

Big Five – velikih pet dimenzija ličnosti

Projekt iz predmeta Statistička analiza podataka, Fakultet elektrotehnike i računarstva

Filip Grebenar, David Konjevod, Lovre Mitrović, Mihaela Bakšić

14.1.2022.

Uvod

Ljudske osobine i njihova manifestacija često su istraživan fenomen. Trenutno dominantni model za klasifikaciju osobina ličnosti je Big Five model. Inicialno je model bio razvijen s ciljem povezivanja osobina ličnosti i akademskih uspjeha i ponašanja. Barem četiri nezavisna skupa istraživača bavili su se problemom klasifikacija osobina ličnosti, te su svi diferencirali pet glavnih osobina ličnosti: ekstraverzija, ugodnost, savjesnost, neuroticizam i otvorenost. Bitno je naglasiti kako se ličnosti ne nalaze u ekstremima već na kontinuumu između ovih pet faktora.

Ekstraverzija se odnosi na interakciju s društvenom okolinom, podrazumjevajući snalaženje i volju osobe za druženjem i poduzetnost. Ugodnost predstavlja empatičnost i kooperativnost s drugima. Savjesnost vrednuje odnos prema obavezama i poslu te spontanost. Neuroticizam opisuje emocionalnu stabilnost osobe. Otvorenost razlikuje osobe prema otvorenosti novim konceptima i idejama.

Ovaj projekt predstavit će generalni pregled podataka, pokušati pronaći pravilnosti u podacima i modelirati ih. Također, pozabavit će se ispitivanjem nekih uvriježenih ideja i mišljenja o osobinama ličnosti.

Skup podataka

Za svakog ispitanika u skupu podatka prisutne su vrijednosti za: državu, godine, spol, faktor ugodnosti, faktor ekstraverzije, faktor otvorenosti, faktor savjesnosti i faktor neuroticizma. Skup podataka ukupno sadrži 307313 zapisa. Prikazani su osnovni podaci deskriptivne statistike za cijeli skup podataka. U nastavku će se dodatno prikazati i analizirati deskriptivna statistika za varijable i podskupove podataka ovisno o potrebi problema koji se obrađuje. Također, vidimo da su podaci potpuni, tj. nema vrijednosti koje nedostaju, što je odlika dobrog skupa podataka.

```
bigfive = read.csv("./big_five_scores.csv")
head(bigfive)
```

```
##   case_id    country age sex agreeable_score extraversion_score openness_score
## 1      1 South Afri  24   1      0.7533333     0.4966667    0.8033333
## 2      3        UK  24   2      0.7333333     0.6800000    0.7866667
## 3      4       USA  36   2      0.8800000     0.7700000    0.8600000
## 4      5        UK  19   1      0.6900000     0.6166667    0.7166667
## 5      6        UK  17   1      0.6000000     0.7133333    0.6466667
## 6      7       USA  17   1      0.6033333     0.5866667    0.6533333
##   conscientiousness_score neuroticism_score
## 1            0.8866667      0.4266667
```

```

## 2           0.7466667    0.5900000
## 3           0.8966667    0.2966667
## 4           0.6366667    0.5633333
## 5           0.6333333    0.5133333
## 6           0.5966667    0.6233333

nrow(bigfive)

## [1] 307313

summary(bigfive)

##      case_id          country        age       sex
##  Min.   :     1   USA      :212625   Min.   :10.00  Min.   :1.000
##  1st Qu.: 83653  Canada   : 21798   1st Qu.:18.00  1st Qu.:1.000
##  Median :166286  UK       : 16489   Median :22.00  Median :2.000
##  Mean   :166682  Australia: 10400   Mean   :25.19  Mean   :1.602
##  3rd Qu.:249627  Netherland: 3469   3rd Qu.:29.00  3rd Qu.:2.000
##  Max.   :334161  India    : 2841   Max.   :99.00  Max.   :2.000
##                  (Other)  : 39691
##      agreeable_score  extraversion_score  openness_score conscientiousness_score
##  Min.   :0.2000      Min.   :0.2000      Min.   :0.2533  Min.   :0.2067
##  1st Qu.:0.6400      1st Qu.:0.6000      1st Qu.:0.6733  1st Qu.:0.6300
##  Median :0.7033      Median :0.6800      Median :0.7367  Median :0.7067
##  Mean   :0.6968      Mean   :0.6723      Mean   :0.7339  Mean   :0.7020
##  3rd Qu.:0.7633      3rd Qu.:0.7500      3rd Qu.:0.7967  3rd Qu.:0.7767
##  Max.   :1.0000      Max.   :0.9933      Max.   :0.9967  Max.   :1.0000
##
##      neuroticism_score
##  Min.   :0.1967
##  1st Qu.:0.4867
##  Median :0.5700
##  Mean   :0.5744
##  3rd Qu.:0.6600
##  Max.   :0.9967
## 

sum(is.na(bigfive$country))

## [1] 0

sum(is.na(bigfive$age))

## [1] 0

sum(is.na(bigfive$sex))

## [1] 0

```

```

sum(is.na(bigfive$agreeable_score))

## [1] 0

sum(is.na(bigfive$extraversion_score))

## [1] 0

sum(is.na(bigfive$openness_score))

## [1] 0

sum(is.na(bigfive$conscientiousness_score))

## [1] 0

sum(is.na(bigfive$neuroticism_score))

## [1] 0

```

Testiranje razlika u otvorenosti kod mladih i starih ispitanika

Često je prisutna pretpostavka kako su mladi ljudi otvoreniji novim iskustvima od starih. Stoga, provest će se testiranje te pretpostavke. Mladim ljudima smatrati će se svi stari 30 ili manje godina, dok će se starim ljudima smatrati oni stari 60 ili više godina. Idemo li korak dalje, može li se uspostaviti jasna veza između dobi ispitanika i njihove otvorenosti? Može li se ta veza modelirati linearnom regresijom?

```

cat('Medijan godina svih ispitanika je ', median(bigfive$age), '\n')

## Medijan godina svih ispitanika je 22

old_subjects = bigfive[bigfive$age >= 60 ,]
young_subjects = bigfive[bigfive$age <= 30 ,]
cat('Srednja vrijednost otvorenosti za mlade ( <= 30 godina ) ispitanike je ', mean(young_subjects$openne

## Srednja vrijednost otvorenosti za mlade ( <= 30 godina ) ispitanike je 0.7338046

cat('Srednja vrijednost otvorenosti za stare ( >= 60 godina ) ispitanike je ', mean(old_subjects$openne

## Srednja vrijednost otvorenosti za stare ( >= 60 godina ) ispitanike je 0.7265828

cat('Varijanca otvorenosti za mlade ( <= 30 godina ) ispitanike je ', var(young_subjects$openness_score)

## Varijanca otvorenosti za mlade ( <= 30 godina ) ispitanike je 0.007601208

```

```

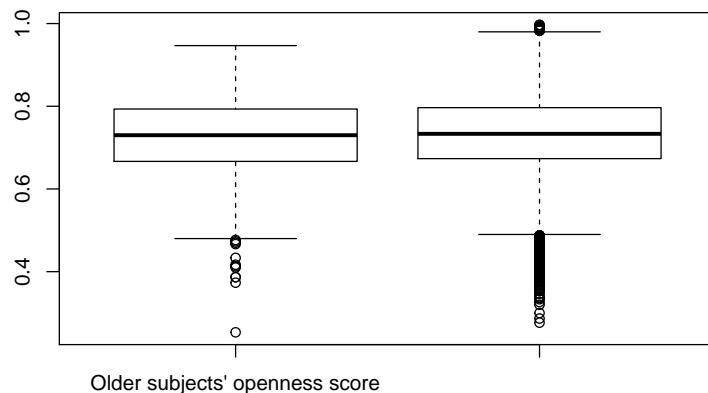
cat('Varijanca otvorenosti za stare ( >= 60 godina ) ispitanike je ', var(old_subjects$openness_score),

## Varijanca otvorenosti za stare ( >= 60 godina ) ispitanike je 0.008904973

boxplot(old_subjects$openness_score, young_subjects$openness_score,
        names = c('Older subjects\' openness score','Younger subjects\' openness score'),
        main='Boxplot of younger and older subjects\' openness score')

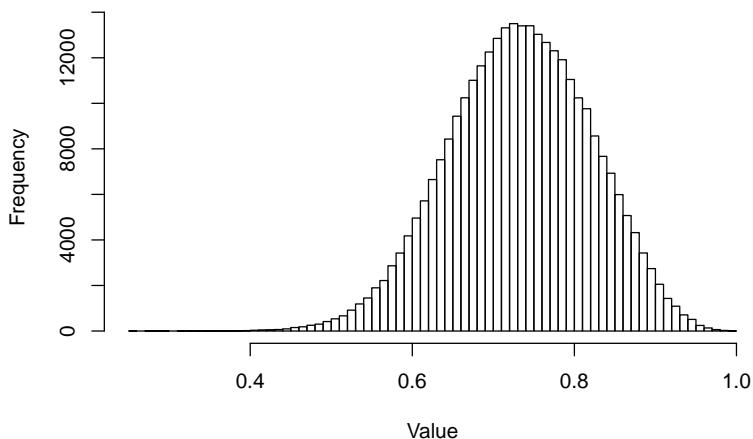
```

Boxplot of younger and older subjects' openness score



```
hist(bigfive$openness_score,main='Openness score histogram',xlab='Value',ylab='Frequency', breaks=100)
```

Openness score histogram

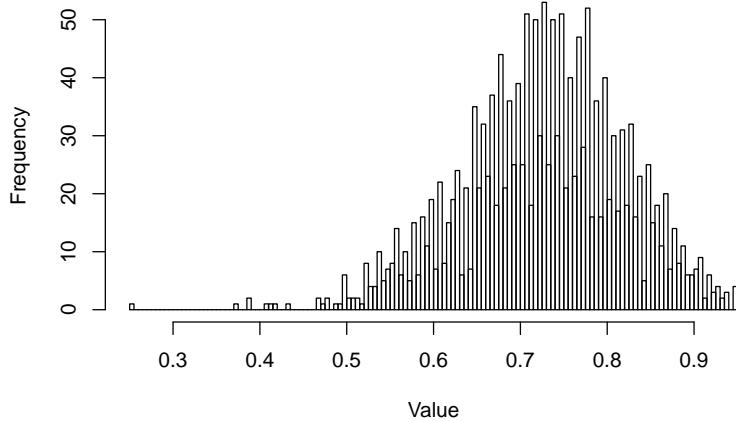


Deskriptivnom analizom podatka dolazimo do saznanja o razlici u srednjim vrijednostima faktora otvorenosti za mlade i stare sudionike.

Normalnost podataka

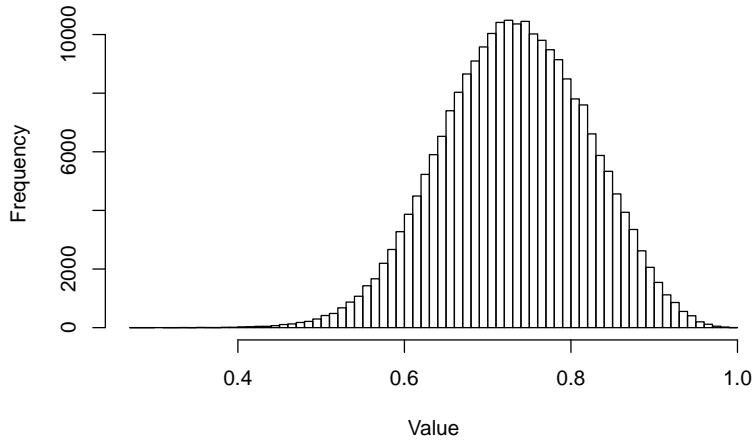
```
hist(old_subjects$openness_score,main='Openness score old subjects histogram',xlab='Value',ylab='Frequency')
```

Openness score old subjects histogram

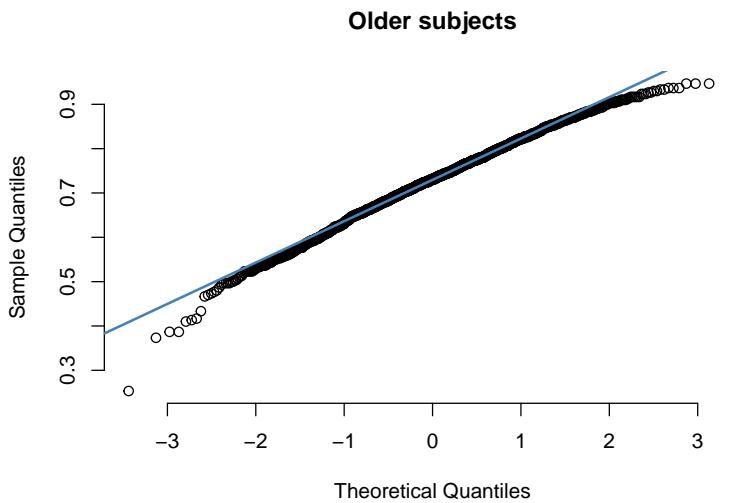


```
hist(young_subjects$openness_score,main='Openness score young subjects histogram',xlab='Value',ylab='Frequency')
```

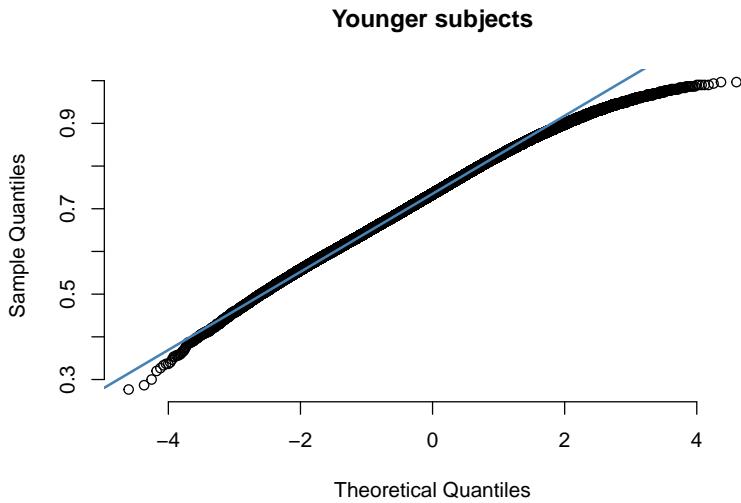
Openness score young subjects histogram



```
qqnorm(old_subjects$openness_score, pch = 1, frame = FALSE,main='Older subjects')
qqline(old_subjects$openness_score, col = "steelblue", lwd = 2)
```



```
qqnorm(young_subjects$openness_score, pch = 1, frame = FALSE, main='Younger subjects')
qqline(young_subjects$openness_score, col = "steelblue", lwd = 2)
```



Iz histograma i qq-plotova za openness_score mladih i starih ispitanika može se naslutiti kako podaci pripadaju normalnoj razdiobi. Ta pretpostavka pokušat će se dodatno potvrditi provođenjem Lillieforsove inačice Kolmogorov-Smirnovljevog testa. Za sve testove unutar ovog podzadatka koristi se razina značajnosti $\alpha = 0.05$.

Testiranje normalnosti - Lillieforsova inačica Kolmogorov-Smirnovljevog testa

Testiranje normalnosti otvorenosti za starije ispitanike

```
lillie.test(old_subjects$openness_score)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
```

```
## data: old_subjects$openness_score
## D = 0.031255, p-value = 0.000517
```

S obzirom da je p-vrijednost $= 0.000517 < \alpha$, nulta hipoteza o pripadnosti podataka o otvorenosti starih ispitanika normalnoj raspodijeli se odbacije u korist alternativne hipoteze.

Testiranje normalnosti faktora otvorenosti za mlađe ispitanike

```
lillie.test(young_subjects$openness_score)
```

```
##
## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: young_subjects$openness_score
## D = 0.018819, p-value < 2.2e-16
```

S obzirom da je p-vrijednost $< 2.2e-16 < \alpha$, nulta hipoteza o pripadnosti podataka o otvorenosti mlađih ispitanika normalnoj raspodjeli se odbacije u korist alternativne hipoteze.

Testiranje jednakosti varijanci otvorenosti za stare i mlade ispitanike

Testiranje jednakosti varijanci provodi se F testom. Pretpostavlja se nezavisnost uzoraka.

