Životni vijek ljudi

Sapali grad

2022-12-15

U ovom izvještaju baviti ćemo se analizom podataka vezanih uz svjetske zdravstvene standarde, životno očekivanje i faktore koji na njih utječu.

Učitavanje i početna analiza podataka

```
head(data)
```

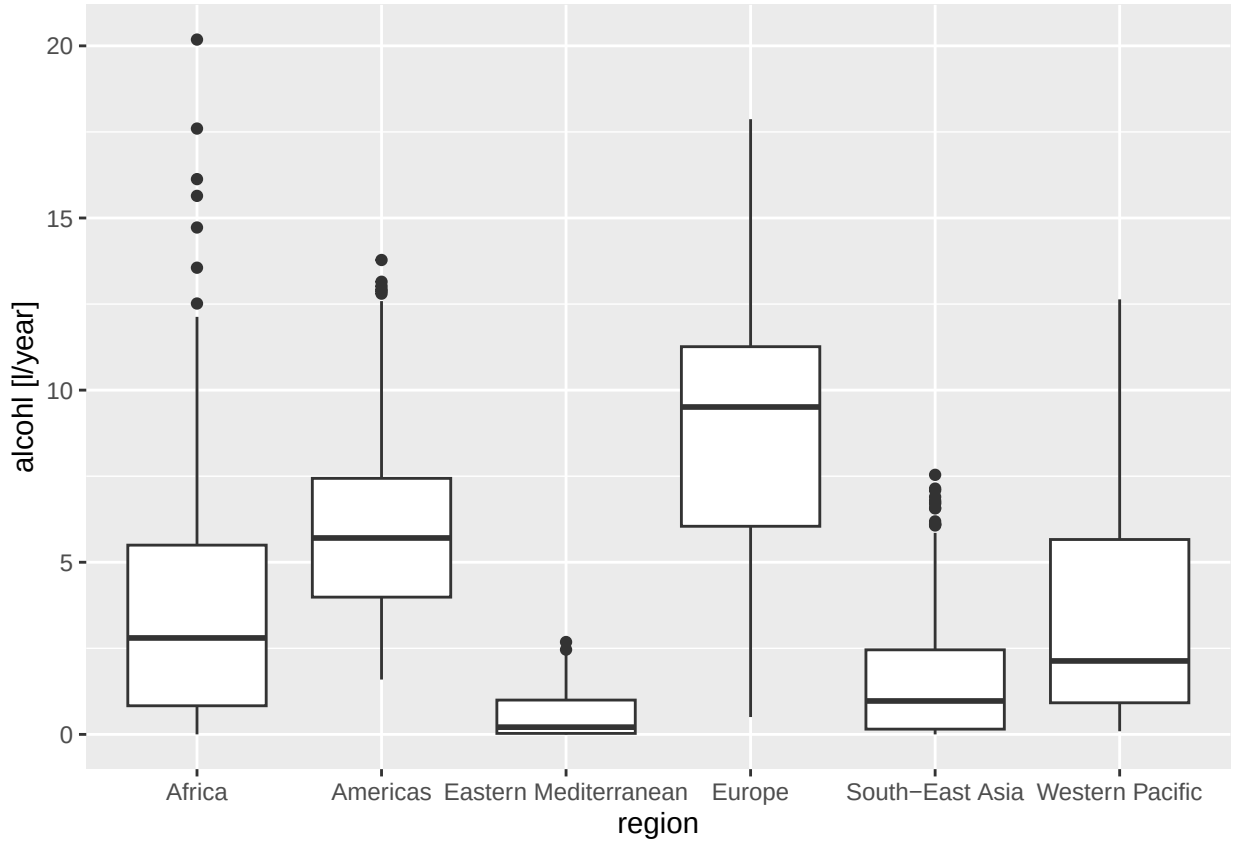
```
## # A tibble: 6 x 32
##   country country~1 region  year life_~2 life_~3 adult~4 infan~5 age1-~6 alcohol
##   <chr>   <chr>     <chr>  <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 Angola  AGO       Africa  2000    47.3    14.7    384.   0.138  0.0257    1.47
## 2 Angola  AGO       Africa  2001    48.2    15.0    372.   0.134  0.0245    1.94
## 3 Angola  AGO       Africa  2002    49.4    15.2    355.   0.128  0.0233    2.08
## 4 Angola  AGO       Africa  2003    50.5    15.4    343.   0.122  0.0219    2.20
## 5 Angola  AGO       Africa  2004    51.5    15.6    334.   0.116  0.0205    2.41
## 6 Angola  AGO       Africa  2005    52.7    15.8    323.   0.109  0.0189    3.49
## # ... with 22 more variables: bmi <dbl>, `age5-19thinness` <dbl>,
## #   `age5-19obesity` <dbl>, hepatitis <dbl>, measles <dbl>, polio <dbl>,
## #   diphtheria <dbl>, basic_water <dbl>, doctors <dbl>, hospitals <dbl>,
## #   gni_capita <dbl>, `gghe-d` <dbl>, che_gdp <dbl>, une_pop <dbl>,
## #   une_infant <dbl>, une_life <dbl>, une_hiv <dbl>, une_gni <dbl>,
## #   une_poverty <dbl>, une_edu_spend <dbl>, une_literacy <dbl>,
## #   une_school <dbl>, and abbreviated variable names 1: country_code, ...
```

1. Postoji li razlika u konzumaciji alkohola tijekom godina medu svjetskim regijama?

Vidimo da podatke trebamo podijeliti na klase na temelju dvaju parametara te zatim testirati nultu hipotezu oblika "sve srednje vrijednosti su jednake".
Uočavamo da se ovdje vrlo vjerojatno radi o dvofaktorskoj ANOVI.
Započinjemo transformacijom i vizualizacijom podataka. Prikazati ćemo podatke grupirane po jednom pa po drugom parametru, tj. regiji pa godinama i zatim po oba parametra.

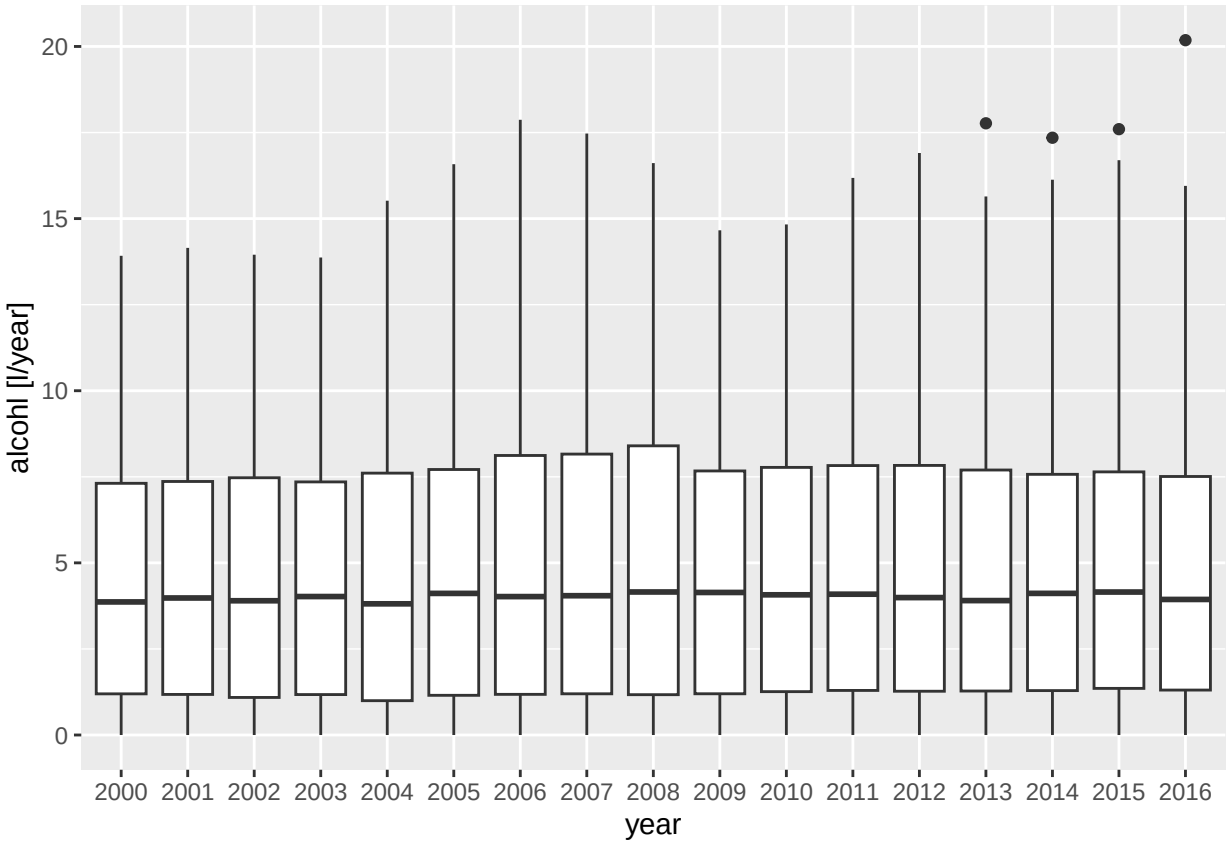
```
#Prikazivanje podataka
g1 <- ggplot(data, aes(x = region, y = alcohol)) + geom_boxplot() + labs(y = "alcohl [l/year]")
g2 <- ggplot(data %>% mutate(year = as.factor(year), alcohol),  aes(x = year, y = alcohol))+
  geom_boxplot() + labs(y = "alcohl [l/year]")

#Jednostavna transformacija i ujedinjavanje redaka za olakšavanje
#daljnje analize
inter = interaction(data$region,data$year)
classData <- data %>% unite(col =class, year, region, sep = "_", remove = T)
g3 <- ggplot(classData, aes(x = class, y = alcohol)) + geom_boxplot() +
  labs(y = "alcohl [l/year]") + theme_void()

g1
```

```
## Warning: Removed 50 rows containing non-finite values (‘stat_boxplot()‘).
```



