

Antropometrijski podatci američkih vojnih snaga

Dario Deković, Emil Huzjak, Mijo Piškur, Janko Vidaković

01/17/2021

Antropometrijski podatci američkih vojnih snaga

Uvod

Podaci kojima se bavimo u ovom projektu dolaze iz dobro poznatog skupa ANSUR II koji sačinjavaju 93 antropometrijske mjere prikupljene na pripadnicima američke vojske. Prikupljeno je 6068 podataka od kojih je 4082 pripadnika muške populacije te 1986 pripadnika ženske populacije. Na danom skupu podataka smo proveli našu analizu podataka koja je obuhvaćala testiranje hipoteza i izgradnju modela za predikciju različitih varijabli. Projekt je podjeljen na odlomke, a svaki odlomak predstavlja jedan donekle zasebni element ovog projekta. Na početku svakog odlomka se nalazi motivacija iza tog dijela istraživanja. Nadalje, kako je skup podataka relativno velik eksploratorna analiza podataka se nalazi u svakome odlomku za dio skupa podataka koji je relevantan za taj dio istraživanja. Bitno je napomenuti kako bi se izbjeglo prepisivanje relevantnih podataka u svakom odlomku se iznova učitava csv datoteka s podacima.

Utjecaj spola na antropometrijske mjere

Motivacija

Kada smo počeli raditi na projektu smo uočili kako dosta antropometrijskih mjera ne slijedi normalnu distribuciju. Daljne istraživanje nam je ukazalo kako je jedan od važnijih razloga upravo varijacija među spolovima. S obzirom koliko antropometrijskih razlika naizgled postoji između muškaraca i žena zaključili smo kako analiza tih razlika zasluguje svoj zasebni odlomak u projektu te smo je stavili na početak projekta s obzirom da dosta daljnih analiza i modela koristi upravo postojanje te razlike.

Istraživanje

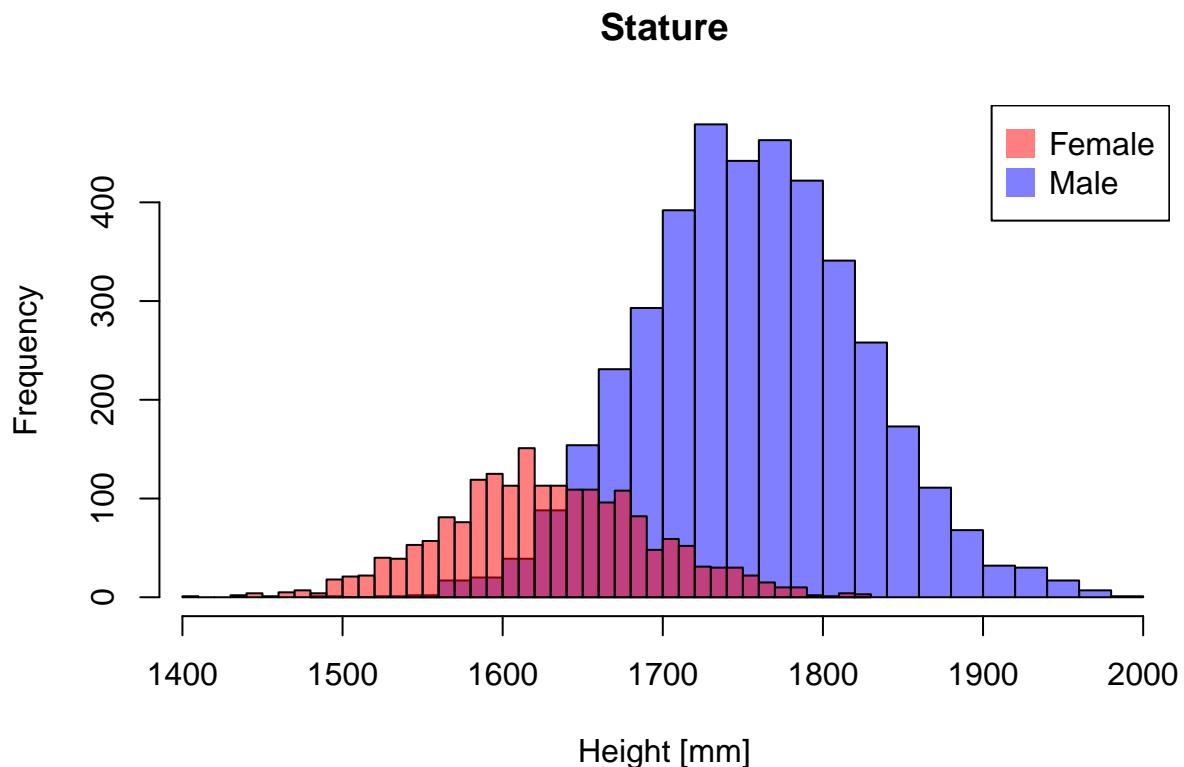
Učitajmo skup podataka i razdvojimo ih po spolu.

```
ansur.II.data = read.csv("ANSUR_II_data.csv") #učitavanje podataka
males <- ansur.II.data[ansur.II.data$Gender=="Male",] #muške vojne snage
females <- ansur.II.data[ansur.II.data$Gender=="Female",] #ženske vojne snage
```

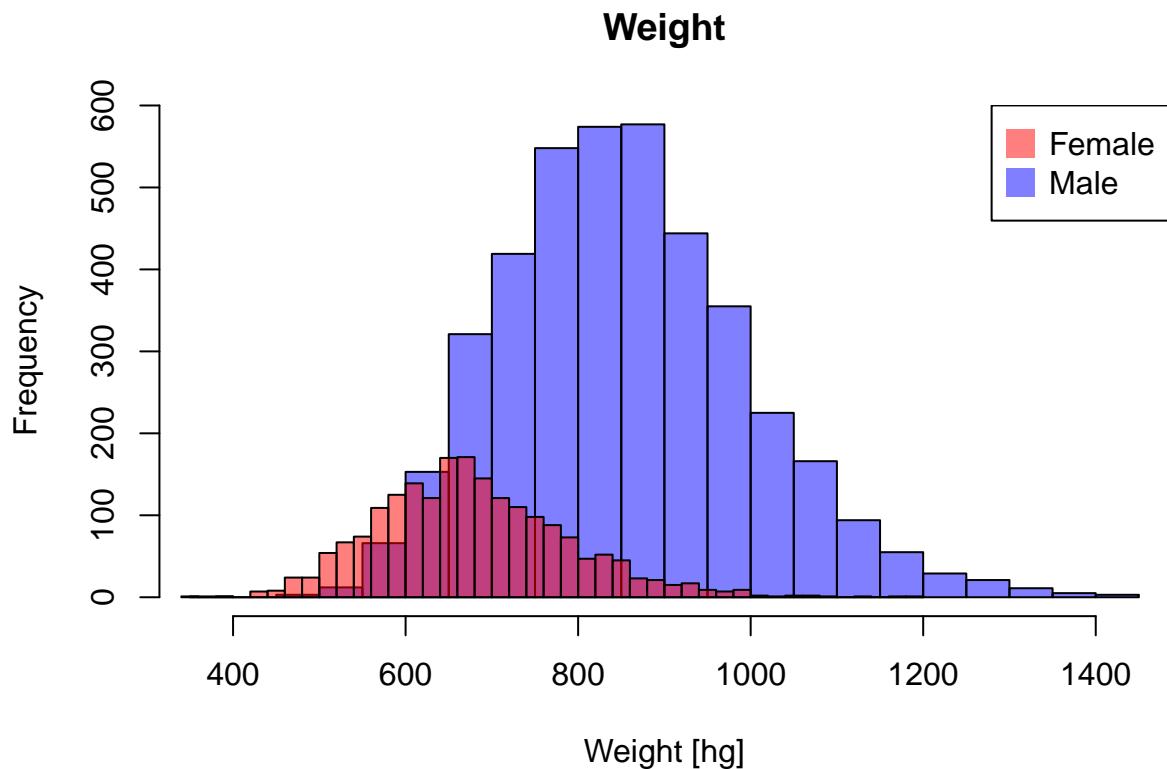
Prikažimo histogramima najznačajnije antropomorfske mjere.

```
#histogrami varijable s obzirom na spol
plot_by_gender <- function(column, main = column, xlab = column) {
  summary(females[[column]])
  summary(males[[column]])
  hist(males[[column]], breaks=30, main=main, xlab=xlab, ylab="Frequency", col=rgb(0,0,1,0.5), xlim = c(0,100))
  hist(females[[column]], breaks=30, main=main, xlab=xlab, ylab="Frequency", col=rgb(1,0,0,0.5), xlim = c(0,100))
  legend(x="topright", c("Female", "Male"), col=c(rgb(1,0,0,0.5), rgb(0,0,1,0.5)), pt.cex = 2, pch = 15)
}

plot_by_gender("stature", "Stature", "Height [mm]")
```

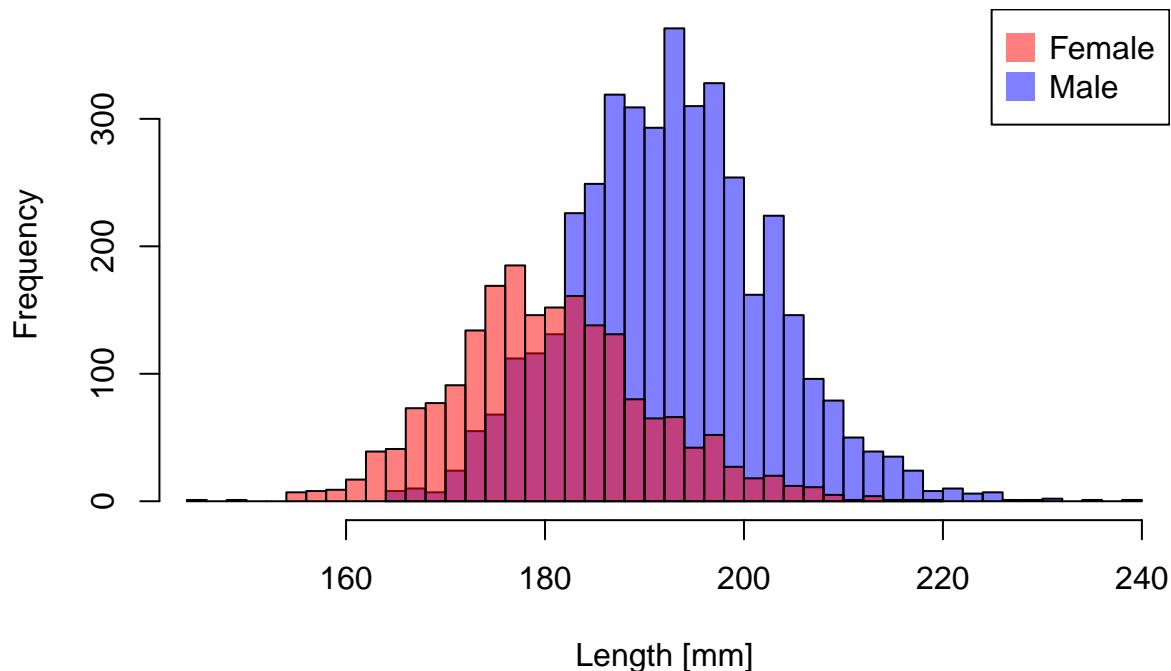


```
plot_by_gender("weightkg", "Weight", "Weight [hg]")
```



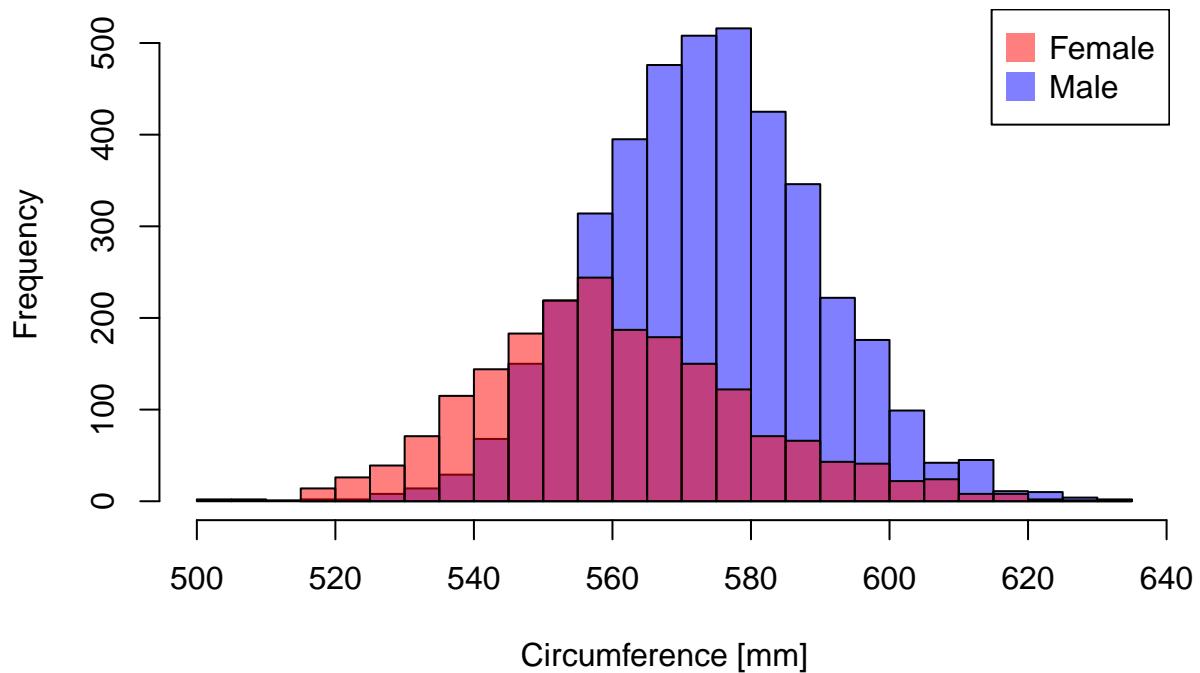
```
plot_by_gender("handlength", "Hand length", "Length [mm]")
```

Hand length



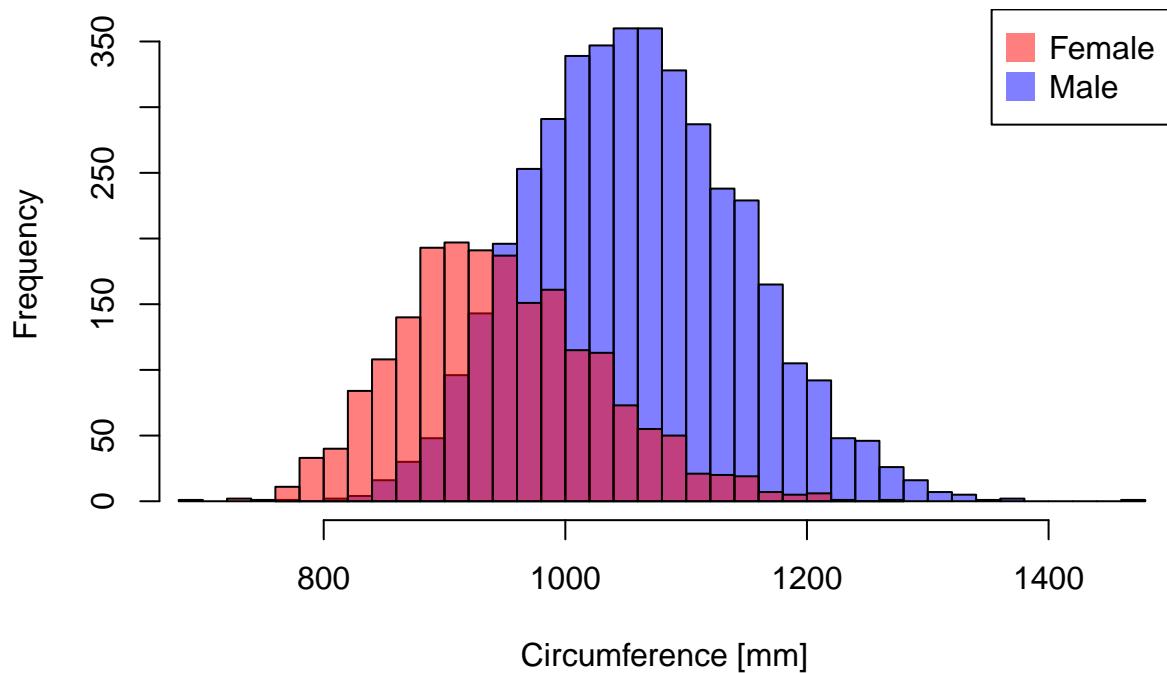
```
plot_by_gender("headcircumference", "Head circumference", "Circumference [mm]")
```

Head circumference



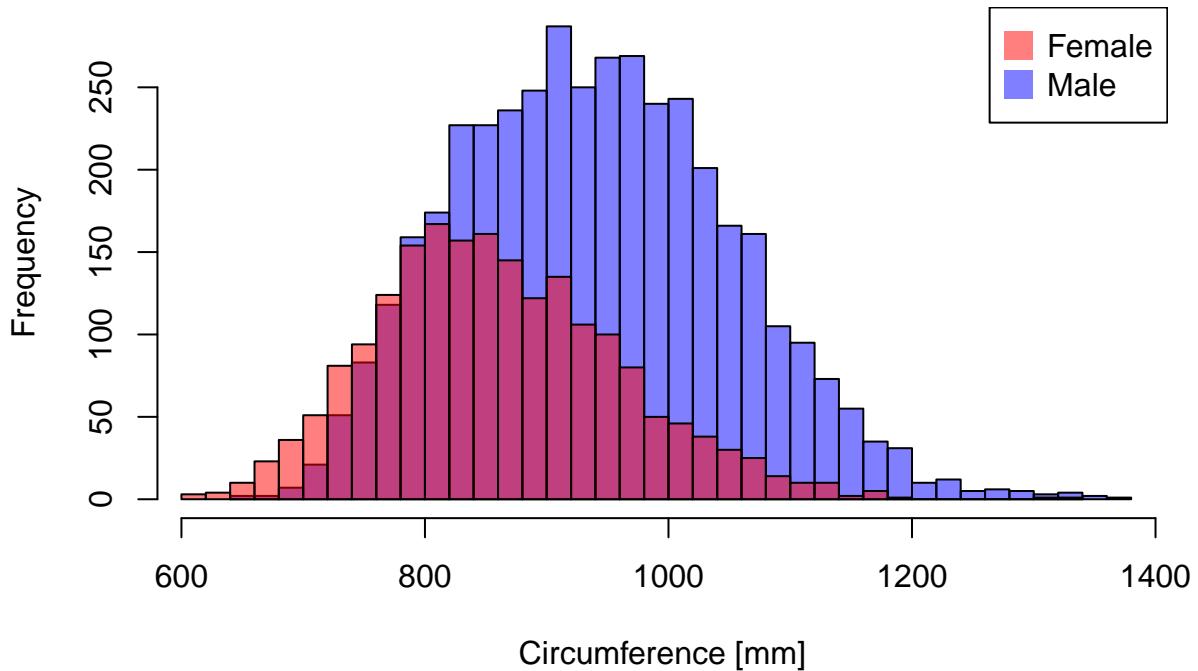
```
plot_by_gender("chestcircumference", "Chest circumference", "Circumference [mm]")
```

Chest circumference



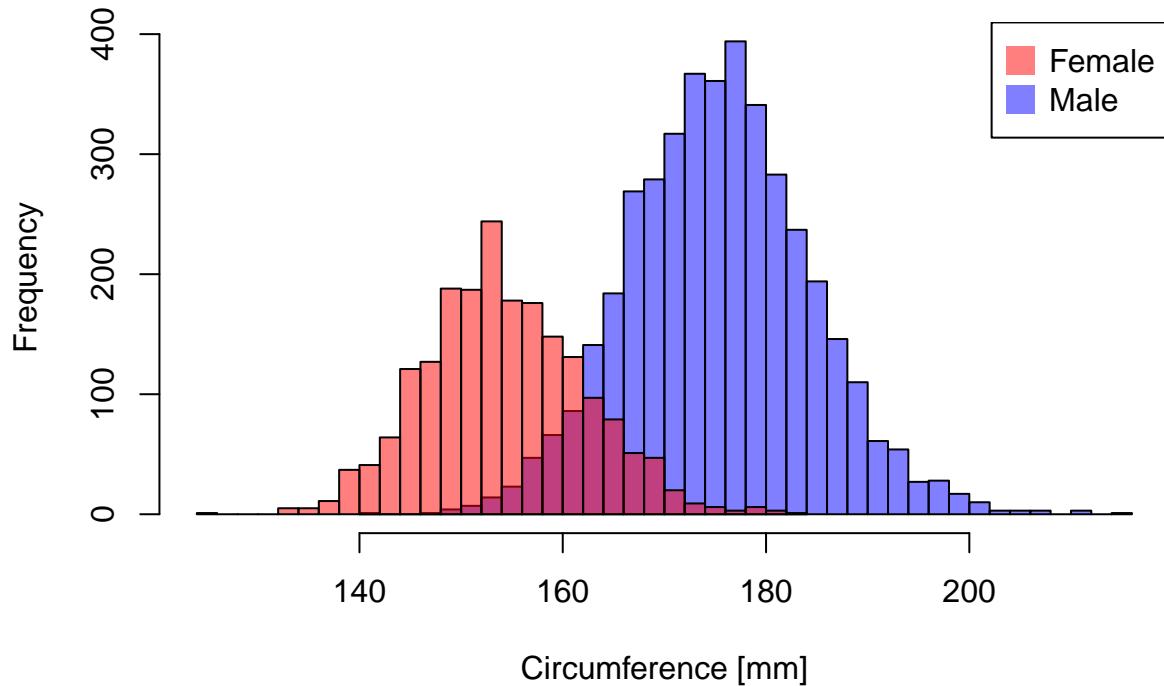
```
plot_by_gender("waistcircumference", "Waist circumference", "Circumference [mm]")
```

Waist circumference



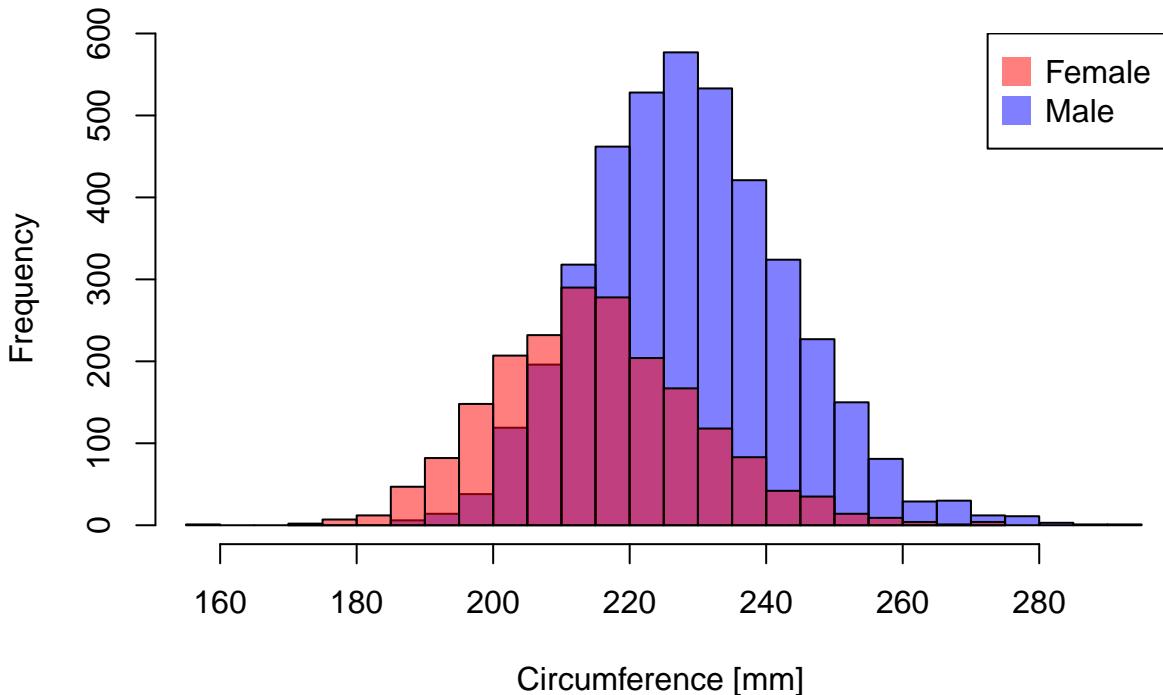
```
plot_by_gender("wrstcircumference", "Wrist circumference", "Circumference [mm]")
```

Wrist circumference



```
plot_by_gender("anklecircumference", "Ankle circumference", "Circumference [mm]")
```

Ankle circumference



Na svim histogramima jasno se vide manje ili veće razlike u antropomorfijskim mjerama s obzirom na spol. Neke varijable, kao opseg zapešća, vizualno pokazuju puno značajniju razliku, dok se na primjer opseg gležnja puno manje razlikuje. Treba imati na umu i činjenicu da je podataka o muškim vojnicima otprilike dvostruko više nego podataka o ženskim vojnicima, što znači da uzorak muških vojnika bolje opisuje populaciju muških vojnih snaga nego što to čini ženski uzorak. Bilo bi zanimljivo pokušati predvidjeti spol vojnika pomoću višestruke logističke regresije, no umjesto toga, promotrimo detaljnije razlike u mjerama centralne tendencije i rasipanja između ova dva uzorka. Pronađimo varijablu za koju se očekivanja razlikuju najviše, tj. najmanje.

```
#računanje relativne razlike očekivanja između spolova s obzirom na rang varijable
get_mean_diff <- function(col) {
  return(abs(mean(females[[col]]) - mean(males[[col]])) / (max(ansur.II.data[col]) - min(ansur.II.data[col])))
}

#nalaženje ekstremne vrijednosti (minimuma ili maksimuma) razlike očekivanja
find_extreme_diff <- function(relation, data, get_variable_diff) {
  extreme_diff <- 0
  column <- ""
  sapply(1:ncol(data), function(i) {
    if (i == 1) {
      column <- colnames(data)[i]
      extreme_diff <- get_variable_diff(colnames(data)[i])
    } else if (relation == "MAX") {
      curr_extreme_diff <- get_variable_diff(colnames(data)[i])
      if (curr_extreme_diff > extreme_diff) {
        column <- colnames(data)[i]
        extreme_diff <- curr_extreme_diff
      } else {}
    }
  })
  return(list(column = column, diff = extreme_diff))
}
```

```

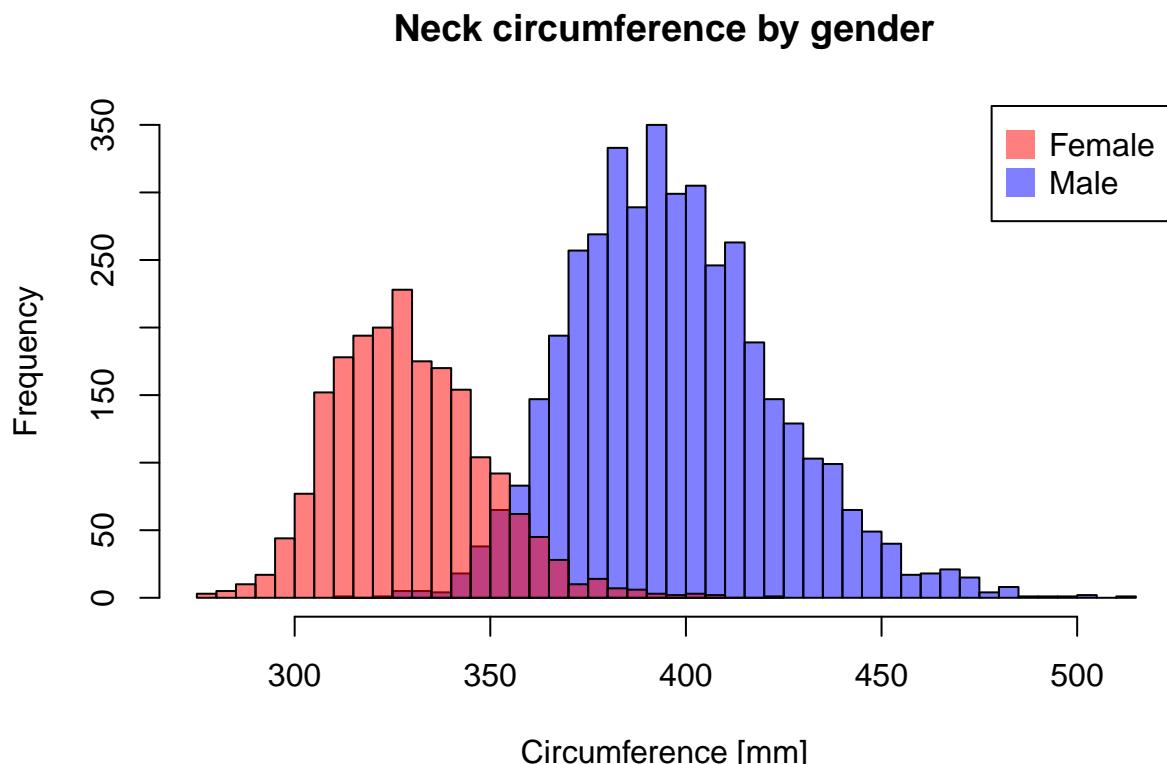
} else if (relation == "MIN") {
  curr_extreme_diff <- get_variable_diff(colnames(data)[i])
  if (curr_extreme_diff < extreme_diff) {
    column <- colnames(data)[i]
    extreme_diff <- curr_extreme_diff
  } else {}
} else {}
})
return(c(column, extreme_diff))
}

ansur.II.data.numeric <- select(ansur.II.data, -c("Gender", "Date", "Installation", "Component", "Branch"))
findExtremeDiff("MAX", ansur.II.data.numeric, get_mean_diff) #najveća razlika
## [1] "neckcircumference" "0.283905345656595"
findExtremeDiff("MIN", ansur.II.data.numeric, get_mean_diff) #najmanja razlika
## [1] "buttockcircumference" "0.00286236893488413"

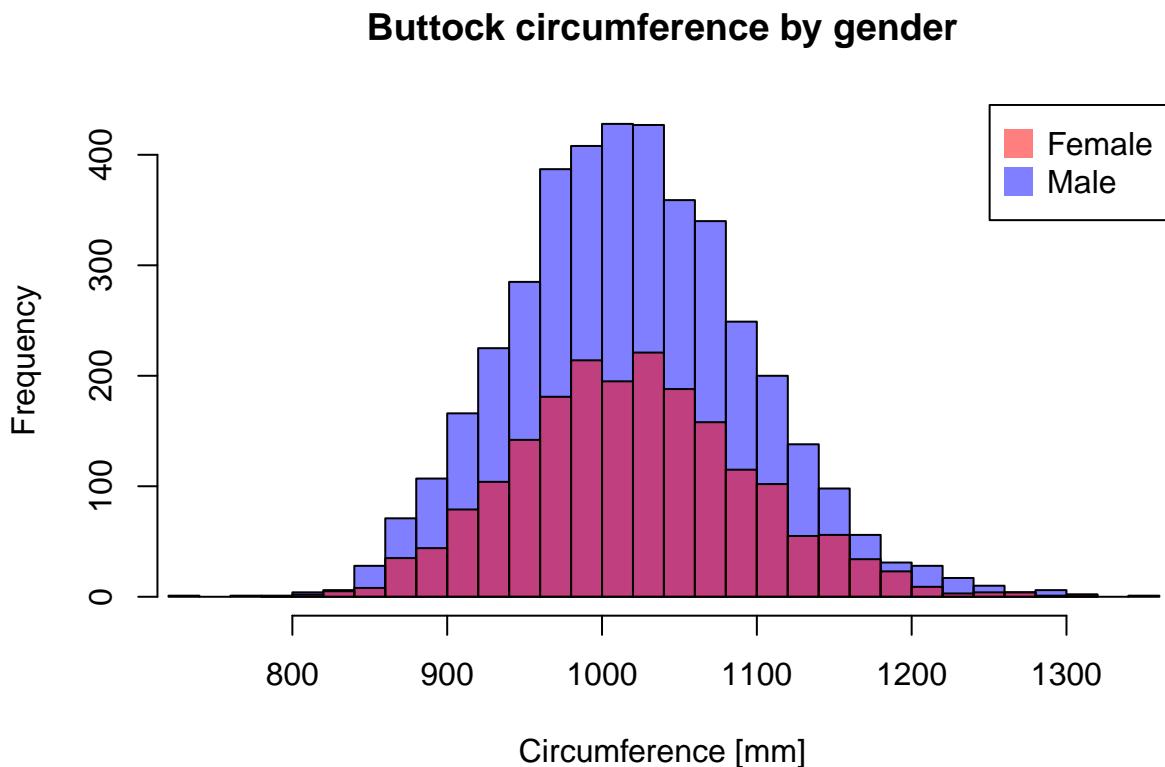
```

Vidimo da je opseg vrata mjera koja se najviše razlikuje s obzirom na spol, s razlikom očekivanja od skoro pa 30%. S druge strane, očekivani opseg zadnjice je skoro jednak u oba uzorka, s razlikom očekivanja od samo 0.2%. Vizualizirajmo razdiobe ovih varijabli

```
plot_by_gender("neckcircumference", "Neck circumference by gender", "Circumference [mm]")
```



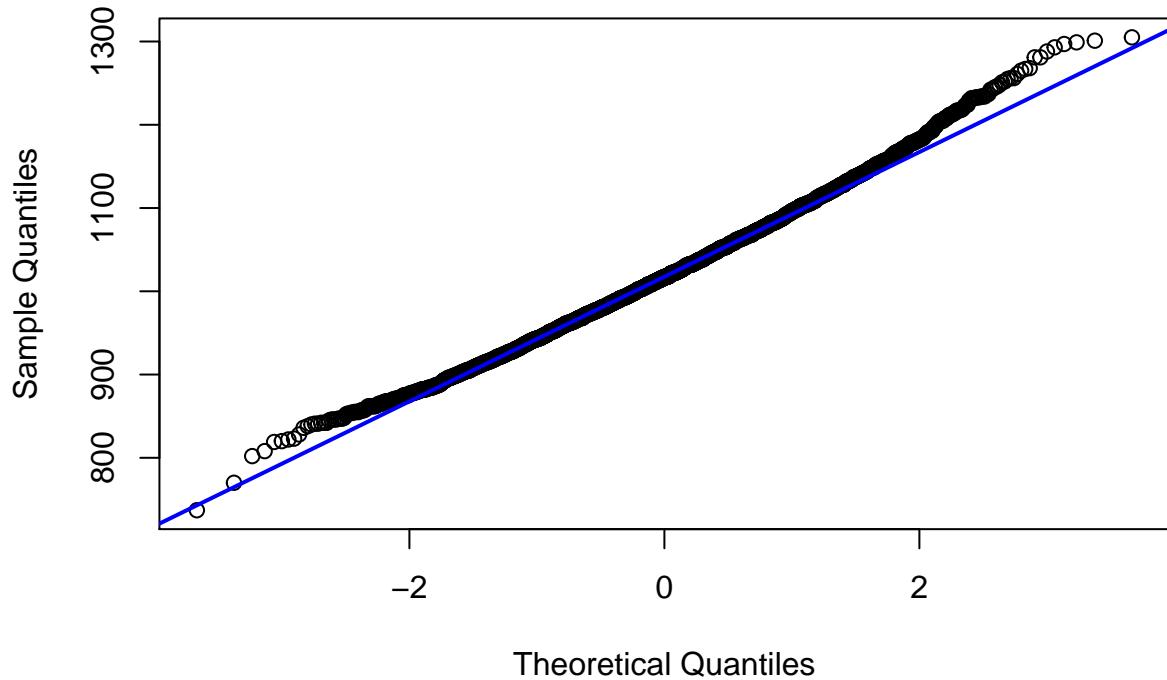
```
plot_by_gender("buttockcircumference", "Buttock circumference by gender", "Circumference [mm]")
```



Histogrami potvrđuju značajne sličnosti, odnosno razlike. Testirajmo hipotezu o jednakosti očekivanja opsega zadnjice između muškaraca i žena. Prije svega, krenimo od provjere ravnaju li se uzorci po normalnoj distribuciji. Histogrami izgledaju poprilično normalno distribuirano, no provjerimo dodatno qq plotovima i Lillieforsovim testom o normalnosti distribucije.

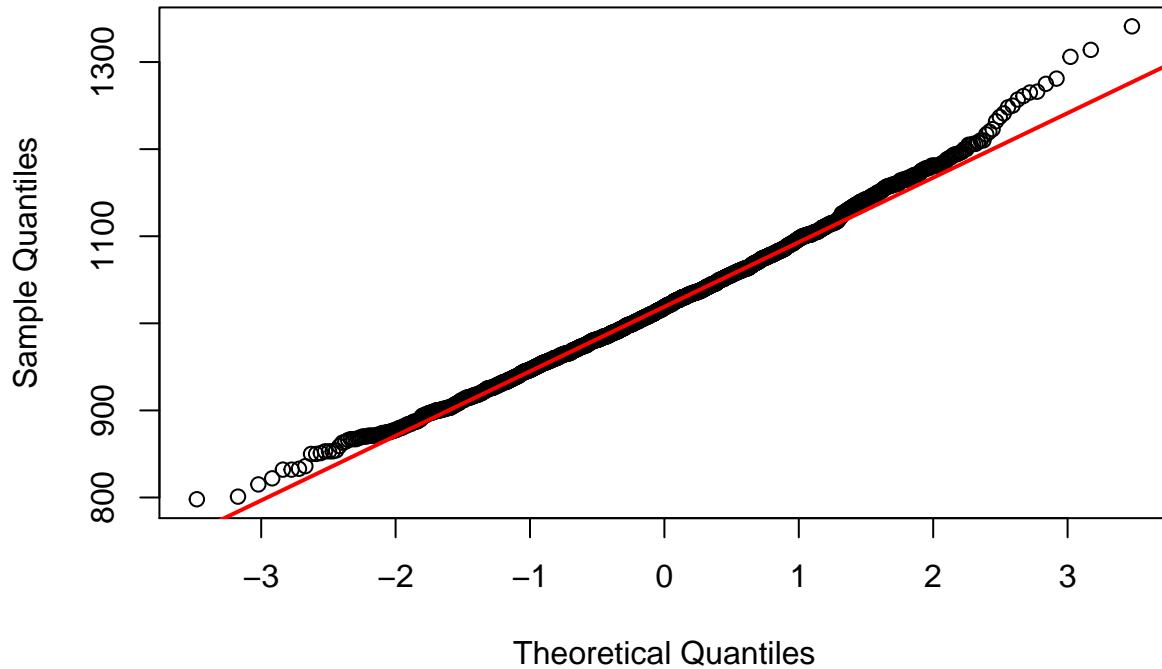
```
test.males <- males$buttockcircumference  
test.females <- females$buttockcircumference  
  
qqnorm(test.males, main="Male buttock circumference")  
qqline(test.males, col="blue", lwd=2)
```

Male buttock circumference



```
qqnorm(test.females, main="Female buttock circumference")
qqline(test.females, col="red", lwd=2)
```

Female buttock circumference



```
require(nortest)

## Loading required package: nortest
lillie.test(test.males)

##
## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: test.males
## D = 0.021627, p-value = 0.0001599
lillie.test(test.females)

##
## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: test.females
## D = 0.023529, p-value = 0.01251
```

qq plotovi pokazuju da se uzorci velikom većinom ravnaju po normalnoj razdiobi, uz malo prevelike repove, kako je uočeno i u histogramima. Lilliefors test pokazuje da distribucija nije normalna sa statističkom značajnošću, no nije daleko od toga i to je prihvatljivo, pogotovo zato što t-test, kojim ćemo testirati hipotezu o očekivanjima, dobro podnosi normalnost.

Druga prepostavka koju moramo provjeriti prije testiranja jednakosti očekivanja je nezavisnost uzorka. Kako su uzorci razdvojeni po spolu, smisleno je prepostaviti da su opservacije nezavisne.

Zadnja stvar koja nam treba prije provođenja testa su informacije o varijancama. Kako varijance populacija nisu poznate, moraju se procijeniti iz uzorka.

```

sd(test.males)

## [1] 76.68107

sd(test.females)

## [1] 75.89471

```

Dobivene vrijednosti sugeriraju da bi varijance, iako nepoznate, mogле biti jednake. Ovu hipotezu možemo provjeriti dvostranim F testom.

$$H_0 : \sigma_m^2 = \sigma_f^2$$

$$H_1 : \sigma_m^2 \neq \sigma_f^2$$

```
var.test(test.males, test.females, alternative = "two.sided")
```

```

##
##  F test to compare two variances
##
## data:  test.males and test.females
## F = 1.0208, num df = 4081, denom df = 1985, p-value = 0.5978
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.9457708 1.1007353
## sample estimates:
## ratio of variances
##                  1.02083

```

Iz rezultata testa, koji je proveden s razinom značajnosti $\alpha = 0.05$, ne možemo reći da postoji statistički značajna razlika u varijancama. Drugim riječima, ne možemo odbaciti nultu hipotezu te nastavljamo s pretpostavkom da su varijance uzoraka jednake.

Koristimo t-test.

$$H_0 : \mu_m = \mu_f$$

$$H_1 : \mu_m \neq \mu_f$$

```
t.test(test.males, test.females, alternative = "two.sided", var.equal = TRUE)
```

```

##
##  Two Sample t-test
##
## data:  test.males and test.females
## t = -0.82686, df = 6066, p-value = 0.4083
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -5.827747  2.370005
## sample estimates:
## mean of x mean of y
## 1019.519 1021.248

```

Na temelju t-testa i visoke p-vrijednosti ne možemo sa statističkom značajnošću odbaciti H_0 . Drugim riječima, ne možemo opovrgnuti pretpostavku da je očekivani opseg zadnjice jednak za muškarce i žene, te dolazimo do zanimljivog i ne nužno očekivanog zaključka da muškarci i žene imaju prosječno jednak opseg zadnjice!

Pogledajmo sad više varijabli s maksimalnim tj. minimalnim razlikama u očekivanjima. Pronađimo 5 varijabli koje se najviše razlikuju, te 5 varijabli koje se najmanje razlikuju.

```

find_n_ex_diff <- function(n = 1, get_variable_diff) {
  ansur.II.data.numeric <- select(ansur.II.data, -c("Gender", "Date", "Installation", "Component", "Brain"))
  mins <- vector("list", length=n)
  maxs <- vector("list", length=n)

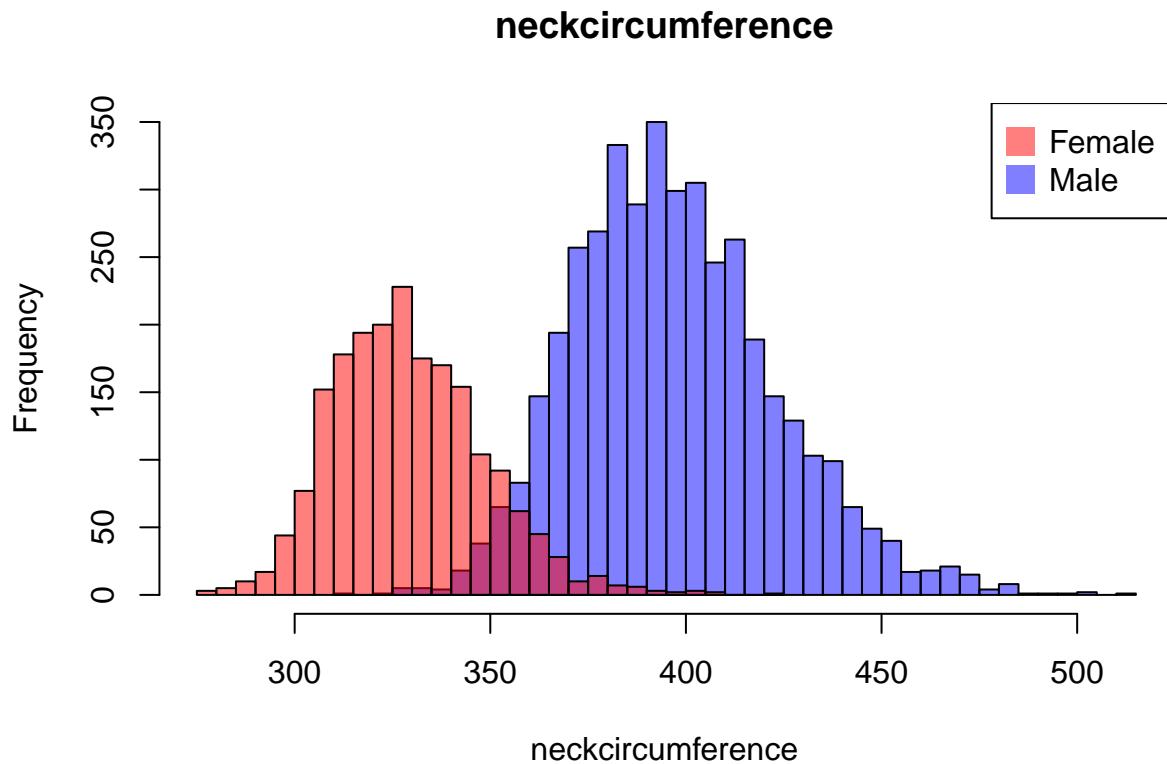
  sapply(1:n, function(i) {
    maxs[[i]] <- find_extreme_diff("MAX", ansur.II.data.numeric, get_variable_diff)
    ansur.II.data.numeric <- select(ansur.II.data.numeric, -c(maxs[[i]][1]))
    maxcol <- maxs[[i]][1]
    plot_by_gender(maxcol)
  })

  sapply(1:n, function(i) {
    mins[[i]] <- find_extreme_diff("MIN", ansur.II.data.numeric, get_variable_diff)
    ansur.II.data.numeric <- select(ansur.II.data.numeric, -c(mins[[i]][1]))
    mincol <- mins[[i]][1]
    plot_by_gender(mincol)
  })
}

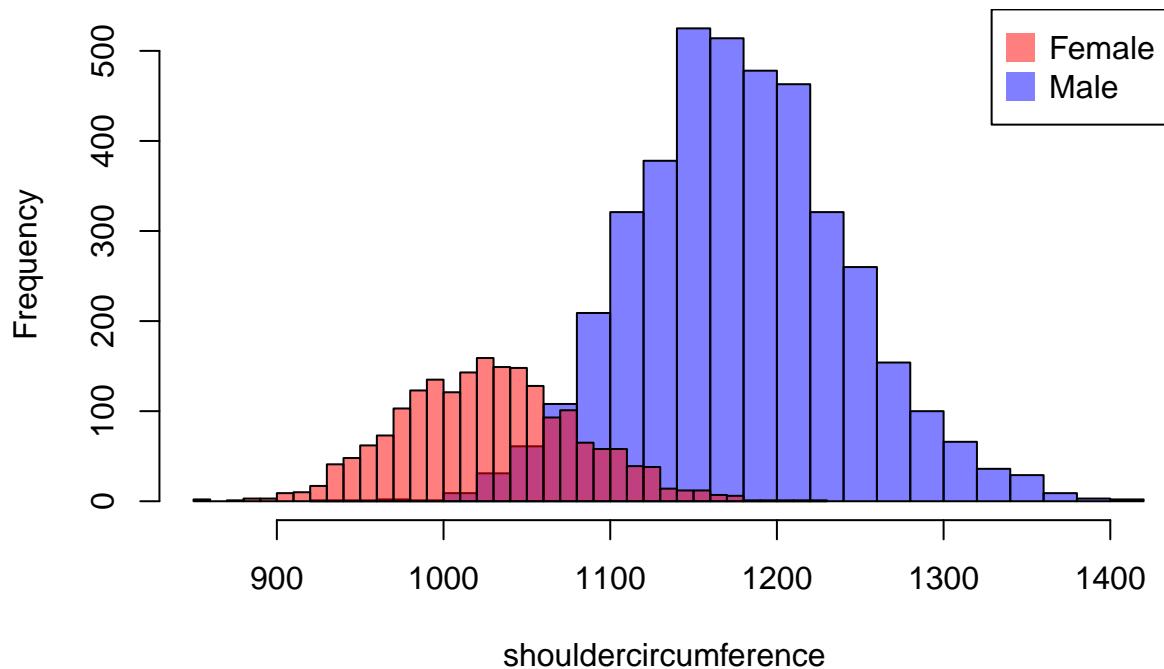
return(list("mins" = mins, "maxs" = maxs))
}

res <- find_n_ex_diff(5, get_mean_diff)

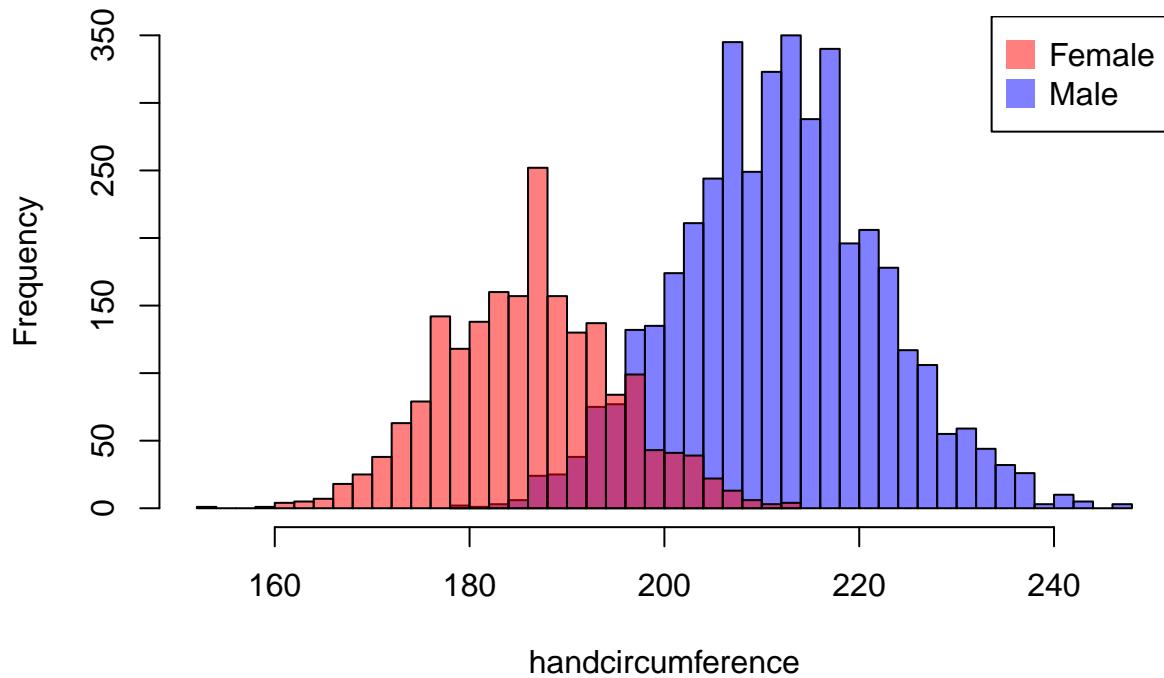
```



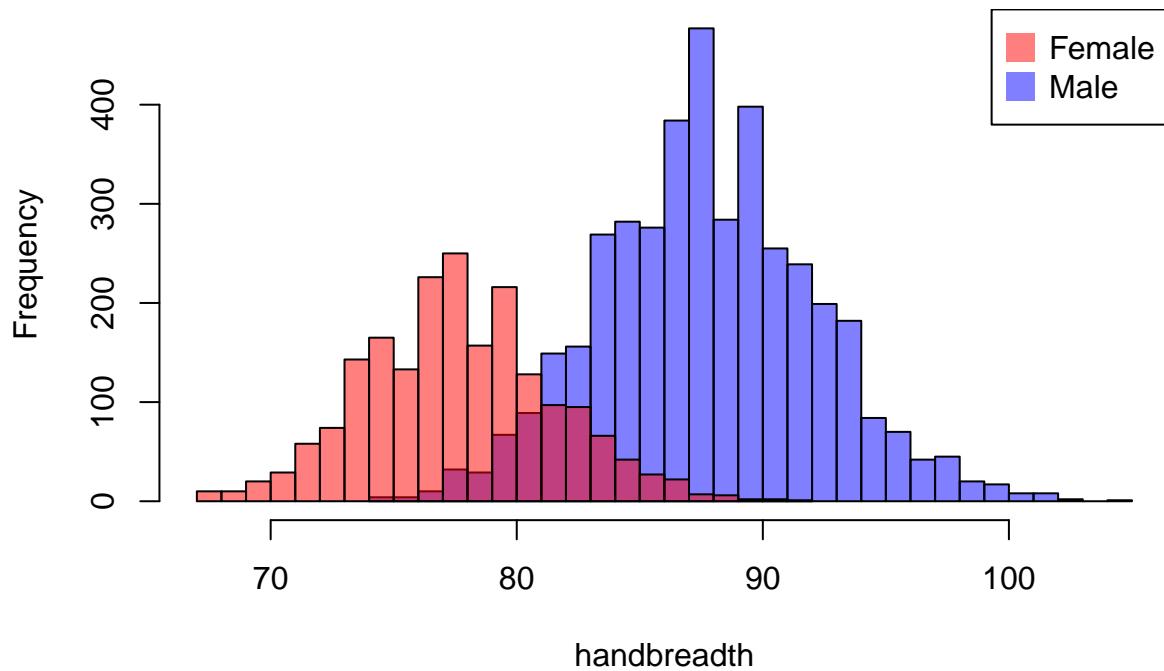
shouldercircumference



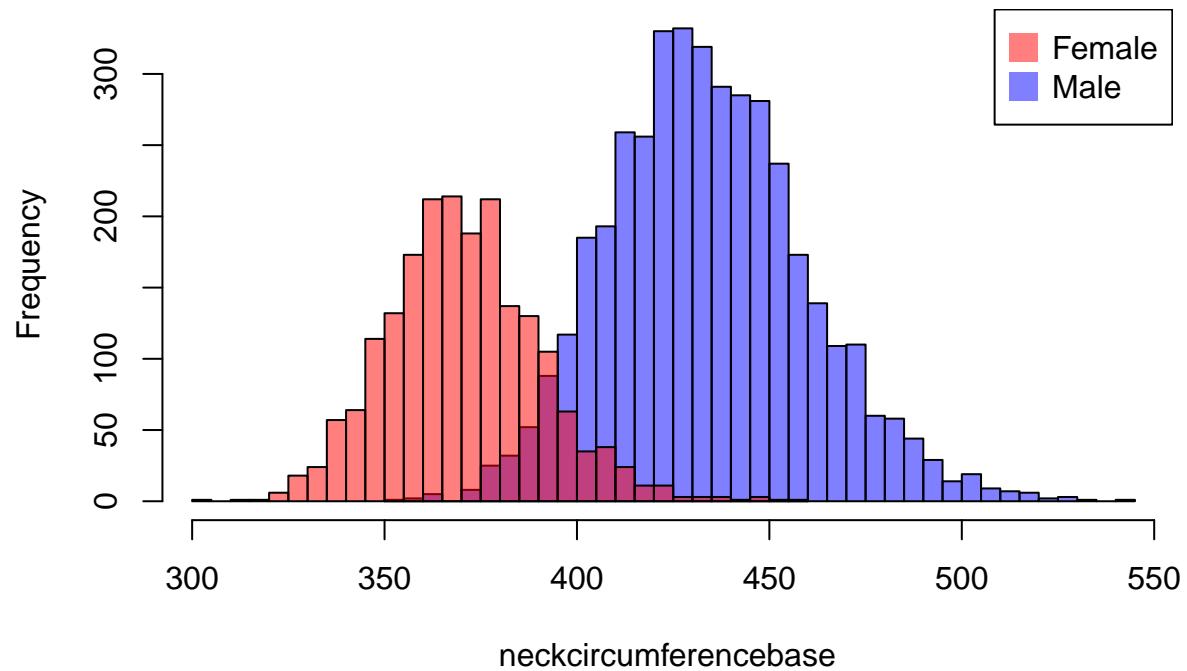
handcircumference



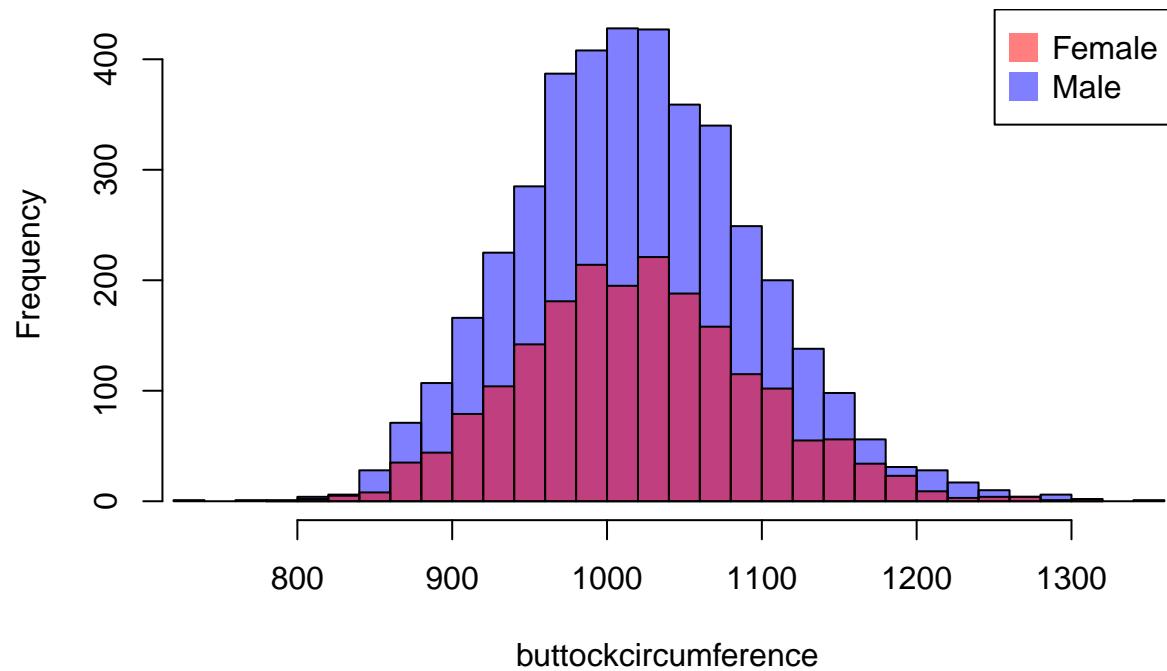
handbreadth



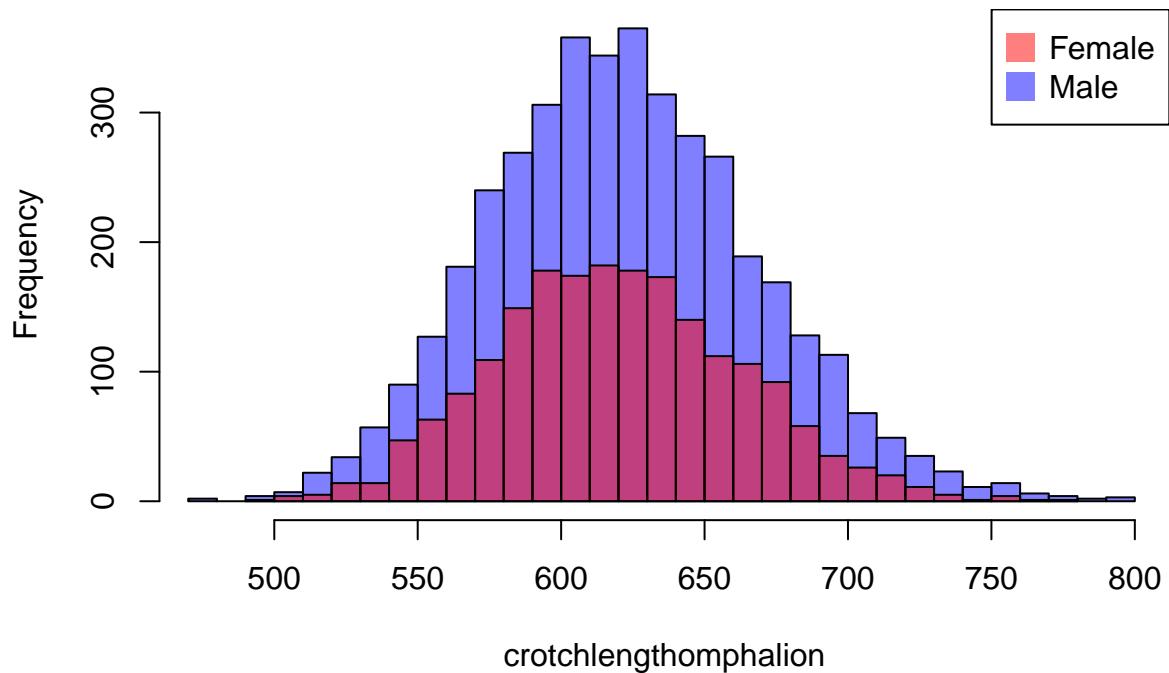
neckcircumferencebase



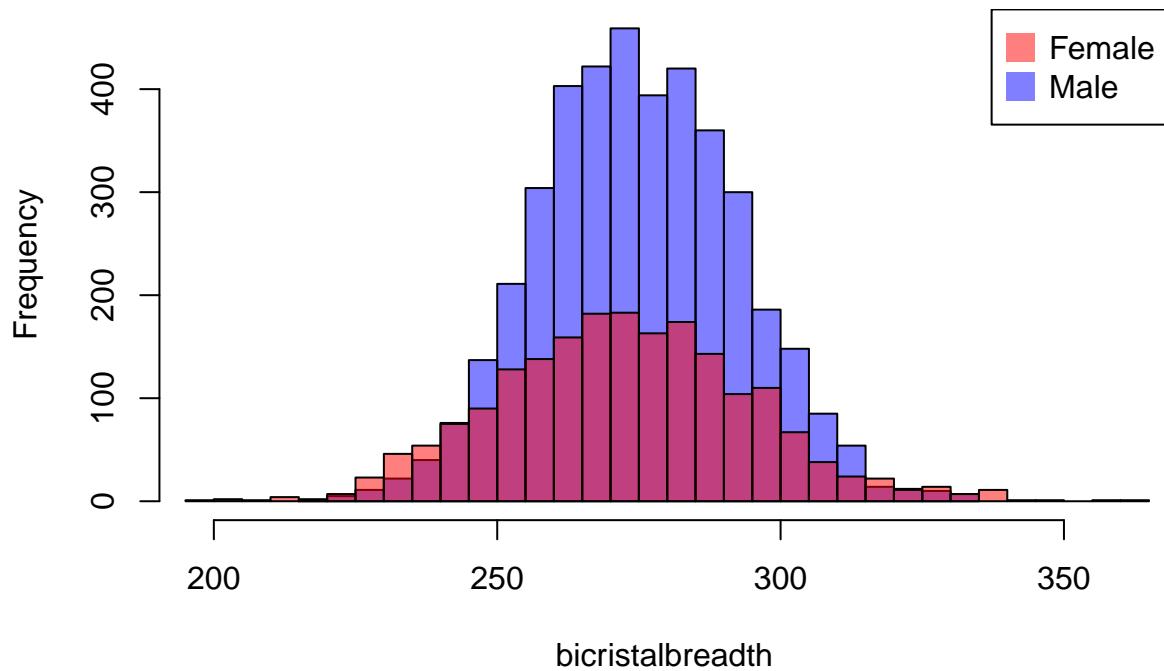
buttockcircumference



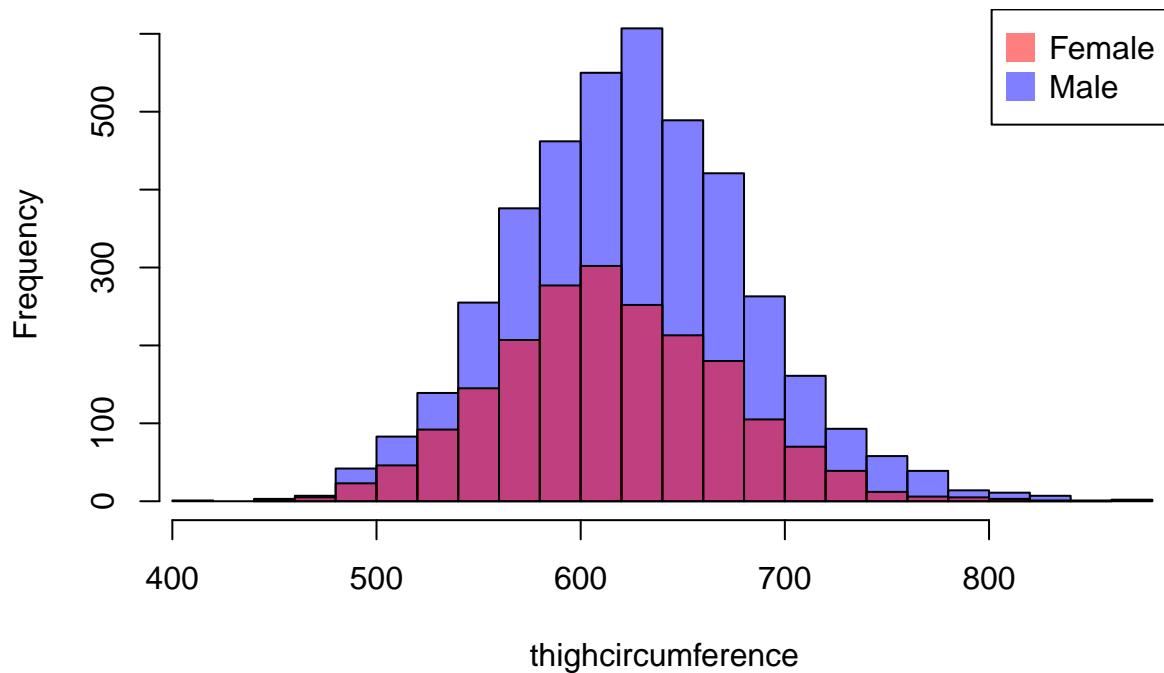
crotchlengthomphalion



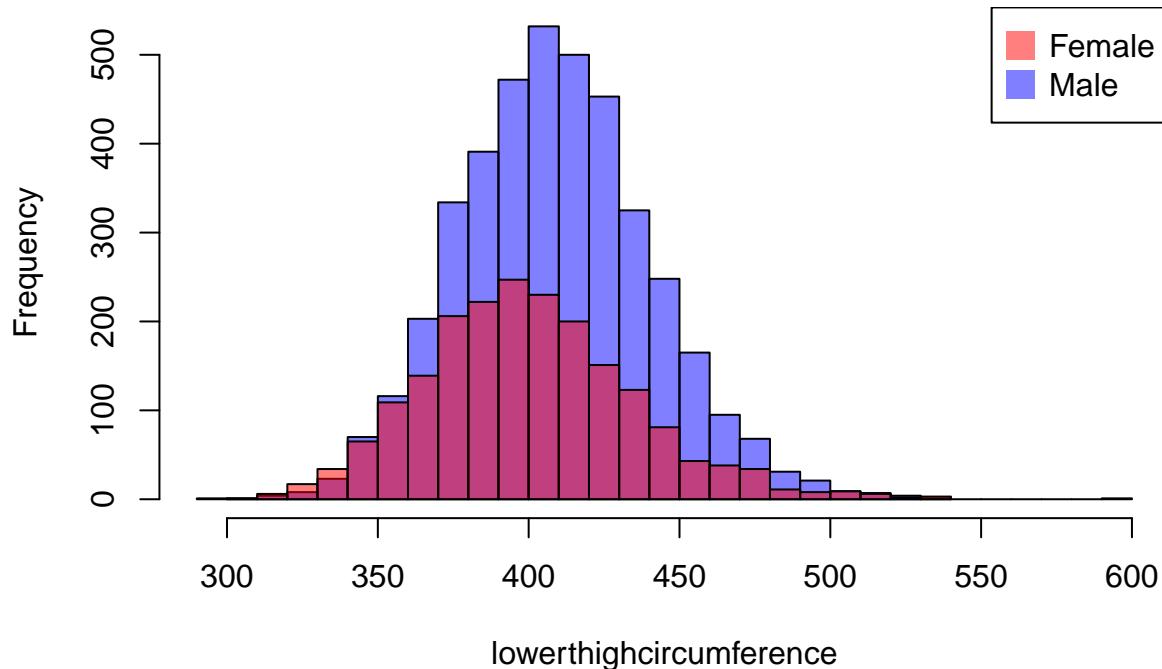
bicristalbreadth



thighcircumference



lowerthighcircumference



Zaključak

Iz dobivenih histograma može se zaključiti da se, s antropometričkog aspekta, muškarci i žene najviše razlikuju u području gornjeg dijela trupa, konkretno oko vrata, ramena i ruku, dok se najmanje razlikuju u području donjeg dijela trupa i gornjeg dijela nogu.

