Kako (ne) zdrave navike utječu na zdravlje

Projektni zadatak iz kolegija Statistička analiza podataka

Dora Doljanin, Filip Pavičić, Marko Andreis, Toni Matić

17/12/2020

1 Motivacija i opis problema

Čovjek oduvijek teži razumjeti odnos svojeg ponašanja i navika sa svojim zdravljem. Navike mogu biti dobre i loše, a moderni način života uvelike utječe na njih. Boljim razumijevanjem kako i koje nezdrave navike utječu na zdravlje, osvještavamo se o potencijalnoj potrebi za promjene ili ustrajnosti na već usvojenim dobrim navikama. Cilj ovog projektnog zadatka je istražiti zdravstvene indikatore koji spadaju u nezdrave navike i zdravstvene tegobe i bolesti. Posebice, zanima nas postoje li veze među njima. Ti indikatori dani su za više gradova, stoga nas također zanima postoje li gradovi koji općenito imaju viši zdravstveni standard.

2 Opis skupa podataka

Dani su podatci jedne godine za 500 američkih gradova. Za svaki grad mjerene su 4 vrste čestih nezdravih navika (kao što je prekomjerno pijenje alkohola ili vođenje fizički neaktivnog života). Te mjere iskazane su kao udio stanovništva pojedinog grada koji ima određenu naviku. Pored toga, mjereno je i 12 zdravstvenih stanja ili bolesti (kao što je artritis, visoki krvni tlak, kronične srčane bolesti itd.), također kao udio stanovništva pojedinog grada koji pati od dane bolesti.

```
knitr::include_graphics('usa_map.jpeg')
```

Figure 1: Prikaz gradova uključenih u skup podataka

Skup podataka koji se koristi u analizi daljnjih problema nalazi se u datoteci `data_health_and_unhealthy_habits.csv` te se pomoću naredbe `read.csv` koja kao argument prima datoteku s podatcima radni podatci pohranjuju u željenu varijablu `habits`. U nastavku je prikazan dio koda koji izvršava navedeno. Dodatno, izdvajaju se samo oni retci koji za vrijednost varijable `DataValueTypeID` poprimaju `CrdPrv`, umjesto alternative `AgeAdj` čime su za daljnju analizu odlučuje za podatke koji nisu prilagođeni starosti stanovništva iz uzorka.

```
habits = read.csv('data_health_and_unhealthy_habits.csv')
crd_habits = habits[habits$DataValueTypeID =="CrdPrv",]
```

Podatci se sastoje od 16000 zapisa (gradovi u kombinaciji s nezdravim navikama ili zdravstvenim tegobama) i 10 njihovih opisa (varijabli).

Deskriptivnu statistiku moguće je dobiti naredbom `summary(habits)`, dodatno prvih nekoliko redaka iz skupa podataka moguće je dohvatiti naredbom `head(habits)`.

Izvođenjem naredbe `sum(is.na(habits$StateDesc))` dobiven je uvid u broj nedostajućih vrijednosti za varijablu `StateDesc` koja identificira pojedinu saveznu državu. Isti postupak izveden je za sve ostale varijable iz skupa podataka. Budući da su navedene vrijednosti bile jednake nuli za svaki izračun, zaključujemo da skup ne sadrži nedostajuće vrijednosti.

1. ZADATAK

Nakon što smo se bolje upoznali s našim podatcima, možemo si postaviti neka zanimljiva pitanja i pokušati odgovoriti na njih koristeći razne statističke alate.

Primjerice, pokušajmo odgovoriti na sljedeće pitanje: postoji li neka nezdrava navika koja je manje ”popularna” u saveznoj državi Arizoni nego u saveznoj državi Tennessee (odnosno, za koju je udio stanovnika savezne države Arizona manji od udjela stanovnika savezne države Tennessee)?

Prvo ćemo iz zadanog skupa podataka izračunati ukupnu veličinu populacije za saveznu državu Arizonu te za saveznu državu Tennessee za dane gradove. Također, za obje savezne države ćemo izračunati veličinu podskupa populacije koji prakticira svaku od 4 nezdrave navike.

Analizu započinje učitavanjem podataka vezanih uz navedene savezne države i izdvajamo retke koji se odnose na nezdrave navike.

```
# svi redovi gdje je država = Arizona ili Tennessee:
data = crd_habits[crd_habits$StateDesc =="Arizona"| crd_habits$StateDesc =="Tennessee",]

# svi redovi gdje je Category = Unhealthy Behaviors:
data = data[data$Category =="Unhealthy Behaviors",]


# potrebno izvesti install.packages("dplyr") u consoli
```

U nastavku analize podatci su grupirani s obzirom na saveznu državu i nezdravu naviku te je za svaki tako dobiveni redak izračunat broj ljudi koji imaju konkretnu nezdravu naviku kao i ukupan broj ljudi u uzroku koji dolaze iz Arizone ili Tennesseeja.

```
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
```

```
groupData = data %>% group_by(StateDesc, Short_Question_Text) %>%
              summarise(UnhealthyCount
```

```
          = round(sum(PopulationCount * Data_Value / 100), digits = 0),
            PopulationCount = sum(PopulationCount))
```

```
## `summarise()` regrouping output by 'StateDesc' (override with `.groups` argument)
```

```
groupData
```

```
## # A tibble: 8 x 4
## # Groups:   StateDesc [2]
##   StateDesc Short_Question_Text UnhealthyCount PopulationCount
##   <chr>     <chr>                       <dbl>           <dbl>
## 1 Arizona   Binge Drinking             585980         3896074
## 2 Arizona   Current Smoking            640296         3896074
## 3 Arizona   Obesity                   1089712         3896074
## 4 Arizona   Physical Inactivity        943338         3896074
## 5 Tennessee Binge Drinking             223269         1836343
## 6 Tennessee Current Smoking            426881         1836343
## 7 Tennessee Obesity                    630928         1836343
## 8 Tennessee Physical Inactivity        555470         1836343
```

Nakon što smo grupirali podatke, iz dobivene tablice izvući ćemo zasebno podatke za svaku nezdravu naviku.

```
UkupnaPopulacija = c(
  sum(data[data$StateDesc == "Arizona","PopulationCount"]) / 4,
  sum(data[data$StateDesc == "Tennessee","PopulationCount"]) / 4
  )

BingeDrinkingPopulacija = c(
  round(groupData[groupData$StateDesc == "Arizona"
          & groupData$Short_Question_Text == "Binge Drinking",
          "UnhealthyCount"][[1]][1], digits = 0),
  round(groupData[groupData$StateDesc == "Tennessee"
          & groupData$Short_Question_Text == "Binge Drinking",
          "UnhealthyCount"][[1]][1], digits = 0)
)

CurrentSmokingPopulacija = c(
  round(groupData[groupData$StateDesc == "Arizona"
          & groupData$Short_Question_Text == "Current Smoking",
          "UnhealthyCount"][[1]][1], digits = 0),
  round(groupData[groupData$StateDesc == "Tennessee"
          & groupData$Short_Question_Text == "Current Smoking",
          "UnhealthyCount"][[1]][1], digits = 0)
)

ObesityPopulacija = c(
  round(groupData[groupData$StateDesc == "Arizona"
          & groupData$Short_Question_Text == "Obesity",
          "UnhealthyCount"][[1]][1], digits = 0),
  round(groupData[groupData$StateDesc == "Tennessee"
          & groupData$Short_Question_Text == "Obesity",
          "UnhealthyCount"][[1]][1], digits = 0)
)

PhysicalInactivityPopulacija = c(
```

```
  round(groupData[groupData$StateDesc == "Arizona"
          & groupData$Short_Question_Text == "Physical Inactivity",
          "UnhealthyCount"][[1]][1], digits = 0),
  round(groupData[groupData$StateDesc == "Tennessee"
          & groupData$Short_Question_Text == "Physical Inactivity",
          "UnhealthyCount"][[1]][1], digits = 0)
)
```