Hipoteze:

$$H_0 : \sigma_{old}^2 = \sigma_{young}^2$$

$$H_1 : \sigma_{old}^2 \neq \sigma_{young}^2$$

```
var.test(old_subjects$openness_score, young_subjects$openness_score)
```

```
##
## F test to compare two variances
##
## data: old_subjects$openness_score and young_subjects$openness_score
## F = 1.1715, num df = 1709, denom df = 238014, p-value = 2.084e-06
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 1.096503 1.254507
## sample estimates:
## ratio of variances
## 1.171521
```

S obzirom da je p-vrijednost $= 2.084e-06 < \alpha$ odbacujemo hipotezu $H_0 : \sigma_{old}^2 = \sigma_{young}^2$ u korist alternativne hipoteze $H_1 : \sigma_{old}^2 \neq \sigma_{young}^2$. Zaključujemo da podatci govore u prilog alternativnoj hipotezi i u testovima koji slijede koristit će se informacija o nejednakosti varijanci dobivena ovim testom.

Testiranje jednakosti srednjih vrijednosti otvorenosti za stare i mlade ispitanike

Testiranje jednakosti srednjih vrijednosti T testom zahtjeva da podaci dolaze iz normalne razdiobe. Prethodno provedenim Lillieforsovim testom normalnosti pokazalo se da to ne vrijedi niti za podatke o otvorenosti mladih niti starih ispitanika. Ipak, histogrami i qq-plot grafovi prikazuju razdiobu koja nije daleko od normalne. Uvezši u obzir oblik grafova i veliku osjetljivost Lillieforsovog testa normalnosti, možemo nastaviti testiranje T testom uz pretpostavku normalnosti podataka.

Testiranje jednakosti srednjih vrijednosti otvorenosti za stare i mlade ispitanike provodi se kao T test za podatke s različitim varijancama, kako je pokazano prethodnim testom. Prepostavlja se nezavisnost uzoraka.

Hipoteze:

$$H_0 : \mu_{old} = \mu_{young}$$

$$H_1 : \mu_{old} \neq \mu_{young}$$

```
t.test(old_subjects$openness_score, young_subjects$openness_score, alt = "two.sided", var.equal = FALSE)
```

```
##  
## Welch Two Sample t-test  
##  
## data: old_subjects$openness_score and young_subjects$openness_score  
## t = -3.155, df = 1730, p-value = 0.001633  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.011711273 -0.002732275  
## sample estimates:  
## mean of x mean of y  
## 0.7265828 0.7338046
```

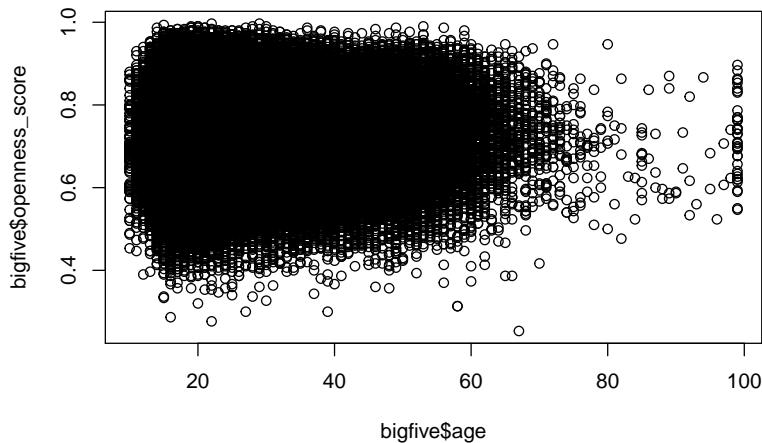
p-vrijednost = 0.00163 < α , stoga se hipoteza $H_0 : \mu_{old} = \mu_{young}$ odbacuje u korist alternativne hipoteze $H_1 : \mu_{old} \neq \mu_{young}$.

Statističkim testiranjem pokazano je kako postoji statistički značajna razlika između srednjih vrijednosti otvorenosti mladih i starih ispitanika, sukladno često prisutnom mišljenju. Ova informacija ponukala nas je da pokušamo jasno modelirati vezu između tih dviju varijabli.

Model linearne regresije za modeliranja odnosa godina i otvorenosti kod ispitanika

S obzirom na činjenicu da smo prethodnim testovima pokazali da je razlika između srednjih vrijednosti otvorenosti za mlade i stare ispitanike statistički značajna, vrijedilo bi se pozabaviti pitanjem odnosa između varijabli godina (age) i otvorenosti (openness_score). Za modeliranje ovisnosti koristi se model linearne regresije. Kako bi se dobio dojam o mogućoj zavisnosti varijable koja ocjenjuje otvorenost ispitanika i njihovih godina, prikazan je grafički prikaz u obliku scatter-plota.

```
plot(bigfive$age,bigfive$openness_score)
```



Ipak, na temelju izgleda grafa moglo bi se zaključiti da veza između varijabli ne postoji ili nije statistički značajna. Kako bi se moglo preciznije diskutirati o tom pitanju, nastaviti će se s modeliranjem njihvog odnosa modelom linearne regresije i provođenjem testova nad modelom, unatoč sugestiji grafičkog prikaza da je korelacija između varijabli vrlo slaba. Na temelju ocjene sposobnosti modela linearne regresije da modelira prikupljene podatke, ocjeniti će se i njihova linearna zavisnost. Izlazna varijabla u modelu je openness_score, dok je varijabla age regresor. S obzirom da je priutan samo jedan regresor, koristi se jednostavna linearna regresija.

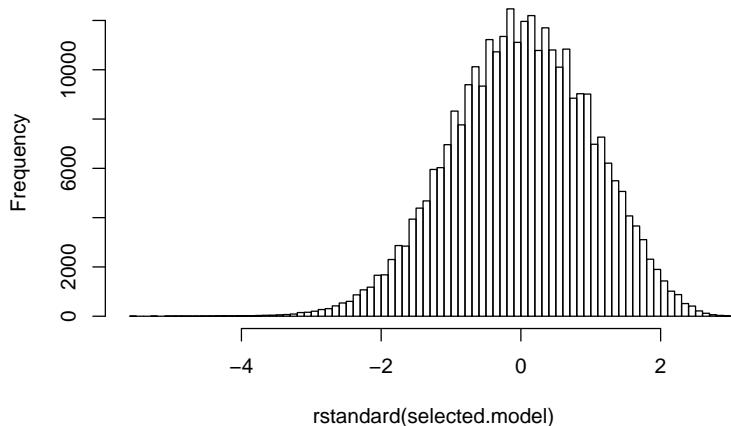
```
selected.model = lm(openness_score~age, data=bigfive)
```

Provjera prepostavki modela

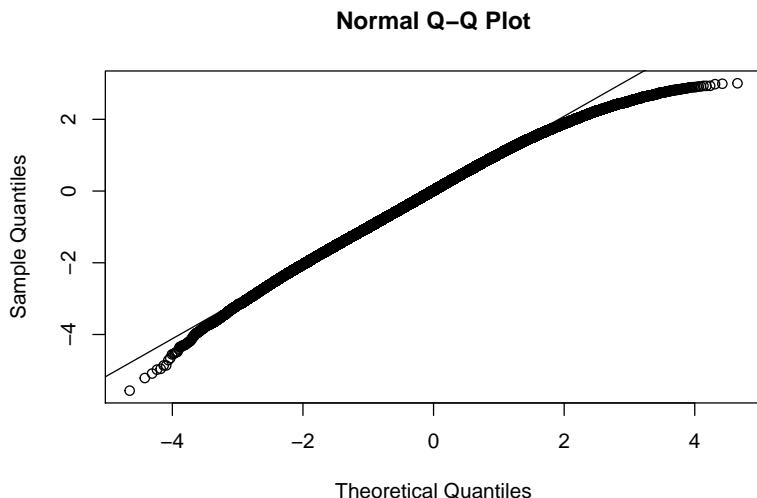
Prije svega treba provjeriti prepostavke modela. Prva prepostavka modela linearne regresije je da su reziduali normalno distribuirani i imaju homogenu varijancu. S obzirom da je korišten model jednostavne regresije, nije potrebno baviti se prepostavkom o slaboj međusobnoj koreliranosti regresora.

```
hist(rstandard(selected.model), breaks=100)
```

Histogram of rstandard(selected.model)



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



Histogram reziduala je vrlo obećava-juć te snažno podupire pretpostavku o normalnosti reziduala. Ipak, potrebno je provesti i Koglomorov-Smirnovljev test kako bi se normalnost insinuirana histogramom potvrdila. Za ovu namjenu, umjesto Lilleforsovog testa koristi se KS test, jer baratamo sa standardiziranim rezidualima, za koje očekujemo da će KS test dati bolje rezultate.

```
ks.test(rstandard(selected.model), 'pnorm')
```

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

##
## One-sample Kolmogorov-Smirnov test
##
## data: rstandard(selected.model)
## D = 0.01445, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

S obzirom da je p-vrijednost $< \alpha$, nulta hipoteza o normalnosti standardiziranih reziduala se odbija u korist alternativnoj hipotezi. S obzirom na izgled histograma, izgled qq-plota i robusnost T-testa, ipak se može govoriti o dovoljnoj normalnosti reziduala za daljnja testiranja i evaluaciju modela.

Za ocjenu kvalitete modela koristi se koeficijent determinacije R^2 , koji odgovara udjelu varijance zavisne varijable koju objašnjava dani model.

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

```
##
## Call:
## lm(formula = openness_score ~ age, data = bigfive)
##
## Residuals:
```

```

##      Min     1Q   Median     3Q    Max
## -0.48752 -0.05958  0.00110  0.06259  0.26325
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.7297757  0.0004282 1704.09 <2e-16 ***
## age         0.0001654  0.0000158   10.46 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08759 on 307311 degrees of freedom
## Multiple R-squared:  0.0003563, Adjusted R-squared:  0.000353
## F-statistic: 109.5 on 1 and 307311 DF, p-value: < 2.2e-16

```

R^2 za procjenjeni model iznosi 0.0003563, što je izrazito niska vrijednost. Oba parametra modela također imaju izuzetno malu p-vrijednost, čime ih se ne može smatrati statistički značajnim. Također, p-vrijednost za F-test signifikantnosti modela je daleko ispod α . Uzimajući ove ocjene modela u obzir, možemo sa velikom sigurnošću tvrditi kako je model linearne regresije neadekvatan odabir za modeliranje odnosa između godina i otvorenosti ispitanika. S obzirom na izgled scatter plota odnosa te dvije varijable, može se reći da je rezultat i očekivan.

Savjesnost između regija

Uvod

U ovom dijelu bavimo se pitanjem imaju li neke regije značajno različite rezultate u određenom faktoru. Primjerice je li opravdan mit o visokoj savjesnosti populacije istočne Azije naspram populacije drugih kontinenata.

Regije sam podijelio na osnovu UN-ove podjele (izvor: "<https://unstats.un.org/sdgs/indicators/regional-groups/>").

```

`%!in%` <- Negate(`%in%`)
region_east_asian =c('Brunei', 'Cambodia', 'China', 'Hong Kong', 'Indonesia', 'Japan', 'Macau', 'Malaysia', 'Philippines', 'Singapore', 'Thailand')
east_asia = bigfive[bigfive$country %in% region_east_asian,]
rest_regions = bigfive[bigfive$country %!in% region_east_asian,]
summary(bigfive$conscientiousness_score)

##      Min. 1st Qu. Median   Mean 3rd Qu.   Max.
##  0.2067  0.6300  0.7067  0.7020  0.7767  1.0000

summary(east_asia$conscientiousness_score)

##      Min. 1st Qu. Median   Mean 3rd Qu.   Max.
##  0.3167  0.6333  0.6967  0.6946  0.7567  1.0000

summary(rest_regions$conscientiousness_score)

##      Min. 1st Qu. Median   Mean 3rd Qu.   Max.
##  0.2067  0.6300  0.7067  0.7023  0.7800  1.0000

```

```

cat('Varijanca savjesnosti za istočnu aziju ', var(east_asia$conscientiousness_score), '\n')

## Varijanca savjesnosti za istočnu aziju  0.009029865

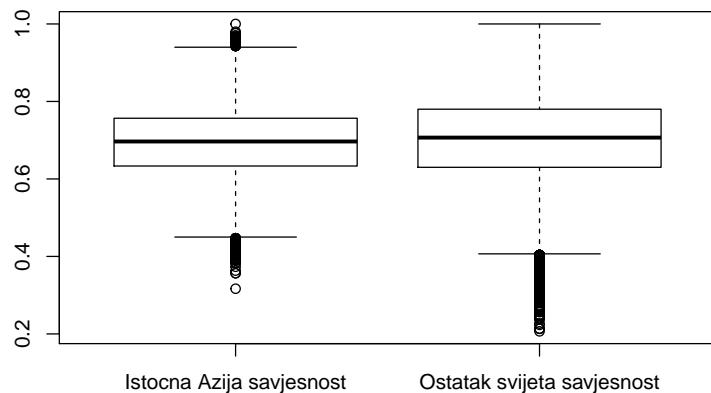
cat('Varijanca savjesnosti za ostatak svijeta ', var(rest_regions$conscientiousness_score), '\n')

## Varijanca savjesnosti za ostatak svijeta  0.01162203

boxplot(east_asia$conscientiousness_score, rest_regions$conscientiousness_score,
        names = c('Istocna Azija\savjesnost','Ostatak svijeta savjesnost'),
        main='Boxplot za istocnu Aziju i ostatak svijeta savjesnost')

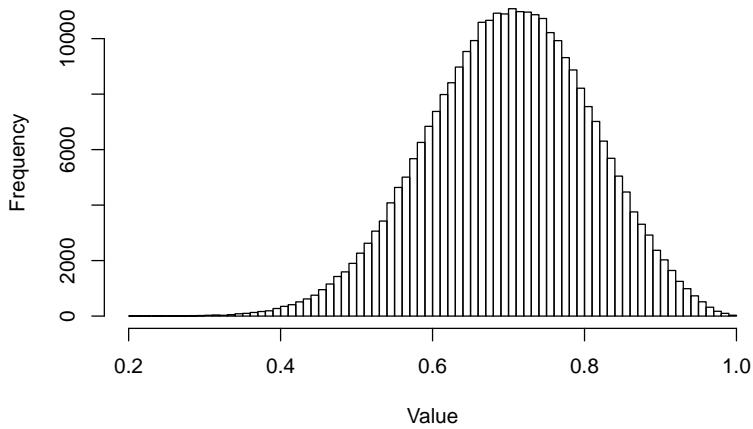
```

Boxplot za istocnu Aziju i ostatak svijeta savjesnost

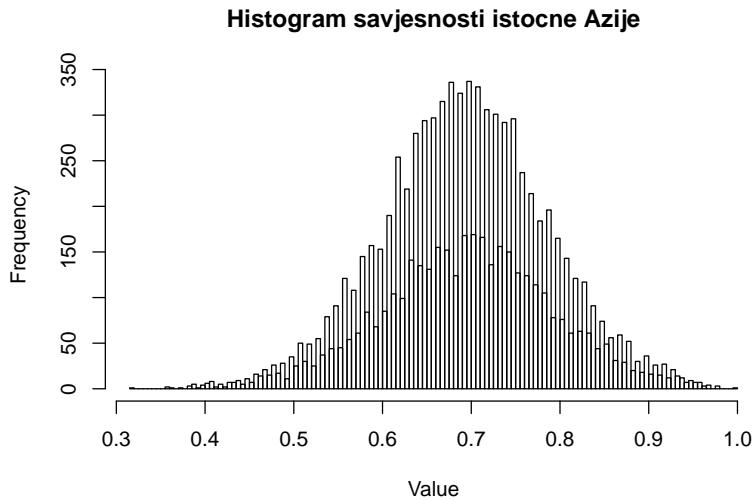


```
hist(bigfive$conscientiousness_score,main='Savjesnost histogram',xlab='Value',ylab='Frequency', breaks=
```

