Projekt - Milijarderi

Samuel Lucić, Luka Glavinić, Ivan Kolar, Dominik Jambrović

2023-01-15

Učitavanje i provjera podataka

Kako bismo započeli s deskriptivnom analizom, kao i provođenjem testova za provjeru određenih hipoteza, potrebno je prvo učitati podatke, provjeriti njihovu ispravnost i po potrebi izbaciti neispravne ili nepotpune podatke.

```
## Broj milijardera:  2614
```

```
## Broj atributa:  22
```

Možemo vidjeti da učitana tablica sadrži podatke o 2614 milijardera - njihova imena, količinu bogatstva, državljanstvo, način zarade bogatstva itd. Koristeći funkciju sapply, možemo vidjeti klase određenih stupaca tablice. Koristeći funkciju summary, za podatak o dobi milijardera možemo uočiti da je minimum negativan - negativna vrijednost dobi označava manjak tog podatka za konkretnog milijardera.

Provjerit ćemo koliko je podataka u tablici označeno s NA, praznim nizom, nulom ili negativnom brojkom (interpretacija takvih oznaka je da su nedostajali potrebni podaci):

```
## Ukupno nedostajućih vrijednosti za varijablu company.founded :  40
## Ukupno nedostajućih vrijednosti za varijablu company.name :  38
## Ukupno nedostajućih vrijednosti za varijablu company.relationship :  46
## Ukupno nedostajućih vrijednosti za varijablu company.sector :  23
## Ukupno nedostajućih vrijednosti za varijablu company.type :  36
## Ukupno nedostajućih vrijednosti za varijablu demographics.age :  385
## Ukupno nedostajućih vrijednosti za varijablu demographics.gender :  34
## Ukupno nedostajućih vrijednosti za varijablu location.gdp :  1665
## Ukupno nedostajućih vrijednosti za varijablu location.region :  1
## Ukupno nedostajućih vrijednosti za varijablu wealth.type :  22
## Ukupno nedostajućih vrijednosti za varijablu wealth.how.category :  86
## Ukupno nedostajućih vrijednosti za varijablu wealth.how.industry :  17
```

Ovisno o potrebama konkretnog zadatka, podatke označene s praznim nizom, nulom ili negativnom brojkom zamijenit ćemo s NA ili u potpunosti izbaciti iz podataka. Da odmah izbacujemo sve retke koji nemaju određeni podatak, uvelike bi si smanjili ukupan broj podataka.

Nakon provjere i pročišćavanja podataka, možemo započeti s deskriptivnom analizom, kao i provjerom proizvoljnih hipoteza. Pri provođenju testiranja, odlučili smo da će razina značajnosti uvijek biti 0.05 - ako je P-vrijednost manja od 0.05, odbijat ćemo nultu hipotezu.

1. zadatak, stara verzija

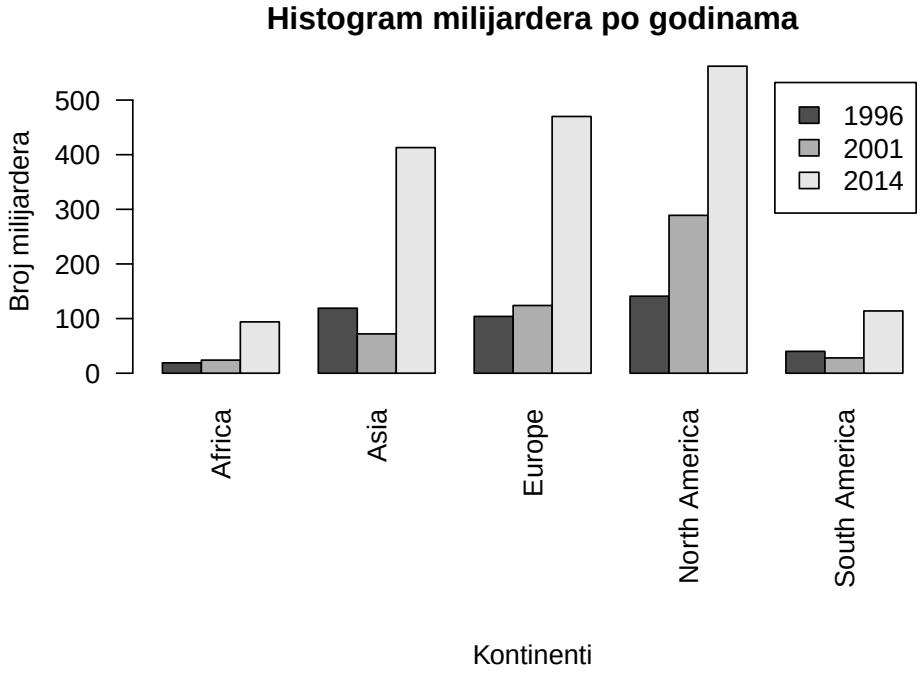
Kako bismo provjerili postoje li kontinenti sa statistički značajno više milijardera, podatke ćemo prikazati koristeći barplot, a pritom ćemo koristiti podatke o regijama milijardera. Prvi korak pri ovakvoj analizi provjera je postoje li podaci o milijarderima koji nemaju podatak o pripadnoj regiji - nedostatak 1 podatka o regiji već smo utvrdili u prethodnom koraku pa tu provjeru nećemo ponavljati. Milijardere s nedostajućim podatkom o regiji izbacit ćemo iz tablice koju koristimo za odgovaranje na ovo pitanje.

```
billionaires.continents <- billionaires[billionaires$location.region != 0, ]
```