Time smo za svaku nezdravu naviku dobili broj ljudi koji ju prakticira u saveznoj državi Arizoni i broj ljudi koji ju prakticira u saveznoj državi Tennessee. Potom ćemo za svaku nezdravu naviku napraviti Z-test o dvije proporcije, koji je u R-u implementiran u funkciji `prop.test()`.

Postavimo hipoteze i provedimo testiranje.

$$H_0 : p_A >= p_T$$
$$H_1 : p_A < p_T$$

U navedenim hipotezama sada p_A predstavlja udio stanovnika savezne države Arizona koji prakticira neku nezdravu naviku, analogno vrijedi i za oznaku p_T koja se odnosi na saveznu državu Tennessee.

Slijedi testiranje postavljenih hipoteza za sljedeće nezdrave navike: konzumiranje alkohola, pušenje, pretilost i fizičku neaktivnost.

```
prop.test(BingeDrinkingPopulacija, UkupnaPopulacija, alternative = "less")
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  BingeDrinkingPopulacija out of UkupnaPopulacija
## X-squared = 8549.6, df = 1, p-value = 1
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.00000000  0.02931567
## sample estimates:
##    prop 1    prop 2
## 0.1504027 0.1215835
```

```
prop.test(CurrentSmokingPopulacija, UkupnaPopulacija, alternative = "less")
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  CurrentSmokingPopulacija out of UkupnaPopulacija
## X-squared = 38224, df = 1, p-value < 2.2e-16
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.00000000 -0.06751972
## sample estimates:
##    prop 1    prop 2
## 0.1643439 0.2324626
```

```
prop.test(ObesityPopulacija, UkupnaPopulacija, alternative = "less")
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  ObesityPopulacija out of UkupnaPopulacija
```

```
## X-squared = 24247, df = 1, p-value < 2.2e-16
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.00000000 -0.06319606
## sample estimates:
##    prop 1    prop 2
## 0.2796949 0.3435785
```

```
prop.test(PhysicalInactivityPopulacija, UkupnaPopulacija, alternative = "less")
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  PhysicalInactivityPopulacija out of UkupnaPopulacija
## X-squared = 23549, df = 1, p-value < 2.2e-16
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.00000000 -0.05969935
## sample estimates:
##    prop 1    prop 2
## 0.2421253 0.3024871
```

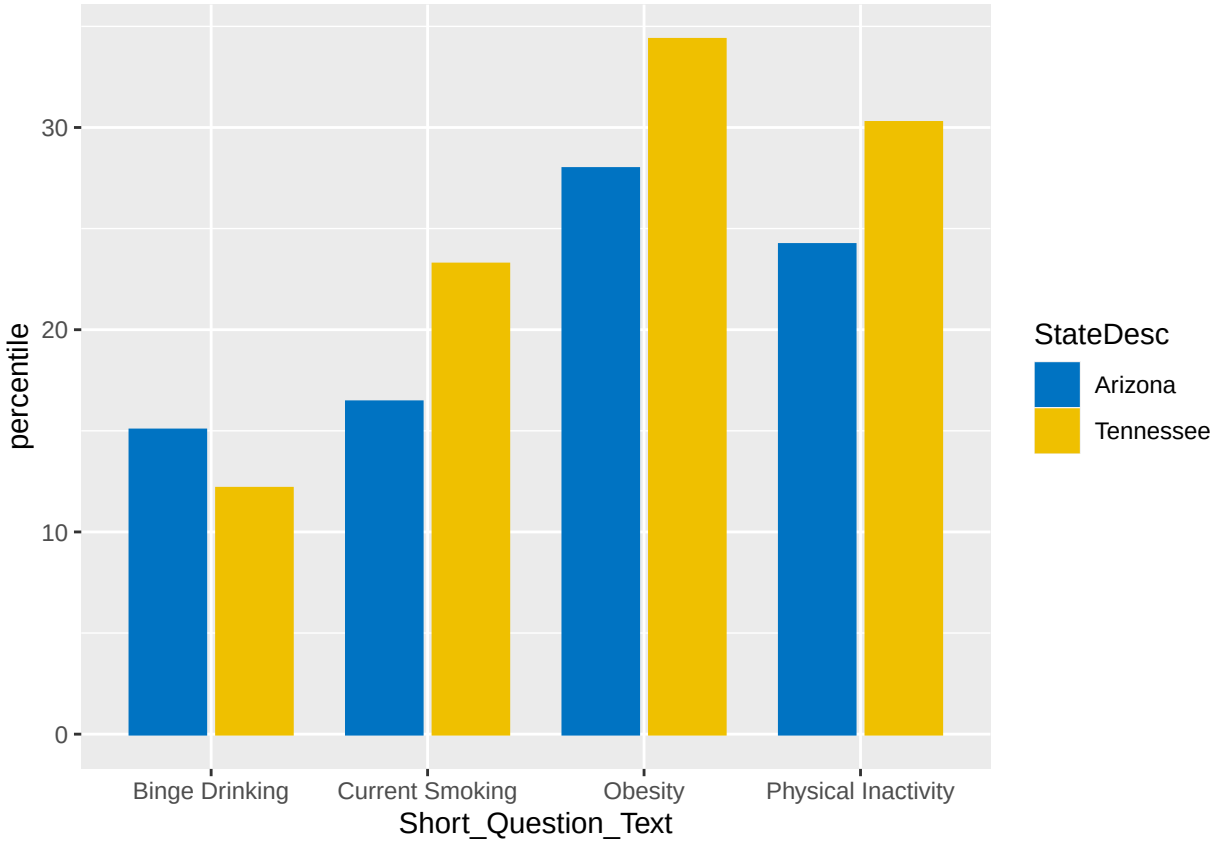
Rezultati testiranja pokazali su da, zbog p-vrijednosti iznosa 1 za nezdravu naviku prekomjernog konzumiranja alkohola, ne možemo odbaciti hipotezu H_0 u korist hipoteze H_1. Za preostale tri navike, međutim, zbog jako male p-vrijednosti iznosa $<$ 2.2e-16 odbacujemo nultu hipotezu te zaključujemo da su pušenje, pretilost i nedovoljno fizičke aktivnosti manje zastupljeni u saveznoj državi Arizoni nego u saveznoj državi Tennessee.