Savjesnost histogram



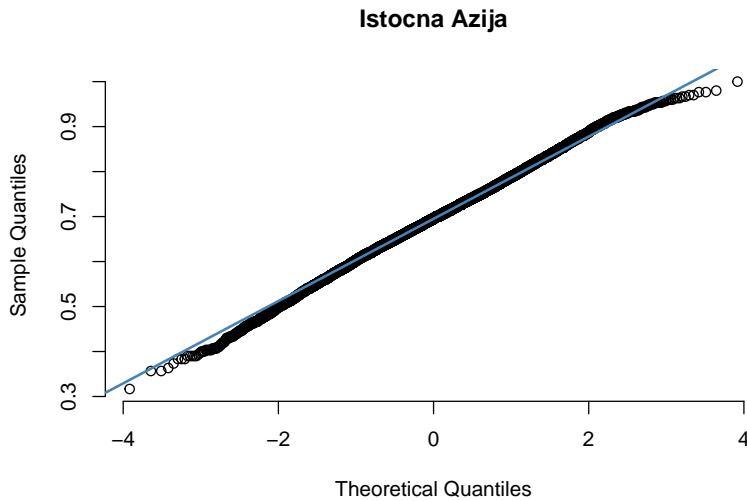
```
hist(east_asia$conscientiousness_score,main='Histogram savjesnosti istocne Azije',xlab='Value',ylab='Fr
```



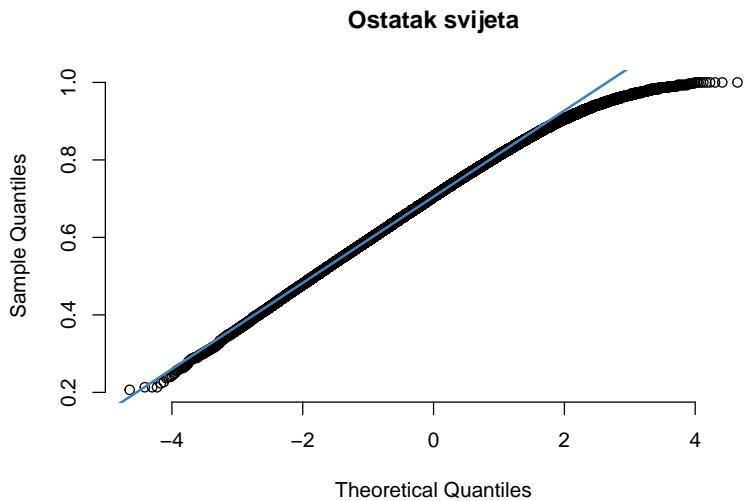
```
hist(rest_regions$conscientiousness_score,main='Histogram savjesnosti ostatka svijeta',xlab='Value',ylab='Fr
```



```
qqnorm(east_asia$conscientiousness_score, pch = 1, frame = FALSE,main='Istocna Azija')
qqline(east_asia$conscientiousness_score, col = "steelblue", lwd = 2)
```



```
qqnorm(rest_regions$conscientiousness_score, pch = 1, frame = FALSE, main='Ostatak svijeta')
qqline(rest_regions$conscientiousness_score, col = "steelblue", lwd = 2)
```



Već iz priloženih grafova može se naslutiti da su obje distribucije normalne te suprotno početnom pitanju izgleda da je savjesnost ispitanika istočne Azije manja nego u ostatku svijeta.

Testiranje normalnosti distribucije

Hipoteze:

H_0 : distribucija `east_asia$conscientiousness_score` pripada normalnoj razdiobi
 H_1 : distribucija `east_asia$conscientiousness_score` NE pripada normalnoj razdiobi

```
lillie.test(east_asia$conscientiousness_score)
```

```
##
```

```

## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: east_asia$conscientiousness_score
## D = 0.017328, p-value = 5.76e-08

```

Na osnovu p vrijednosti od 5.76e-08 odbacujemo nultu hipotezu u korist alternative da se savjesnost ispitanika istočne Azije ne ravna po normalnoj distribuciji.

Hipoteze:

H_0 : distribucija rest_regions\$conscientiousness_score pripada normalnoj razdiobi H_1 : distribucija rest_regions\$conscientiousness_score NE pripada normalnoj razdiobi

```
lillie.test(rest_regions$conscientiousness_score)
```

```

##
## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: rest_regions$conscientiousness_score
## D = 0.020199, p-value < 2.2e-16

```

Na osnovu p vrijednosti od 2.2e-18 odbacujem nultu hipotezu u korist alternative da se savjesnost ispitanika iz ostalih regija ne ravna po normalnoj distibuciji.

Uzevši u obzir preveliku osjetljivost ovog testa i činjenicu da je iz histograma vidljiva normalnost u nastavku prepostavljam normalnost za obje populacije.

Testiranje jednakosti varijance savjesnosti

Prepostavljamo nezavisnost uzorka i izvodim F test.

Hipoteze:

$$H_0 : \sigma_{eastAsia}^2 = \sigma_{rest}^2$$

$$H_1 : \sigma_{eastAsia}^2 \neq \sigma_{rest}^2$$

```
var.test(east_asia$conscientiousness_score, rest_regions$conscientiousness_score)
```

```

##
## F test to compare two variances
##
## data: east_asia$conscientiousness_score and rest_regions$conscientiousness_score
## F = 0.77696, num df = 11152, denom df = 296159, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 0.7565887 0.7981497
## sample estimates:
## ratio of variances
## 0.7769608

```

Na osnovu male p vrijednosti odbacujemo nultu hipotezu u korist alternative da su varijance distribucije različite

Testiranje jednakosti razine savjesnosti

Koristimo T test i pretpostavljam nezavisnost varijabli i različitost varijanci te normalnu razdiobu.

Hipoteze:

$$H_0 : \mu_{eastAsia} = \mu_{rest}$$

$$H_1 : \mu_{eastAsia} > \mu_{rest}$$

```
t.test(east_asia$conscientiousness_score, rest_regions$conscientiousness_score, alt = "greater", var.equal = TRUE)

##
##  Welch Two Sample t-test
##
## data: east_asia$conscientiousness_score and rest_regions$conscientiousness_score
## t = -8.3476, df = 12258, p-value = 1
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
## -0.009206648 Inf
## sample estimates:
## mean of x mean of y
## 0.6945865 0.7022776
```

Uz veliku p vrijednost nismo uspjeli odbaciti nullu hipotezu i zaključujem nije opravdan mit o visokoj savjesnosti populacije istočne Azije.

Savjesnost između više regija

Uvod

Pošto se nismo uvjerili u ispravnost mita da su stanovnici istočne Azije savjesniji od ostatka svijeta pitam postoji li uopće razlike u savjesnosti između stavnika različitih regija. Za primjer ćemo uzeti tri regije za koje imamo sličan broj podataka. Naše pitanje glasi: "Postoji li razlika u savjesnosti između latinske Amerike, sjeverne Afrike i zapadne Azije(arapskog svijeta) i centralne Azije"

```
region_latin_america = c('Argentina', 'Anguilla', 'Trinidad a', 'Chile', 'Mexico', 'Brazil', 'Bahamas',
region_north_africa_and_west_asia = c('Syria', 'Egypt', 'Turkey', 'Algeria', 'Kuwait', 'Israel', 'Lebanon',
region_central_asia = c('India', 'Afghanista', 'Bangladesh', 'Pakistan', 'Iran', 'Turkmenist', 'Nepal',

latin_america = bigfive[bigfive$country %in% region_latin_america,]
arabic_world = bigfive[bigfive$country %in% region_north_africa_and_west_asia,]
central_asia = bigfive[bigfive$country %in% region_central_asia,]

summary(latin_america$conscientiousness_score)

##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
##  0.3167  0.6467  0.7267  0.7234  0.8067  0.9900

summary(arabic_world$conscientiousness_score)
```

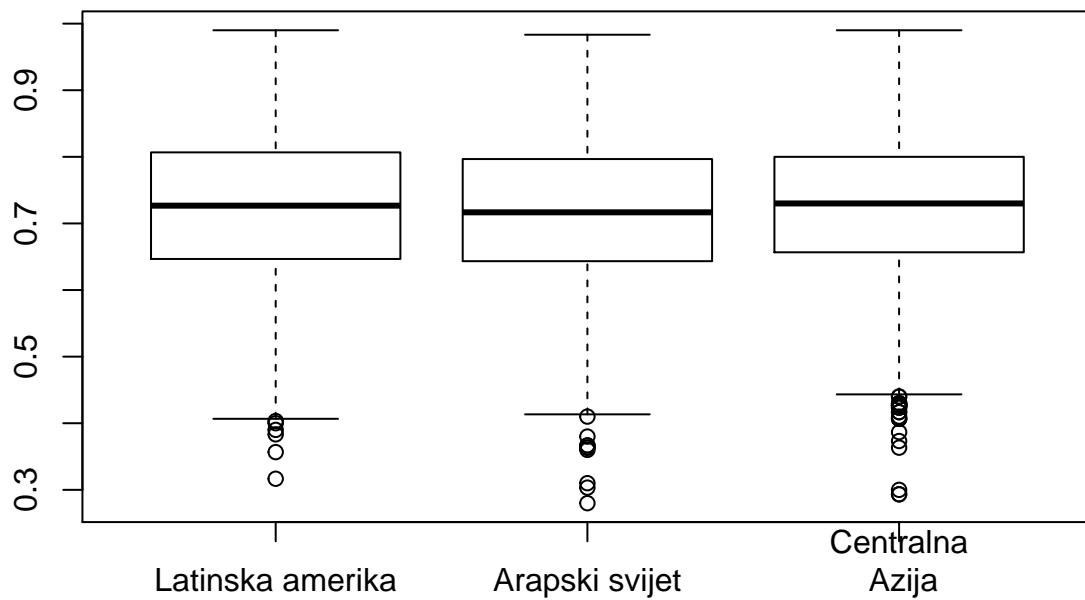
```
##      Min. 1st Qu. Median     Mean 3rd Qu.     Max.  
##  0.2800  0.6433  0.7167  0.7157  0.7958  0.9833
```

```
summary(central_asia$conscientiousness_score)
```

```
##      Min. 1st Qu. Median     Mean 3rd Qu.     Max.  
##  0.2933  0.6567  0.7300  0.7253  0.8000  0.9900
```

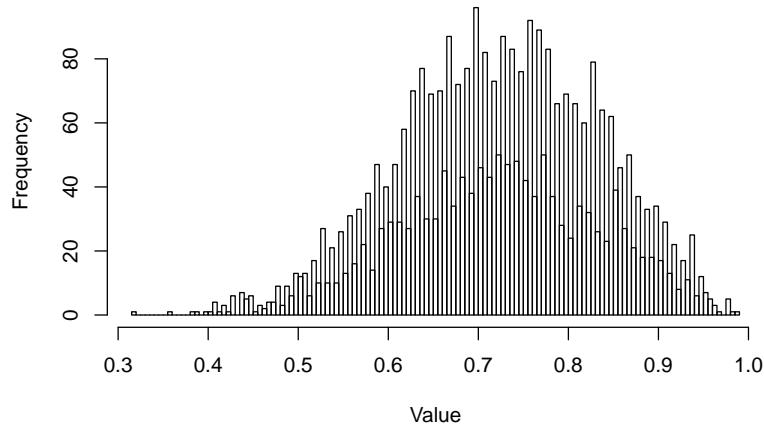
```
boxplot(latin_america$conscientiousness_score,arabic_world$conscientiousness_score,central_asia$conscientiousness_score,  
       names = c('Latinska amerika','Arapski svijet','Centralna\n Azija'),  
       main='Boxplot savjesnosti za zadane regije')
```

Boxplot savjesnosti za zadane regije



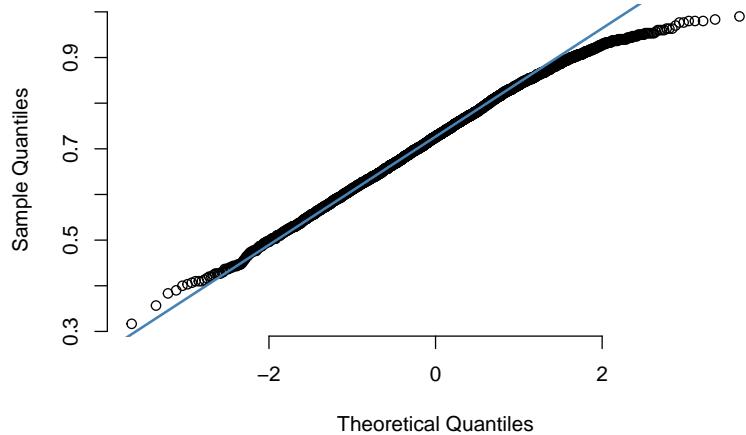
```
hist(latin_america$conscientiousness_score,main='Histogram savjesnosti Latinske Amerike',xlab='Value',ylab='Frequency')
```

Histogram savjesnosti Latinske Amerike



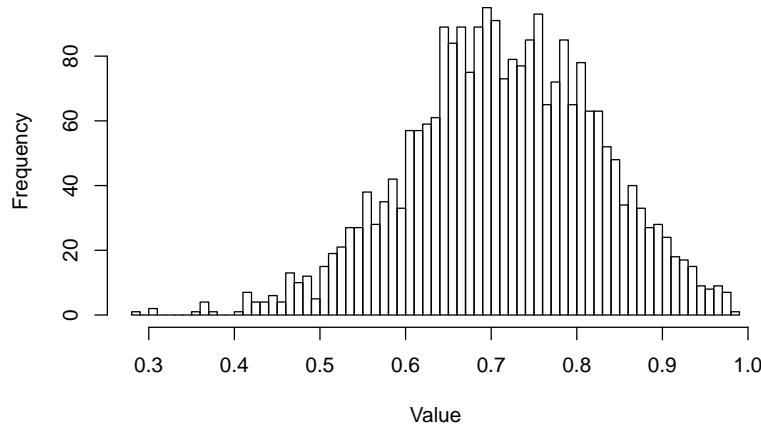
```
qqnorm(latin_america$conscientiousness_score, pch = 1, frame = FALSE, main='Savjesnost Latinske Amerike')
qqline(latin_america$conscientiousness_score, col = "steelblue", lwd = 2)
```

Savjesnost Latinske Amerike



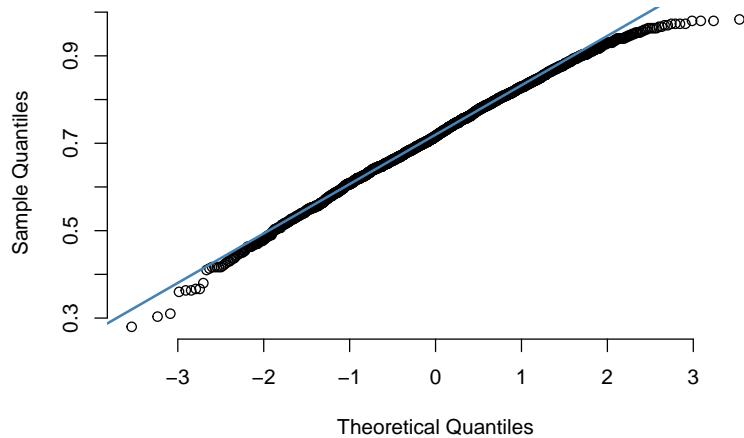
```
hist(arabic_world$conscientiousness_score, main='Histogram savjesnosti Arapskog svijeta', xlab='Value', ylab='Frequency')
```

Histogram savjesnosti Arapskog svijeta



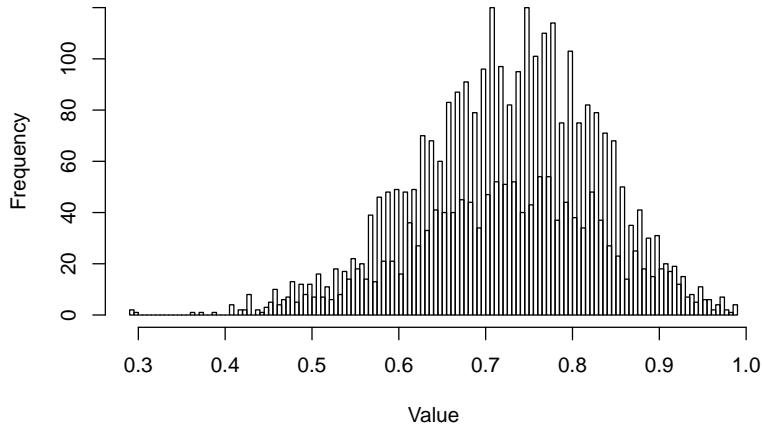
```
qqnorm(arabic_world$conscientiousness_score, pch = 1, frame = FALSE, main='Savjesnost Arapski svijet')
qqline(arabic_world$conscientiousness_score, col = "steelblue", lwd = 2)
```

Savjesnost Arapski svijet



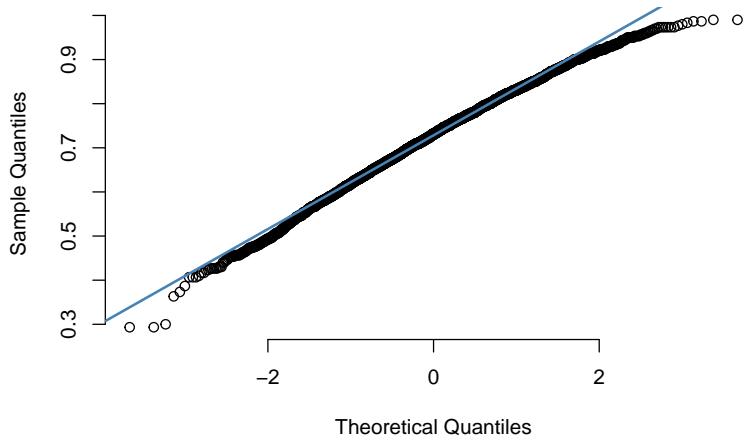
```
hist(central_asia$conscientiousness_score, main='Histogram savjesnosti centralne Azije', xlab='Value', ylab='Frequency')
```

Histogram savjesnosti centralne Azije



```
qqnorm(central_asia$conscientiousness_score, pch = 1, frame = FALSE, main='Savjesnost centralne Azije')
qqline(central_asia$conscientiousness_score, col = "steelblue", lwd = 2)
```

Savjesnost centralne Azije



Iz histograma i qq plota je vidljivo da se savjesnost svih zadanih regija ravna po normalnoj razdiobi.

