, 9, 8, 16, 10, 13, 14, 6, 15, 17, 18, 19)
health_overview <- health_overview[, colorder]
head(health_overview, 1)
```

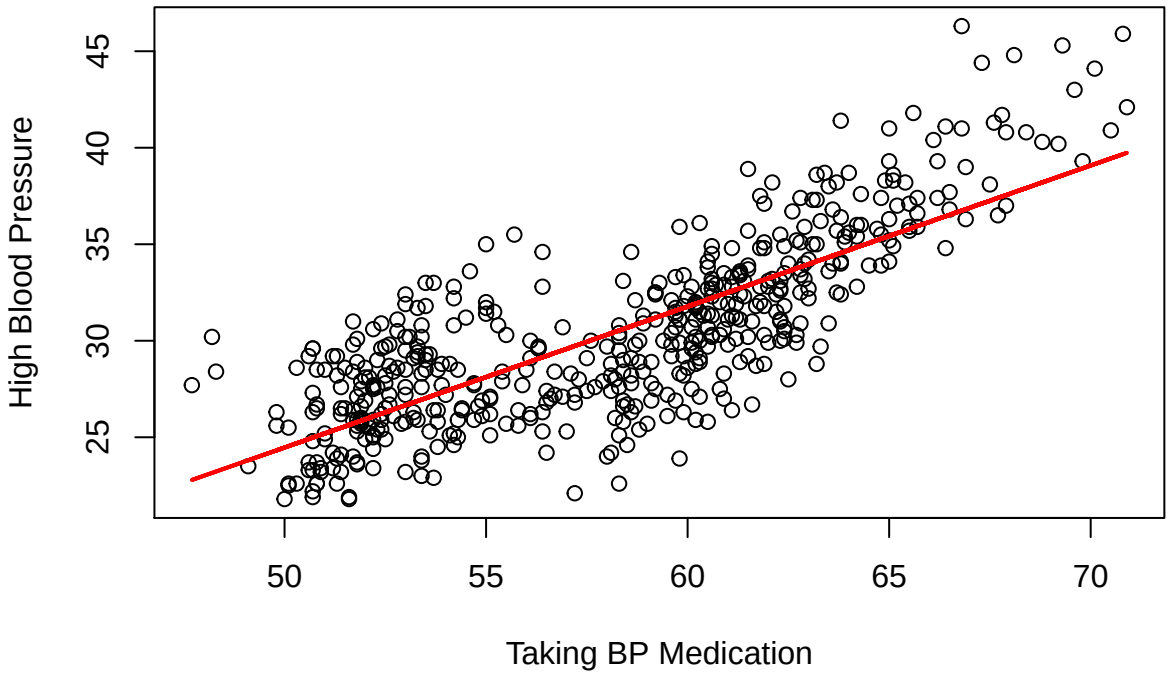
```
##   StateDesc   CityName PopulationCount Health.Insurance Taking.BP.Medication
## 1   Alabama Birmingham          212237             80.2                 70.1
##   Annual.Checkup Cholesterol.Screening Arthritis Asthma Cancer
## 1           76.8                  75.8        31   11.5    5.7
##   Chronic.Kidney.Disease Coronary.Heart.Disease COPD Diabetes
## 1                    3.5                   7.6    9     16.8
```

```
##   High.Blood.Pressure High.Cholesterol Mental.Health.Issues
## 1                44.1             35.3                 15.6
##   Physical.Health.Issues Stroke
## 1                   16.4    5.2
```

Krenimo redom te promotrimo povezanost između uzimanja lijekova za regulaciju tlaka i broja ljudi koji imaju problema s krvnim tlakom.

Kako zasada promatramo utjecaj samo jedne nezavisne varijable (uzimanje lijekova) na zavisnu varijablu (udio stanovnika s povišenim krvnim tlakom), za vizualizaciju će nam vrlo dobro poslužiti scatter plot.

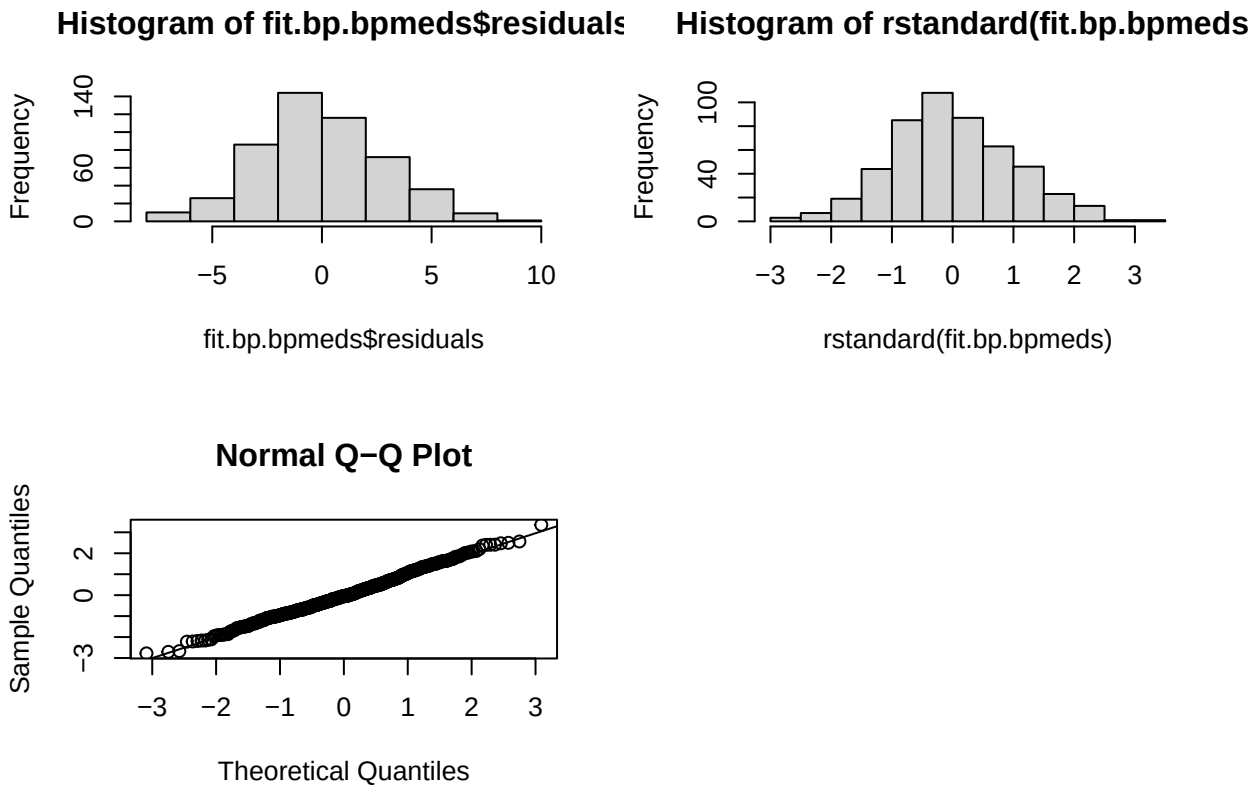
```
plot(health_overview$Taking.BP.Medication,
     health_overview$High.Blood.Pressure,
     xlab = "Taking BP Medication",
     ylab = "High Blood Pressure")
fit.bp.bpmeds = lm(High.Blood.Pressure~Taking.BP.Medication,
                   data = health_overview)
lines(health_overview$Taking.BP.Medication,fit.bp.bpmeds$fitted.values,
      col='red',
      lwd = 2)
```



Na prvi pogled možda se ovakav graf čini iznenađujućim, ali kad bolje promislimo "povišen tlak" je relativan pojam te je izgledno da dio ljudi koji imaju samo blago povišen tlak neće piti lijekove, ali oni s izrazito visokim tlakom sigurno hoće. Logično je da je u gradovima gdje ima općenito više ljudi s povišenim tlakom, vrlo vjerojatno veći i broj ljudi s jako visokim tlakom pa ovakav odnos te dvije varijable ima smisla. Prije nego detaljnije pogledamo ovaj model, provjerimo pretpostavke modela: *normalnost* i *homoskedastičnost* reziduala.

Slijede histogrami reziduala i standardiziranih reziduala te qq-plot standardiziranih reziduala.