Vizualizirat ćemo dobivene podatke stupčastim dijagramom.

```
groupDataPercentile = data %>% group_by(StateDesc, Short_Question_Text) %>%
  summarise(percentile =
              sum(PopulationCount * Data_Value) / sum(PopulationCount))
```

```
## `summarise()` regrouping output by 'StateDesc' (override with `.groups` argument)
```

```
library(ggplot2)
p <- ggplot(groupDataPercentile, aes(x = Short_Question_Text, y = percentile)) +
  geom_bar(
    aes(color = StateDesc, fill = StateDesc),
    stat = "identity", position = position_dodge(0.8),
    width = 0.7
    ) +
  scale_color_manual(values = c("#0073C2FF", "#EFC000FF")) +
  scale_fill_manual(values = c("#0073C2FF", "#EFC000FF"))
p
```



Sada kada pogledamo prikaz podataka u stupčastom dijagramu, jasno nam je zašto smo u prvom testu dobili p-vrijednost = 1, s obzirom na to da i sam postotak jasno pokazuje da je nezdrava navika prekomjernog konzumiranja alkohola zastupljenija u saveznoj državi Arizoni.

2. ZADATAK

U sljedećem koraku analize dobivenih podataka o stanovništvu saveznih država i njihovim nezdravim navikama i zdravstvenim tegobama, postavljamo sljedeće pitanje: za određenu zdravstvenu tegobu i proizvoljan broj odabranih saveznih država, je li postotak stanovništva koji boluje od odabrane bolesti jednak u svima od tih država?

Konkretno, u ovoj analizi odlučujemo se za sljedeće savezne države: New York, Floridu i Illinois i pokušavamo odrediti je li postotak stanovništva koji boluje od dijabetesa jednak za sve tri navedene države.

Za početak, izvlačimo podatke potrebne u daljnjem tijeku analize. Iz početno selektiranih podataka izvlačimo one koji se odnose na saveznu državu New York, odnosno one kojima varijabla `StateDesc` poprima vrijednost `New York`. Budući da smo za daljnju analizu odabrali dijabetes kao bolest po kojoj radimo usporedbu, dodatno izvlačimo retke kojima varijabla `Short_Question_Text` poprima vrijednost `Diabetes`.

```
NewYork = crd_habits[crd_habits['StateDesc'] == 'New York' &
          crd_habits['Short_Question_Text'] == 'Diabetes',
          c('StateDesc', 'CityName', 'Category', 'Measure',
            'Data_Value', 'PopulationCount')]


sumNYDiabetes = round(sum(NewYork$Data_Value * NewYork$PopulationCount / 100), digits = 0)
sumNYTotal = sum(NewYork$PopulationCount)
```

```
percentageNY = sumNYDiabetes / sumNYTotal
```

Isti postupak ponavljamo za saveznu državu Floridu.

```
Florida = crd_habits[crd_habits['StateDesc'] == 'Florida' &
          crd_habits['Short_Question_Text'] == 'Diabetes', c('StateDesc',
                     'CityName', 'Category', 'Measure', 'Data_Value',
                     'PopulationCount')]

sumFLDiabetes = round(sum(Florida$Data_Value * Florida$PopulationCount / 100), digits = 0)
sumFLTotal = sum(Florida$PopulationCount)
percentageFL = sumFLDiabetes / sumFLTotal
```

Također, postupak se ponavlja za saveznu državu Illinois.

```
Illinois = crd_habits[crd_habits['StateDesc'] == 'Illinois' &
           crd_habits['Short_Question_Text'] == 'Diabetes', c('StateDesc',
                     'CityName', 'Category', 'Measure', 'Data_Value',
                     'PopulationCount')]


sumILDiabetes = round(sum(Illinois$Data_Value * Illinois$PopulationCount / 100), digits = 0)
sumILTotal = sum(Illinois$PopulationCount)
percentageIL = sumILDiabetes / sumILTotal
```

```
observed = matrix(c(sumNYDiabetes, sumNYTotal - sumNYDiabetes, sumFLDiabetes,
                    sumFLTotal - sumFLDiabetes, sumILDiabetes,
                    sumILTotal - sumILDiabetes), nrow = 2)
colnames(observed) = c("New York", "Florida", "Illinois")
rownames(observed) = c("Yes", "No")
```

Iz tako dobivenih podataka pripremamo kontingencijsku tablicu koja nam služi za provođenje testa homogenosti. U nastavku se nalazi kontingencijska tablica kojoj su stupci savezne države New York, Florida i Illinois, a pojedini redak predstavlja broj ljudi koji imaju (`Yes`) i broj ljudi koji nemaju dijabetes (`No`).

```
observed
```

```
##     New York Florida Illinois
## Yes   998089  592019   468971
## No   8298410 4574468  3978621
```

Sada imamo pripremljeno sve za obavljanje testa homogenosti. Postavljamo nul-hipotezu i alternativnu hipotezu:
H_0 : *postotak stanovistva koji boluje od dijabetesa jednak za sve tri drzave*
H_1 : *postotak stanovistva koji boluje od dijabetesa nije jednak za sve tri drzave*

Pozivamo metodu `chisq.test()` koja prima pripremljenu kontingencijsku tablicu i na temelju kontingencijske tablice računa vrijednost χ^2 statistike te u konačnici i samu p-vrijednost na temelju koje donosimo zaključke.

```
res=chisq.test(observed)
res
```

```
##
##  Pearson's Chi-squared test
##
## data:  observed
## X-squared = 2497.1, df = 2, p-value < 2.2e-16
```