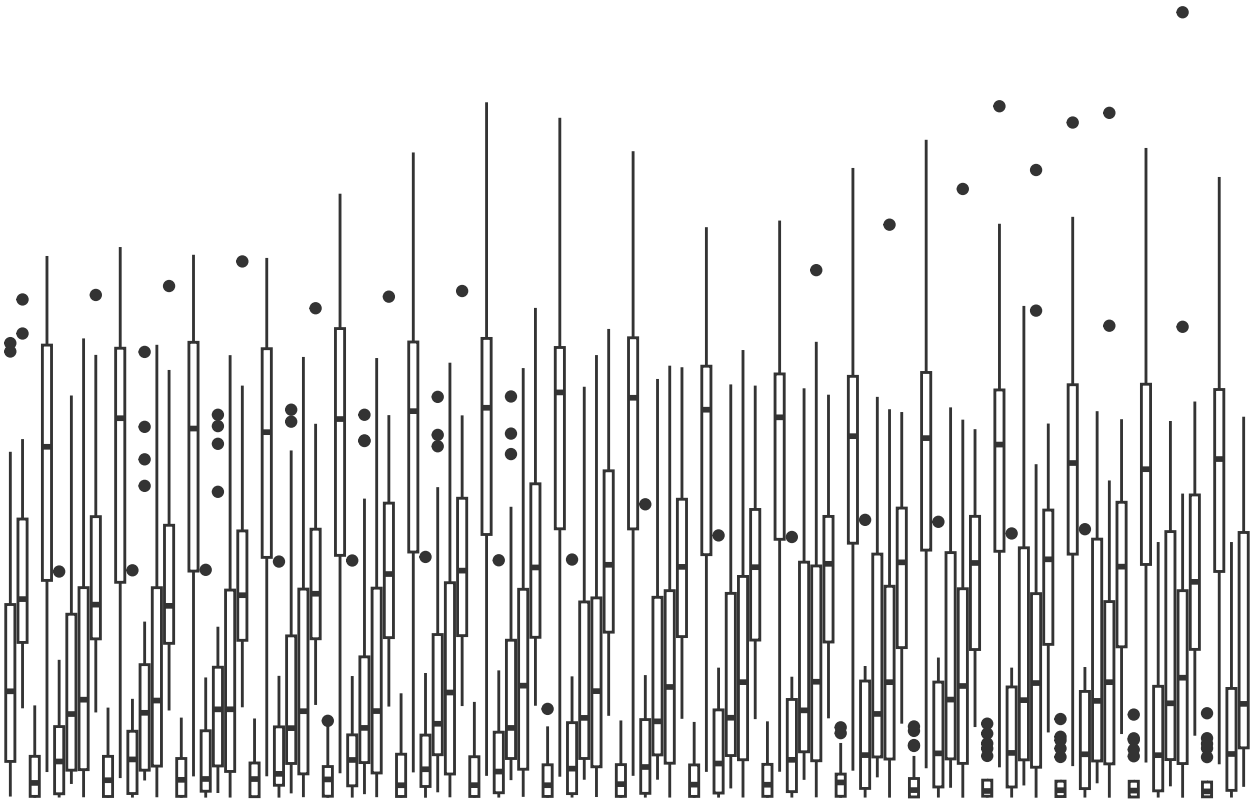
```
g2
```

```
## Warning: Removed 50 rows containing non-finite values (‘stat_boxplot()‘).
```




```
g3
```

```
## Warning: Removed 50 rows containing non-finite values (‘stat_boxplot()‘).
```



Primjećujemo da se podaci ponašaju izuzetno lijepo kad podatke klasificiramo po godinama, ali veoma neujednačeno u klasifikacije po regijama.
Uz to, makar je ispis boxplota koji je podjeljen po svim klasama veoma nečitljiv, možemo uočiti uzorak koji se ponavlja iz godine u godinu kako se regije izmijenjuju. Što je i za očekivati jer nismo uočili veliku varijabilnosti pri klasifikaciji po godinama, ali je ona prisutna kada klasificiramo po regijama
Sljedeći je korak provjera normalnosti unutar svake od 96 i klasa i provjera homoskedastičnosti među klasama.

Nismo uključili ispis 96 lillieforova testa, ali podaci pretežito nisu normalno distribuirani.
ANOVA na našu sreću, jest relativno robusna na kršenje uvjeta normalnosti, ali ju puno više smeta narušavanje homoskedastičnosti (pogotovo uz nejednake veličine uzoraka kao u našem slučaju).
Provjeravamo homoskedastičnost bartlettovim testom.

```
bartlett.test(classData$alcohol~classData$class)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  classData$alcohol by classData$class
## Bartlett's K-squared = 891.26, df = 101, p-value < 2.2e-16
```

Dobivamo p-value reda 10^{-16} te smo primorani odbaciti nul hipotezu da su varijance među klasama jednake.
Kako nam osnovne pretpostavke ne vrijede, okrećemo se neparametarskoj ANOVI, Kruskal-Wallis testu.
Valjalo bi napomenuti da makar ovdje klasificiramo po dva parametra, ustvari ćemo testiranje hipoteza svesti na jednostruku ANOVU koja se klasificira po jednom parametru, ali je taj paramtear uređenja dvojka (godina, regija).

Testiramo hipotezu:

$$H_0 : \mu_{2000,Afrika} = \mu_{2000,Americas} = ... = \mu_{2016,WesternPacific}$$
$$H_1 : baremjedan\mu_{i,j}nijejednak$$
$$\alpha = 0.05$$
$$i \in N, i \in [2000, 2016]$$
$$j \in \{Africa, Americas, ..., WesternPacific\}$$

```
kruskal.test(classData$alcohol~classData$class)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  classData$alcohol by classData$class
## Kruskal-Wallis chi-squared = 1519.9, df = 101, p-value < 2.2e-16
```

Provođenjem Kruskal-Wallis testa dobivamo p-value koji je debelo zakoračio u kritični odsječak te odbacujemo nul hipotezu i zaključujemo da postoji razlika u konzumaciji alkohola među svjetskim regijama i godinama.
Međutim, boxplotovi koje smo nacrtali nameću pitanje: Utječe li na naš zaključak uopće prisustvo klasifikacije po godinama?
Stoga provodimo jedan dodatan Kruskal-Wallis test kojim provjeravamo je li srednja vrijednost ista u ovisnosti o godinama.

```
kruskal.test(data$alcohol~data$year)
```

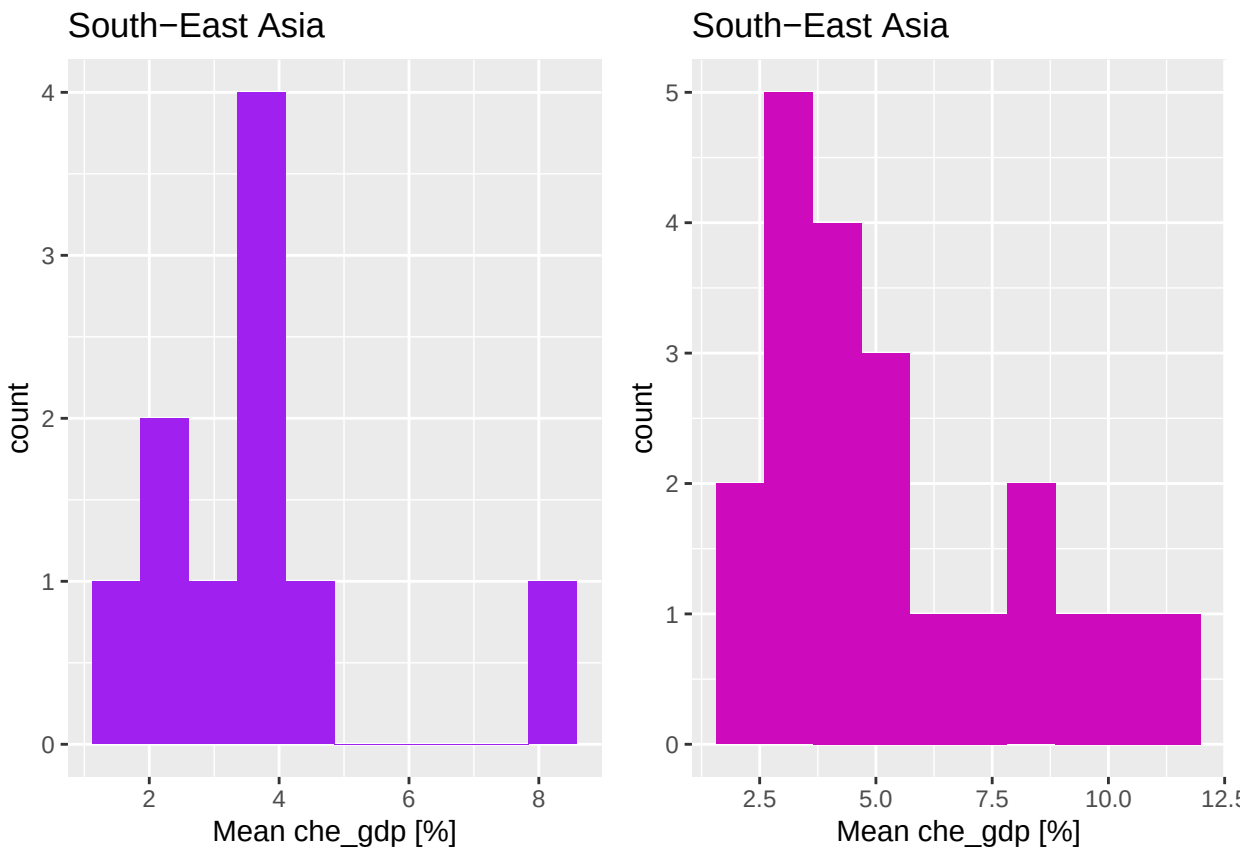
```
##
##  Kruskal-Wallis rank sum test
##
## data:  data$alcohol by data$year
## Kruskal-Wallis chi-squared = 1.29, df = 16, p-value = 1
```