Model za procjenu kilaže vojnika iz dostupnih podataka

Motivacija

O tjelesnoj težini ovise brojni postupci u medicini. Jedan od najvažnijih postupaka u modernoj medicini je anestezija, a doza anestetika ovisi primarno o težini osobe koja ju prima. Naravno, u nekim situacijama vaga nije dostupna, pogotovo kada pričamo o situacijama u kojima se najčešće nalazi vojska. Zato smatramo kako bi korist modela koji uspješno predviđa težinu osobe na temelju podataka koji se mogu dobiti najobičnjim metrom velika.

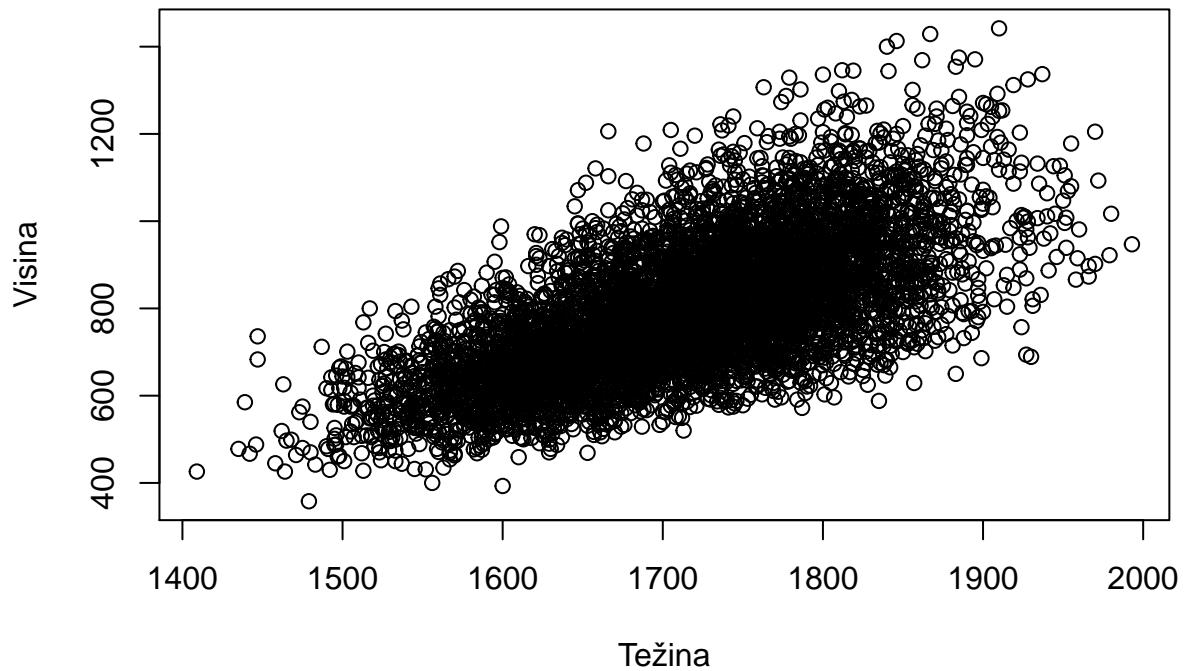
Istraživanje

```
antrData = read.csv("ANSUR_II_data.csv")
```

Prvi korak će nam biti izbor i vizualizacija izabranih parametara. Inicijalni izbor parametara je moja slobodna procjena koji parametri bi mogli imati značajan utjecaj na kilažu vojnika uzimajući u obzir dostupnost tih podataka jednostavnim krojačkim metrom i ispitivanjem. Naravno konačni izbor parametara će ovisiti o performansama modela.

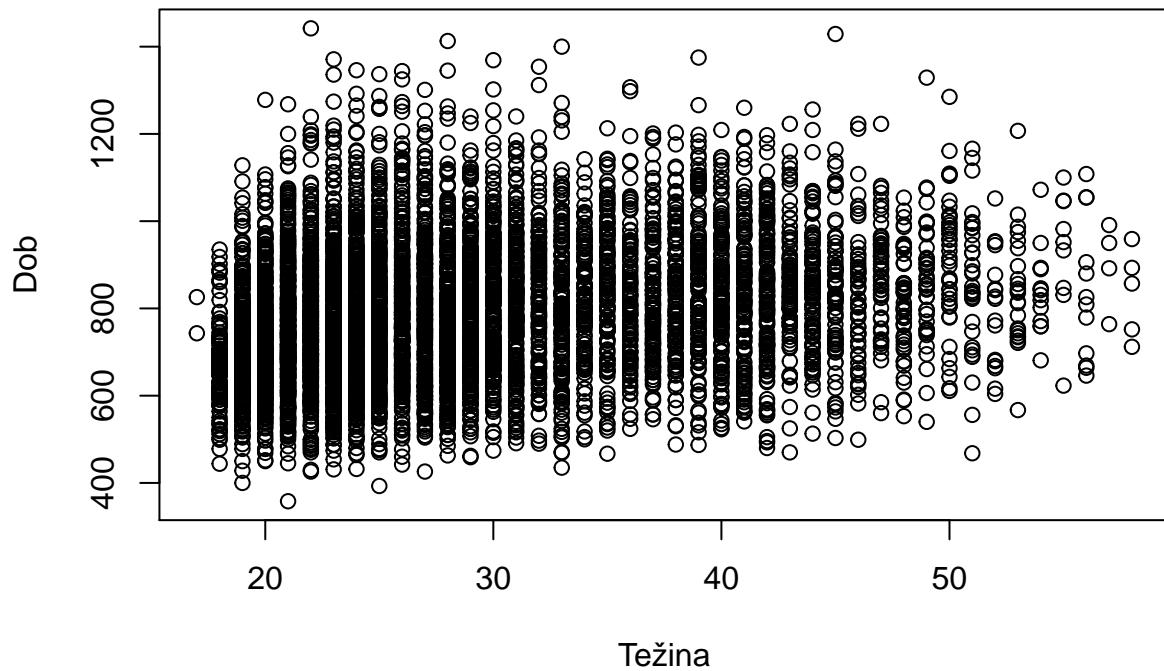
```
plot(antrData$stature, antrData$weightkg, xlab = "Težina", ylab = "Visina",
     main = "Distribucija težina u ovisnosti o visini")
```

Distribucija težina u ovisnosti o visini



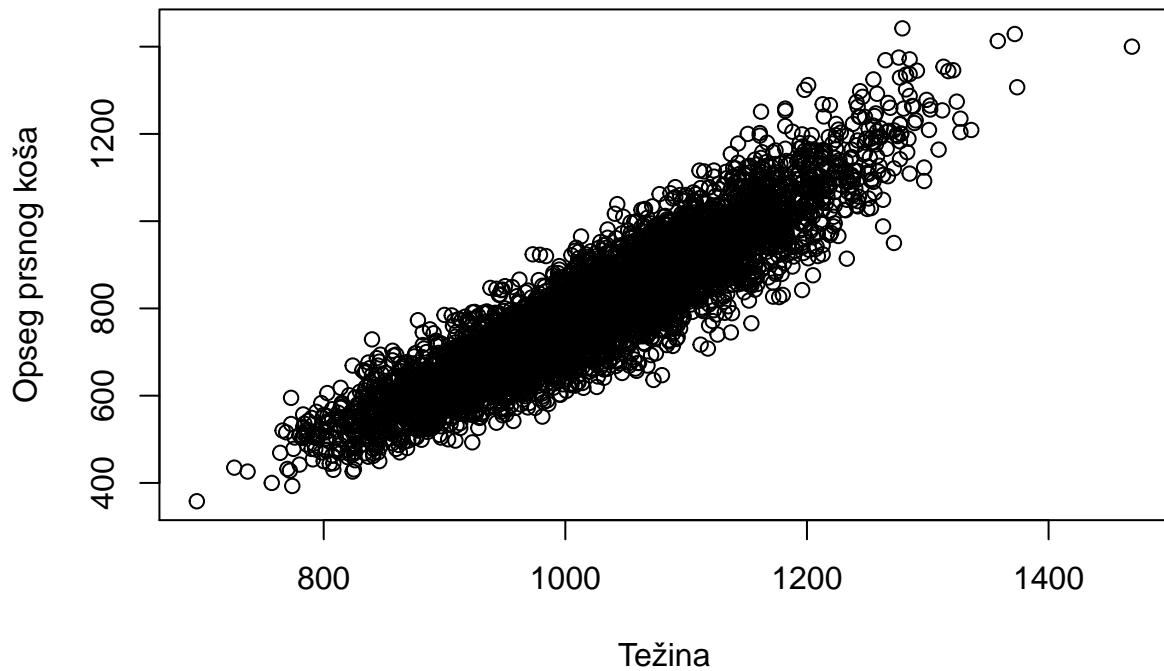
```
plot(antrData$Age, antrData$weightkg, xlab = "Težina", ylab = "Dob",
     main = "Distribucija težina u ovisnosti o dobi")
```

Distribucija težina u ovisnosti o dobi



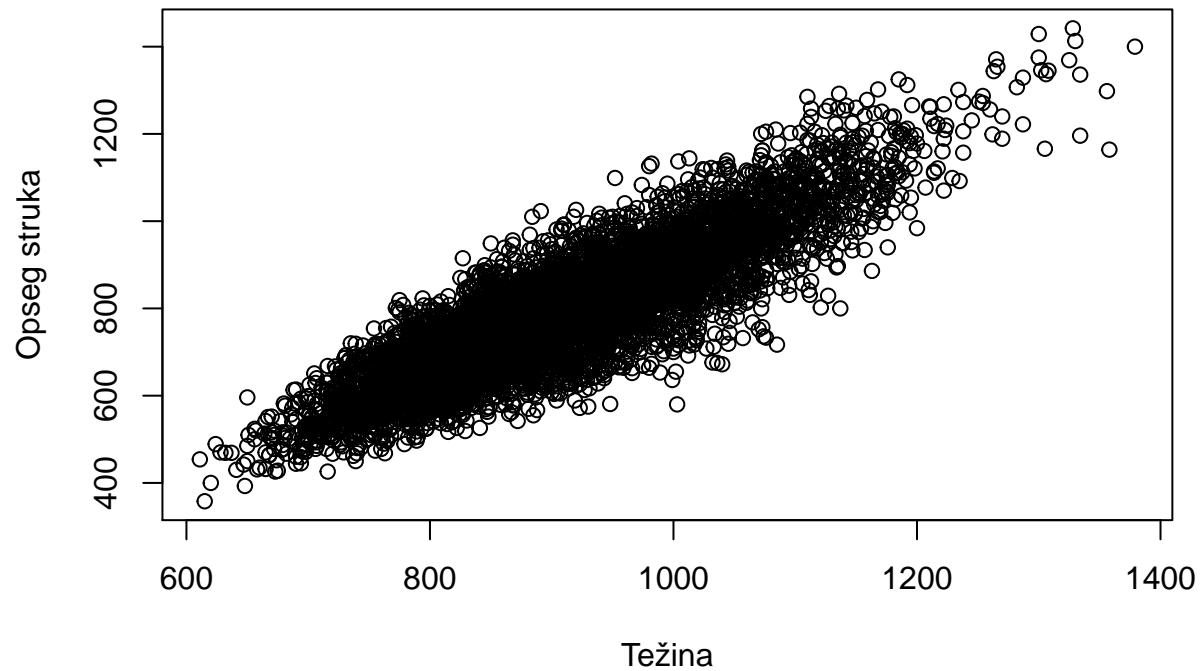
```
plot(antrData$chestcircumference, antrData$weightkg, xlab = "Težina", ylab = "Opseg prsnog koša",
     main = "Distribucija težina u ovisnosti o opsegu prsnog koša")
```

Distribucija težina u ovisnosti o opsegu prsnog koša



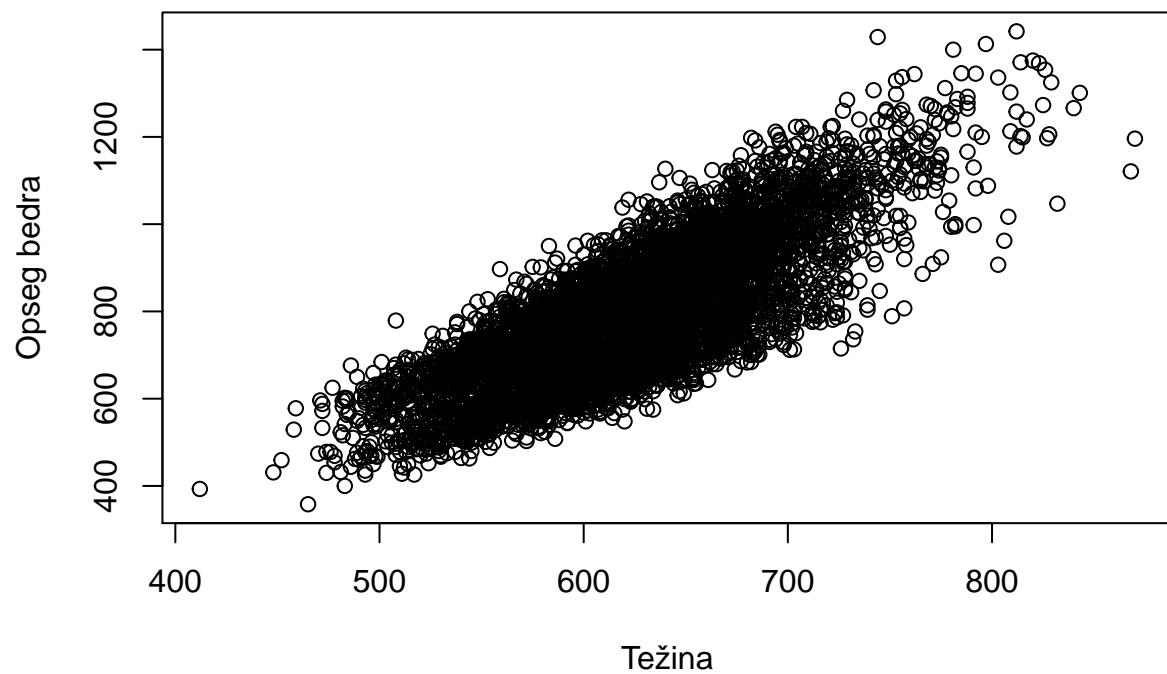
```
plot(antrData$waistcircumference, antrData$weightkg, xlab = "Težina", ylab = "Opseg struka",
main = "Distribucija težina u ovisnosti o opsegu struka")
```

Distribucija težina u ovisnosti o opsegu struka



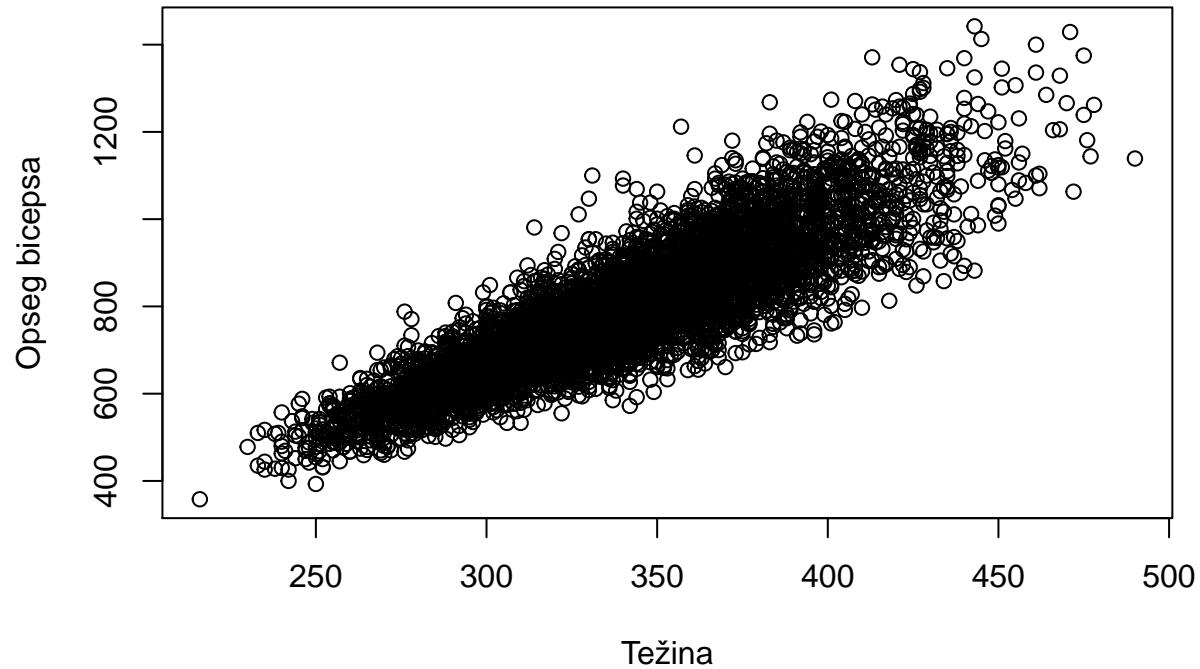
```
plot(antrData$thighcircumference, antrData$weightkg, xlab = "Težina", ylab = "Opseg bedra",
main = "Distribucija težina u ovisnosti o opsegu bedra")
```

Distribucija težina u ovisnosti o opsegu bedra



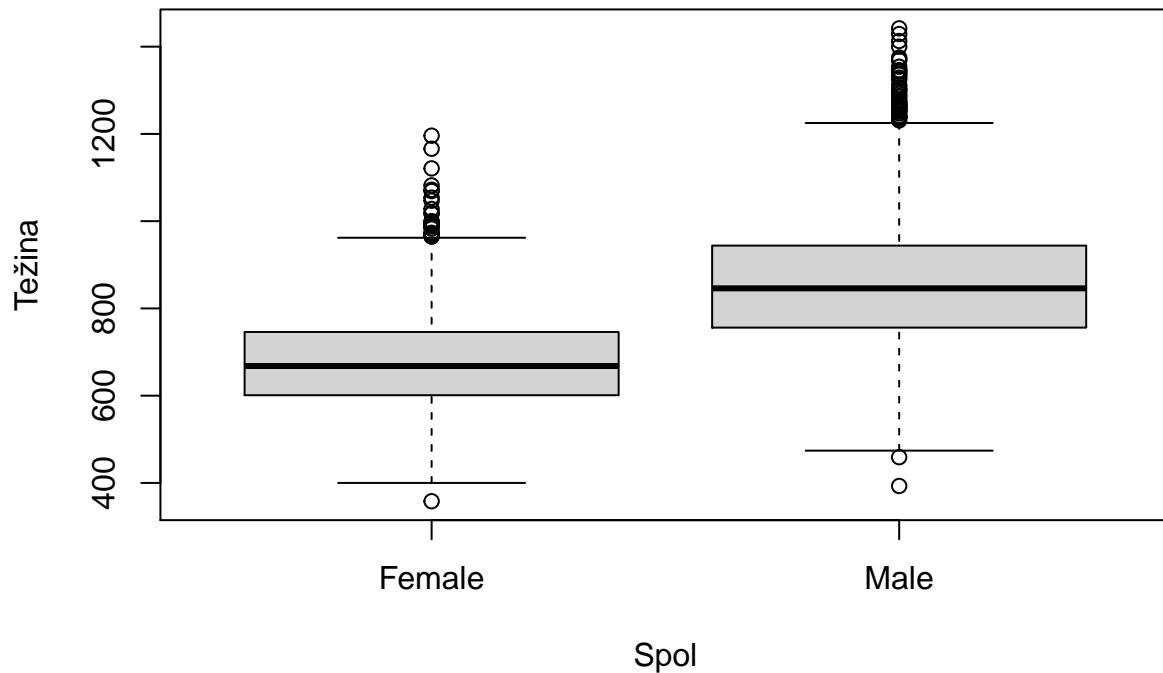
```
plot(antrData$bicepscircumferenceflexed, antrData$weightkg, xlab = "Težina", ylab = "Opseg bicepsa",
     main = "Distribucija težina u ovisnosti o opsegu bicepsa")
```

Distribucija težina u ovisnosti o opsegu bicepsa



```
boxplot(antrData$weightkg~antrData$Gender, xlab = "Spol", ylab = "Težina",
        main = "Distribucija težina u ovisnosti o spolu")
```

Distribucija težina u ovisnosti o spolu



Vidimo da većina izabranih varijabli na grafovima pokazuje jasnu zavisnost s kilažom. Varijabla koja ne pokazuje lijepi linearni trend je dob(Age) ispitanika. Sada pristupamo kostrukciji jednostavnih modela linearne regresije za svaki od izabranih parametara.

#Priprema kategorijskih varijabli

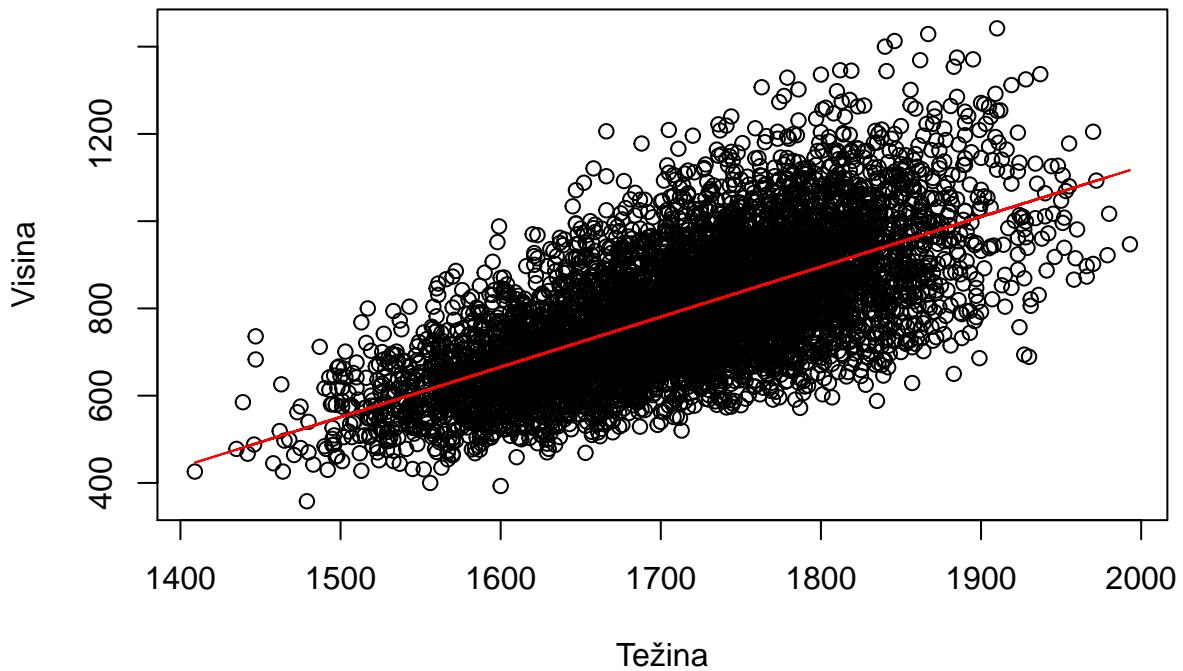
```
require(fastDummies)

## Loading required package: fastDummies
antrData = dummy_cols(antrData, select_columns = c("Gender"))

fitHeight = lm(weightkg ~ stature, data = antrData)
fitAge = lm(weightkg ~ Age, data = antrData)
fitChest = lm(weightkg ~ chestcircumference, data = antrData)
fitWaist = lm(weightkg ~ waistcircumference, data = antrData)
fitThigh = lm(weightkg ~ thighcircumference, data = antrData)
fitBiceps = lm(weightkg ~ bicepscircumferenceflexed, data = antrData)
fitGender = lm(weightkg ~ Gender_Male, data = antrData)

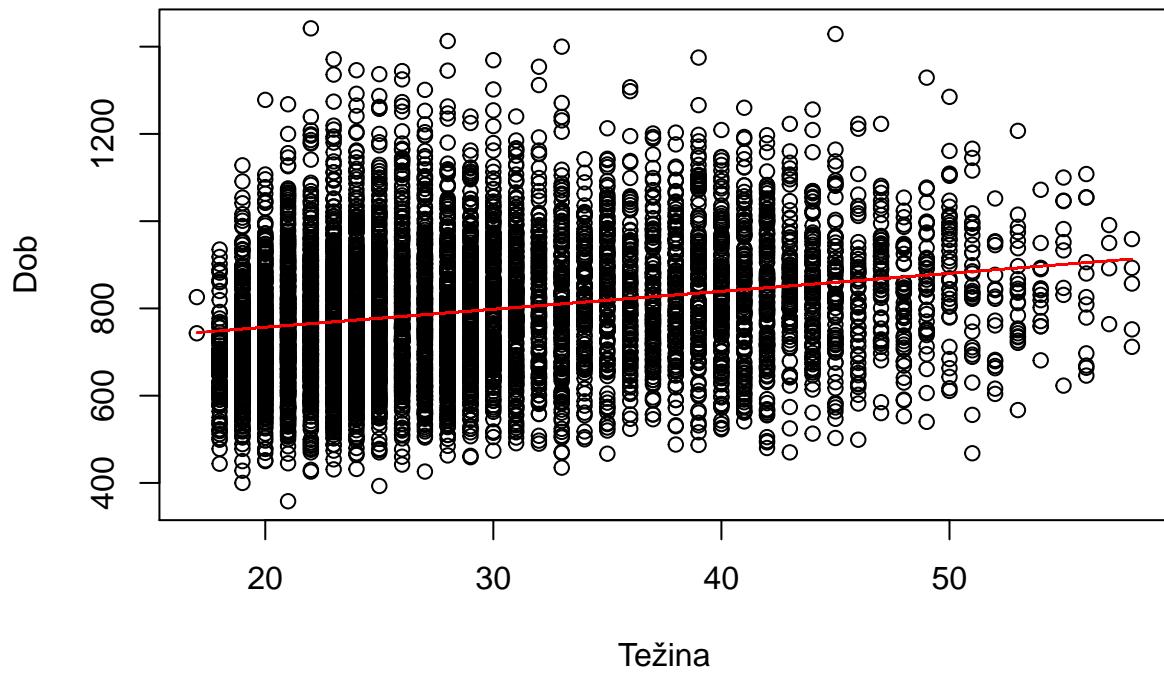
plot(antrData$stature, antrData$weightkg, xlab = "Težina", ylab = "Visina",
     main = "Distribucija težina u ovisnosti o visini")
lines(antrData$stature, fitHeight$fitted.values, col = 'red')
```

Distribucija težina u ovisnosti o visini



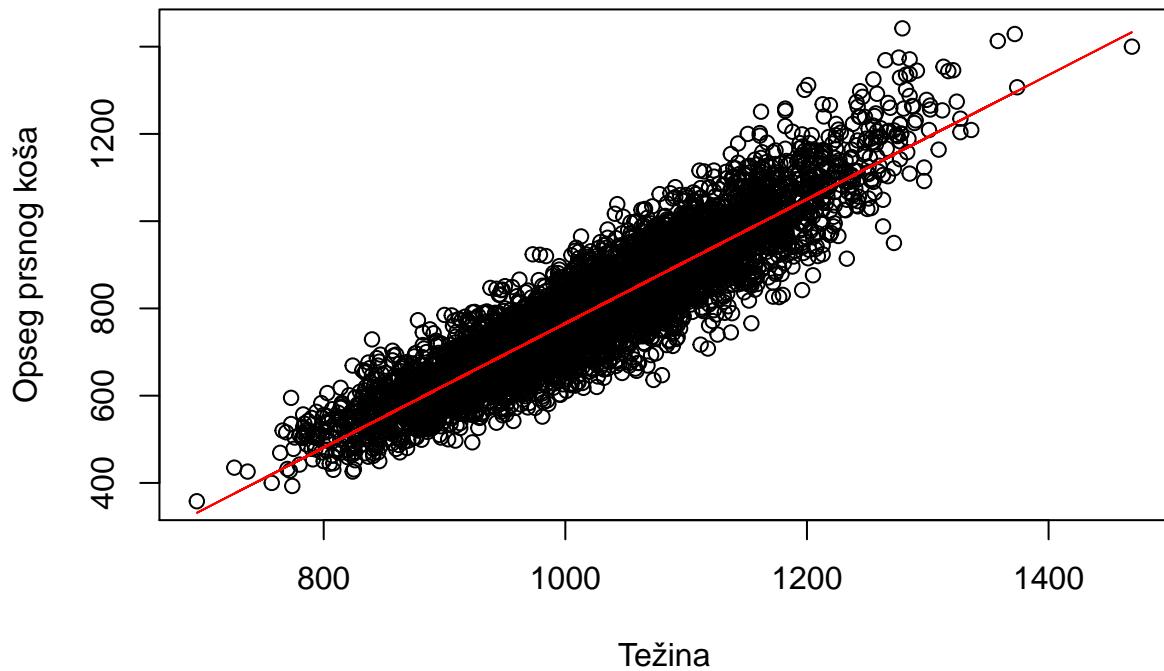
```
plot(antrData$Age, antrData$weightkg, xlab = "Težina", ylab = "Dob",
      main = "Distribucija težina u ovisnosti o dobi")
lines(antrData$Age, fitAge$fitted.values, col = 'red')
```

Distribucija težina u ovisnosti o dobi



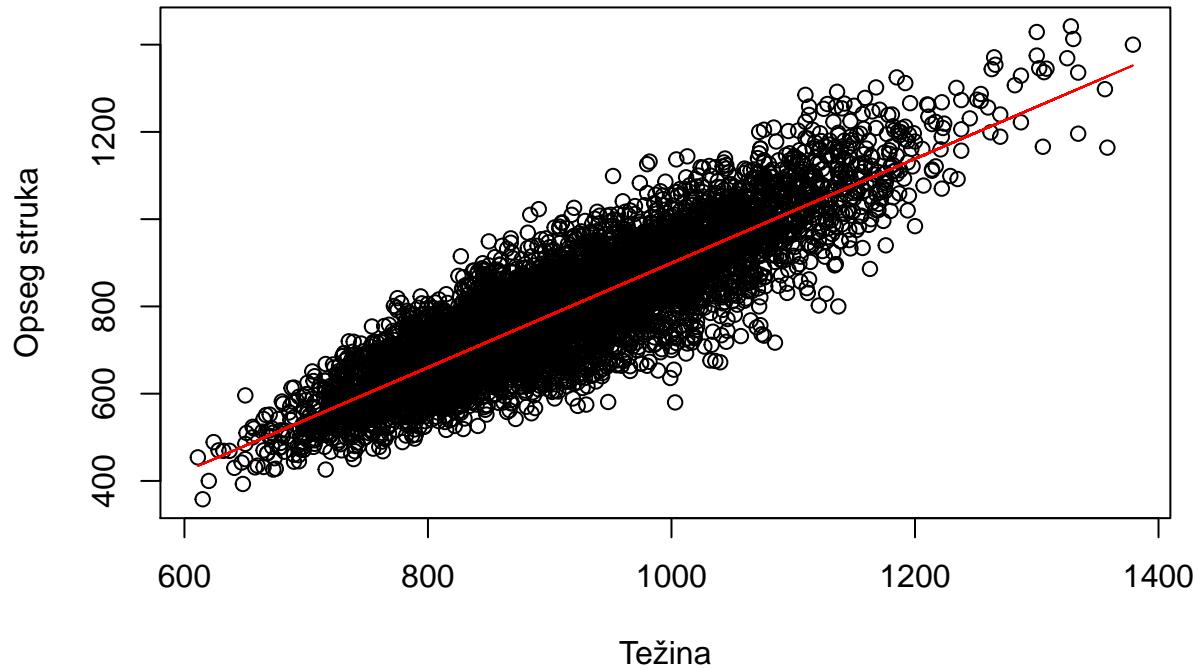
```
plot(antrData$chestcircumference, antrData$weightkg, xlab = "Težina", ylab = "Opseg prsnog koša",
      main = "Distribucija težina u ovisnosti o opsegu prsnog koša")
lines(antrData$chestcircumference, fitChest$fitted.values, col = 'red')
```

Distribucija težina u ovisnosti o opsegu prsnog koša



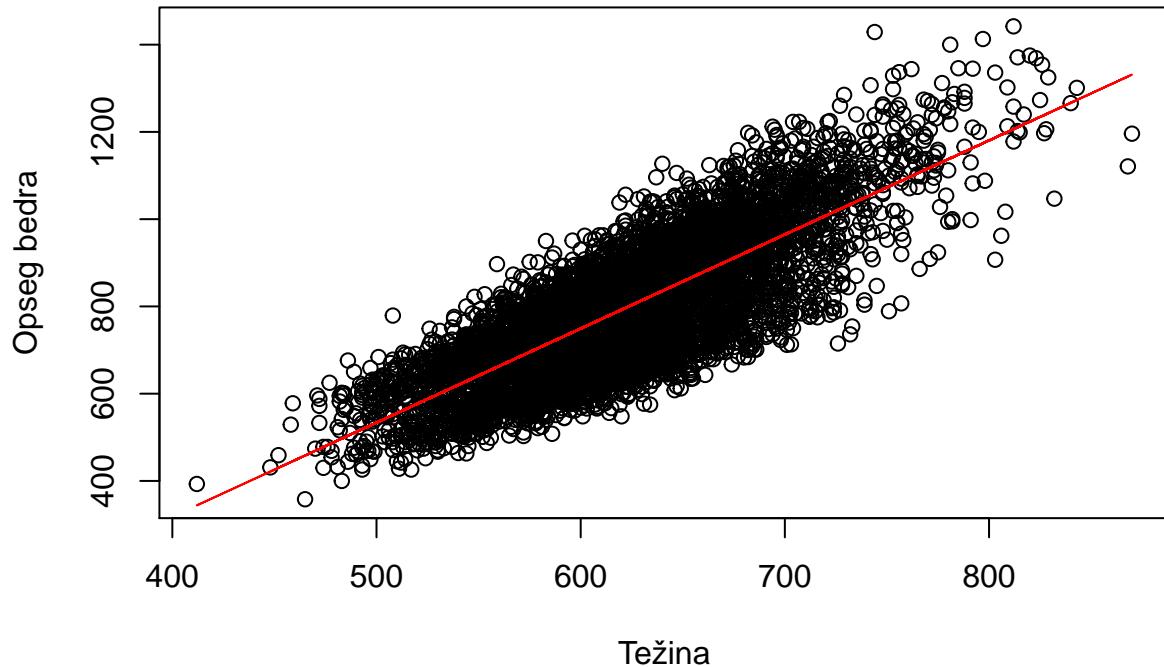
```
plot(antrData$waistcircumference, antrData$weightkg, xlab = "Težina", ylab = "Opseg struka",
      main = "Distribucija težina u ovisnosti o opsegu struka")
lines(antrData$waistcircumference, fitWaist$fitted.values, col = 'red')
```

Distribucija težina u ovisnosti o opsegu struka



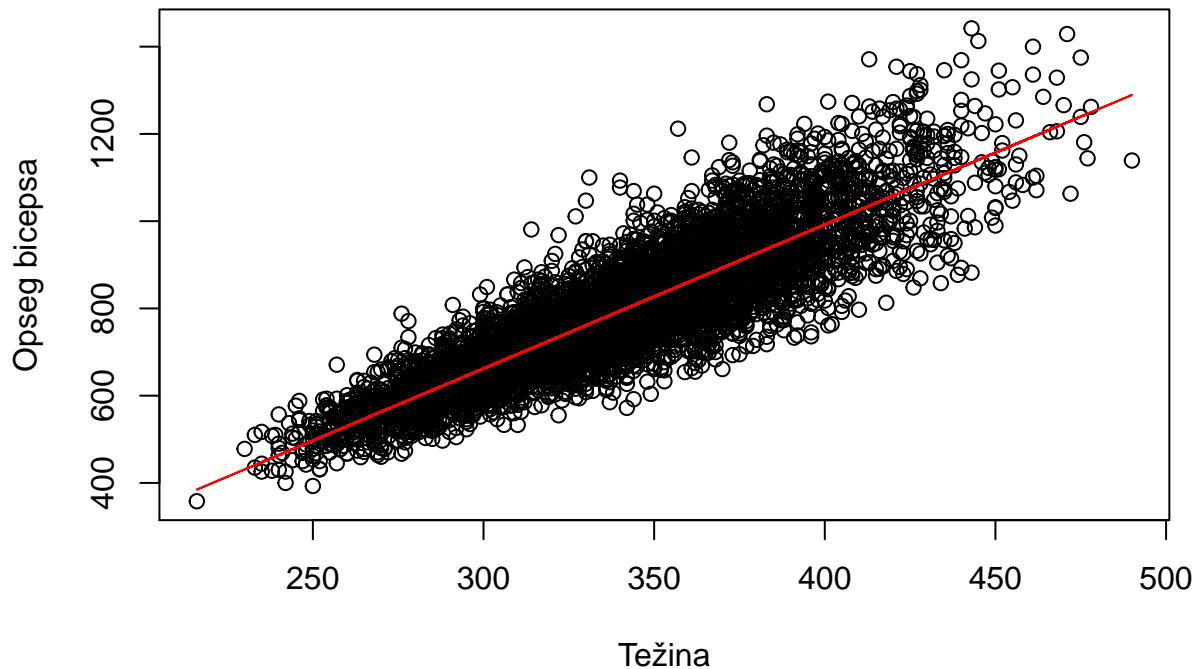
```
plot(antrData$thighcircumference, antrData$weightkg, xlab = "Težina", ylab = "Opseg bedra",
      main = "Distribucija težina u ovisnosti o opsegu bedra")
lines(antrData$thighcircumference, fitThigh$fitted.values, col = 'red')
```

Distribucija težina u ovisnosti o opsegu bedra



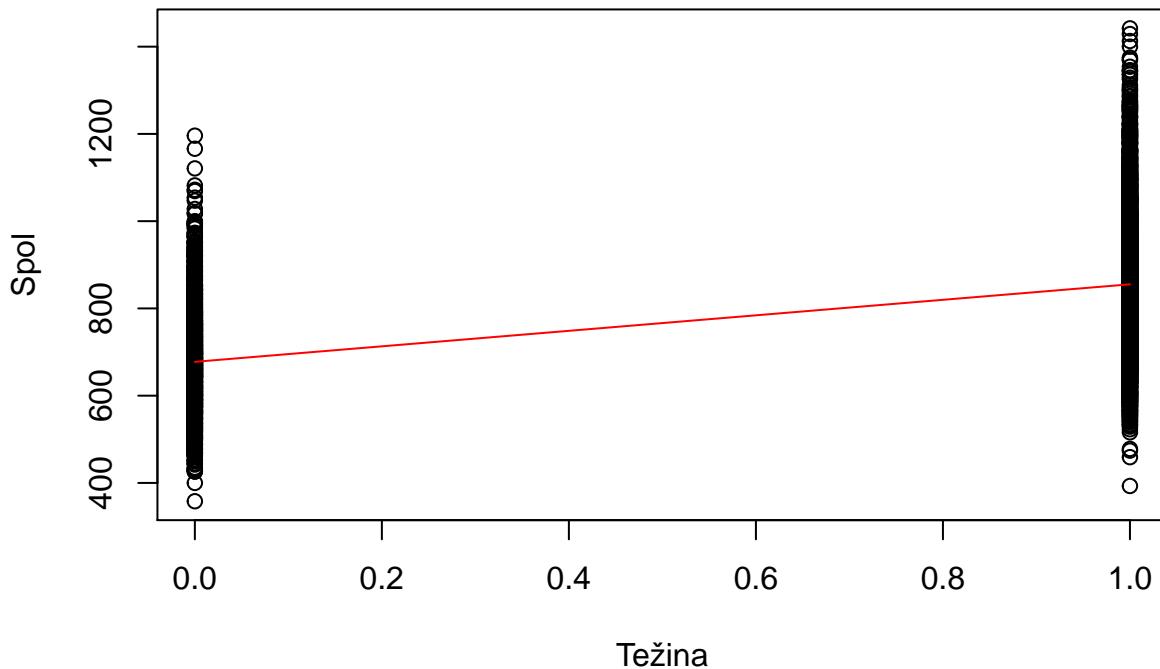
```
plot(antrData$bicepscircumferenceflexed, antrData$weightkg, xlab = "Težina", ylab = "Opseg bicepsa",
     main = "Distribucija težina u ovisnosti o opsegu bicepsa")
lines(antrData$bicepscircumferenceflexed, fitBiceps$fitted.values, col = 'red')
```

Distribucija težina u ovisnosti o opsegu bicepsa



```
plot(antrData$Gender_Male, antrData$weightkg, xlab = "Težina", ylab = "Spol",
      main = "Distribucija težina u ovisnosti o spolu")
lines(antrData$Gender_Male, fitGender$fitted.values, col = "red")
```

Distribucija težina u ovisnosti o spolu



```
summary(fitHeight)

##
## Call:
## lm(formula = weightkg ~ stature, data = antrData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -355.63  -81.63   -5.94   72.39  464.82 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -1.171e+03  2.879e+01 -40.69   <2e-16 ***
## stature      1.148e+00  1.677e-02   68.47   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 117.6 on 6066 degrees of freedom
## Multiple R-squared:  0.4359, Adjusted R-squared:  0.4359 
## F-statistic:  4688 on 1 and 6066 DF,  p-value: < 2.2e-16

summary(fitAge)

##
## Call:
## lm(formula = weightkg ~ Age, data = antrData)
##
```

```

## Residuals:
##      Min     1Q Median     3Q    Max
## -416.51 -109.10 -11.38  96.30 676.85
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 674.6014    6.9926  96.47 <2e-16 ***
## Age         4.1158    0.2256  18.25 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 152.4 on 6066 degrees of freedom
## Multiple R-squared:  0.05203, Adjusted R-squared:  0.05187
## F-statistic: 332.9 on 1 and 6066 DF, p-value: < 2.2e-16
summary(fitChest)

##
## Call:
## lm(formula = weightkg ~ chestcircumference, data = antrData)
##
## Residuals:
##      Min     1Q Median     3Q    Max
## -233.552 -40.286 -1.407  39.711 279.205
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -657.87136   8.24720 -79.77 <2e-16 ***
## chestcircumference 1.42351   0.00803 177.27 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 62.97 on 6066 degrees of freedom
## Multiple R-squared:  0.8382, Adjusted R-squared:  0.8382
## F-statistic: 3.143e+04 on 1 and 6066 DF, p-value: < 2.2e-16
summary(fitWaist)

##
## Call:
## lm(formula = weightkg ~ waistcircumference, data = antrData)
##
## Residuals:
##      Min     1Q Median     3Q    Max
## -322.89 -50.10 -0.17  51.79 257.08
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.961e+02  7.913e+00 -37.42 <2e-16 ***
## waistcircumference 1.195e+00  8.586e-03 139.22 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 76.43 on 6066 degrees of freedom
## Multiple R-squared:  0.7616, Adjusted R-squared:  0.7616

```

```

## F-statistic: 1.938e+04 on 1 and 6066 DF, p-value: < 2.2e-16
summary(fitThigh)

##
## Call:
## lm(formula = weightkg ~ thighcircumference, data = antrData)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -305.88 -72.16  12.08  67.47 369.32
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -544.28737   13.17970  -41.3   <2e-16 ***
## thighcircumference  2.15587    0.02109   102.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 94.88 on 6066 degrees of freedom
## Multiple R-squared:  0.6327, Adjusted R-squared:  0.6326
## F-statistic: 1.045e+04 on 1 and 6066 DF, p-value: < 2.2e-16
summary(fitBiceps)

##
## Call:
## lm(formula = weightkg ~ bicepscircumferenceflexed, data = antrData)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -252.04 -49.25 -1.95  45.07 361.87
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -328.43475   8.02766  -40.91   <2e-16 ***
## bicepscircumferenceflexed  3.30131   0.02337  141.24   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 75.59 on 6066 degrees of freedom
## Multiple R-squared:  0.7668, Adjusted R-squared:  0.7668
## F-statistic: 1.995e+04 on 1 and 6066 DF, p-value: < 2.2e-16
summary(fitGender)

##
## Call:
## lm(formula = weightkg ~ Gender_Male, data = antrData)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -462.24 -91.24 - 9.24  80.76 586.76
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
```

```

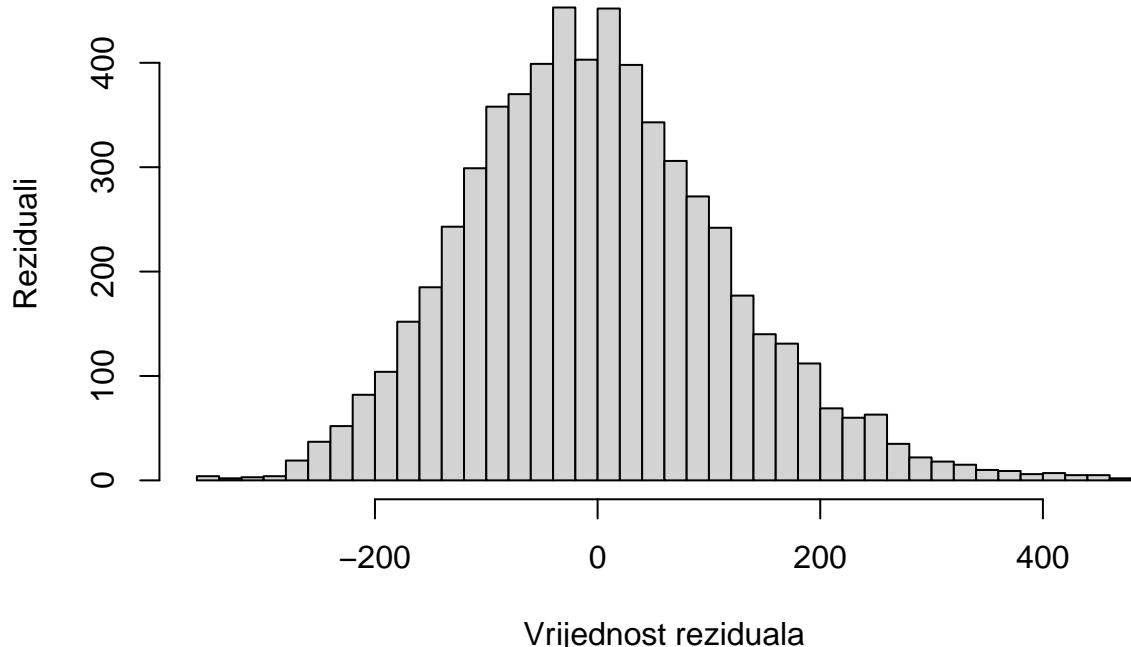
## (Intercept) 677.582      2.973 227.91    <2e-16 ***
## Gender_Male 177.658      3.625 49.01    <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 132.5 on 6066 degrees of freedom
## Multiple R-squared:  0.2837, Adjusted R-squared:  0.2836
## F-statistic:  2402 on 1 and 6066 DF,  p-value: < 2.2e-16

```

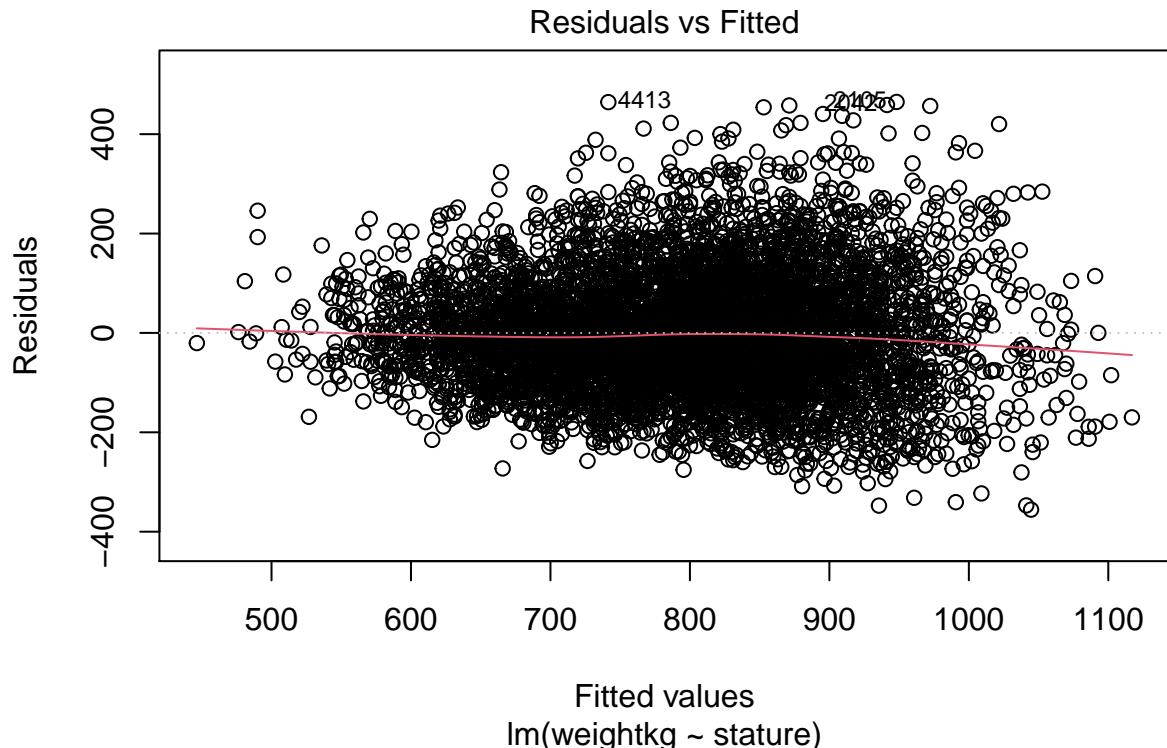
Univarijatnim modelima za sve inicijalno izabrane varijable vidimo kako je inicijalni izbor parametara bio zadovoljavajući. T testovi nad dobivenim koeficijentima pokazuju visoku razinu značajnosti, a koeficijent determinacije R^2 objašnjava dobar dio varijance u podacima. Univarijatni modeli koji ne postižu zadovoljavajuće rezultate su dob i spol. R^2 vrijednost u modelu u kojem je regresor dob je 0.05302 što ukazuje na to da objašnjava jako mali dio varijance u podacima. Isto tako spol ima malu R^2 vrijednost te ćemo u daljoj analizi razmotriti ako nam se više isplati imati odvojene liniarne modele za muškarce i žene.

```
#Provjeri normalnosti
hist(fitHeight$residuals, xlab = "Vrijednost reziduala", ylab = "Reziduali", breaks = 30, main = "Reziduali modela s regresorom visina")
```