```
par(mfrow=c(2,2))
hist(fit.bp.bpmeds$residuals)
hist(rstandard(fit.bp.bpmeds))
qqnorm(rstandard(fit.bp.bpmeds))
qqline(rstandard(fit.bp.bpmeds))
```



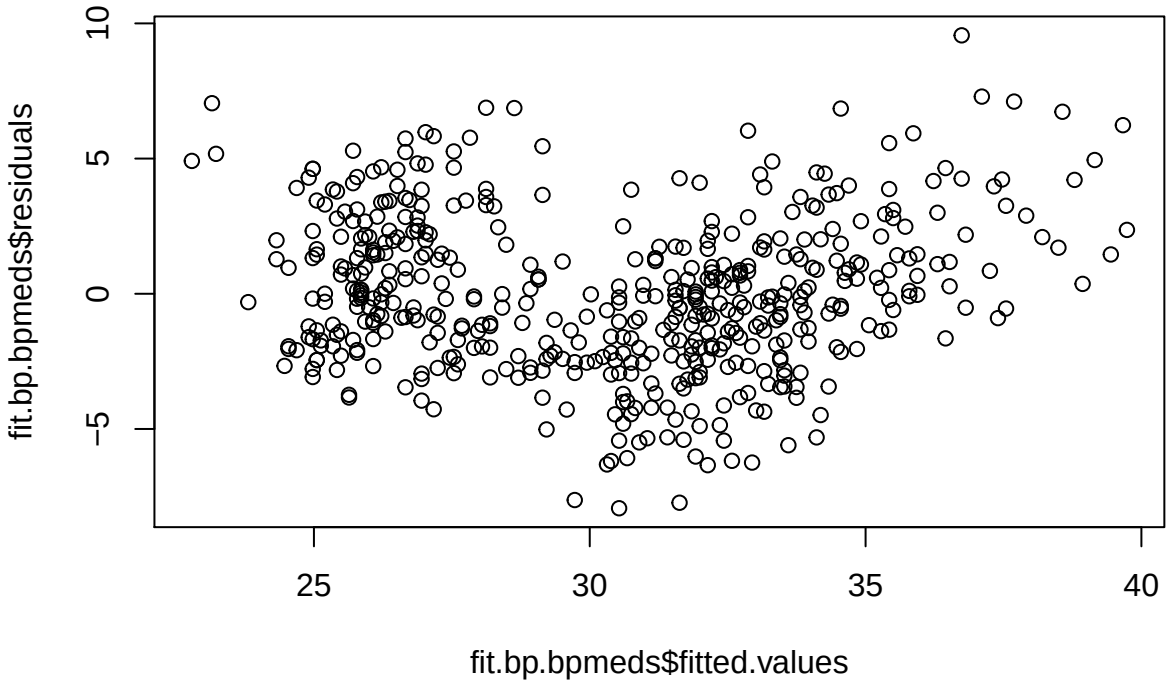
Svi priloženi grafovi ukazuju na normalnost reziduala, ali možemo i provesti recimo Lillieforsov test nad njima da se u to uvjerimo:

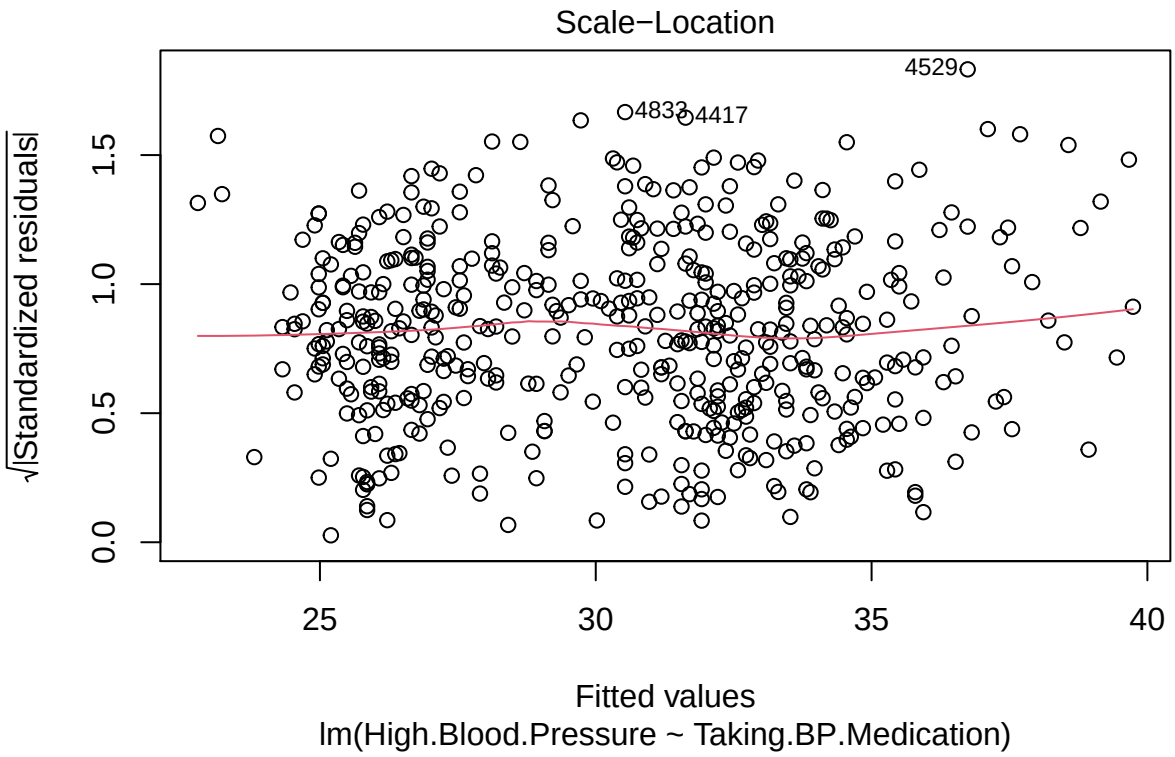
```
lillie.test(rstandard(fit.bp.bpmeds))
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rstandard(fit.bp.bpmeds)
## D = 0.033965, p-value = 0.1738
```

p-vrijednost od 0.1738 nam potvrđuje ono što smo i očekivali: ne možemo odbaciti nultu hipotezu tj. možemo pretpostaviti normalnost reziduala.

Sada ostaje pokazati homogenost varijance reziduala, tj. reziduali se ne bi smjeli "širiti" s povećanjem $\hat{y}$. Za to su korisni sljedeći scatter-plotovi:


fit.bp.bpmeds$residuals
10
5
0
−5
25
30
35
40
fit.bp.bpmeds$fitted.values



Kada promotrimo reziduale u ovisnosti o procijenjenim vrijednostima na prvom grafu, vidimo naznake heteroskedastičnosti - za najveće vrijednosti reziduali su uglavnom veći od nule. Međutim, pogledamo li drugi graf na kojem su apsolutne vrijednosti standardiziranih reziduala, vidimo da je stanje ipak prihvatljivo pa zaključujemo da su pretpostavke modela zadovoljene.