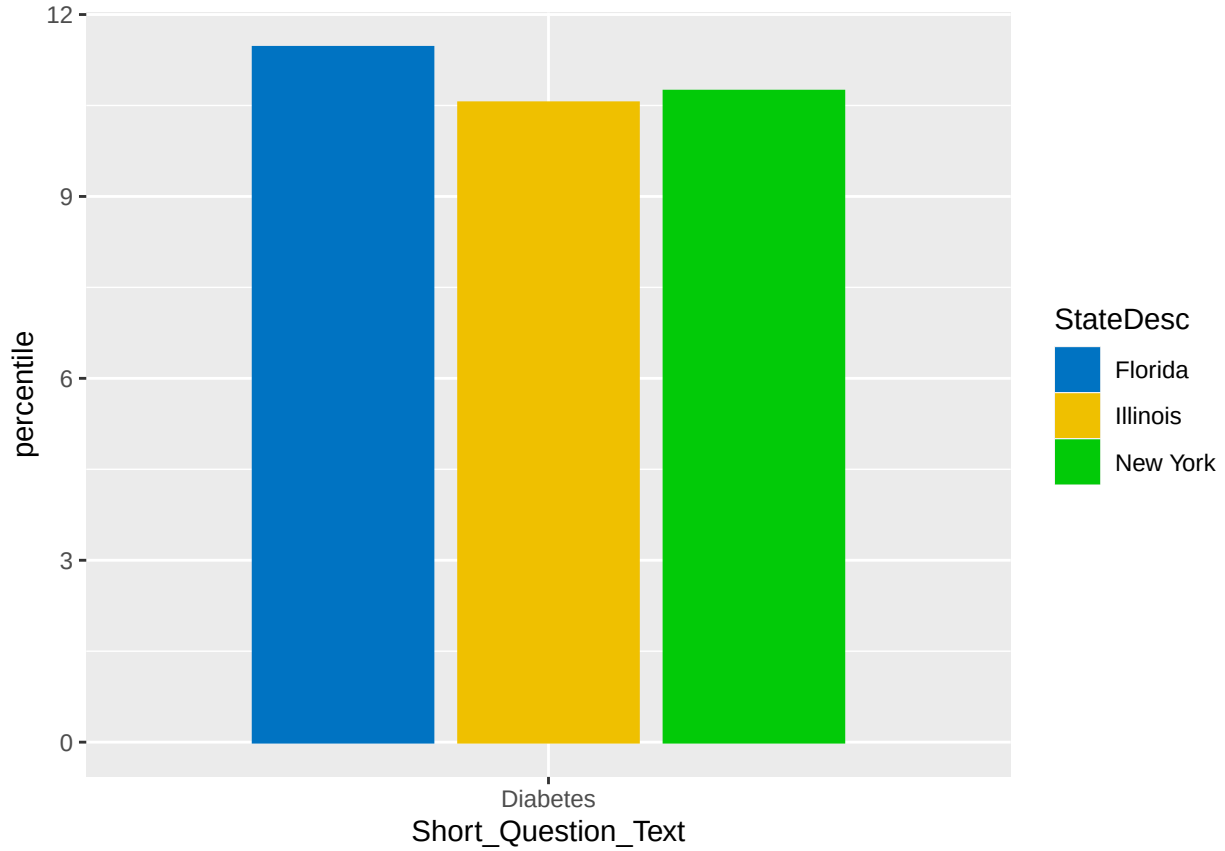
Na temelju izračunate p-vrijednosti čiji je iznos skoro jednak nuli i postavljenih hipoteza možemo odbaciti

nul-hipotezu H_0 u korist alternativne hipoteze H_1 i zaključiti kako postotak stanovništva koji boluje od dijabetesa nije jednak za sve tri države. Međutim, valja biti oprezan pri zaključivanju budući da se broj jedinki u uzorku broji u milijunima zbog čega dolazi do efekta da se odbacuju pretpostavke koje su možda točne. Konkretno, iz grafa koji slijedi u nastavku, razumno bi bilo zaključiti kako je postotak stanovništva koji boluje od dijabetesa ipak jednak za sve tri države s obzirom na prikazane vrijednosti unatoč rezultatu testa.

```
# svi redovi gdje je država = New York ili Florida ili Illinois:
data = crd_habits[crd_habits$StateDesc =="New York"| crd_habits$StateDesc =="Florida" |
               crd_habits$StateDesc =="Illinois",]

# svi redovi gdje je Short_Question_Text = Diabetes:
data = data[data$Short_Question_Text =="Diabetes",]


groupDataPercentile = data %>% group_by(StateDesc, Short_Question_Text) %>%
  summarise(percentile = sum(PopulationCount * Data_Value) /
              sum(PopulationCount))
```

```
## `summarise()` regrouping output by 'StateDesc' (override with `.groups` argument)
```

```
p <- ggplot(groupDataPercentile, aes(x = Short_Question_Text, y = percentile)) +
  geom_bar(
    aes(color = StateDesc, fill = StateDesc),
    stat = "identity", position = position_dodge(0.8),
    width = 0.7
    ) +
  scale_color_manual(values = c("#0073C2FF", "#EFC000FF", "#02CA08FF"))+
  scale_fill_manual(values = c("#0073C2FF", "#EFC000FF", "#02CA08FF"))
p
```



Na temelju stupčastog dijagrama sada vidimo kako je postotak stanovništva koji boluje od dijabetesa nešto viši u saveznoj državi Floridi, nego što je to slučaj u saveznim državama Illinois i New York. Budući da su vrijednosti koje predstavljaju postotak stanovništva koji boluje od dijabetesa za Illinois i New York nešto bliže nego što je to slučaj za kombinacije spomenutih država i savezne države Floride, u nastavku analize ćemo provjeriti vrijedi li hipoteza da su postotci stanovništva koji boluje od dijabetesa jednaki za savezne države Illinois i New York.

Postavljamo nul-hipotezu i alternativnu hipotezu:
H_0 : *postotak stanovistva koji boluje od dijabetesa jednak je za Illinois i New York*
H_1 : *postotak stanovistva koji boluje od dijabetesa nije jednak za Illinois i New York*

```
observed = matrix(c(sumILDiabetes, sumILTotal - sumILDiabetes,
                    sumNYDiabetes, sumNYTotal - sumNYDiabetes), nrow = 2)


colnames(observed) = c("Illinois", "New York")
rownames(observed) = c("Yes", "No")

observed
```

```
##     Illinois New York
## Yes   468971   998089
## No   3978621  8298410
```

```
res=chisq.test(observed)
res
```

```
##
```

```
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  observed
## X-squared = 116.05, df = 1, p-value < 2.2e-16
```

Na temelju izračunate p-vrijednosti čiji je iznos skoro jednak nuli i postavljenih hipoteza možemo odbaciti nul-hipotezu H_0 u korist alternativne hipoteze H_1 i zaključiti kako postotak stanovništva koji boluje od dijabetesa ipak nije jednak za Illinois i New York. Međutim, ponovno, valja biti oprezan pri zaključivanju, zbog već spomenutog velikog broja jedinki u uzorku. Konkretno, iz grafa koji je prethodno prikazan, razumno bi bilo zaključiti kako je postotak stanovništva koji boluje od dijabetesa ipak jednak za Illinois i New York s obzirom na prikazane vrijednosti unatoč rezultatu testa.

3 Zadatak

Nakon što smo ispitali veze između ovih nezdravih navika i pojedinih bolesti, odlučili smo detaljnije proučiti bolest COPD (Kronična opstruktivna plućna bolest) i ispitati koje nezdrave navike imaju najveći utjecaj na nju.

```
Imebolest = 'COPD'
```

Izdvojit ćemo nekoliko nama najzanimljivih modela i detaljno ćemo ih predstaviti.

Pogledajmo kakva je veza između bolesti COPD i svake od 4 nezdrave navike.