Reziduali modela s regresorom visina

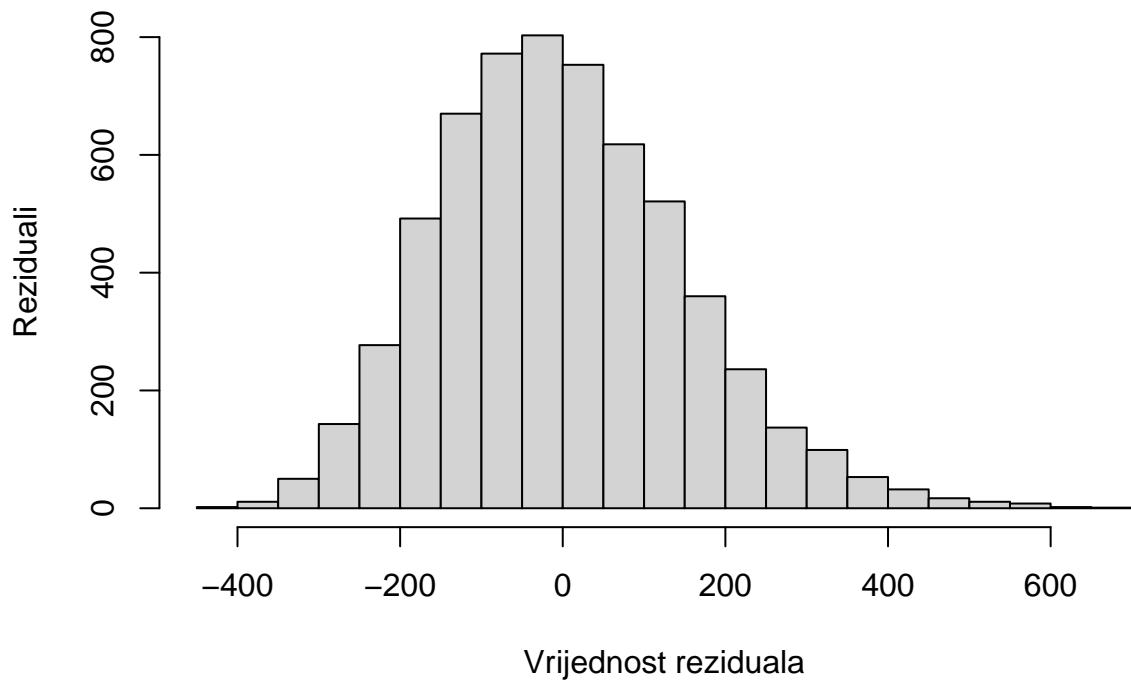


```
plot(fitHeight, which = 1)
```

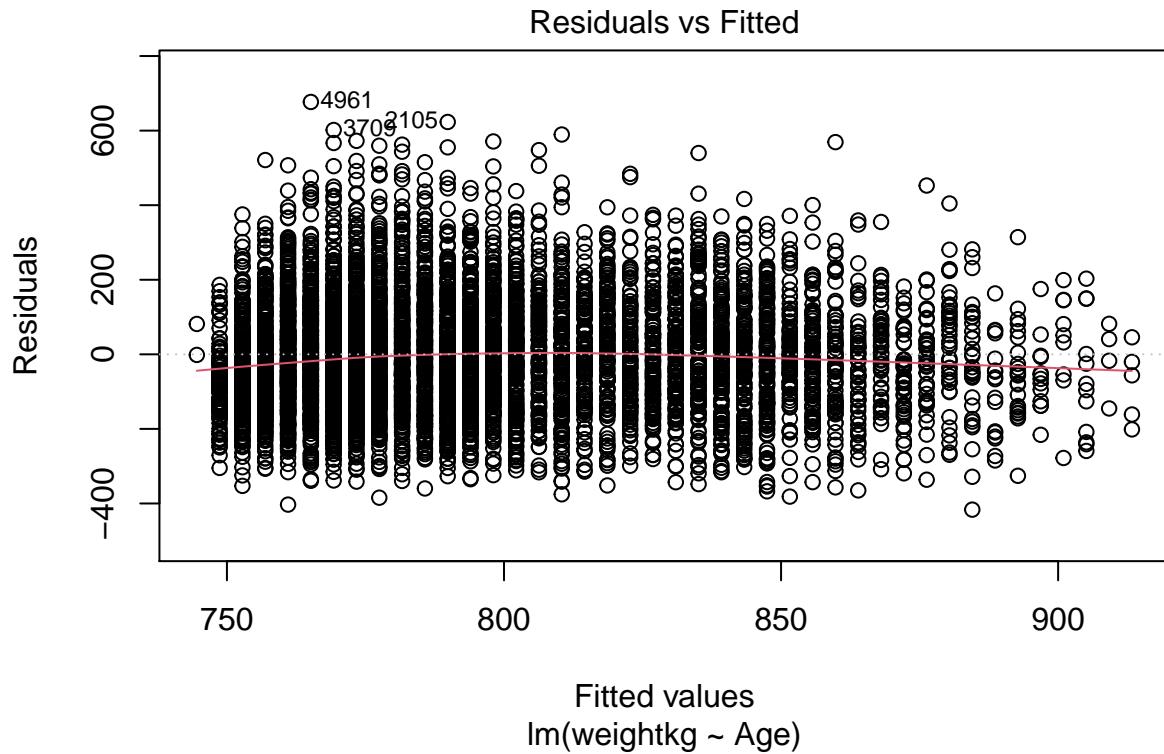


```
hist(fitAge$residuals, xlab = "Vrijednost reziduala", ylab = "Reziduali", breaks = 30, main = "Reziduali")
```

Reziduali modela s regresorom dob

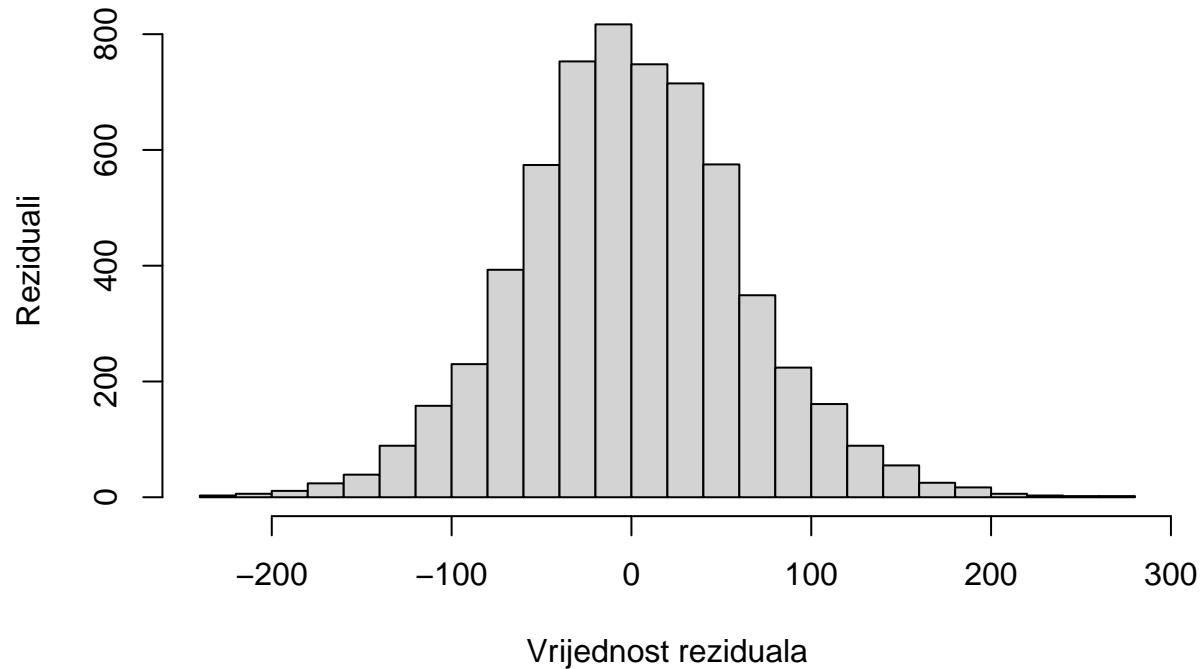


```
plot(fitAge, which = 1)
```

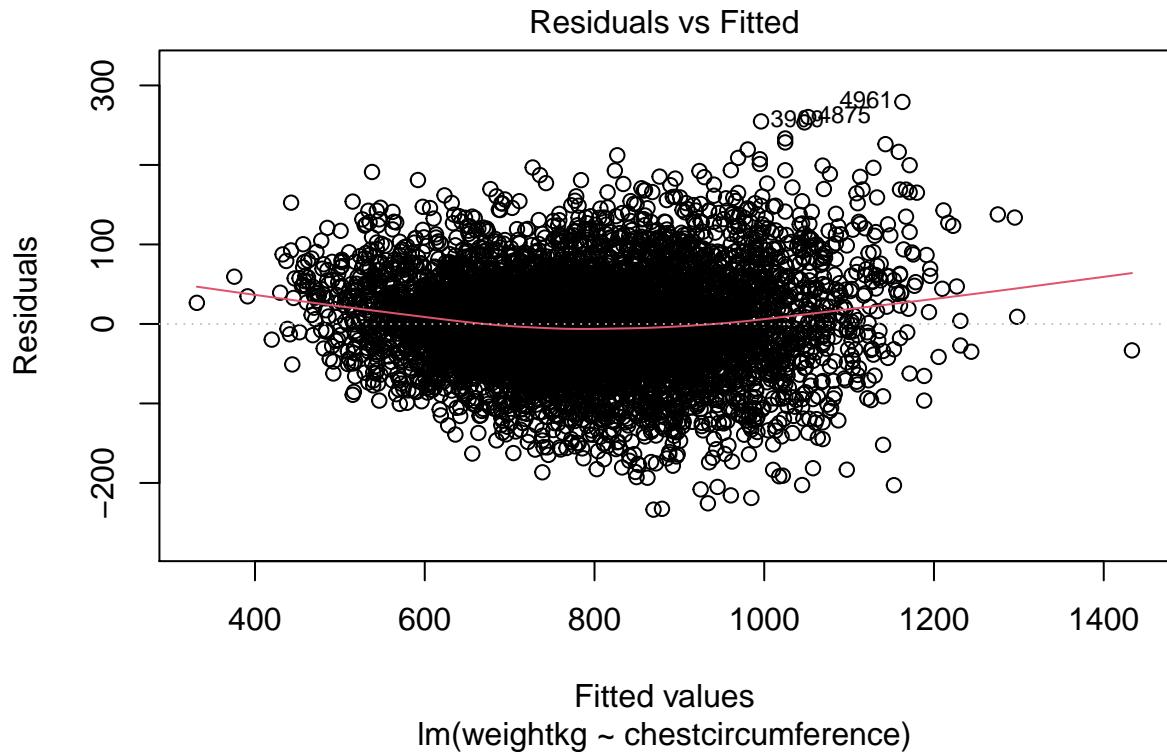


```
hist(fitChest$residuals, xlab = "Vrijednost reziduala", ylab = "Reziduali", breaks = 30, main = "Reziduali")
```

Reziduali modela s regresorom opseg prsnog koša

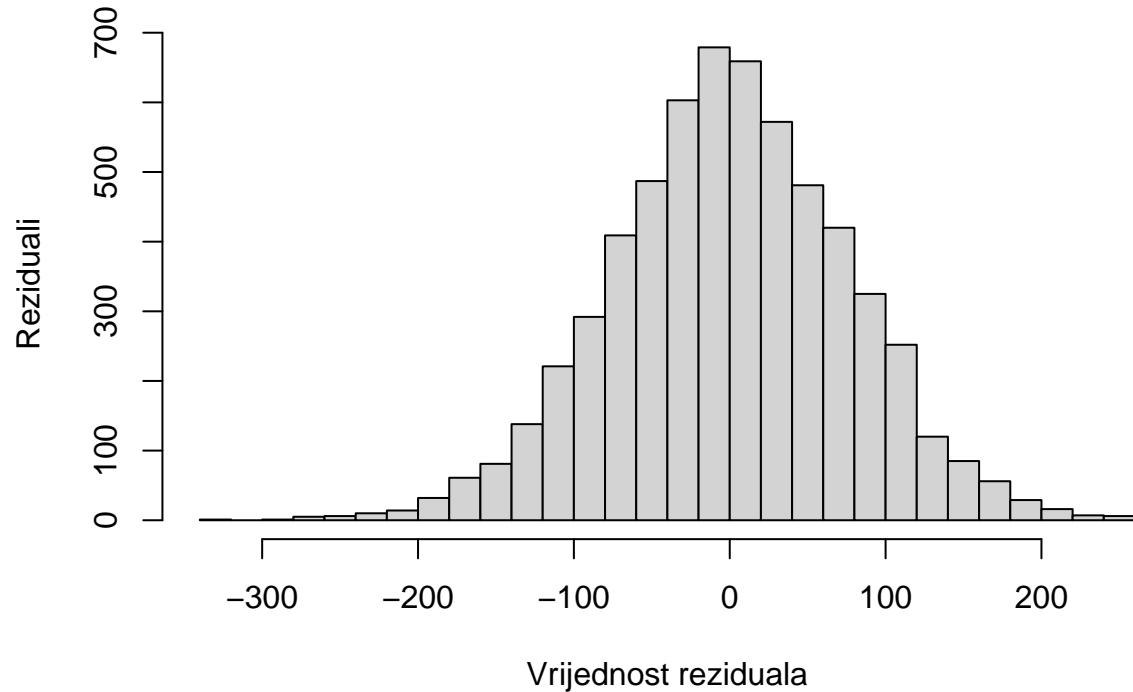


```
plot(fitChest, which = 1)
```

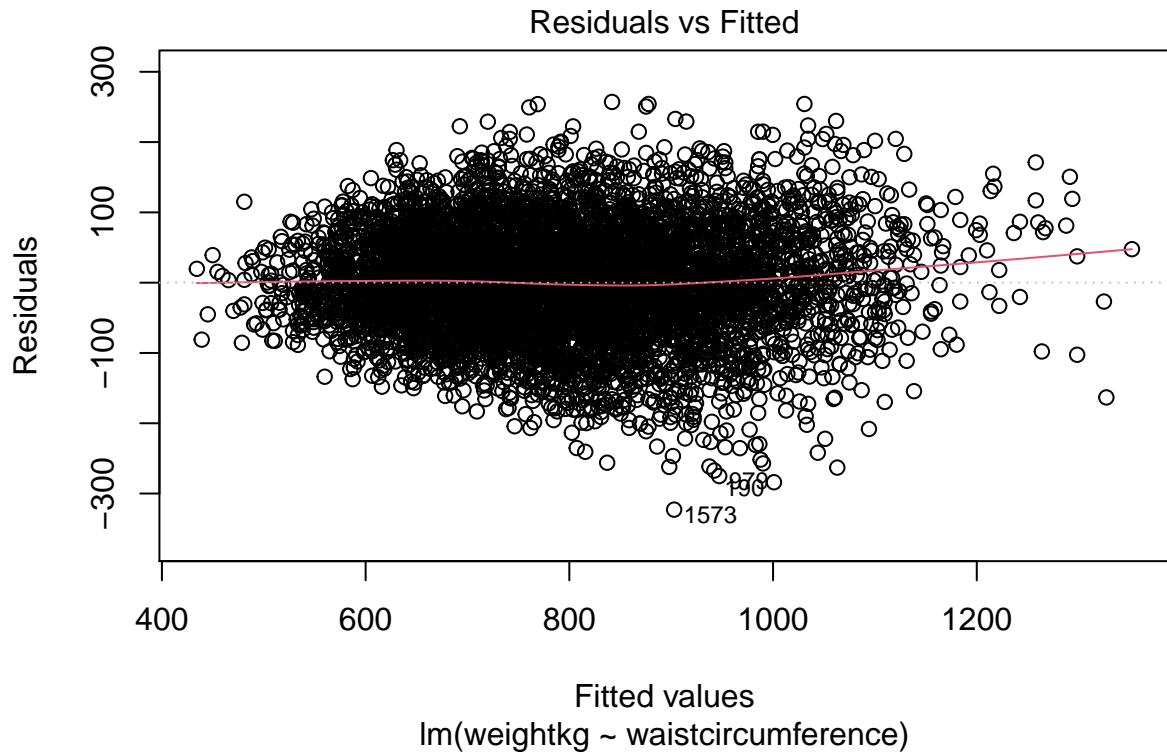


```
hist(fitWaist$residuals, xlab = "Vrijednost reziduala", ylab = "Reziduali", breaks = 30, main = "Reziduali")
```

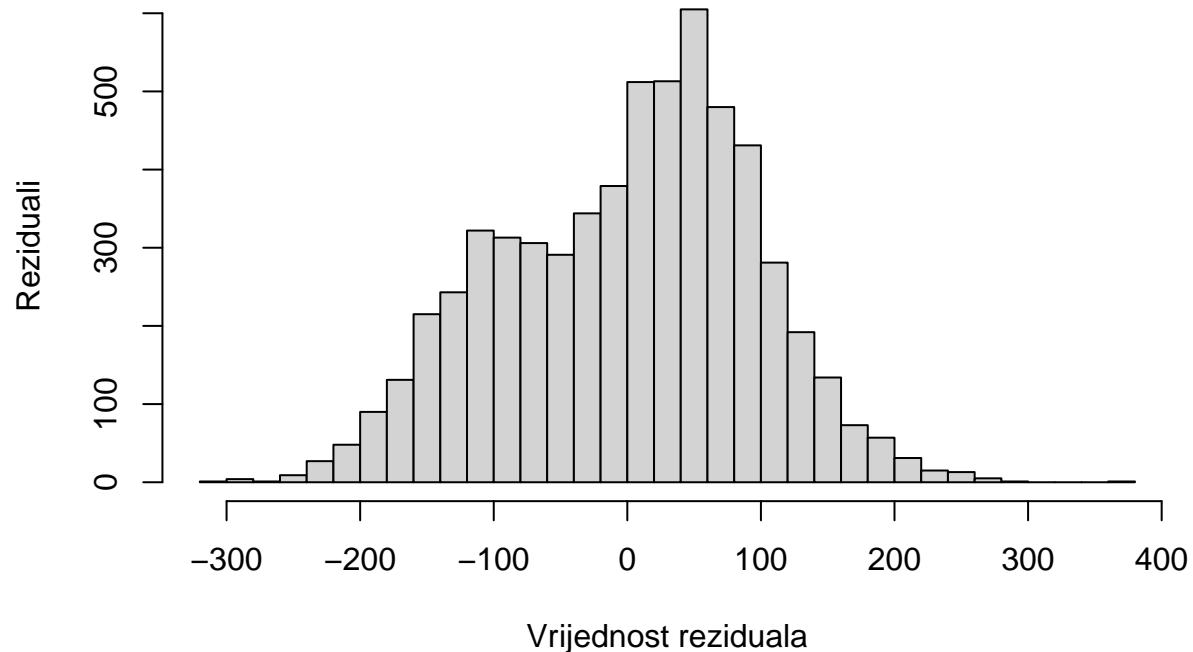
Reziduali modela s regresorom opseg struka



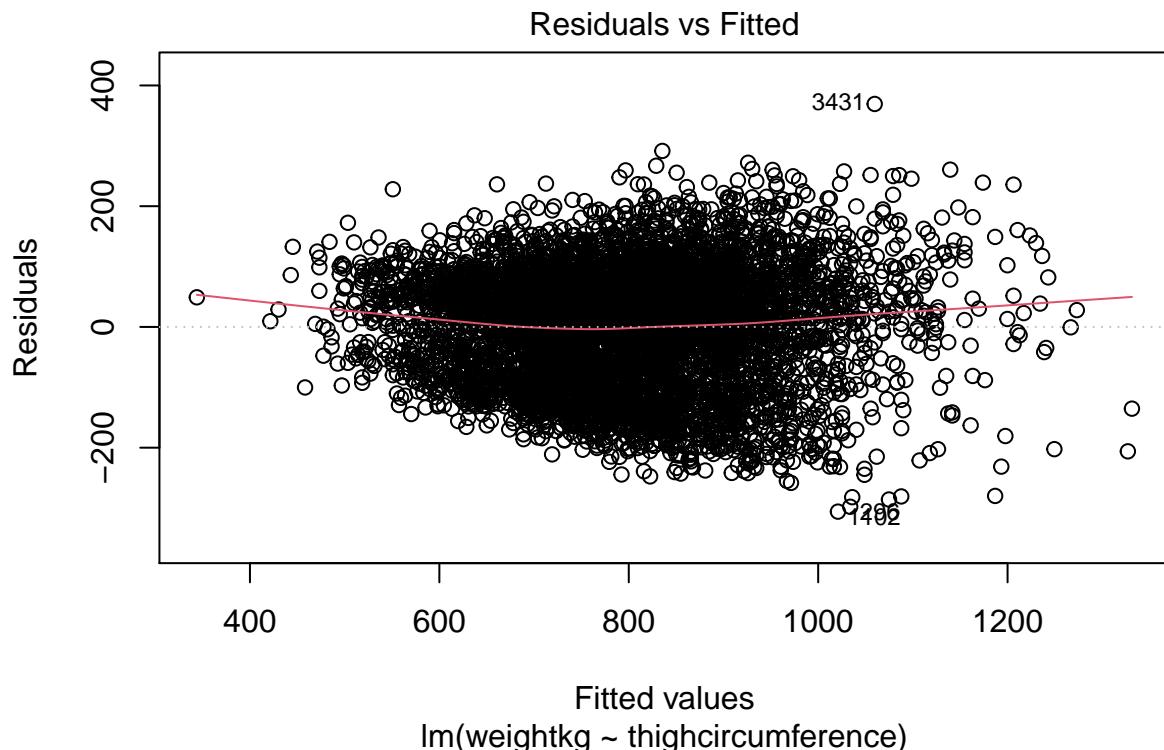
```
plot(fitWaist, which = 1)
```



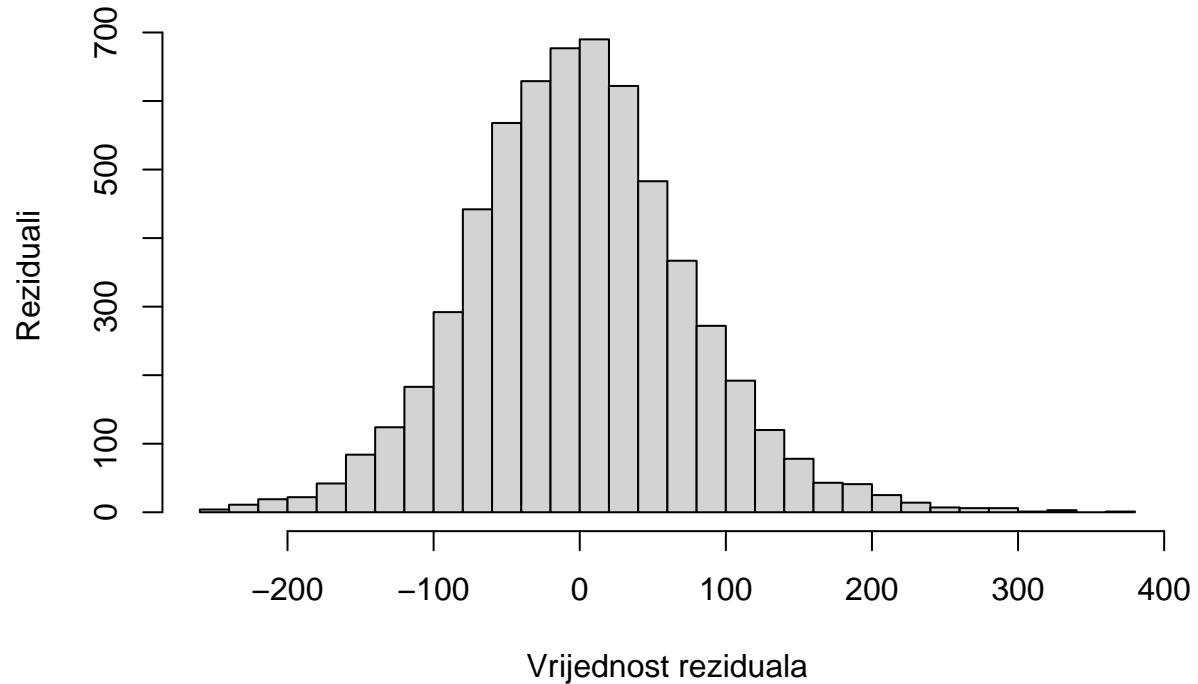
Reziduali modela s regresorom opseg bedra



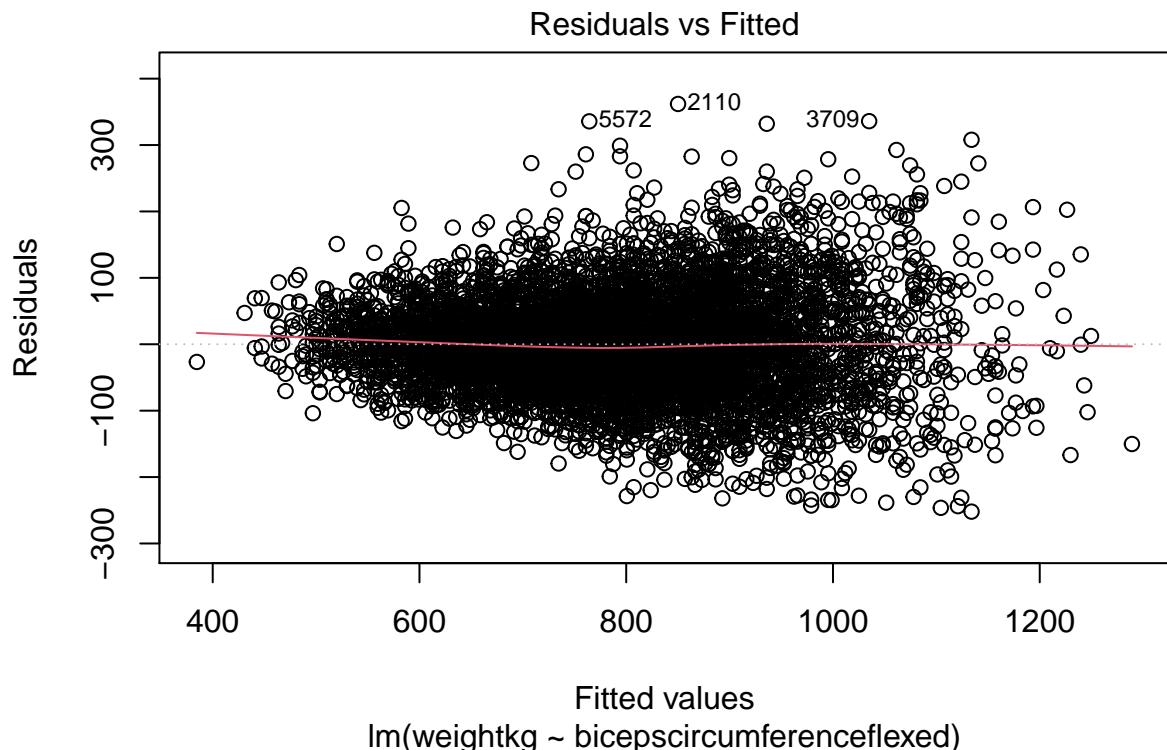
```
plot(fitThigh, which = 1)
```



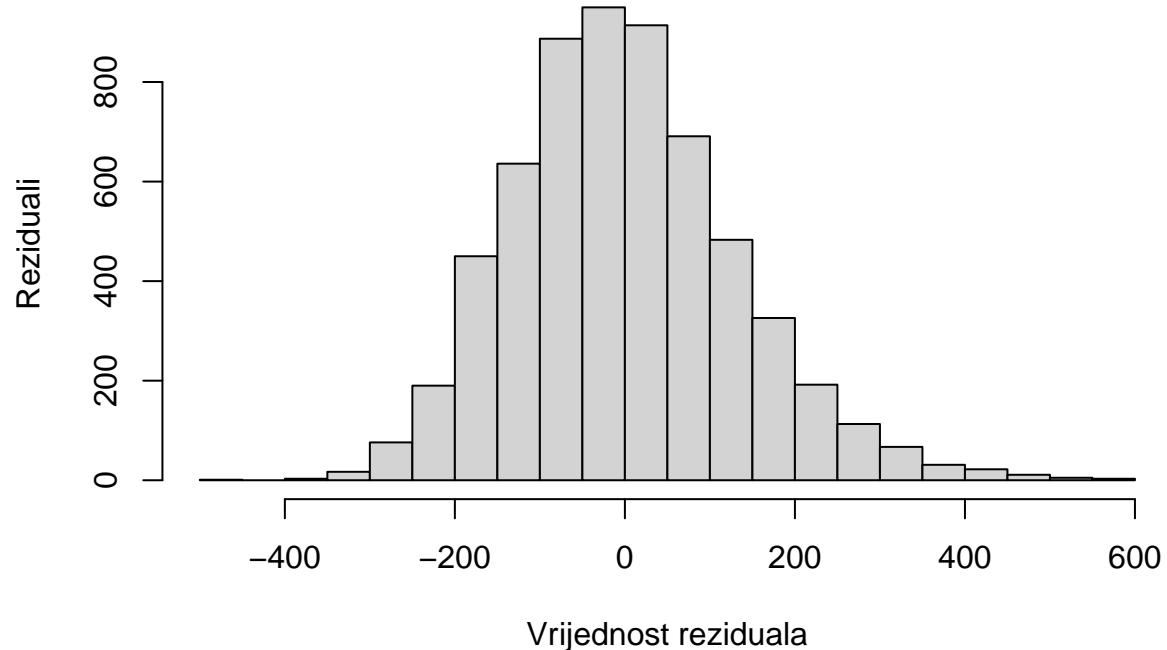
Reziduali modela s regresorom opseg bicepsa



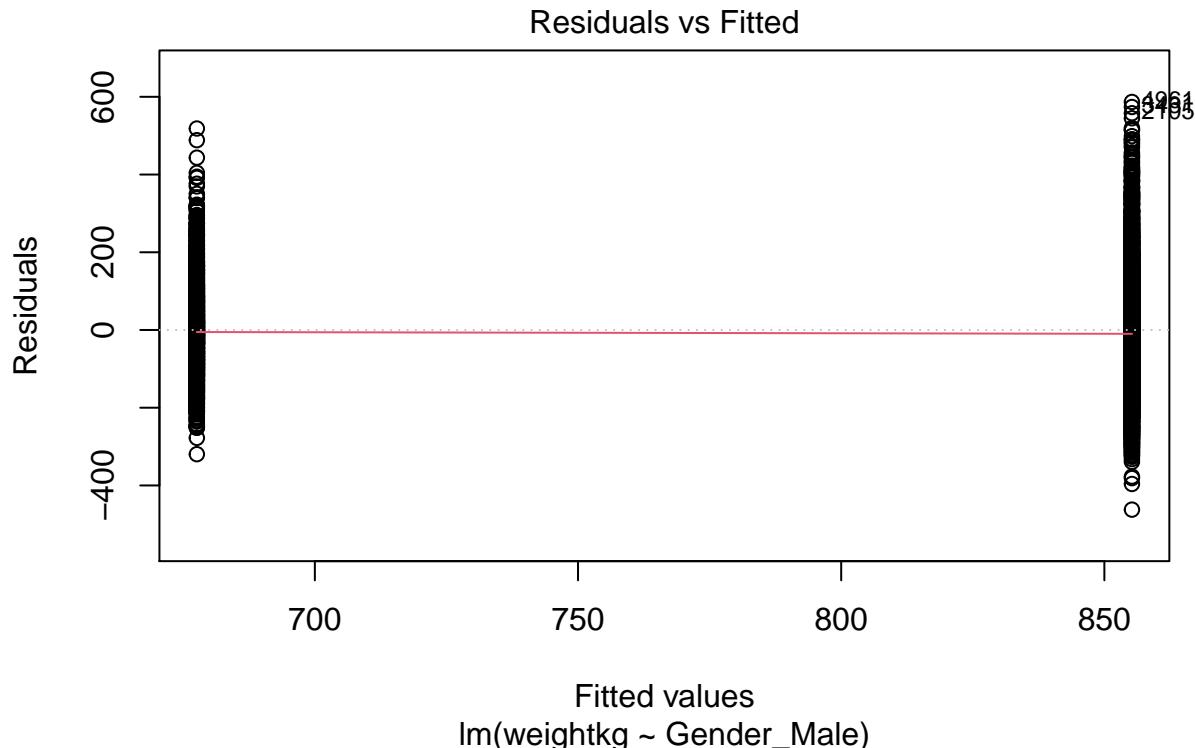
```
plot(fitBiceps, which = 1)
```



Reziduali modela s regresorom spol



```
plot(fitGender, which = 1)
```



Iz histograma vidimo kako distribucija reziduala nema teških repova i kako izgleda realtivno normalno distribuirano tako da možemo iz toga ako uzmemo u obzir robustnost t-testa na normalnost zaključiti kako je početna nužna pretpostavka o normalnosti distribucija zadovoljena. Isto tako je bitna pretpostavka o homogenosti varijanci modela. Iz grafova reziduala i fitanih vrijednosti vidimo kako u većini slučajeva pretpostavka vrijedi. Opsezi pokazuje određenu zavisnost koja bi se mogla objasniti varijacijama u spolu. S tim idejama krećemo u daljnju analizu u kojoj ćemo konstruirati multivarijatni regresijski model za predviđanje težine.

Ipak prije nastavka moramo proučiti neke pretpostavke modela. Kako imamo 4 regresora koja se tiču opsega dijelova ljudskog tijela za očekivati je kako bi ti regresori mogli biti jako korelirani i samim time u međusobnom prisutstvu smanjiti svoju značajnost u modelu.

```
cor(cbind(antrData$waistcircumference, antrData$bicepscircumflexed, antrData$chestcircumference,
          antrData$thighcircumference, antrData$stature, antrData$Age))

##           [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
## [1,] 1.0000000 0.7266864 0.8749186 0.7830274 0.36873362 0.36197157
## [2,] 0.7266864 1.0000000 0.8413394 0.6825308 0.52029558 0.21722676
## [3,] 0.8749186 0.8413394 1.0000000 0.6921765 0.50701012 0.32219486
## [4,] 0.7830274 0.6825308 0.6921765 1.0000000 0.25241652 0.14108139
## [5,] 0.3687336 0.5202956 0.5070101 0.2524165 1.00000000 0.04044715
## [6,] 0.3619716 0.2172268 0.3221949 0.1410814 0.04044715 1.00000000
```

Kao što je i očekivano regresori koji se tiču opsega su međusobno jako korelirani, tako da, prije njihovog uključivanja u ukupni model, ćemo konstruirati model koji se sastoji samo od ta 4 regresora.

```
fitCircumference = lm(weightkg ~ waistcircumference + bicepscircumflexed + chestcircumference +
                        + thighcircumference, data = antrData)
```

```

summary(fitCircumference)

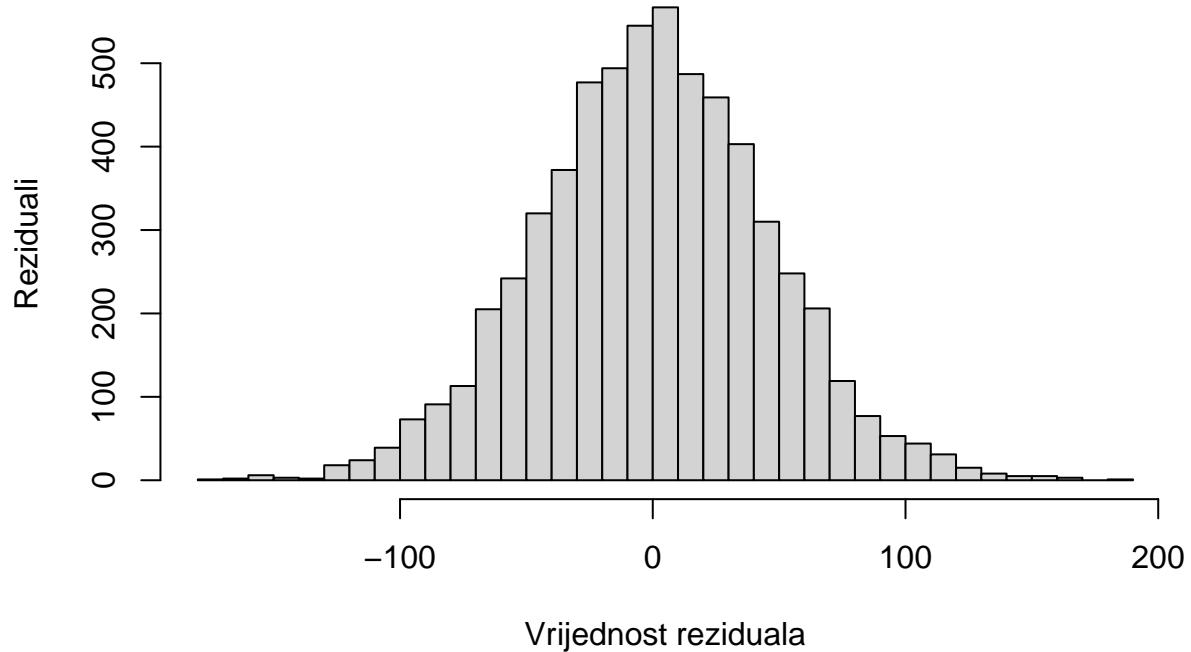
##
## Call:
## lm(formula = weightkg ~ waistcircumference + bicepscircumferenceflexed +
##      chestcircumference + thighcircumference, data = antrData)
##
## Residuals:
##       Min     1Q   Median     3Q    Max 
## -179.966 -29.846   0.153  30.262 181.298 
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)             -733.09334   7.64967 -95.83   <2e-16 ***
## waistcircumference      0.23347   0.01266  18.45   <2e-16 ***
## bicepscircumferenceflexed 1.12923   0.02782  40.59   <2e-16 ***
## chestcircumference      0.59021   0.01570  37.60   <2e-16 ***
## thighcircumference       0.52786   0.01738  30.38   <2e-16 ***  
## ---                        
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
##
## Residual standard error: 46.05 on 6063 degrees of freedom
## Multiple R-squared:  0.9135, Adjusted R-squared:  0.9134 
## F-statistic: 1.601e+04 on 4 and 6063 DF,  p-value: < 2.2e-16

```

Suprotno očekivanjima svi koeficijenti uz regresore se razlikuju od nule uz zanemarivo malu p vrijednost. Sami model objašnjava preko 91% varijance u podacima. Možda najbitnija činjenica je kako je R^2_{adj} neznatno manji od R^2 što je dobar indikator da ne koristimo suviše regresore u modelu.

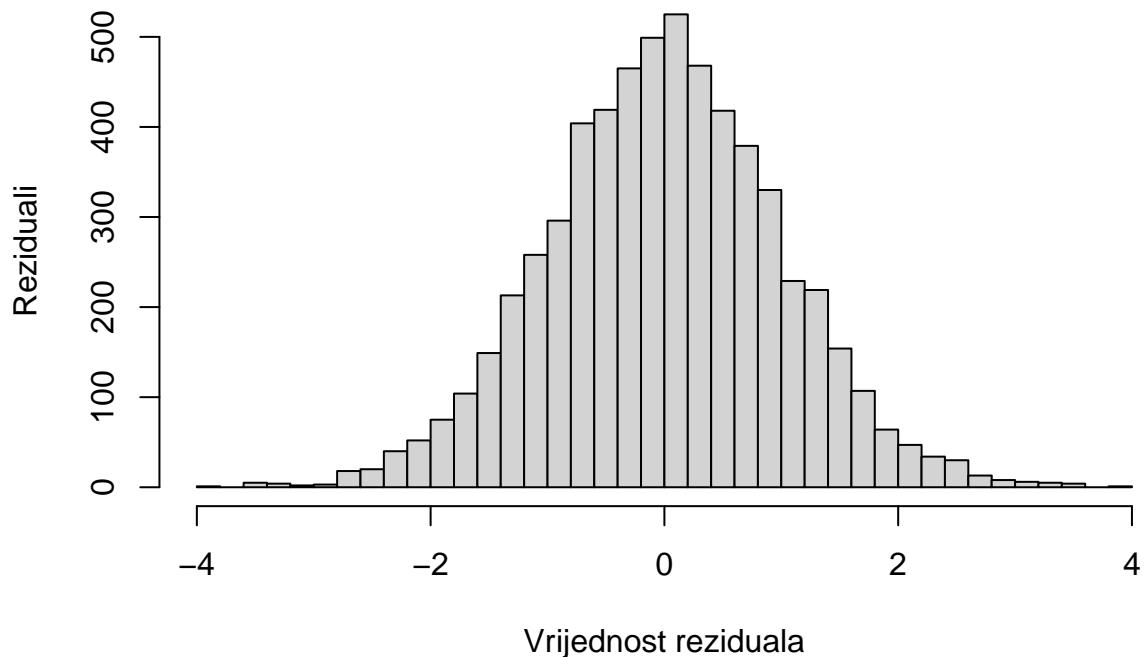
```
hist(fitCircumference$residuals, xlab = "Vrijednost reziduala", ylab = "Reziduali", breaks = 30, main = "Histogram reziduala")
```

Reziduali modela s regresorima koji imaju veze s opsegom

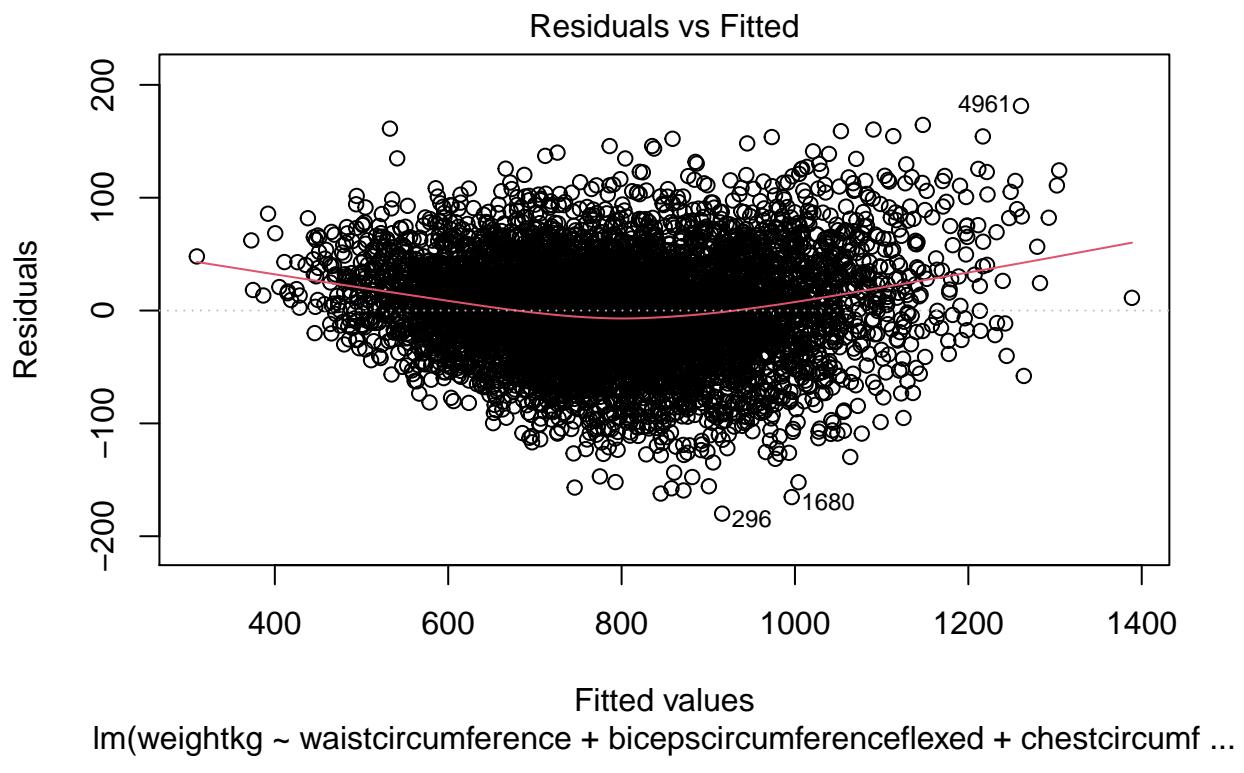


```
hist(rstandard(fitCircumference), xlab = "Vrijednost reziduala", ylab = "Reziduali", breaks = 30, main = "Reziduali modela s regresorima koji imaju veze s opsegom")
```

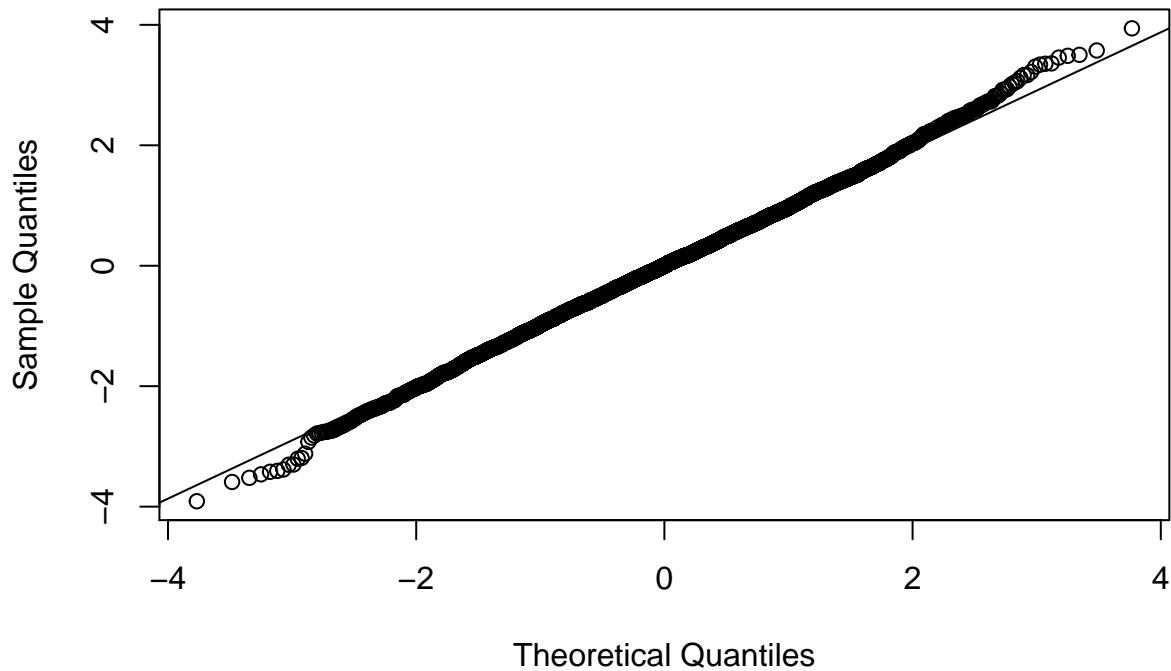
Standardizirani reziduali modela s regresorima koji imaju veze s opsegom



```
plot(fitCircumference, which = 1)
```



Normal Q-Q Plot



```
require(nortest)
lillie.test(rstandard(fitCircumference))

##
## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: rstandard(fitCircumference)
## D = 0.0092922, p-value = 0.2314
```

Ovaj model pokazuje jako dobre rezultate i vidimo kako je pretpostavka o normalnosti zadovoljena s obzirom da histogram ima lijep oblike i Q-Q graf nema teške repove. Vizualne pretpostavke dodatno potvrđuju Lillieforsov test na normalnost u kojem na relativno velika p vrijednost govori kako ne možemo odbaciti H_0 = Podaci dolaze iz normalne distribucije. S druge strane pretpostavka o homogenosti varijance pokazuje jasnu polinomijalnu(kvadratnu) zavisnost.

U sljedećem koraku nam preostaje konstruirati model za sve prethodno spomenute regresore.

```
fitMulti = lm(weightkg ~ waistcircumference + bicepscircumferenceflexed + chestcircumference +
+ thighcircumference + stature + Age + Gender_Male, data = antrData)

summary(fitMulti)

##
## Call:
## lm(formula = weightkg ~ waistcircumference + bicepscircumferenceflexed +
##     chestcircumference + thighcircumference + stature + Age +
##     Gender_Male, data = antrData)
##
```

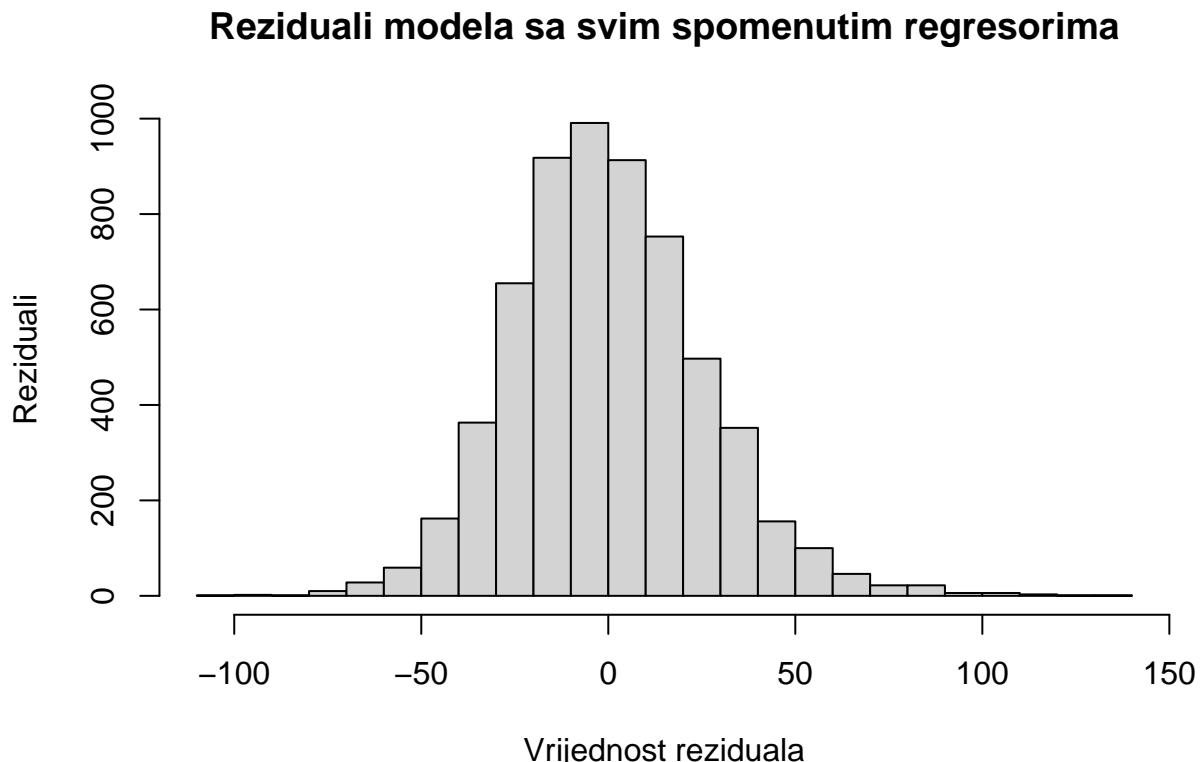
```

## Residuals:
##      Min       1Q   Median      3Q     Max
## -102.949  -17.186  -1.709  15.334 138.765
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -1.376e+03 8.320e+00 -165.396 < 2e-16 ***
## waistcircumference  3.019e-01 7.328e-03  41.196 < 2e-16 ***
## bicepscircumference 6.313e-01 1.831e-02  34.471 < 2e-16 ***
## chestcircumference   3.795e-01 8.987e-03  42.228 < 2e-16 ***
## thighcircumference   7.282e-01 1.221e-02  59.627 < 2e-16 ***
## stature              4.926e-01 5.129e-03  96.028 < 2e-16 ***
## Age                  -3.213e-01 4.292e-02 -7.487 8.03e-14 ***
## Gender_Male          8.904e+00 1.282e+00   6.946 4.16e-12 ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.59 on 6060 degrees of freedom
## Multiple R-squared:  0.9733, Adjusted R-squared:  0.9733
## F-statistic: 3.157e+04 on 7 and 6060 DF, p-value: < 2.2e-16

```

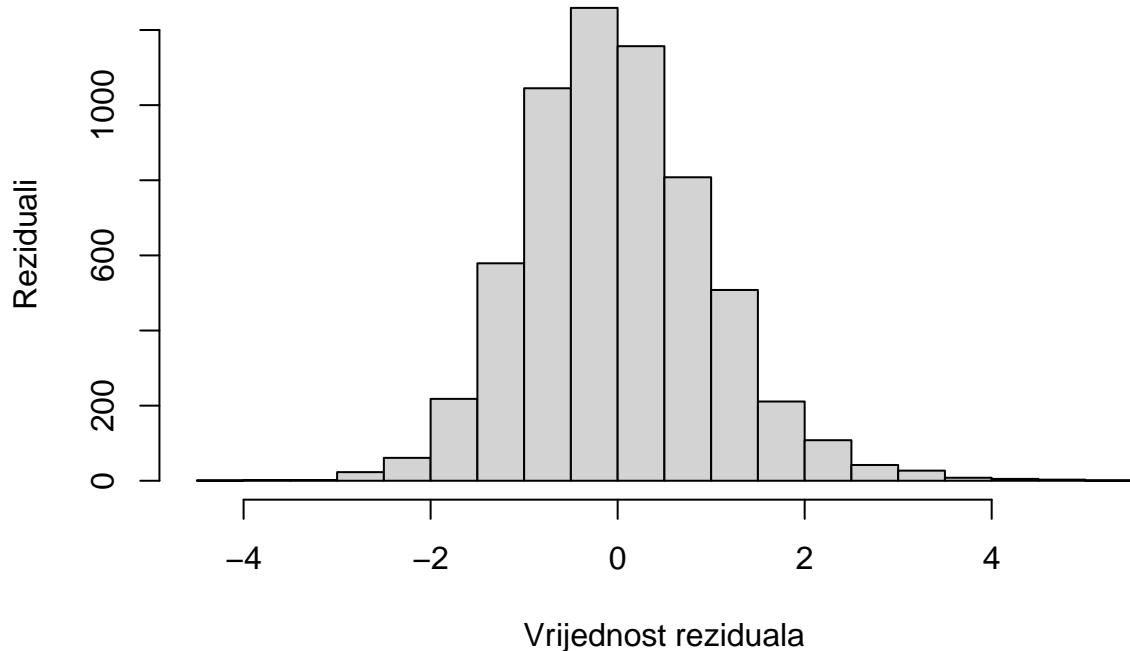
Ovaj model isto tako postiže i više nego zadovoljavajuće rezultate. Svi koeficijenti su statistički značajni te je $R^2_{adj} = R^2$ što nam ukazuje na to da nemamo suvišnih regresora. Nadalje vrijednosti R^2_{adj} i R^2 je 0.9733 što nam govori kako model objašnjava 97.33% varijance u podacima. Preostaje nam ispitati pretpostavku o normalnosti.

```
hist(fitMulti$residuals, xlab = "Vrijednost reziduala", ylab = "Reziduali", breaks = 30, main = "Reziduali modela sa svim spomenutim regresorima")
```

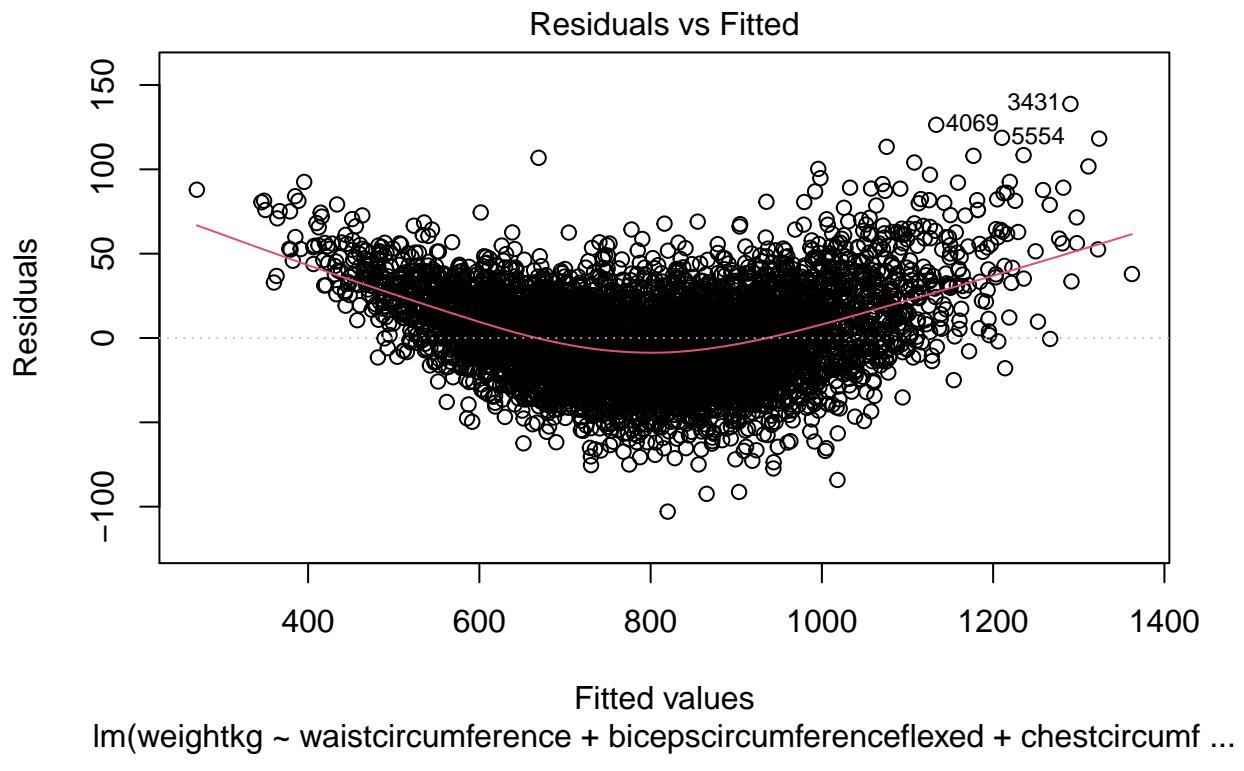


```
hist(rstandard(fitMulti), xlab = "Vrijednost reziduala", ylab = "Reziduali", breaks = 30, main = "Standar
```

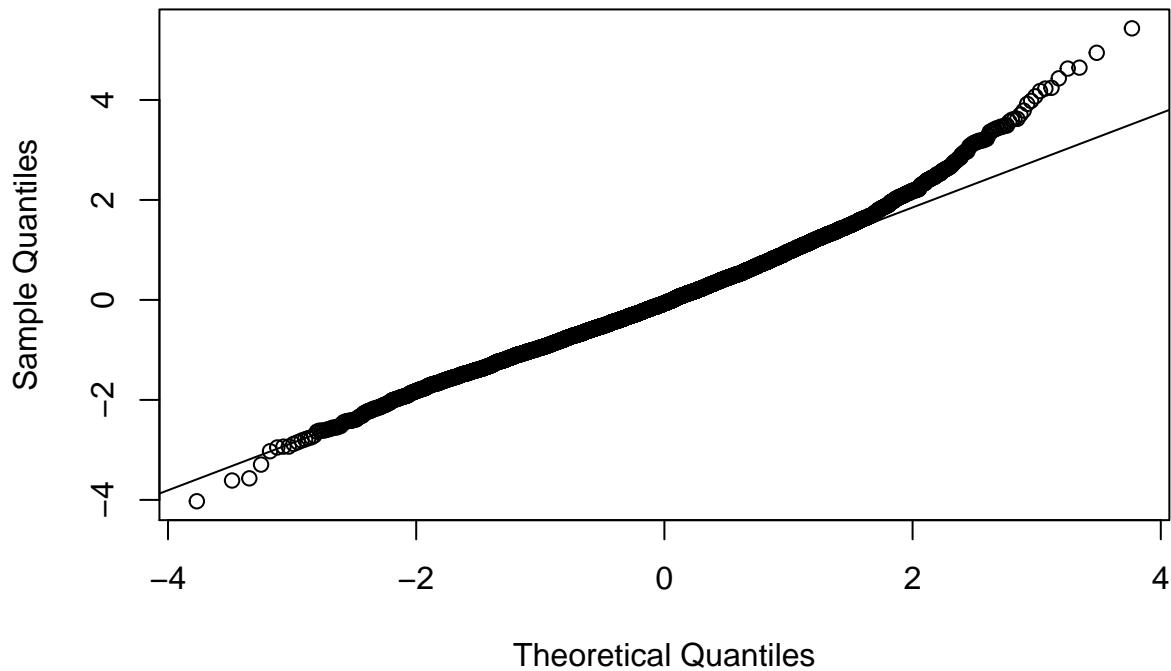
Standardizirani reziduali modela sa svim spomenutim regresorima



```
plot(fitMulti, which = 1)
```



Normal Q-Q Plot



```
require(nortest)
lillie.test(rstandard(fitMulti))

##
## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: rstandard(fitMulti)
## D = 0.028482, p-value = 4.091e-12
```

Vidimo kako nam unatoč povećanoj R^2 vrijednosti pretpostavka o normalnostivije ne vrijedi. Tu su najindikativniji Q-Q graf i Lillieforsov test. Q-Q plot ima težak rep u gornjim kvartilima, a Lillieforsov test izbacuje jako malu p vrijednost što nam nalaže odbacivanje H_0 = “Podaci dolaze iz normalne distribucije”. Nadalje pretpostavka o homogenosti varijance nije zadovoljena s obzirom da residuali i dalje pokazuju jasan polinomijalni trend. Sada bi se mogli pozvati na robusnost t-testa na normalnost ali iz prethodnih analiza na ovom projektu znamo koliko varijacije unose razlike koje postoje među spolovima. Stoga ćemo izgraditi dva odvojena modela za muškarce i žene te analizirati kako to utječe na normalnost i homogenost varijanci.

```
antrData_male <- antrData[antrData$Gender_Male == 1,]
antrData_female <- antrData[antrData$Gender_Male == 0,]

fitMulti_male = lm(weightkg ~ waistcircumference + bicepscircumferenceflexed +
                     + thighcircumference + stature + Age, data = antrData_male)

summary(fitMulti_male)

##
## Call:
## lm(formula = weightkg ~ waistcircumference + bicepscircumferenceflexed +
```

```

##      chestcircumference + thighcircumference + stature + Age,
##      data = antrData_male)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -90.803 -16.192  -0.807  14.571 110.707
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -1.463e+03  9.916e+00 -147.53 < 2e-16 ***
## waistcircumference     3.506e-01  8.815e-03   39.77 < 2e-16 ***
## bicepscircumferenceflexed 6.192e-01  2.018e-02   30.68 < 2e-16 ***
## chestcircumference      4.271e-01  1.094e-02   39.02 < 2e-16 ***
## thighcircumference      6.960e-01  1.480e-02   47.02 < 2e-16 ***
## stature                  5.071e-01  5.610e-03   90.40 < 2e-16 ***
## Age                     -3.742e-01  5.003e-02   -7.48 9.04e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.52 on 4075 degrees of freedom
## Multiple R-squared:  0.9727, Adjusted R-squared:  0.9726
## F-statistic: 2.418e+04 on 6 and 4075 DF,  p-value: < 2.2e-16
fitMulti_female = lm(weightkg ~ waistcircumference + bicepscircumferenceflexed + chestcircumference
                      + thighcircumference + stature + Age, data = antrData_female)

summary(fitMulti_female)

##
## Call:
## lm(formula = weightkg ~ waistcircumference + bicepscircumferenceflexed +
##      chestcircumference + thighcircumference + stature + Age,
##      data = antrData_female)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -67.669 -14.473  0.185  13.696  97.655
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -1.174e+03  1.235e+01 -95.044 < 2e-16 ***
## waistcircumference      1.643e-01  1.002e-02  16.388 < 2e-16 ***
## bicepscircumferenceflexed 7.643e-01  2.987e-02  25.585 < 2e-16 ***
## chestcircumference      3.054e-01  1.138e-02  26.847 < 2e-16 ***
## thighcircumference      7.199e-01  1.624e-02  44.335 < 2e-16 ***
## stature                  4.640e-01  7.837e-03  59.203 < 2e-16 ***
## Age                     -4.093e-01  6.083e-02  -6.728 2.25e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.1 on 1979 degrees of freedom
## Multiple R-squared:  0.9632, Adjusted R-squared:  0.9631
## F-statistic: 8633 on 6 and 1979 DF,  p-value: < 2.2e-16

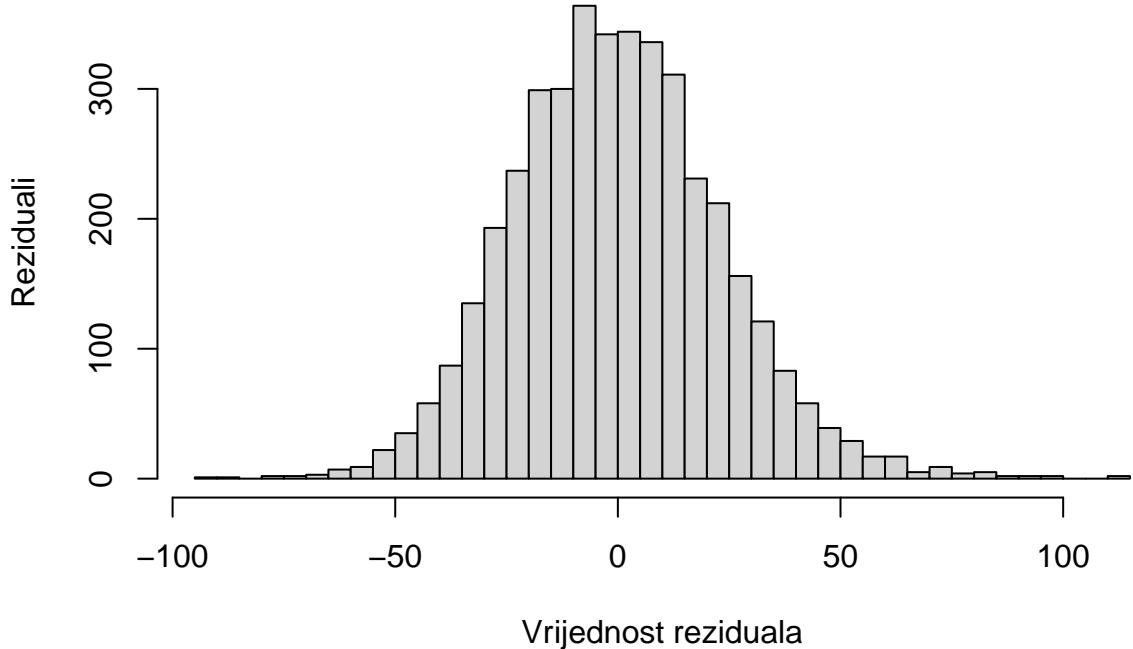
```

Na prvi pogled nije se puno promjenilo ali primarni razlog ovog koraka je bilo popravljanje pretpostavke o

normalnosti tako da u sljedećem koraku provjeravamo upravo to.