```
summary(fit.bp.bpmeds)
```

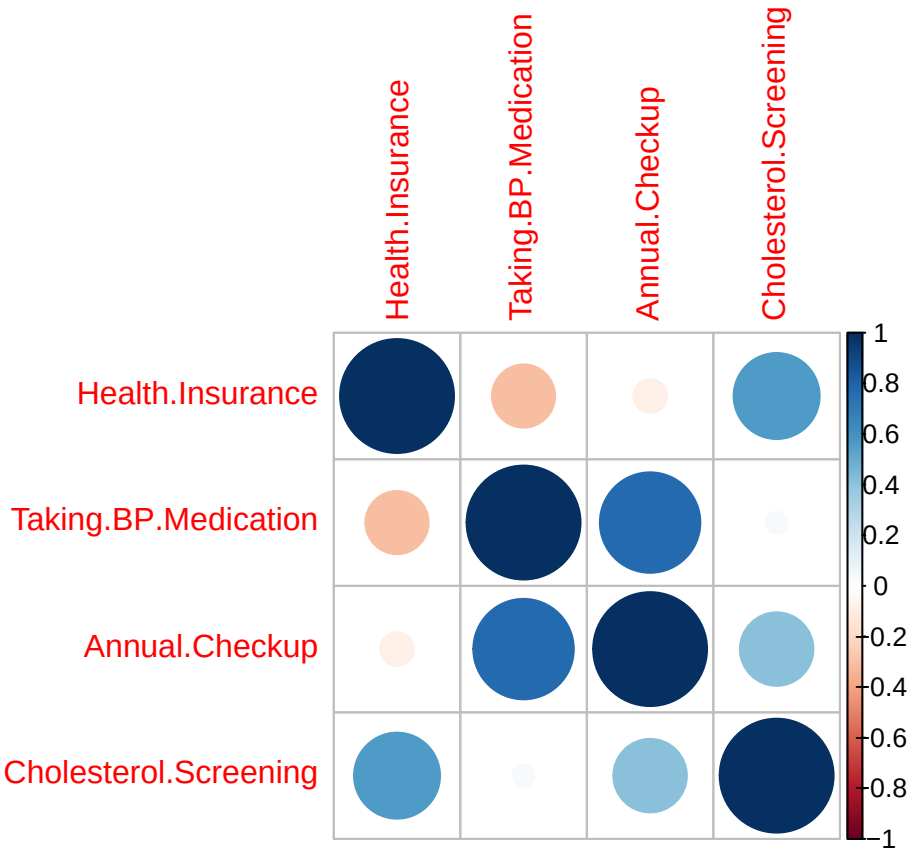
```
##
## Call:
## lm(formula = High.Blood.Pressure ~ Taking.BP.Medication, data = health_overview)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.9318 -1.9813 -0.1335  1.8429  9.5576
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -12.06585    1.47056  -8.205 1.98e-15 ***
## Taking.BP.Medication   0.73066    0.02521  28.981  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.859 on 498 degrees of freedom
## Multiple R-squared:  0.6278, Adjusted R-squared:  0.627
## F-statistic: 839.9 on 1 and 498 DF,  p-value: < 2.2e-16
```

Očekivano vidimo da je uzimanje lijekova za regulaciju tlaka statistički značajan regresor. S obzirom na

kontekst problema možemo biti poprilično zadovoljni s koeficijentom korelacije od skoro 63%, ali vrijedi u analizu uključiti i ostale metode preventivne zaštite i pogledati takav model višestruke regresije.

Provjerimo koreliranost između različitih metoda preventivne zaštite.