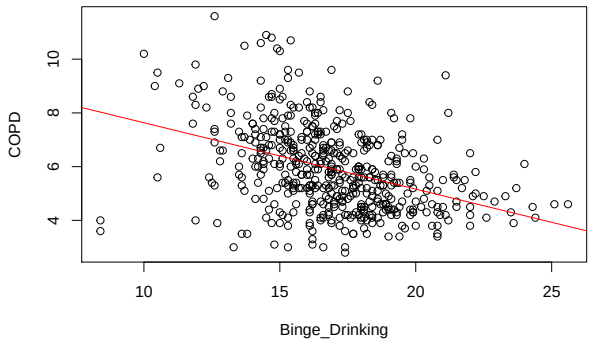
```
COPD = crd_habits[crd_habits['Short_Question_Text'] == Imebolest, 'Data_Value']
Binge_Drinking = crd_habits[crd_habits['Short_Question_Text']
                            == 'Binge Drinking', 'Data_Value']
Current_Smoking = crd_habits[crd_habits['Short_Question_Text']
                             == 'Current Smoking', 'Data_Value']
Obesity = crd_habits[crd_habits['Short_Question_Text'] == 'Obesity', 'Data_Value']
Physical_Inactivity = crd_habits[crd_habits['Short_Question_Text']
                                 == 'Physical Inactivity', 'Data_Value']
```

Na sljedećim grafovima, x os označava postotak ljudi koji prakticira određenu nezdravu naviku, a y os označava postotak ljudi koji boluje od bolesti COPD. Svaka točka predstavlja podatak za jedan od gradova iz ulaznog skupa podataka. Izgled grafova upućuju na to da postoji neka vrsta ovisnosti između varijabli Current_Smoking, Obesity i Physical_Inactivity s bolešću COPD, što ćemo u nastavku detaljnije istražiti. Što se tiče varijable Binge_Drinking, njezin graf pokazuje potencijalni blagi negativni trend, što nas je iznenadilo.

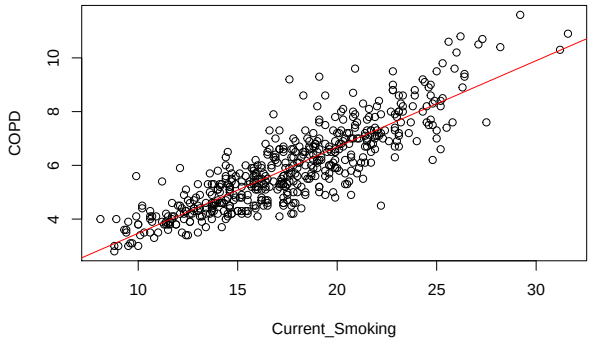
Pogledajmo kako izgledaju modeli jednostruke regresije za svaku od 4 nezdrave navike.

```
fit.Binge_Drinking = lm(COPD ~ Binge_Drinking)
fit.Current_Smoking = lm(COPD ~ Current_Smoking)
fit.Obesity = lm(COPD ~ Obesity)
fit.Physical_Inactivity = lm(COPD ~ Physical_Inactivity)
```

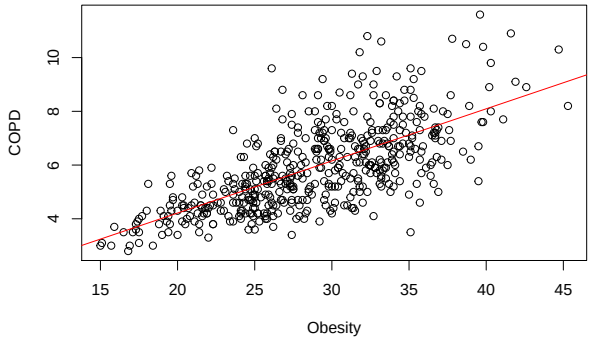
```
plot(Binge_Drinking, COPD)
abline(fit.Binge_Drinking, col='red')
```



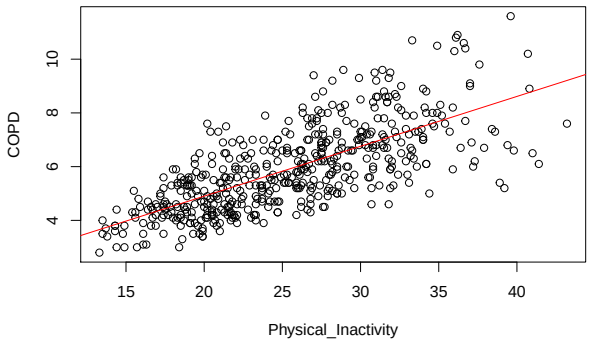

```
plot(Current_Smoking, COPD)
abline(fit.Current_Smoking, col='red')
```




```
plot(Obesity, COPD)
abline(fit.Obesity, col='red')
```



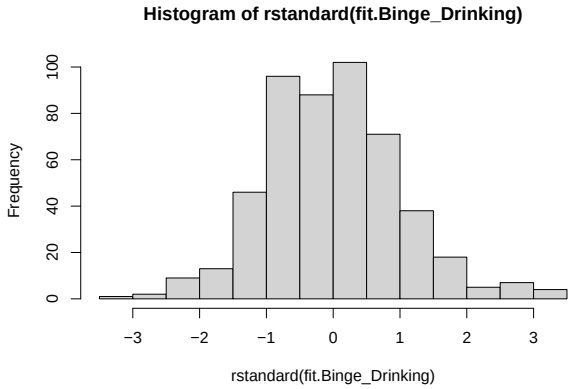

```
plot(Physical_Inactivity, COPD)
abline(fit.Physical_Inactivity, col='red')
```



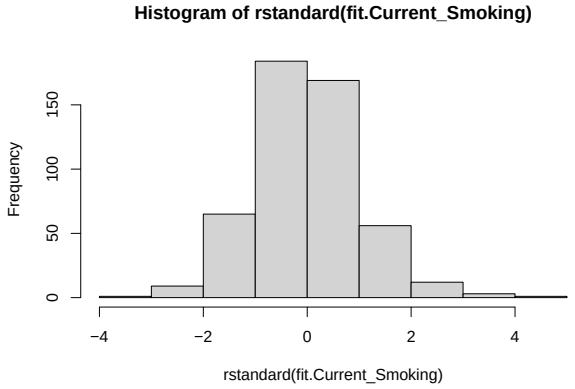
Provjerimo da pretpostavke modela (normalnost reziduala i homogenost varijance) nisu jako narušene.

Normalnost reziduala ispitat ćemo pomoću histograma i QQ-plota, koji nam mogu nam dati njihov okvirni distribucijski dojam. Reziduale je dobro prikazati u ovisnosti o procijenjenim modelima kako bismo ispitali homogenost varijance.

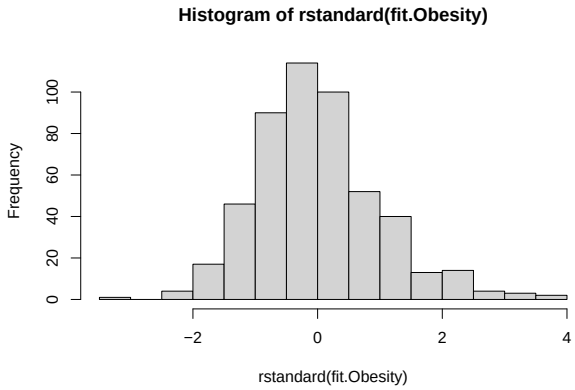
```
# Normalnost reziduala

#histogrami normalnosti

hist(rstandard(fit.Binge_Drinking))
```