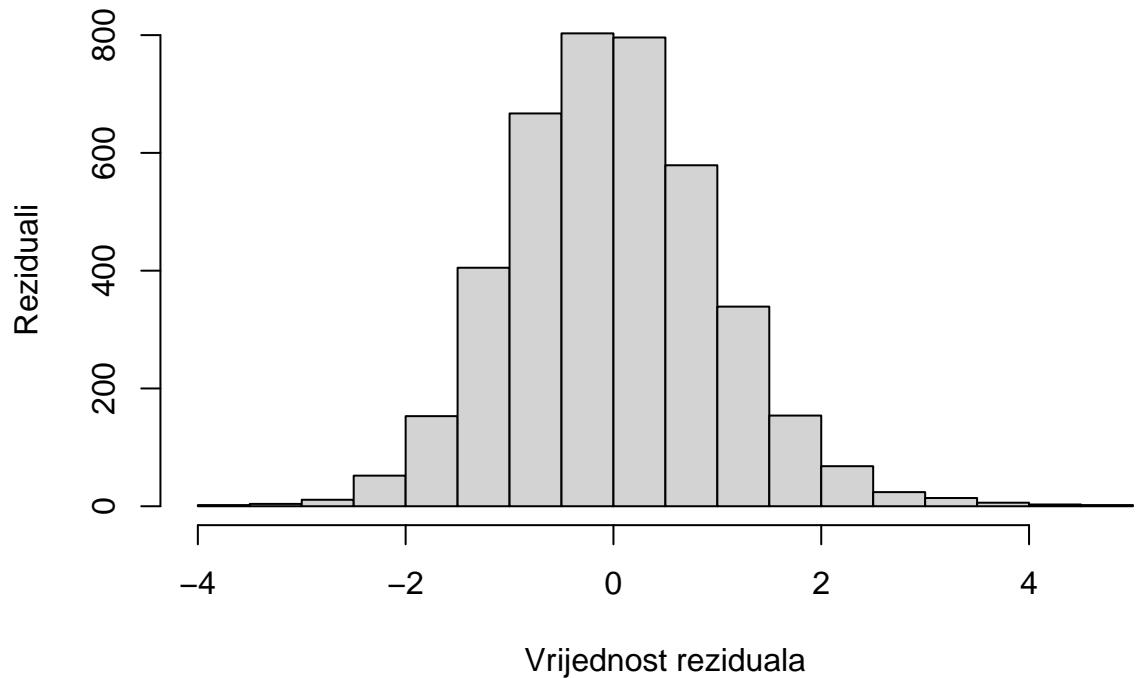
```
hist(fitMulti_male$residuals, xlab = "Vrijednost reziduala", ylab = "Reziduali", breaks = 30, main = "Reziduali modela sa svim spomenutim regresorima za muškarce")
```

Reziduali modela sa svim spomenutim regresorima za muškarce

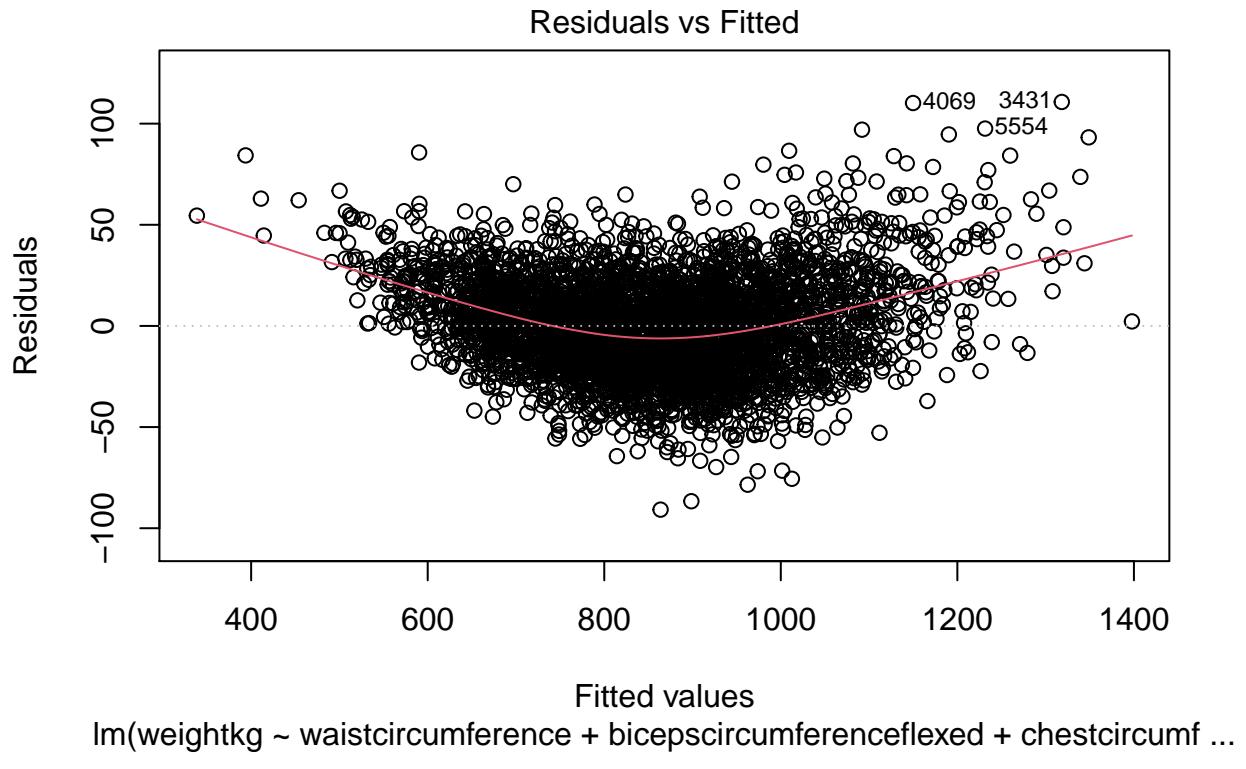


```
hist(rstandard(fitMulti_male), xlab = "Vrijednost reziduala", ylab = "Reziduali", breaks = 30, main = "Reziduali modela sa svim spomenutim regresorima za muškarce")
```

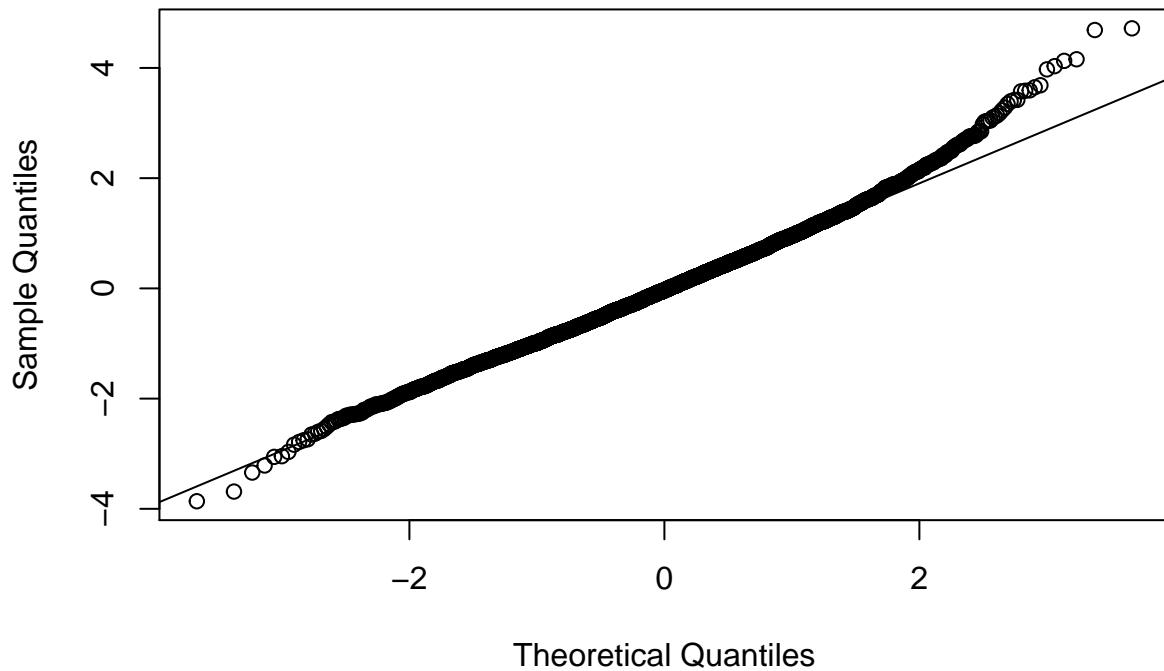
andardizirani reziduali modela sa svim spomenutim regresorima za mu



```
plot(fitMulti_male, which = 1)
```



Normal Q-Q Plot

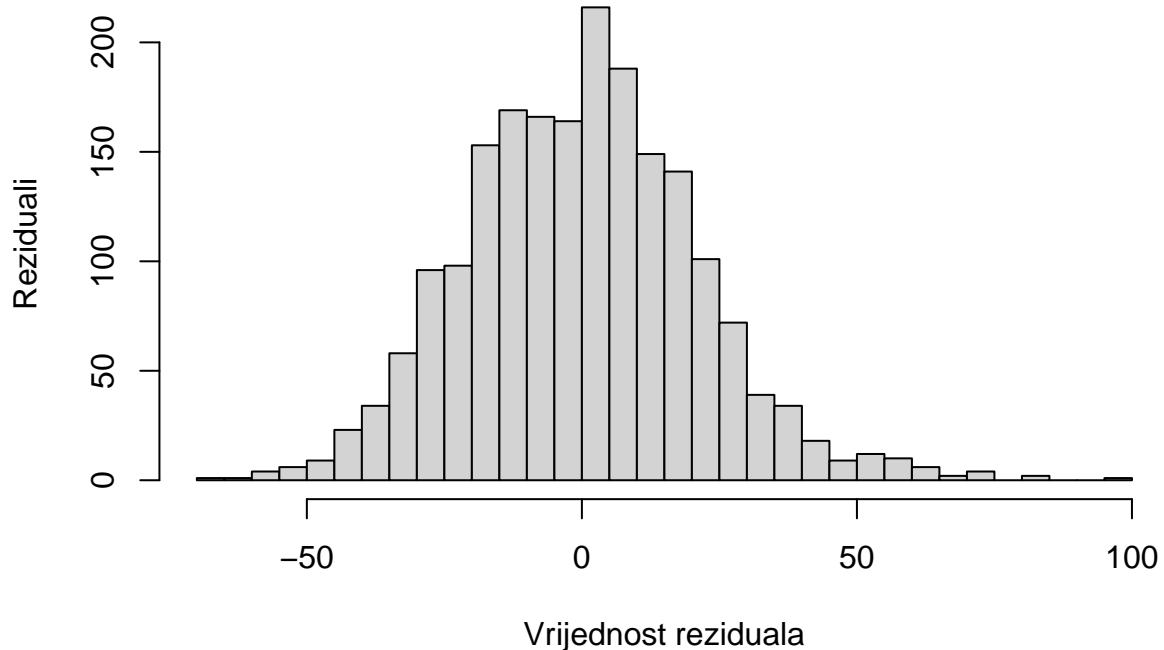


```
require(nortest)
lillie.test(rstandard(fitMulti_male))
```

```
##
## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: rstandard(fitMulti_male)
## D = 0.018782, p-value = 0.002171
```

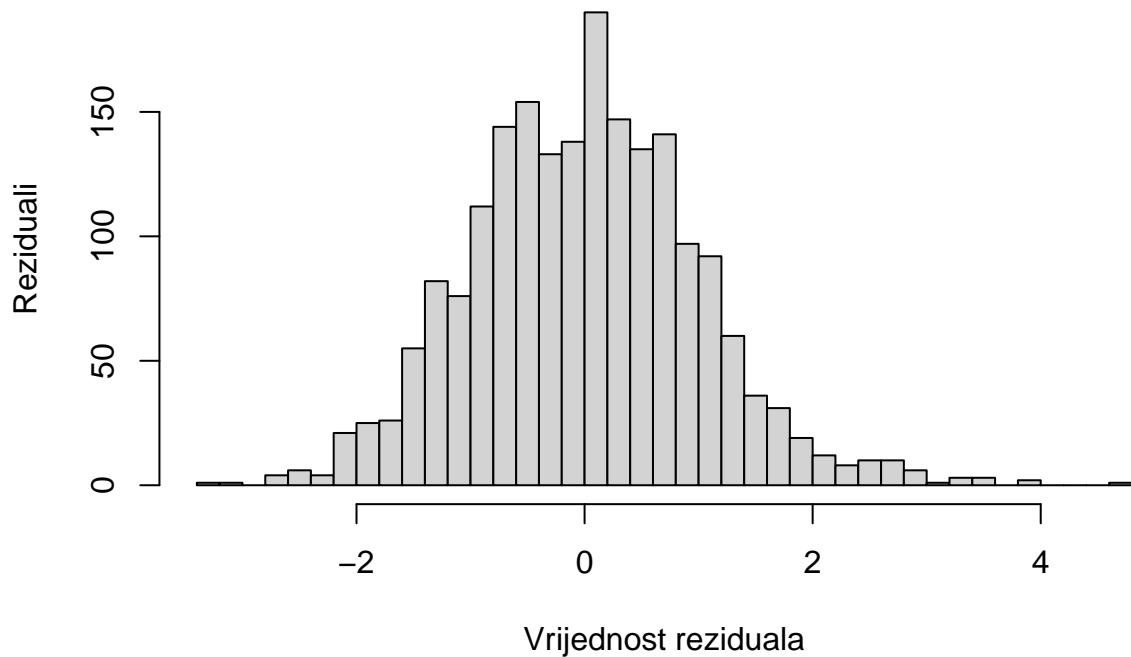
```
hist(fitMulti_female$residuals, xlab = "Vrijednost reziduala", ylab = "Reziduali", breaks = 30, main =
```

Reziduali modela sa svim spomenutim regresorima za žene

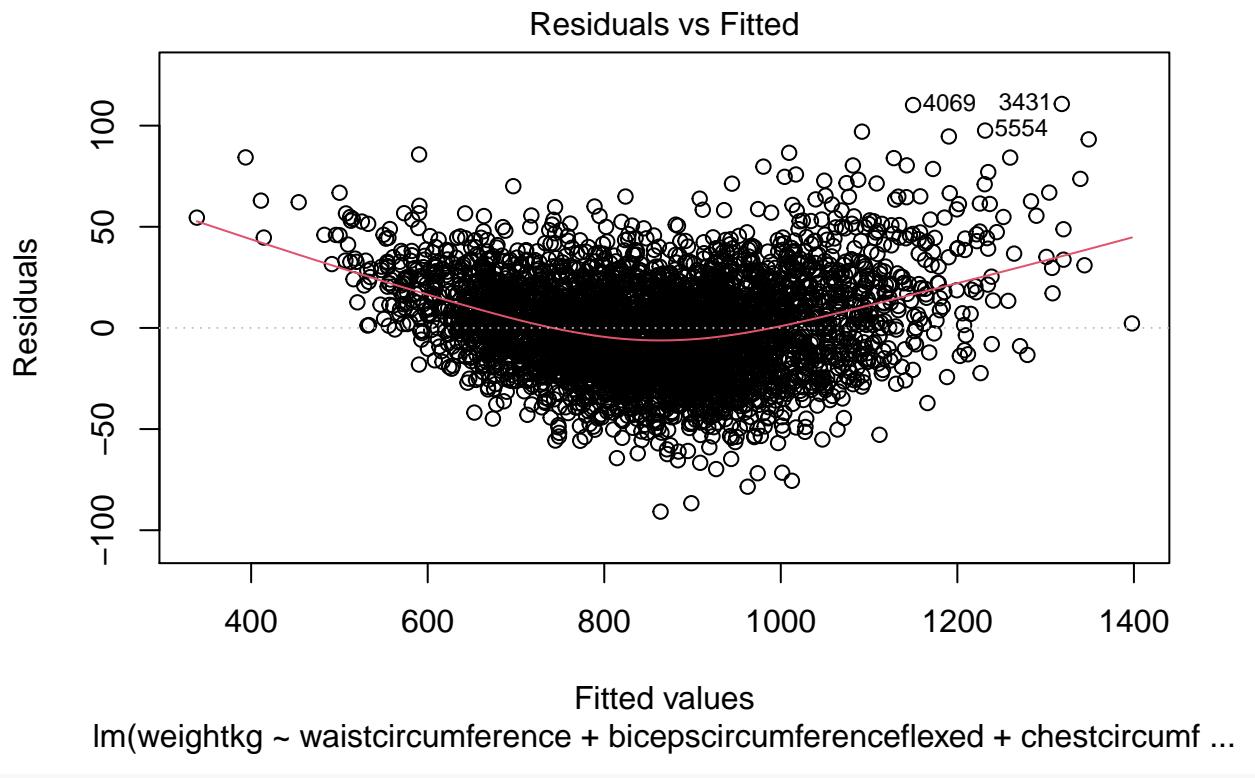


```
hist(rstandard(fitMulti_female), xlab = "Vrijednost reziduala", ylab = "Reziduali", breaks = 30, main =
```

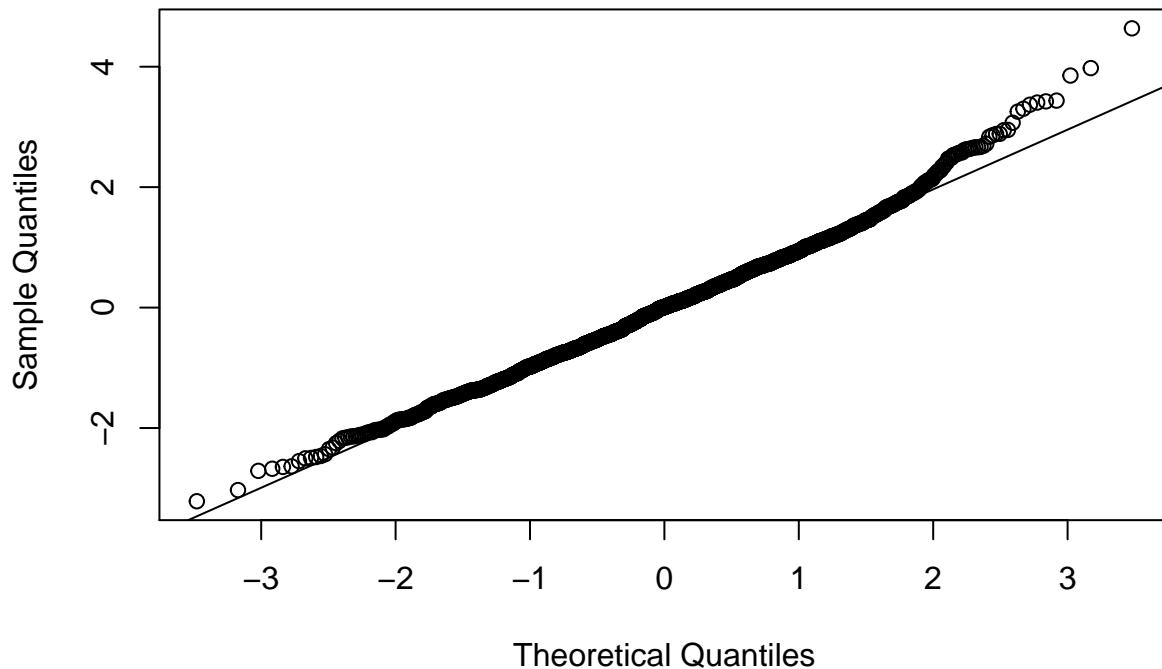
Standardizirani reziduali modela sa svim spomenutim regresorima za žene



```
plot(fitMulti_male, which = 1)
```



Normal Q-Q Plot



```
require(nortest)
lillie.test(rstandard(fitMulti_female))
```

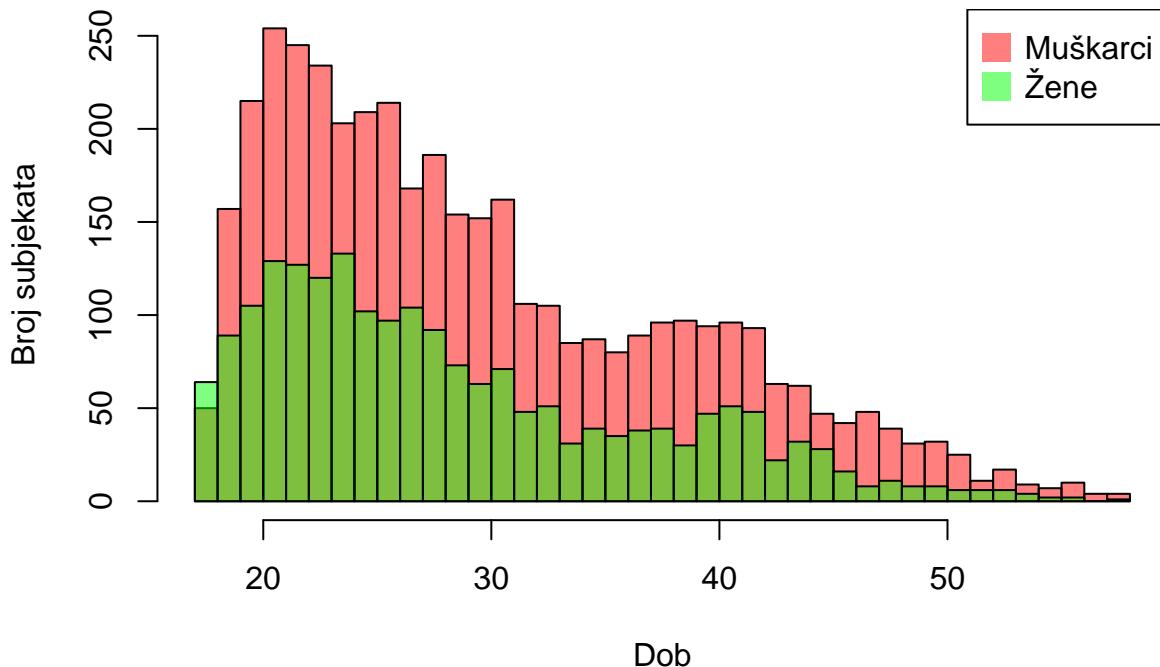
```
##
## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: rstandard(fitMulti_female)
## D = 0.017626, p-value = 0.1415
```

Iako je p vrijednost u oba slučaja značajno narasla, dapače za žensku populaciju ne odbacujemo H_0 , problem je i dalje prisutan u muškoj populaciji. Nadalje homogenost varijanci se nije popravila te i dalje indicira polinomijalnu ovisnost. Ipak ako razmislimo o tome kakvi ljudi služe u vojsci lako možemo doći do zaključka o tome gdje bi se problem mogao nalaziti.

```
hist(antrData_male$Age, breaks = 30 ,col = rgb(1,0,0,0.5), xlab = "Dob", ylab = "Broj subjekata", main =
hist(antrData_female$Age, breaks = 30, col =rgb(0,1,0,0.5), add = T)

legend('topright', legend = c("Muškarci", "Žene"), col = c(rgb(1,0,0,0.5), rgb(0,1,0,0.5)), pt.cex=2, p
```

Dob među muškom populacijom



I upravo ovdje nailazimo na problem, kako je biti vojnik izuzetno naporan fizički posao populacija je većinski sastavljena od mladih ljudi. Kao što je vidljivo iz histograma distribucija dobi među populacijom nije normalna tje vjerovatni uzročnik pada našeg modela na testu normalnosti.

Kako bi to provjerili izgraditi ćemo model koji ne uključuje dob te ga testirati na normalnost. Zbog male R^2 vrijednosti linearne modela koji je uključivao samo dob ne očekujemo preveliki gubitak u tom području.

```
fitMulti_male = lm(weightkg ~ waistcircumference + bicepscircumferenceflexed + chestcircumference
+ thighcircumference + stature, data = antrData_male)

summary(fitMulti_male)

##
## Call:
## lm(formula = weightkg ~ waistcircumference + bicepscircumferenceflexed +
##     chestcircumference + thighcircumference + stature, data = antrData_male)
##
## Residuals:
##      Min      1Q   Median      3Q      Max 
## -90.453 -16.267 - 0.584  14.993 111.686 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -1.472e+03 9.906e+00 -148.60 <2e-16 ***
## waistcircumference 3.266e-01 8.266e-03 39.51 <2e-16 ***
## bicepscircumferenceflexed 6.006e-01 2.016e-02 29.79 <2e-16 ***
## chestcircumference 4.228e-01 1.100e-02 38.42 <2e-16 ***
```

```

## thighcircumference      7.407e-01  1.363e-02   54.33  <2e-16 ***
## stature                  5.093e-01  5.640e-03   90.30  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.68 on 4076 degrees of freedom
## Multiple R-squared:  0.9723, Adjusted R-squared:  0.9723
## F-statistic: 2.862e+04 on 5 and 4076 DF,  p-value: < 2.2e-16

fitMulti_female = lm(weightkg ~ waistcircumference + bicepscircumferenceflexed + chestcircumference
                      + thighcircumference + stature, data = antrData_female)

summary(fitMulti_female)

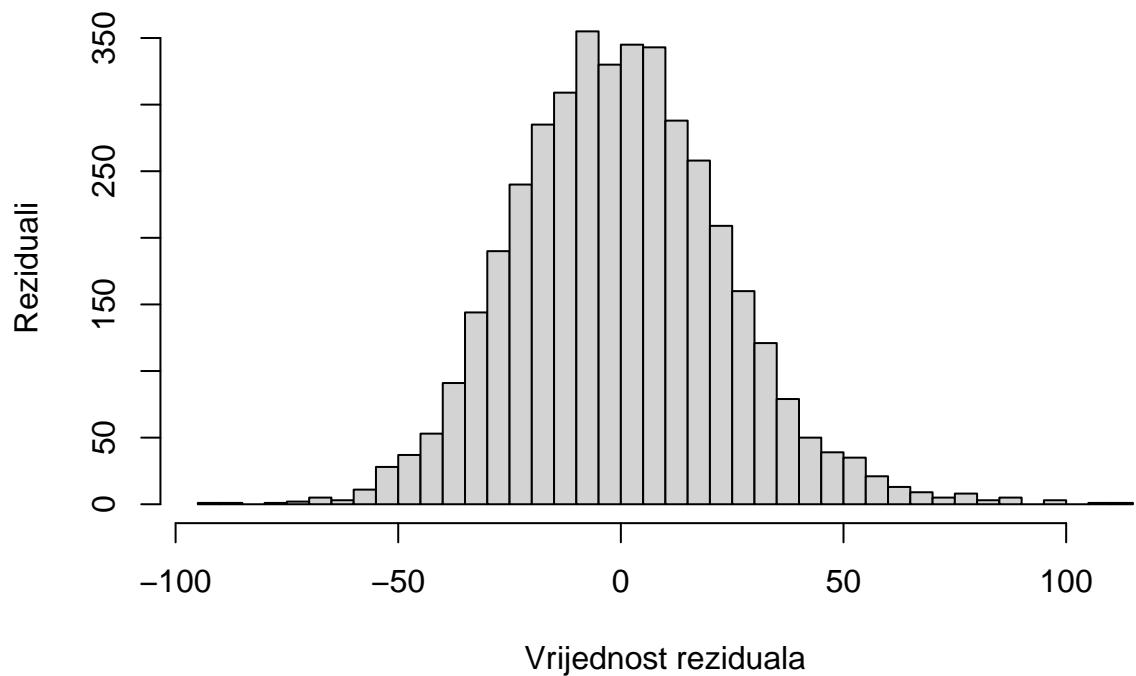
##
## Call:
## lm(formula = weightkg ~ waistcircumference + bicepscircumferenceflexed +
##     chestcircumference + thighcircumference + stature, data = antrData_female)
##
## Residuals:
##    Min      1Q  Median      3Q      Max
## -72.006 -14.769   0.122  13.998 101.542
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.179e+03  1.246e+01 -94.64  <2e-16 ***
## waistcircumference 1.591e-01  1.010e-02  15.74  <2e-16 ***
## bicepscircumferenceflexed 7.482e-01  3.011e-02  24.85  <2e-16 ***
## chestcircumference  2.956e-01  1.141e-02  25.91  <2e-16 ***
## thighcircumference 7.315e-01  1.633e-02  44.81  <2e-16 ***
## stature          4.671e-01  7.910e-03  59.05  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.34 on 1980 degrees of freedom
## Multiple R-squared:  0.9624, Adjusted R-squared:  0.9623
## F-statistic: 1.012e+04 on 5 and 1980 DF,  p-value: < 2.2e-16

```

Vidimo kako je smanjenje R^2 vrijednosti zanemarivo tako da je isključivanje dobi iz modela i s te strane opravdano. Provjerimo sad normlanost naših novih modela.

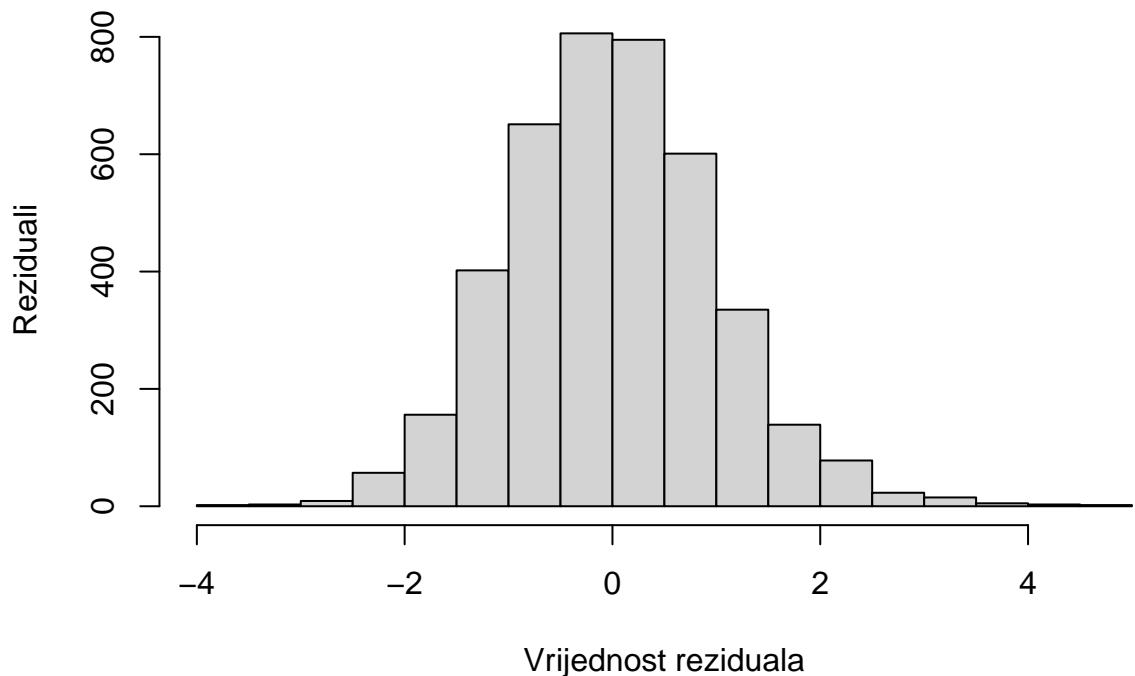
```
hist(fitMulti_male$residuals, xlab = "Vrijednost reziduala", ylab = "Reziduali", breaks = 30, main = "Rezidualni histogram za muške")
```

Reziduali modela sa svim spomenutim regresorima za muškarce

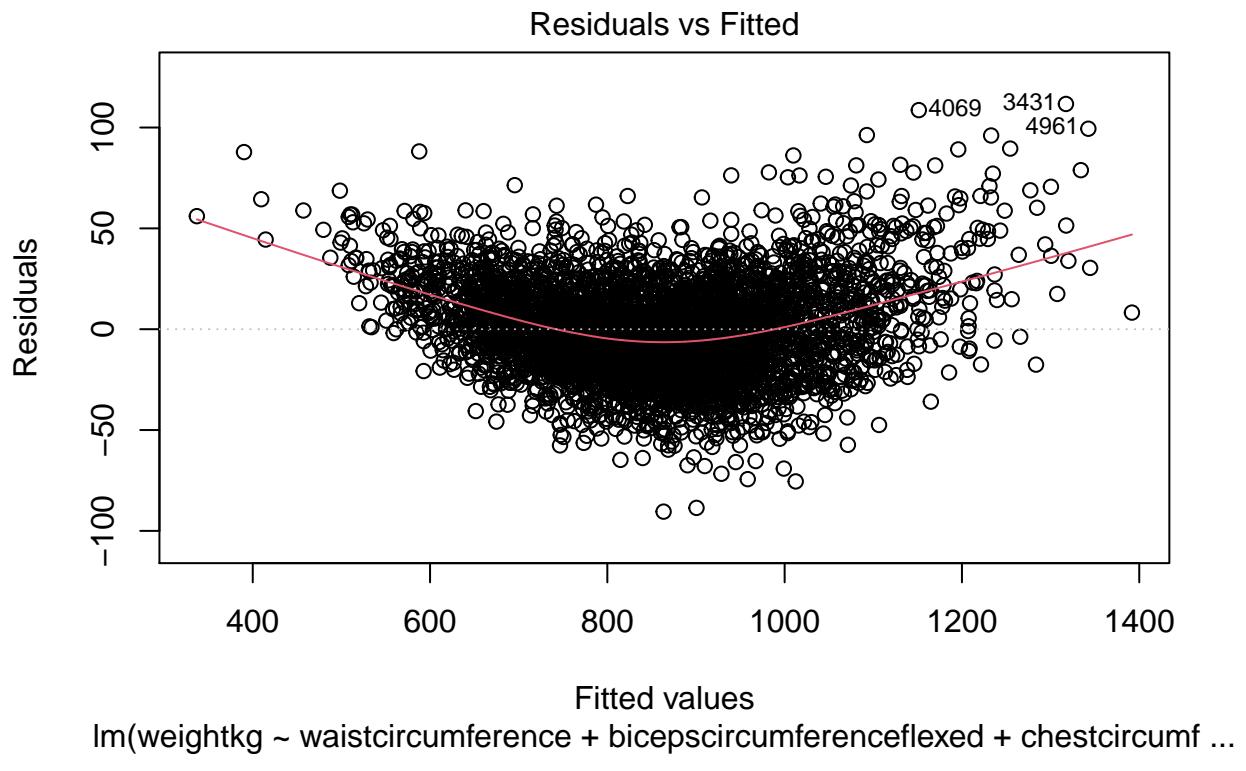


```
hist(rstandard(fitMulti_male), xlab = "Vrijednost reziduala", ylab = "Reziduali", breaks = 30, main = "")
```

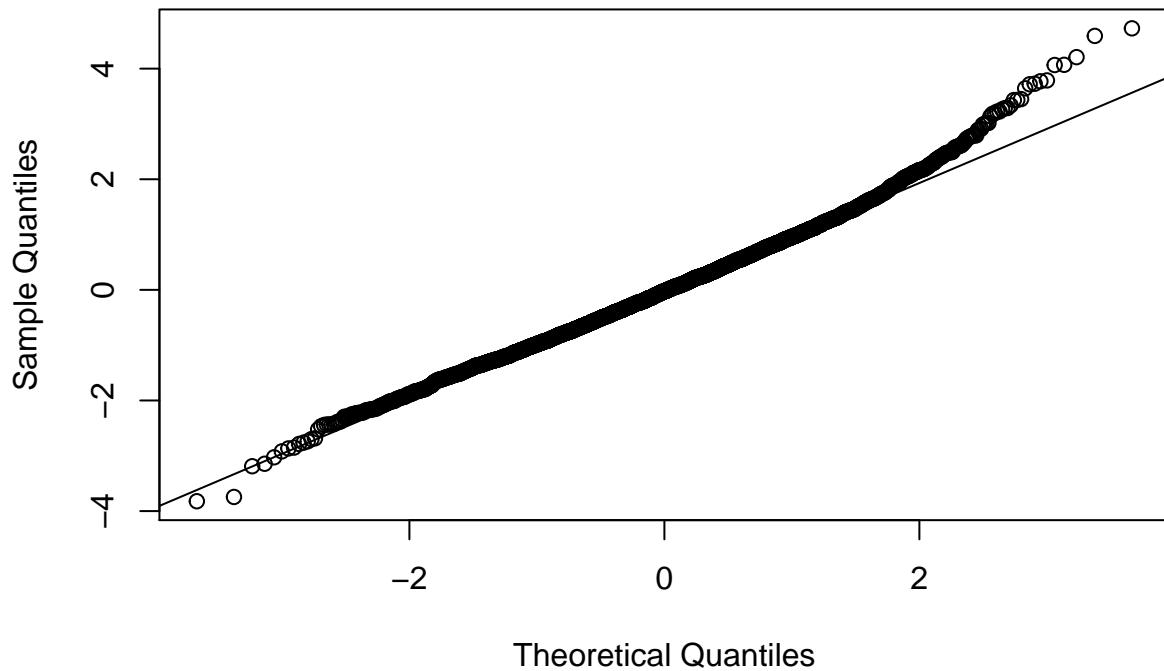
andardizirani reziduali modela sa svim spomenutim regresorima za mu



```
plot(fitMulti_male, which = 1)
```



Normal Q-Q Plot

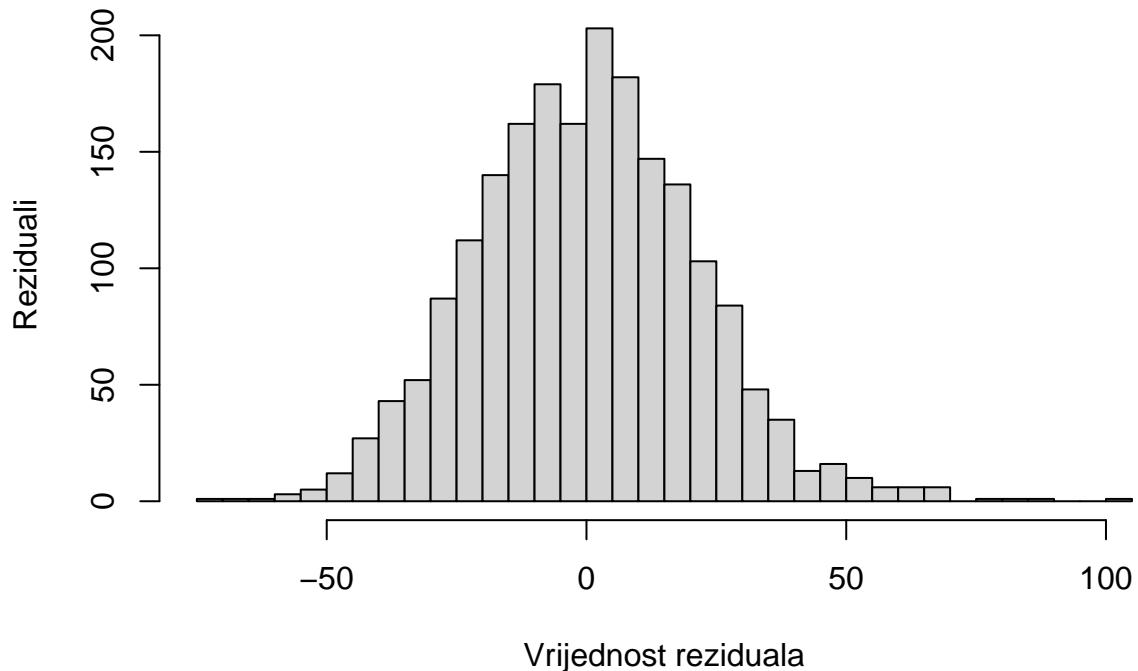


```
require(nortest)
lillie.test(rstandard(fitMulti_male))
```

```
##
## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: rstandard(fitMulti_male)
## D = 0.020825, p-value = 0.0003487
```

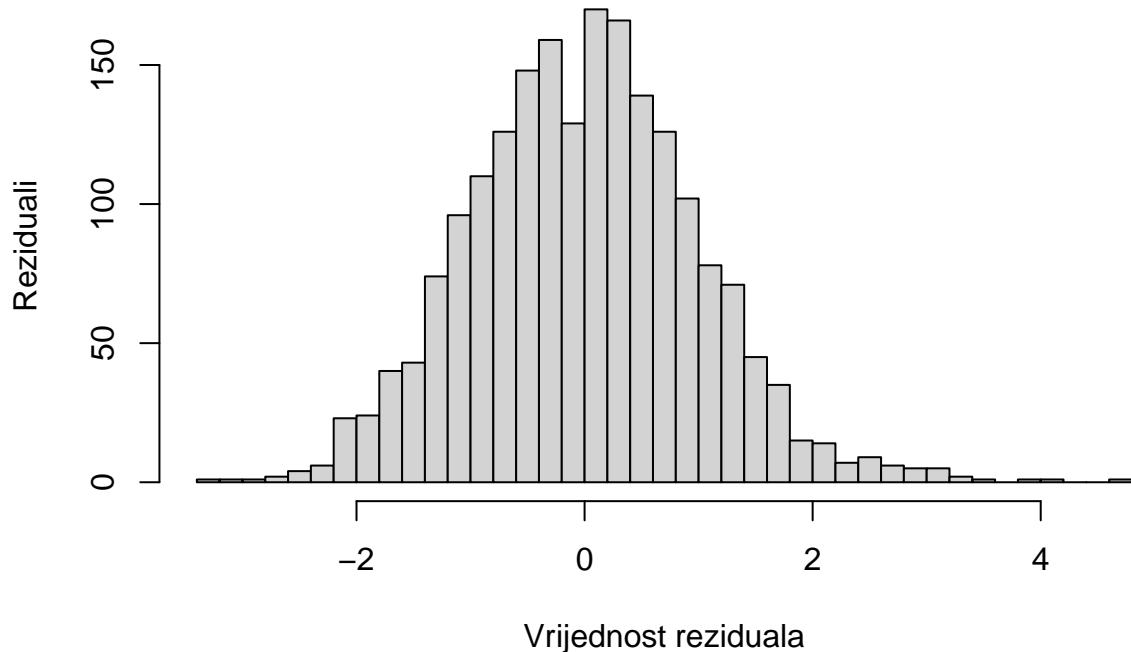
```
hist(fitMulti_female$residuals, xlab = "Vrijednost reziduala", ylab = "Reziduali", breaks = 30, main =
```

Reziduali modela sa svim spomenutim regresorima za žene

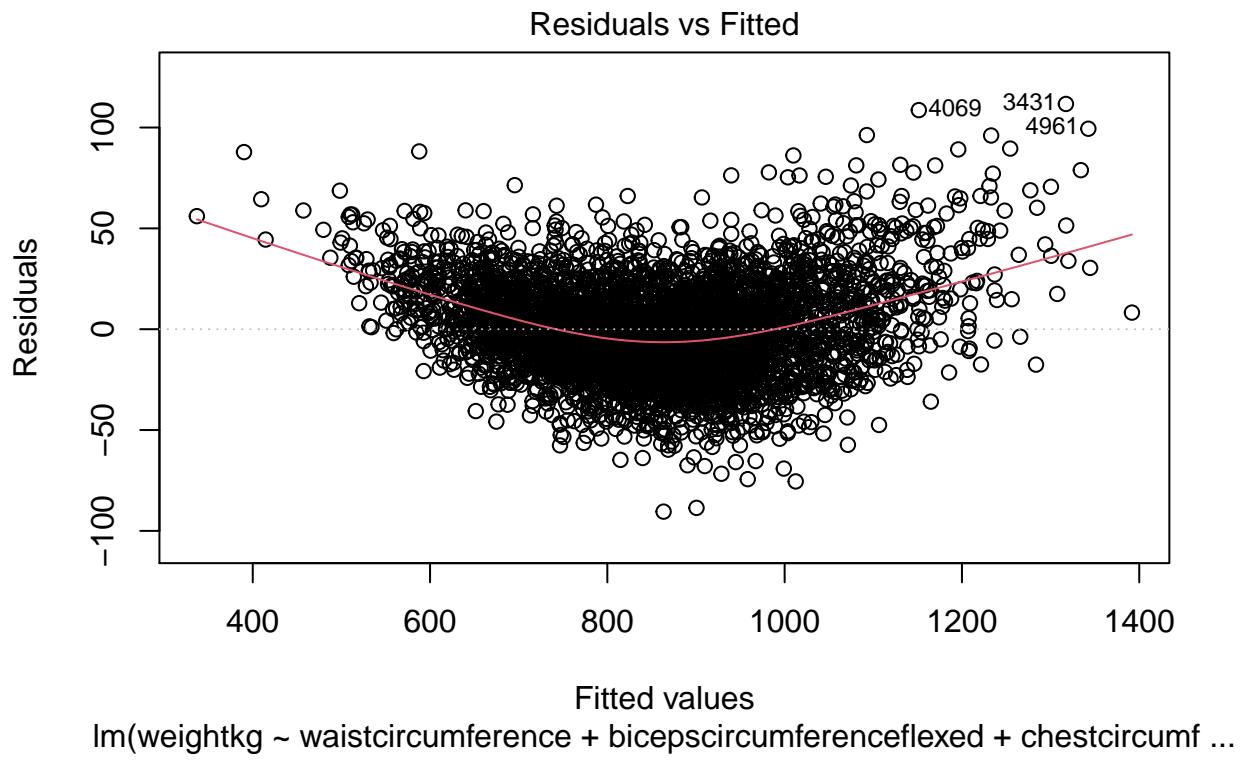


```
hist(rstandard(fitMulti_female), xlab = "Vrijednost reziduala", ylab = "Reziduali", breaks = 30, main =
```

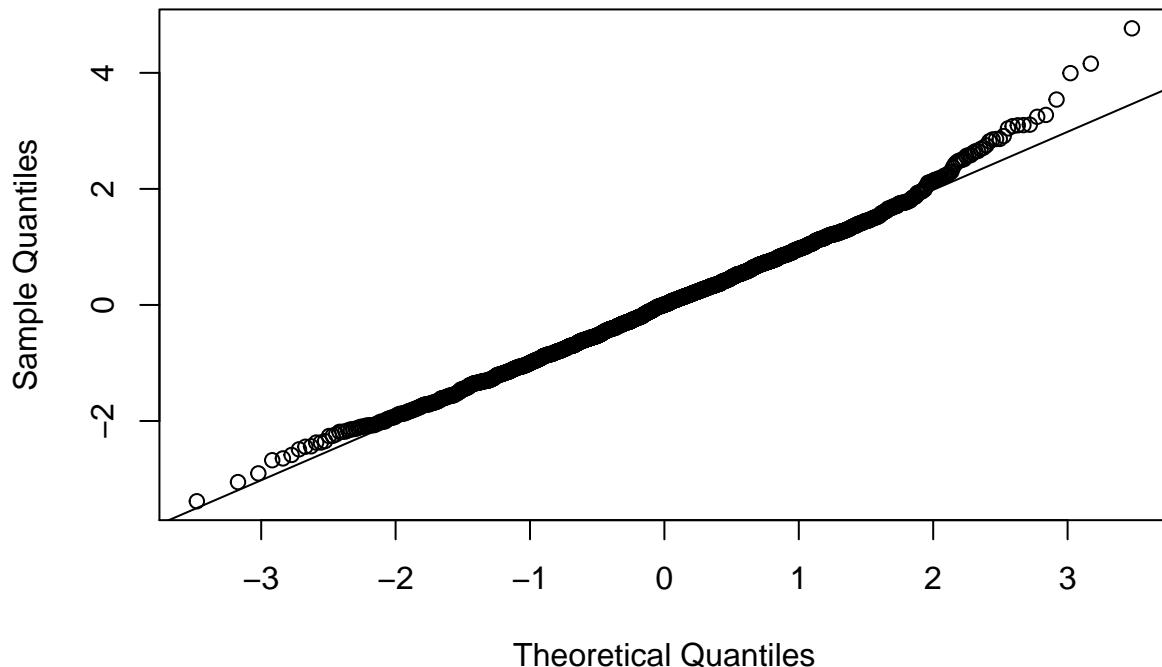
Standardizirani reziduali modela sa svim spomenutim regresorima za žene



```
plot(fitMulti_male, which = 1)
```



Normal Q-Q Plot



```
require(nortest)
lillie.test(rstandard(fitMulti_female))

##
## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: rstandard(fitMulti_female)
## D = 0.017049, p-value = 0.1745
```

Tu ćemo stati s daljnjim poboljšavanje modela. Za ženski dio populacije ne odbacujemo H_0 te zaključujemo kako nema dovoljno dokaza da bi odbacili hipotezu o normalnosti. Nažalost za muški dio populacije ne možemo donjeti isti zaključak na temelju Lillieforsovog testa. S druge strane možemo se pozvati na robusnost t-testa na normalnost, s obzirom na to da su histogrami relativno urednog oblika te nemaju teških repova. Ipak ovo je indikacija kako u težini među muškarcima postoji varijacija koju naš model ne objašnjava. Isto tako daljna poboljšanja modela su svakako moguća s obzirom na graf reziduala i prilagođenih vrijednosti iz kojeg jasno vidimo kako pretpostavka o homogenosti varijance nije zadovoljena te reziduali pokazuju polinomijalnu ovisnost.

Zaključak

Izgradili smo dva multivarijatna liniarna modela za predviđanje težine osobe. Modeli su odvojeni po spolu i oba modela objašnjavaju preko 95% varijance u podacima. Ovakav prediktivni model bi mogao biti izuzetno koristan u medicini pogotovo u područjima gdje vaga nije dostupna. Iz tog razloga je svih 5 regresora moguće dobiti korištenjem jednostavnog krojačkog metra. Isto tako podaci ne zahtjevaju nikakvo ispitivanje bolesnika, što može biti izuzetno korisno u hitnim situacijama kada je pacijent bez svijesti.

Utjecaj dominantne ruke na antropometrijske mjere

Motivacija

Za očekivati je da ljudi koji pišu desnom rukom imaju veću desnu ruku od onih koji pišu lijevom. Na raspolaganju imamo listu opsega desnog bicepsa koju možemo podijeliti na one od dešnjaka, lijevakata, i onih koji su podjednako spretni s obje ruke. S obzirom da je vojska profesija u kojoj je fizička spremna izrazito važna, a za očekivati je za ljude koji su izrazito fizički spremni da im je mišićna masa dobro izbalansirana. Ova analiza može ukazati na moguće manjkavosti u treningu vojnika.

Istraživanje

```
antrData = read.csv("ANSUR_II_data.csv")
summary(antrData$bicepscircumferenceflexed)

##      Min. 1st Qu. Median     Mean 3rd Qu.      Max.
##    216.0   311.0  341.0   340.9   370.0   490.0

ansurLeft <- antrData[ which(antrData$WritingPreference=='Left hand'),]
ansurRight <- antrData[ which(antrData$WritingPreference=='Right hand'),]
ansurAmbi <- antrData[ which(antrData$WritingPreference=='Either hand (No preference)'),]

summary(ansurLeft$bicepscircumferenceflexed)

##      Min. 1st Qu. Median     Mean 3rd Qu.      Max.
##    244.0   311.0  341.0   340.4   368.2   472.0

summary(ansurRight$bicepscircumferenceflexed)

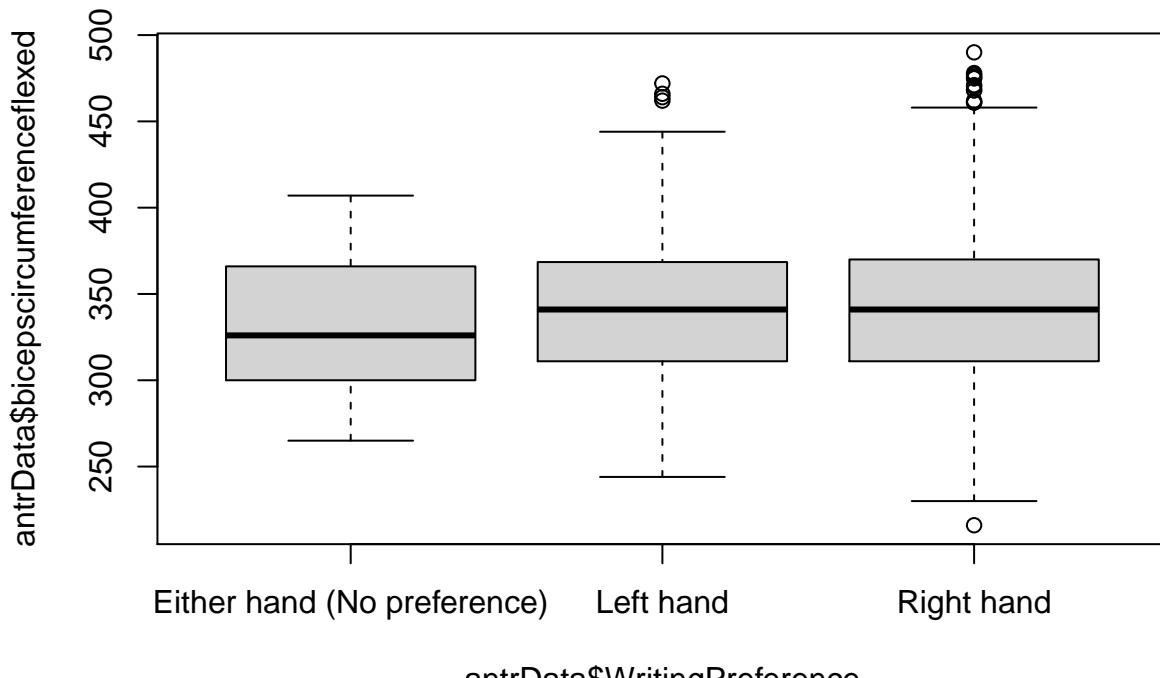
##      Min. 1st Qu. Median     Mean 3rd Qu.      Max.
##    216.0   311.0  341.0   341.1   370.0   490.0

summary(ansurAmbi$bicepscircumferenceflexed)

##      Min. 1st Qu. Median     Mean 3rd Qu.      Max.
##    265.0   301.2  326.0   332.3   365.5   407.0

boxplot(antrData$bicepscircumferenceflexed~antrData$WritingPreference,
        main = "Opseg bicepsa s ")
```

Opseg bicepsa s



```
summary(ansurLeft$bicepscircumferenceflexed)
```

```
##      Min. 1st Qu. Median     Mean 3rd Qu.    Max.
##  244.0   311.0  341.0  340.4  368.2  472.0
```

```
summary(ansurRight$bicepscircumferenceflexed)
```

```
##      Min. 1st Qu. Median     Mean 3rd Qu.    Max.
##  216.0   311.0  341.0  341.1  370.0  490.0
```

```
summary(ansurAmbi$bicepscircumferenceflexed)
```

```
##      Min. 1st Qu. Median     Mean 3rd Qu.    Max.
##  265.0   301.2  326.0  332.3  365.5  407.0
```

Vizualizacija sugerira da nema značajne razlike između opsega desnog bicepsa ljevaka i dešnjaka. Vojnici koji nemaju samo jednu dominantnu ruku imaju malo manji prosjek opsega bicepsa u ovom uzorku. Osobe čije su mjere u pitanju rade intenzivne simetrične vježbe i važna im je snaga obje ruke. Zbog takvog načina treniranja se smanjuje razmjer između veličine dominantne i nedominantne ruke. Za generalnu populaciju bi razlika mogla biti veća.