Kao što smo i sumnjali, provođenjem drugog Kruskal-Wallis testa dobivamo p-value koji teži u 1 i zaključujemo da su srednje vrijednosti konzumacije alkohola jednake po godinama. Iz čega donosimo zaključak da je regija parametar koji zapravo utječe na razliku među srednjim vrijednostima.

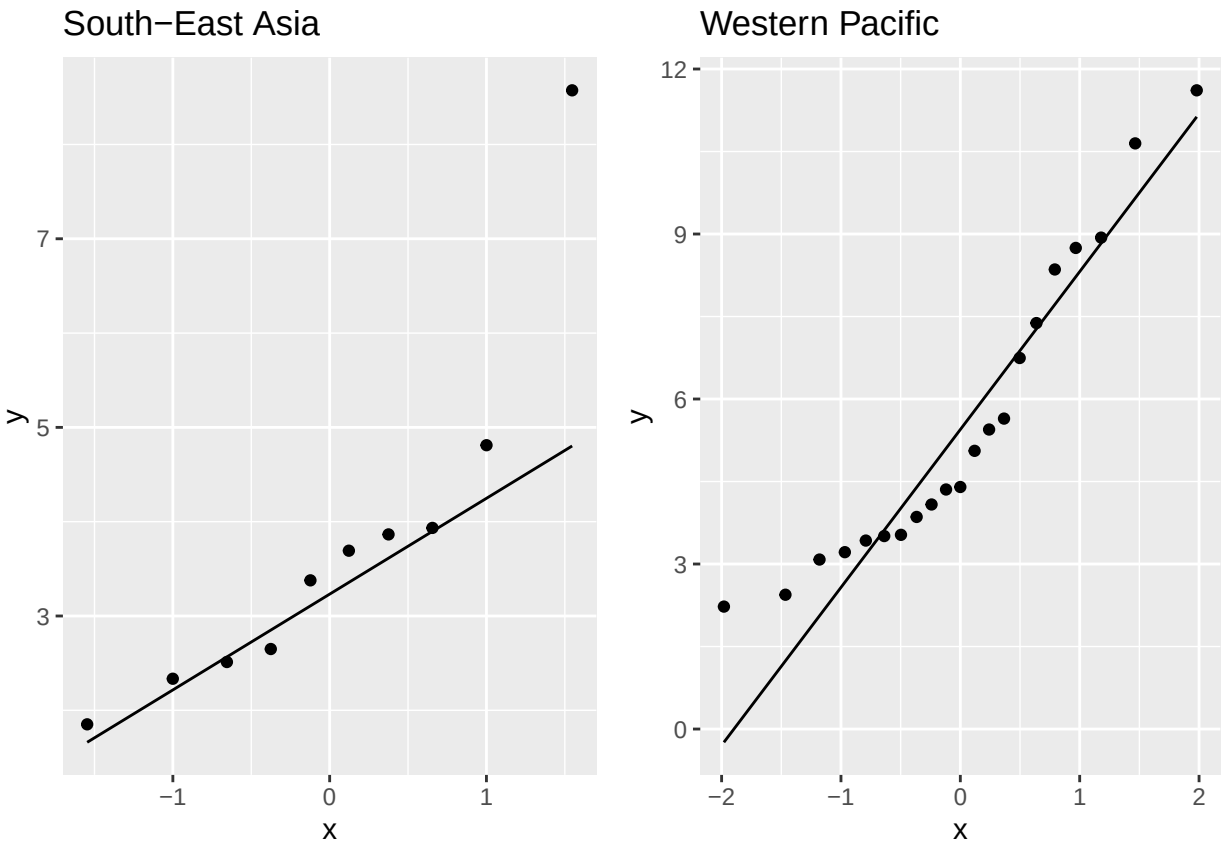
2. Razlikuje li se količina uloženog novca u zdravstvenu skrb izmedu zemalja jugoistočne Azije i zemalja zapadnog Pacifika? dodaj drugi parametar za zdravstvo

Pitanje nas odma navodi na ideju provođenja standardnog t-testa, ali prvo transformiramo podatke, provjeravamo pretpostavku normalnosti i pokušavamo ukloniti potencijalnu zavisnost.
Trebamo srednju vrijednost uloženog novca u zdravstvenu skrb u regijama Southeast Asia i Western Pacific. Kako naš uzorak sadrži očitanja iz više država u vremenskom periodu od više godina, bilo bi naivno očekivati da su rezultati iz iste države nezavisni kroz godine. Osim toga, neke države nemaju zabilježene tražene podatke u svim godinama te bi zbog toga došlo do disbalansa.
Stoga je naš prvi korak nakon grupiranja po regijama, grupiranje i po država i računanje srednje vrijednosti za svaku od njih.
Nakon toga provjeravamo jesu li podaci aproksimativno normalni u prirodi.

```
#Grupiranje podataka
seData <- filter(data, region == "South-East Asia" & !is.na(che_gdp)) %>%
  group_by(country) %>% summarise(mean_che = mean(che_gdp, na.rm = T)) %>%
  select(mean_che) %>% ungroup

wpData <- filter(data, region == "Western Pacific" & !is.na(che_gdp)) %>%
  group_by(country) %>% summarise(mean_che = mean(che_gdp, na.rm = T)) %>%
  select(mean_che) %>% ungroup

#Vizualizacija podataka
g1 <- ggplot(seData, aes(x = mean_che)) +
  geom_histogram(bins = 10, fill = "purple") +
  labs(title = "South-East Asia", x = "Mean che_gdp [%]")
g2 <- ggplot(wpData, aes(x = mean_che)) +
  geom_histogram(bins = 10, fill = 6) +
  labs(title = "South-East Asia", x = "Mean che_gdp [%]")

#Uređivanje ispisa
grid.arrange(g1, g2, ncol = 2)
```




```
#Generiranje qq grafova
g1 <- ggplot(seData, aes(sample = mean_che)) + geom_qq() +
  geom_qq_line() + labs(title = "South-East Asia")
g2 <- ggplot(wpData, aes(sample = mean_che)) + geom_qq() +
  geom_qq_line() + labs(title = "Western Pacific")

#Uređivanje ispisa
grid.arrange(g1, g2, ncol = 2)
```



Vidimo da se podaci ponašaju pretežito normalno s pokojim outlierom, ali t-test nije ništa ako ne robustan tako da postavljamo hipotezu.

$$H_0 : \mu_0 = \mu_1$$
$$H_1 : \mu_0 \neq \mu_1$$
$$\alpha = 0.05$$

Provodimo t-test sa razinom značajnosti 5%.

```
#Samostalno računamo sve podatke potrebne za provođenje testa kako se ne bi potkrala pogreška

length(seData) -> n0
length(wpData) -> n1

seData <- seData %>% unlist(use.names = F)
wpData <- wpData %>% unlist(use.names = F)

seData %>% mean(na.rm = T) -> mi0
wpData %>% mean(na.rm = T) -> mi1

seData %>% sd(na.rm = T) -> s0
wpData %>% sd(na.rm = T) -> s1

var.test(seData, wpData, alternative = "two.sided")
```

```
##
```

```
##  F test to compare two variances
##
## data:  seData and wpData
## F = 0.48757, num df = 9, denom df = 20, p-value = 0.2683
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.1718892 1.7878796
## sample estimates:
## ratio of variances
##          0.4875718
```

```
#Provodeći f-test donosimo zaključak da nema značajne razlike među varijancama
#dvaju uzoraka i stoga koristimo "Pooled t-statistiku"

cat("mi0:", mi0, " s0: ", s0, " n0:", n0, "\n")
```

```
## mi0: 3.760256  s0:  1.912992  n0: 1
```

```
cat("mi1:", mi1, " s1: ", s1, " n1:", n1, "\n")
```

```
## mi1: 5.556877  s1:  2.739643  n1: 1
```

```
cat("stupnjevi slobode:", as.character(n0 + n1 - 2) , "\n")
```

```
## stupnjevi slobode: 0
```

```
#Provodimo t-test
t.test(seData, wpData, conf.level = 0.95, var.equal = T)
```

```
##
##  Two Sample t-test
##
## data:  seData and wpData
## t = -1.8612, df = 29, p-value = 0.07288
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.770851  0.177608
## sample estimates:
## mean of x mean of y
##  3.760256  5.556877
```

Kako smo proveli test s razinom značajnosti 5%, a naš p-value iznosi 7.28%, ne odbacujemo početnu hipotezu i donosimo zaključak da se količina uloženog novca u zdravstvenu skrb ne razlikuje među dvije regije. Ipak, ovo je odličan pokazatelj toga zašto se rezultati statističih istraživanja trebaju uzimati sa zrnom soli. Da smo samo malo podesili razinu značajnosti ili odbacili pokoje stršeće vrijednosti u ime očuvanja distribucije podataka defintivno smo mogli donijeti oprečan zaključak kada bi nam to bilo u interesu.