Nakon što smo iz podataka izbacili takve milijardere, određene ćemo regije mijenjanjem vrijednosti grupirati u nove skupine: regije "Middle East/North Africa" i "Sub-Saharan Africa" bit će zamijenjene vrijednošću "Africa", regije "East Asia" i "South Asia" bit će zamijenjene vrijednošću "Asia", a regija "Latin America" bit će zamijenjena vrijednošću "South America". Time će za svakog milijardera biti dostupan podatak o kontinentu s kojeg potječe. Ipak, treba uzeti u obzir i činjenicu da bi za određene regije (Latin America i Middle East/North Africa) bilo moguće odabrati i drugačije grupiranje.

```
for (i in 1:length(billionaires.continents$location.region)) {
  if (billionaires.continents$location.region[i] == "Middle East/North Africa" | billionaires.continents
    billionaires.continents$location.region[i] <- "Africa"
  } else if (billionaires.continents$location.region[i] == "East Asia" | billionaires.continents$locati
    billionaires.continents$location.region[i] <- "Asia"
  } else if (billionaires.continents$location.region[i] == "Latin America") {
    billionaires.continents$location.region[i] <- "South America"
  }
}
```

Nakon eliminacije milijardera bez podataka o regiji i zamjene podataka o regiji s pripadnim kontinentima, tablicu ćemo podijeliti na 3 tablice ovisno o godini kojoj pripada određeni rank. Ovime osiguravamo da podaci iz zasebnih tablica vjerno prikazuju stvarno stanje pripadnih godina, kao i da naši prikazi i zaključci nisu pod utjecajem višestrukog pojavljivanja istih ljudi. Kako bismo što bolje mogli usporediti raspodjelu milijardera po kontinentima za zasebne godine, prikazat ćemo podatke kroz sve 3 godine koristeći 1 barplot.



Uočavamo velik porast broja milijardera u razdoblju 2001 - 2014 za sve kontinente, kao i podjelu u dvije grupe - Europa, Azija i Sjeverna Amerika imaju značajno veći broj milijardera naspram Afrike i Južne Amerike. Ova podjela veoma je vidljiva za podatke iz 2014. godine, ali je isti odnos grupa prisutan i za preostale dvije godine. Da imamo podatke o ukupnom broju stanovnika po regiji, mogli bismo provesti test nezavisnosti da provjerimo jesu li proporcije milijardera po kontinentu jednake za sve regije.

Nažalost, nemamo dostupan taj podatak. Iako zbog nejednakosti broja stanovnika po kontinentu test goodness-of-fit korišten za usporedbu s uniformnom distribucijom nije najsmisleniji izbor, provest ćemo ga za svaku od godina.

```
## Broj milijardera po kontinentu 1996. godine:  19 119 104 141 40
```

```
##
##  Chi-squared test for given probabilities
##
## data:  v1
## X-squared = 130.42, df = 4, p-value < 2.2e-16
```

```
## Broj milijardera po kontinentu 2001. godine:  24 72 124 289 28
```

```
##
##  Chi-squared test for given probabilities
##
## data:  v2
## X-squared = 444.76, df = 4, p-value < 2.2e-16
```