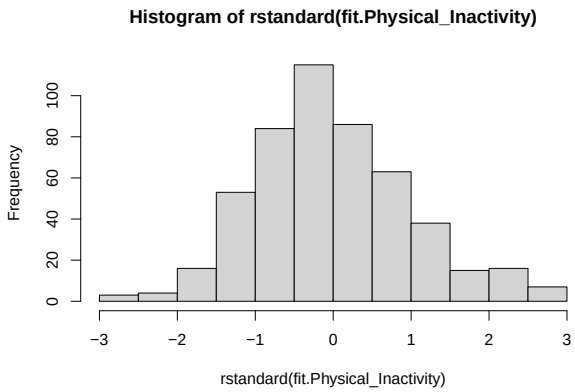

```
hist(rstandard(fit.Current_Smoking))
```



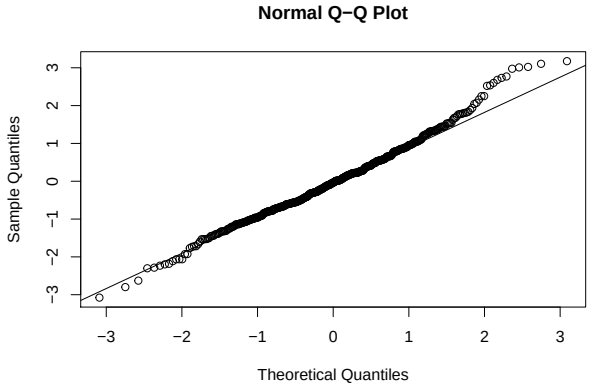

```
hist(rstandard(fit.Obesity))
```



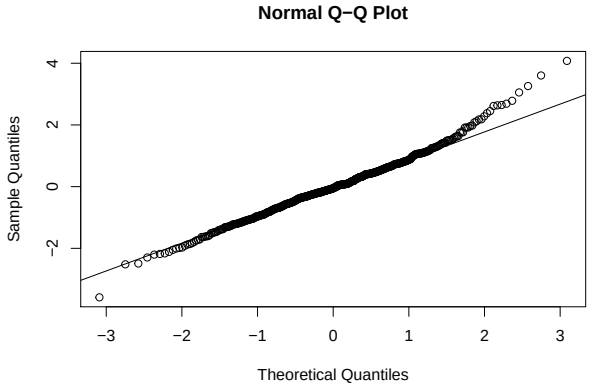

```
hist(rstandard(fit.Physical_Inactivity))
```



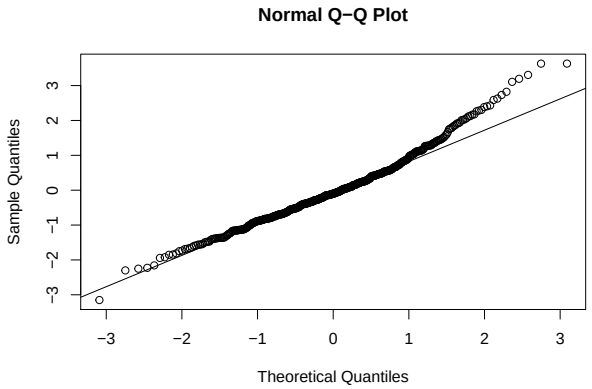

```
## QQ plot

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



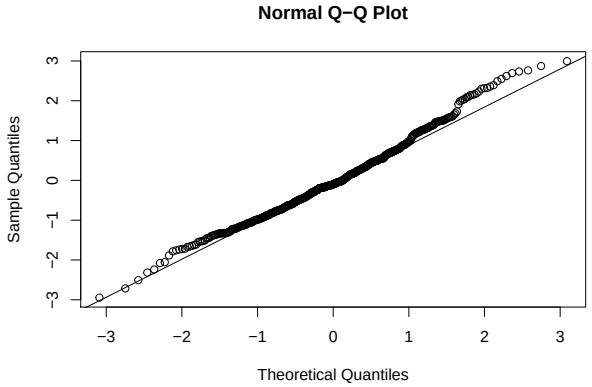

```
qqnorm(rstandard(fit.Current_Smoking))
qqline(rstandard(fit.Current_Smoking))
```



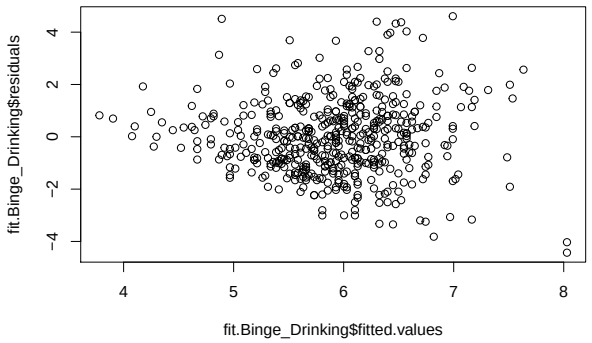

```
qqnorm(rstandard(fit.Obesity))
qqline(rstandard(fit.Obesity))
```



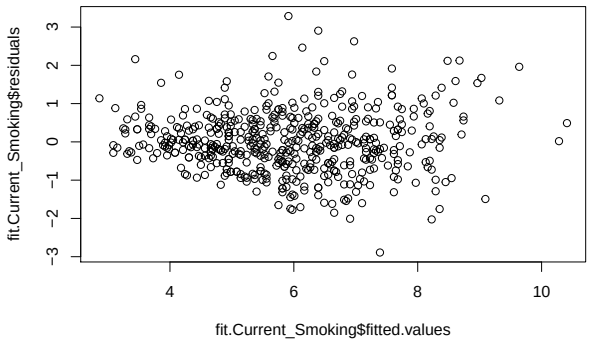

```
qqnorm(rstandard(fit.Physical_Inactivity))
qqline(rstandard(fit.Physical_Inactivity))
```



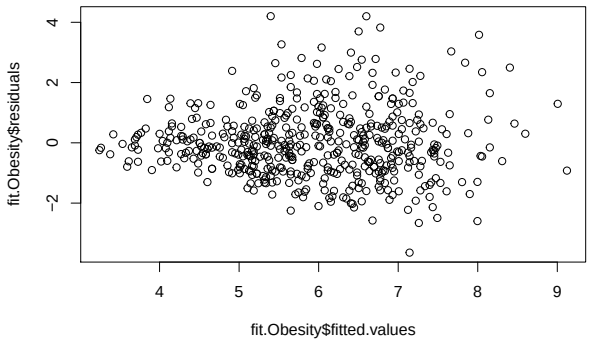

```
# Homogenost Varijance

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



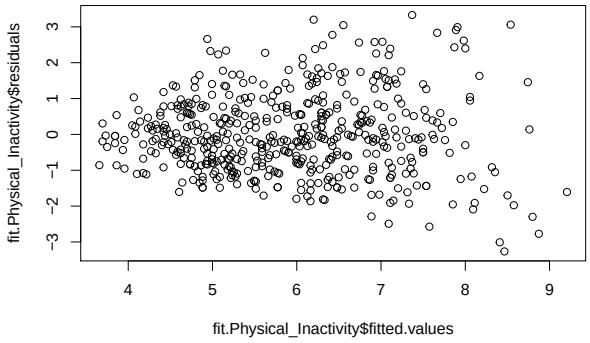

```
plot(fit.Current_Smoking$fitted.values, fit.Current_Smoking$residuals)
```




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




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



Histogrami pokazuju da reziduali nemaju previše iskrivljene distribucije, a iz QQ-plotova vidimo da nemaju previše teške repove, iz čega možemo zaključiti da reziduali imaju približno normalnu distribuciju, što je u redu jer su testovi koje ovdje koristimo u analizi regresijskih modela ionako robusni na normalnost.