```
temp <- health_overview[,c(4,5,6,7)]
temp %>% cor %>% corrplot
```




```
## [1] "Correlation between:"
```

```
## [1] "Annual checkups and taking BP medication 0.777800472385448"
```

```
## [1] "Insurance and cholesterol screening: 0.569948426783744"
```

Vidimo da postoji poprilično velika korelacija između sistematskih pregleda i uzimanja lijekova(78%) što nam ima smisla s obzirom da se pacijentima na pregledu ustanovi povišen tlak i zatim propišu lijekovi. Također vidimo koreliranost između zdravstvenog osiguranja i pregleda kolesterola(57%). Imajući na umu da se radi o američkim gradovima gdje je zdravstvena skrb izrazito skupa, za očekivati je da ljudi bez zdravstvenog osiguranja uglavnom neće ići na preglede kolesterola.

Pogledajmo sada linearni model povišenog krvnog tlaka u ovisnosti o svim metodama zaštite.

```
bp_model <- health_overview[,c(4,5,6,7,15)]
fit.bp.all <- lm(High.Blood.Pressure ~ ., data = bp_model)
summary(fit.bp.all)
```

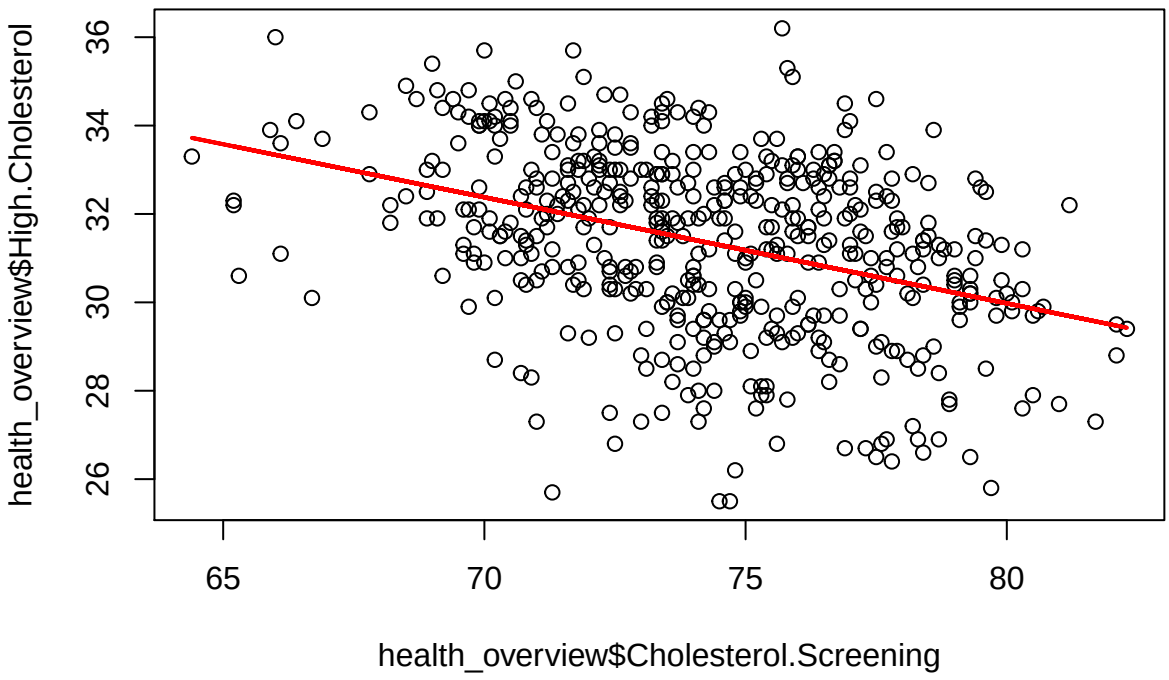
```
##
## Call:
## lm(formula = High.Blood.Pressure ~ ., data = bp_model)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5478 -1.5775 -0.1655  1.4384  7.3778
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           24.33484    2.78267   8.745  < 2e-16 ***
## Health.Insurance      -0.11260    0.02230  -5.049 6.27e-07 ***
## Taking.BP.Medication   0.50406    0.03862  13.051  < 2e-16 ***
## Annual.Checkup         0.25948    0.04526   5.733 1.72e-08 ***
## Cholesterol.Screening -0.42402    0.05128  -8.268 1.26e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.372 on 495 degrees of freedom
## Multiple R-squared:  0.7453, Adjusted R-squared:  0.7432
## F-statistic: 362.1 on 4 and 495 DF,  p-value: < 2.2e-16
```

Svi su regresori statistički značajni, a R^2 i R^2_{adj} koji penalizira dodatne parametre su veći nego u slučaju jednostavne regresije i iznose oko 74%.

Druga povezanost koju očekujemo je između pregleda kolesterola i udjela ljudi s povišenim kolesterolom.

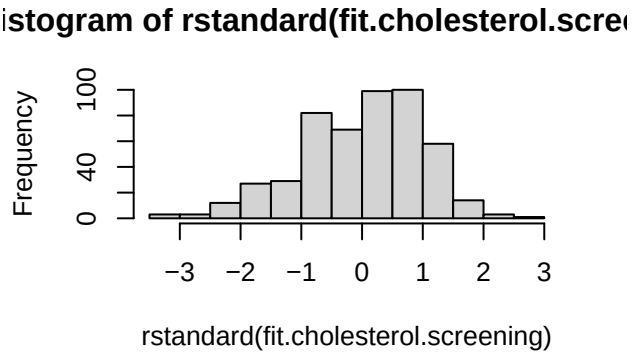
```
plot(health_overview$Cholesterol.Screening, health_overview$High.Cholesterol)
fit.cholesterol.screening <- lm(High.Cholesterol ~ Cholesterol.Screening,
                                data = health_overview)
lines(health_overview$Cholesterol.Screening, fit.cholesterol.screening$fitted.values,
      col = "red",
      lwd = 2)
```

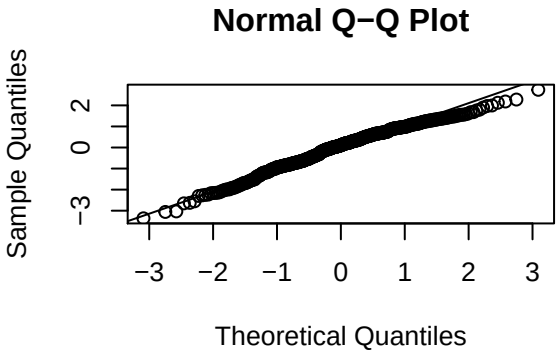


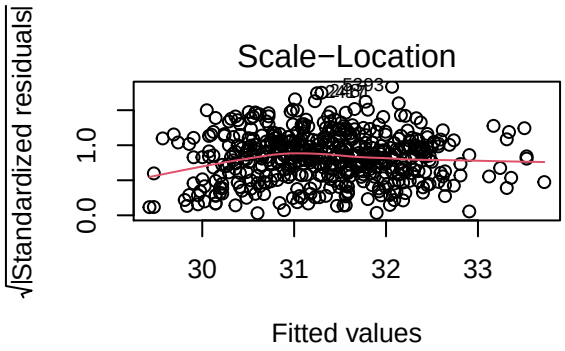
Negativni koeficijent smjera pravca ima smisla i u skladu je s očekivanim - ljudi koji su u posljednjih 5 godina provjerili kolesterol i ustanovili da im je povišen, vjerojatno će promjeniti svoje životne navike i eventualno početi piti lijekove te samim time udio ljudi s povišenim kolesterolom opada.