Pogledajmo raspodjelu korištenja ruku u uzorku.

```
nrow(ansurLeft)
```

```
## [1] 656
```

```
nrow(ansurRight)
```

```
## [1] 5350
```

```

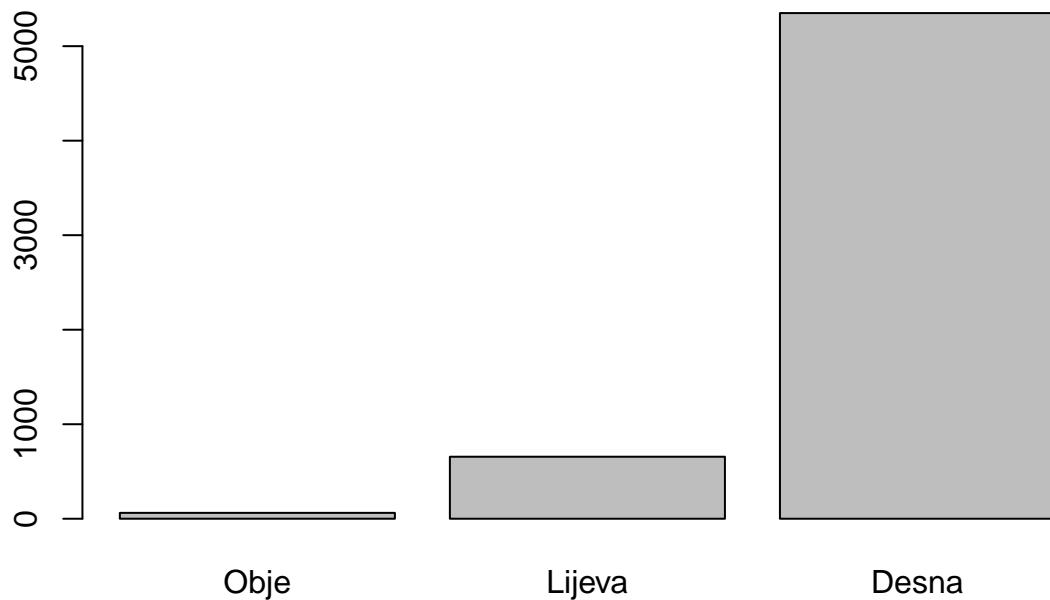
nrow(ansurAmbi)

## [1] 62

counts <- table(antrData$WritingPreference)
barplot(counts, main="Dominantna ruka", horiz=FALSE,
  names.arg=c("Obje", "Lijeva", "Desna"))

```

Dominantna ruka

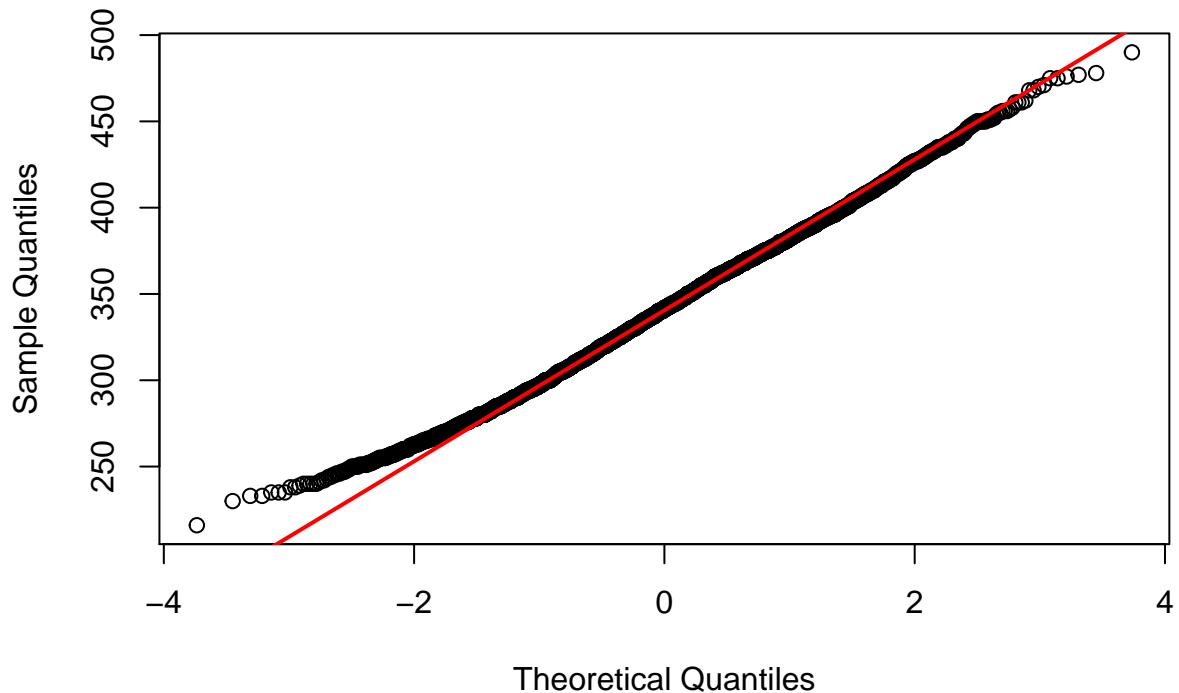


```

require(nortest)
qqnorm(ansurRight$bicepscircumferenceflexed, main="Biceps circumference, right-handed")
qline(ansurRight$bicepscircumferenceflexed, col="red", lwd=2)

```

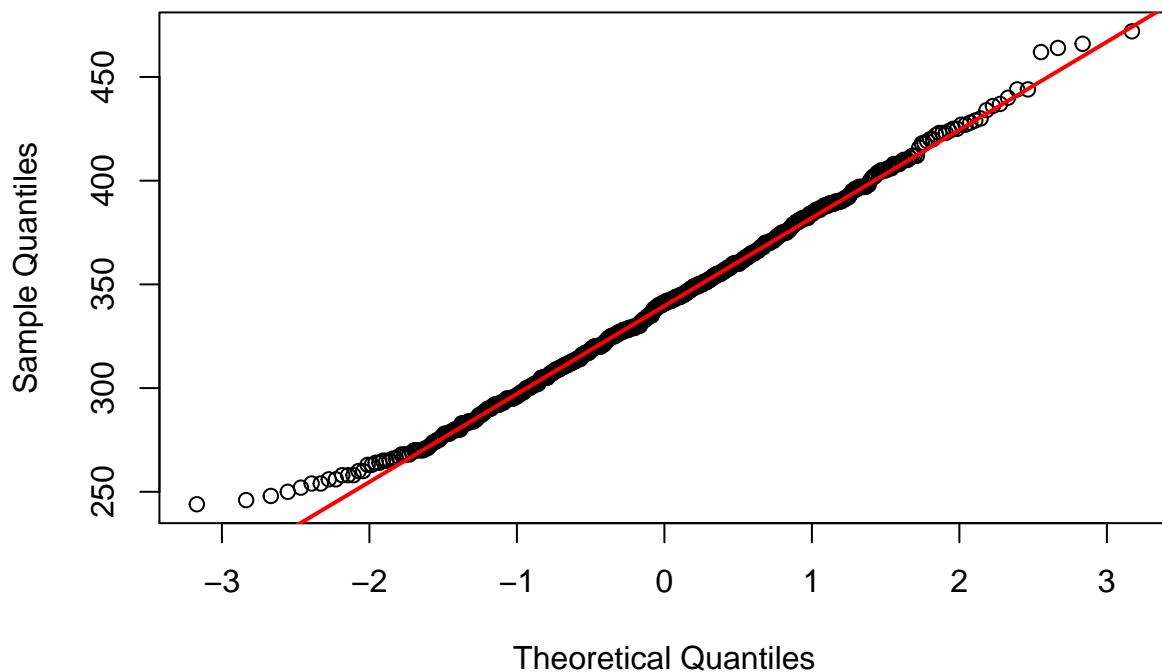
Biceps circumference, right-handed



```
lillie.test(ansurRight$bicepscircumferenceflexed)

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: ansurRight$bicepscircumferenceflexed
## D = 0.020079, p-value = 4.198e-05
qqnorm(ansurLeft$bicepscircumferenceflexed, main="Biceps circumference, left-handed")
qqline(ansurLeft$bicepscircumferenceflexed, col="red", lwd=2)
```

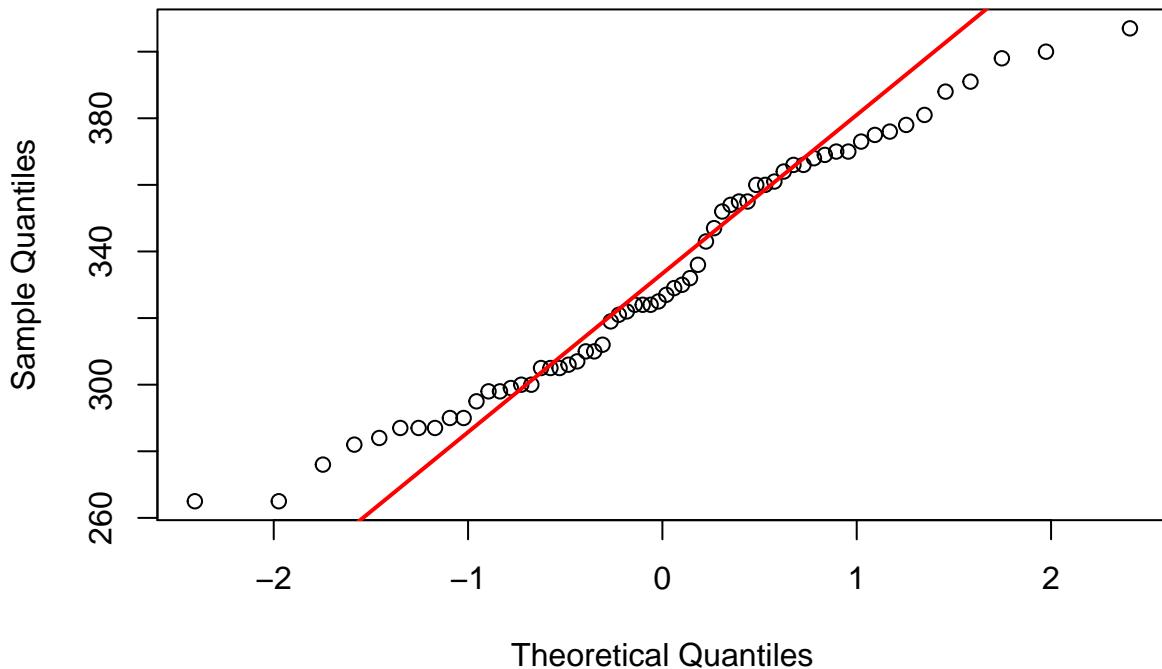
Biceps circumference, left-handed



```
lillie.test(ansurRight$bicepscircumferenceflexed)

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: ansurRight$bicepscircumferenceflexed
## D = 0.020079, p-value = 4.198e-05
qqnorm(ansurAmhi$bicepscircumferenceflexed, main="Biceps circumference, either hand")
qqline(ansurAmhi$bicepscircumferenceflexed, col="red", lwd=2)
```

Biceps circumference, either hand



```
lillie.test(ansurAmbi$bicepscircumferenceflexed)
```

```
##  
## Lilliefors (Kolmogorov-Smirnov) normality test  
##  
## data: ansurAmbi$bicepscircumferenceflexed  
## D = 0.097485, p-value = 0.1517
```

Raspodjela je približno normalna, i mjerena su nezavisna. Testirajmo jednakost varijanci.

```
var.test(ansurLeft$bicepscircumferenceflexed, ansurRight$bicepscircumferenceflexed)
```

```
##  
## F test to compare two variances  
##  
## data: ansurLeft$bicepscircumferenceflexed and ansurRight$bicepscircumferenceflexed  
## F = 1.0242, num df = 655, denom df = 5349, p-value = 0.6708  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 0.9154918 1.1518323  
## sample estimates:  
## ratio of variances  
## 1.024208
```

Nadalje prepostavljamo da su varijance jednake. Pokrenimo t-test

```
t.test(ansurLeft$bicepscircumferenceflexed, ansurRight$bicepscircumferenceflexed, var.equal=TRUE)
```

```
##
```

```

## Two Sample t-test
##
## data: ansurLeft$bicepscircumferenceflexed and ansurRight$bicepscircumferenceflexed
## t = -0.39645, df = 6004, p-value = 0.6918
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -4.051947 2.688750
## sample estimates:
## mean of x mean of y
## 340.4162 341.0978

```

Ne možemo prihvati hipotezu da su srednje vrijednosti različite sa 95%-tnom sigurnošću. Usporedimo sad ambideksterne vojnike i dešnjake

```

var.test(ansurAmbi$bicepscircumferenceflexed, ansurRight$bicepscircumferenceflexed)

##
## F test to compare two variances
##
## data: ansurAmbi$bicepscircumferenceflexed and ansurRight$bicepscircumferenceflexed
## F = 0.79618, num df = 61, denom df = 5349, p-value = 0.2517
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 0.573551 1.177803
## sample estimates:
## ratio of variances
## 0.7961773

t.test(ansurAmbi$bicepscircumferenceflexed, ansurRight$bicepscircumferenceflexed, var.equal=TRUE)

##
## Two Sample t-test
##
## data: ansurAmbi$bicepscircumferenceflexed and ansurRight$bicepscircumferenceflexed
## t = -1.6601, df = 5410, p-value = 0.09694
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -19.172668 1.590057
## sample estimates:
## mean of x mean of y
## 332.3065 341.0978

```

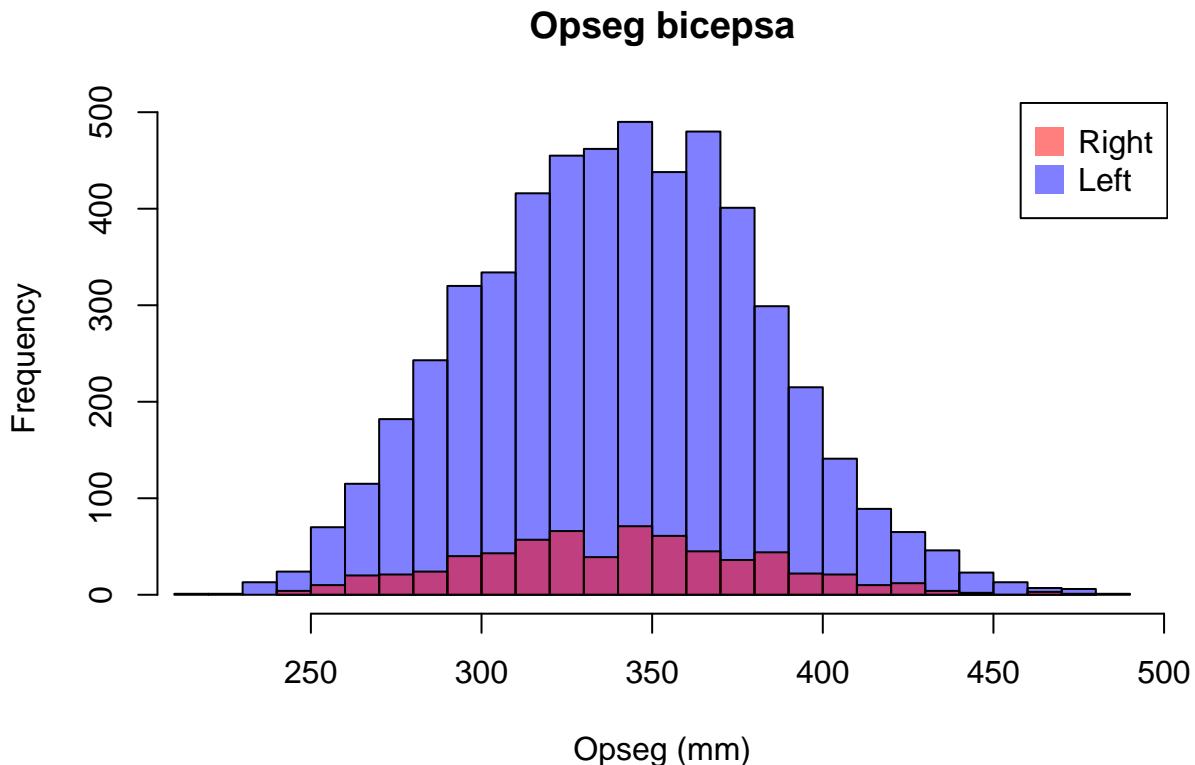
Srednje vrijednosti su dalje nego između ljevaka i dešnjaka, no i dalje ne možemo zaključiti da su različite.

```
plot_by_hand <- function(column, main = column, xlab = column) {
```

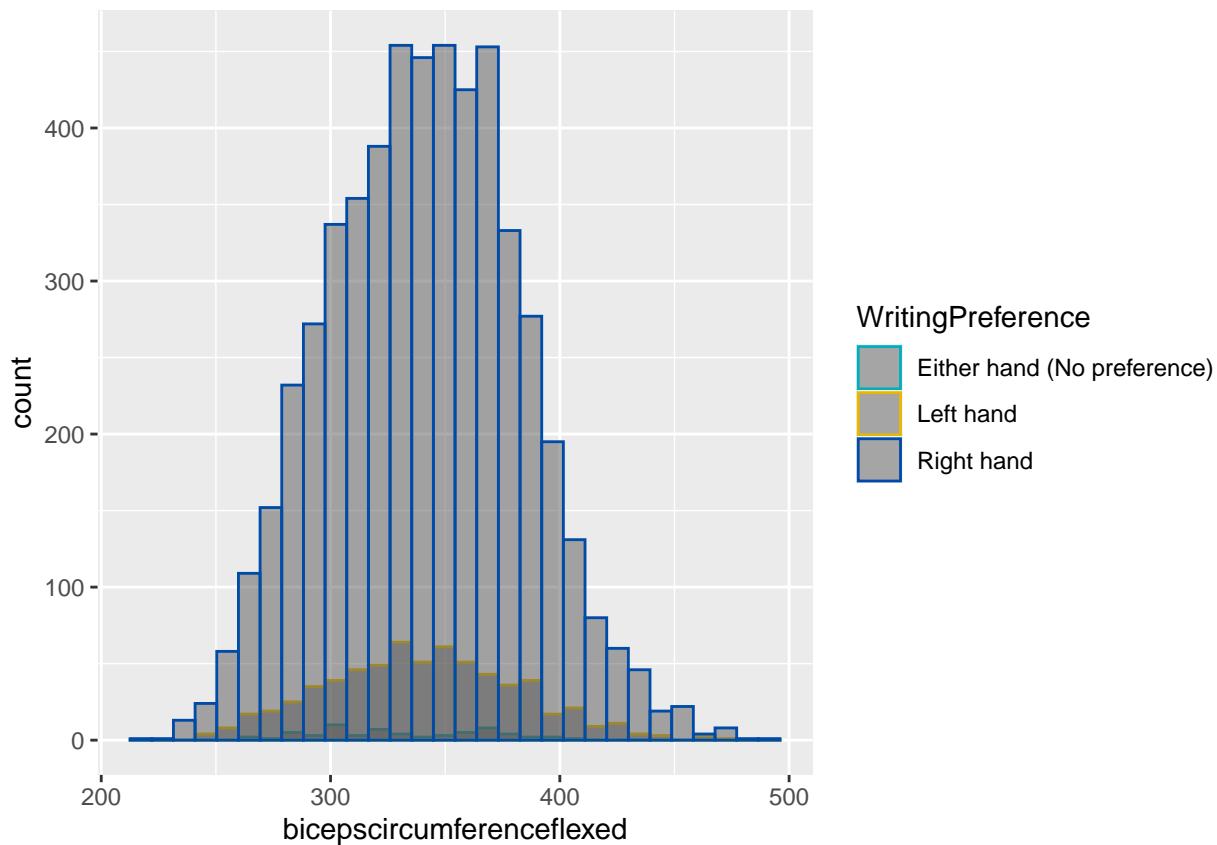
```

  hist(ansurRight[[column]], breaks=30, main=main, xlab=xlab, ylab="Frequency", col=rgb(0,0,1,0.5), xlim=c(0,100)
  hist(ansurLeft[[column]], breaks=30, main=main, xlab=xlab, ylab="Frequency", col=rgb(1,0,0,0.5), xlim=c(0,100)
  #abline(v=mean(ansur.II.data[[column]]), col="red", lwd=2)
  #abline(v=median(ansur.II.data[[column]]), col="blue", lwd=2)
  legend(x="topright", c("Right", "Left"), col=c(rgb(1,0,0,0.5), rgb(0,0,1,0.5)), pt.cex = 2, pch = 15)
}
```

```
plot_by_hand("bicepscircumferenceflexed", "Opseg bicepsa", "Opseg (mm)")
```



```
ggplot_by_property <- function(wdata) {  
  ggplot(wdata, aes(x = bicepscircumferenceflexed)) +  
  geom_histogram(aes(color = WritingPreference),  
                 position = "identity", bins = 30, alpha = 0.5) +  
  scale_color_manual(values = c("#00AFBB", "#E7B800", "#004BA8"))  
}  
  
ggplot_by_property(antrData)
```

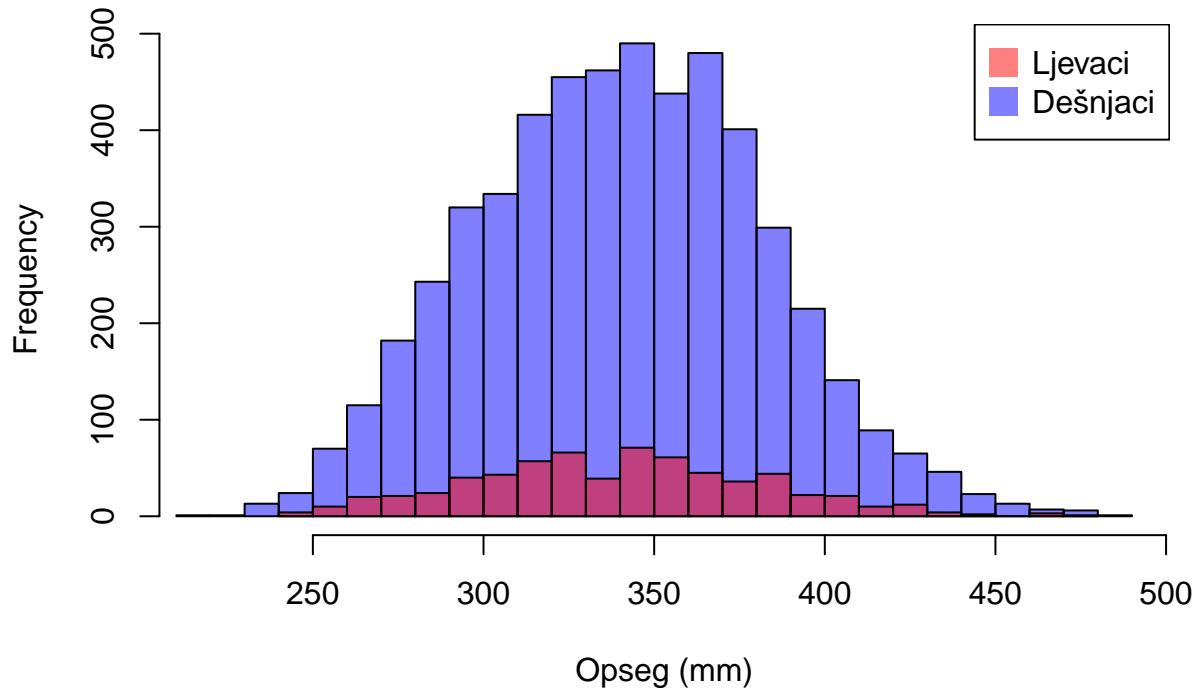


```
plotajDL <- function() {

  hist(ansurRight$bicepscircumferenceflexed, breaks=30, main="Opseg bicepsa", xlab="Opseg (mm)", ylab="Fr
  hist(ansurLeft$bicepscircumferenceflexed, breaks=30, main="Opseg bicepsa", xlab="Opseg (mm)", ylab="Fr
  ##hist(ansurAmbi$bicepscircumferenceflexed, breaks=30, main="Opseg bicepsa", xlab="Opseg (mm)", ylab=
  #abline(v=mean(ansur.II.data[[column]]), col="red", lwd=2)
  #abline(v=median(ansur.II.data[[column]]), col="blue", lwd=2)
  legend(x="topright", c("Ljevaci", "Dešnjaci"), col=c(rgb(1,0,0,0.5), rgb(0,0,1,0.5)), pt.cex = 2, pch=
}

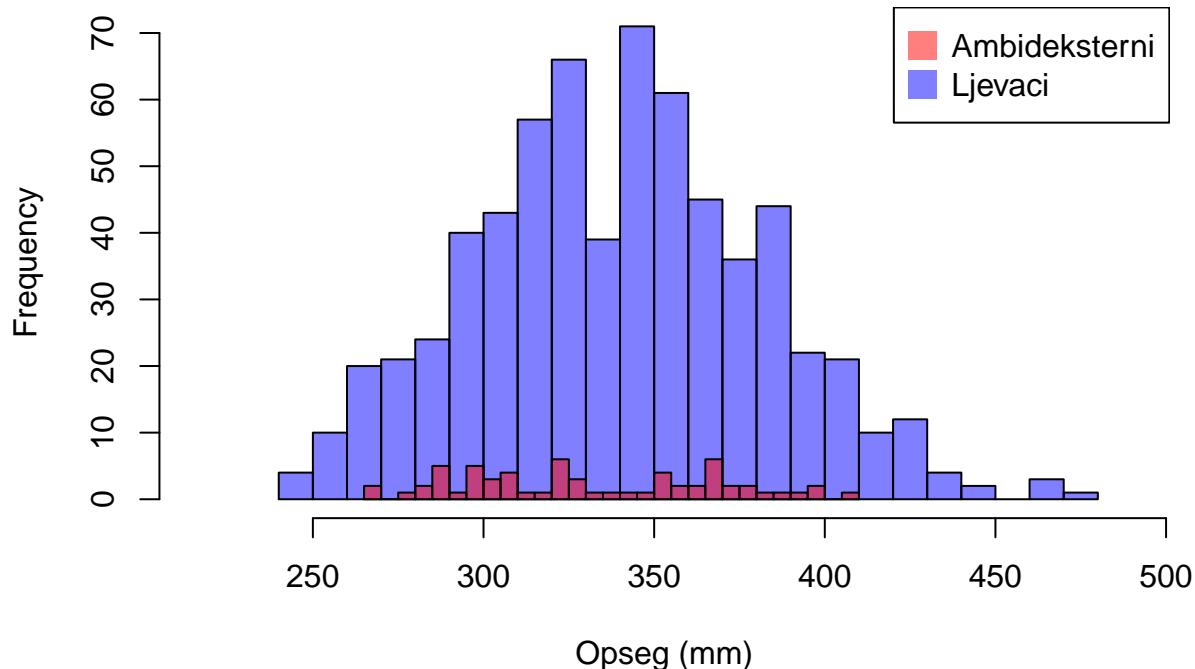
plotajDL()
```

Opseg bicepsa



```
plotajAmbiLeft <- function() {  
  
  hist(ansurLeft$bicepscircumferenceflexed, breaks=30, main="Opseg bicepsa", xlab="Opseg (mm)", ylab="Frequency")  
  hist(ansurAmbi$bicepscircumferenceflexed, breaks=30, main="Opseg bicepsa", xlab="Opseg (mm)", ylab="Frequency")  
  ##hist(ansurAmbi$bicepscircumferenceflexed, breaks=30, main="Opseg bicepsa", xlab="Opseg (mm)", ylab="Frequency")  
  #abline(v=mean(ansur.II.data[[column]]), col="red", lwd=2)  
  #abline(v=median(ansur.II.data[[column]]), col="blue", lwd=2)  
  legend(x="topright", c("Ambideksterni", "Ljevaci"), col=c(rgb(1,0,0,0.5), rgb(0,0,1,0.5)), pt.cex = 2)  
  
}  
  
plotajAmbiLeft()
```

Opseg bicepsa



Nismo pronašli statistički značajnu ovisnost opsega bicepsa o dominantnoj ruci. Uzrok je vjerojatno to što je pisaća ruka faktor sa komparativno malenim utjecajem. Puno veći utjecaj na veličinu desne ruke imaju drugi tjelesni parametri, primjerice težina. Gledajmo stoga kao slučajnu varijablu omjer opsega bicepsa i tjelesne težine kako bi pokušali izolirati utjecaj dominantne ruke.

```
weightVec = ansurLeft$weightkg
bicVec = ansurLeft$bicepscircumferenceflexed
ratioLeft = bicVec
for (i in 1:length(bicVec)) {
  ratioLeft[i] <- (bicVec[i]/weightVec[i])
}
weightVec = ansurRight$weightkg
bicVec = ansurRight$bicepscircumferenceflexed
ratioRight = bicVec
for (i in 1:length(bicVec)) {
  ratioRight[i] <- (bicVec[i]/weightVec[i])
}
weightVec = ansurAmbe$weightkg
bicVec = ansurAmbe$bicepscircumferenceflexed
ratioAmbe = bicVec
for (i in 1:length(bicVec)) {
  ratioAmbe[i] <- (bicVec[i]/weightVec[i])
}
print("Ljevaci:")
## [1] "Ljevaci:"
```

```

summary(ratioLeft)

##      Min. 1st Qu. Median     Mean 3rd Qu.    Max.
## 0.3150  0.3955  0.4273  0.4283  0.4589  0.5847

print("Dešnjaci:")

## [1] "Dešnjaci:"
```

```

summary(ratioRight)

##      Min. 1st Qu. Median     Mean 3rd Qu.    Max.
## 0.2946  0.4040  0.4340  0.4359  0.4658  0.6361

print("Lješnjaci:")

## [1] "Lješnjaci:"
```

```

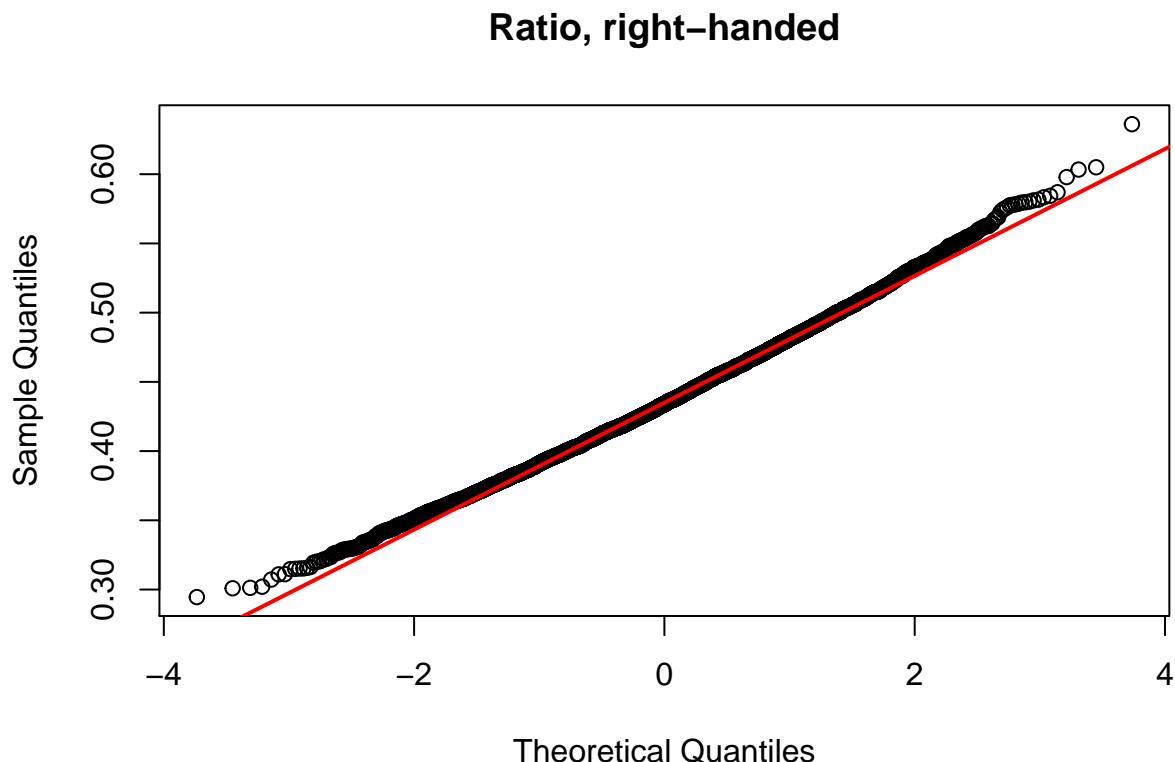
summary(ratioAmbi)

##      Min. 1st Qu. Median     Mean 3rd Qu.    Max.
## 0.3682  0.4119  0.4354  0.4430  0.4728  0.5910
```

Provjerimo stupanj normalnost

```

qqnorm(ratioRight, main="Ratio, right-handed")
qqline(ratioRight, col="red", lwd=2)
```



```

lillie.test(ratioRight)
```

```

##  

## Lilliefors (Kolmogorov-Smirnov) normality test  

##  

## data: ratioRight  

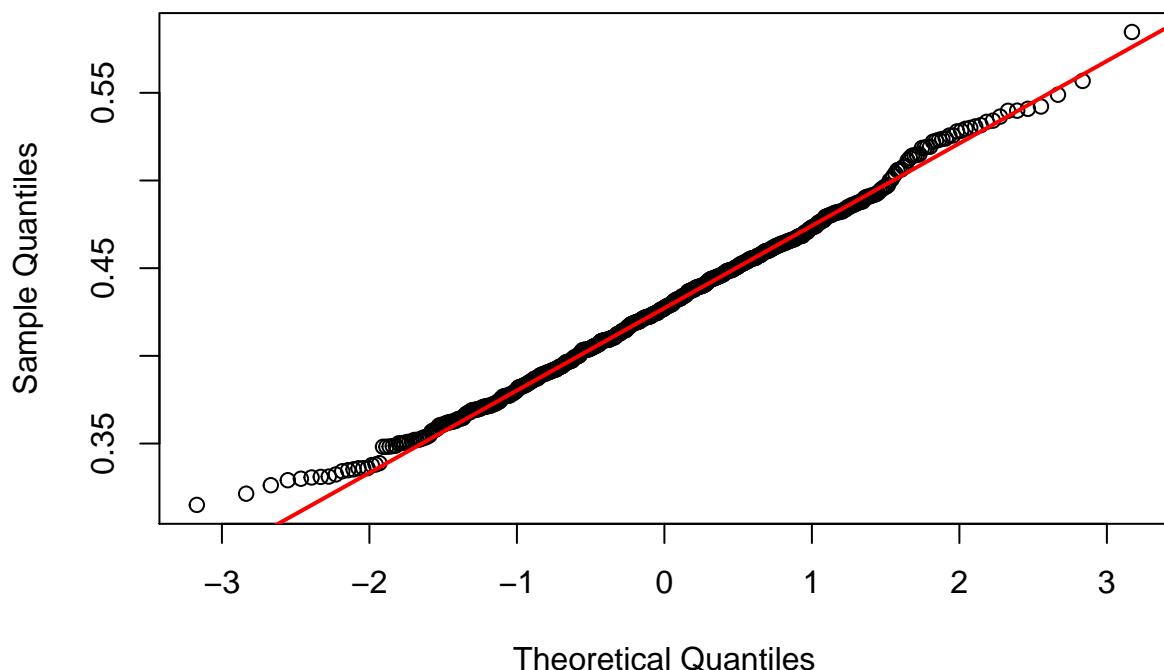
## D = 0.021706, p-value = 5.282e-06  

qqnorm(ratioLeft, main="Ratio, left-handed")  

qqline(ratioLeft, col="red", lwd=2)

```

Ratio, left-handed



```

lillie.test(ratioLeft)

##  

## Lilliefors (Kolmogorov-Smirnov) normality test  

##  

## data: ratioLeft  

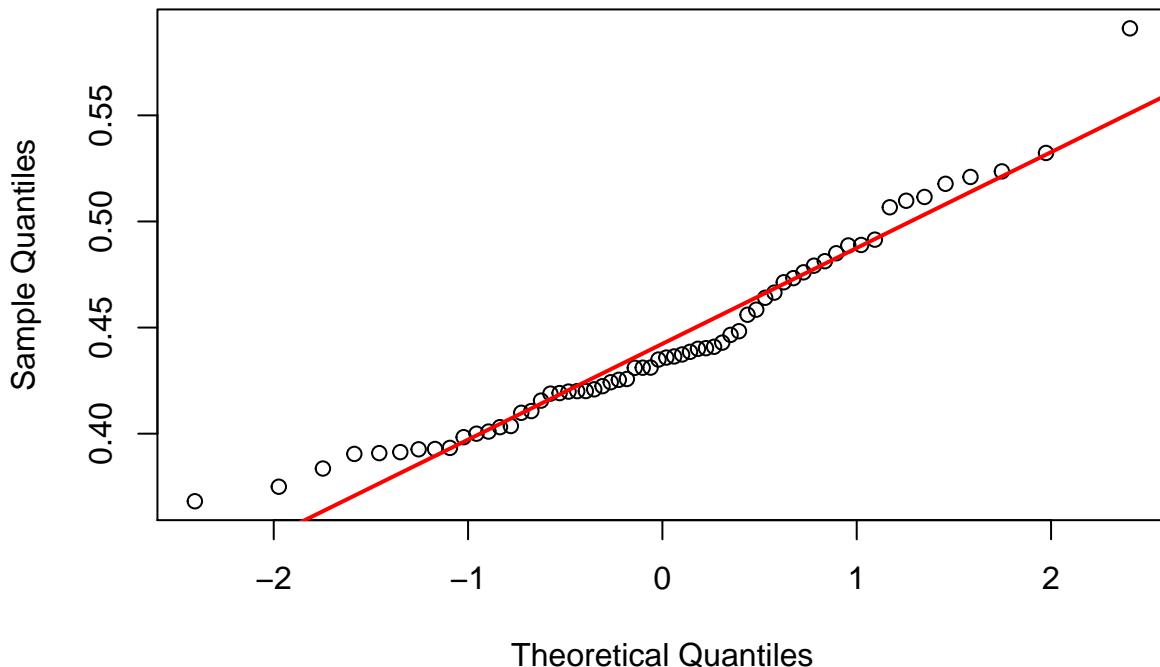
## D = 0.017871, p-value = 0.8788  

qqnorm(ratioAmbi, main="Ratio, either hand")  

qqline(ratioAmbi, col="red", lwd=2)

```

Ratio, either hand



```
lillie.test(ratioAmbe)
```

```
##  
## Lilliefors (Kolmogorov-Smirnov) normality test  
##  
## data: ratioAmbe  
## D = 0.13105, p-value = 0.00992
```

Oblik razdiobe je zadovoljavajuć za korištenje t-testa.

```
var.test(ratioLeft, ratioRight)
```

```
##  
## F test to compare two variances  
##  
## data: ratioLeft and ratioRight  
## F = 1.0303, num df = 655, denom df = 5349, p-value = 0.5986  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 0.9209053 1.1586433  
## sample estimates:  
## ratio of variances  
## 1.030264
```

```
t.test(ratioLeft, ratioRight, var.equal=TRUE)
```

```
##  
## Two Sample t-test
```

```

## 
## data: ratioLeft and ratioRight
## t = -4.033, df = 6004, p-value = 5.575e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.011260891 -0.003894288
## sample estimates:
## mean of x mean of y
## 0.4283016 0.4358792

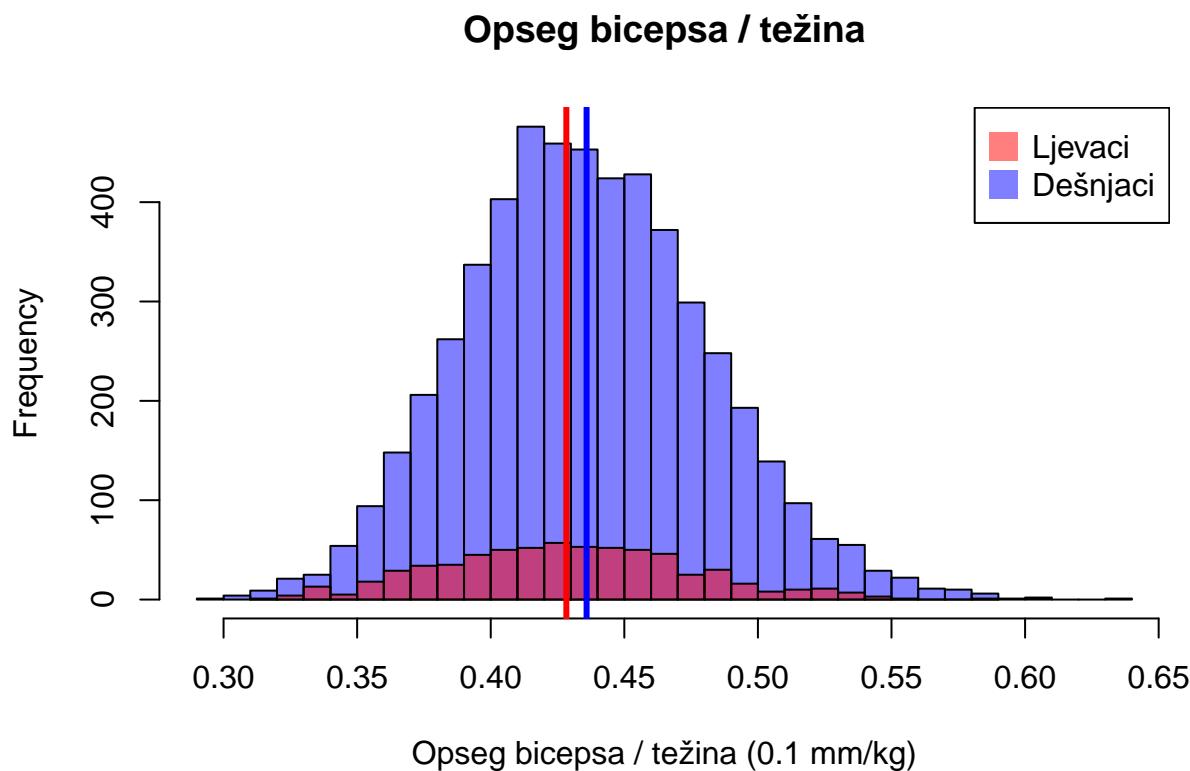
```

Zaista, opseg desnog bicepsa pokazuje statistički značajnu ovisnost o tome je li dominantna ruka lijeva ili desna uzevši u obzir težinu. Vizualizirajmo.

```

hist(ratioRight, breaks=30, main="Opseg bicepsa / težina", xlab="Opseg bicepsa / težina (0.1 mm/kg)", yl
hist(ratioLeft, breaks=30, main="Opseg bicepsa / težina", xlab="Opseg bicepsa / težina (0.1 mm/kg)", yl
abline(v = mean(ratioLeft), col=rgb(1,0,0,1), lwd = 3)
abline(v = mean(ratioRight), col=rgb(0,0,1,1), lwd = 3)
legend(x="topright", c("Ljevaci", "Dešnjaci"), col=c(rgb(1,0,0,0.5), rgb(0,0,1,0.5)), pt.cex = 2, pch =

```



Ispitajmo odnos za ambideksterne vojнике.

```
var.test(ratioRight, ratioAmbi)
```

```

## 
## F test to compare two variances

```

```

##
## data: ratioRight and ratioAmber
## F = 1.0201, num df = 5349, denom df = 61, p-value = 0.9607
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 0.6895525 1.4160154
## sample estimates:
## ratio of variances
## 1.020071
t.test(ratioAmber, ratioRight, var.equal=TRUE)

##
## Two Sample t-test
##
## data: ratioAmber and ratioRight
## t = 1.2295, df = 5410, p-value = 0.2189
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.004232848 0.018473915
## sample estimates:
## mean of x mean of y
## 0.4429997 0.4358792

```

U ovom slučaju ne nalazimo statističko značajnu razliku.

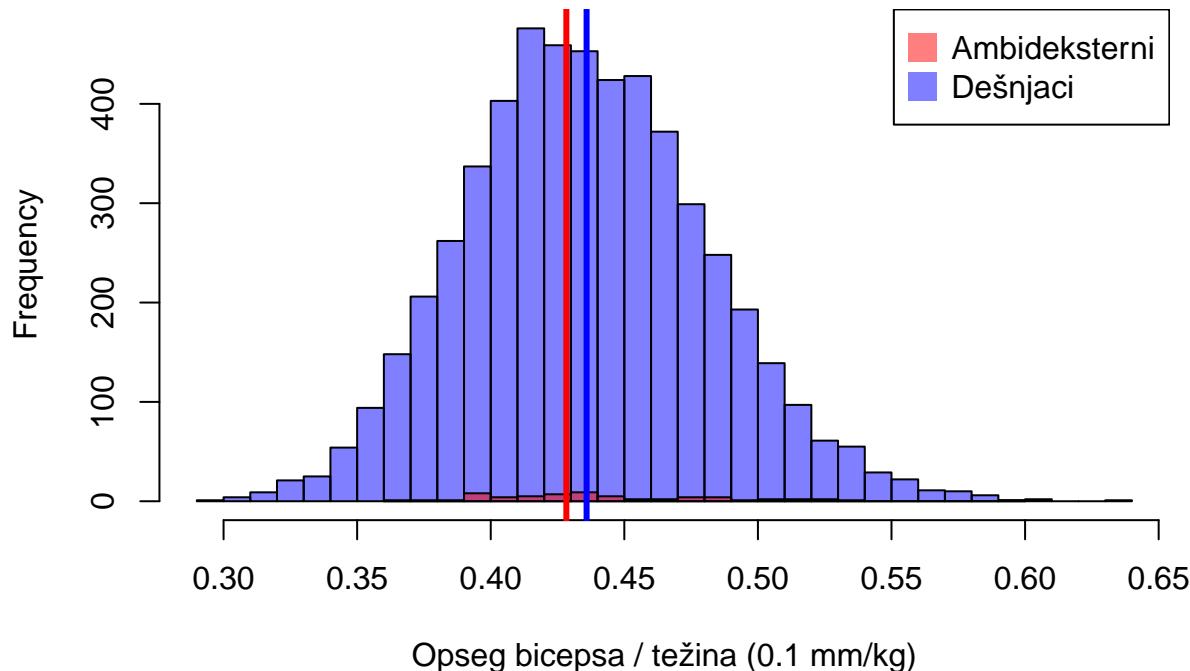
```

hist(ratioRight, breaks=30, main="Opseg bicepsa / težina", xlab="Opseg bicepsa / težina (0.1 mm/kg)", yl
hist(ratioAmber, breaks=30, main="Opseg bicepsa / težina", xlab="Opseg bicepsa / težina (0.1 mm/kg)", yl
abline(v = mean(ratioLeft), col=rgb(1,0,0,1), lwd = 3)
abline(v = mean(ratioRight), col=rgb(0,0,1,1), lwd = 3)

legend(x="topright", c("Ambideksterni", "Dešnjaci"), col=c(rgb(1,0,0,0.5), rgb(0,0,1,0.5)), pt.cex = 2,

```

Opseg bicepsa / težina



Koje još varijable pokazuju korelaciju s pisaćom rukom? Odaberimo varijable handbreadth i handcircumference kao mjere koje su povezane s rukom te bi mogle pokazati zavisnost.

```
weightVec = ansurLeft$weightkg
vec = ansurLeft$handbreadth
ratioLeft = vec
print(length(vec))

## [1] 656
for (i in 1:length(vec)) {
  ratioLeft[i] <- (vec[i]/weightVec[i])
}
weightVec = ansurRight$weightkg
vec = ansurRight$handbreadth
ratioRight = vec
for (i in 1:length(vec)) {
  ratioRight[i] <- (vec[i]/weightVec[i])
}
var.test(ratioLeft, ratioRight)

##
## F test to compare two variances
##
## data: ratioLeft and ratioRight
## F = 1.042, num df = 655, denom df = 5349, p-value = 0.4706
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
```

```

##  0.9313866 1.1718305
## sample estimates:
## ratio of variances
##                 1.04199
t.test(ratioLeft, ratioRight, var.equal=TRUE)

##
## Two Sample t-test
##
## data: ratioLeft and ratioRight
## t = -2.8307, df = 6004, p-value = 0.00466
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.0033434520 -0.0006073775
## sample estimates:
## mean of x mean of y
## 0.1078182 0.1097936

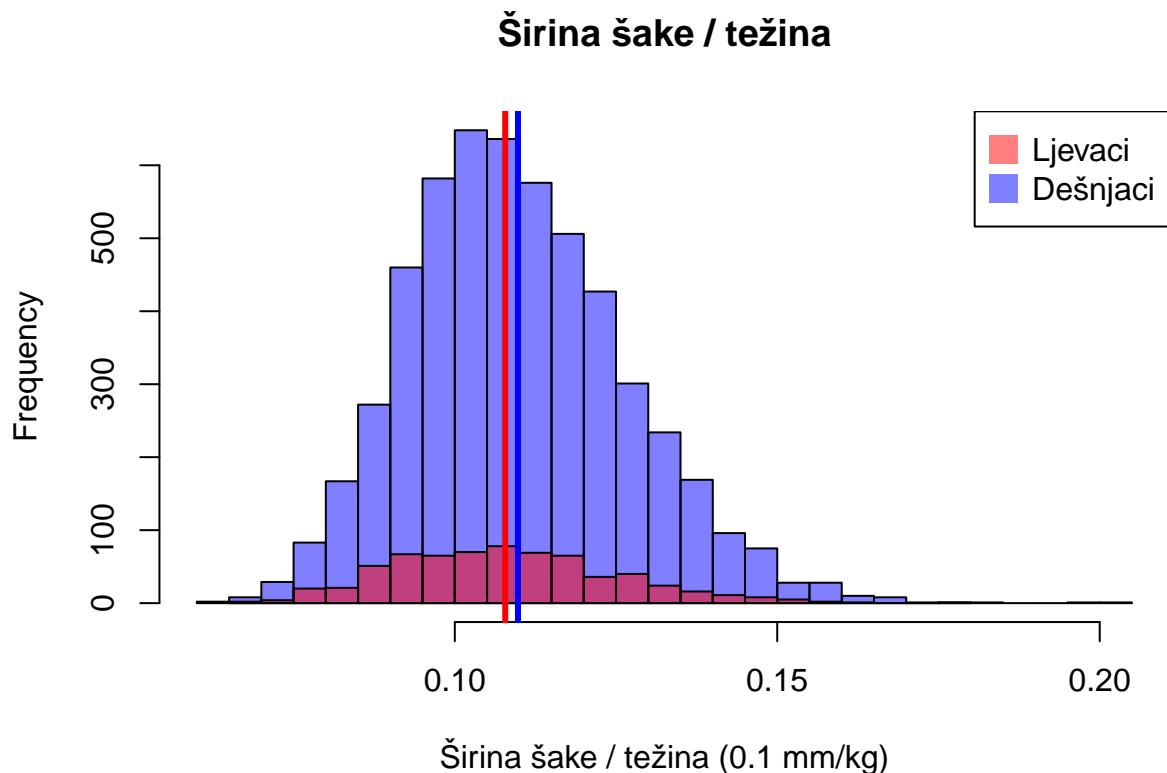
hist(ratioRight, breaks=30, main="Širina šake / težina", xlab="Širina šake / težina (0.1 mm/kg)", ylab="Frequency")

hist(ratioLeft, breaks=30, main="Širina šake / težina", xlab="Širina šake / težina (0.1 mm/kg)", ylab="Frequency")

abline(v = mean(ratioLeft), col=rgb(1,0,0,1), lwd = 3)
abline(v = mean(ratioRight), col=rgb(0,0,1,1), lwd = 3)

legend(x="topright", c("Ljevaci", "Dešnjaci"), col=c(rgb(1,0,0,0.5), rgb(0,0,1,0.5)), pt.cex = 2, pch = 15)

```



```

weightVec = ansurLeft$weightkg
vec = ansurLeft$handcircumference
ratioLeft = vec
print(length(vec))

## [1] 656

for (i in 1:length(vec)) {
  ratioLeft[i] <- (vec[i]/weightVec[i])
}
weightVec = ansurRight$weightkg
vec = ansurRight$handcircumference
ratioRight = vec
for (i in 1:length(vec)) {
  ratioRight[i] <- (vec[i]/weightVec[i])
}
var.test(ratioLeft, ratioRight)

## 
## F test to compare two variances
##
## data: ratioLeft and ratioRight
## F = 1.0312, num df = 655, denom df = 5349, p-value = 0.5877
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.9217482 1.1597039
## sample estimates:
## ratio of variances
##               1.031207
t.test(ratioLeft, ratioRight, var.equal=TRUE)

##
## Two Sample t-test
##
## data: ratioLeft and ratioRight
## t = -3.0889, df = 6004, p-value = 0.002018
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.008235202 -0.001840653
## sample estimates:
## mean of x mean of y
## 0.2582769 0.2633149

hist(ratioRight, breaks=30, main="Opseg šake / težina", xlab="Opseg šake / težina (0.1 mm/kg)", ylab="Fr")

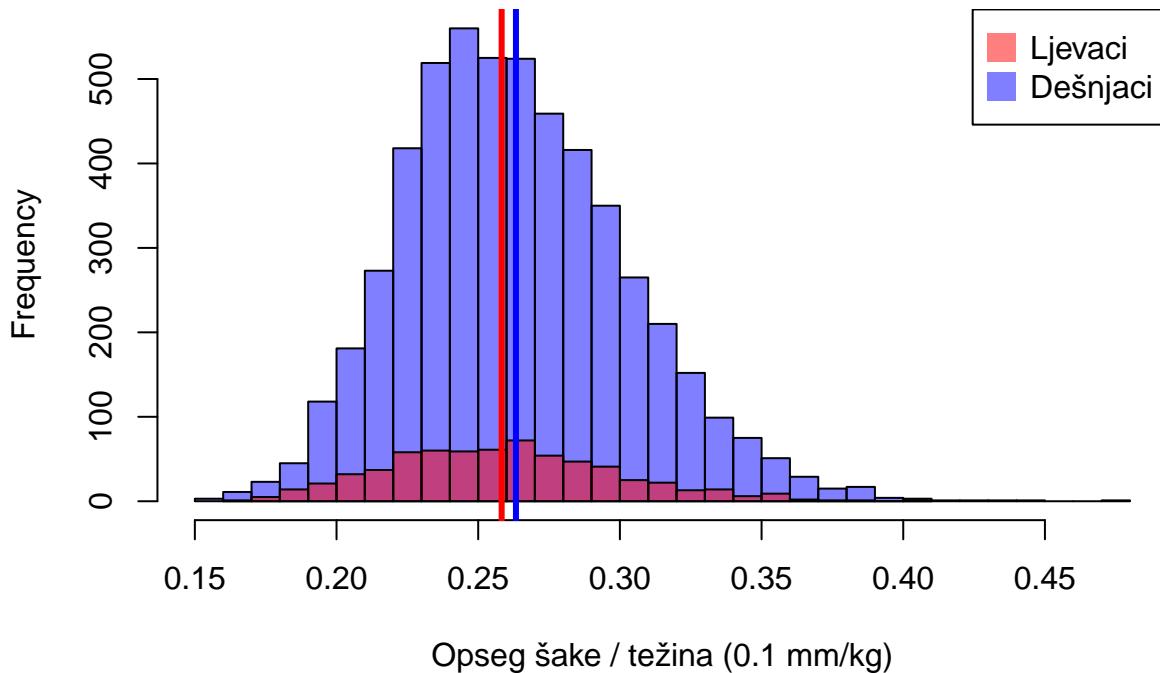
hist(ratioLeft, breaks=30, main="Opseg šake / težina", xlab="Opseg šake / težina (0.1 mm/kg)", ylab="Fr")

abline(v = mean(ratioLeft), col=rgb(1,0,0,1), lwd = 3)
abline(v = mean(ratioRight), col=rgb(0,0,1,1), lwd = 3)

legend(x="topright", c("Ljevaci", "Dešnjaci"), col=c(rgb(1,0,0,0.5), rgb(0,0,1,0.5)), pt.cex = 2, pch =

```

Opseg šake / težina



Za te varijable nalazimo ovisnost, no ne tako jaku kao za opseg bicepsa.

Zaključak

Češće korištenje jedne ruke svakako do značajne mjere povećava veličinu njenih mišića. U uzorku koji je reprezentativan za cijelu populaciju bi našli jaču korelaciju jer vojnici vježbaju obje ruke podjednako i to dovodi do simetrije u veličini desne i lijeve ruke. Stupanj nesimetričnosti tijela se može koristiti kao pokazatelj uspješnosti funkcionalnog treninga. Kampovi s većom razlikom između ruku bi mogli to iskoristiti da poboljšaju trening i pojačaju snagu nedominantne strane tijela u svojih vojnika.

Logistički model za predviđanje pripadnika bijele ili crne rase na temelju omjera raspona ruku i visine

Motivacija

Svakom obožavatelju američkog sporta poznate su brojne antropometrijske mjere koje se uzimaju u obzir prilikom evaluacije igrača. Jedna od najbitnijih mjera za procjenu igrača je raspon ruku, najviše je to naglašeno u košarci. Kako sam i sam vjerni pratitelj NBA i NFL lige primjetio sam kako u tim evaluacijama crni igrači najčešće imaju veći raspon ruku nego bijeli igrači, pogotovo u odnosu na visinu. U sljedećem odlomku slijedi analiza omjera raspona ruku i visine između bijele i crne rase. Kako bi se smanjio utjecaj dodatnih faktora, promatrati ćemo muškarce i žene odvojeno. Na kraju ćemo pristupiti izgradnji logističkog modela za predviđanje rase na temelju prije spomenutog omjera.

Istraživanje

```
antrData = read.csv("ANSUR_II_data.csv")
```

```

antrData$colorData = "black"
antrData$colorData[antrData$DODRace != 1] = "red"

require(fastDummies)
antrData = dummy_cols(antrData, select_columns = c("Gender"))

```

Nakon dodavanja stupca za boju različitih rasa i stvaranju dummy varijable za prikaz spola, pristupamo odvajanju podataka na muške i ženske te izdavajanje pripadnika bijele i crne rase.

```

antrData_male <- antrData[antrData$Gender_Male == 1,]
antrData_blackwhite <- antrData_male[antrData_male$DODRace == 1 | antrData_male$DODRace == 2,]

antrData_female <- antrData[antrData$Gender_Male == 0,]
antrData_blackwhite_female <- antrData_female[antrData_female$DODRace == 1 | antrData_female$DODRace == 2,]

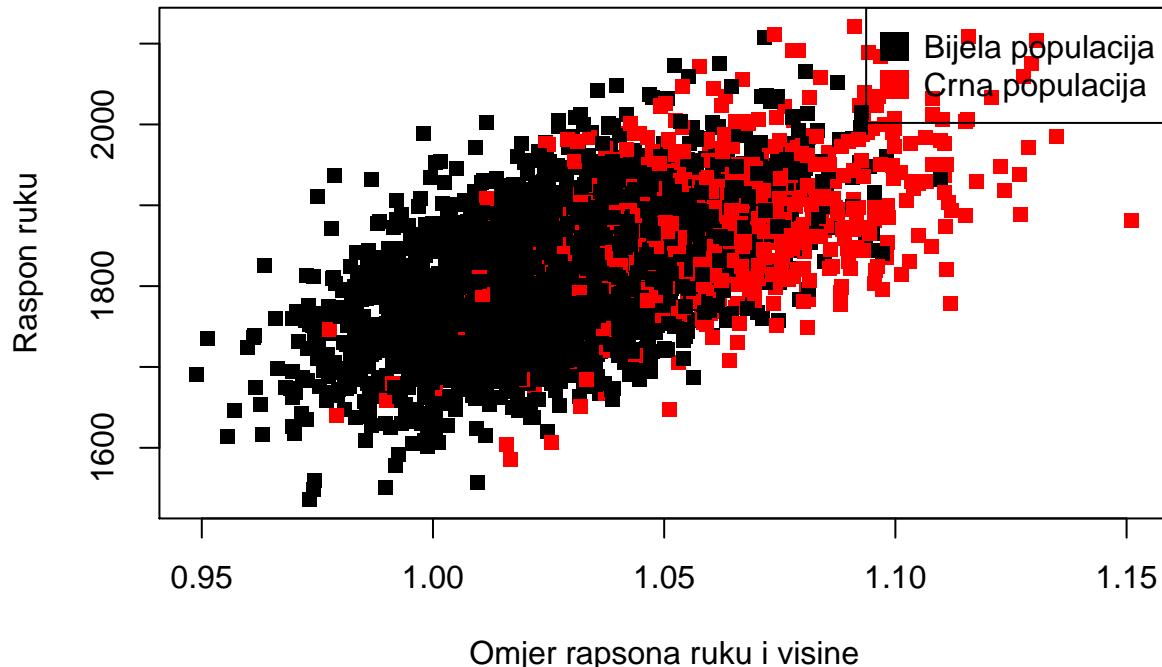
Nakon razdvajanja stvaramo stupac u kojem će se nalaziti omjer nad kojim ćemo vršiti analizu.

antrData_blackwhite$spanheightratio <- antrData_blackwhite$span / antrData_blackwhite$stature
antrData_blackwhite_female$spanheightratio <- antrData_blackwhite_female$span / antrData_blackwhite_female$stature

plot(antrData_blackwhite$spanheightratio, antrData_blackwhite$span, pch=15, xlab = 'Omjer rapsona ruku i visine',
      legend('topright', legend = c("Bijela populacija", "Crna populacija"), col = c("black", "red"), pt.cex=1.5)

```

Omjer medu muškom populacijom

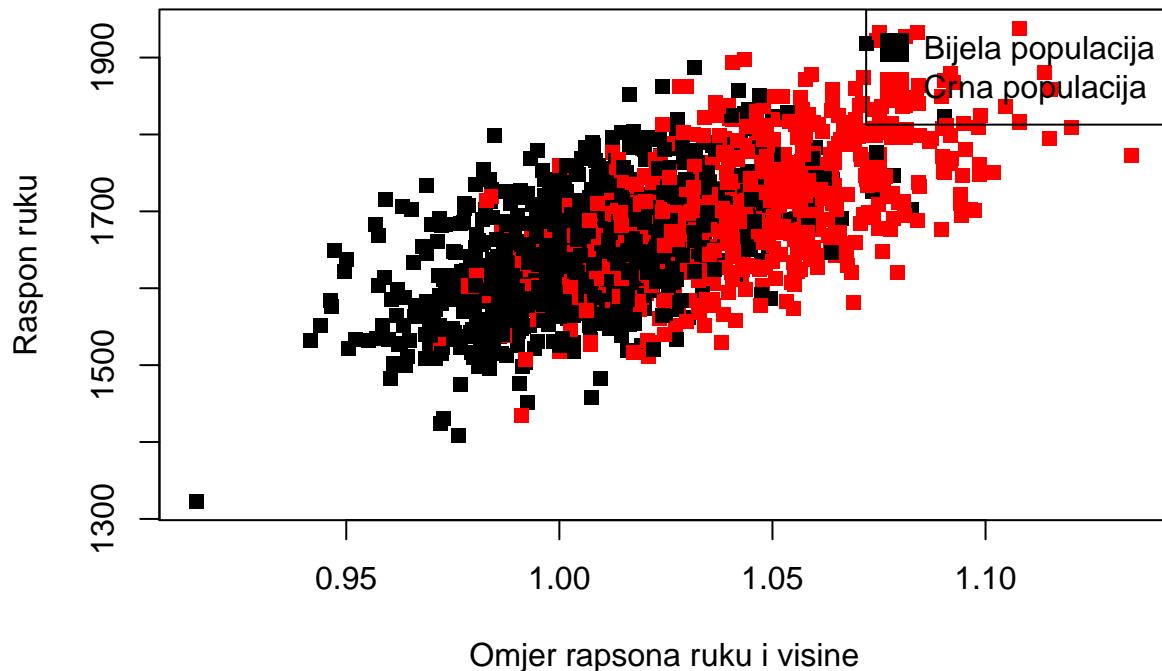


```

plot(antrData_blackwhite_female$spanheightratio, antrData_blackwhite_female$span, pch=15, xlab = 'Omjer rapsona ruku i visine',
      legend('topright', legend = c("Bijela populacija", "Crna populacija"), col = c("black", "red"), pt.cex=1.5)

```

Omjer među ženskom populacijom



Dijagram raspršenja nam pruža određenu informaciju o tome kako se omjer razlikuje između pripadnika bijele i crne rase. Ipak s obzirom na broj podataka histogram će nam možda pružiti jasniji uvod u prirodu dviju distribucija.