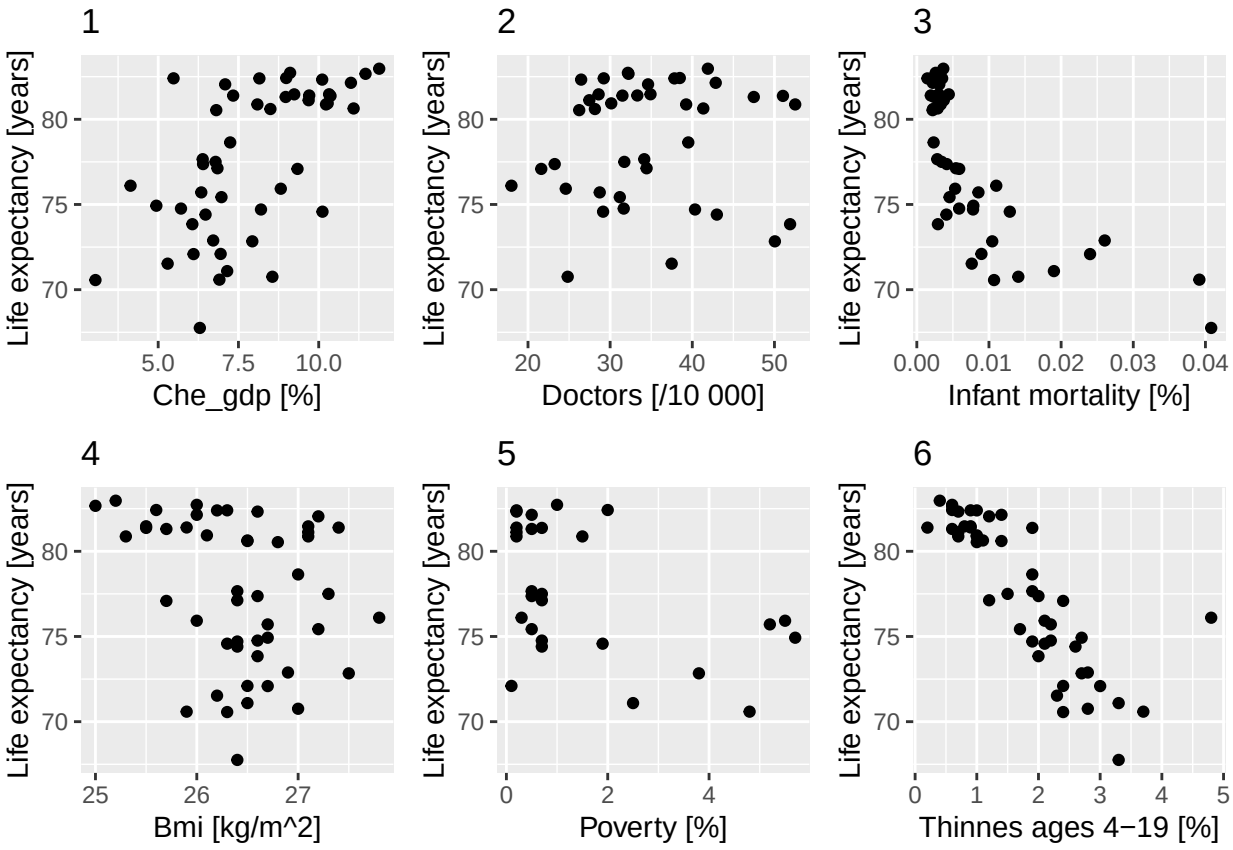
3. Može li se na temelju zadanih parametara objasniti očekivani životni vijek ljudi u Europi u 2015. godini?

Pitanje nas navodi na traženje neke međuovisnosti između životnog očekivanja i nekih od parametara. Kao i uvijek počinjemo vizualizacijom podataka i nadamo se da ćemo uočiti neku linearnu vezu ili vezu koju ćemo nekom transforamcijom moći učiniti linearnom.
Neki od podataka koje intuitivno očekujemo da bi mogli imati utjecaj na životno očekivanje su, ulaganje u zdravstvo, količina doktora, infantilni mortalitet, i slični.

```
#Filtriramo podatke
linRegData <- data %>% filter(year == "2015", region == "Europe",
                              !is.na(life_expect), !is.na(che_gdp))
linRegData$obesity <- linRegData$`age5-19obesity`
linRegData$thinness <- linRegData$`age5-19thinness`

#Stvaramo scatter-plotove
g1 <- ggplot(linRegData, aes(x = che_gdp, y = life_expect)) +
  geom_point() + labs(title = 1, y = "Life expectancy [years]"
                      , x = "Che_gdp [%]")
g2 <- ggplot(linRegData, aes(x = doctors, y = life_expect)) +
  geom_point() + labs(title = 2, y = "Life expectancy [years]"
                      , x = "Doctors [/10 000]")
g3 <- ggplot(linRegData, aes(x = infant_mort, y = life_expect)) +
  geom_point() + labs(title = 3, y = "Life expectancy [years]"
                      , x = "Infant mortality [%]")
g4 <- ggplot(linRegData, aes(x = bmi, y = life_expect)) +
  geom_point() + labs(title = 4, y = "Life expectancy [years]"
                      , x = "Bmi [kg/m^2]")
g5 <- ggplot(linRegData, aes(x = une_poverty, y = life_expect)) +
  geom_point() + labs(title = 5, y = "Life expectancy [years]"
                      , x = "Poverty [%]")
g6 <- ggplot(linRegData, aes(x = thinness, y = life_expect)) +
  geom_point() + labs(title = 6, y = "Life expectancy [years]"
                      , x = "Thinnes ages 4-19 [%]")

#Redamo ih na lakše pregledan grid
grid.arrange(g1, g2, g3, g4, g5, g6, ncol = 3)
```

```
## Warning: Removed 8 rows containing missing values (‘geom_point()‘).
```

```
## Warning: Removed 21 rows containing missing values (‘geom_point()‘).
```



Uočavamo da se u prvom grafu nazire nekakva linearna veza koju bi očekivali u stvarnim podacima.
U drugom i četvrtom grafu je varijabilnost životnog očekivanja pri fiksiranju vrijednosti na x osi maltene jednaka za svaki x što je jako dobar indikator da ne postoji linearna veza.
U trećem pak grafu isto uočavamo vezu koja je potencijalno polinomijalne prirode, i nečemo ju ubacivati u naš linearni model.
Peti graf ne pokazuje obećavajuće znakove linearne ovisnosti.
U šestom grafu vidimo veoma očite naznake linearne ovisnosti.

Fokusirati ćemo se na prvi i šesti graf, tj kako thinness i ulaganje države u zdravstvo utječe na životno očekivanje. Koristimo linearni model s dva regresora.
Pri tome bi trebali provjeriti pretpostavke Linearna regresije da su reziduali iz normalne distribuirane s jednakom varijancom i očekivanjem 0. Nažalost homogenost varijance ćemo teško provjeriti na skupu ove veličine.

```
#Radimo model
linModel <- lm(life_expect~che_gdp+thinness, linRegData)
summary(linModel)
```

```
##
## Call:
## lm(formula = life_expect ~ che_gdp + thinness, data = linRegData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2587 -0.9687 -0.0173  1.0454  9.6818
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  80.0099     2.1853  36.613  < 2e-16 ***
## che_gdp       0.3900     0.2085   1.870    0.068 .
## thinness     -3.1674     0.4220  -7.505 1.83e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.337 on 45 degrees of freedom
## Multiple R-squared:  0.7178, Adjusted R-squared:  0.7052
## F-statistic: 57.23 on 2 and 45 DF,  p-value: 4.343e-13
```

```
#Provjeravamo pretpostavaku linearne regresije da su reziduali iz normalne
#razdiobe
lillie.test(linModel$residuals)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  linModel$residuals
## D = 0.11223, p-value = 0.1363
```

Sada testiramo hipotezu:

$$H_0 : \mu_e = 0$$
$$H_1 : \mu_e \neq 0$$
$$\alpha = 0.05$$

```
#Provjeravamo je li očekivanje 0 sa razinom značajnosti 0.05
t.test(linModel$residuals, conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  linModel$residuals
## t = -3.9251e-16, df = 47, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.663875  0.663875
## sample estimates:
##     mean of x
## -1.295283e-16
```