```
par(mfrow = c(2, 2))
hist(rstandard(fit.cholesterol.screening))

qqnorm(rstandard(fit.cholesterol.screening))
qqline(rstandard(fit.cholesterol.screening))

plot(fit.cholesterol.screening,3)

lillie.test(rstandard(fit.cholesterol.screening))
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rstandard(fit.cholesterol.screening)
## D = 0.05325, p-value = 0.001757
```



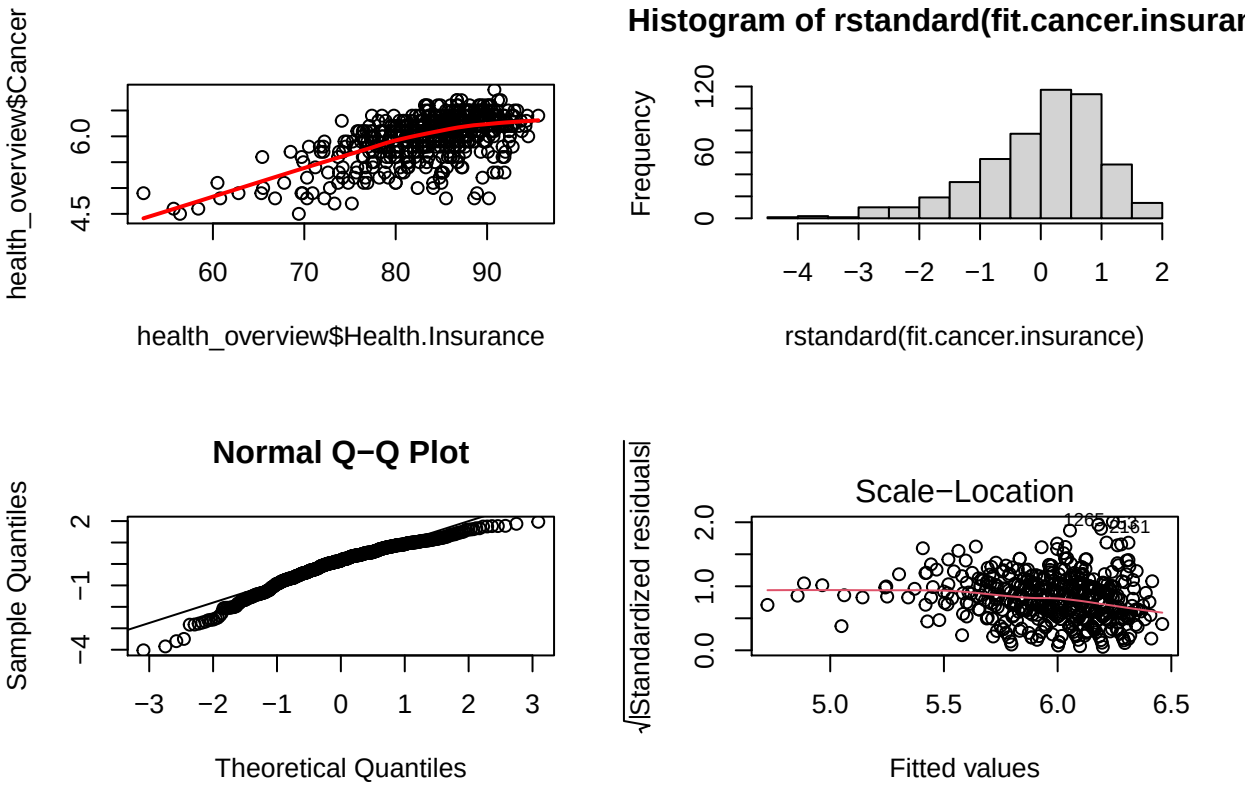




Reziduali u ovom slučaju nisu normalno distribuirani pa pretpostavke modela nisu zadovoljene.

Sljedeći odnos koji ima smisla analizirati je između zdravstvenog osiguranja i raka. Ljudi koji imaju zdravstveno osiguranje imat će bolju zdravstvenu skrb, bolju dijagnostiku i samim time veću šansu otkrivanja raka.

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

scatter.smooth(health_overview$Health.Insurance, health_overview$Cancer,
               lpars=list(col="red", lwd=2))
fit.cancer.insurance = lm(Cancer~Health.Insurance, data= health_overview)

hist(rstandard(fit.cancer.insurance))

qqnorm(rstandard(fit.cancer.insurance))
qqline(rstandard(fit.cancer.insurance))

plot(fit.cancer.insurance,3)
```