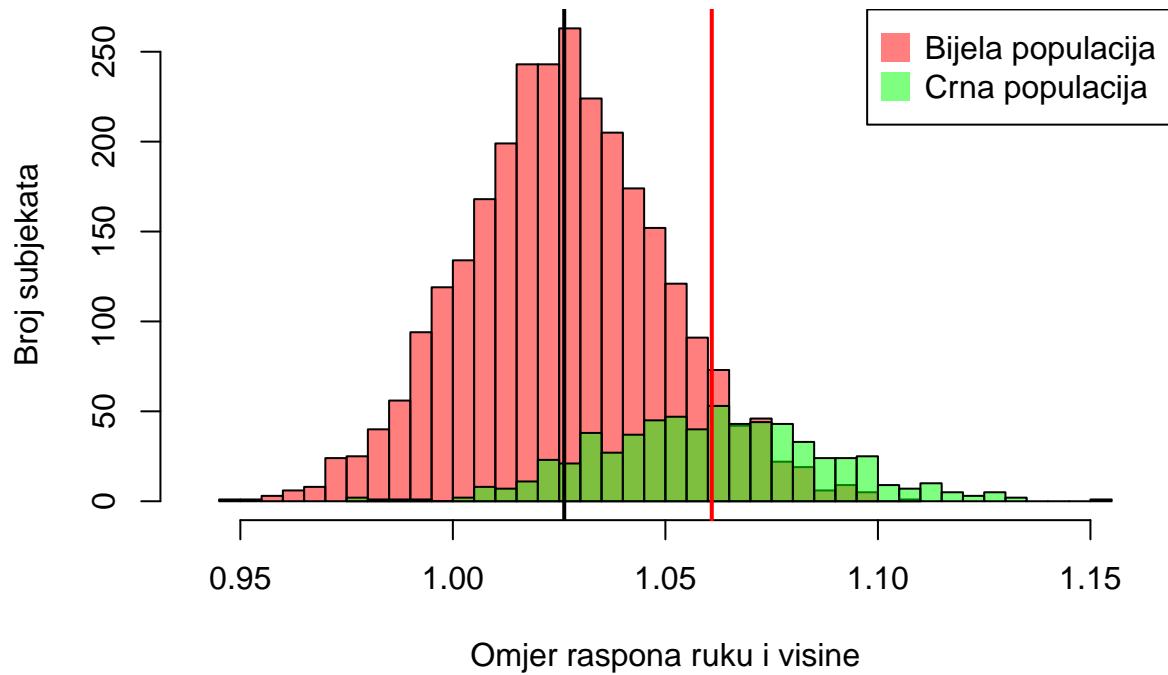
```
mean_male_white <- mean(antrData_blackwhite$spanheightratio[antrData_blackwhite$DODRace == 1])
mean_male_black <- mean(antrData_blackwhite$spanheightratio[antrData_blackwhite$DODRace == 2])

hist(antrData_blackwhite$spanheightratio[antrData_blackwhite$DODRace == 1], breaks = 30, xlim=c(0.94,1.10))
hist(antrData_blackwhite$spanheightratio[antrData_blackwhite$DODRace == 2], breaks = 30, xlim=c(0.94,1.10))

abline(v = mean_male_white, col = "black", lwd = 2)
abline(v = mean_male_black, col = "red", lwd = 2)

legend('topright', legend = c("Bijela populacija", "Crna populacija"), col = c(rgb(1,0,0,0.5), rgb(0,1,0,0.5)))
```

Omjer medu muškom populacijom



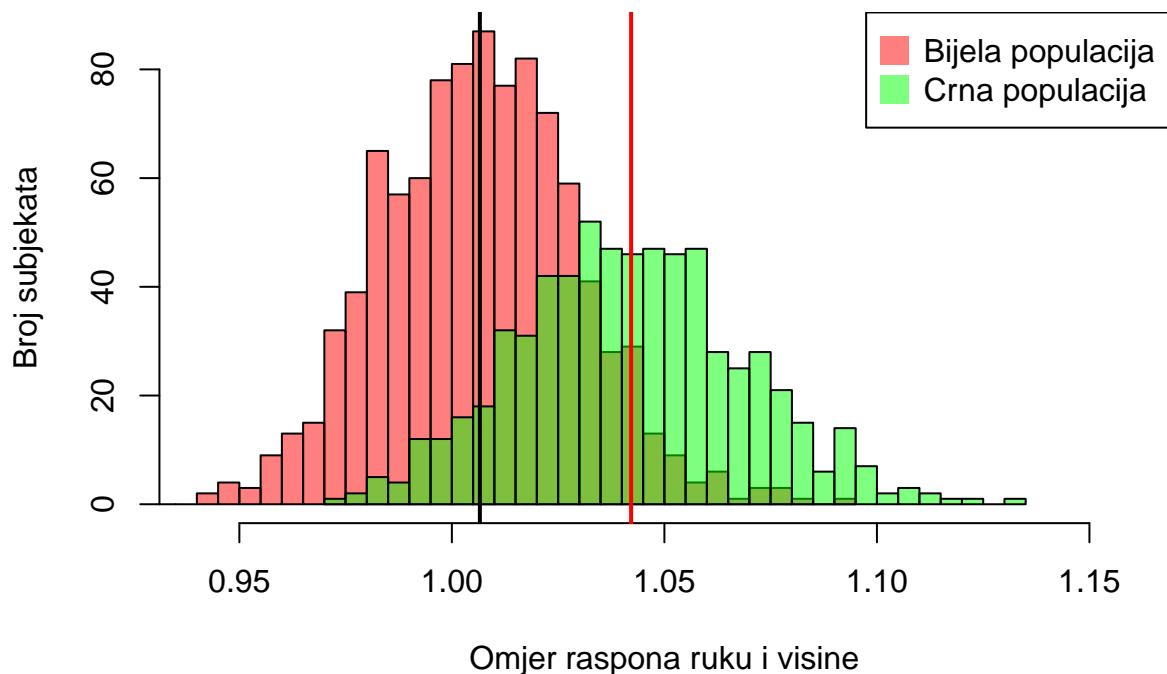
```
mean_female_white <- mean(antrData_blackwhite_female$spanheightratio[antrData_blackwhite_female$DODRace == 1])
mean_female_black <- mean(antrData_blackwhite_female$spanheightratio[antrData_blackwhite_female$DODRace == 2])

hist(antrData_blackwhite_female$spanheightratio[antrData_blackwhite_female$DODRace == 1], breaks = 30, xlab = "Omjer raspona ruku i visine", col = "red")
hist(antrData_blackwhite_female$spanheightratio[antrData_blackwhite_female$DODRace == 2], breaks = 30, xlab = "Omjer raspona ruku i visine", col = "green")

abline(v = mean_female_white, col = "black", lwd = 2)
abline(v = mean_female_black, col = "red", lwd = 2)

legend('topright', legend = c("Bijela populacija", "Crna populacija"), col = c(rgb(1,0,0,0.5), rgb(0,1,0,0.5)))
```

Omjer među ženskom populacijom



Iz histograma se jako lijepo vidi razlika između dvije populacije te razlika njihovih srednjih vrijednosti. Testiranje hipoteze ako je razlika tih dviju statistika statistički značajna ćemo ostaviti za kasnije, a sada kada smo se uvjerili da određena razina razlike postoji ćemo pristupiti izgradnji prvog logističkog modela. Prvi model ćemo izgraditi samo za mušku populaciju.

```

antrData_blackwhite = dummy_cols(antrData_blackwhite, select_columns = c("DODRace"))
antrData_blackwhite_female = dummy_cols(antrData_blackwhite_female, select_columns = c("DODRace"))

logitSpan <- glm(DODRace_1 ~ spanheightratio, data=antrData_blackwhite, family='binomial')
summary(logitSpan)

##
## Call:
## glm(formula = DODRace_1 ~ spanheightratio, family = "binomial",
##      data = antrData_blackwhite)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -3.2154   0.1657   0.3439   0.5542   2.1830
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  60.291     2.473  24.38  <2e-16 ***
## spanheightratio -56.391     2.356 -23.94  <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: 3319.1 on 3458 degrees of freedom
## Residual deviance: 2432.8 on 3457 degrees of freedom
## AIC: 2436.8
##
## Number of Fisher Scoring iterations: 5

```

Zasada stvari izgledaju dobro oba koeficijenta su statistički značajna, a residual deviance koja predstavlja razliku između zasićenog modela(svaki podatak ima jedan parametar) i našeg modela pokazuje značajnu razliku od null deviance koja predstavlja razliku između zasićenog modela i null modela(modela koji koristi samo jedan parametar) što se smatra poželjnim. Ipak već u sljedećem koraku uočiti ćemo bitan problem s našim modelom.

```

logitSpan_probs <- predict(logitSpan, type="response")
logitSpan_predictions <- ifelse(logitSpan_probs > 0.5, 1, 0)
confMatrix <- as.data.frame(table(logitSpan_predictions, antrData_blackwhite$DODRace_1))
print(confMatrix)

## logitSpan_predictions Var2 Freq
## 1                      0    0 241
## 2                      1    0 401
## 3                      0    1 118
## 4                      1    1 2699

accuracy <- (confMatrix[1, 3] + confMatrix[4, 3])/length(logitSpan_predictions)
print(accuracy)

## [1] 0.8499566

```

Ovdje nailazimo na čest problem kod logističke regresije, ne balansirani skupovi kategorija. S obzirom da je pripadnika bijele rase skoro 5 puta više od pripadnika crne rase modelu se puno više "ispłati" predviđati pripadnost bijeloj rasi nego crnoj. Naravno ako gledamo samo preciznost modela nećemo doći do tog zaključka te je bitno proučiti matricu zabune koja nam otkriva ovaj problem u našem modelu. Naime jasno vidimo

```

table(antrData_blackwhite$DODRace)

##
##      1    2
## 2817  642

table(antrData_blackwhite_female$DODRace)

##
##      1    2
## 975  656

```

Isto tako je jasno da je u ženskim podacima isti problem značajno manji i ta razlika možda predstavlja temelje za zanimljivu analizu mogući uzroka ali to je van domene ove analize.

S obzirom na više nego dovoljan broj podataka umjetno ćemo stvoriti balansirane skupove tako što ćemo ispremješati redove u tablici te nausmješno izabrati retke koji će ulaziti u model. Kako bi dobili ponovljive podatke seed ćemo postaviti na fiksnu vrijednost

```

equalData <- function(df){
  size <- nrow(df[df$DODRace == 2, ])
  count <- 0
  indices <- c()

```

```

for(i in 1:nrow(df)){
  if(df$DODRace[i] == 1){
    if(count >= size){
      indices[i] <- FALSE
    } else{
      count <- count + 1
      indices[i] <- TRUE
    }
  } else {
    indices[i] <- TRUE
  }
}
return(indices)
}

set.seed(42)

rows <- sample(nrow(antrData_blackwhite))
rows_female <- sample(nrow(antrData_blackwhite_female))

antrData_blackwhite_shuffle <- antrData_blackwhite[rows, ]
antrData_blackwhite_female_shuffle <- antrData_blackwhite_female[rows_female, ]

indices <- equalData(antrData_blackwhite_shuffle)
indices_female <- equalData(antrData_blackwhite_female_shuffle)

antrData_blackwhite_shuffle <- antrData_blackwhite_shuffle[indices, ]
nrow(antrData_blackwhite_shuffle[antrData_blackwhite_shuffle$DODRace == 2, ])

## [1] 642
nrow(antrData_blackwhite_shuffle[antrData_blackwhite_shuffle$DODRace == 1, ])

## [1] 642
antrData_blackwhite_female_shuffle <- antrData_blackwhite_female_shuffle[indices_female, ]
nrow(antrData_blackwhite_female_shuffle[antrData_blackwhite_female_shuffle$DODRace == 2, ])

## [1] 656
nrow(antrData_blackwhite_female_shuffle[antrData_blackwhite_female_shuffle$DODRace == 1, ])

## [1] 656

```

Nakon balansiranja broj podataka iz obije kategorije pristupamo ponovnoj konstrukciji histograma koji bi u ovom slučaju trebali biti još indikativniji.

```

mean_male_white <- mean(antrData_blackwhite_shuffle$spanheightratio[antrData_blackwhite_shuffle$DODRace == 1])
mean_male_black <- mean(antrData_blackwhite_shuffle$spanheightratio[antrData_blackwhite_shuffle$DODRace == 2])

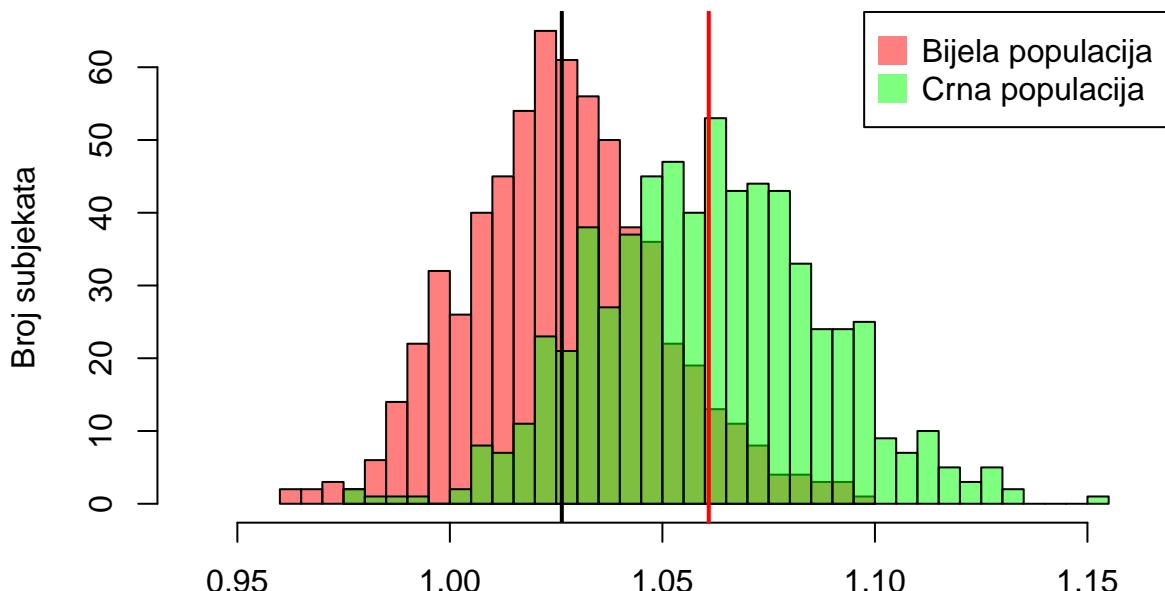
hist(antrData_blackwhite_shuffle$spanheightratio[antrData_blackwhite_shuffle$DODRace == 1], breaks = 30)
hist(antrData_blackwhite_shuffle$spanheightratio[antrData_blackwhite_shuffle$DODRace == 2], breaks = 30)

abline(v = mean_male_white, col = "black", lwd = 2)
abline(v = mean_male_black, col = "red", lwd = 2)

```

```
legend('topright', legend = c("Bijela populacija", "Crna populacija"), col = c(rgb(1,0,0,0.5), rgb(0,1,0,0.5)),
```

Omjer medu muškom populacijom



Omjer raspona ruku i visine

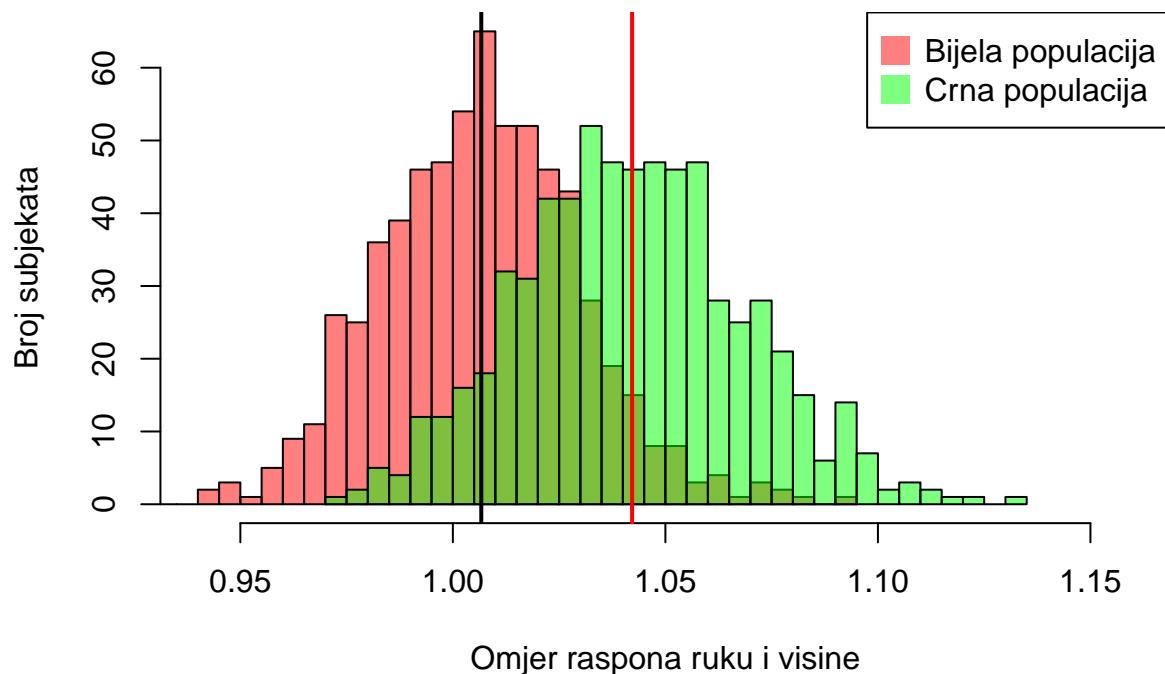
```
mean_female_white <- mean(antrData_blackwhite_female_shuffle$spanheightratio[antrData_blackwhite_female_shuffle$DODRace == 1])
mean_female_black <- mean(antrData_blackwhite_female_shuffle$spanheightratio[antrData_blackwhite_female_shuffle$DODRace == 2])

hist(antrData_blackwhite_female_shuffle$spanheightratio[antrData_blackwhite_female_shuffle$DODRace == 1], col = "black")
hist(antrData_blackwhite_female_shuffle$spanheightratio[antrData_blackwhite_female_shuffle$DODRace == 2], col = "red")

abline(v = mean_female_white, col = "black", lwd = 2)
abline(v = mean_female_black, col = "red", lwd = 2)

legend('topright', legend = c("Bijela populacija", "Crna populacija"), col = c(rgb(1,0,0,0.5), rgb(0,1,0,0.5)),
```

Omjer među ženskom populacijom



Sada se u histogramu jasnije vidi razlika između dviju populacija i jasnija je distribucija dviju populacija.

Sada imamo jednak broj nausimčno odabralih pripadnika bijele rase kao i pripadnika crne rase. S time eliminiramo pristranost modela prema bijeloj rasi i možemo bolje proučavati kako se omjer ponaša te ćemo sad izgraditi i model za ženski dio populacije.

```
logitSpan_corrected <- glm(DODRace_1 ~ spanheightratio, data=antrData_blackwhite_shuffle, family='binomial')
summary((logitSpan_corrected))
```

```
##
## Call:
## glm(formula = DODRace_1 ~ spanheightratio, family = "binomial",
##      data = antrData_blackwhite_shuffle)
##
## Deviance Residuals:
##      Min        1Q        Median         3Q        Max
## -2.73836   -0.78047    0.03691    0.80041    2.51341
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 59.644     3.445   17.31  <2e-16 ***
## spanheightratio -57.199     3.306  -17.30  <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: 1780.0  on 1283  degrees of freedom
## Residual deviance: 1268.2  on 1282  degrees of freedom
## AIC: 1272.2
##
## Number of Fisher Scoring iterations: 5
logitSpan_female <- glm(DODRace_1 ~ spanheightratio, data=antrData_blackwhite_female_shuffle, family='binomial')
summary(logitSpan_female)

##
## Call:
## glm(formula = DODRace_1 ~ spanheightratio, family = "binomial",
##      data = antrData_blackwhite_female_shuffle)
##
## Deviance Residuals:
##      Min        1Q     Median        3Q       Max
## -2.45955 -0.78503  0.00171  0.80658  2.77810
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 58.940     3.379   17.44 <2e-16 ***
## spanheightratio -57.578     3.303  -17.43 <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: 1818.8  on 1311  degrees of freedom
## Residual deviance: 1285.4  on 1310  degrees of freedom
## AIC: 1289.4
##
## Number of Fisher Scoring iterations: 5

```

Sada su rezultati puno kvalitetniji, primjećujemo prvo kako je medijan reziduala jako blizu nuli što je dobar indikator da model nije pristrand. Isto tako devijacija reziduala je značajno manja od devijacije nul modela za oba modela. Nadalje obije devijacije reziduala su jako blizu broju stupnjeva slobode što je dobar indikator za kvalitetu modela.

```

logitSpan_probs_corrected <- predict(logitSpan_corrected, type="response")
logitSpan_probs_female <- predict(logitSpan_female, type="response")

logitSpan_predictions_corrected <- ifelse(logitSpan_probs_corrected > 0.5, 1, 0)
conffMatrix <- as.data.frame(table(logitSpan_predictions_corrected, antrData_blackwhite_shuffle$DODRace))
print(conffMatrix)

##    logitSpan_predictions_corrected Var2 Freq
## 1                               0   0  476
## 2                               1   0  166
## 3                               0   1  142
## 4                               1   1  500

logitSpan_predictions_female <- ifelse(logitSpan_probs_female > 0.5, 1, 0)
conffMatrix_female <- as.data.frame(table(logitSpan_predictions_female, antrData_blackwhite_female_shuffle))
print(conffMatrix_female)

##    logitSpan_predictions_female Var2 Freq
## 1                           0   0  495

```

```

## 2          1     0   161
## 3          0     1   150
## 4          1     1   506
accuracy <- (confMatrix[1, 3] + confMatrix[4, 3])/length(logitSpan_predictions_corrected)
print(accuracy)

## [1] 0.7601246

accuracy_female <- (confMatrix_female[1, 3] + confMatrix_female[4, 3])/length(logitSpan_predictions_corrected)
print(accuracy_female)

## [1] 0.7629573

Iako nam dolazi manja točnost nego u ne prilagodenom modelu gdje nije jednak broj elemenata svake klase treba uzeti u obzir kako točnost sama po sebi nije dobar prediktor kvalitete logističkog modela te je ovaj model neosporno kvalitetniji od prethodnog te možemo vidjeti da model postiže značajno bolju točnost od nasumičnog predviđanja i za muškarce i za žene.

Sljedeći korak nam može biti dodatno filtriranje podataka s genetske strane jer DODRace stupac predstavlja osobnu preferiranu rasu. S obzirom da velika većina ljudi u ovom modernom vremenu globalizacije potječe iz različitih rasnih skupina možemo filtrirati po stupcu SubjectNumericRace koji predstavlja potpuni prikaz podrijetla osobe tako da uzmemo osobe koje su prijavile samo crnu ili bijelu rasu kao svoje podrijetlo.

antrData_blackwhiteonly <- antrData_male[antrData_male$SubjectNumericRace == 1 | antrData_male$SubjectNumericRace == 2]
antrData_blackwhiteonly_female <- antrData_female[antrData_female$SubjectNumericRace == 1 | antrData_female$SubjectNumericRace == 2]

antrData_blackwhiteonly$spanheightratio <- antrData_blackwhiteonly$span / antrData_blackwhiteonly$status
antrData_blackwhiteonly_female$spanheightratio <- antrData_blackwhiteonly_female$span / antrData_blackwhiteonly_female$status

rows <- sample(nrow(antrData_blackwhiteonly))
rows_female <- sample(nrow(antrData_blackwhiteonly_female))

antrData_blackwhiteonly_shuffle <- antrData_blackwhiteonly[rows, ]
antrData_blackwhiteonly_female_shuffle <- antrData_blackwhiteonly_female[rows_female, ]

indices <- equalData(antrData_blackwhiteonly_shuffle)
indices_female <- equalData(antrData_blackwhiteonly_female_shuffle)

antrData_blackwhiteonly_shuffle <- antrData_blackwhiteonly_shuffle[indices, ]
nrow(antrData_blackwhiteonly_shuffle[antrData_blackwhiteonly_shuffle$DODRace == 2, ])

## [1] 502
nrow(antrData_blackwhiteonly_shuffle[antrData_blackwhiteonly_shuffle$DODRace == 1, ])

## [1] 502
antrData_blackwhiteonly_female_shuffle <- antrData_blackwhiteonly_female_shuffle[indices_female, ]
nrow(antrData_blackwhiteonly_female_shuffle[antrData_blackwhiteonly_female_shuffle$DODRace == 2, ])

## [1] 525
nrow(antrData_blackwhiteonly_female_shuffle[antrData_blackwhiteonly_female_shuffle$DODRace == 1, ])

## [1] 525
mean_male_white <- mean(antrData_blackwhiteonly_shuffle$spanheightratio[antrData_blackwhiteonly_shuffle$DODRace == 1])
mean_male_black <- mean(antrData_blackwhiteonly_shuffle$spanheightratio[antrData_blackwhiteonly_shuffle$DODRace == 2])

```

```

hist(antrData_blackwhiteonly_shuffle$spanheightratio[antrData_blackwhiteonly_shuffle$SubjectNumericRace == "Bijela populacija"], col = "red", main = "Omjer medu muškom populacijom", xlab = "Omjer raspona ruku i visine", ylab = "Broj subjekata", xaxt = "top", xaxs = "dfrac", xaxp = c(0.95, 1.15, 10), yaxt = "top", yaxs = "dfrac", yaxp = c(0, 50, 10), density = 10, border = "black", breaks = 10, plot = TRUE)
hist(antrData_blackwhiteonly_shuffle$spanheightratio[antrData_blackwhiteonly_shuffle$SubjectNumericRace == "Crna populacija"], col = "green", main = "Omjer medu muškom populacijom", xlab = "Omjer raspona ruku i visine", ylab = "Broj subjekata", xaxt = "top", xaxs = "dfrac", xaxp = c(0.95, 1.15, 10), yaxt = "top", yaxs = "dfrac", yaxp = c(0, 50, 10), density = 10, border = "black", breaks = 10, plot = TRUE)

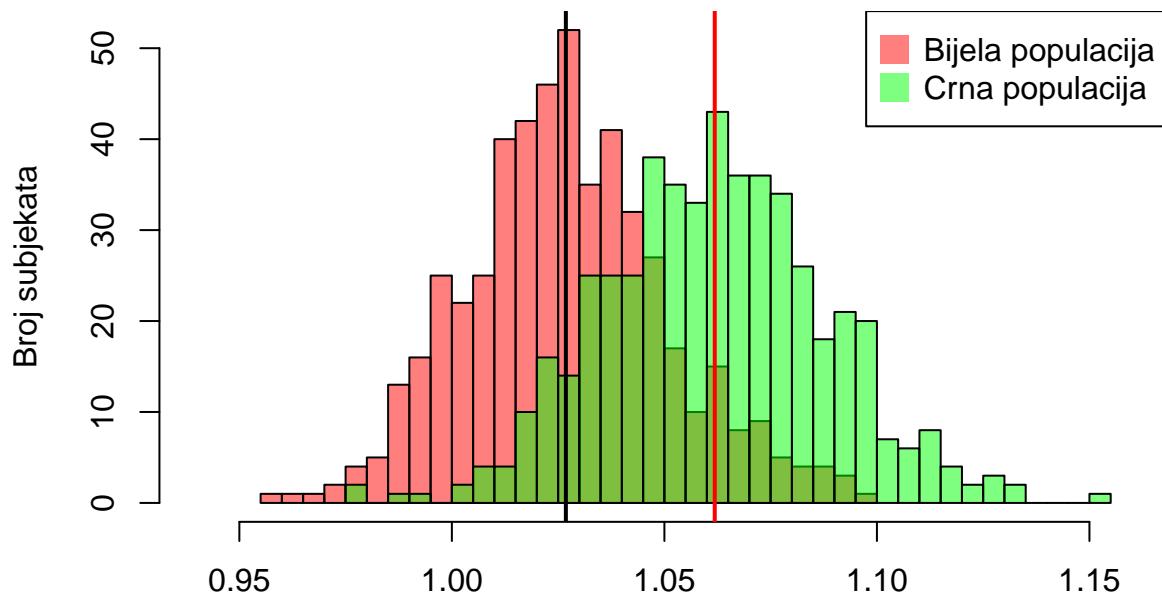
mean_male_white <- mean(antrData_blackwhiteonly_shuffle$spanheightratio[antrData_blackwhiteonly_shuffle$SubjectNumericRace == "Bijela populacija"])
mean_male_black <- mean(antrData_blackwhiteonly_shuffle$spanheightratio[antrData_blackwhiteonly_shuffle$SubjectNumericRace == "Crna populacija"])

abline(v = mean_male_white, col = "black", lwd = 2)
abline(v = mean_male_black, col = "red", lwd = 2)

legend('topright', legend = c("Bijela populacija", "Crna populacija"), col = c(rgb(1,0,0,0.5), rgb(0,1,0,0.5)), bty = "n", border = "black", horiz = FALSE)

```

Omjer medu muškom populacijom



Omjer raspona ruku i visine

```

mean_female_white <- mean(antrData_blackwhiteonly_female_shuffle$spanheightratio[antrData_blackwhiteonly_shuffle$SubjectNumericRace == "Bijela populacija"])
mean_female_black <- mean(antrData_blackwhiteonly_female_shuffle$spanheightratio[antrData_blackwhiteonly_shuffle$SubjectNumericRace == "Crna populacija"])

hist(antrData_blackwhiteonly_female_shuffle$spanheightratio[antrData_blackwhiteonly_female_shuffle$SubjectNumericRace == "Bijela populacija"], col = "red", main = "Omjer raspona ruku i visine", xlab = "Omjer raspona ruku i visine", ylab = "Broj subjekata", xaxt = "top", xaxs = "dfrac", xaxp = c(0.95, 1.15, 10), yaxt = "top", yaxs = "dfrac", yaxp = c(0, 50, 10), density = 10, border = "black", breaks = 10, plot = TRUE)
hist(antrData_blackwhiteonly_female_shuffle$spanheightratio[antrData_blackwhiteonly_female_shuffle$SubjectNumericRace == "Crna populacija"], col = "green", main = "Omjer raspona ruku i visine", xlab = "Omjer raspona ruku i visine", ylab = "Broj subjekata", xaxt = "top", xaxs = "dfrac", xaxp = c(0.95, 1.15, 10), yaxt = "top", yaxs = "dfrac", yaxp = c(0, 50, 10), density = 10, border = "black", breaks = 10, plot = TRUE)

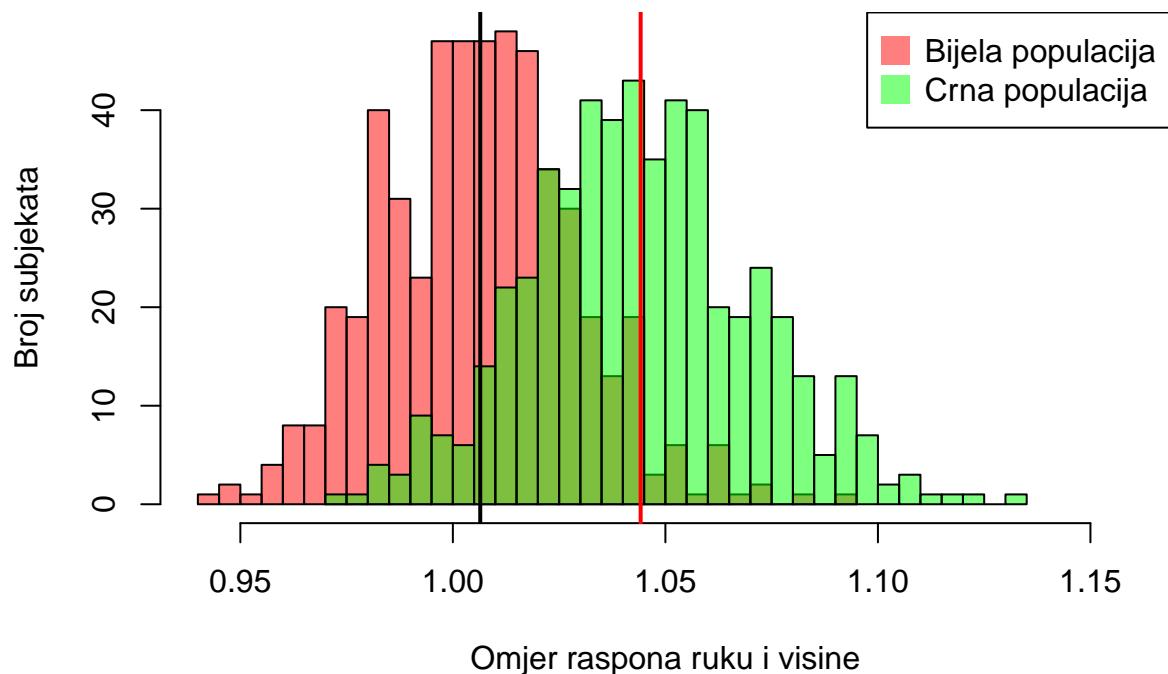
mean_female_white <- mean(antrData_blackwhiteonly_female_shuffle$spanheightratio[antrData_blackwhiteonly_female_shuffle$SubjectNumericRace == "Bijela populacija"])
mean_female_black <- mean(antrData_blackwhiteonly_female_shuffle$spanheightratio[antrData_blackwhiteonly_female_shuffle$SubjectNumericRace == "Crna populacija"])

abline(v = mean_female_white, col = "black", lwd = 2)
abline(v = mean_female_black, col = "red", lwd = 2)

legend('topright', legend = c("Bijela populacija", "Crna populacija"), col = c(rgb(1,0,0,0.5), rgb(0,1,0,0.5)), bty = "n", border = "black", horiz = FALSE)

```

Omjer među ženskom populacijom



Zasada ne primjećujemo neku značajnu razliku iz samih histograma, ali nastavljamo s izgradnjom logističkog modela.

```

antrData_blackwhiteonly_shuffle = dummy_cols(antrData_blackwhiteonly_shuffle, select_columns = c("SubjectNumericRace_1", "spanheightratio"))
antrData_blackwhiteonly_female_shuffle = dummy_cols(antrData_blackwhiteonly_female_shuffle, select_columns = c("SubjectNumericRace_1", "spanheightratio"))

logitSpan_only <- glm(SubjectNumericRace_1 ~ spanheightratio, data=antrData_blackwhiteonly_shuffle, family = "binomial")
summary(logitSpan_only)

## 
## Call:
## glm(formula = SubjectNumericRace_1 ~ spanheightratio, family = "binomial",
##      data = antrData_blackwhiteonly_shuffle)
## 
## Deviance Residuals:
##      Min        1Q     Median        3Q       Max
## -2.7178   -0.7873    0.1538    0.7972    2.4043
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 57.800     3.761   15.37   <2e-16 ***
## spanheightratio -55.372     3.605  -15.36   <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: 1397.4  on 1007  degrees of freedom
## Residual deviance: 1000.3  on 1006  degrees of freedom
## AIC: 1004.3
##
## Number of Fisher Scoring iterations: 5
logitSpan_only_female <- glm(SubjectNumericRace_1 ~ spanheightratio, data=antrData_blackwhiteonly_female)
summary((logitSpan_only_female))

##
## Call:
## glm(formula = SubjectNumericRace_1 ~ spanheightratio, family = "binomial",
##      data = antrData_blackwhiteonly_female_shuffle)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.5880  -0.7284   0.1263   0.7464   2.8858
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 64.610     4.031   16.03 <2e-16 ***
## spanheightratio -63.063     3.937  -16.02 <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: 1458.37  on 1051  degrees of freedom
## Residual deviance:  975.31  on 1050  degrees of freedom
## AIC: 979.31
##
## Number of Fisher Scoring iterations: 5

```

I dalje ne primjećujemo značajnu razliku između filtriranih rasa i preferiranih rasa

```

logitSpan_probs_only <- predict(logitSpan_only, type="response")
logitSpan_probs_only_female <- predict(logitSpan_only_female, type="response")

logitSpan_predictions_only <- ifelse(logitSpan_probs_only > 0.5, 1, 0)
conffMatrix <- as.data.frame(table(logitSpan_predictions_only, antrData_blackwhiteonly_shuffle$Subject))
print(conffMatrix)

##   logitSpan_predictions_only Var2 Freq
## 1                           0   0 379
## 2                           1   0 123
## 3                           0   1 106
## 4                           1   1 400

logitSpan_predictions_only_female <- ifelse(logitSpan_probs_only_female > 0.5, 1, 0)
conffMatrix_female <- as.data.frame(table(logitSpan_predictions_only_female, antrData_blackwhiteonly_female))
print(conffMatrix_female)

##   logitSpan_predictions_only_female Var2 Freq
## 1                               0   0 407
## 2                               1   0 117
## 3                               0   1 104
## 4                               1   1 424

```

```

accuracy <- (confMatrix[1, 3] + confMatrix[4, 3])/length(logitSpan_predictions_only)
print(accuracy)

## [1] 0.7728175

accuracy <- (confMatrix_female[1, 3] + confMatrix_female[4, 3])/length(logitSpan_predictions_only_fem)
print(accuracy)

## [1] 0.789924

```

Vidimo kako se preciznost modela neznatno povećala ali kao što je već naznačeno preciznost nije uvijek najbolja mjera kvalitete modela. Iz priloženoga možemo vidjeti kako je preferirana rasa najčešće jednaka samoj rasi te kako bi za detaljniju analizu ovog omjera bilo potrebno detaljnije genetsko ispitivanje podrijetla koje je van domene ovog istraživanja.

Kao završnu zanimljivost testirati ćemo naš model na neviđenim podacima. Podaci su prikupljeni iz službenih mjera igrača američkog nogometa prilikom ulaska u ligu. Podaci su prikupljeni sa sljedećih stranica:

<https://www.cbssports.com/nfl/draft/news/nfl-combine-2020-tracker-measurements-results-of-the-best-individual-performances-from-indianapolis/> <https://nflcombineresults.com/playerpage.php?i=22415>

Podatke učitavamo i testiramo naš model na njima

```

testingData = read.csv("AddedData.csv")

testingData$spanheightratio <- testingData$span/testingData$height

logitSpan_probs_testing <- predict(logitSpan_only, newdata = testingData ,type="response")

logitSpan_predictions_testing <- ifelse(logitSpan_probs_testing > 0.5, 1, 0)
confMatrix <- as.data.frame(table(logitSpan_predictions_testing, testingData$race))
print(confMatrix)

##   logitSpan_predictions_testing Var2 Freq
## 1                           0     0    4
## 2                           1     0    1
## 3                           0     1    1
## 4                           1     1    4

accuracy <- (confMatrix[1, 3] + confMatrix[4, 3])/length(logitSpan_predictions_testing)
print(accuracy)

## [1] 0.8

```

Vidimo kako smo na ovom malom skupu podataka postigli točnost od 80%. Naravno ovako mali skup za testiranje nije relevantan za izvlačenje neki ozbiljnijih zaključaka ali je lijepo vidjeti kako naš model radi na dosad nevidenim podacima.

Zaključak

Izgradili smo logistički model za predviđanje rase osobe na temelju omjera raspona ruku i visine. Nadalje pokazano je postojanje razlike između bijele i crne rase u pogledu ovog omjera što predstavlja zanimljivu činjenicu s antropološkog i evolucijskog stajališta. Postojanje ove razlike bi moglo biti korisno proizvođačima vojne opreme. Isto tako ova razlika je vjerojatno jedan od razloga dominacije pripadnika crne rase u sportovima u kojima raspon ruku igra značajnu ulogu, kao što je košarka.

Određivanje optimalnih mjera kaciga na temelju antropometrijskih podataka američke

Motivacija

Motivacija za ovaj problem je prilično očita. Kao i bilo koji poslodavac koji je zadužen za opremanje svojih vojnika i vojsku zanima koliko je koje opreme potrebno te mjere same opreme. Određivanje broja potrebnih mjera kaciga i veličine tih potrebnih mjera svakako spada pod prije spomenuti interes vojske.

Istraživanje

Mjere kaciga trebale bi se odrediti na temelju antropometrijskih podataka koji imaju veze s glavom. Izdvojimo iz skupa podataka antropometrijske mjere koje bi mogle utjecati na odluku kategorija veličina. Te mjere su:

- bitragion chin arc - duljina linije koja se proteže od vrha jednog uha do vrha drugog uha, a putuje po licu prolazeći vrhom brade
- bitragion submandibular arc - duljina linije od vrha jednog uha do vrha drugog uha, ali ovaj put ispod vilice (kao zaštitni remen za kacigu)
- bizygomatic breadth - maksimalna horizontalna širina glave s prednje strane, od najispupčenijeg vrha jednog obraza do istog tog vrha drugo obraza
- ear breadth - maksimalna širina desnog uha okomito u odnosu na pravac koji prati uho po duljini
- ear length - maksimalna duljina desnog uha
- ear protrusion - maksimalna udaljenost između vanjskog ruba uha i glave ("mjera stršenja uha")
- head breadth - maksimalna horizontalna širina glave, mjerena između točaka glave koje se nalaze direktno iznad gornjeg ruba uha
- head circumference - maksimalan opseg glave, mjerjen po liniji koja se proteže od točke iznad gornjeg ruba uha, preko čela, do točke iznad gornjeg ruba uha na drugoj strani i preko najispupčenijeg stražnjeg dijela glave
- head length - maksimalna duljina glave, od najispupčenije točke na čelu do najispupčenije točke na stražnjem dijelu glave
- menton-sellion lenght - udaljenost od podnožja nosa između obrvi do najispupčenijeg dijela brade

napomena: sve veličine izražene su u milimetrima

```
ansur.II.data = read.csv("ANSUR_II_data.csv")
head = select(ansur.II.data, "bitragionchinarc", "bitragionsubmandibulararc", "bizygomaticbreadth", "earbreadth", "earlength", "earprotrusion", "headbreadth", "headcircumference")
summary(head)
```

```
##   bitragionchinarc bitragionsubmandibulararc bizygomaticbreadth   earbreadth
## Min.    :267.0      Min.    :245.0          Min.    :116.0      Min.    :25.00
## 1st Qu.:314.0      1st Qu.:292.0          1st Qu.:135.0      1st Qu.:33.00
## Median :326.0      Median :307.0          Median :140.0      Median :35.00
## Mean    :324.9      Mean    :306.6          Mean    :139.7      Mean    :35.13
## 3rd Qu.:336.0      3rd Qu.:321.0          3rd Qu.:145.0      3rd Qu.:37.00
## Max.    :385.0      Max.    :390.0          Max.    :174.0      Max.    :46.00
##   earlength     earprotrusion     headbreadth     headcircumference
## Min.    :46.00      Min.    :13.00      Min.    :131.0      Min.    :500
## 1st Qu.:59.00      1st Qu.:20.00      1st Qu.:148.0      1st Qu.:557
## Median :62.00      Median :22.00      Median :152.0      Median :570
## Mean    :62.63      Mean    :22.26      Mean    :152.2      Mean    :570
## 3rd Qu.:66.00      3rd Qu.:24.00      3rd Qu.:156.0      3rd Qu.:582
## Max.    :81.00      Max.    :34.00      Max.    :180.0      Max.    :635
```

```

##      headlength      mentonsellionlength
##  Min.   :168.0      Min.   : 91.0
##  1st Qu.:191.0      1st Qu.:114.0
##  Median :197.0      Median :120.0
##  Mean   :196.3      Mean   :119.5
##  3rd Qu.:202.0      3rd Qu.:125.0
##  Max.   :225.0      Max.   :156.0

sapply(head, class)

##          bitragionchinarc bitragionsubmandibulararc      bizygomaticbreadth
##                  "integer"           "integer"             "integer"
##          earbreadth          earlength            earprotrusion
##                  "integer"           "integer"             "integer"
##          headbreadth         headcircumference      headlength
##                  "integer"           "integer"             "integer"
##          mentonsellionlength
##                  "integer"

```

Vidimo da su sve varijable numeričkog tipa. Za određivanje kategorija veličina kacige, izdvojimo dodatno iz skupa podataka samo neke, koji se čine najznačajnijima za veličinu kacige. Potrebni su nam duljina, širina i opseg glave za određivanje mjera same kacige. Ostale varijable nisu toliko značajne u određivanju veličina kacige, stoga ćemo u skupu podataka zadržati samo navedene.

```

helmet = select(head, "headbreadth", "headcircumference", "headlength")
summary(helmet)

```

```

##   headbreadth   headcircumference   headlength
##  Min.   :131.0    Min.   :500       Min.   :168.0
##  1st Qu.:148.0    1st Qu.:557       1st Qu.:191.0
##  Median :152.0    Median :570       Median :197.0
##  Mean   :152.2    Mean   :570       Mean   :196.3
##  3rd Qu.:156.0    3rd Qu.:582       3rd Qu.:202.0
##  Max.   :180.0    Max.   :635       Max.   :225.0

```

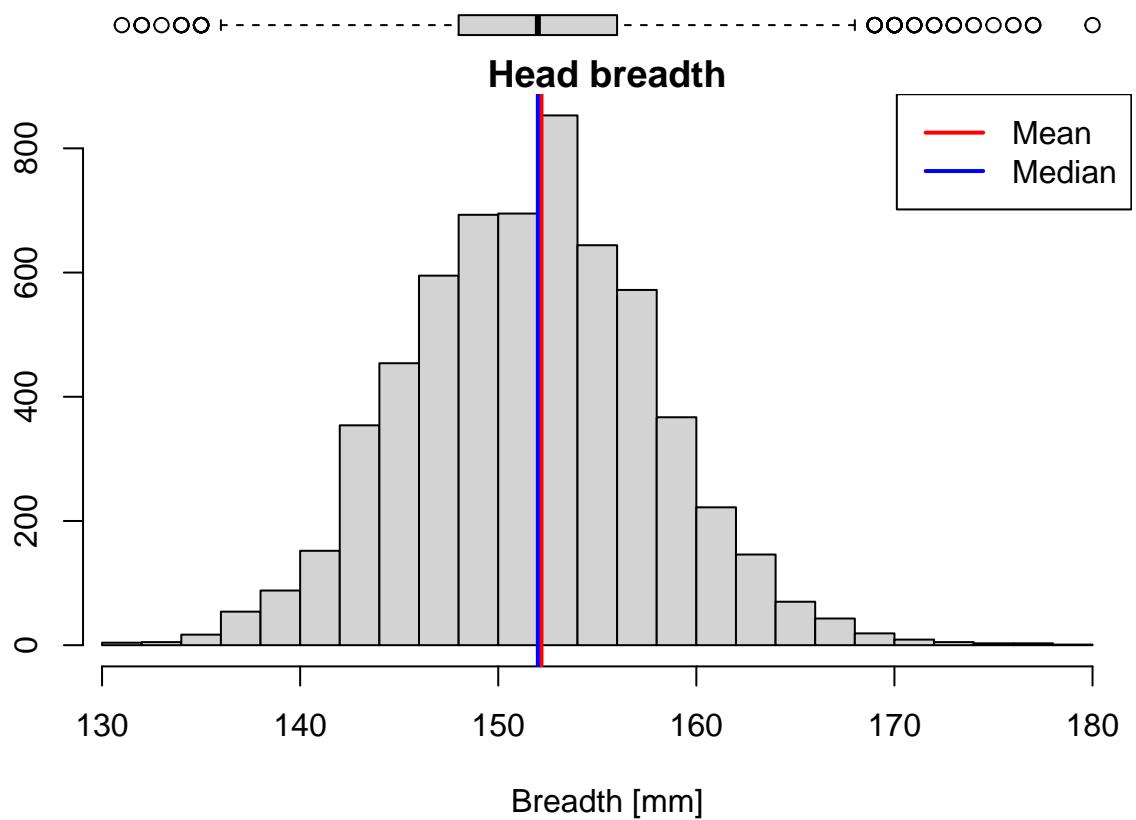
Provodimo deskriptivnu analizu varijabli prema kojima modeliramo veličine kaciga.

```

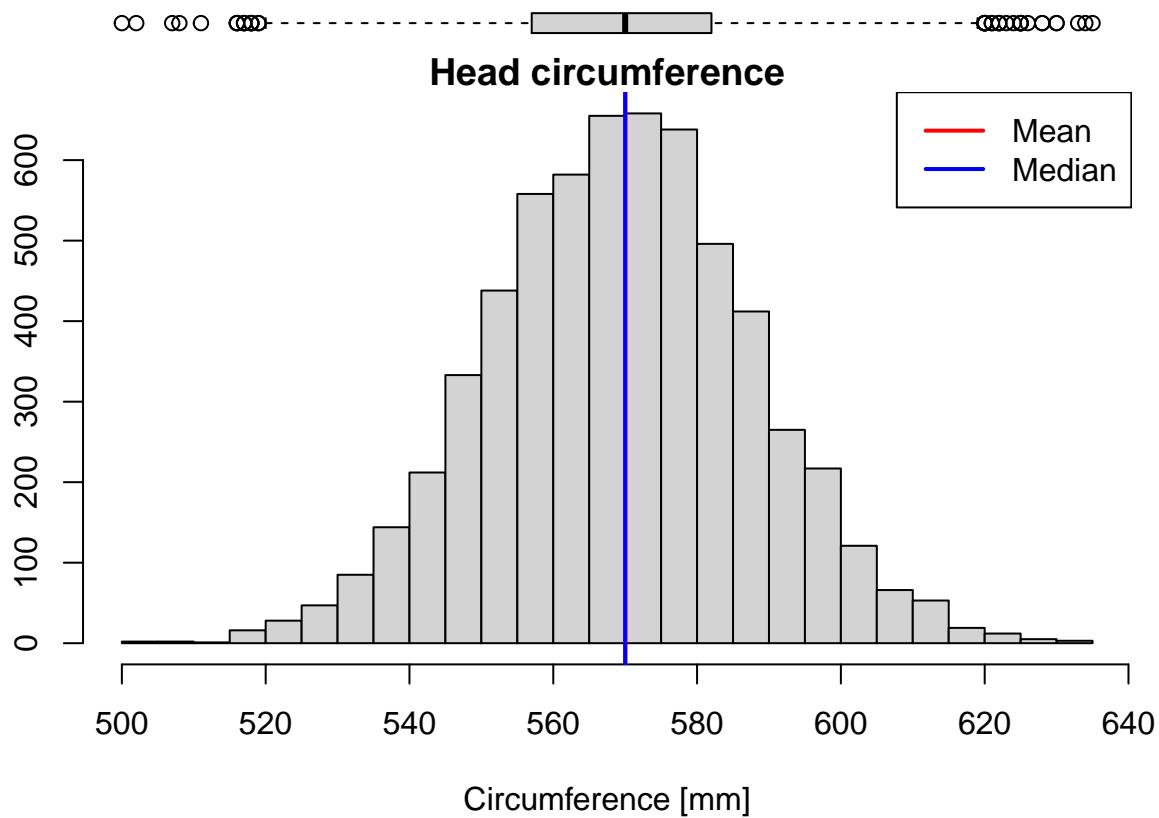
plot_summary <- function(data, main, xlab) {
  layout(mat=matrix(c(1,2),2,1,byrow=TRUE), height=c(1,8))
  par(mar=c(0, 3.1, 1.1, 2.1))
  boxplot(data, horizontal=TRUE, xaxt="n", frame=FALSE, ylim=c(min(data), max(data)))
  par(mar=c(4, 3.1, 1.1, 2.1))
  hist(data, breaks=30, main=main, xlab=xlab, ylab="Frequency", xlim = c(min(data), max(data)))
  abline(v=mean(data), col="red", lwd=2)
  abline(v=median(data), col="blue", lwd=2)
  legend(x="topright", c("Mean", "Median"), col=c("red", "blue"), lwd=c(2,2))
}

plot_summary(data = helmet$headbreadth, main="Head breadth", xlab="Breadth [mm]")

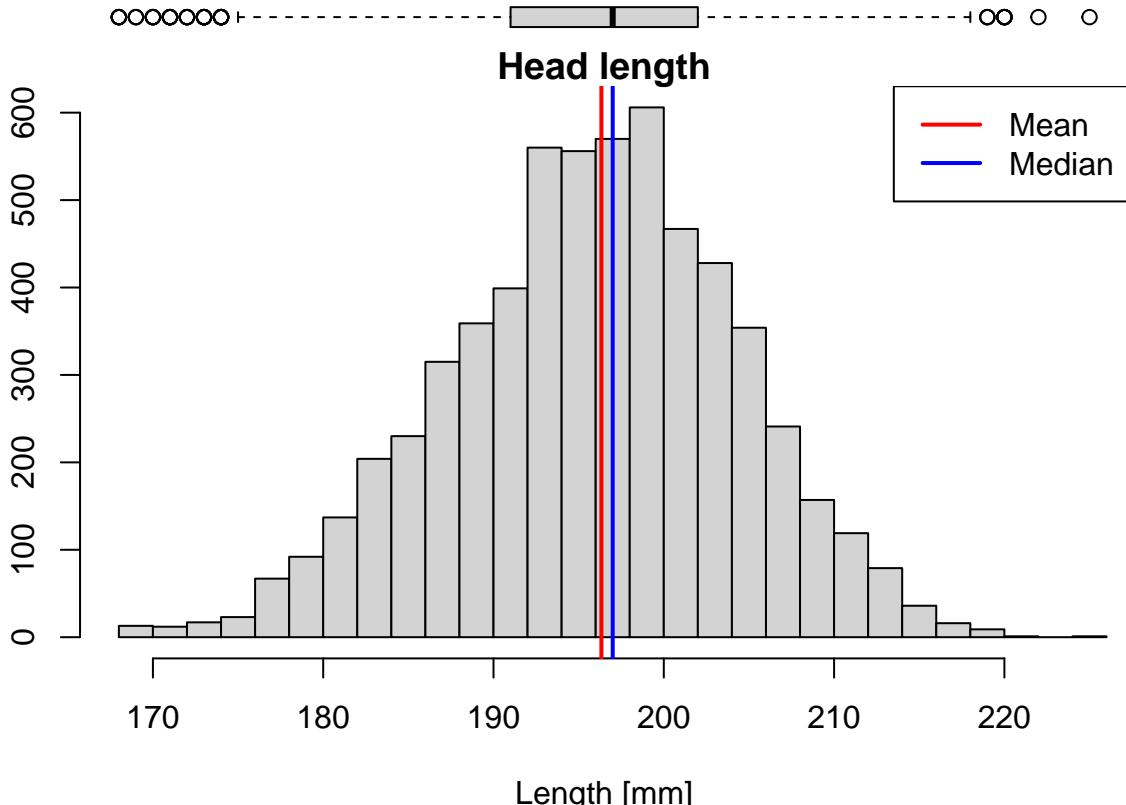
```



```
plot_summary(data = helmet$headcircumference, main="Head circumference", xlab="Circumference [mm]")
```



```
plot_summary(data = helmet$headlength, main="Head length", xlab="Length [mm] ")
```



Iz vizualizacije varijabli lijepo se vidi da su sve distribucije unimodalne i simetrične, s relativno uskim interkvartilnim rangom u odnosu na širinu distribucije.

Za određivanje veličina kacige, recimo da za svakog vojnika treba postojati kaciga koja odgovara njegovim mjerama glave, s maksimalnim dozvoljenim odstupanjem od 2 cm. Uz to, valja naglasiti da je puno veći problem kaciga koja je premala nego kaciga koja je prevelika, jer premašu kacigu vojnik fizički ne može obući. Zato ćemo kao najveću veličinu uzeti kacigu koja odgovara najvećim vrijednostima varijabli koje predstavljaju mjere glave.

Valja odrediti i broj različitih veličina kacige. Varijabli s najvećim rangom bit će potrebno najviše različitih veličina kacige, dok se ostale varijable s manjim rangom mogu uže podijeliti, radi ravnomjernosti podjele.

Bitna pretpostavka ovog postupka određivanja veličina kaciga je da su varijable koje predstavljaju mjere glave međusobno ovisne. Na primjer, ako napravimo 5 različitih veličina kaciga, te ako je nekom vojniku opseg glave u rangu 3. veličine, znači li to da će mu nužno i širina i duljina glave odgovarati kategoriji 3. veličine?

Smisleno je prepostaviti da postoji značajna korelacija između tih varijabli, konkretno da duljina i širina glave dobro određuju opseg. Ako glavu modeliramo kao elipsoid i prepostavimo da su sve varijable izmjerene u istoj ravnini, onda u toj ravnini opseg glave predstavlja opseg elipse, čija glavna os leži na pravcu duž kojeg je izmjerena duljina glave, a sporedna os leži na pravcu duž kojeg je izmjerena širina.

U egzaktnom slučaju, opseg elipse može se izračunati pomoću integrala

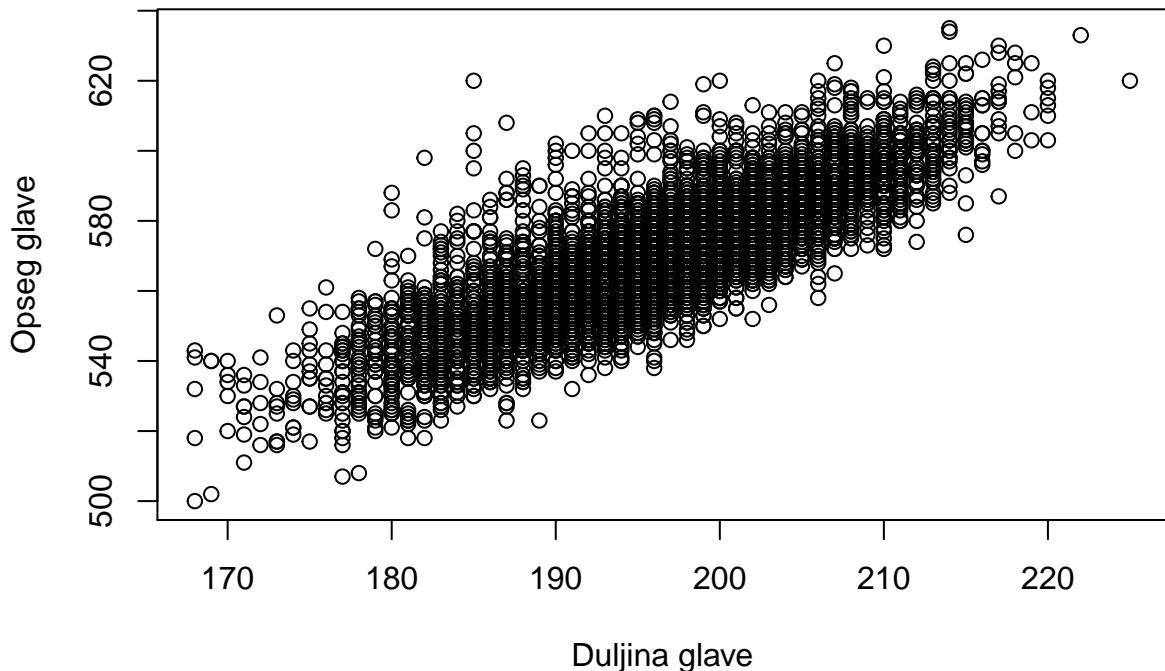
$$4a \int_a^{2\pi} \sqrt{1 - e^2 \sin^2 \theta} d\theta$$

, gdje je a duljina glavne poluos (u ovom slučaju polovica duljine glave), a $e = \sqrt{a^2 - b^2}$.