Iz grafova procijenjenih vrijednosti i reziduala vidimo da varijanca ne varira toliko jako da bismo trebali posumnjati u njenu homogenost.

Zaključujemo da su pretpostavke normalnosti reziduala i homogenosti varijance zadovoljene.

Analizirajmo detaljnije procijenjene modele.

```
summary(fit.Binge_Drinking)
```

```
##
## Call:
## lm(formula = COPD ~ Binge_Drinking)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4295 -0.9766 -0.0542  0.8520  4.6082
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    10.10470    0.43640  23.155   <2e-16 ***
## Binge_Drinking -0.24705    0.02549  -9.692   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.457 on 498 degrees of freedom
## Multiple R-squared:  0.1587, Adjusted R-squared:  0.157
## F-statistic: 93.94 on 1 and 498 DF,  p-value: < 2.2e-16
```

```
summary(fit.Current_Smoking)
```

```
##
## Call:
## lm(formula = COPD ~ Current_Smoking)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8904 -0.5137 -0.0358  0.4669  3.2872
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.259365   0.153995   1.684   0.0928 .
## Current_Smoking 0.321216   0.008491  37.828   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8072 on 498 degrees of freedom
## Multiple R-squared:  0.7418, Adjusted R-squared:  0.7413
## F-statistic:  1431 on 1 and 498 DF,  p-value: < 2.2e-16
```

```
summary(fit.Obesity)
```

```
##
## Call:
## lm(formula = COPD ~ Obesity)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6431 -0.7862 -0.1199  0.6144  4.2027
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.334353   0.272069   1.229     0.22
## Obesity     0.193982   0.009272  20.922   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.159 on 498 degrees of freedom
## Multiple R-squared:  0.4678, Adjusted R-squared:  0.4667
## F-statistic: 437.7 on 1 and 498 DF,  p-value: < 2.2e-16
```

```
summary(fit.Physical_Inactivity)
```

```
##
## Call:
## lm(formula = COPD ~ Physical_Inactivity)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -3.2647 -0.7948 -0.1073  0.6400  3.3303
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.189795   0.214755    5.54  4.9e-08 ***
## Physical_Inactivity 0.185584   0.008191   22.66  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.115 on 498 degrees of freedom
## Multiple R-squared:  0.5076, Adjusted R-squared:  0.5066
## F-statistic: 513.3 on 1 and 498 DF,  p-value: < 2.2e-16
```

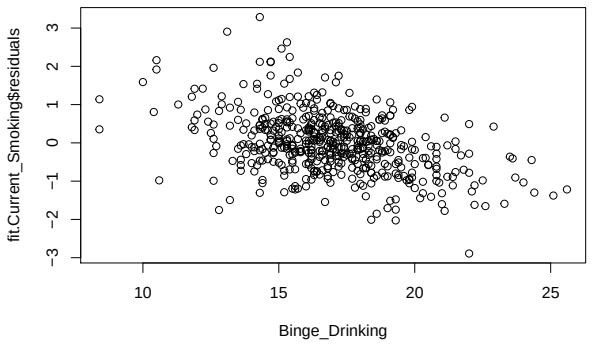
Vidimo da na temelju dobivene p-vrijednosti ne možemo odbaciti nultu hipotezu da je kod modela s regresorom Obesity i modela s regresorom Current_Smoking slobodni parametar jednak nuli.

Jako male p-vrijednosti kod F-testova za svaki od modela govore nam da možemo odbaciti hipotezu da su modeli neznačajni.

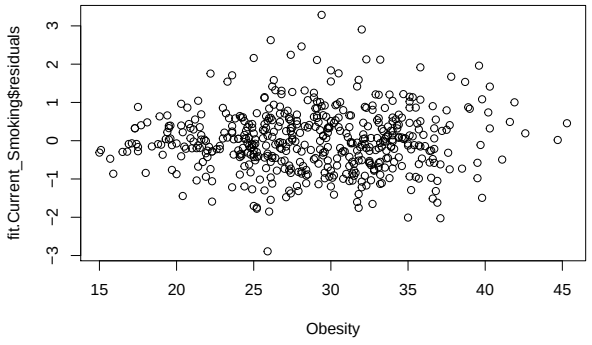
Na temelju R^2 vrijednosti vidimo da je model s regresorom Current_Smoking najbolji od 4 ispitana modela jednostruke regresije, dok regresor Binge_Drinking najslabije objašnjava varijancu u podacima s $R^2 = 0.157$.

Prije nego donesemo odluku o najboljem modelu, provjerit ćemo ovisnost reziduala modela s regresorom Current_Smoking (fit.Current_Smoking$residuals) s ostalim varijablama.

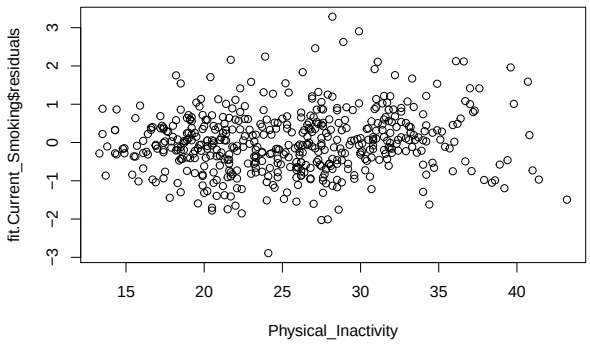
```
plot(Binge_Drinking, fit.Current_Smoking$residuals)
```




```
plot(Obesity, fit.Current_Smoking$residuals)
```



```
plot(Physical_Inactivity, fit.Current_Smoking$residuals)
```



Vidimo, međutim, da postoji neka zavisnost između promatranih reziduala i varijable Binge_Drinking, što nas potiče na razmišljanje da postoji ovisnost o još nekim regresorima, stoga ćemo dalje u nastavku isprobati nekoliko višeregresorskih modela.

Isprobat ćemo neke višeregresorske modele, no prije toga ćemo provjeriti korelacijske koeficijente među regresorima pomoću matrice korelacija.

```
cor(cbind(Binge_Drinking, Current_Smoking, Obesity,Physical_Inactivity))
```

```
##                     Binge_Drinking Current_Smoking    Obesity
## Binge_Drinking           1.0000000      -0.2067240 -0.2765743
## Current_Smoking         -0.2067240       1.0000000  0.7749085
## Obesity                 -0.2765743       0.7749085  1.0000000
## Physical_Inactivity     -0.4509752       0.7544748  0.8173297
##                     Physical_Inactivity
## Binge_Drinking                -0.4509752
## Current_Smoking                0.7544748
## Obesity                        0.8173297
## Physical_Inactivity            1.0000000
```

Vidimo da postoji velika korelacija između regresora Obesity i Physical_Inactivity, što smo i očekivali s obzirom da je poznato da manjak fizičke aktivnosti povećava rizik od pretilosti. Međutim, nismo očekivali ovako veliku korelaciju između regresora Current_Smoking i Obesity te Current_Smoking i Physical_Inactivity. Za najveću dopuštenu granicu korelacije među ulaznim varijablama uzeli smo korelaciju od 80%.