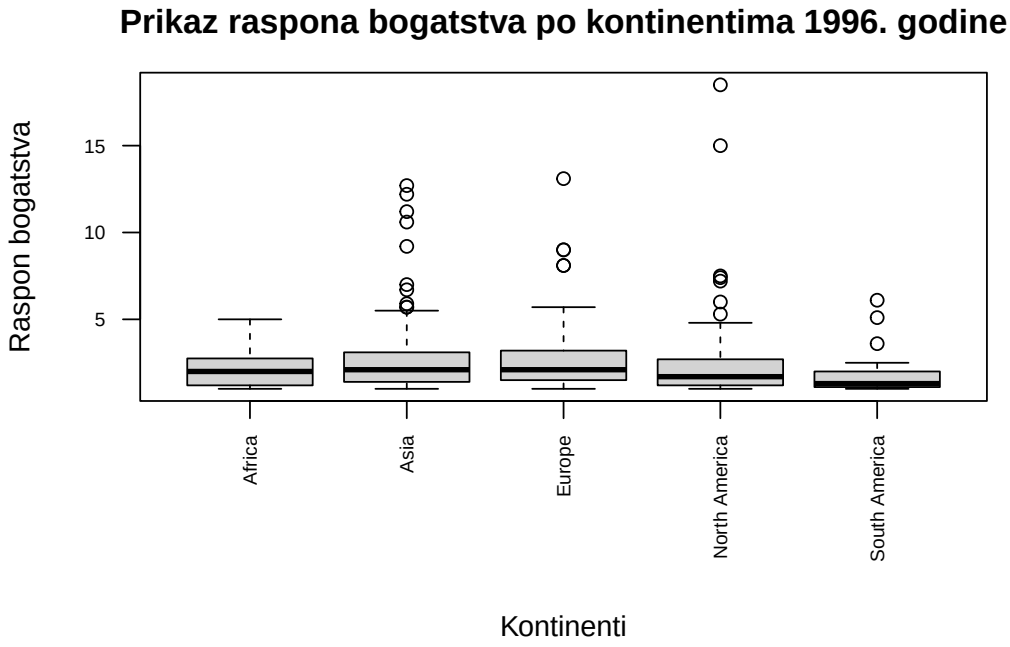
```
## Broj milijardera po kontinentu 2014. godine:  94 413 470 562 114
```

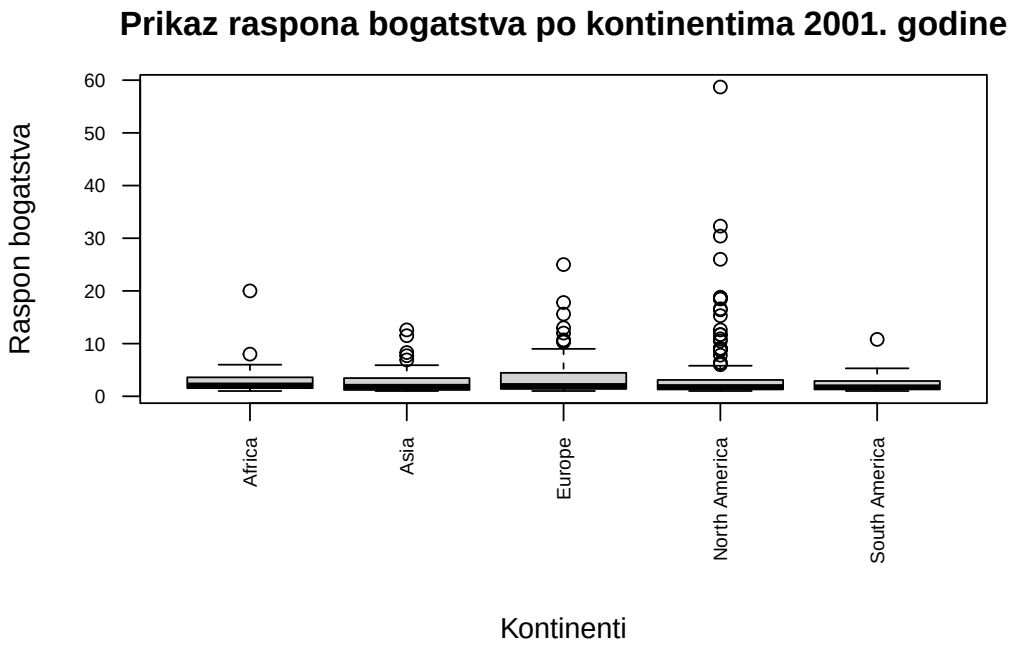
```
##
##  Chi-squared test for given probabilities
##
## data:  v3
## X-squared = 552.52, df = 4, p-value < 2.2e-16
```