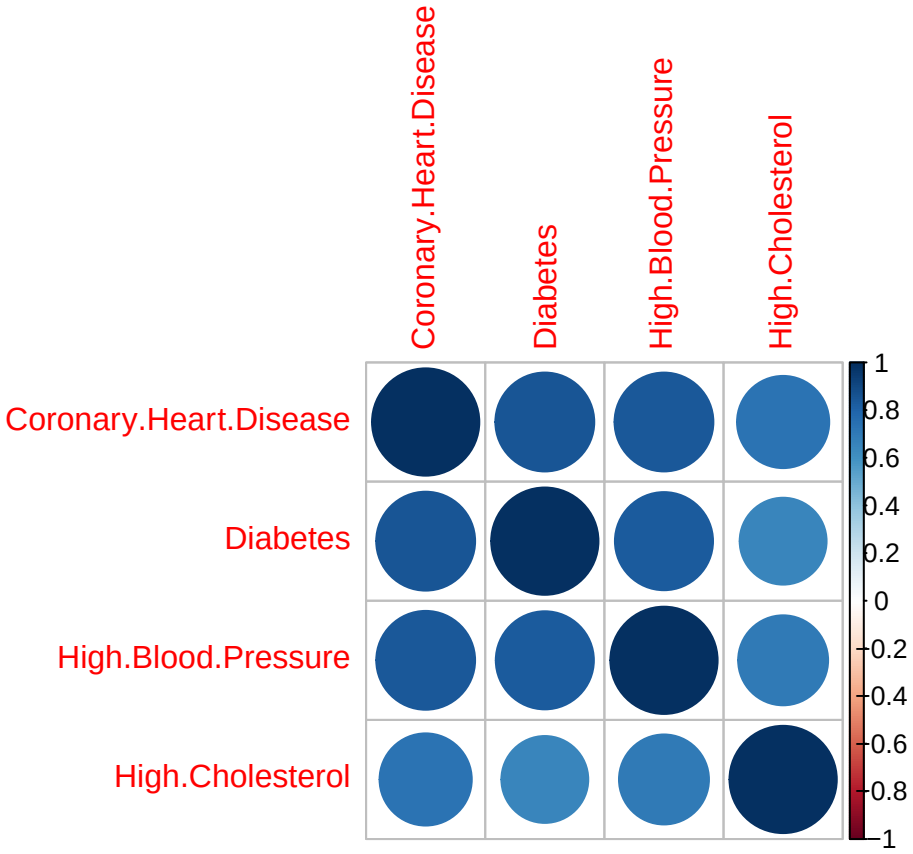

```
lillie.test(rstandard(fit.cancer.insurance))
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rstandard(fit.cancer.insurance)
## D = 0.087255, p-value = 9.79e-10
```

Iako određena povezanost postoji, reziduali u ovom slučaju nisu normalno distribuirani niti homoskedastični te stoga odbacujemo ovaj model.

Osim utjecaja pojedinih metoda zaštite na bolesti, zanimljivo je provjeriti postoji li možda povezanost između nekih parova bolesti. Koronarna bolest, visoki kolesterol, visoki tlak i dijabetes usko su povezani s nezdravim načinom prehrane, starosti osobe, konzumiranjem alkohola te pušenjem. Vrijedi ispitati njihove odnose! Radi sažetosti, usredotočimo se na ovisnost koronarne bolesti o ostalim prethodno navedenima.

```
chd_model <- health_overview[,c(12,14,15,16)]
chd_model %>% cor %>% corrplot
```



Vidimo da postoji snažna korelacija između svih ovih bolesti, pa možda naš model višestruke regresije ne bude valjan. Provjerimo kvalitetu modela jednostavne regresije između koronarne bolesti i ostalih pojedinačno.

```
fit.chd.bp <- lm(Coronary.Heart.Disease ~ High.Blood.Pressure, data = chd_model)
fit.chd.chol <- lm(Coronary.Heart.Disease ~ High.Cholesterol, data = chd_model)
fit.chd.diab <- lm(Coronary.Heart.Disease ~ Diabetes, data = chd_model)
summary(fit.chd.bp)
```

```
##
## Call:
## lm(formula = Coronary.Heart.Disease ~ High.Blood.Pressure, data = chd_model)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.65471 -0.36442 -0.00516  0.33221  1.73810
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.249414   0.154550   1.614    0.107
## High.Blood.Pressure 0.180177   0.005026  35.848   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5256 on 498 degrees of freedom
## Multiple R-squared:  0.7207, Adjusted R-squared:  0.7201
## F-statistic:  1285 on 1 and 498 DF,  p-value: < 2.2e-16
```

```r
summary(fit.chd.chol)
```

```
##
## Call:
## lm(formula = Coronary.Heart.Disease ~ High.Cholesterol, data = chd_model)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.70650 -0.42007 -0.04177  0.41670  2.49381
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -5.47929    0.45967  -11.92   <2e-16 ***
## High.Cholesterol  0.35736    0.01463   24.43   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6708 on 498 degrees of freedom
## Multiple R-squared:  0.5451, Adjusted R-squared:  0.5442
## F-statistic: 596.7 on 1 and 498 DF,  p-value: < 2.2e-16
```

```r
summary(fit.chd.diab)
```

```
##
## Call:
## lm(formula = Coronary.Heart.Disease ~ Diabetes, data = chd_model)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.51322 -0.32902 -0.00533  0.34005  1.73414
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.218580   0.100045   22.18   <2e-16 ***
## Diabetes    0.342096   0.009489   36.05   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5234 on 498 degrees of freedom
## Multiple R-squared:  0.723,  Adjusted R-squared:  0.7224
## F-statistic:  1300 on 1 and 498 DF,  p-value: < 2.2e-16
```

Vidimo da i modeli jednostavne regresije koronarne bolesti u ovisnosti o povišenom tlaku ili dijabetesu imaju poprilično visok R^2 od oko 72%. Prikažimo sada model višestruke regresije.