Proučimo sada neke višeregresorske modele.

Odlučili smo se za kombinacije regresora među kojima je Current_Smoking jer se pokazao najznačajnijim regresorom kod jednostrukih modela (međutim, isprobali smo i kombinacije u kojima nema tog regresora, za svaki slučaj).

Prikažimo neke od modela za koje smatramo da bi mogli biti dobri.

```
# Kombinacija regresora Current_Smoking i Obesity
fit.multi_1_2 = lm(COPD ~ Current_Smoking + Obesity)
summary(fit.multi_1_2)
```

```
##
## Call:
## lm(formula = COPD ~ Current_Smoking + Obesity)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -2.8016 -0.5110 -0.0270  0.4681  3.2799
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.13211    0.18964   0.697    0.486
## Current_Smoking  0.30926    0.01343  23.027   <2e-16 ***
## Obesity          0.01174    0.01021   1.149    0.251
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.807 on 497 degrees of freedom
## Multiple R-squared:  0.7425, Adjusted R-squared:  0.7415
## F-statistic: 716.6 on 2 and 497 DF,  p-value: < 2.2e-16
```

```
#plot(fit.multi_1_2$fitted.values, fit.multi_1_2$residuals)


# Kombinacija regresora Current_Smoking i Physical_Inactivity
fit.multi_1_3 = lm(COPD ~ Current_Smoking + Physical_Inactivity)
summary(fit.multi_1_3)
```

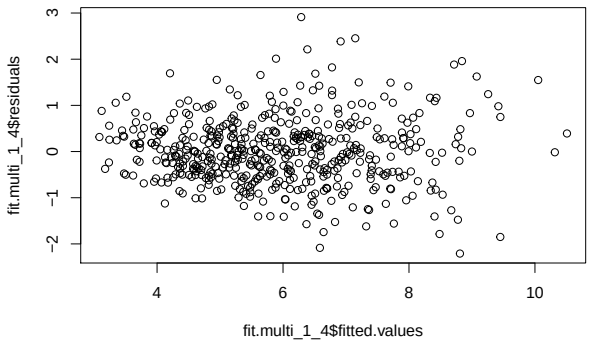
```
##
## Call:
## lm(formula = COPD ~ Current_Smoking + Physical_Inactivity)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6504 -0.5163 -0.0344  0.4389  3.1838
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.014867   0.161915   0.092    0.927
## Current_Smoking     0.280318   0.012720  22.037  < 2e-16 ***
## Physical_Inactivity 0.037862   0.008885   4.261 2.43e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7937 on 497 degrees of freedom
## Multiple R-squared:  0.7509, Adjusted R-squared:  0.7499
## F-statistic: 749.2 on 2 and 497 DF,  p-value: < 2.2e-16
```

```
#plot(fit.multi_1_3$fitted.values, fit.multi_1_3$residuals)

# Kombinacija regresora Current_Smoking i Binge_Drinking
fit.multi_1_4 = lm(COPD ~ Current_Smoking + Binge_Drinking)
summary(fit.multi_1_4)
```

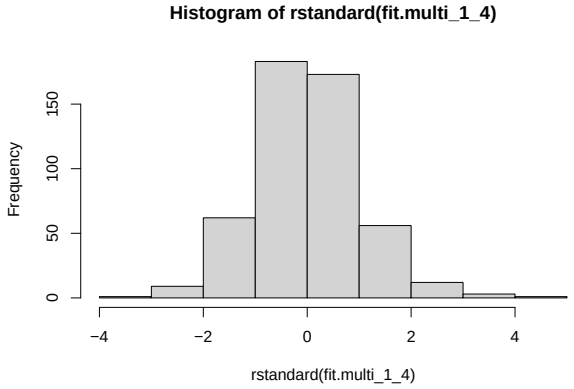
```
##
## Call:
## lm(formula = COPD ~ Current_Smoking + Binge_Drinking)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.20900 -0.47998 -0.01721  0.41242  2.91149
```

```
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.988499   0.283560   10.54   <2e-16 ***
## Current_Smoking   0.303471   0.007788   38.97   <2e-16 ***
## Binge_Drinking   -0.142733   0.012950  -11.02   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7244 on 497 degrees of freedom
## Multiple R-squared:  0.7925, Adjusted R-squared:  0.7917
## F-statistic: 949.3 on 2 and 497 DF,  p-value: < 2.2e-16
```

```
plot(fit.multi_1_4$fitted.values, fit.multi_1_4$residuals)
```




```
hist(rstandard(fit.multi_1_4))
```




```
# Kombinacija regresora Current_Smoking i Physical_Inactivity i Binge_Drinking
fit.multi_1_3_4 = lm(COPD ~ Current_Smoking + Physical_Inactivity + Binge_Drinking)
summary(fit.multi_1_3_4)
```

```
##
## Call:
## lm(formula = COPD ~ Current_Smoking + Physical_Inactivity + Binge_Drinking)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -2.23905 -0.47678 -0.01872  0.41510  2.91473
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3.071504   0.339956   9.035   <2e-16 ***
## Current_Smoking      0.307479   0.011933  25.767   <2e-16 ***
## Physical_Inactivity -0.004052   0.009137  -0.443    0.658
## Binge_Drinking      -0.145706   0.014591  -9.986   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7249 on 496 degrees of freedom
## Multiple R-squared:  0.7926, Adjusted R-squared:  0.7914
## F-statistic: 631.9 on 3 and 496 DF,  p-value: < 2.2e-16
```

```
# plot(fit.multi_1_3_4$fitted.values, fit.multi_1_3_4$residuals)
```

Podatci o modelu s tri regresora pokazuju da na razini značajnosti od 5% možemo odbaciti tvrdnju da je varijabla Physical_Inactivity značajna, što nismo očekivali s obzirom na to da je u modelu jednostruke regresije davala bolji model od modela s regresorom Binge_Drinking. Isti rezultat daje i troregresijski model s regresorom Obesity. Na temelju toga zaključujemo da modele s ta tri regresora nema smisla koristiti.

Zanimljivo je primijetiti da je Binge_Drinking kao samostalni regresor davao najlošiji model, ali u kombinaciji s regresorom Current_Smoking daje najveću vrijednost R^2. To možemo objasniti činjenicom da je korelacijski koeficijent za regresore Current_Smoking i Binge_Drinking puno manji od koeficijenata ostalih kombinacija regresora, stoga oni opisuju različite efekte na promatranu bolest.

Promatrajući R^2 vrijednosti dobivenih modela, zaključujemo da je najbolji model s regresorima Current_Smoking i Binge_Drinking jer ima najveću R^2 vrijednost te objašnjava približno 79.17% varijance u podatcima.