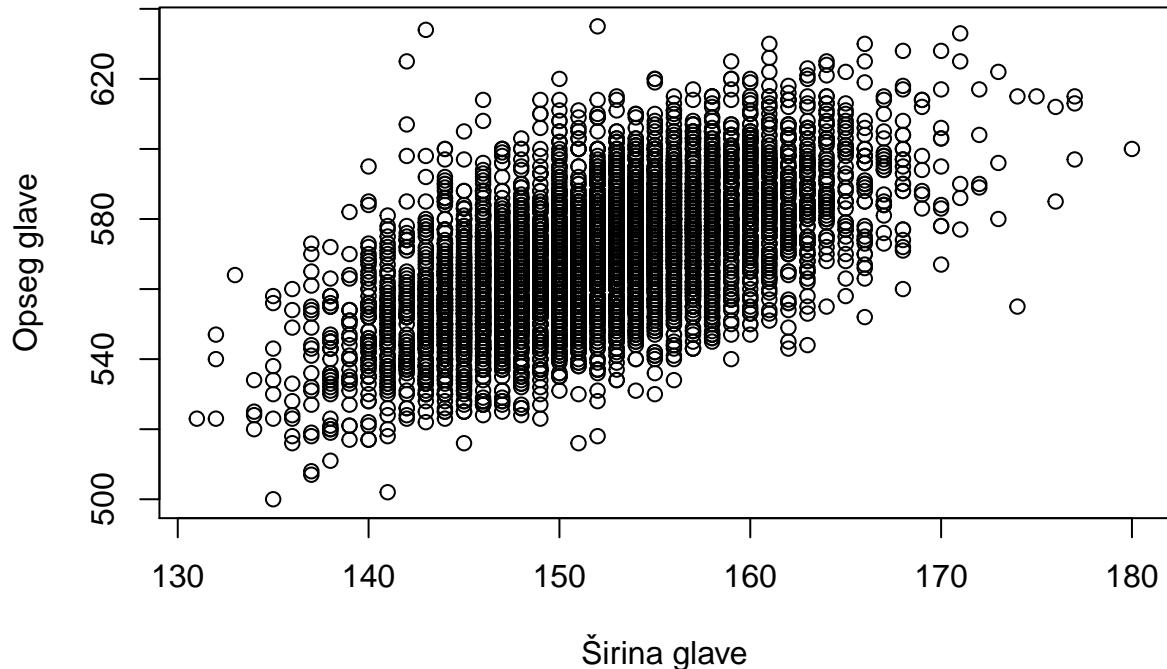
Iz formula je vidljivo da opseg definitivno ovisi o objema poluosima, ali varijable koje imamo ne odgovaraju egzaktnom slučaju, kao prvo zato što glava u stvarnom svijetu nije savršeni elipsoid, a kao drugo zato što nisu sve veličine izmjerene u istoj ravnini. Zato ćemo pomoću linearne regresije probati odrediti koliko u stvarnosti duljina i širina glave određuju opseg.

Pogledajmo dijagram raspršenja posebno za duljinu i opseg, te širinu i opseg.

```
plot(x = helmet$headlength, y=helmet$headcircumference, xlab="Duljina glave", ylab="Opseg glave")
```



```
plot(x = helmet$headbreadth, y=helmet$headcircumference, xlab="Širina glave", ylab="Opseg glave")
```



U oba slučaja vidljiv je pozitivan trend, te se čini da duljina glave puno jasnije opisuje opseg glave nego njena širina, ali obje su varijable dobri kandidati za regresore.

Da bismo vidjeli koliko su varijable značajne u procjenjivanju opsega glave, napravimo jednostavnu regresiju.

```
fit.length = lm(headcircumference~headlength, data=helmet)
plot(helmet$headlength, helmet$headcircumference)
#line(helmet$headlength, fit.length$fitted.values) #why is this not working tho
summary(fit.length)
```

```
##
## Call:
## lm(formula = headcircumference ~ headlength, data = helmet)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -34.157  -7.005  -0.761   6.041  69.867 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 225.30735    3.15526   71.41 <2e-16 ***
## headlength     1.75582    0.01606  109.36 <2e-16 ***
## ---      
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 10.61 on 6066 degrees of freedom
## Multiple R-squared:  0.6635, Adjusted R-squared:  0.6634
```

```

## F-statistic: 1.196e+04 on 1 and 6066 DF, p-value: < 2.2e-16
fit.breadth = lm(headcircumference~headbreadth, data=helmet)
summary(fit.breadth)

##
## Call:
## lm(formula = headcircumference ~ headbreadth, data = helmet)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -53.456  -9.290  -0.191   9.243  80.136 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 302.04056   4.58756   65.84   <2e-16 ***
## headbreadth    1.76101   0.03012   58.47   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.63 on 6066 degrees of freedom
## Multiple R-squared:  0.3604, Adjusted R-squared:  0.3603 
## F-statistic: 3418 on 1 and 6066 DF, p-value: < 2.2e-16

```

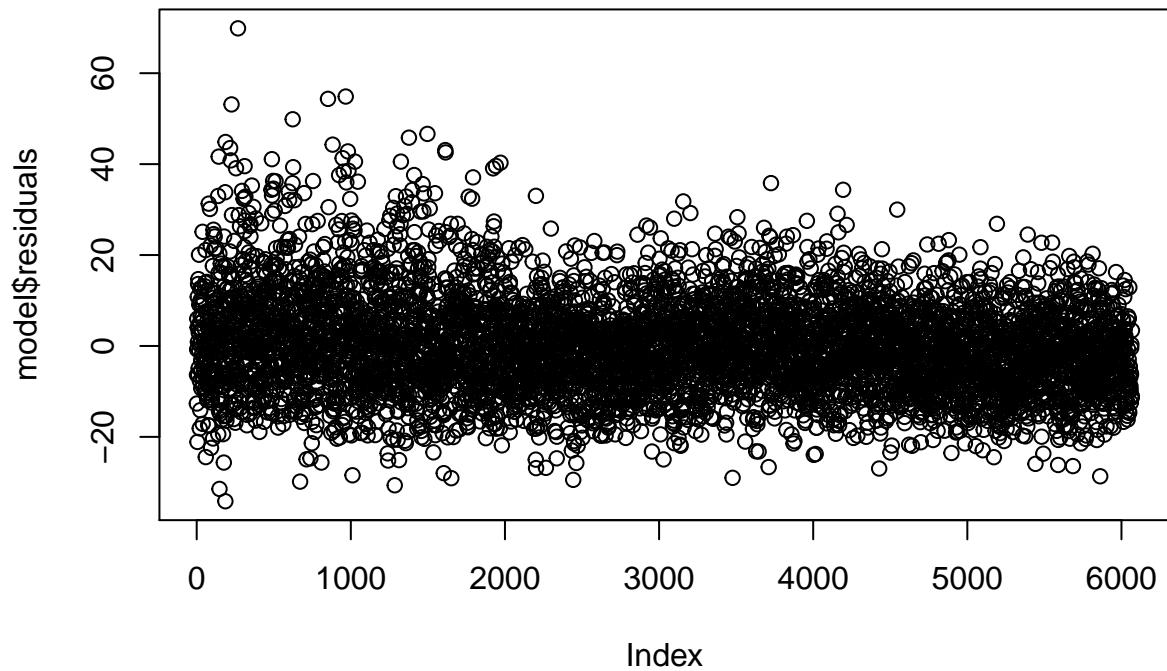
Prije interpretacije rezultata modela, valja provjeriti pretpostavke o normalnosti reziduala i homogenosti varijance.

```

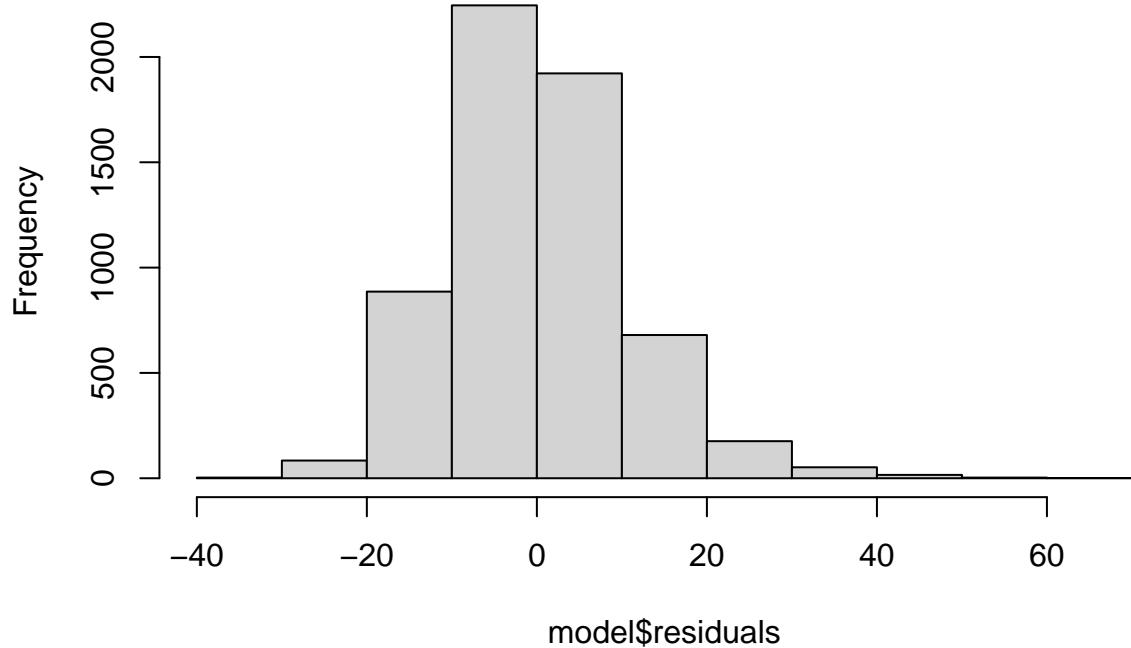
check_assumptions <- function(model) {
  plot(model$residuals)
  hist(model$residuals, main="Histogram reziduala")
  hist(rstandard(model), main="Histogram standardiziranih reziduala")
  qqnorm(rstandard(model), main="QQ plot")
  qqline(rstandard(model), col="red")
  plot(model$fitted.values, model$residuals)
  require(nortest)
  lillie.test(rstandard(model))
}

check_assumptions(fit.length)

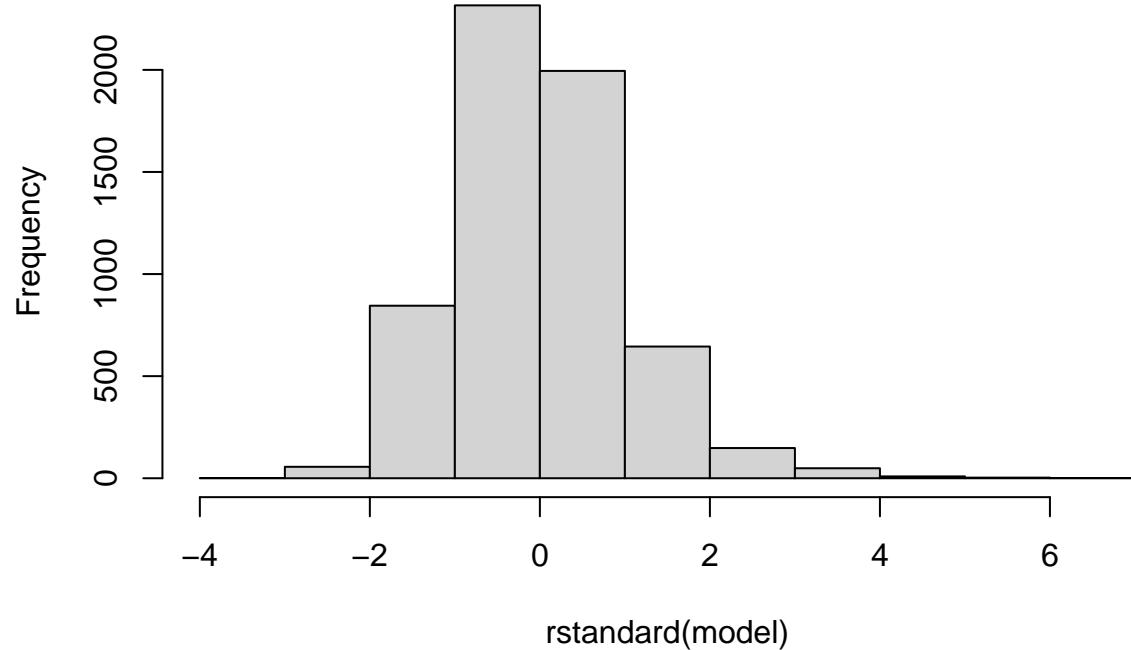
```



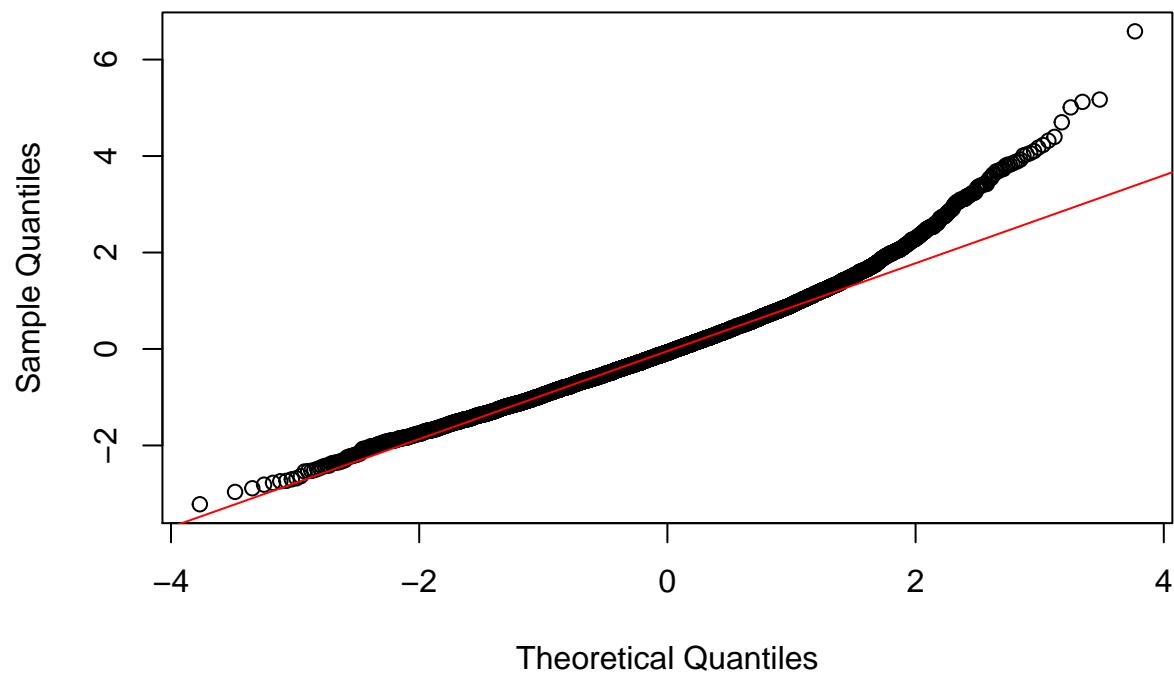
Histogram reziduala

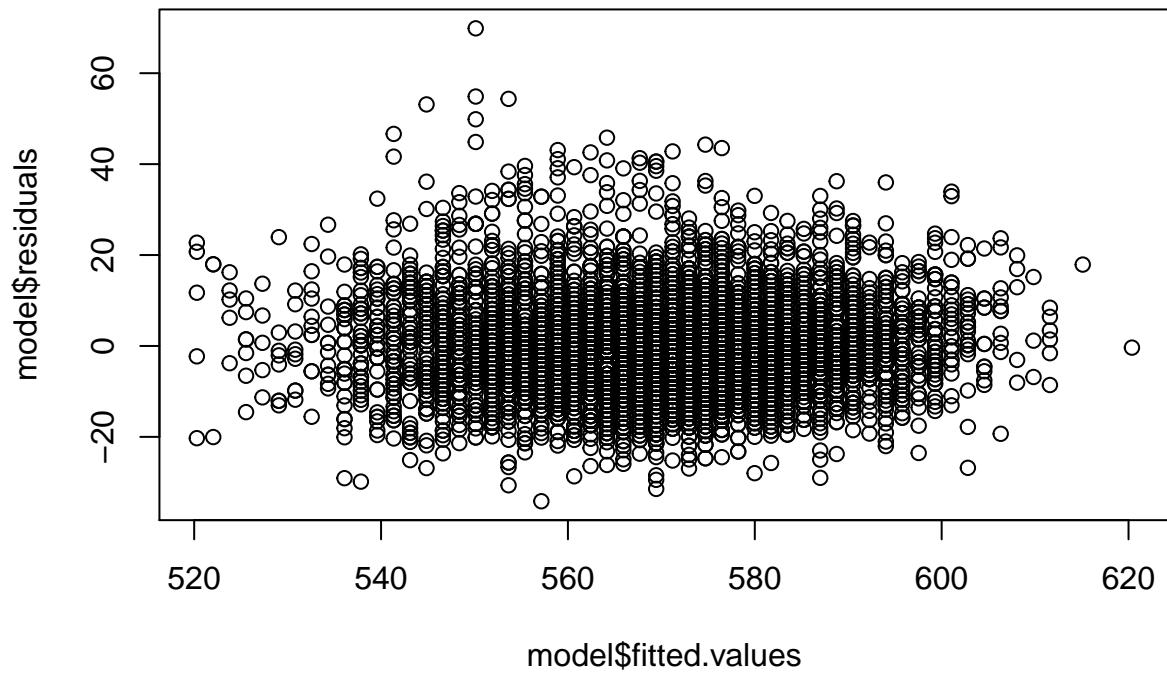


Histogram standardiziranih reziduala

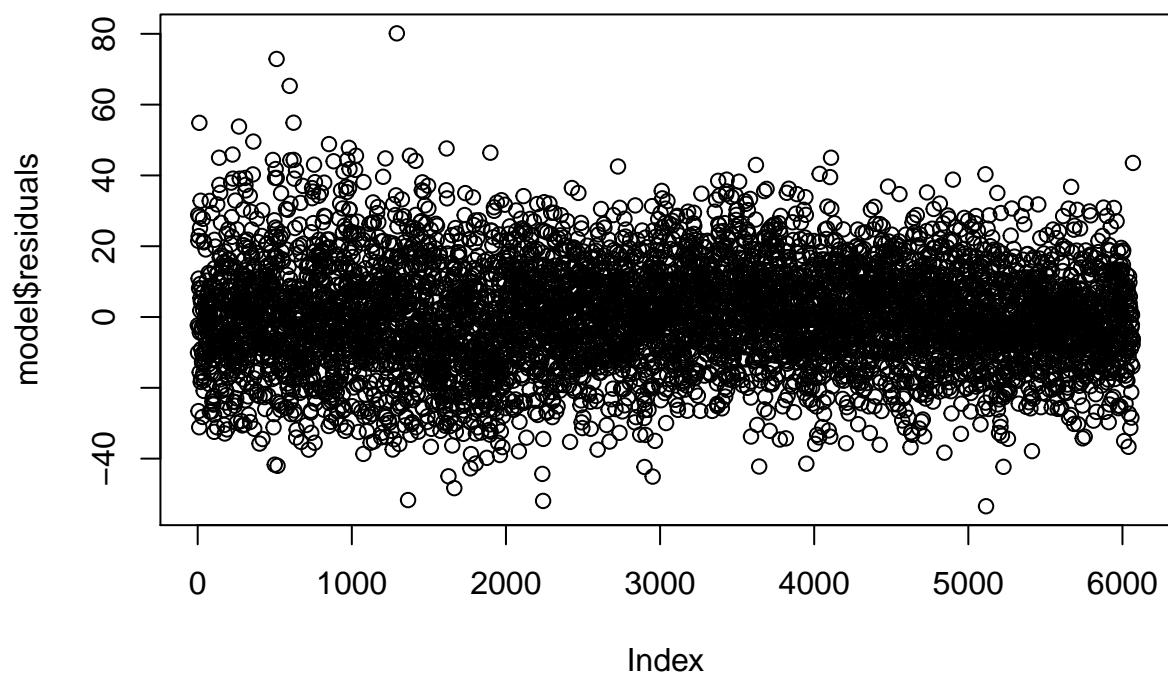


QQ plot

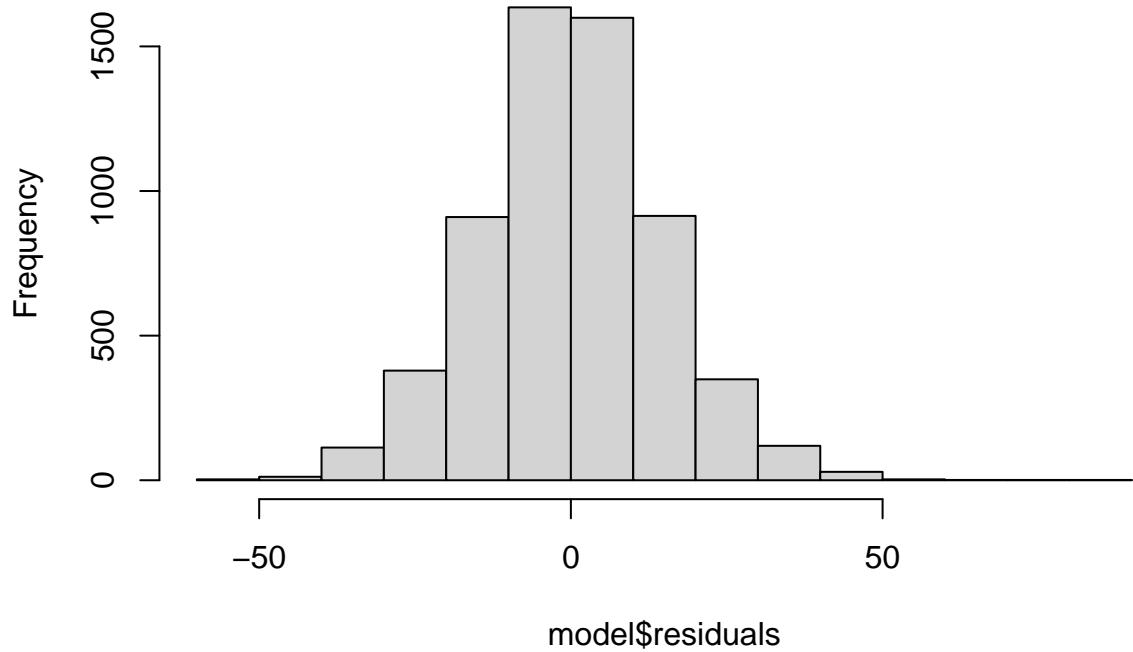




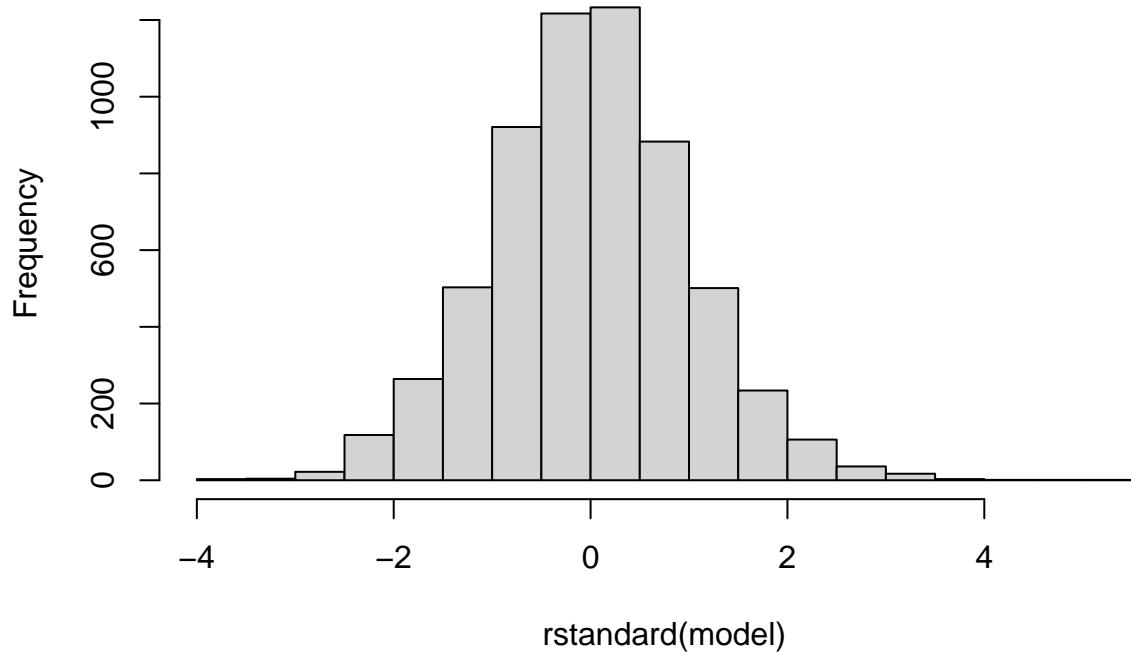
```
##  
## Lilliefors (Kolmogorov-Smirnov) normality test  
##  
## data: rstandard(model)  
## D = 0.039538, p-value < 2.2e-16  
check_assumptions(fit.breadth)
```



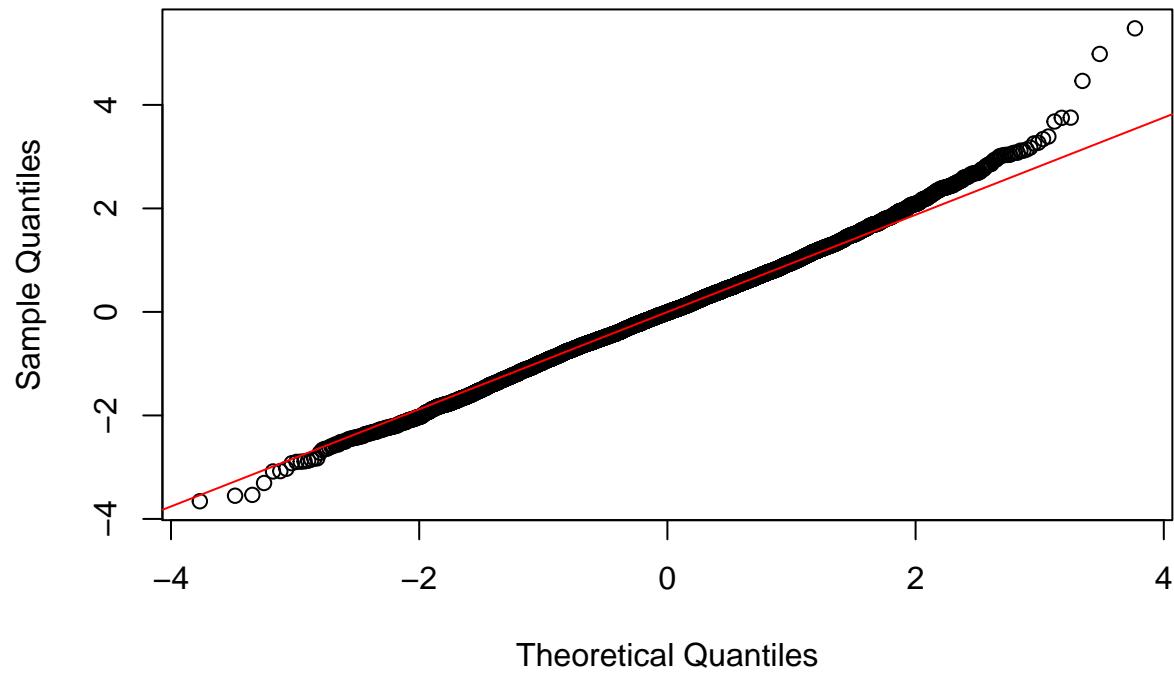
Histogram reziduala

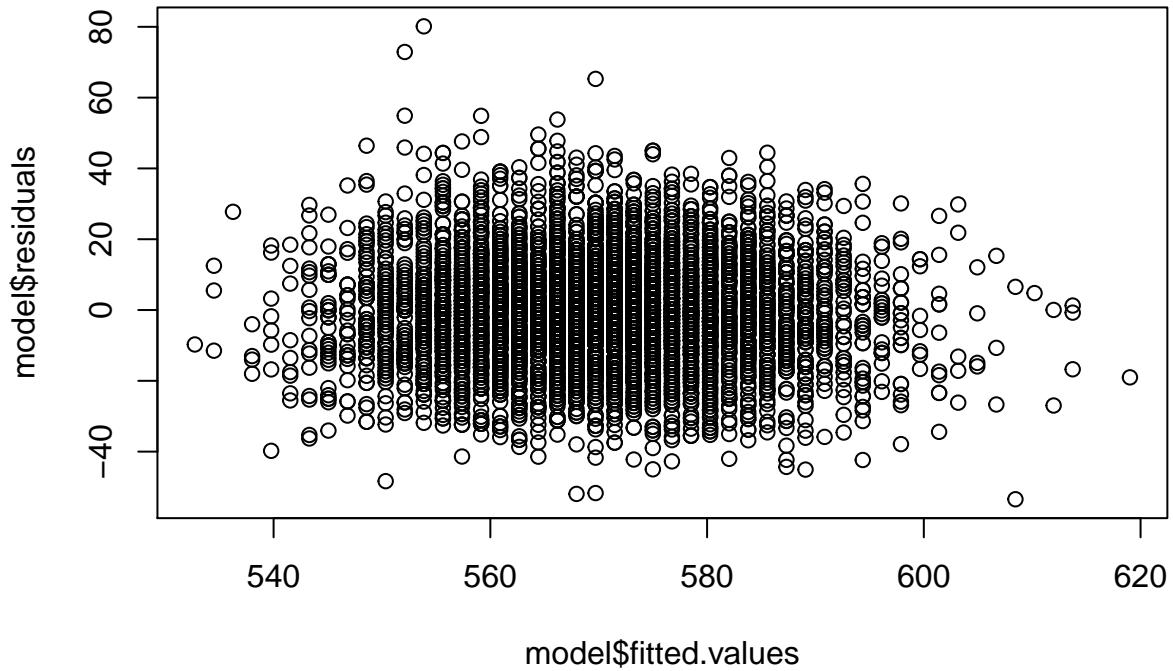


Histogram standardiziranih reziduala



QQ plot





```
##  
## Lilliefors (Kolmogorov-Smirnov) normality test  
##  
## data: rstandard(model)  
## D = 0.016578, p-value = 0.0006649
```

Provjera pretpostavki modela daje podnošljive rezultate. Iako histogrami, pogotovo za reziduale modela u kojem je regresor duljina glave, imaju jako dugačak desni rep, te Lillieforsov test normalnosti daje jako loše rezultate, probat ćemo svejedno statistički interpretirati rezultate, jer t-test, koji se koristi za procjenu parametara modela, dobro podnosi ne-normalnost.

Modeli su zadovoljavajući, duljina glave dosta dobro određuje opseg glave, a širina nešto manje, ali definitivno nezanemarivo.

Također je zanimljivo uočiti da u se u dijagramu reziduala po indeksima jasno vidi razlika između prve trećine podataka i ostatka. Razlog ove pojave je jednostavan - podatci su veoma osjetljivi na spol, a otprilike prva trećina podataka predstavlja mjere ženskih vojnih snaga, dok su ostatak mjere muških vojnih snaga. Prikažimo to grafički:

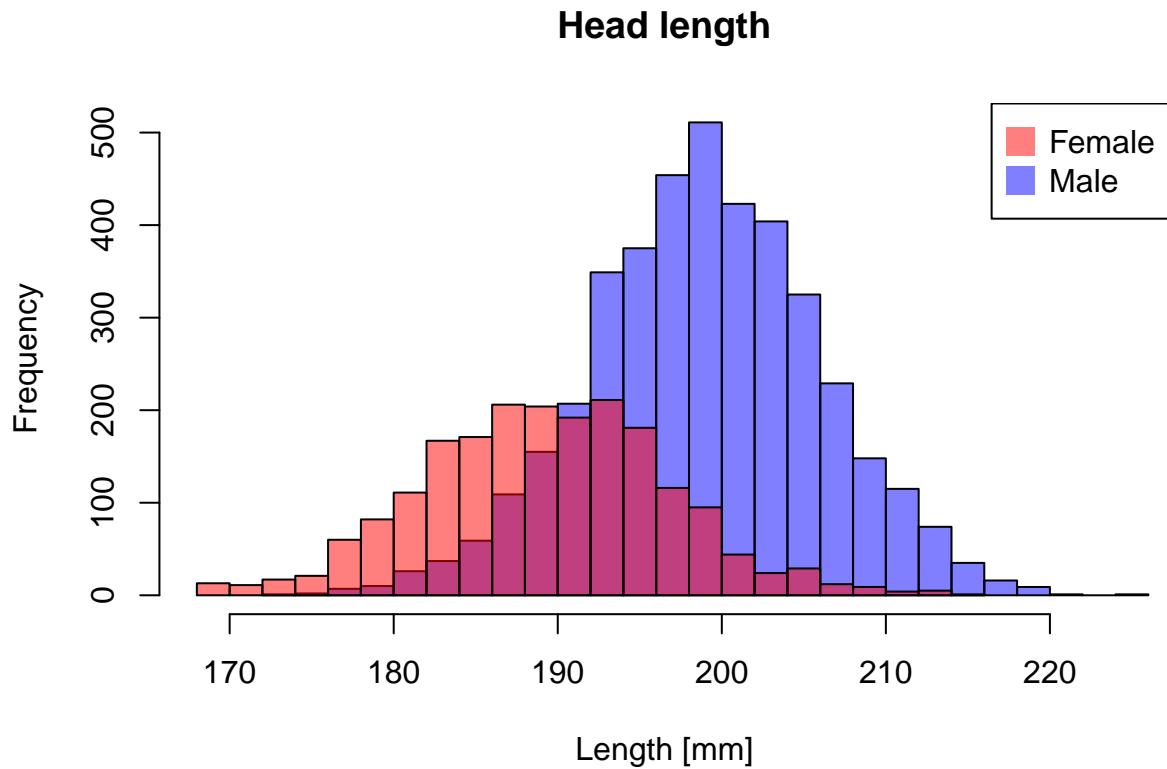
```
males <- ansur.II.data[ansur.II.data$Gender=="Male",]  
females <- ansur.II.data[ansur.II.data$Gender=="Female",]  
  
plot_by_gender <- function(column, main = column, xlab = column) {  
  #summary(females[[column]])  
  #summary(males[[column]])  
  hist(males[[column]], breaks=30, main=main, xlab=xlab, ylab="Frequency", col=rgb(0,0,1,0.5), xlim = c(535, 625))  
  hist(females[[column]], breaks=30, main=main, xlab=xlab, ylab="Frequency", col=rgb(1,0,0,0.5), xlim = c(535, 625))  
  #abline(v=mean(ansur.II.data[[column]]), col="red", lwd=2)
```

```

#abline(v=median(ansur.II.data[[column]]), col="blue", lwd=2)
legend(x="topright", c("Female", "Male"), col=c(rgb(1,0,0,0.5), rgb(0,0,1,0.5)), pt.cex = 2, pch = 15)
}

plot_by_gender("headlength", main="Head length", xlab="Length [mm]")

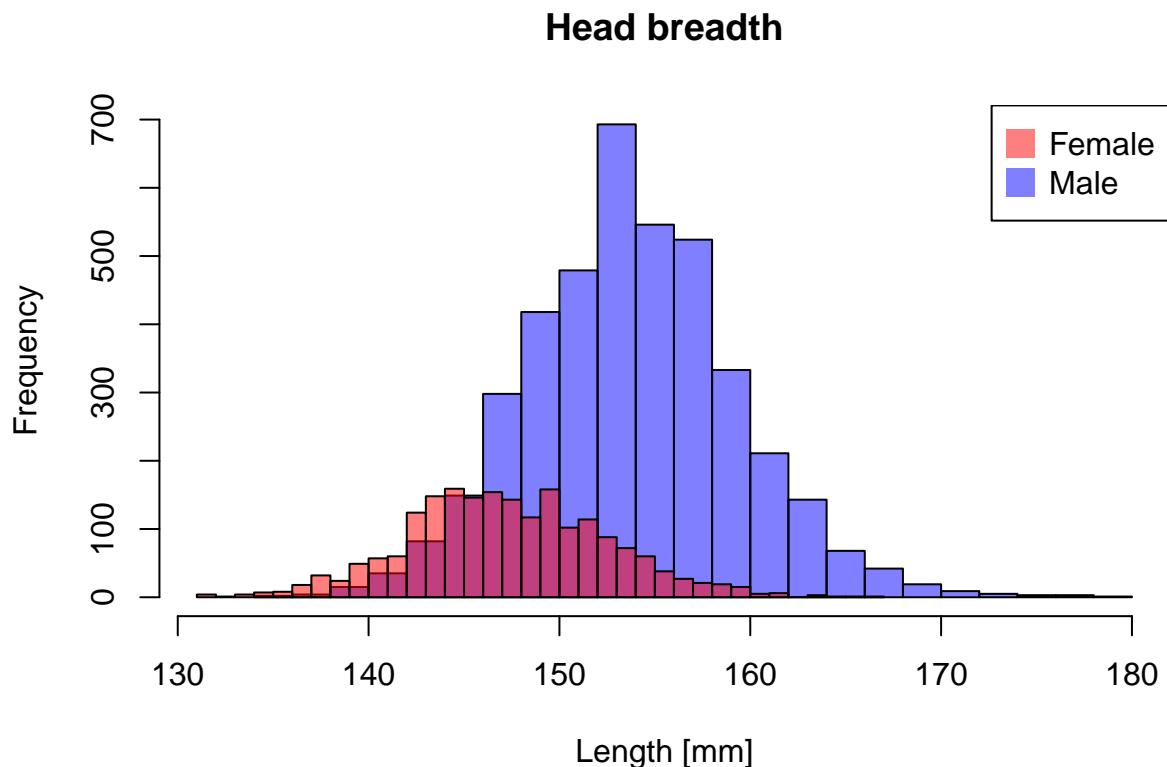
```



```

plot_by_gender("headbreadth", main="Head breadth", xlab="Length [mm]")

```



Kao što se vidi na histogramima, i duljina i širina glave imaju vrlo različita očekivanja s obzirom na spol. Kad bi se to uzealo u obzir prilikom modeliranja, možda bi se dobili nešto precizniji rezultati, no time se za sada nećemo zamarati.

Povjerimo koliko dobro duljina i širina glave zajedno opisuju opseg glave. Prije procjene parametara, provjerimo koliko su duljina i širina glave korelirane. Očekivano je da jesu korelirane, ali svejedno obje utječu na opseg glave i obje će biti bitne za procjenu parametara modela.

```
cor(ansur.II.data$headlength, ansur.II.data$headbreadth)
```

```
## [1] 0.4039947
cor.test(ansur.II.data$headlength, ansur.II.data$headbreadth)

##
##  Pearson's product-moment correlation
##
## data: ansur.II.data$headlength and ansur.II.data$headbreadth
## t = 34.397, df = 6066, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3827234 0.4248379
## sample estimates:
##        cor
## 0.4039947
```

Postoji korelacija, no ona nije prevelika. Provedimo višestruku regresiju.

```

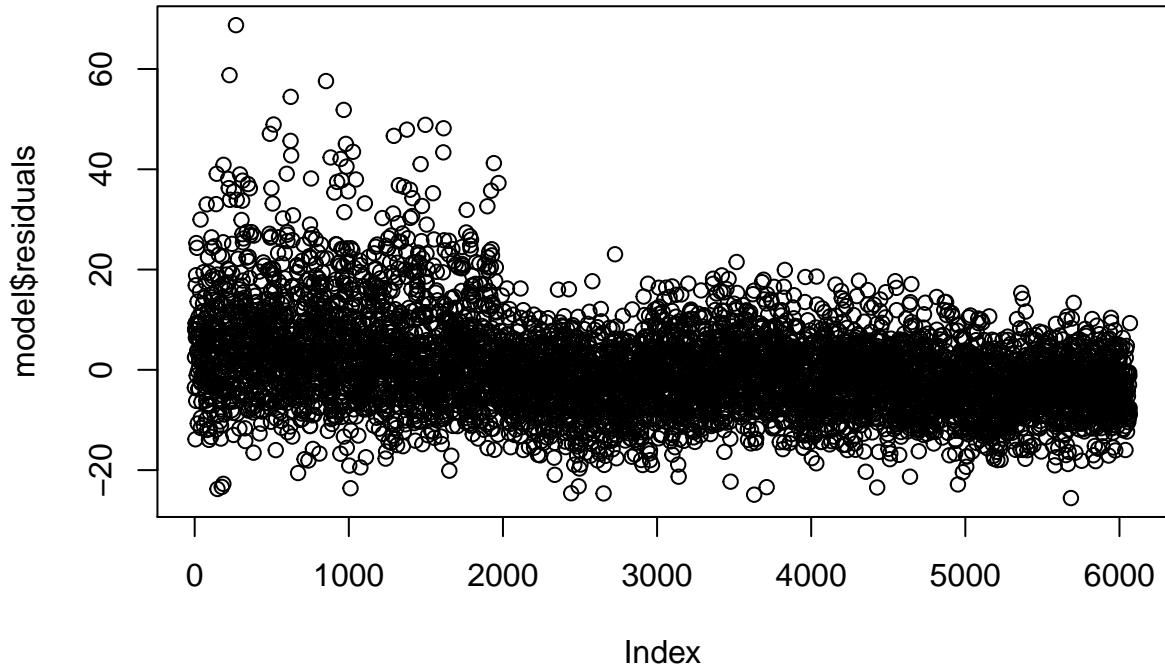
fit.multi = lm(headcircumference~headlength+headbreadth, ansur.II.data)
summary(fit.multi)

##
## Call:
## lm(formula = headcircumference ~ headlength + headbreadth, data = ansur.II.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -25.572  -5.863  -1.095   4.344  68.742 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 136.01806   3.32725  40.88   <2e-16 ***
## headlength    1.47349   0.01509   97.67   <2e-16 ***
## headbreadth   0.95096   0.02053   46.32   <2e-16 ***  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.121 on 6065 degrees of freedom
## Multiple R-squared:  0.7514, Adjusted R-squared:  0.7513 
## F-statistic:  9167 on 2 and 6065 DF,  p-value: < 2.2e-16

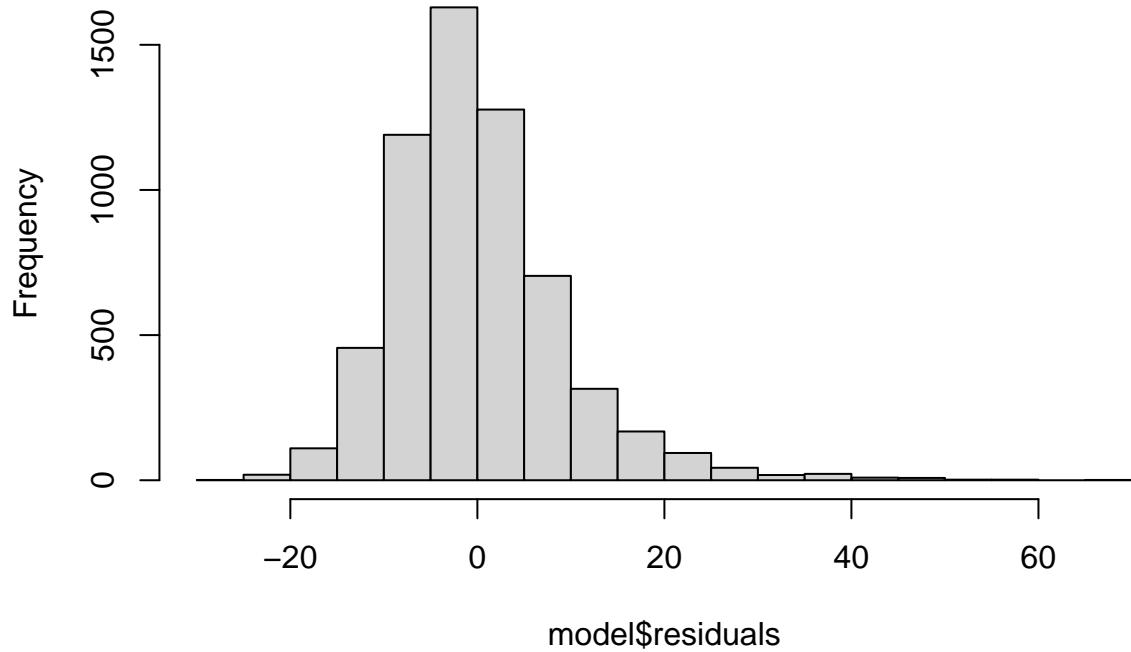
Provjerimo i da pretpostavke modela nisu previše narušene.

check_assumptions(fit.multi)

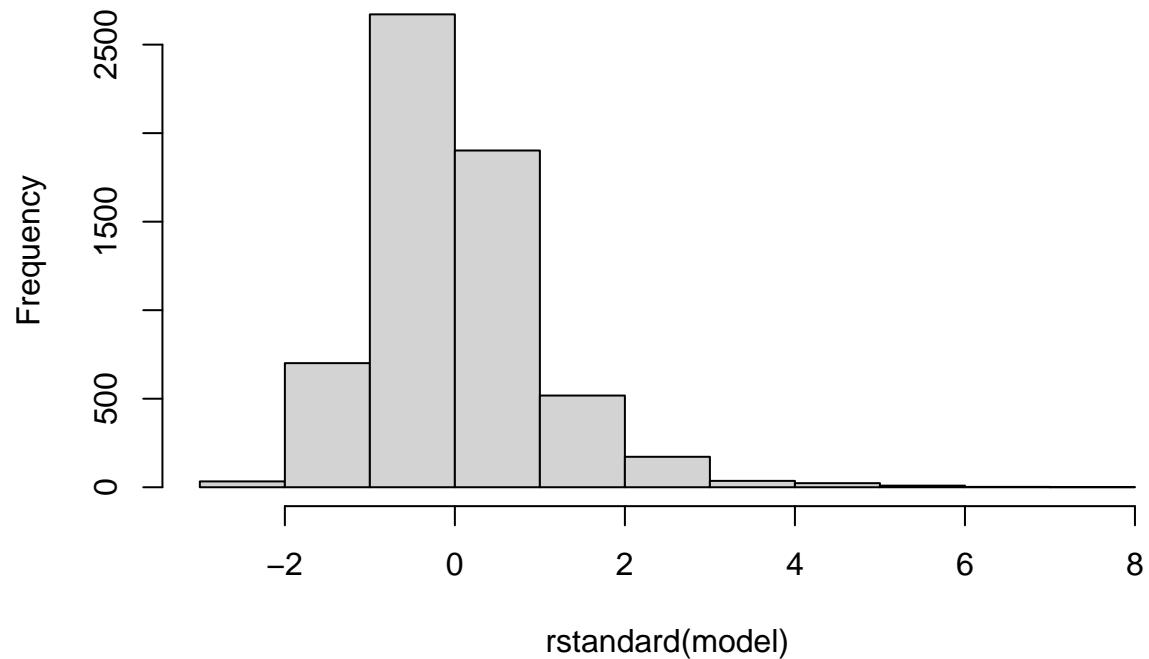
```



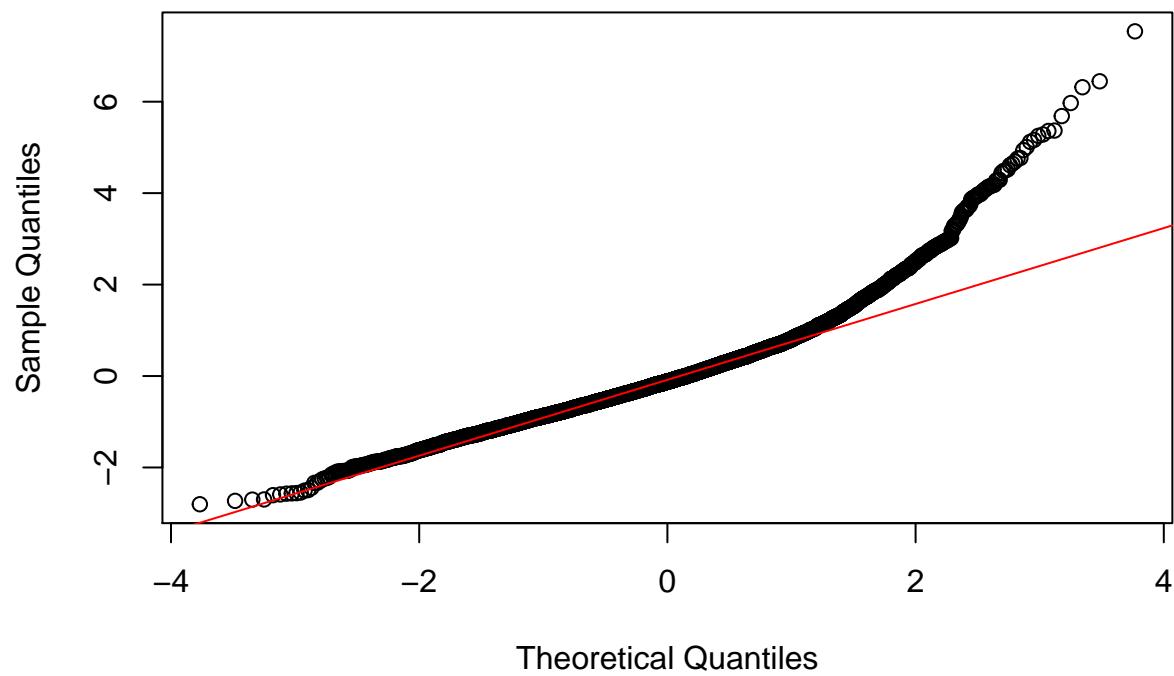
Histogram reziduala

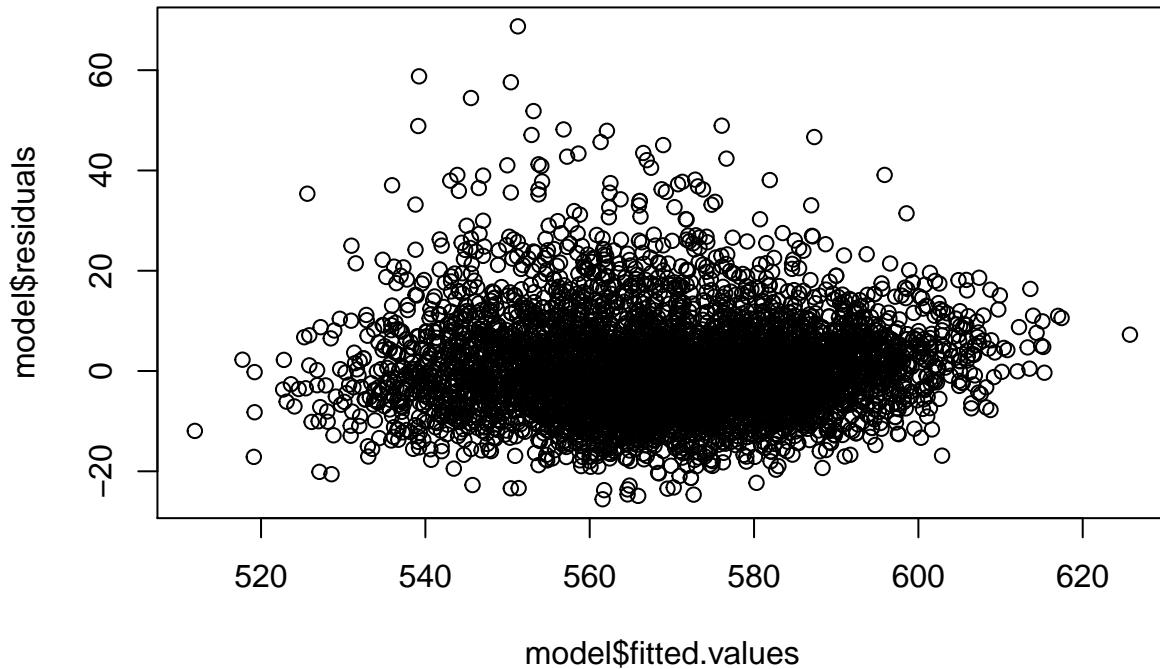


Histogram standardiziranih reziduala



QQ plot





```
##  
## Lilliefors (Kolmogorov-Smirnov) normality test  
##  
## data: rstandard(model)  
## D = 0.070264, p-value < 2.2e-16
```

Normalnost reziduala je vidljivo lošija nego kod linearnih modela, što se pogotovo vidi u desnom repu distribucije reziduala. Također, dijagram reziduala u ovisnosti o procjenjenim vrijednostima modela pokazuje da reziduali postaju "gušći" s većim vrijednostima. Ovo se može interpretirati na način da duljina i širina glave bolje opisuju opseg glave kod manjih opservacija, dok se kod većih opsega glave trebaju uzeti u obzir dodatni parametri koji nisu procjenjeni modelom.