Sa šokantnom p-vrijednošću iznosa 1, ne odbacujemo srednju nul hipotezu i donosimo zaključak da su pretpostavke modela ispunjene.
Model s dva regresora che_gdp i thinness nam daje $R^2 = 0.72$ što je izuzetno dobro. Uz to lillieforsov test normalnosti proveden nad rezidualima daje p-value od 0.1363 te nemama potrebe odbaciti nultu hipotezu tog testa koja jest da su reziduali normalno distribuirani.
Ipak, moramo bit svijesni činjenice da dodavanjem regresora u model skoro sigurno povećavamo faktor R^2 neovisno o tome ima li taj regresor zapravo ikakav utjecaj na promatrani parametar. Stoga moramo biti oprezni pri dodavanju modela jer dodavanjem "garbage" podataka možemo stvoriti iluziju da gradimo dobar model. Iznos
Tu u igru ulazi $adjusted R^2$ čija je zadaća penaliziranje ocjene modela pri dodavanju regresora koji ne

povećavaju R^2 onoliko koliko bi bilo očekivano. Stoga bi kako ubacujemo regresore u naš model, trebali sve više i više pozornosti obraćati na $adjustedR^2$.
Naš $adjustedR^2$ iznosi otprilike 0.71 što je dovoljno blizu iznosu našeg R^2 te nas navodi na zaključak da naš model ima dobru snagu predviđanja

Svoj model s dva regresora tada možemo opisati na sljedeči način:

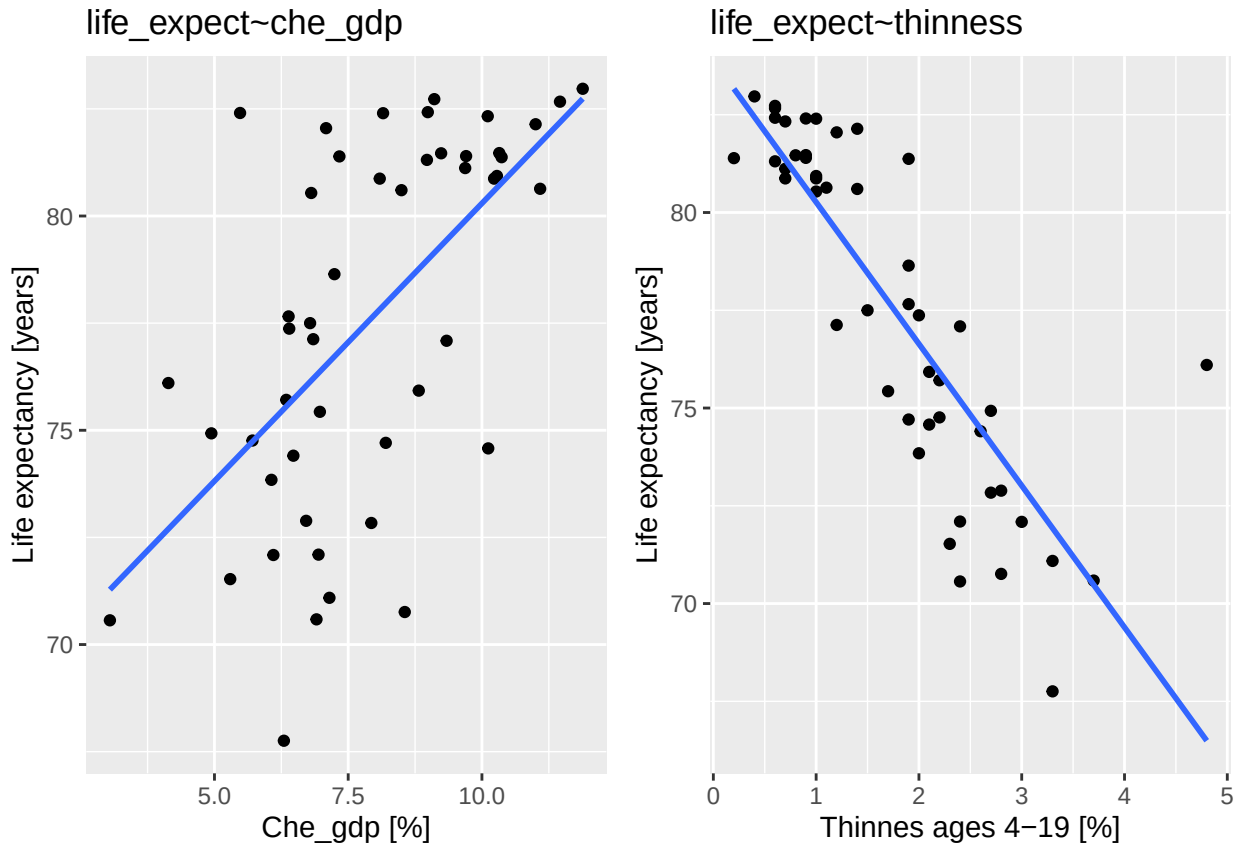
$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + e_i$$

gdje x_1 predstavlja gdp_che, x_2 thinness, a e rezidual. S uvrštenim koeficijentima koje vidimo kao izlaz funkcije summary(linModel) model izgleda ovako:

$$\hat{y}_i = 80 + 0.39 x_1 - 3.17 x_2 + e_i$$

Iz formule i grafova koji slijede je jasno da postoji pozitivna linearna veza između državnog ulaganja u zdravstveni sustav i očekivanog trajanja života, te negativna linearna veza između stopa mršavosti i očekivanog trajanja života.

```
g1 <- g1 + stat_smooth(formula = y~x, method = "lm", se = F) + labs(title = "life_expect~che_gdp")
g2 <- g6 + stat_smooth(formula = y~x, method = "lm", se = F) + labs(title = "life_expect~thinness")
grid.arrange(g1, g2, ncol = 2)
```



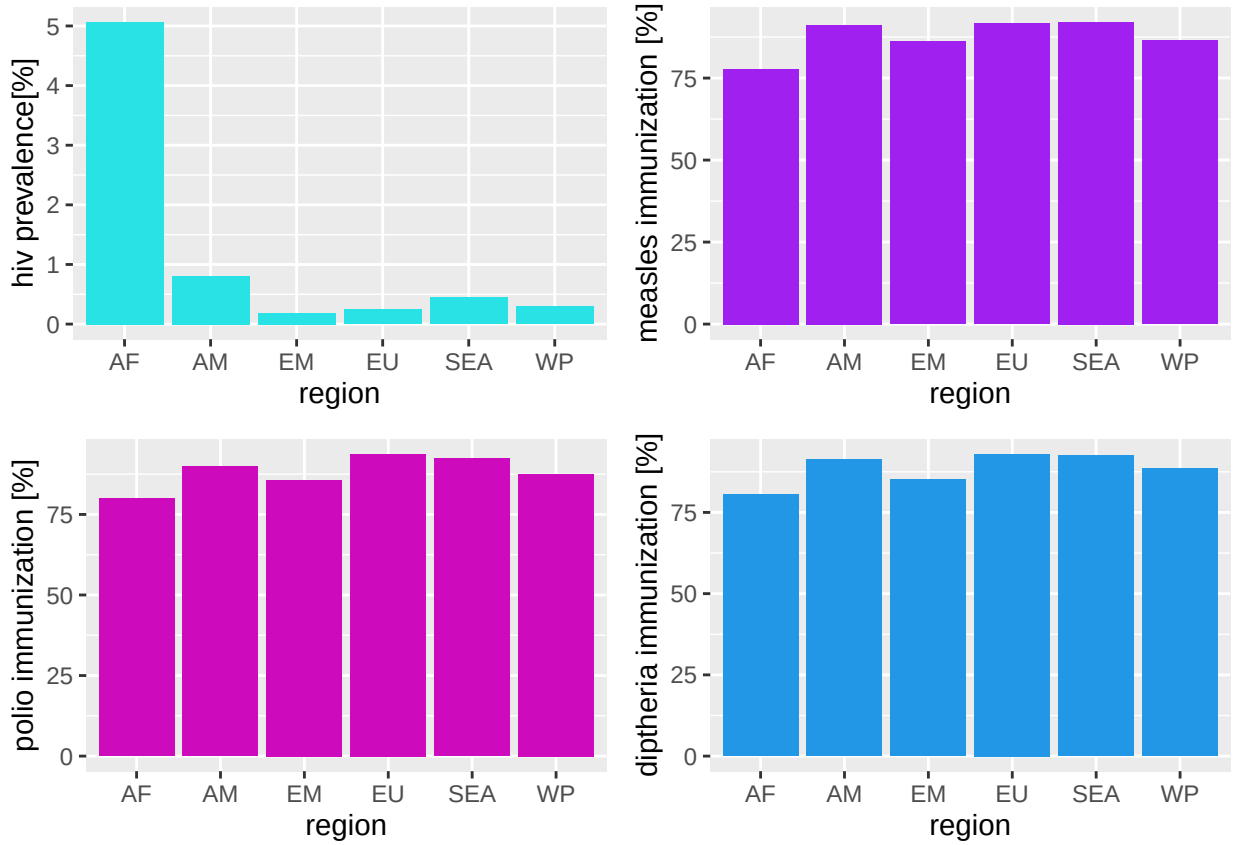
Ovo je ustvari veoma zdravorazumski rezultat.
Države koje više ulažu u javno zdravstvo pružaju prosječnom građaninu bolju zdravstvenu skrb, lijekove i uslugu. Dok pak pretjerana mršavost ima znatne negativne učinke na zdravlje kao što su veći rizik visokog tlaka, dijabetesa, raznih drugih srčanih problema, pada imuniteta itd.