Testiranje normalnosti distribucije

Hipoteze:

H_0 : distribucija regije pripada normalnoj razdiobi H_1 : distribucija regije NE pripada normalnoj razdiobi

```
lillie.test(latin_america$conscientiousness_score)
```

```
##  
## Lilliefors (Kolmogorov-Smirnov) normality test  
##
```

```

## data: latin_america$conscientiousness_score
## D = 0.026175, p-value = 3.018e-06

lillie.test(arabic_world$conscientiousness_score)

##
## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: arabic_world$conscientiousness_score
## D = 0.022587, p-value = 0.005327

lillie.test(central_asia$conscientiousness_score)

##
## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: central_asia$conscientiousness_score
## D = 0.032478, p-value = 1.122e-10

```

Za savjesnost svake regije na Lillieforovom testu smo dobili izuzetno malenu p vrijednost te odbacujemo nultu hipotezu za svaku regiju. Zaključak na osnovu Lillieforova testa jest da se nijedna distribucija savjesnosti ne ravna po normalnoj razdiobi. Ali na osnovu histograma i qq plota u nastavku pretpostavljamo da se svaka distibucija savjesnosti ravna po normalnoj razdiobi.

Testiranje homogenosti varijance između populacija

Nulta hipoteza je da su varijance savjesnosti svih regija jednake. Alternativa je da nisu odnosno da se barem jedan par varijanci razlikuje.

$$H_0 : \sigma_{latinAmerica}^2 = \sigma_{arabicWorld}^2 = \sigma_{centralAsia}^2 \\ H_1 : \neg H_0.$$

Za testiranje ove hipoteze koristimo Bartlettov test.

```

score = c(latin_america$conscientiousness_score,arabic_world$conscientiousness_score,central_asia$conscientiousness_score)

la_vec = rep(c("latin_america"),length(latin_america$conscientiousness_score))
arb_vec = rep(c("arabic_world"),length(arabic_world$conscientiousness_score))
asi_vec = rep(c("central_asia"),length(central_asia$conscientiousness_score))
region = c(la_vec,arb_vec,asi_vec)

df = data.frame(region,score)

bartlett.test(df$score ~ df$region)

##
## Bartlett test of homogeneity of variances
##
## data: df$score by df$region
## Bartlett's K-squared = 9.0166, df = 2, p-value = 0.01102

```

```

var(latin_america$conscientiousness_score)

## [1] 0.01216558

var(arabic_world$conscientiousness_score)

## [1] 0.01235618

var(central_asia$conscientiousness_score)

## [1] 0.01124714

```

Na osnovu p vrijednosti od 0.01102 odbacujemo nultu hipotezu i zaključujem da varijance ove tri populacije nisu homogene. Ali za potrebe testiranja u nastavku prepostavljam da jesu.

Testiranje jednakosti savjesnosti

Provjeravamo postoje li razlike u savjesnosti između regija. Kao nultu hipotezu prepostavljam da sve odabранe regije imaju jednako očekivanje razinu savjesnosti naspram alternativne hipoteze da nemaju to jest da barem jedan par regija nema jednako očekivanje za razinu savjesnosti.

$$H_0 : \mu_{latinAmerica}^2 = \mu_{arabicWorld}^2 = \mu_{centralAsia}^2 \\ H_1 : \neg H_0.$$

U svrhu testiranja naših hipoteza koristit ćemo ANOVA test. Zbog prepostavki ANOVA testa prepostavljam da su distribucije normalne, varijance homogene i podaci nezavisni.

```

# Test
a = aov(df$score ~ df$region)
summary(a)

##                               Df Sum Sq Mean Sq F value    Pr(>F)
## df$region          2   0.15  0.07578   6.396 0.00167 ***
## Residuals      10374 122.91  0.01185
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Na osnovu p vrijednosti od 0.00167 odbacujemo nultu hipotezu u korist alternative da očekivanje razine savjesnosti nije isto za svaku odabranu regiju.

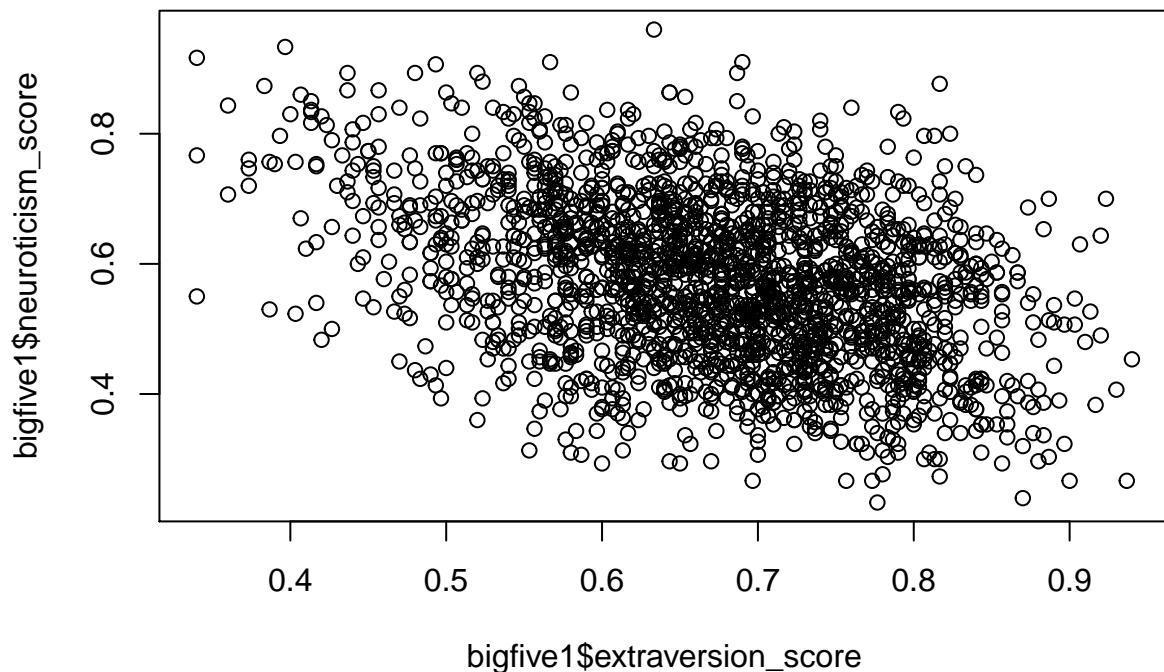
Motivacija: Možemo li zaključiti nešto o jednoj osobini uzimajući kao ulazne variable jednu ili više preostalih osobina?

Ono što očekujemo je da ne bi smjela postojati visoka korelacija između pojedinih osobina ličnosti pošto svaka od njih na jedinstven način opisuje određeni element ljudske osobnosti. Uz ovih 5, poznat je i velik broj drugih osobina, ali ovih 5 je izabrano na način da u idealnom slučaju u potpunosti definiraju osobnost neke osobe i razlikuju je od osobnosti drugih pojedinaca.

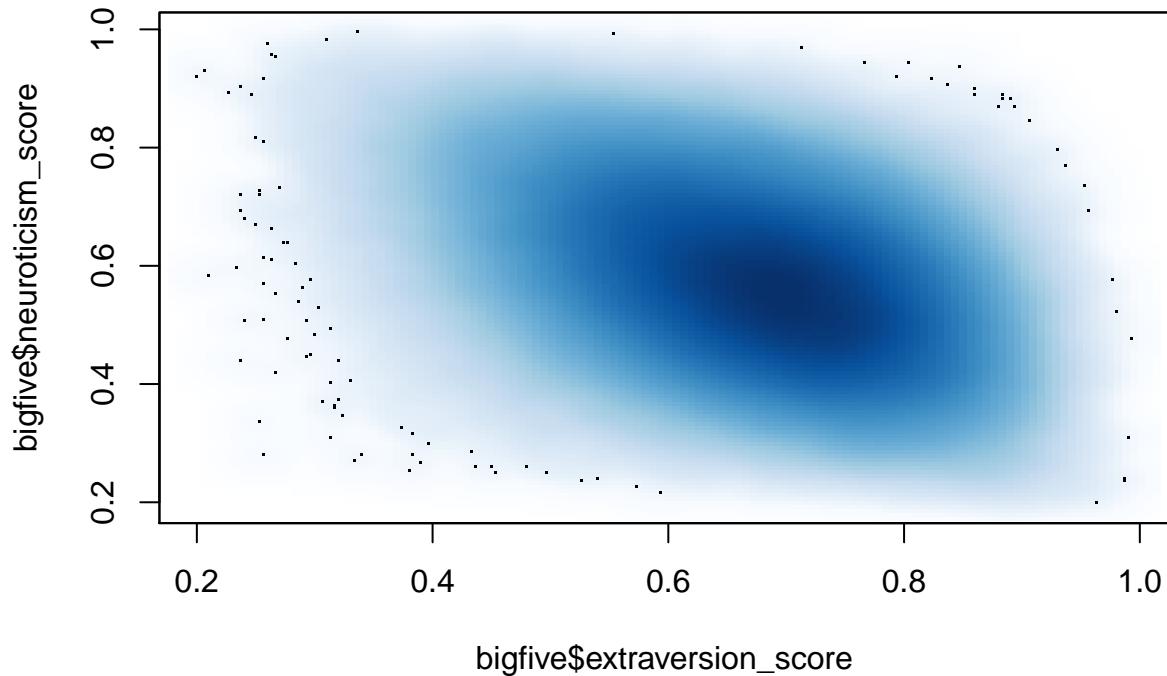
Kao izlaznu varijablu proizvoljno smo odabrali neuroticizam, a pregledavajući znanstvenu literaturu došli smo do zaključka da bi ekstraverzija mogla u najvećoj mjeri objasniti varijaciju u neuroticizmu što donekle i ima smisla jer osobe koje su introverti teže artikuliraju svoje emocije te su skloniji anksioznosti i depresiji.

Kako bismo mogli bolje uočiti potencijalnu vezu dviju varijabli, uzet ćemo slučajan poduzorak našeg ulaznog skupa podataka.

```
bigfive1 = bigfive[sample(nrow(bigfive), 2000),]  
plot(bigfive1$extraversion_score, bigfive1$neuroticism_score)
```



```
smoothScatter(bigfive$extraversion_score, bigfive$neuroticism_score)
```

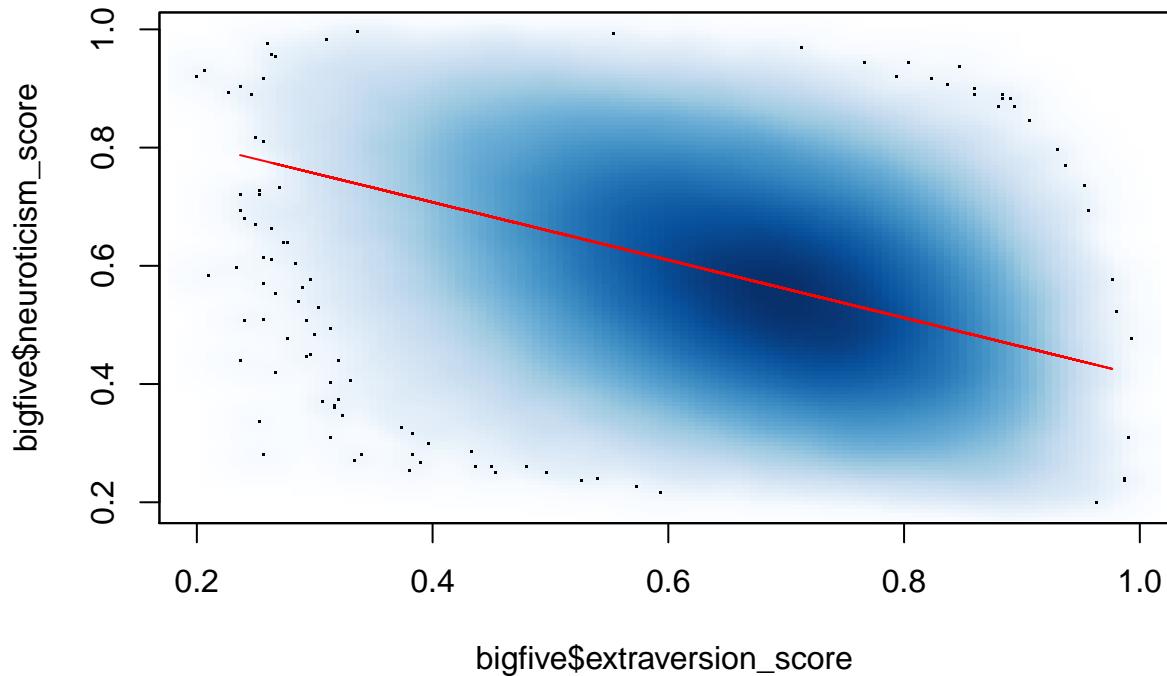


Scatter plot pokazuje da postoji negativan efekt varijable extraversion_score na izlaznu varijablu neuroticism_score, odnosno osobe koje postiže nizak score na ekstraverziji, postižu visok score na neuroticizmu. Zato ćemo pokušati kreirati model jednostavne regresije koji uključuje spomenute varijable:

$$Y = \beta_0 + \beta_1 x_1 + \epsilon$$

Pritom pretostavljamo da je ϵ normalno distribuirana slučajna varijabla (pogreška) s homogenom varijancom te da su pritom $\epsilon_1, \epsilon_2, \dots, \epsilon_n$ međusobno nezavisne, odnosno nekorelirane slučajne varijable.

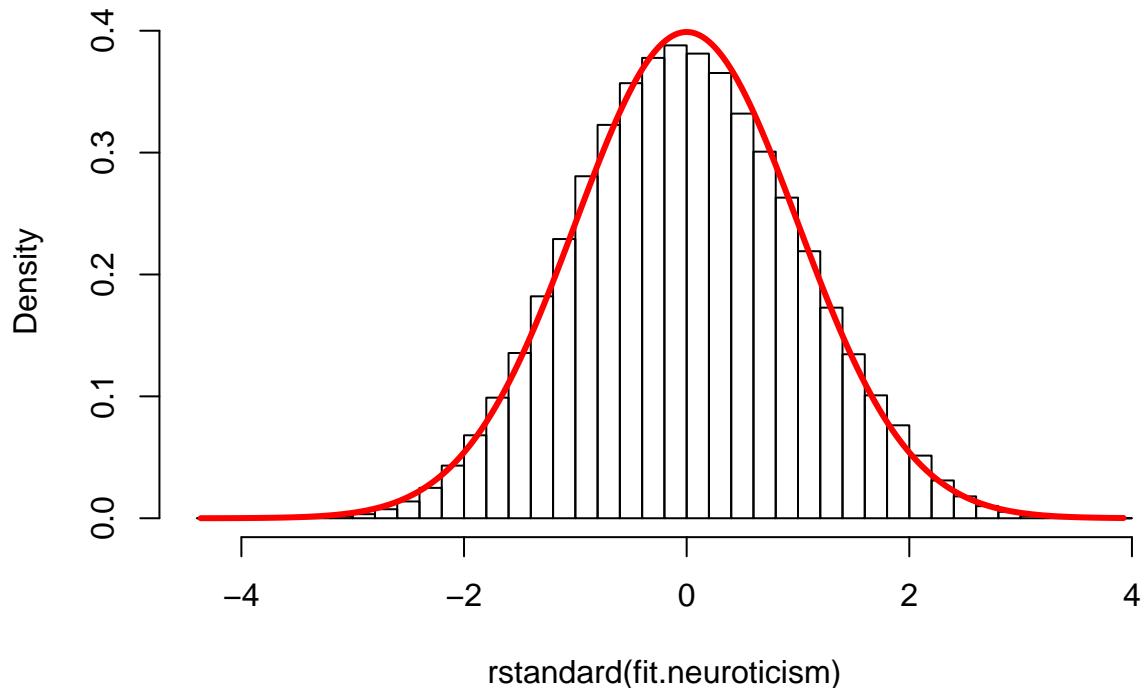
```
fit.neuroticism = lm(neuroticism_score ~ extraversion_score, data = bigfive)
smoothScatter(bigfive$extraversion_score, bigfive$neuroticism_score)
lines(bigfive$extraversion_score, fit.neuroticism$fitted.values, col = "red")
```



Sada ćemo provjeriti pretpostavke o normalnosti reziduala i homogenosti varijance. To ćemo učiniti grafički pomoću histograma i q-q plota te statističkim testovima - Kolmogorov-Smirnovljev test i Lillieforsova korekcija.