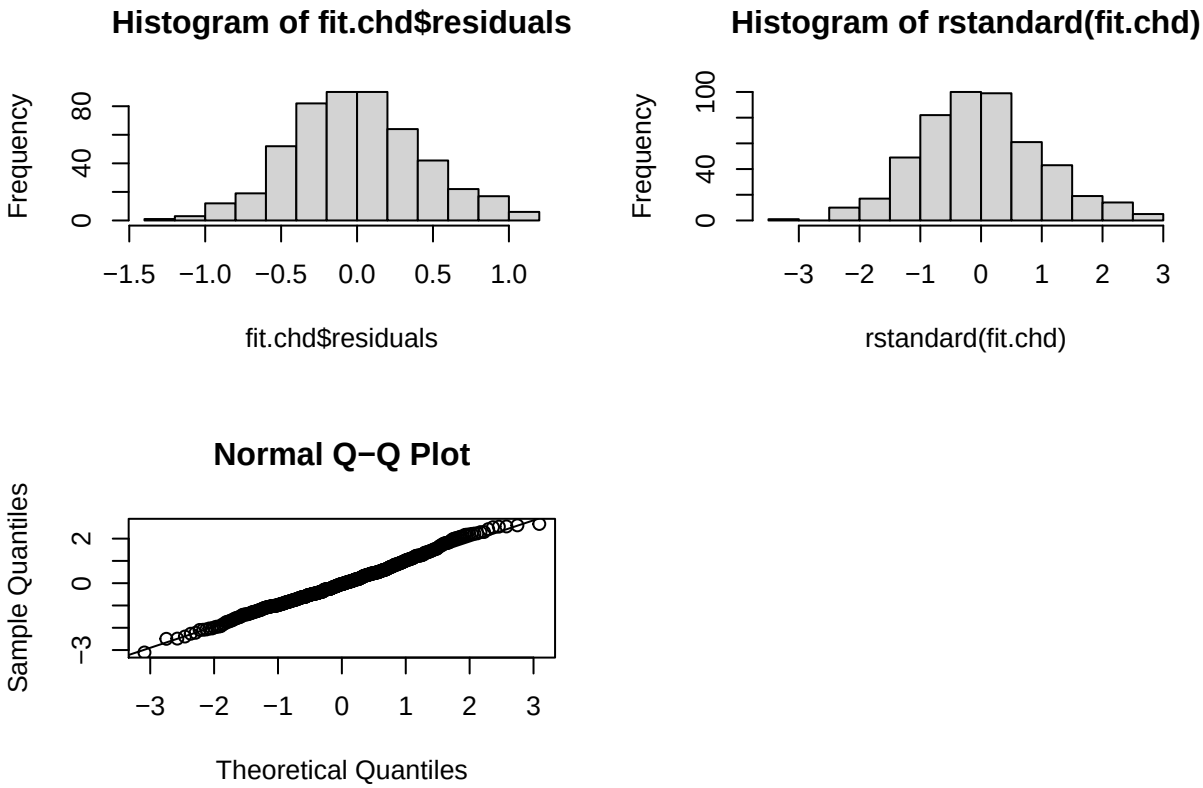
```r
fit.chd <- lm(Coronary.Heart.Disease ~ ., data = chd_model)
summary(fit.chd)
```

```
##
## Call:
## lm(formula = Coronary.Heart.Disease ~ ., data = chd_model)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.34471 -0.29482 -0.01838  0.26119  1.14274
##
```

```
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -1.600030   0.333622  -4.796 2.14e-06 ***
## Diabetes             0.167476   0.014559  11.504  < 2e-16 ***
## High.Blood.Pressure  0.072932   0.008153   8.945  < 2e-16 ***
## High.Cholesterol     0.108188   0.013509   8.009 8.33e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4336 on 496 degrees of freedom
## Multiple R-squared:  0.8107, Adjusted R-squared:  0.8095
## F-statistic:   708 on 3 and 496 DF,  p-value: < 2.2e-16
```

Unatoč koreliranosti, vidimo da su svi regresori značajni s vrlo niskom p-vrijednosti i koeficijent determinacije nam je >80%, što je bolje nego u modelima jednostavne linearne regresije. Provjerimo pretpostavke modela.

```
par(mfrow=c(2,2))
hist(fit.chd$residuals)
hist(rstandard(fit.chd))
qqnorm(rstandard(fit.chd))
qqline(rstandard(fit.chd))
```



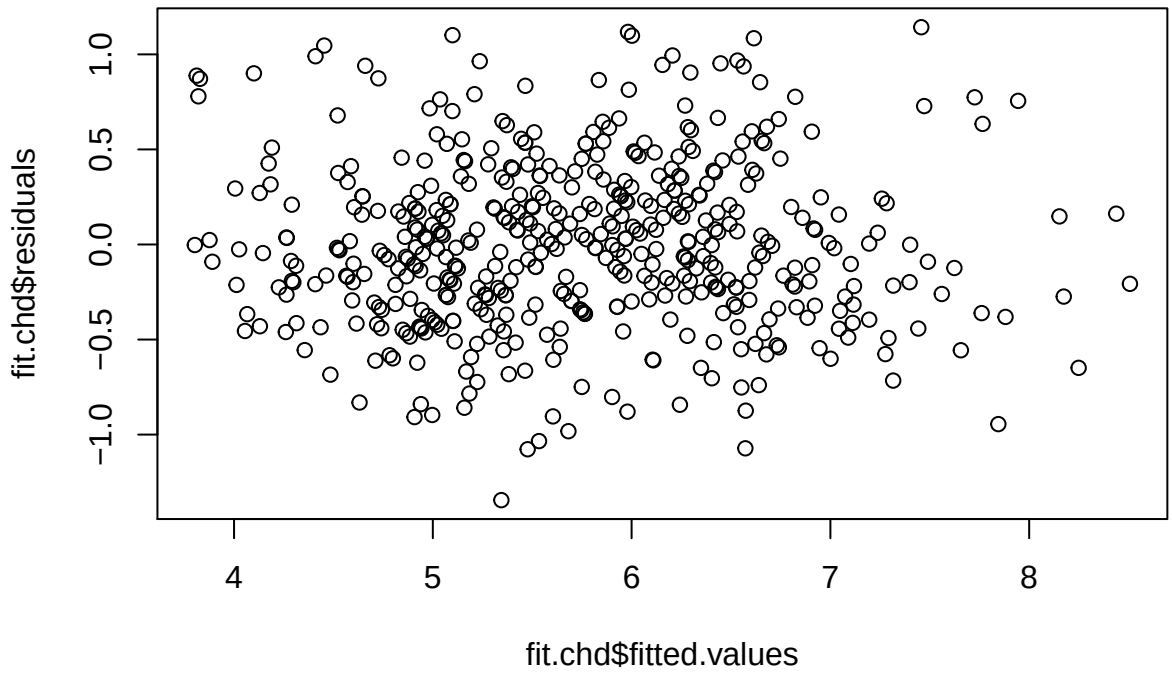
Grafovi nam izgledaju prihvatljivo, distribucije su zvonolike i nisu previše zakrivljene. Uvjerimo se u normalnost Lillieforsovim testom.