Prije završetka modeliranja kaciga, provjerimo još jednu sitnicu. Naime, u opisu postupaka mjerjenja ovih podataka, navedeno je kako je u mjerjenje duljine glave, a i opsega glave, ulazila kosa vojnih snaga. Za očekivati je da žene u prosjeku imaju frizure većeg obujma, što unosi dodatnu varijancu u mjerjenja koja ne može biti objašnjena "fiksnim" antropometrijskim značajkama kao što su duljina i širina same lubanje, koje su praktički nepromjenjive u odrasloj dobi. Provjerimo kako se ponašaju varijance duljine i širine glave s obzirom na spol.

```
var(females$headlength)  
## [1] 55.48991  
var(males$headlength)  
## [1] 49.34389  
var(females$headbreadth)
```

```

## [1] 26.95563
var(males$headbreadth)

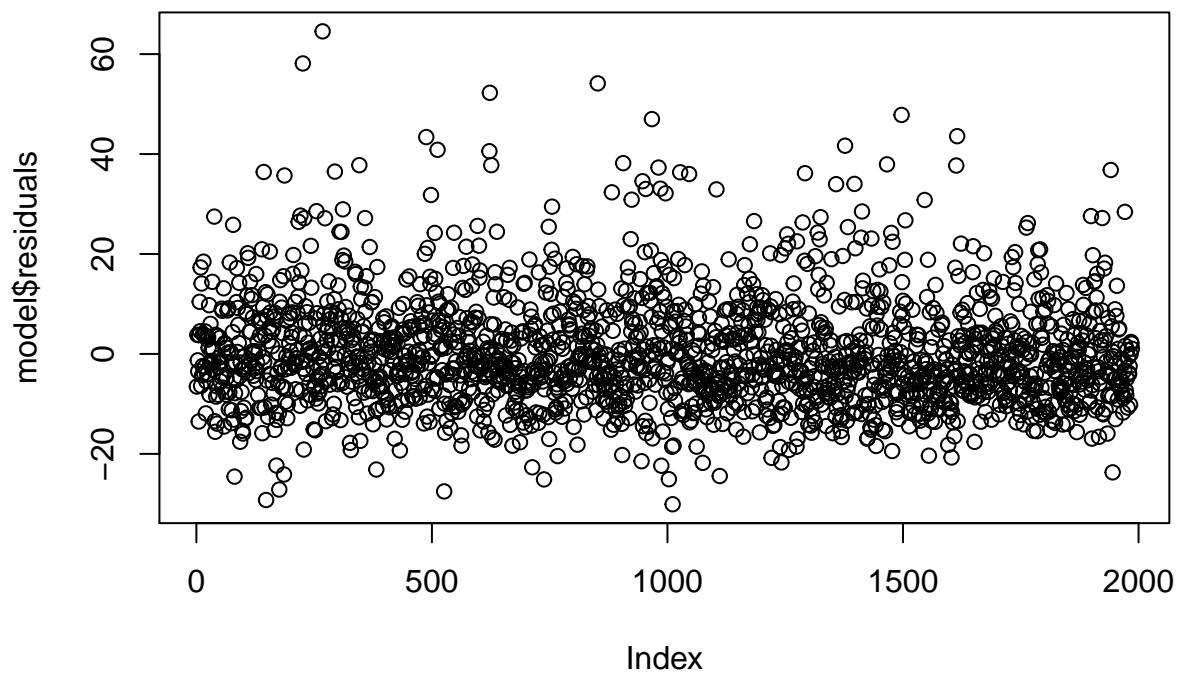
## [1] 30.60153

Vidimo da je varijanca širine glave veća u muškaraca, ali duljina glave je raspršenija kod žena, što bi se moglo objasniti kosom. Ako pretpostavimo da muškarci prosječno imaju manje kose, i da nezanemarivi dio varijance u duljini glave ženskih vojnih snaga dolazi iz kose, model koji procjenjuje opseg glave na temelju duljine i širine trebao bi davati bolje rezultate za muškarce, a lošije za žene. Provjerimo ove hipoteze.

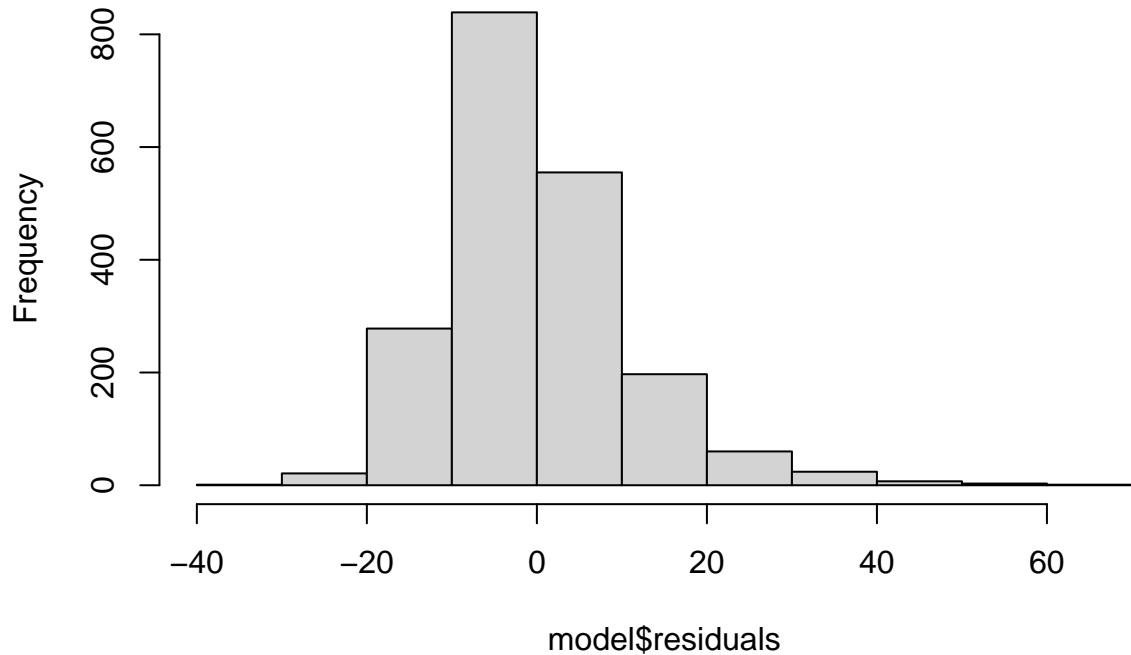
fit.females = lm(headcircumference~headlength+headbreadth, ansur.II.data[ansur.II.data$Gender=="Female"]
summary(fit.females)

##
## Call:
## lm(formula = headcircumference ~ headlength + headbreadth, data = ansur.II.data[ansur.II.data$Gender
##     "Female", ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -30.081  -6.954  -1.722   5.288  64.555
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 35.28554    8.43352   4.184 2.99e-05 ***
## headlength   1.77136    0.03352  52.846 < 2e-16 ***
## headbreadth  1.28306    0.04809  26.679 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.85 on 1983 degrees of freedom
## Multiple R-squared:  0.6862, Adjusted R-squared:  0.6859
## F-statistic:  2168 on 2 and 1983 DF,  p-value: < 2.2e-16
check_assumptions(fit.females)

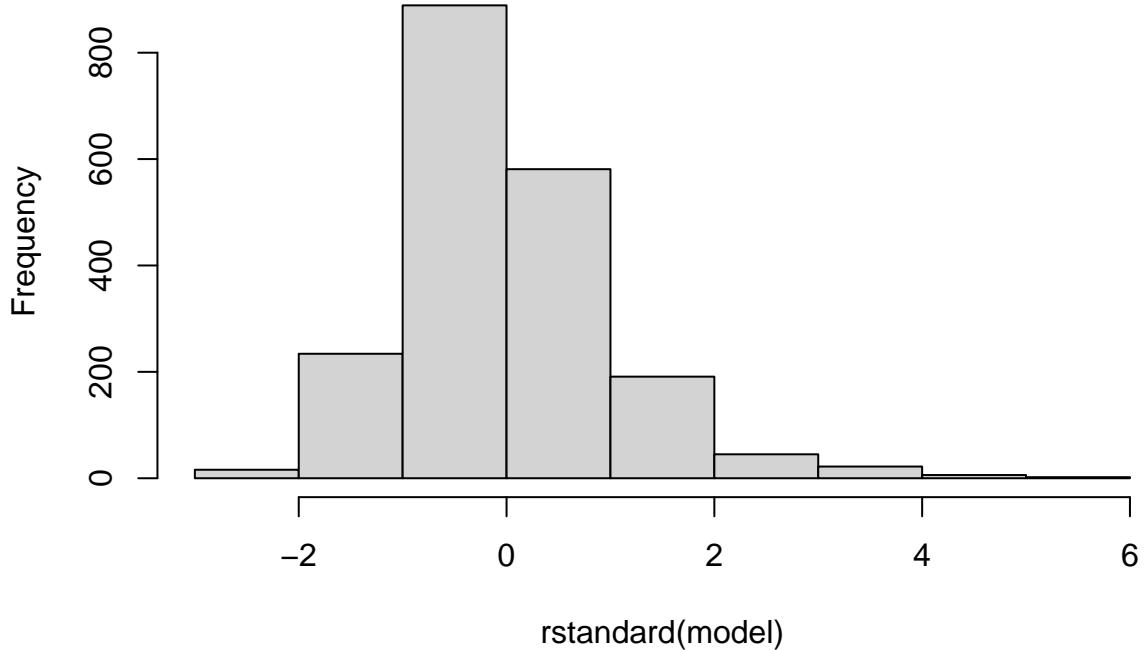
```



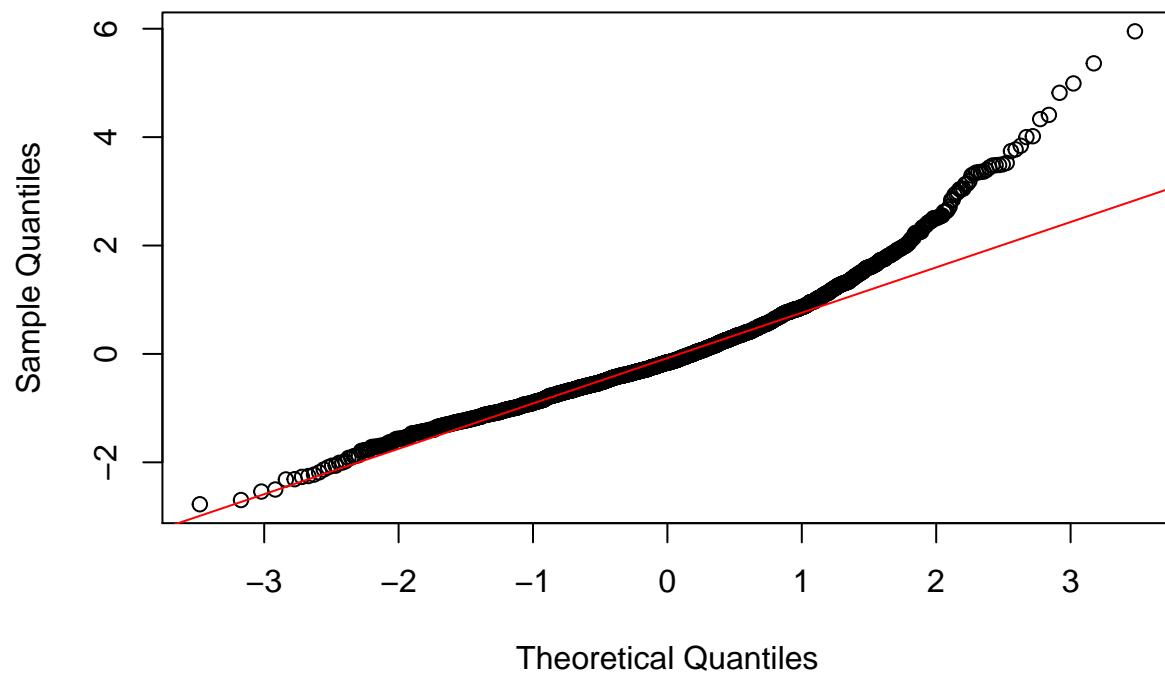
Histogram reziduala

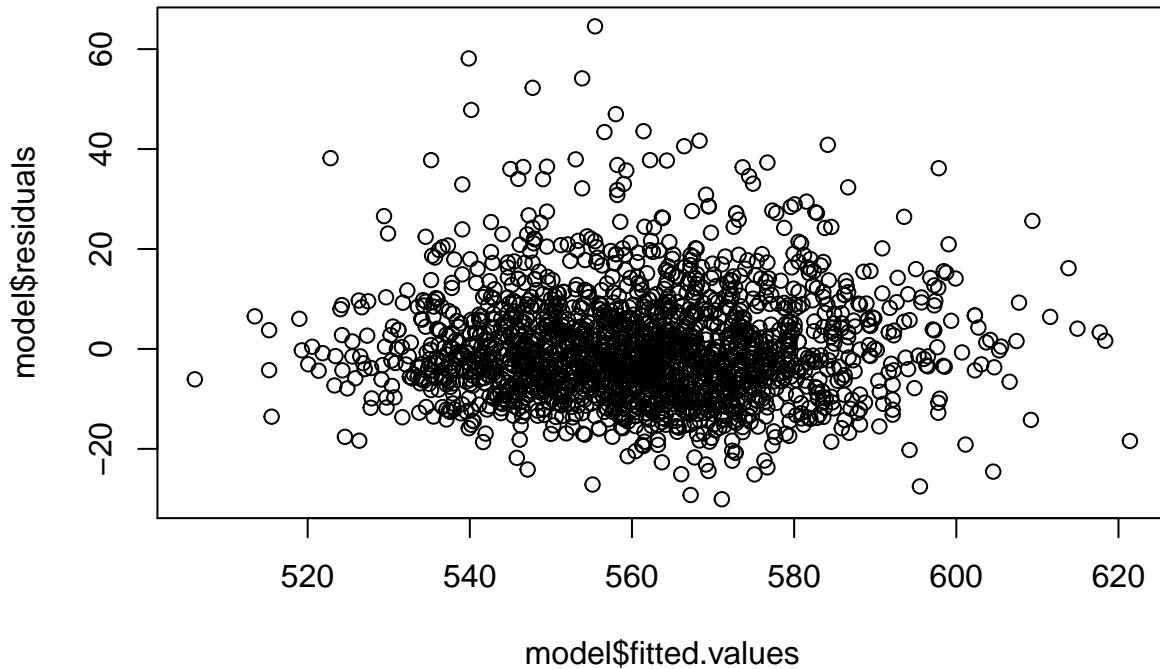


Histogram standardiziranih reziduala



QQ plot





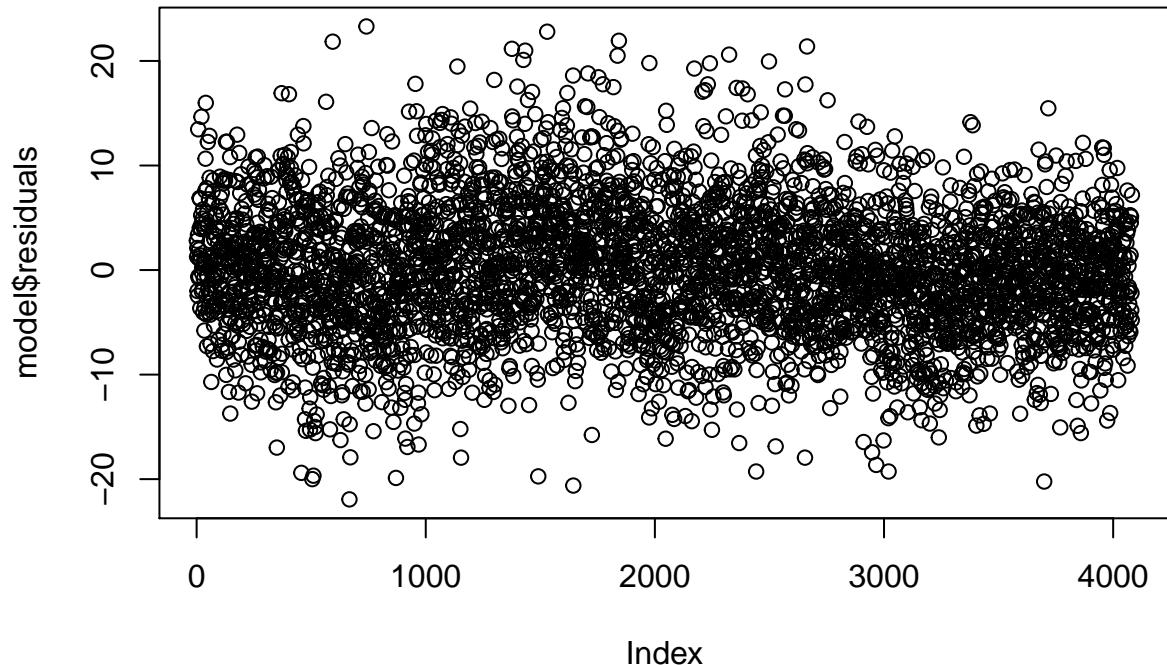
```

## 
## Lilliefors (Kolmogorov-Smirnov) normality test
## 
## data: rstandard(model)
## D = 0.077165, p-value < 2.2e-16
fit.males = lm(headcircumference~headlength+headbreadth, ansur.II.data[ansur.II.data$Gender=="Male",])
summary(fit.males)

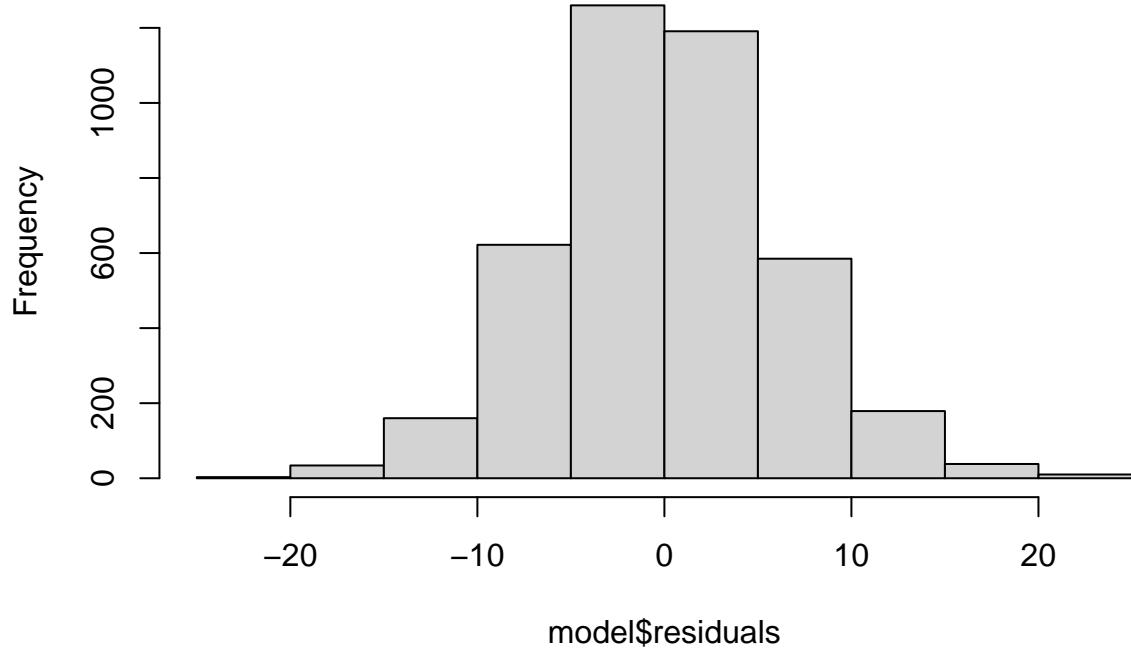
## 
## Call:
## lm(formula = headcircumference ~ headlength + headbreadth, data = ansur.II.data[ansur.II.data$Gender
##     "Male", ])
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -21.9347  -4.0612  -0.1322   3.8964  23.2979 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 43.99733   3.57091  12.32   <2e-16 ***
## headlength   1.71108   0.01406 121.70   <2e-16 ***
## headbreadth  1.22470   0.01785  68.60   <2e-16 ***
## ---      
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 6.211 on 4079 degrees of freedom

```

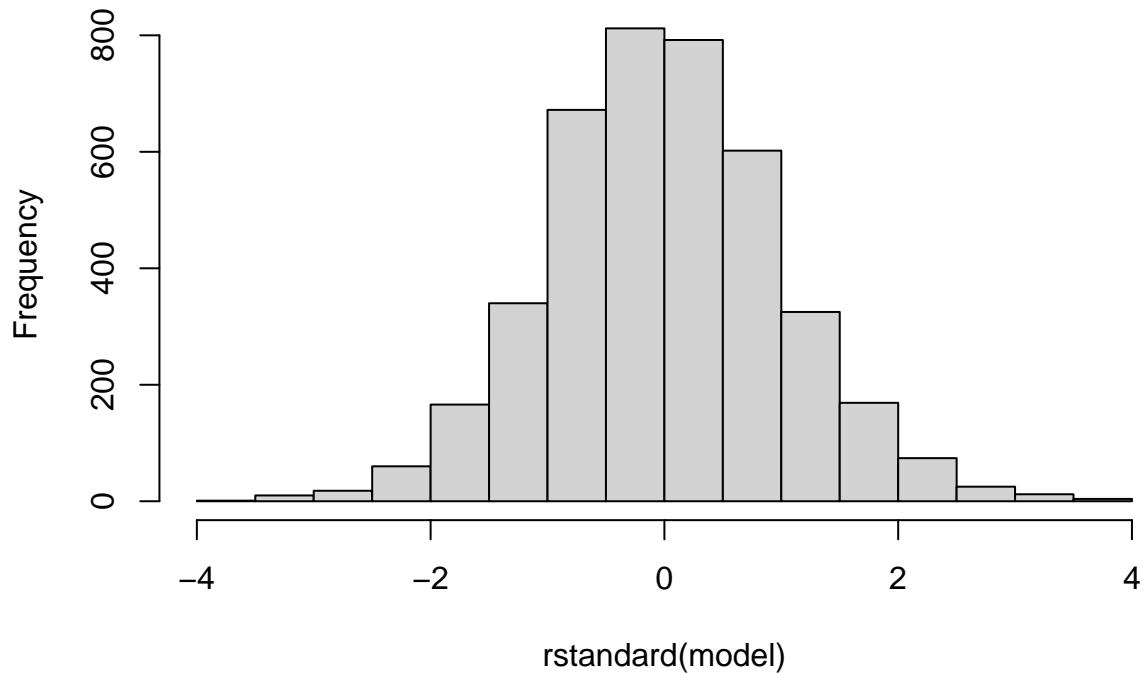
```
## Multiple R-squared:  0.8503, Adjusted R-squared:  0.8502  
## F-statistic: 1.158e+04 on 2 and 4079 DF,  p-value: < 2.2e-16  
check_assumptions(fit.males)
```

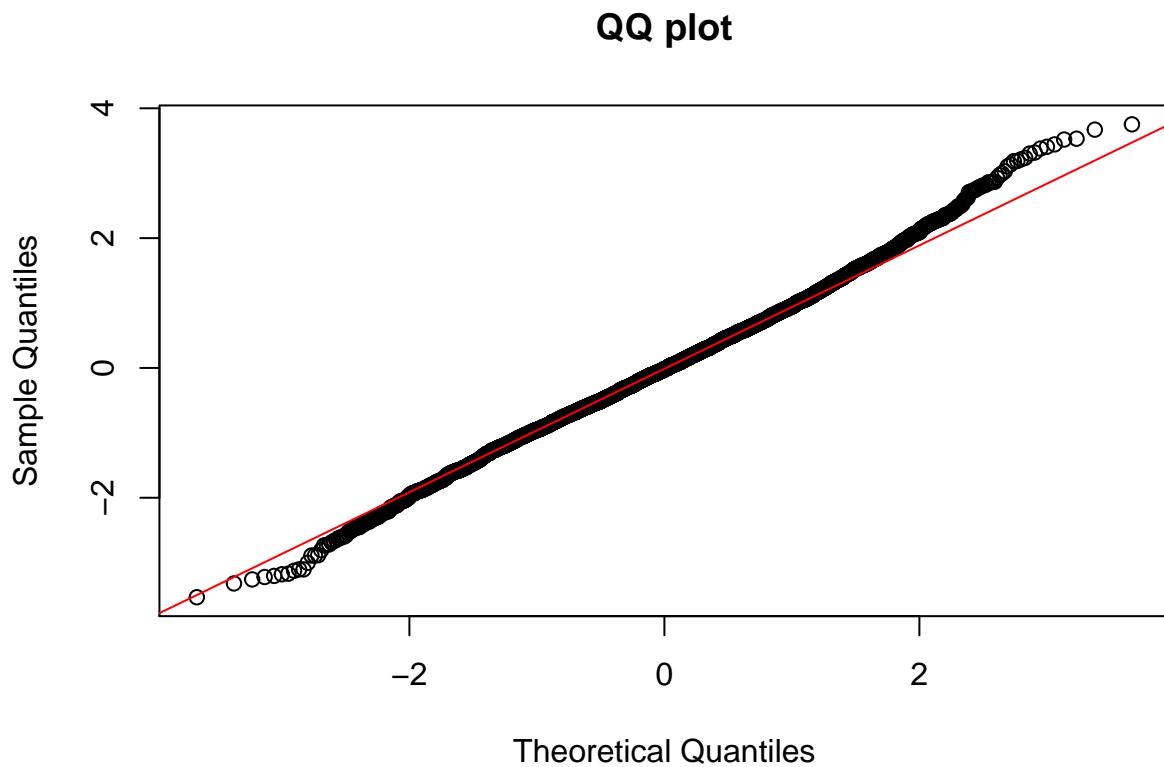


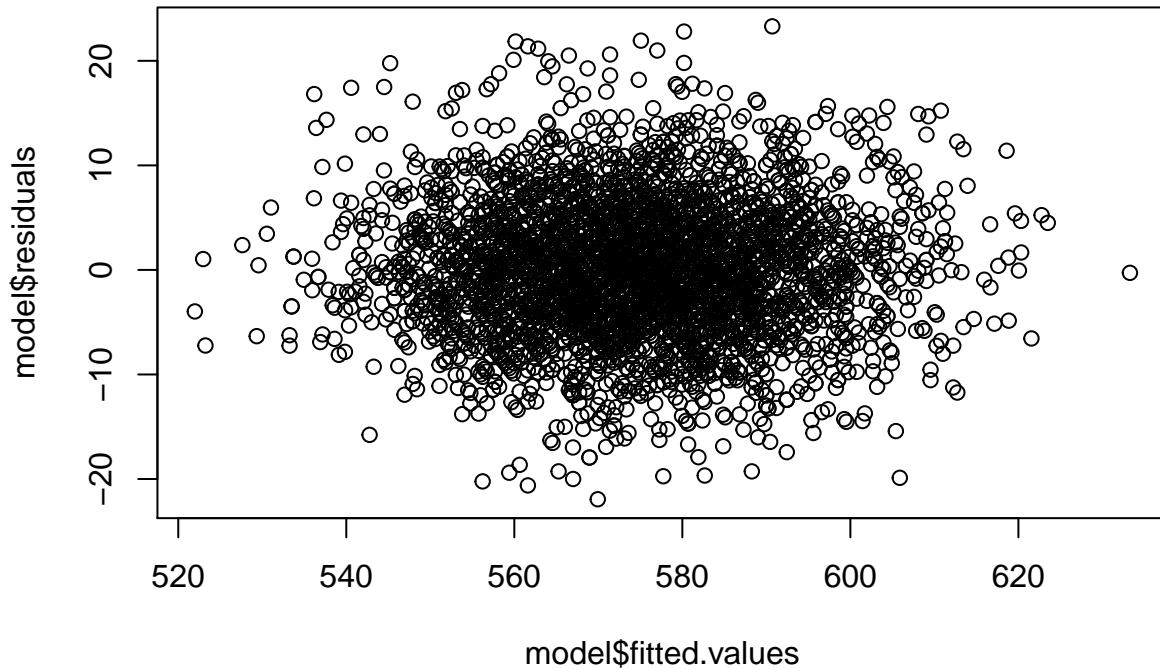
Histogram reziduala



Histogram standardiziranih reziduala







```
##  
## Lilliefors (Kolmogorov-Smirnov) normality test  
##  
## data: rstandard(model)  
## D = 0.016193, p-value = 0.01588
```

Prepostavke su dale jako dobre rezultate. Model se ponaša puno bolje ako se primjeni samo na podatke muškaraca, te objašnjava čak 85 varijance u opsegu glave, što je jako dobar rezultat, pogotovo s obzirom na to da se model bazira samo na dva regresora. S druge strane, za podatke žena, model se ponaša očekivano lošije, što je vjerojatno zbog toga što varijanca zbog kose dolazi više do izražaja.

Prepostavimo da će vojnici noseći kacige morati imati frizuru koja ih neće značajno ometati u nošenju kacige. Drugim riječima, vodeći se višestrukom regresijom nad podatcima muških vojnih snaga, prepostavimo da su duljina, širina i opseg glave međusobno uskladene mjere, te da ako vojniku odgovara neka veličina kacige s obzirom na opseg, analogno će mu odgovarati s obzirom na duljinu i širinu.

Pogledajmo koja od ovih mjerima ima najveći rang.

```
length.range = max(ansur.II.data$headlength) - min(ansur.II.data$headlength)  
breadth.range = max(ansur.II.data$headbreadth) - min(ansur.II.data$headbreadth)  
circumference.range = max(ansur.II.data$headcircumference) - min(ansur.II.data$headcircumference)  
print(length.range)  
  
## [1] 57  
print(breadth.range)  
  
## [1] 49
```

```
print(circumference.range)
```

```
## [1] 135
```

Zaključak

Očekivano, opseg glave ima najveći rang, ali je i najmanje osjetljiv na veličinu kacige(krug kojem se polumjer poveća za 1 cm, opseg se poveća za 2π). Zato ćemo uzeti duljinu glave kao najbitniju mjeru. Ako rang od 57 mm podijelimo na 4 dijela, dobijemo 4 različite veličine kaciga čija se duljina razlikuje za 19 mm. To znači da će, u najgorem slučaju, vojnik imati 9.5 mm “praznog prostora” s prednje i stražnje strane kacige, što je potpuno prihvatljivo. Ostale mjere analogno podijelimo na 4 jednakih dijela i konačno dobivamo sljedeće veličine kaciga:

```
length.inc = length.range / 3
breadth.inc = breadth.range / 3
circumference.inc = circumference.range / 3

lengths = c(min(ansur.II.data$headlength), min(ansur.II.data$headlength) + length.inc, max(ansur.II.data$headlength))
breadths = c(min(ansur.II.data$headbreadth), min(ansur.II.data$headbreadth) + breadth.inc, max(ansur.II.data$headbreadth))
circumferences = c(min(ansur.II.data$headcircumference), min(ansur.II.data$headcircumference) + circumference.inc, max(ansur.II.data$headcircumference))

helmets = data.frame("Veličina"=c("S", "M", "L", "XL"), "Duljina kacige [mm]" = lengths, "Širina kacige [mm]" = breadths, "Opseg kacige [mm]" = circumferences)
print(helmets)

##   Velicina Duljina.kacige..mm. Širina.kacige..mm. Opseg.kacige..mm.
## 1      S            168        131.0000          500
## 2      M            187        147.3333          545
## 3      L            206        163.6667          590
## 4     XL            225        180.0000          635

##Analiza razlike u težini između američkih vojarni
```

Motivacija

Težina može biti zanimljiv indikator raznih proces u ljudskom tijelu. Od prehrane do količine treninga, analiza težina može pokazati zanimljive rezultate između vojarni

Istraživanje

```
ansur = read.csv("ANSUR_II_data.csv")

uniqueCamps = unique(ansur$Installation)

attributes = names(ansur)

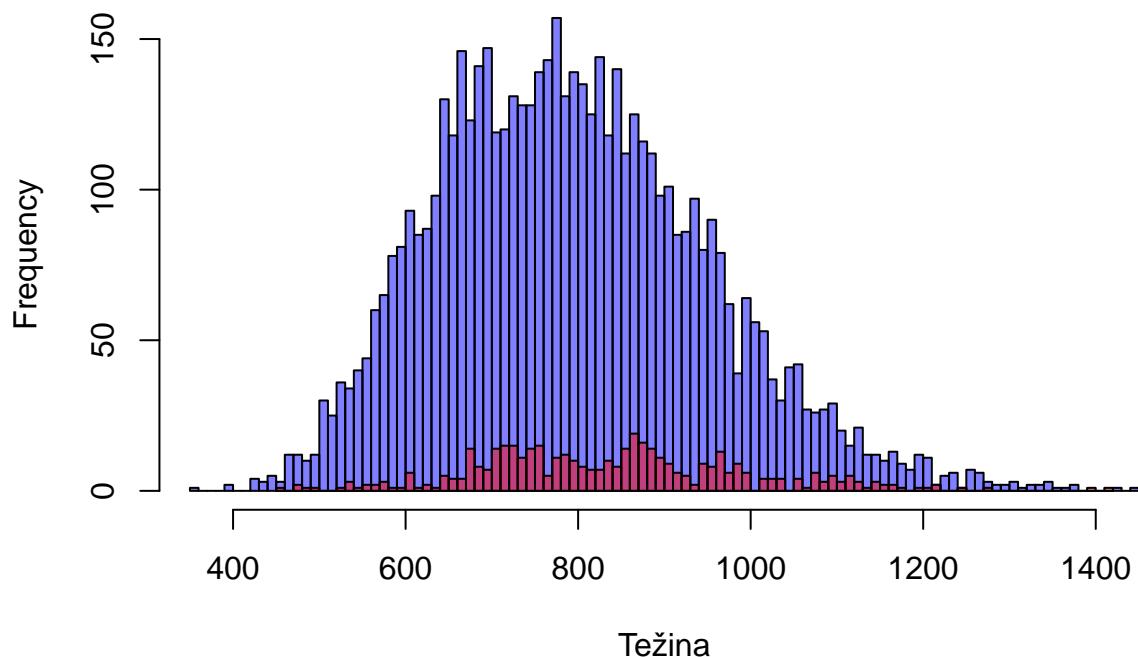
numByCamp = c()

for (camp in uniqueCamps) {
  inCamp = ansur[ which(ansur$Installation==camp),]
  ##ansurByCamps.append(inCamp);

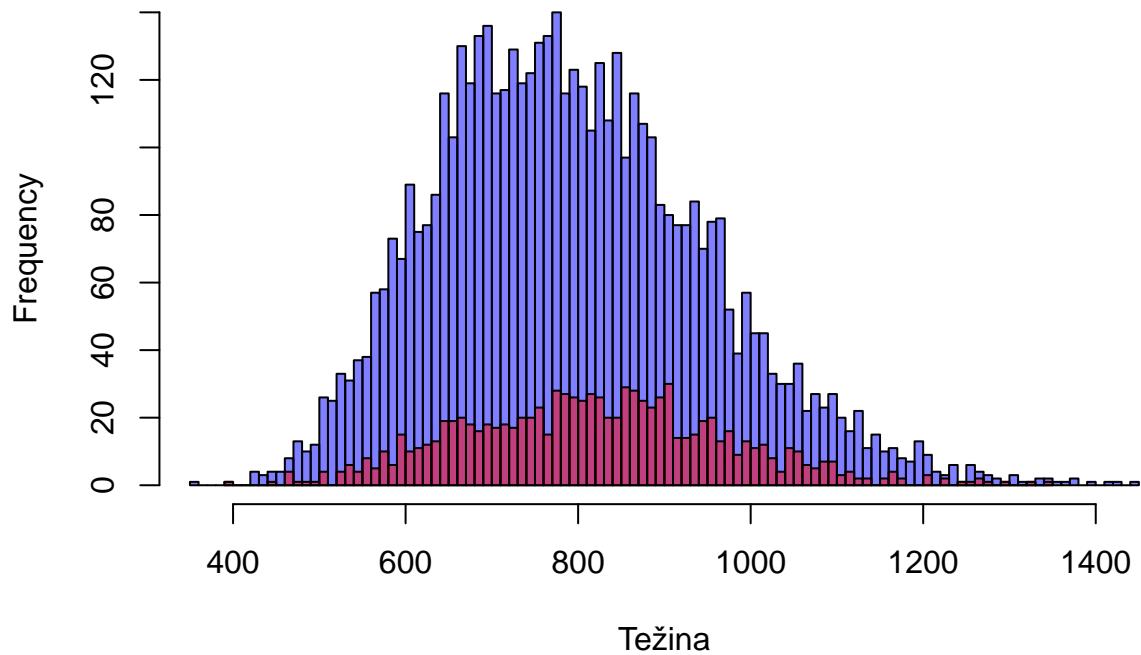
  if (nrow(inCamp) < 50) {
    next
  }
  exceptCamp = ansur[ which(ansur$Installation!=camp),]
```

```
hist(exceptCamp$weightkg, breaks=100, main=paste(camp, " po težini"), xlab="Težina", ylab="Frequency")  
  
hist(inCamp$weightkg, breaks=100, main=paste(camp, " po težini"), xlab="Težina", ylab="Frequency", col="red")  
}
```

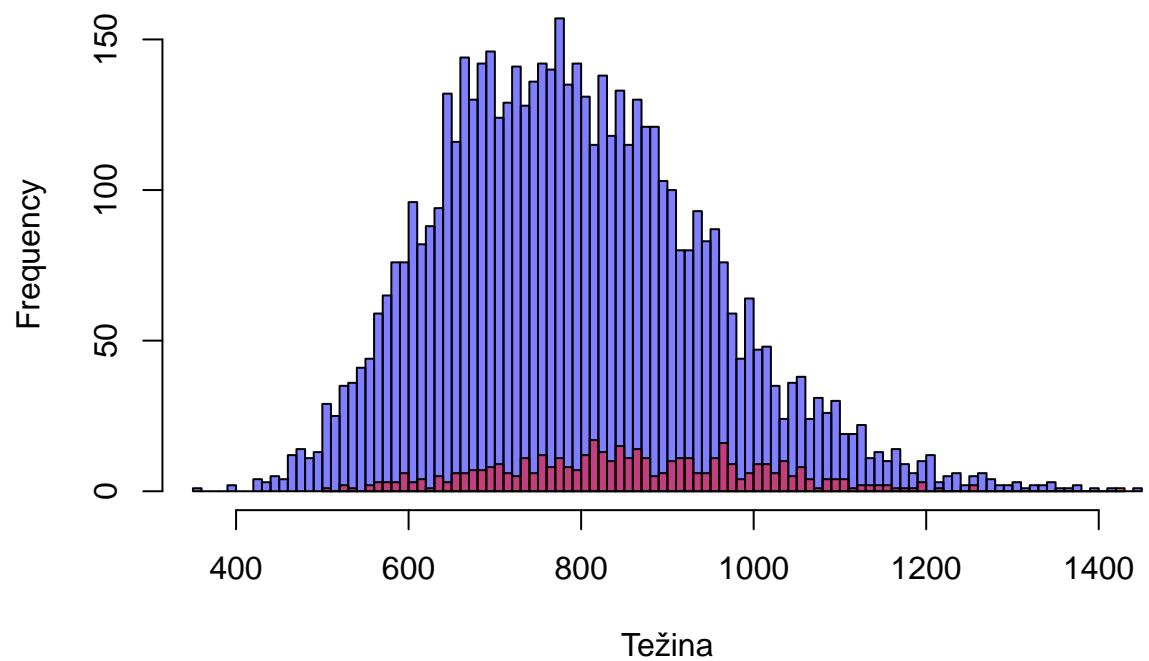
Fort Hood po težini



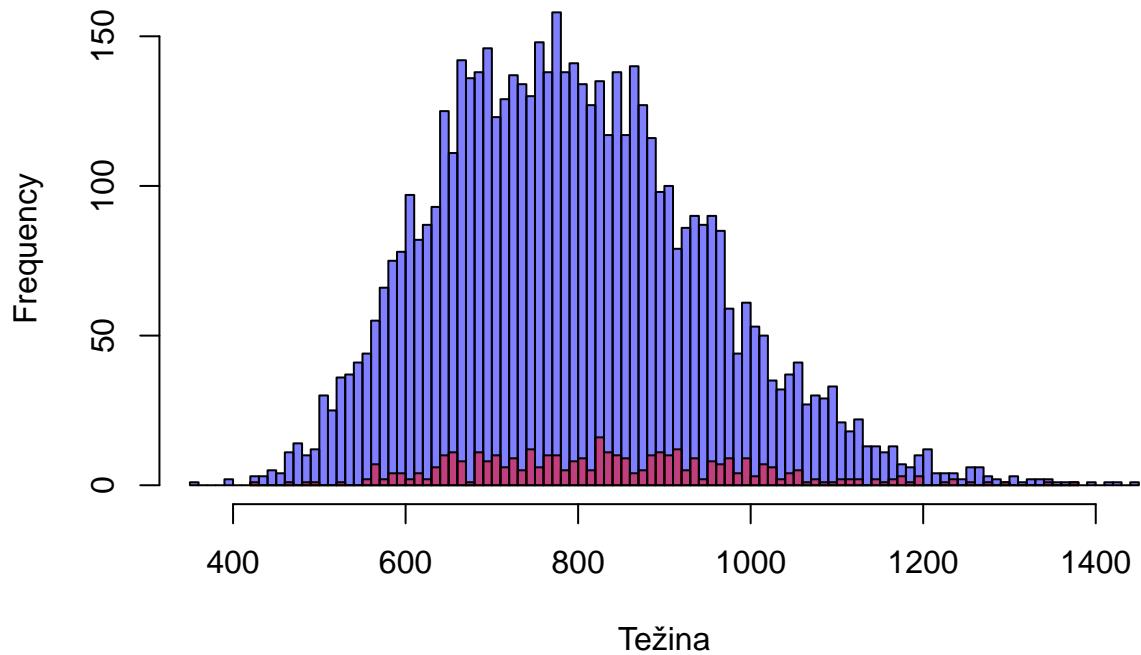
Fort Bliss po težini



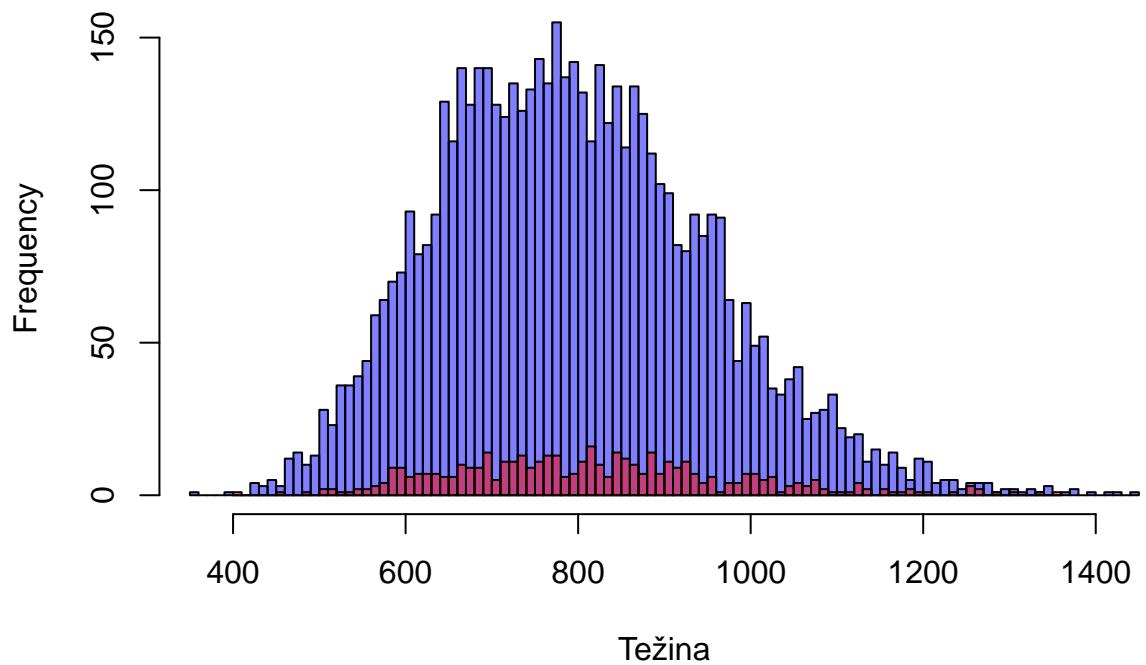
Camp Atterbury po težini



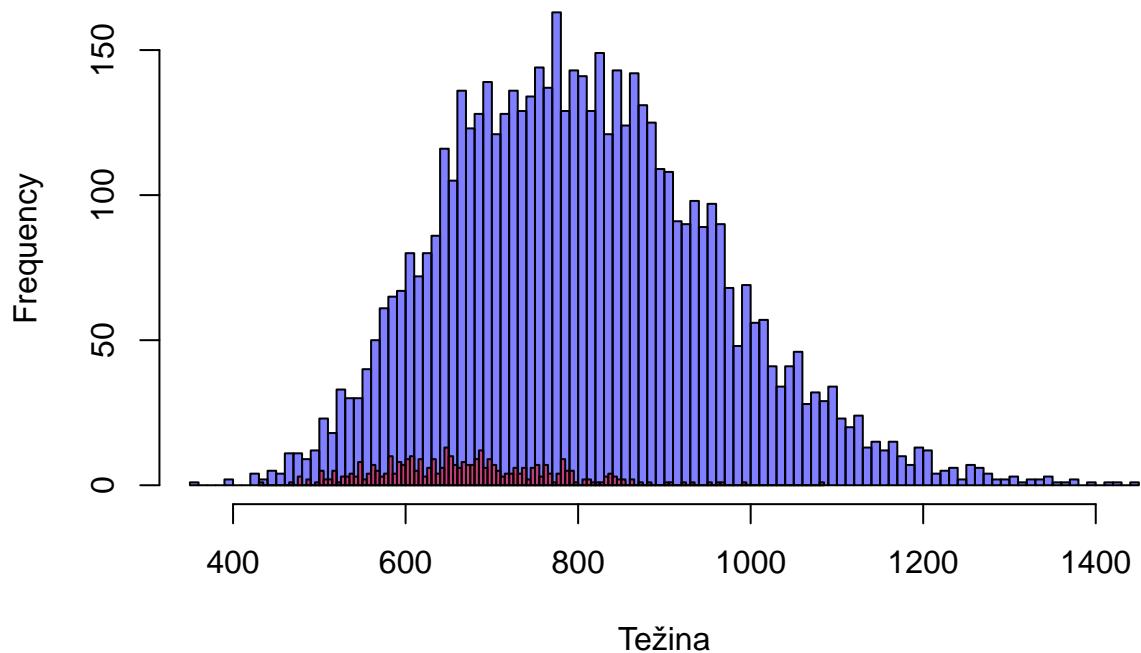
Fort Drum po težini



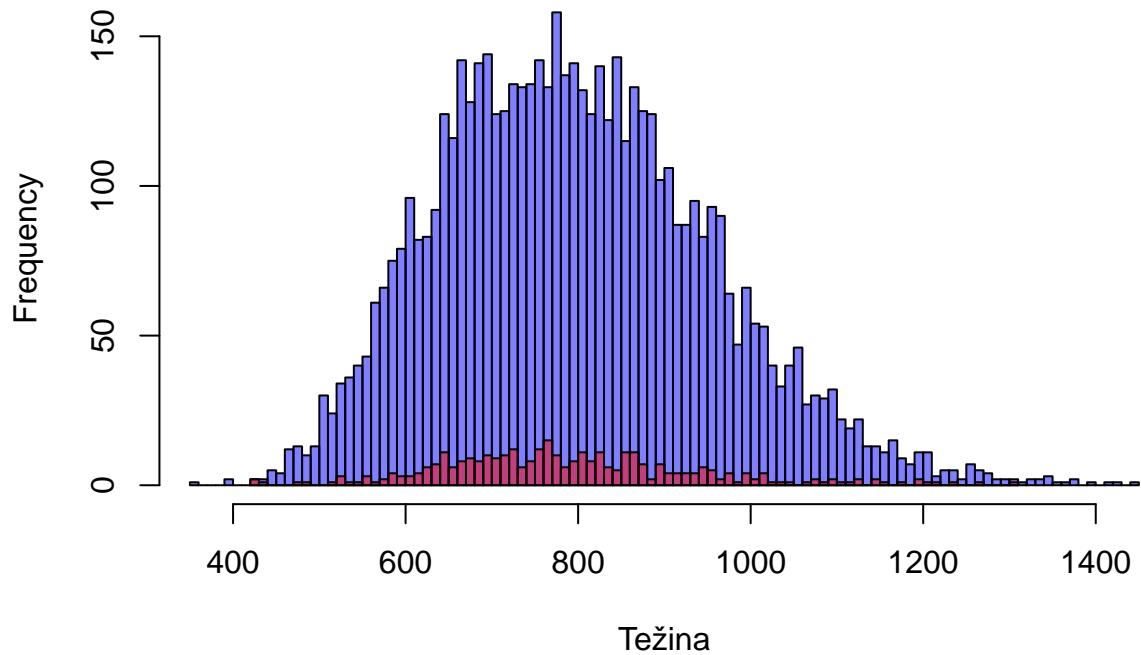
Fort McCoy po težini



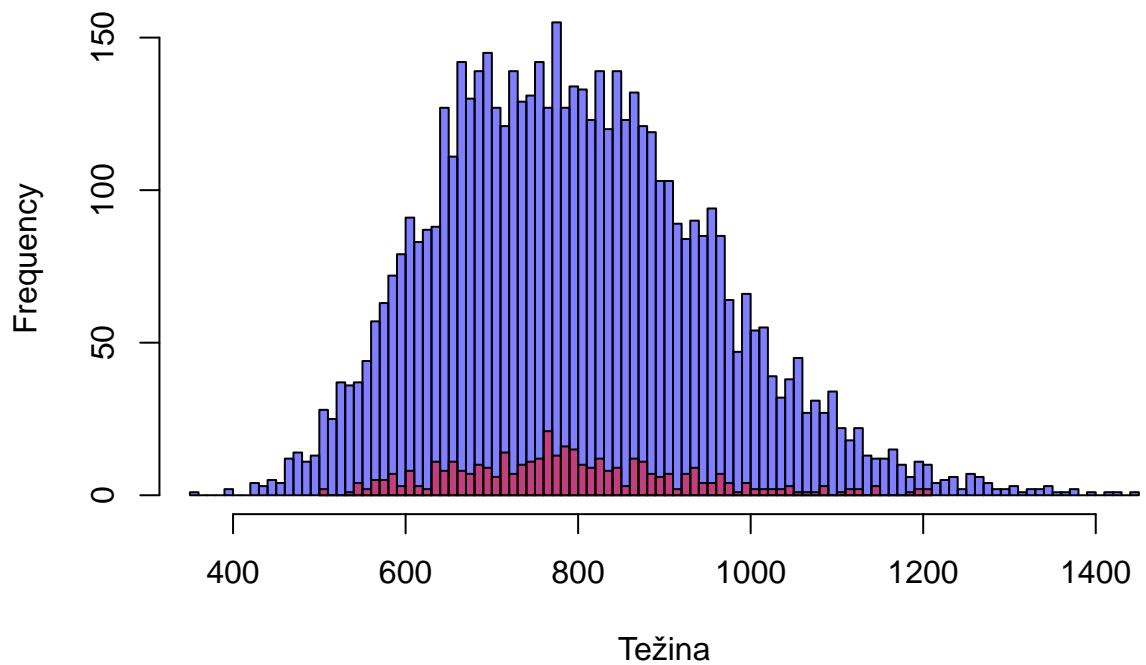
Fort Lee po težini



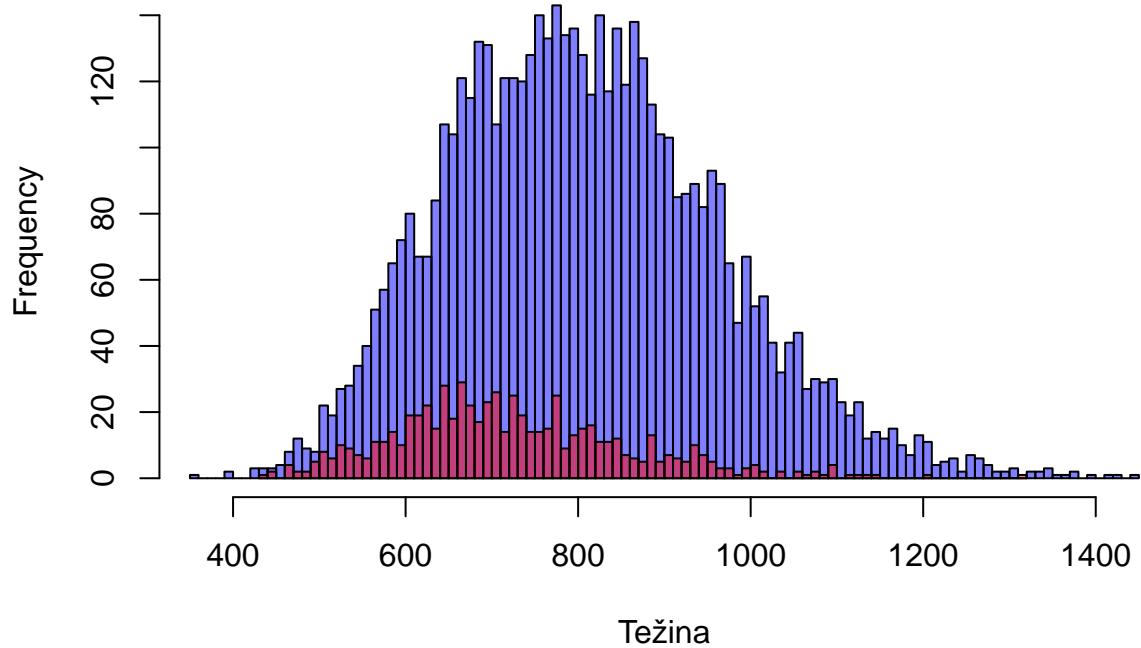
Fort Stewart po težini



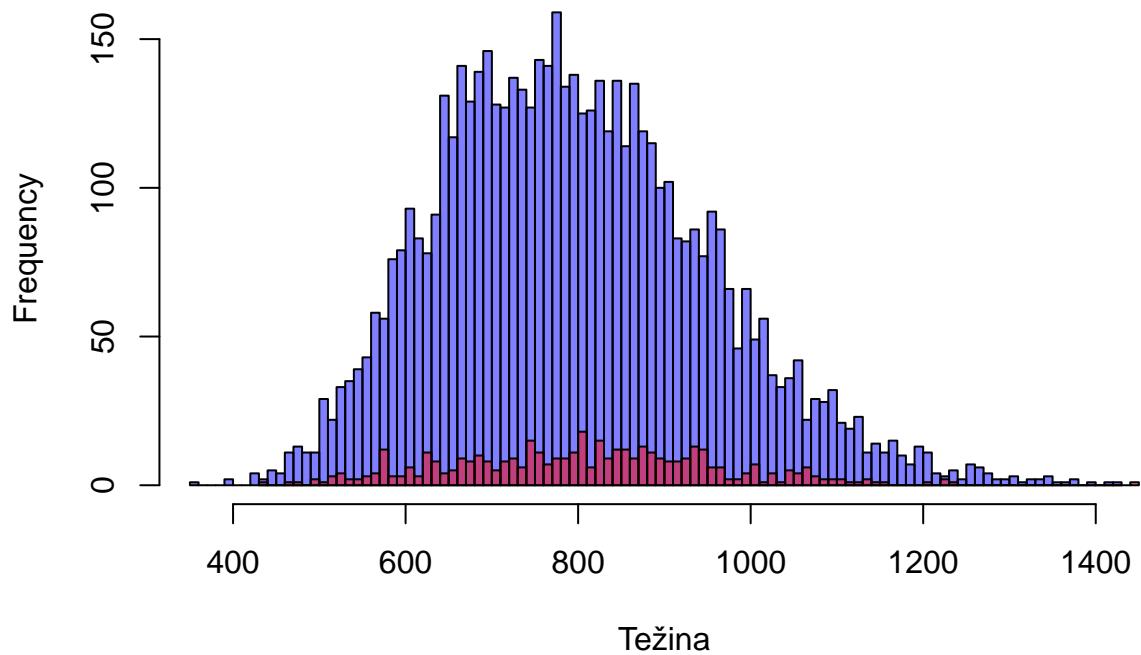
Fort Bragg po težini



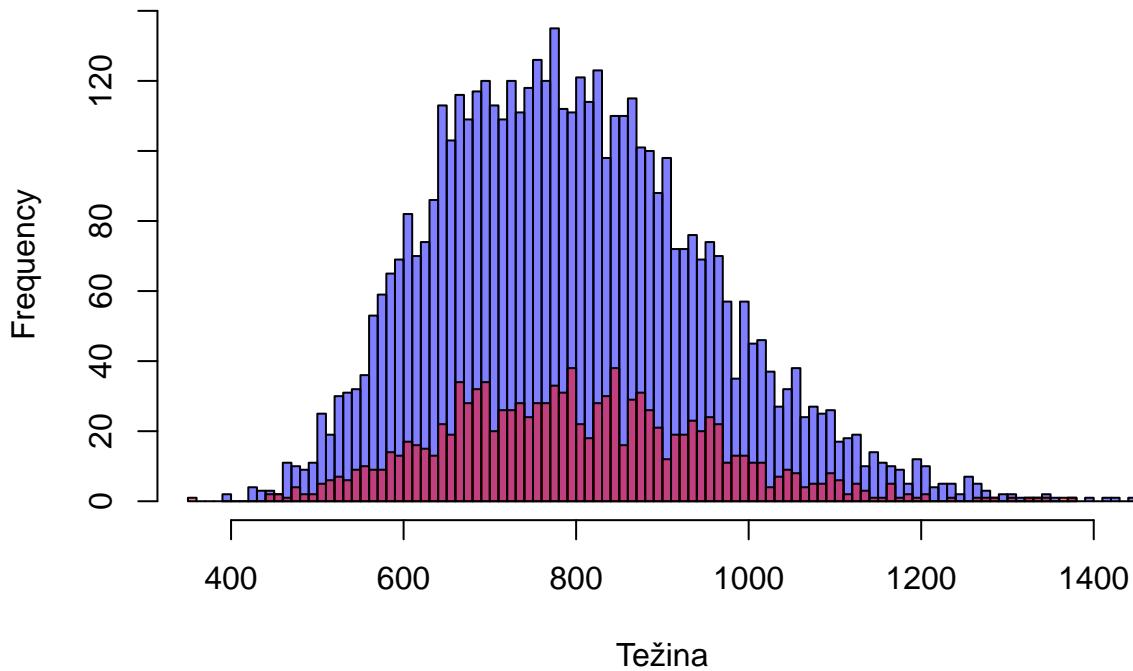
Fort Gordon po težini



Fort Huachuca po težini



Camp Shelby po težini



Iz histograma težine je vidljivo koliko neki kampovi imaju raličiti broj ispitanika. Za daljnju analizu izabratiti ćemo kampove s najvećim brojem pripadnika.

```
table(ansur$Installation)
```

```
##  
## Camp Atterbury      Camp Shelby       Fort Bliss        Fort Bragg      Fort Drum  
##          441           1160            963           397           391  
## Fort Gordon         Fort Hood        Fort Huachuca    Fort Lee       Fort McCoy  
##          669           439             436           380           452  
## Fort Rucker        Fort Stewart  
##          1              339
```

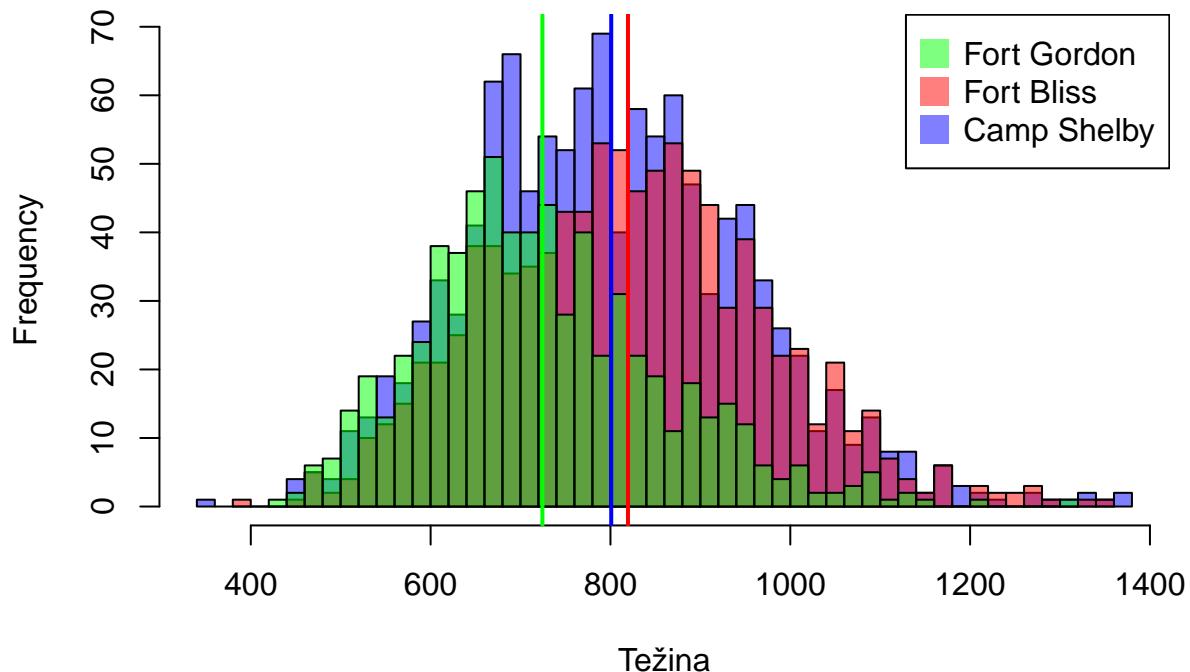
Vidimo kako su najveći kampovi vojnika Camp Shelby, Fort Bliss i Fort Gordon. Za daljnju analizu srednje vrijednosti težine ćemo koristiti upravo ta tri kampa.

```
fortGordon = ansur[ansur$Installation=="Fort Gordon", ]  
fortBliss = ansur[ansur$Installation=="Fort Bliss", ]  
campShelby = ansur[ansur$Installation=="Camp Shelby", ]  
  
hist(campShelby$weightkg, breaks=50, main="Usporedba težina među kampovima", xlab="Težina", ylab="Freque  
hist(fortBliss$weightkg, breaks=50, col=rgb(1,0,0,0.5), add=T)  
  
hist(fortGordon$weightkg, breaks=50, col=rgb(0,1,0,0.5), add=T)  
  
abline(v = mean(campShelby$weightkg), col = rgb(0,0,1), lwd = 2)  
abline(v = mean(fortGordon$weightkg), col = rgb(0,1,0), lwd = 2)
```

```
abline(v = mean(fortBliss$weightkg), col = rgb(1,0,0), lwd = 2)
```

```
legend('topright', legend = c("Fort Gordon", "Fort Bliss", "Camp Shelby"), col = c(rgb(0,1,0,0.5), rgb(1,0,0,0.5),
```

Usporedba težina medu kampovima



Vidimo kako postoji određena razlika između srednjih vrijednosti tri najveća kampa. Ostaje nam za smanjiti uzorke na veličinu najmanjeg uzorka i provesti analizu.

```
set.seed(42)

row_shelby <- sample(nrow(campShelby))
row_gordon <- sample(nrow(fortGordon))
row_bliss <- sample(nrow(fortBliss))

campShelby <- campShelby[row_shelby, ]
fortGordon <- fortGordon[row_gordon, ]
fortBliss <- fortBliss[row_bliss, ]

campShelby <- campShelby[1:nrow(fortGordon),]
fortBliss <- fortBliss[1:nrow(fortGordon),]

nrow(campShelby)

## [1] 669
nrow(fortBliss)

## [1] 669
```

```

nrow(fortGordon)

## [1] 669

Nakon nasumičnog balansiranja skupova podataka provodimo analizu varijance kako bi provjerili pretpostavke t-testa za analizu jednakosti dviju srednjih vrijednosti

var.test(fortGordon$weightkg, fortBliss$weightkg)

## 
## F test to compare two variances
##
## data: fortGordon$weightkg and fortBliss$weightkg
## F = 0.77302, num df = 668, denom df = 668, p-value = 0.0008962
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 0.6641469 0.8997307
## sample estimates:
## ratio of variances
## 0.7730158

var.test(fortGordon$weightkg, campShelby$weightkg)

## 
## F test to compare two variances
##
## data: fortGordon$weightkg and campShelby$weightkg
## F = 0.79018, num df = 668, denom df = 668, p-value = 0.002376
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 0.6788907 0.9197044
## sample estimates:
## ratio of variances
## 0.7901764

var.test(campShelby$weightkg, fortBliss$weightkg)

## 
## F test to compare two variances
##
## data: campShelby$weightkg and fortBliss$weightkg
## F = 0.97828, num df = 668, denom df = 668, p-value = 0.7767
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 0.8405046 1.1386454
## sample estimates:
## ratio of variances
## 0.9782825

```

Iz varijance možemo zaključiti kako za Fort Gordon ne možemo koristiti t-test u kojem uzimamo da su nam varijance jednake. S druge strane F-test između Fort Blissa i Camp Shelbya ima p vrijednost 0.7767 te sa velikom sigurnošću možemo tvrditi da su varijance jednake.

```

t.test(fortGordon$weightkg, fortBliss$weightkg, alternative = "two.sided", var.equal = FALSE)

## 
## Welch Two Sample t-test
##

```

```

## data: fortGordon$weightkg and fortBliss$weightkg
## t = -12.082, df = 1314.5, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -111.25808 -80.17391
## sample estimates:
## mean of x mean of y
## 724.3318 820.0478

t.test(fortGordon$weightkg, campShelby$weightkg, alternative = "two.sided", var.equal = FALSE)

##
## Welch Two Sample t-test
##
## data: fortGordon$weightkg and campShelby$weightkg
## t = -8.8519, df = 1317.9, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -85.14462 -54.25149
## sample estimates:
## mean of x mean of y
## 724.3318 794.0299

t.test(campShelby$weightkg, fortBliss$weightkg, alternative = "two.sided", var.equal = TRUE)

##
## Two Sample t-test
##
## data: campShelby$weightkg and fortBliss$weightkg
## t = -3.109, df = 1336, p-value = 0.001917
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -42.434815 -9.601059
## sample estimates:
## mean of x mean of y
## 794.0299 820.0478

```

T-testom u sva tri slučaja odbacujemo H_0 = Kampovi imaju jednaku srednju vrijednost težine.

Zaključak

Odbacivanje nulte hipoteze nam govori kako postoji ne zanemariva razlika između najvećih kampova američke vojske u težini. Takav zaključak otvara područje za dodatna istraživanja uzroka postojanja te razlike ali takvo istraživanje je van domene ovog projekta.

Zaključak

Nakon provedenih svih analiza stvarno možemo reći da smo došli do nekih neočekivanih otkrića. Za kraj važno je napomenuti kako svi ovdje predstavljeni modeli i analize nipošto nisu dobar predstavnik općenite ljudske populacije. Pogotovo kada se uzme u obzir fizička specifičnost koju nosi profesija vojnika.