```
hist(rstandard(fit.neuroticism), probability = TRUE, breaks = 30)
y <- seq(min(rstandard(fit.neuroticism)), max(rstandard(fit.neuroticism)), 0.01)
lines(y, dnorm(y, mean(rstandard(fit.neuroticism)), sd(rstandard(fit.neuroticism))), col = "red", lwd =
```

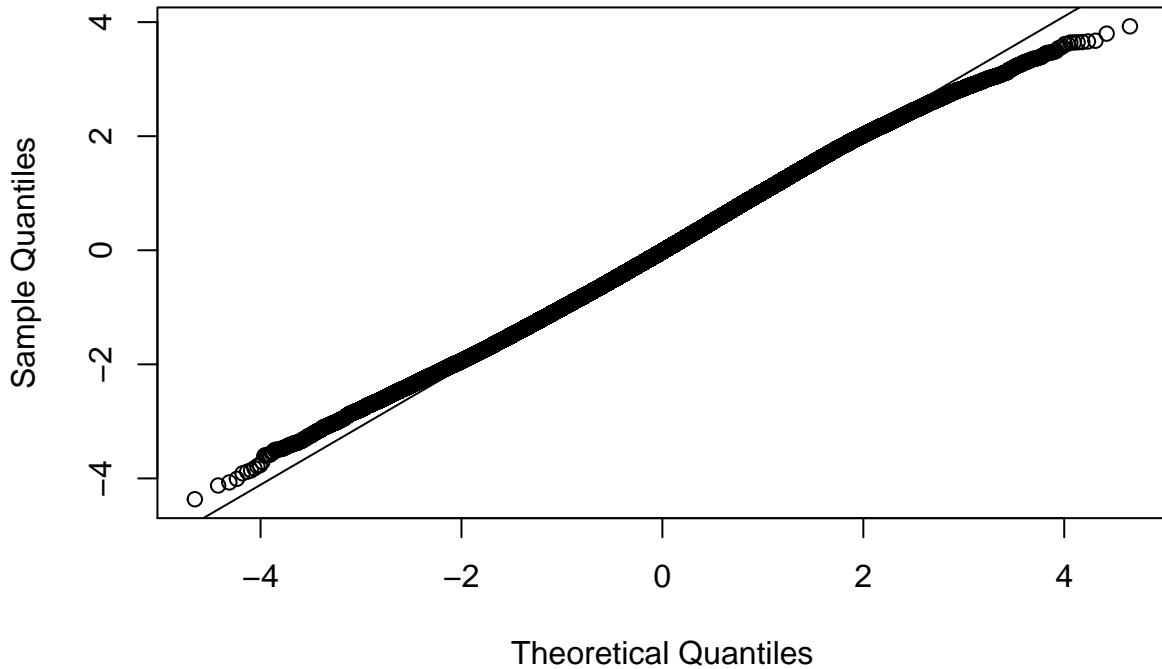
Histogram of rstandard(fit.neuroticism)



Iz histograma se može naslutiti normalna distribucija. Vidimo da se graf funkcije gustoće vjerojatnosti normalne slučajne varijable s parametrima uzoračke distribucije prilično dobro podudara s funkcijom gustoće vjerojatnosti varijable neuroticism_score.

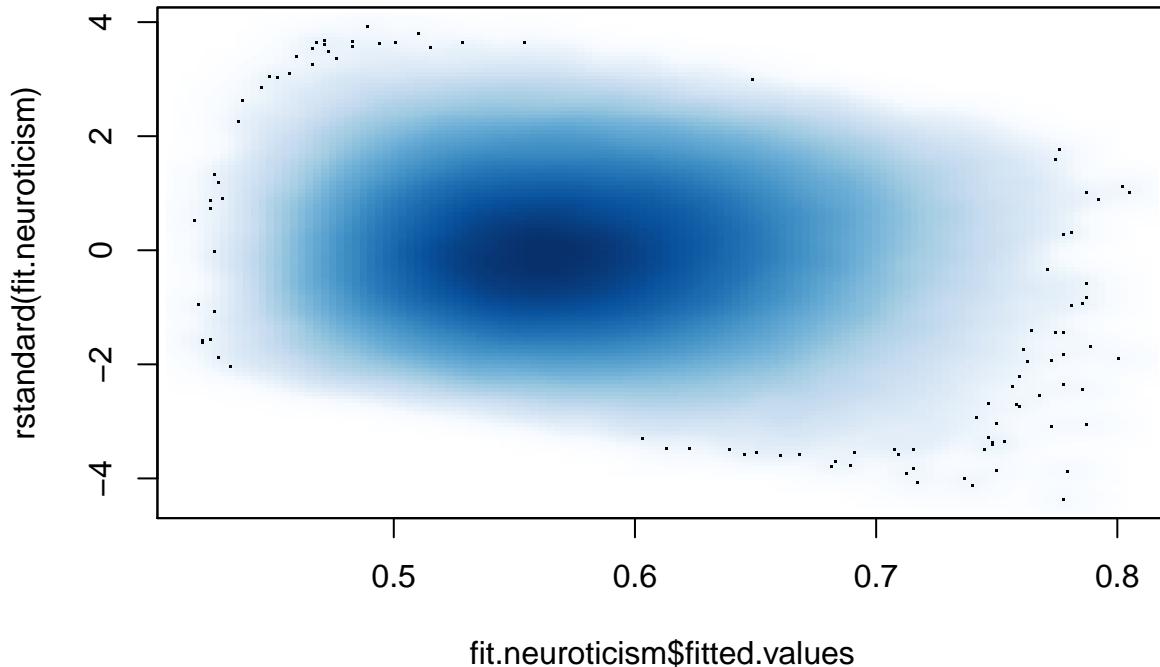
```
qqnorm(rstandard(fit.neuroticism))
qqline(rstandard(fit.neuroticism))
```

Normal Q-Q Plot



q-q plot pokazuje da se kvantili uzoračke distribucije u velikoj mjeri slažu s teorijskim kvantilima normalne distribucije. U repovima distribucije postoje manja odstupanja što možemo objasniti mogućim postojanjem outliera.

```
smoothScatter(fit.neuroticism$fitted.values, rstandard(fit.neuroticism))
```



Iz scatter plota koji pokazuje standardizirane reziduale u odnosu na procjene modela uočavamo da nema naznake nehomogenosti pogreške.

```
ks.test(rstandard(fit.neuroticism), "pnorm")

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

##
## One-sample Kolmogorov-Smirnov test
##
## data: rstandard(fit.neuroticism)
## D = 0.0095406, p-value < 2.2e-16
## alternative hypothesis: two-sided

require(nortest)
lillie.test(rstandard(fit.neuroticism))

##
## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: rstandard(fit.neuroticism)
## D = 0.0095403, p-value < 2.2e-16
```

Iz p vrijednost KS-testa možemo zaključiti da uz nivo značajnosti od 5% ne možemo odbaciti nullu hipotezu o normalnosti, dok iz Lillieforsove inačice možemo izvući isti zaključak.

Važno je napomenuti da u ovom slučaju KS-test je bolji izbor s obzirom da standardizacijom reziduala njihovo teoretsko očekivanje je nula i varijanca jedan. Mi u ovom slučaju KS-testom uspoređujemo uzoračku distribuciju s jediničnom normalnom razdiobom. Lillieforsova inačica dizajnirana je više za slučaj kada ne poznajemo ni varijancu ni očekivanje varijabli, već ih procjenjujemo iz uzorka.

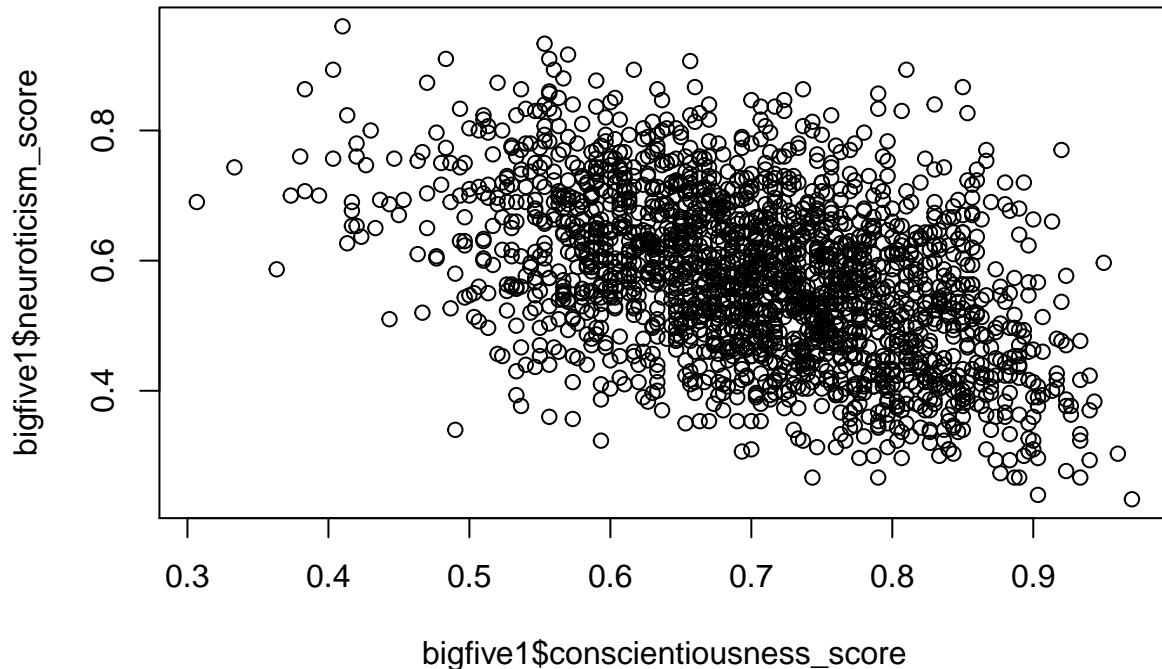
```
fit.neuroticism = lm(neuroticism_score ~ extraversion_score, data = bigfive)
summary(fit.neuroticism)
```

```
##
## Call:
## lm(formula = neuroticism_score ~ extraversion_score, data = bigfive)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.49754 -0.07955 -0.00210  0.07830  0.44747
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 0.902976  0.001288 700.8   <2e-16 ***
## extraversion_score -0.488714  0.001892 -258.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.114 on 307311 degrees of freedom
## Multiple R-squared:  0.1784, Adjusted R-squared:  0.1784 
## F-statistic: 6.673e+04 on 1 and 307311 DF,  p-value: < 2.2e-16
```

Iz vrijednost t testova možemo zaključiti da su oba koeficijenta statistički značajna uz bilo koju razinu značajnosti. Uz to f test je pokazao da je i cijeli model signifikantan. Ipak, nizak koeficijent determinacije pokazuje da je samo mali dio promjena zavisne varijable, konkretno 17.84% objašnjeno dobivenim modelom jednostavne linearne regresije.

Osim ekstraverzije kao potencijalne osobine iz koje bismo mogli nešto zaključiti o neuroticizmu osobe, pokazalo se da bi još jedna osobina u Big Five modelu mogla objašnjavati rezultat koji osoba postiže na neuroticizmu, a to je savjesnost. To možemo interpretirati time da osobe koje su nediscplinirane, neuredne i nemaju rutinu mogu biti sklonije neurotičnom ponašanju od savjesnih i organiziranih ljudi.

```
plot(bigfive1$conscientiousness_score, bigfive1$neuroticism_score)
```



Iz scatter plota možemo uočiti naznaku negativne korelacije, odnosno zavisnosti dviju varijabli što odgovara našoj početnoj tezi. Kako bismo kvantificirali navedenu korelaciju, možemo izračunati Pearsonov korelacijski koeficijent za ove dvije varijable, a dodatno možemo pogledati korelacijsku tablicu svih 5 faktora što će nam pomoći i u kreiranju modela višestruke regresije i provjere njenih prepostavki.

```

correlations <- cor(cbind(bigfive$agreeable_score, bigfive$extraversion_score, bigfive$openness_score, b

colnames(correlations) <- c("agreeable_score", "extraversion_score", "openness_score", "conscientiousne
rownames(correlations) <- c("agreeable_score", "extraversion_score", "openness_score", "conscientiousne

library(corrplot)

## corrplot 0.92 loaded

corrplot(correlations, method = "number")

```



```
cor.test(bigfive$extraversion_score, bigfive$neuroticism_score)
```

```
##
## Pearson's product-moment correlation
##
## data: bigfive$extraversion_score and bigfive$neuroticism_score
## t = -258.32, df = 307311, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## -0.4252762 -0.4194666
## sample estimates:
##       cor
## -0.4223758
```

Iz korelacijske tablice naslućujemo da postoji određena veza između varijabli neuroticism_score i conscientiousness_score. Ono što također možemo uočiti je da nema kršenja prepostavki, tj. da nema značajne korelacije između dviju varijabli extraversion_score i conscientiousness_score što nam ujedno pokazuje i t-test Pearsonovog korelacijskog koeficijenta tih dviju varijabli.