4. Ima li razlike u trendovima zaraženosti/imunizacije različitim bolestima medu svjetskim regijama?

Iz postavljenog pitanja, očito je da radimo jednoparametarsku analizu varijance.
Za ispitivanje ovog pitanja uzimati ćemo najsvježije podatke iz 2016. Prije početka ispitivanja filtrirat ćemo i vizualizirati podatke iz svih regija barplotovima kako bi dobili bolji uvid u cijelu situaciju.

```
medDataFiltered <- filter(data, year == 2016)
hivData <- medDataFiltered %>% group_by(region) %>%
  summarise(hiv = mean(une_hiv, na.rm = T)) %>% ungroup()
measlesData <- medDataFiltered %>% group_by(region) %>%
  summarise(measles = mean(measles, na.rm = T)) %>% ungroup()
polioData <- medDataFiltered %>% group_by(region) %>%
  summarise(polio = mean(polio, na.rm = T)) %>% ungroup()
diphtheriaData <- medDataFiltered %>% group_by(region) %>%
  summarise(diphtheria = mean(diphtheria, na.rm = T)) %>% ungroup()

#pomocna funkcija za jednokratnu transformaciju regija kako bi ispis
#grafova bio ljepsi

transformLabs <- function(labs){
  for(i in 1:length(labs)){
    if(labs[i] == "Africa"){
      labs[i] = "AF"
    } else if(labs[i] == "Americas"){
      labs[i] = "AM"
    } else if(labs[i] == "Eastern Mediterranean") {
      labs[i] = "EM"
    } else if(labs[i] == "Europe"){
      labs[i] = "EU"
    } else if(labs[i] == "South-East Asia"){
      labs[i] = "SEA"
    } else if(labs[i] == "Western Pacific"){
      labs[i] = "WP"
    }
  }

  return (labs)
}

g1 <- hivData %>% mutate(region = transformLabs(region), hiv) %>%
  ggplot(aes(x = region, y = hiv)) + geom_bar(stat = "identity", fill = 5) +
  labs(y = "hiv prevalence[%]")

g2 <- measlesData %>% mutate(region = transformLabs(region), measles) %>%
  ggplot(aes(x = region, y = measles)) +
  geom_bar(stat = "identity", fill = "purple") +
  labs(y = "measles immunization [%]")

g3 <- polioData %>% mutate(region = transformLabs(region), polio) %>%
  ggplot(aes(x = region, y = polio)) +
  geom_bar(stat = "identity", fill = 6) + labs(y = "polio immunization [%]")

g4 <- diphtheriaData %>% mutate(region = transformLabs(region), diphtheria) %>%
```

```
  ggplot(aes(x = region, y = diphtheria)) +
  geom_bar(stat = "identity", fill = 4) +
  labs(y = "diptheria immunization [%]")

grid.arrange(g1, g2, g3, g4, ncol = 2)
```

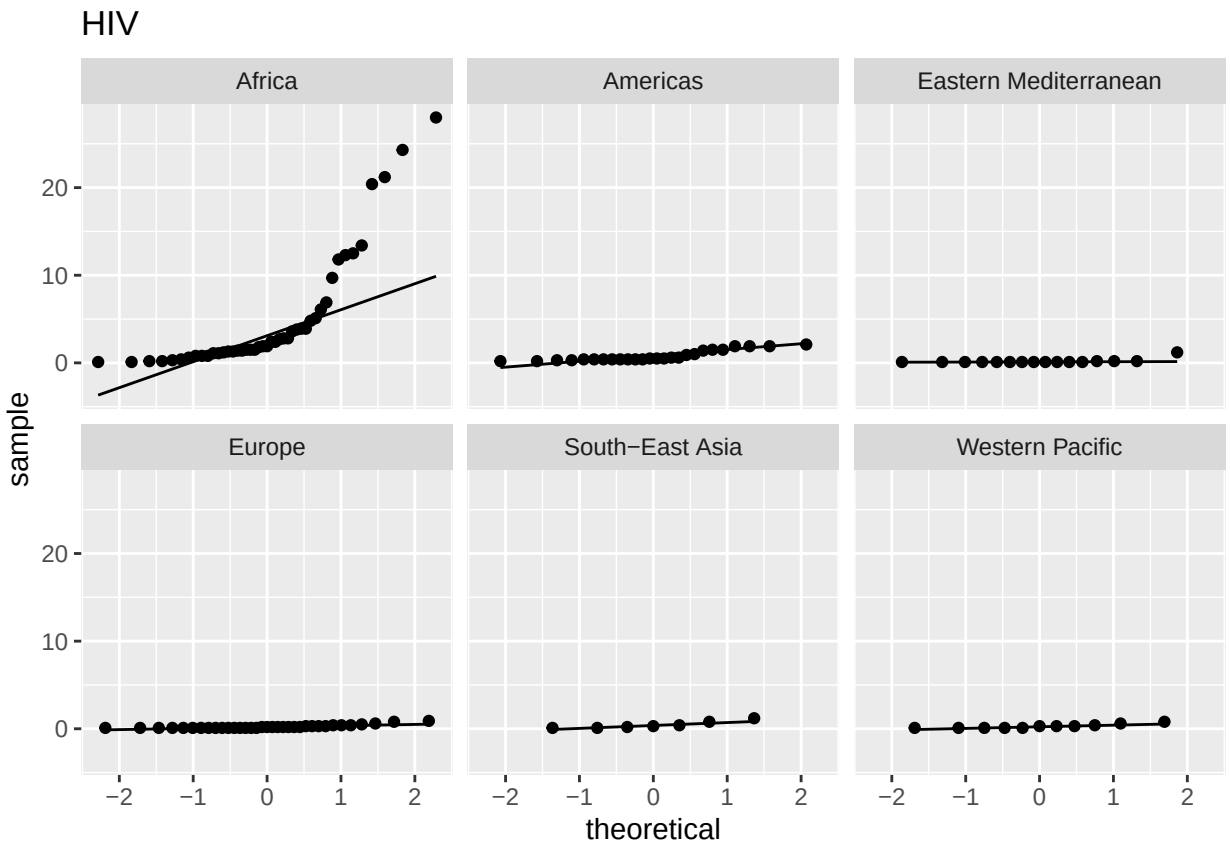


Gledajući na ovakvoj skali, podaci se, osim onih za hiv, čine relativno ujednačenima. Ipak tako se čine, zbog skale. U stvarnosti podaci nisu pretjerano ujednačeni i već sad očekujemo da će ANOVA pokazati da postoji razlika u trendu zaraženosti/imunizacije u svijetskim regijama.
Prije nego što to pokažemo, na redu je provjeravanje pretpostavki ANOVE tj. normalna razdioba i homoskedastičnost (iste varijance među klasama).

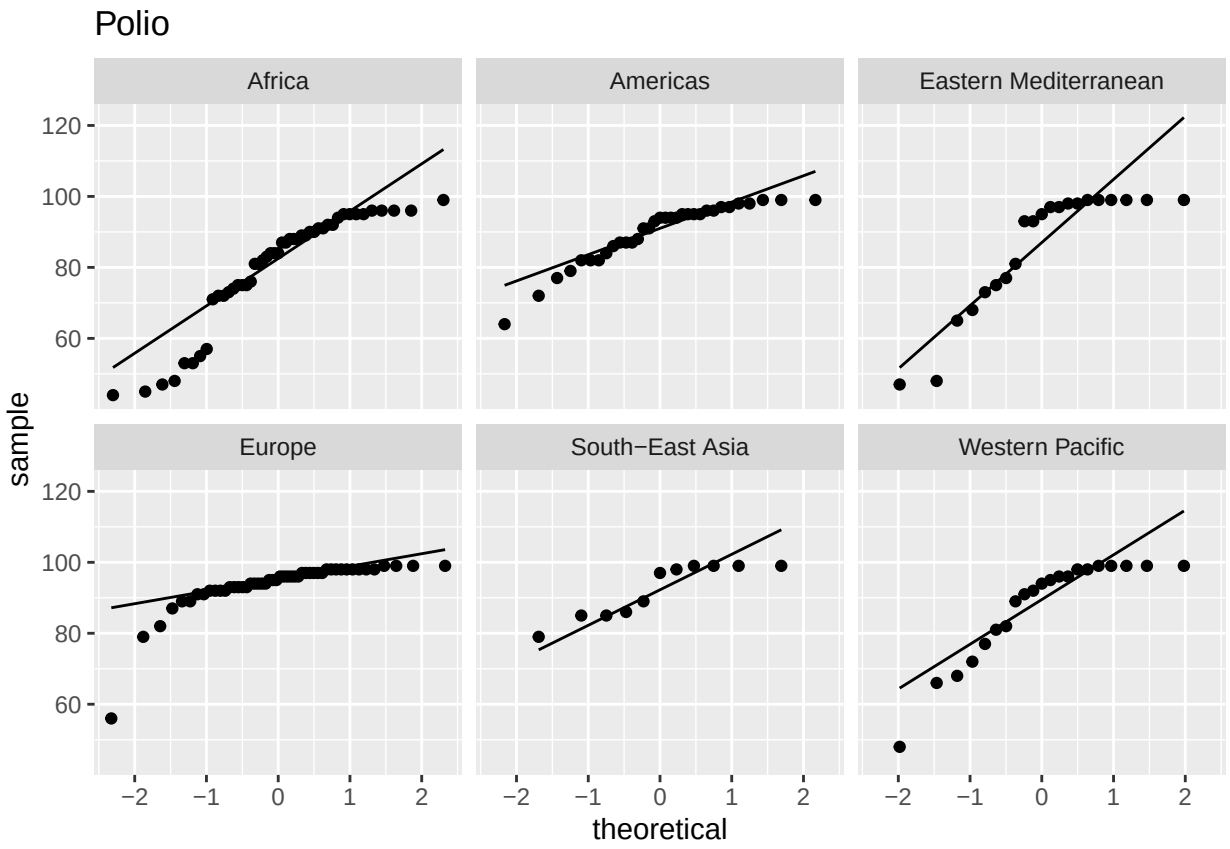
```
#Provjeravamo normalnost podataka qqplotom jer bi bilo puno previse ispisa kada bi smo radili
#Lillieforsov test na svakoj od klasa
ggplot(medDataFiltered, aes(sample = une_hiv, na.rm = T)) +
  stat_qq_line() + stat_qq()+ facet_wrap(~region) + labs(title = "HIV")
```

```
## Warning: Removed 43 rows containing non-finite values (‘stat_qq_line()‘).
```