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

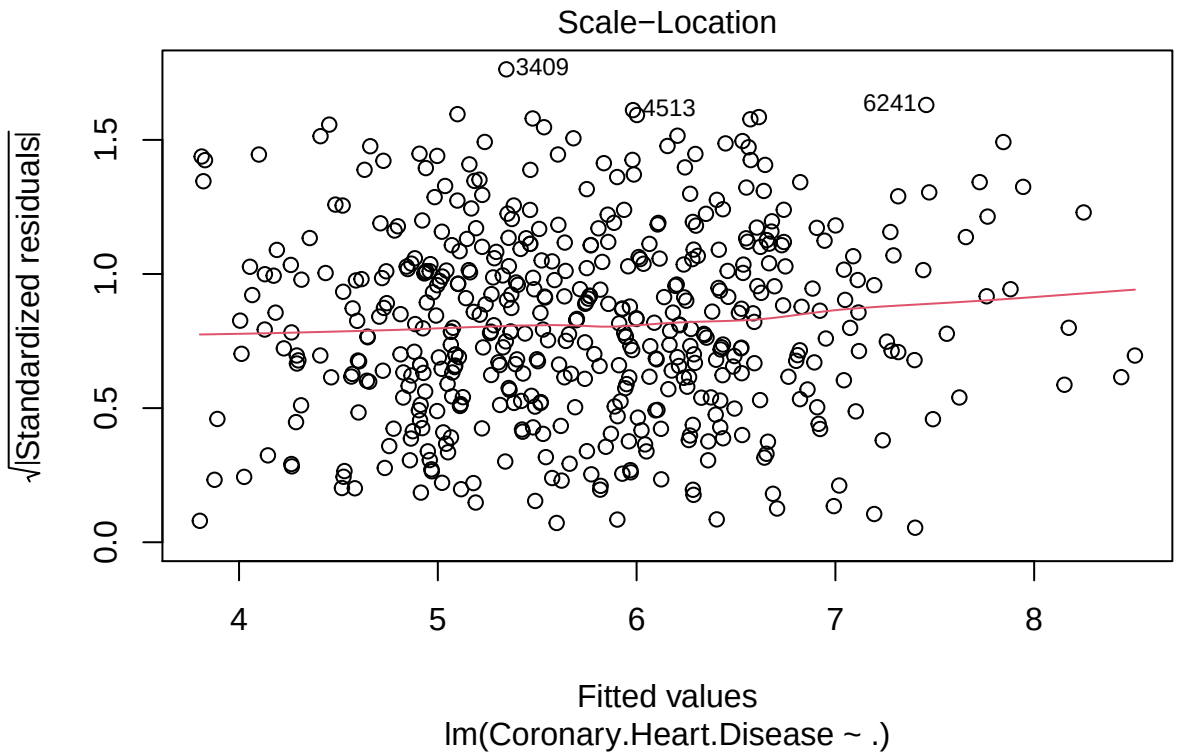
```
##
```

```
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rstandard(fit.chd)
## D = 0.030888, p-value = 0.2933
```

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




```
plot(fit.chd, 3)
```



Iz grafova odlučujemo prihvatiti pretpostavku homogenosti varijance. Možemo sada izračunati predikciju udjela stanovništva s koronarnom bolesti u ovisnosti o udjelima svake od bolesti koje su regresori u modelu. Provjerimo očekivani udio ljudi s koronarnom bolesti u slučaju da polovica stanovništva ima problema s tlakom, povišen kolesterol i pati od dijabetesa.

```
test_data <- data.frame(High.Blood.Pressure = 50, High.Cholesterol = 50, Diabetes = 50)
mean.resp <- predict(fit.chd, test_data, interval = "confidence")
pred.value <- predict(fit.chd, test_data, interval = "prediction")
```

```
## [1] "95%-tni interval pouzdanosti za srednju vrijednost: [ 14.9206920090254 ,  16.738876903915 ]"
```

```
## [1] "Predikcija za zadane vrijednosti: 15.8297844564702"
```

```
## [1] "95%-tni interval pouzdanosti za predikciju: [ 14.5839339957584 ,  17.075634917182 ]"
```

Valja uočiti da je interval pouzdanosti širi za predikciju jedne vrijednosti nego za srednju vrijednost, što je u skladu s očekivanim.

Zaključak

Ovim projektom smo kroz tri glavna zadatka proveli statističku analizu podataka o preventivnoj zdravstvenoj zaštiti. Najprije smo usporedili popularnost pojedinih metoda u saveznim državama Ohio i Florida koristeći se testom o proporcijama, zatim smo testom o homogenosti analizirali koliki udio ljudi boluje od kroničnih plućnih bolesti te na kraju kroz modele linearne regresije (jednostavne i višestruke) pokazali odnos između bolesti i metoda.