```
fit.neuroticism_m = lm(neuroticism_score ~ extraversion_score + conscientiousness_score, bigfive)
summary(fit.neuroticism_m)
```

```
##
## Call:
```

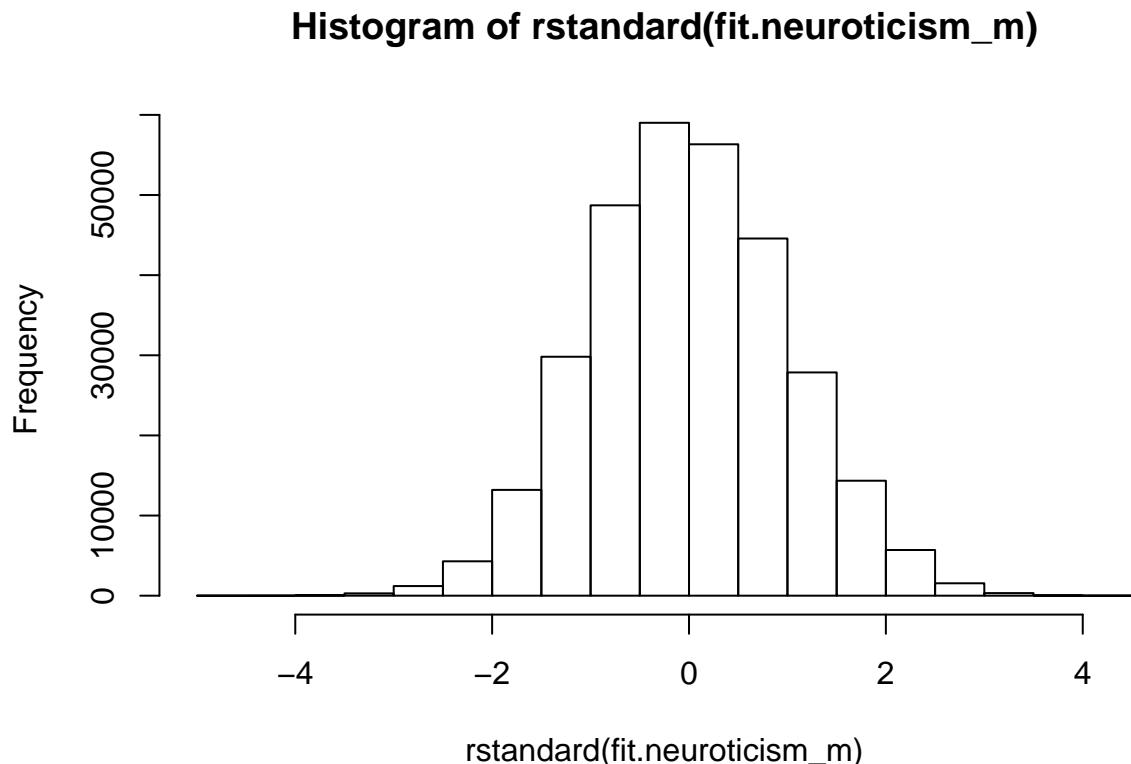
```

## lm(formula = neuroticism_score ~ extraversion_score + conscientiousness_score,
##     data = bigfive)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.50744 -0.07071 -0.00254  0.06919  0.42177
## 
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.187790  0.001553  764.8 <2e-16 ***
## extraversion_score     -0.414690  0.001717 -241.5 <2e-16 ***
## conscientiousness_score -0.476616  0.001738 -274.2 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.1022 on 307310 degrees of freedom
## Multiple R-squared:  0.3399, Adjusted R-squared:  0.3399
## F-statistic: 7.914e+04 on 2 and 307310 DF,  p-value: < 2.2e-16

```

Iz modela multivariatne regresije možemo uočiti da su svi koeficijenti, a uz to i sam model signifikantni uz bilo koju razinu značajnosti. Također, koeficijent determinacije se povećao u odnosu na model jednostavne regresije što znači da je conscientiousness_score doprinio poboljšanju funkcije ovisnosti izlazne varijable o regresorima, odnosno da je veća proporcija varijacije varijable Y objašnjena ovim modelom.

```
hist(rstandard(fit.neuroticism_m))
```



```

ks.test(rstandard(fit.neuroticism_m), "pnorm")

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

## 
## One-sample Kolmogorov-Smirnov test
##
## data: rstandard(fit.neuroticism_m)
## D = 0.010805, p-value < 2.2e-16
## alternative hypothesis: two-sided

```

Iz grafičkog prikaza i KS-testa zaključujemo da nema kršenja pretpostavke normalnosti reziduala. Možemo probati uključiti i preostala dva faktora u model multivarijantne regresije.

```

fit.multi = lm(neuroticism_score ~ extraversion_score + conscientiousness_score + agreeable_score + openness_score, data = bigfive)
summary(fit.multi)

```

```

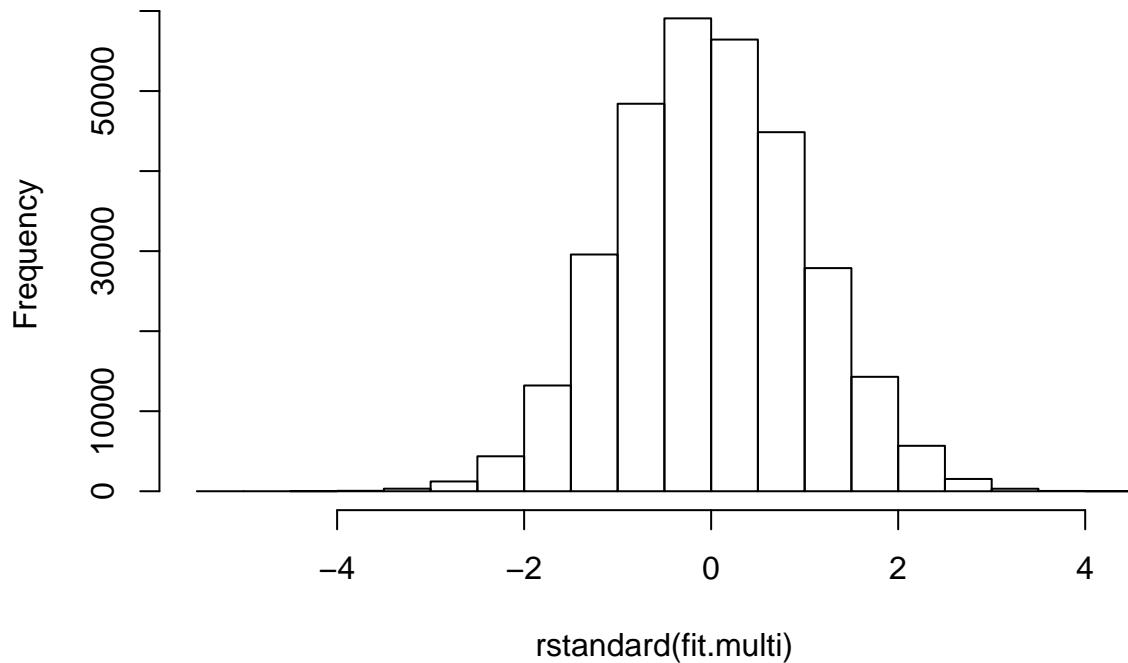
## 
## Call:
## lm(formula = neuroticism_score ~ extraversion_score + conscientiousness_score +
##      agreeable_score + openness_score, data = bigfive)
##
## Residuals:
##       Min     1Q     Median      3Q     Max
## -0.51118 -0.07058 -0.00223  0.06913  0.42830
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.170481  0.002231 524.55   <2e-16 ***
## extraversion_score     -0.425744  0.001773 -240.10   <2e-16 ***
## conscientiousness_score -0.462549  0.001843 -250.98   <2e-16 ***
## agreeable_score         -0.037587  0.002109  -17.82   <2e-16 ***
## openness_score          0.055941  0.002212   25.29   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.102 on 307308 degrees of freedom
## Multiple R-squared:  0.3417, Adjusted R-squared:  0.3417
## F-statistic: 3.988e+04 on 4 and 307308 DF,  p-value: < 2.2e-16

```

Vidimo da su opet svi koeficijenti kao i sam model signifikantni do na bilo koju razinu značajnosti. Uz to, iz procjena nepoznatih parametara zaključujemo da najveći efekt na izlaznu varijablu imaju regresori extraversion_score i conscientiousness_score. Povećanje koeficijenta determinacije postoji, ali je zanemarivo pa možemo zaključiti da dodani faktori ne objašnjavaju u značajnijoj mjeri neobjašnjenu varijaciju modela s postojeća dva faktora. Ono što bismo u ovom slučaju očekivali je da se prilagođeni koeficijent determinacije koji penalizira dodatne parametre smanji, ali pošto se radi o velikom uzorku izostao je takav efekt.

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

Histogram of rstandard(fit.multi)



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

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

##
## One-sample Kolmogorov-Smirnov test
##
## data: rstandard(fit.multi)
## D = 0.010145, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

Iz grafičkog prikaza i KS-testa zaključujemo da nema kršenja pretpostavke normalnosti reziduala.

Sada ćemo pokušati primijeniti isti model na muškarce starije od 22 godina.

```
fit.multi = lm(neuroticism_score ~ extraversion_score + conscientiousness_score + agreeable_score + open
summary(fit.multi)

##
## Call:
## lm(formula = neuroticism_score ~ extraversion_score + conscientiousness_score +
##     agreeable_score + openness_score, data = bigfive[bigfive$sex ==
##     1 & bigfive$age > 22, ])
```

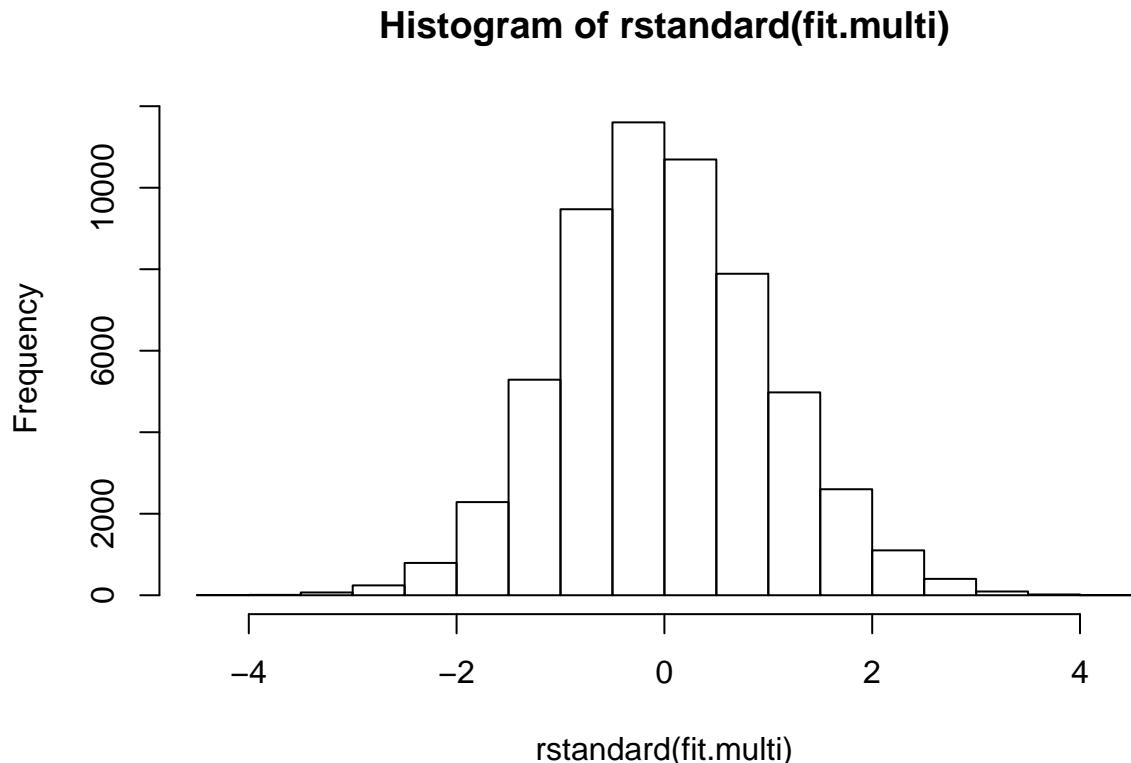
```

## 
## Residuals:
##      Min       1Q   Median      3Q      Max
## -0.42838 -0.06545 -0.00413  0.06258  0.41472
## 
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.242882  0.004819 257.91 <2e-16 ***
## extraversion_score   -0.454399  0.003933 -115.54 <2e-16 ***
## conscientiousness_score -0.496085  0.004188 -118.44 <2e-16 ***
## agreeable_score        -0.146280  0.004605  -31.77 <2e-16 ***
## openness_score         0.063095  0.004825   13.08 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.09616 on 57541 degrees of freedom
## Multiple R-squared:  0.45, Adjusted R-squared:  0.4499
## F-statistic: 1.177e+04 on 4 and 57541 DF, p-value: < 2.2e-16

```

Signifikantnost koeficijenata i modela ostala je statistički značajna, ali vidimo da se koeficijent determinacije povećao. Razlog tome je što se za osobnost mlađih ljudi smatra da je još u razvoju te je sukladno tome sklona većim varijacijama od one kod starijih ljudi (<https://www.tandfonline.com/doi/abs/10.1080/10478401003596626>). Uz to, pokazalo se da ukoliko se promatraju samo ispitanici muškog spola, da model još bolje objašnjava varijaciju izlazne varijable.

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



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

```
##  
##  One-sample Kolmogorov-Smirnov test  
##  
## data: rstandard(fit.multi)  
## D = 0.017863, p-value = 2.22e-16  
## alternative hypothesis: two-sided
```

Iz grafičkog prikaza i KS-testa zaključujemo da nema kršenja pretpostavke normalnosti reziduala.

Karakteristike prema spolu ispitanika Postoje li razlike u karakteristikama prema spolu ispitanika? Možemo li temeljem nekih drugih varijabli odrediti spol ispitanika? Želimo ispitati postoje li dovoljne razlike između rezultata testiranja između pripadnika različitih spolova da bi mogli na temelju rezultata ispitivanja odrediti je li test rješava žena ili muškarac.

Kako bismo mogli predvidjeti spol ispitanika, možemo ispitati različite varijable koje bi mogle utjecati na spol:

- Neuroticism_score
- Agreeableness_score
- Conscientiousness_score
- Extraversion_score
- Openness/Intellect
- Age

Prvi korak je razvrstati podatke u dvije grupe - muškarci i žene. Konkretno u ovome skupu 1 predstavlja muški spol, dok 2 predstavlja ženski spol.

```
male_subjects = bigfive[bigfive$sex == 1 ,]  
female_subjects = bigfive[bigfive$sex == 2 ,]
```

Zatim možemo usporediti srednje vrijednosti za 5 kategorija kako bi uočili po kojim karakteristikama se najviše razlikuju spolovi:

```
cat("Razlika između srednji vrijednosti:\n")
```

```
## Razlika između srednji vrijednosti:
```

```
srednja_vrijednost <- matrix(c(  
  mean(female_subjects$neuroticism_score),  
  mean(female_subjects$extraversion_score),  
  mean(female_subjects$agreeable_score),  
  mean(female_subjects$conscientiousness_score),  
  mean(female_subjects$openness_score),  
  mean(male_subjects$neuroticism_score),  
  mean(male_subjects$extraversion_score),  
  mean(male_subjects$agreeable_score),  
  mean(male_subjects$conscientiousness_score),  
  mean(male_subjects$openness_score)),ncol=2)
```

```

colnames(srednja_vrijednost) <- c("Žene", "Muškarci")
rownames(srednja_vrijednost) <- c("neuroticism_score", "extraversion_score", "agreeable_score", "conscientiousness_score", "openness_score")
print(srednja_vrijednost)

##                                     Žene   Muškarci
## neuroticism_score      0.5944653 0.5439874
## extraversion_score     0.6792409 0.6618528
## agreeable_score        0.7140516 0.6706726
## conscientiousness_score 0.7059135 0.6960649
## openness_score         0.7410141 0.7232220

cat("\nRazlika između varijanci:\n")

## 
## Razlika između varijanci:

varijance <- matrix(c(
  var(female_subjects$neuroticism_score),
  var(female_subjects$extraversion_score),
  var(female_subjects$agreeable_score),
  var(female_subjects$conscientiousness_score),
  var(female_subjects$openness_score),
  var(male_subjects$neuroticism_score),
  var(male_subjects$extraversion_score),
  var(male_subjects$agreeable_score),
  var(male_subjects$conscientiousness_score),
  var(male_subjects$openness_score)), ncol=2)
colnames(varijance) <- c("Žene", "Muškarci")
rownames(varijance) <- c("neuroticism_score", "extraversion_score", "agreeable_score", "conscientiousness_score", "openness_score")
varijance <- as.table(varijance)
print(varijance)

##                                     Žene   Muškarci
## neuroticism_score      0.014981522 0.015537636
## extraversion_score     0.011221689 0.012522746
## agreeable_score        0.007872459 0.008956176
## conscientiousness_score 0.011353278 0.011739488
## openness_score         0.007302675 0.008049347