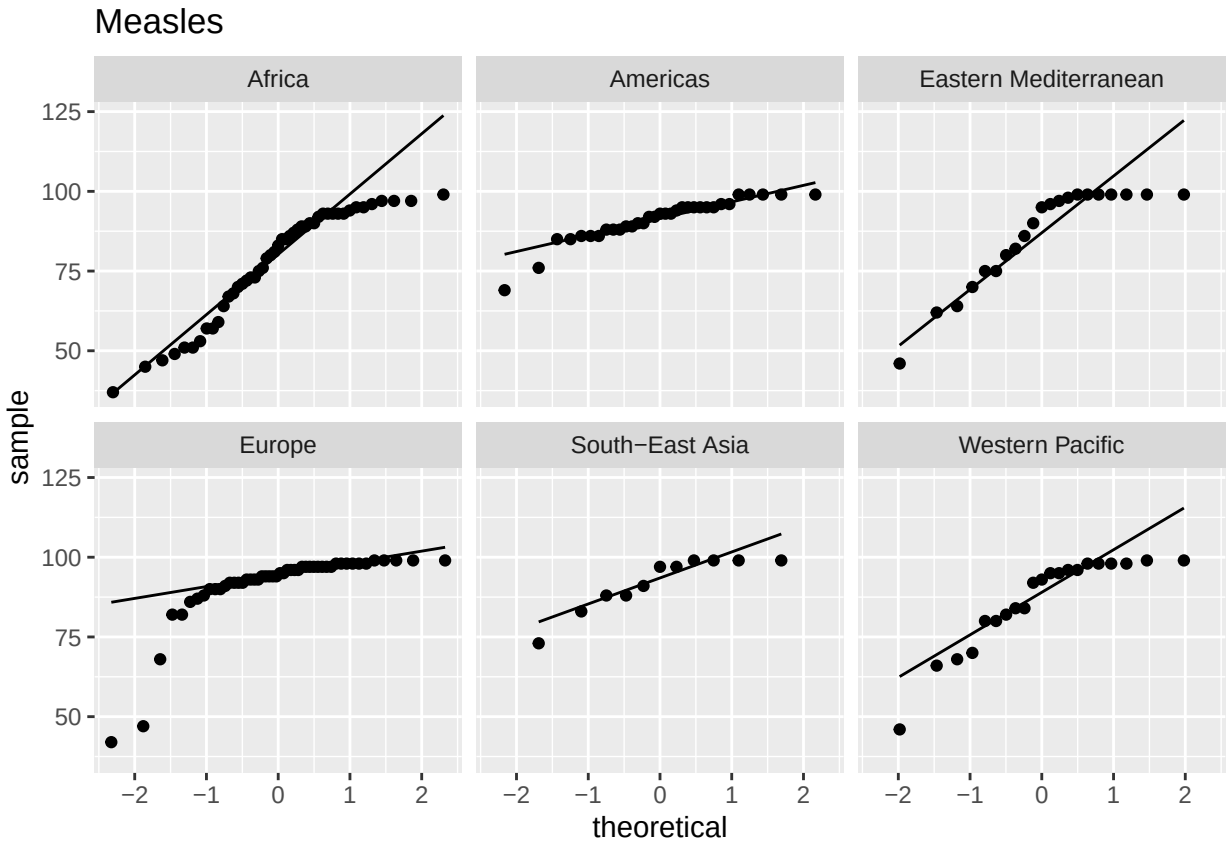
```
## Warning: Removed 43 rows containing non-finite values (‘stat_qq()‘).
```



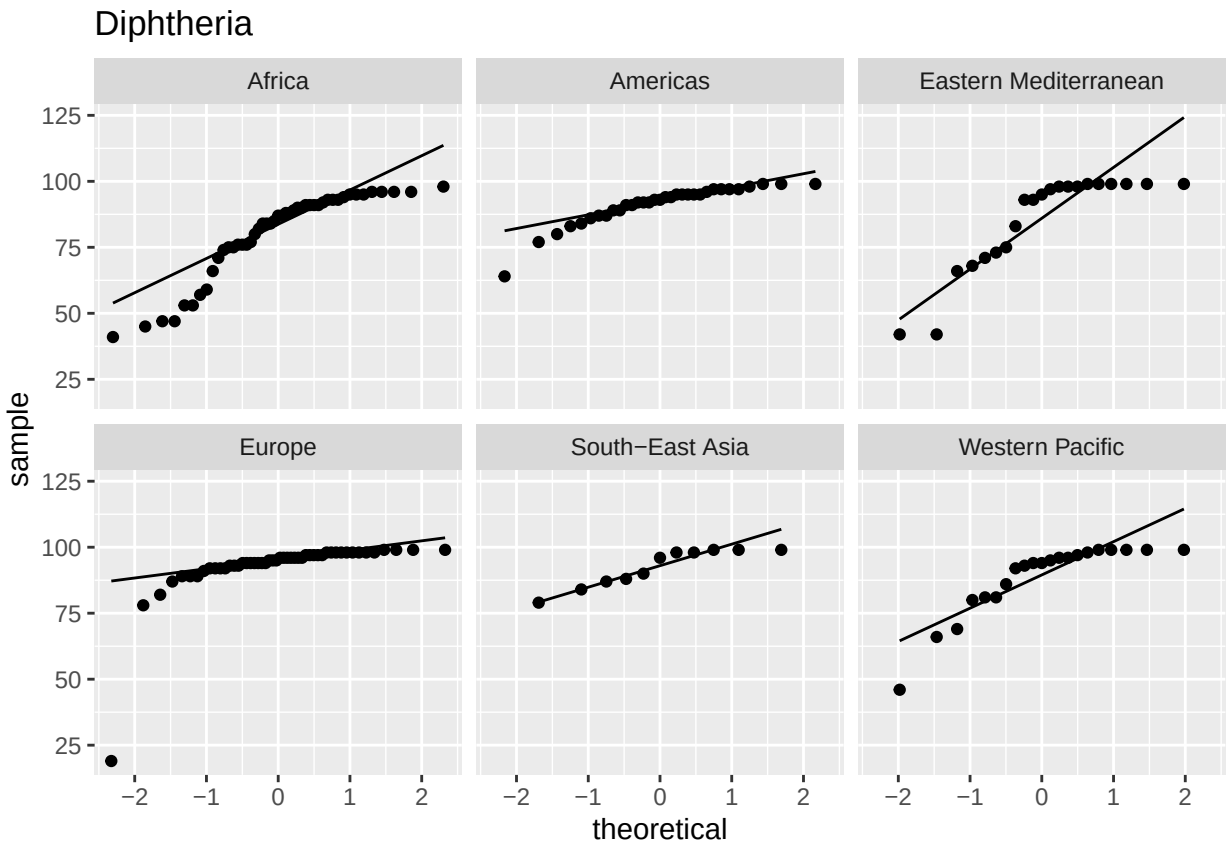

```
ggplot(medDataFiltered, aes(sample = polio, na.rm = T)) +
  stat_qq_line() + stat_qq()+ facet_wrap(~region) + labs(title = "Polio")
```




```
ggplot(medDataFiltered, aes(sample = measles, na.rm = T)) +
  stat_qq_line() + stat_qq()+ facet_wrap(~region) + labs(title = "Measles")
```




```
ggplot(medDataFiltered, aes(sample = diphtheria, na.rm = T)) +
  stat_qq_line() + stat_qq()+ facet_wrap(~region) + labs(title = "Diphtheria")
```




```
#Provest ćemo Lillieforsov test samo na podacima za HIV kao primjer,
#koristimo ga umjesto KS testa jer ne znamo srednju vrijednost i stdev
#populacije
for(name in unique(data$region)){
  print(name)
  print(medDataFiltered %>% filter(region == name) %>%
          select(une_hiv) %>% unlist %>% lillie.test())
}
```

```
## [1] "Africa"
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  .
## D = 0.27876, p-value = 1.68e-09
##
## [1] "Americas"
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  .
## D = 0.28537, p-value = 8.167e-06
##
## [1] "Eastern Mediterranean"
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
```

```
##
## data:  .
## D = 0.41924, p-value = 2.568e-08
##
## [1] "Europe"
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  .
## D = 0.26486, p-value = 1.219e-06
##
## [1] "South-East Asia"
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  .
## D = 0.25574, p-value = 0.1798
##
## [1] "Western Pacific"
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  .
## D = 0.24693, p-value = 0.05952
```

```
#Provodimo Bartlettov test na svakom od uzoraka kako bi smo provjerili
#homoskedastičnost
bartlett.test(medDataFiltered$hepatitis~medDataFiltered$region)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  medDataFiltered$hepatitis by medDataFiltered$region
## Bartlett's K-squared = 24.593, df = 5, p-value = 0.0001669
```

```
bartlett.test(medDataFiltered$une_hiv~medDataFiltered$region)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  medDataFiltered$une_hiv by medDataFiltered$region
## Bartlett's K-squared = 398.81, df = 5, p-value < 2.2e-16
```

```
bartlett.test(medDataFiltered$measles~medDataFiltered$region)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  medDataFiltered$measles by medDataFiltered$region
## Bartlett's K-squared = 34.104, df = 5, p-value = 2.27e-06
```

```
bartlett.test(medDataFiltered$diphtheria~medDataFiltered$region)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  medDataFiltered$diphtheria by medDataFiltered$region
## Bartlett's K-squared = 31.287, df = 5, p-value = 8.221e-06
```