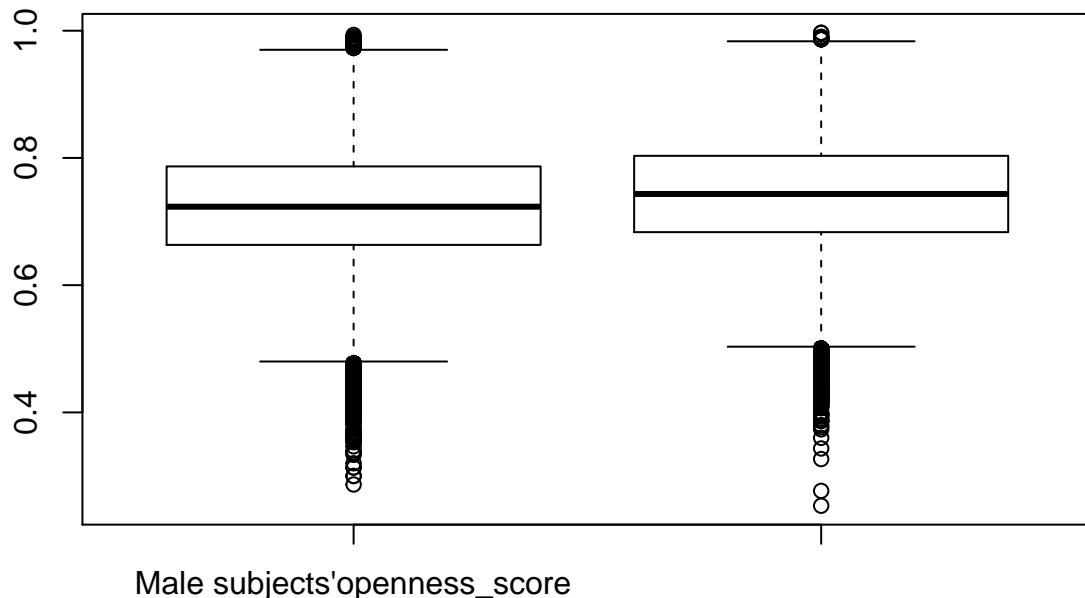
```

Nakon što usporedimo srednje vrijednosti prema svih 5 faktora, može se uočiti da između nekih postoji veća, odnosno manja razlika. Iako razlike nisu velike želimo ispitati postoji li mogućnost procjene spola na temelju ovih faktora. Također, najveća razlika se uočava između neuroticizma, ugodnosti i otvorenosti dok je između savjesnosti i ekstraverzije ona gotovo zanemariva. Prikazali smo podatke grafički kako bismo provjerili njihovu normalnost te pokušali uočiti vizualne karakteristike.

Vizualizacija podataka

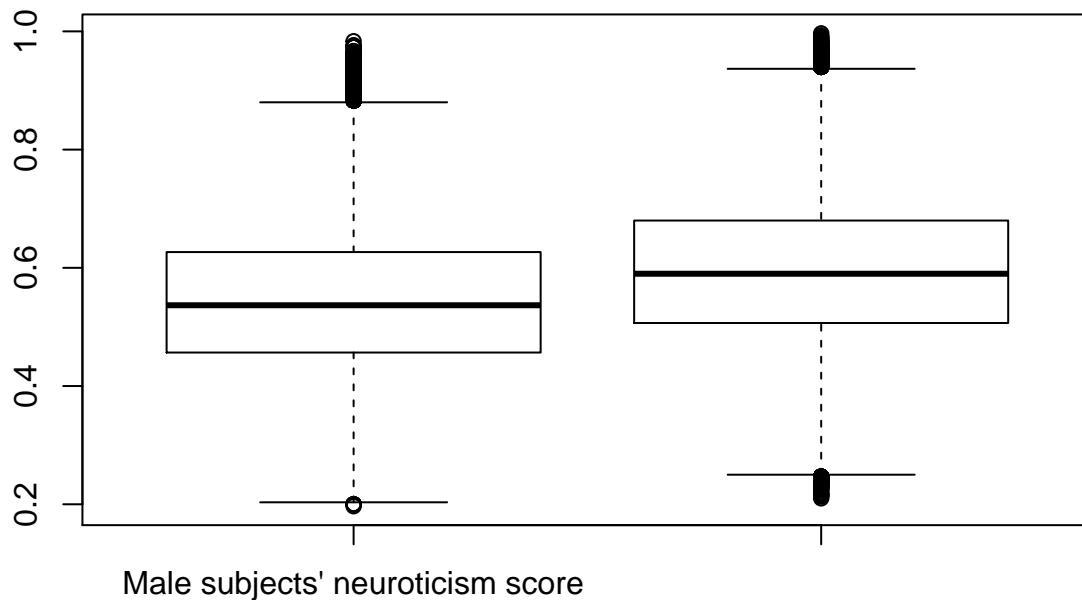
```
boxplot(male_subjects$openness_score, female_subjects$openness_score,
        names = c('Male subjects\' openness_score', 'Female subjects\' openness_score'),
        main='Boxplot of male and female subjects\' openness_score')
```

Boxplot of male and female subjects' openness_score



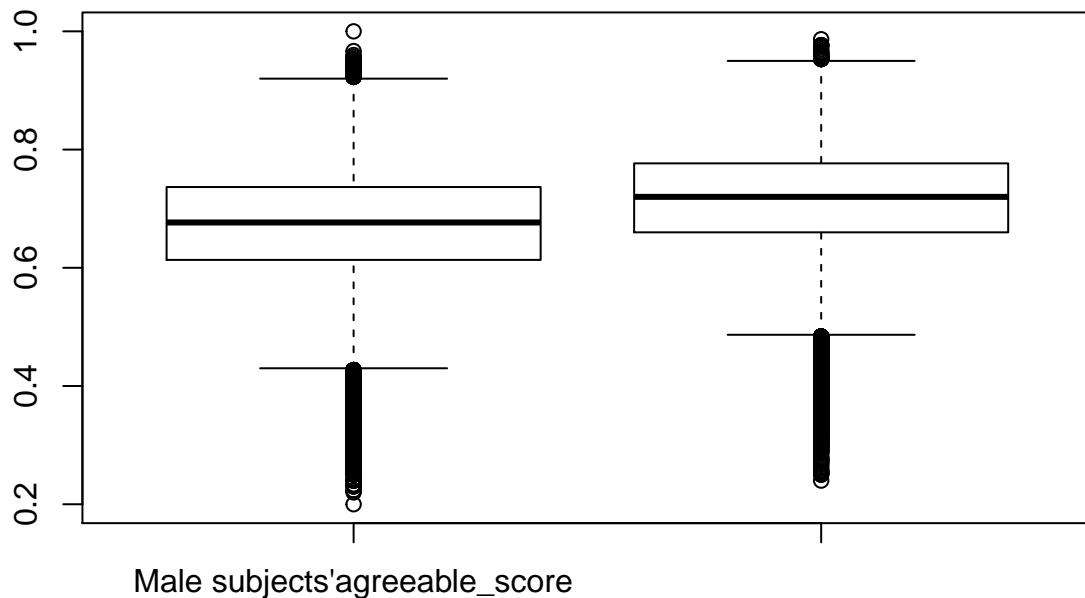
```
boxplot(male_subjects$neuroticism_score, female_subjects$neuroticism_score,
        names = c('Male subjects\' neuroticism score', 'Female subjects\' neuroticism score'),
        main='Boxplot of male and female subjects\' neuroticism_score')
```

Boxplot of male and female subjects' neuroticism_score



```
boxplot(male_subjects$agreeable_score, female_subjects$agreeable_score,
        names = c('Male subjects\' agreeable_score', 'Female subjects\' agreeable_score'),
        main='Boxplot of male and female subjects\' agreeable_score')
```

Boxplot of male and female subjects' agreeable_score

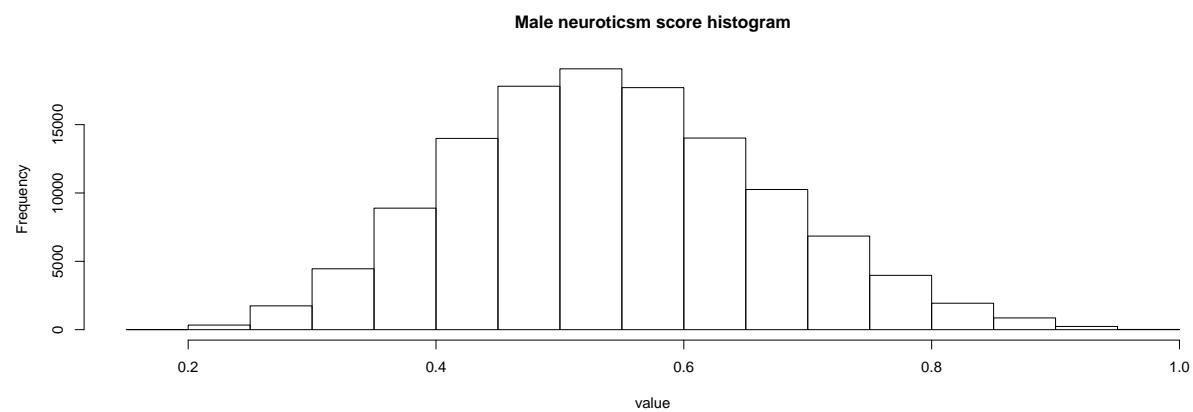


Možemo uočiti razliku u srednjoj vrijednosti neuroticizma između pripadnika ženskog i muškog spola. Stoga postoje indikacije da bi neuroticism_score trebao biti viši kod žena nego muškaraca pa ćemo tu hipotezu testirati t-testom. Prije samog testiranja moramo provjeriti pretpostavke nezavisnosti i normalnosti.

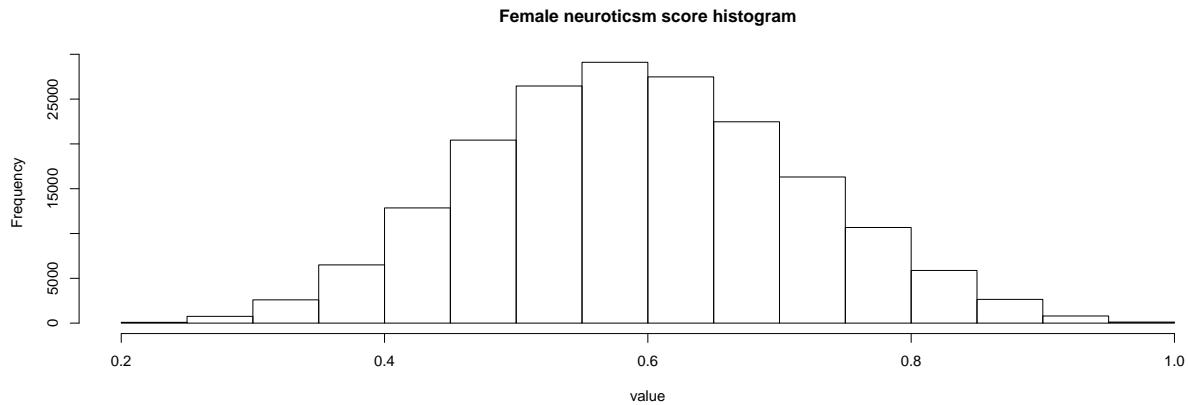
Pošto je u ovom slučaju nezavisnost zadovoljena, sljedeći korak je provjeriti normalnost podataka koju najčešće provjeravamo: histogramom, qq-plotom te KS-testom kojim provjeravamo pripadnost podataka distribuciji.

Sljedeće testove radimo samo za neuroticizam uz pretpostavku da se za ostale faktore podaci slično ponšaju.

```
hist(male_subjects$neuroticism_score, main='Male neuroticism score histogram', xlab='value', ylab='Frequency')
```

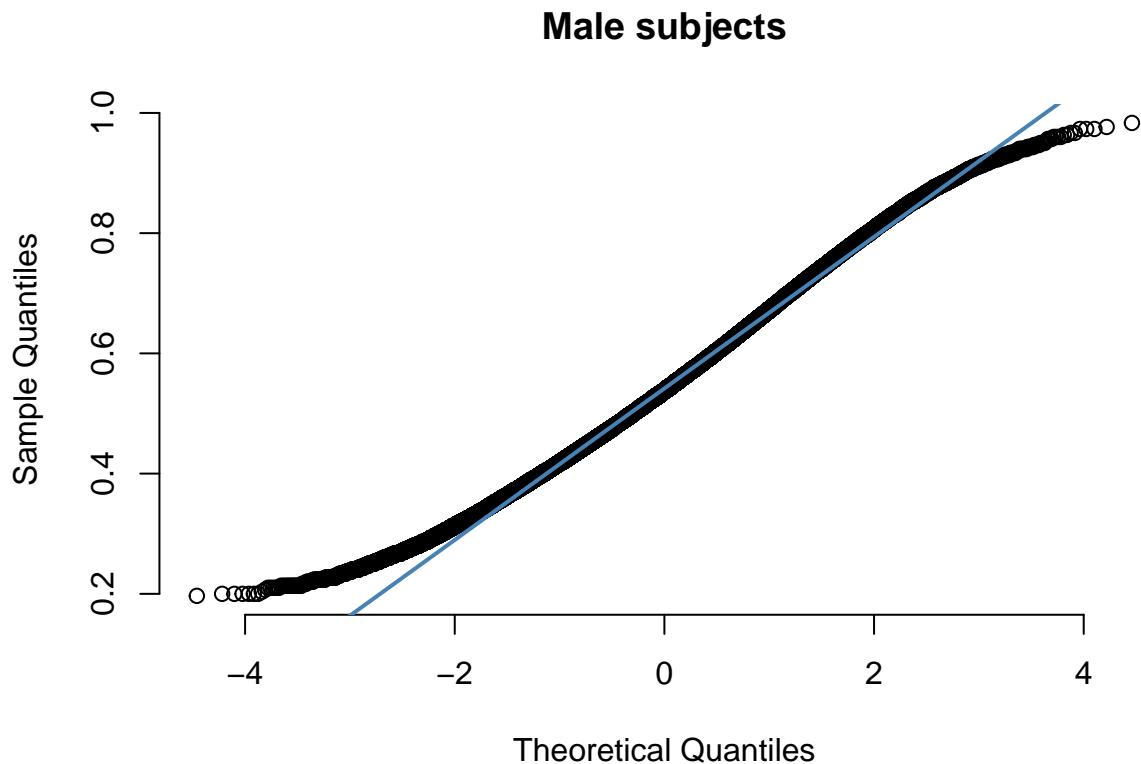


```
hist(female_subjects$neuroticism_score, main='Female neuroticism score histogram', xlab='value', ylab='Frequency')
```

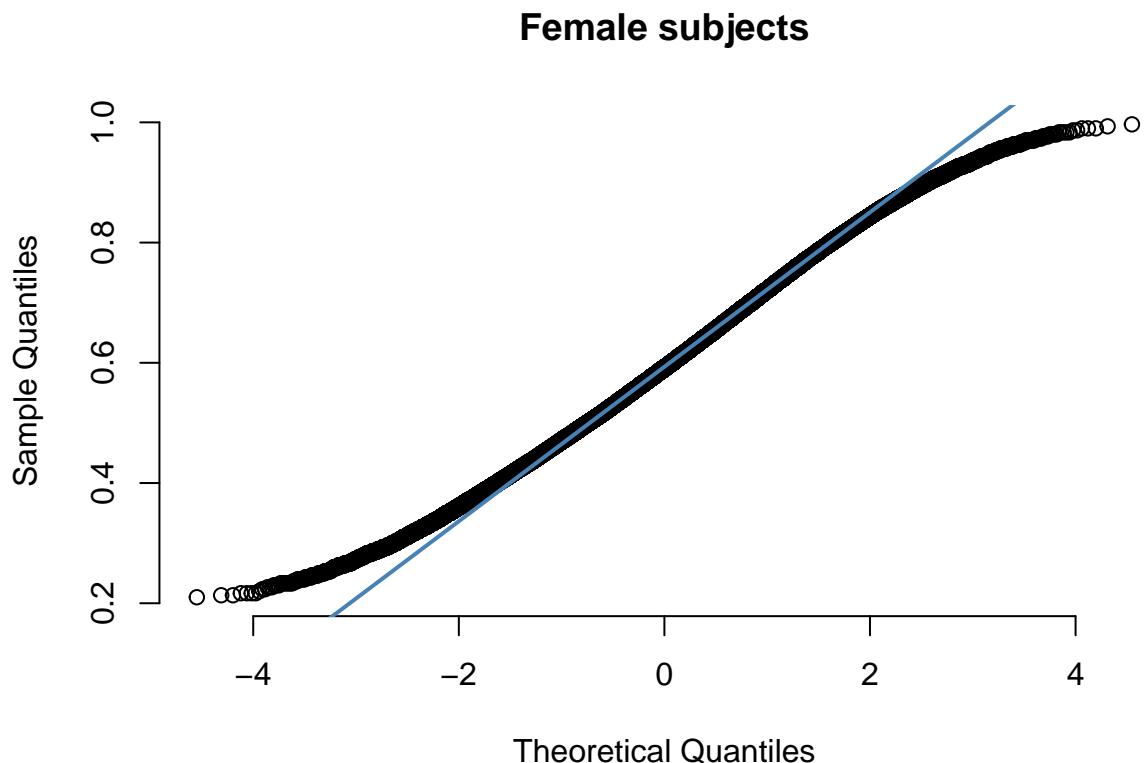


Histogrami upućuju na normalnost podataka.

```
qqnorm(male_subjects$neuroticism_score, pch = 1, frame = FALSE, main='Male subjects')
qqline(male_subjects$neuroticism_score, col = "steelblue", lwd = 2)
```



```
qqnorm(female_subjects$neuroticism_score, pch = 1, frame = FALSE, main='Female subjects')
qqline(female_subjects$neuroticism_score, col = "steelblue", lwd = 2)
```



Uzorački kvantili uglavnom se slažu s teorijskim kvantilima normalne distribucije osim u repovima distribucije gdje postoji manja odstupanja.

Testiranje normalnosti - Lillieforsova inačica Kolmogorov-Smirnovljevog testa

Hipoteze:

H_0 : distribucija male_subjects\$neuroticism_score pripada normalnoj razdiobi H_1 : distribucija male_subjects\$neuroticism_score NE pripada normalnoj razdiobi

```
lillie.test(male_subjects$neuroticism_score)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
##  data:  male_subjects$neuroticism_score
##  D = 0.025863, p-value < 2.2e-16
```

Hipoteze:

H_0 : distribucija female_subjects\$neuroticism_score pripada normalnoj razdiobi H_1 : distribucija female_subjects\$neuroticism_score NE pripada normalnoj razdiobi

```

lillie.test(female_subjects$neuroticism_score)

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: female_subjects$neuroticism_score
## D = 0.018533, p-value < 2.2e-16

```

S obzirom da je p-vrijednost $< \alpha$, odbacujemo nultu hipotezu u korist alternative da se neuroticizam muških/ženskih ispitanika ne ravna po normalnoj distibuciji, što je suprotno od onoga što smo uočili u grafovima. Uvezši u obzir oblik grafova, osjetljivost Lillieforsovog testa normalnosti i robusnosti t-testa na nenormalnost u slučaju velikog uzorka, možemo nastaviti testiranje uz pretpostavku normalnosti podataka.

Testiranje jednakosti varijanci neuroticizma za muške/ženske ispitanike

Pretpostavljam nezavisnost uzorka i izvodom F-test.

Hipoteze:

$$H_0 : \sigma_{Male}^2 = \sigma_{Female}^2$$

$$H_1 : \sigma_{Male}^2 \neq \sigma_{Female}^2$$

```

var.test(male_subjects$neuroticism_score, female_subjects$neuroticism_score)

```

```

##
##  F test to compare two variances
##
## data: male_subjects$neuroticism_score and female_subjects$neuroticism_score
## F = 1.0371, num df = 122163, denom df = 185148, p-value = 2.546e-12
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 1.026583 1.047776
## sample estimates:
## ratio of variances
## 1.03712

```

Na osnovu male p vrijednosti odbacujem nultu hipotezu u korist alternative da su varijance distribucije različite.

Testiranje jednakosti srednjih vrijednosti neuroticizma za muške/ženske ispitanike

Kako bismo mogli provoditi t-test uzorci moraju biti nezavisni i dolaziti iz normalne distribucije. Nezavisnost možemo utvrditi sa sigurnošću pošto je svaki ispitanik neovisan o drugom, odnosno svaki čovjek je zaseban pojedinac te njegovi rezultati ne ovise o rezultatima drugog pojedinca. Normalnost podataka smo ispitivali histogramima, qq-plotovima i statističkim testovima.

Koristimo t-test i prepostavljamo nezavisnost varijabli i nejednakost varijanci.

Hipoteze:

$$H_0 : \mu_{Female} = \mu_{Male}$$

$$H_1 : \mu_{Female} > \mu_{Male}$$

```
t.test(female_subjects$neuroticism_score ,male_subjects$neuroticism_score, alt = "greater", var.equal = TRUE)

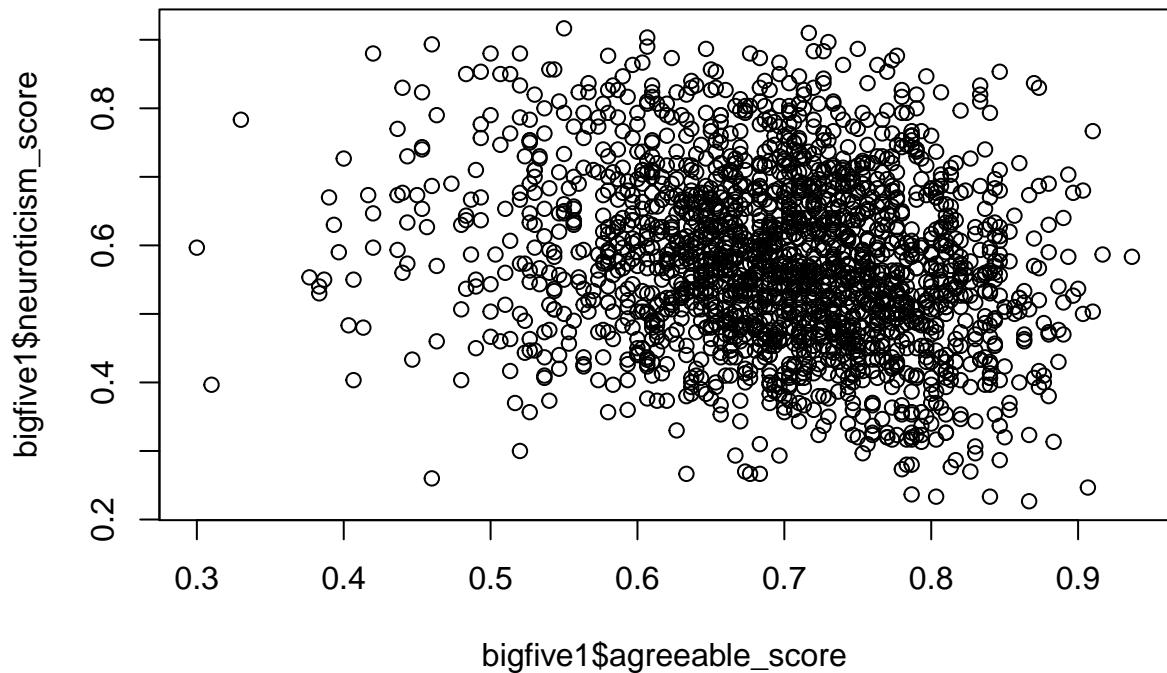
##
##  Welch Two Sample t-test
##
## data:  female_subjects$neuroticism_score and male_subjects$neuroticism_score
## t = 110.65, df = 258116, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.04972757      Inf
## sample estimates:
## mean of x mean of y
## 0.5944653  0.5439874
```

p-vrijednost je manja od 0.05, stoga se hipoteza $H_0 : \mu_{Female} = \mu_{Male}$ odbacuje u korist alternativne hipoteze $H_1 : \mu_{Female} > \mu_{Male}$

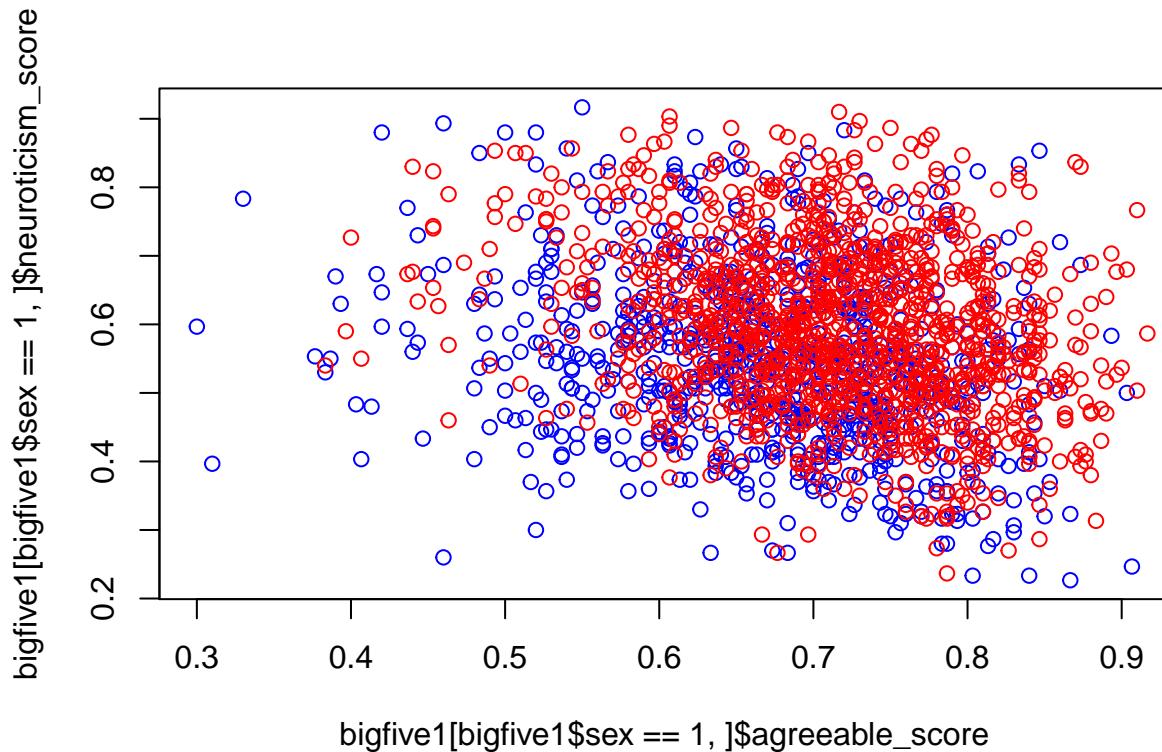
Osim neuroticizma za koji smo statističkim testiranjem utvrdili da se razlikuje između spolova, osobina koja bi se potencijalno također mogla pokazati kao razlikovni faktor između spolova jest ugodnost (<https://www.frontiersin.org/articles/10.3389/fpsyg.2011.00178/full>). Možemo grafički prikazati ovisnost tih dviju varijabli te istaknuti razliku u spolovima.

```
bigfive1 = bigfive[sample(nrow(bigfive), 2000),]

plot(bigfive1$agreeable_score, bigfive1$neuroticism_score )
```



```
plot(bigfive1[bigfive1$sex == 1,]$agreeable_score, bigfive1[bigfive1$sex == 1,]$neuroticism_score, col = "black", pch = 1, cex = 0.5)
points(bigfive1[bigfive1$sex == 2,]$agreeable_score, bigfive1[bigfive1$sex == 2,]$neuroticism_score, col = "red", pch = 1, cex = 0.5)
```



Iz scatter plota možemo uočiti razliku u osobinama između muškog i ženskog spola, odnosno da žene generalno postižu veći score na neuroticizmu i ugodnosti.

Ovo nam daje tračak nade da bismo možda iz velikog skupa podataka mogli kreirati logistički model regresije pomoću kojega ćemo pomoći varijabli neuroticism_score i agreeableness_score pokušati predvidjeti kojeg je spola ispitanik.

Model logističke regresije za klasifikaciju spola ispitanika

```
require(ISLR)
```

```
## Loading required package: ISLR
```

Kako bismo dobili binarnu varijablu za spol prilagodit ćemo varijablu spola tako da 0 predstavlja muški, a 1 ženski spol.

```
data = bigfive
data$sex = data$sex - 1

model <- glm(sex ~ neuroticism_score + agreeable_score, data = data, family = binomial)
summary(model)
```

```
##
## Call:
```

```

## glm(formula = sex ~ neuroticism_score + agreeable_score, family = binomial,
##      data = data)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.4886 -1.1518  0.6908  0.9637  2.5378
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.65978   0.04153 -160.4 <2e-16 ***
## neuroticism_score 4.45536   0.03374  132.1 <2e-16 ***
## agreeable_score   6.54909   0.04569  143.3 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 413025  on 307312  degrees of freedom
## Residual deviance: 377841  on 307310  degrees of freedom
## AIC: 377847
##
## Number of Fisher Scoring iterations: 4

Rsq = 1 - model$deviance/model>null.deviance
Rsq

```

```
## [1] 0.08518719
```

Iz rezultata modela možemo uočiti da su svi koeficijenti statistički signifikantni uz bilo koju razinu značajnosti. Usporedbom devijanci null-modela i našeg modela pomoću R^2 možemo zaključiti da je naš model bolji, ali ne pretjerano dobar.

U procjeni kvalitete modela koristit ćemo i tzv. matricu zabune iz koje ćemo očitati kolika je točnost našeg modela.

```

yHat <- model$fitted.values > 0.5
tab <- table(data$sex, yHat)
tab

##      yHat
##      FALSE  TRUE
## 0 49426 72738
## 1 29118 156031

accuracy = sum(diag(tab)) / sum(tab)
accuracy

```

```
## [1] 0.6685594
```

Vidimo da je udio točno klasificiranih primjera oko 67%.

Možemo pokušati u model uvrstiti i preostale ulazne varijable našeg dataseta kako bismo dobili kvalitetniji model.

```

model1 = glm(sex ~ agreeable_score + extraversion_score + openness_score + conscientiousness_score + neuroticism_score, family = binomial)
summary(model1)

##
## Call:
## glm(formula = sex ~ neuroticism_score + agreeable_score, family = binomial,
##      data = data)
##
## Deviance Residuals:
##    Min      1Q   Median      3Q     Max
## -2.4886 -1.1518  0.6908  0.9637  2.5378
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.65978   0.04153 -160.4 <2e-16 ***
## neuroticism_score 4.45536   0.03374  132.1 <2e-16 ***
## agreeable_score  6.54909   0.04569  143.3 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 413025  on 307312  degrees of freedom
## Residual deviance: 377841  on 307310  degrees of freedom
## AIC: 377847
##
## Number of Fisher Scoring iterations: 4

Rsq = 1 - model1$deviance/model$null.deviance
Rsq

##
## [1] 0.1216226

Možemo uočiti da je svaki score ponovno signifikantan, dok je varijabla age uz razinu značajnosti od 5% statistički nesignifikantna. Vidimo da se  $R^2$  povećao u odnosu na prvotni model što je pozitivno, a daljnjom analizom utvrdit ćemo je li kvalitetniji.

yHat <- model$fitted.values > 0.5
tab <- table(data$sex, yHat)
tab

##
##      yHat
##      FALSE   TRUE
## 0 49426 72738
## 1 29118 156031

accuracy = sum(diag(tab)) / sum(tab)
accuracy

##
## [1] 0.6685594

```

Model daje točnost od oko 69% što je svakako bolje od slučajnog pogađanja spola, ali i povećanje u odnosu na prvotni model pa naslućujemo da je ovaj model kvalitetniji. To ćemo provjeriti tako što ćemo testirati devijance ova dva modela.

```
anova(model, model1, test = "LRT")

## Analysis of Deviance Table
##
## Model 1: sex ~ neuroticism_score + agreeable_score
## Model 2: sex ~ agreeable_score + extraversion_score + openness_score +
##            conscientiousness_score + neuroticism_score + age
##      Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1     307310    377841
## 2     307306    362792  4     15049 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Likelihood ratio test (LRT) pokazao je da postoji značajna razlika u devijancama te možemo zaključiti kako uz dodatne preostale faktore, naš model ima veću kvalitetu. Međutim, ono što ne možemo je tvrditi da sa velikom vjerojatnosti na temelju samo rezultata faktora možemo predvidjeti spol.