Podaci nisu pretjerano normalni, ipak to možda i ne bi bio pretjerani problem jer je ANOVA relativno robusna na kršenje uvjeta normalnosti. No naši problemi ne staju tu, jer naši podaci ne samo da nisu homoskedastični, već je p-vrijednosti provedenih bartlettovih testa za jednakost varijance sežu između redova 10^-3 i 10^-16. To nam, blago rečeno, ne ide u prilog. Jednoparametarska ANOVA čak i je relativno robusna metoda glede kršenja homoskedastičnosti (ne onoliko koliko nama treba), ali samo ako su uzorci jednakih veličina, što nisu. Tako da se kao i u prvom zadatku moramo okrenuti neparametarskoj ANOVI, Kruskal-Wallis testu kako bi provjerili sljedeću hipotezu koja se nameće iz postavljenog pitanja:

$$H_0 : \mu_0 = \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$$
$$H_1 : barem jedan \mu_i nije jednak$$
$$\alpha = 0.05$$

```
kruskal.test(une_hiv~region, medDataFiltered)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  une_hiv by region
## Kruskal-Wallis chi-squared = 80.493, df = 5, p-value = 6.617e-16
```

```
kruskal.test(hepatitis~region, medDataFiltered)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  hepatitis by region
## Kruskal-Wallis chi-squared = 24.587, df = 5, p-value = 0.0001674
```

```
kruskal.test(measles~region, medDataFiltered)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  measles by region
## Kruskal-Wallis chi-squared = 28.35, df = 5, p-value = 3.109e-05
```

```
kruskal.test(diphtheria~region, medDataFiltered)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  diphtheria by region
## Kruskal-Wallis chi-squared = 30.995, df = 5, p-value = 9.387e-06
```

Vidimo da p-vrijednosti donose dosta čvrst zaključak u svakom od 4 provedena testa za hiv, ospice, polio i difteriju, a taj zaključak jest odbacivanje nul hipoteze da su sve stope zaraženosti/imunizacije jednake u korist alternativne da barem jedna nije. Ovo naravno, nije veliki šok jer smo iz prvog pogleda na podatke uočili veliko odstupanje u stopama zaraženosti/imunizacije u Africi, koju je svijet iskorištavao stoljećima i onda ostavio da se sama nosi s posljedicama.

5. Postoji li korelacija između konzumacije alkohola i prosječne količine obrazovanja

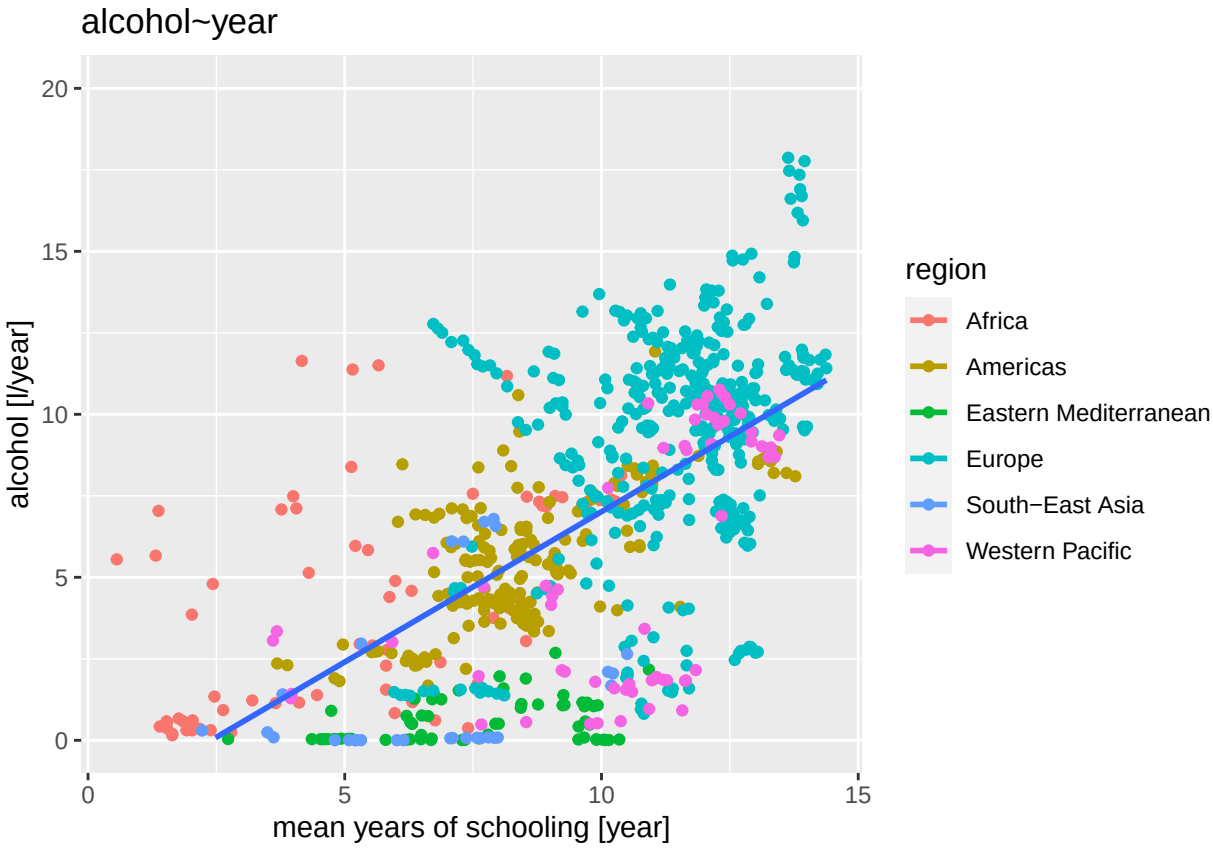
```
ggplot(data, aes(x = une_school, y = alcohol, color = region)) + geom_point() +
  stat_smooth(formula = y~x, method = "lm", aes(group = 1), se = F) +
  scale_y_continuous(limits = c(0, 20))+
  labs(x = "mean years of schooling [year]", y = "alcohol [l/year]",
       title = "alcohol~year")
```

```
## Warning: Removed 2310 rows containing non-finite values (‘stat_smooth()‘).
```

```
## Warning: The following aesthetics were dropped during statistical transformation: colour
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a ‘group‘ aesthetic or to convert a numerical
##   variable into a factor?
```

```
## Warning: Removed 2310 rows containing missing values (‘geom_point()‘).
```

```
## Warning: Removed 11 rows containing missing values (‘geom_smooth()‘).
```




```
linModAl <- lm(alcohol~une_school, data)
summary(linModAl)
```

```
##
## Call:
## lm(formula = alcohol ~ une_school, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.5460 -1.8221  0.1801  1.7792 10.0085
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.21047    0.40179  -5.502 5.07e-08 ***
## une_school   0.92229    0.03968  23.243  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.204 on 799 degrees of freedom
##   (2310 observations deleted due to missingness)
## Multiple R-squared:  0.4034, Adjusted R-squared:  0.4026
## F-statistic: 540.2 on 1 and 799 DF,  p-value: < 2.2e-16
```

Model nije pretjerano uvjerljiv, ali ne možemo osporiti postojanje nekakve linearne veze s faktorom $R^2 = 0.4$.
Interesantno je uočiti da države sa većom stopom obrazovanja konzumiraju više alkohola.
Razlozi k tome su sigurno razni i vjerojatno sežu od veće kupovne moći sve do nezdravog nošenja sa stresom.