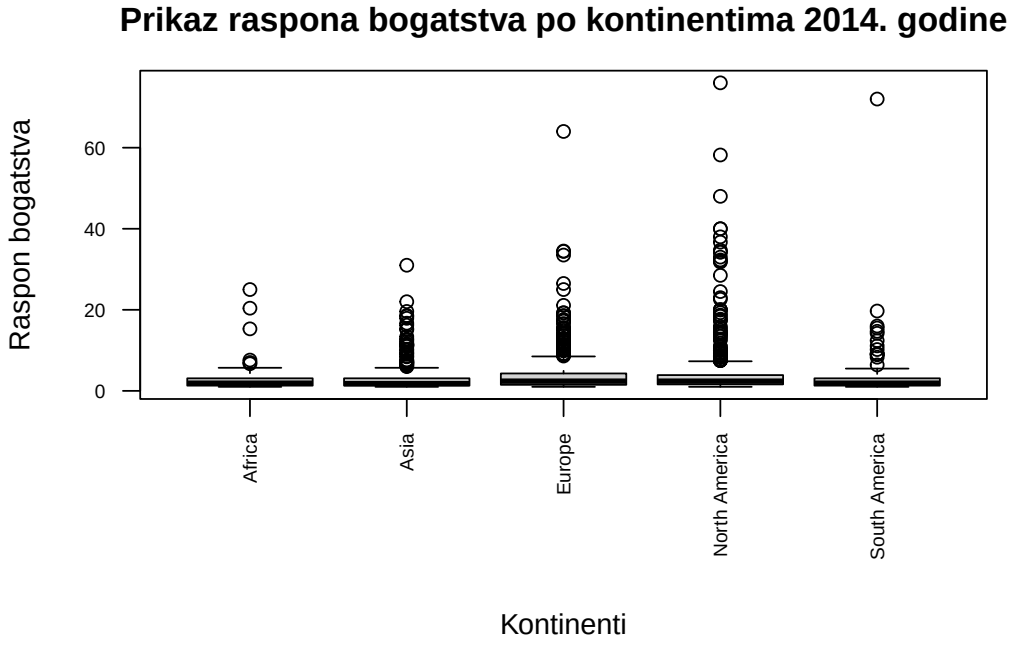
Na temelju veoma malih p-vrijednosti (puno manje od razine značajnosti 0.05), zaključujemo da broj milijardera po kontinentu ne možemo usporediti s uniformnom distribucijom - broj milijardera na određenim kontinentima značajno je veći od drugih (za sve tri godine).

1. zadatak, nova verzija

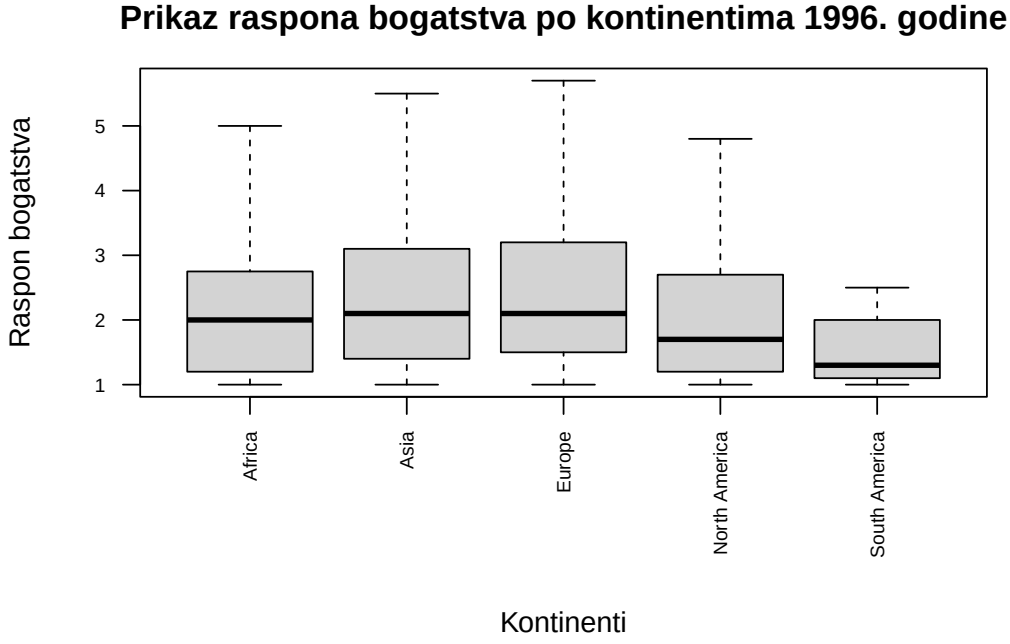
Kako bismo provjerili ima li neki kontinent značajno više milijardi tj. postoji li kontinent s prosjekom iznosa milijardi po milijarderu koji značajno odstupa od ostalih, prvo ćemo raspon bogatstva po kontinentima prikazati koristeći 3 boxplota, svaki za zasebnu godinu.

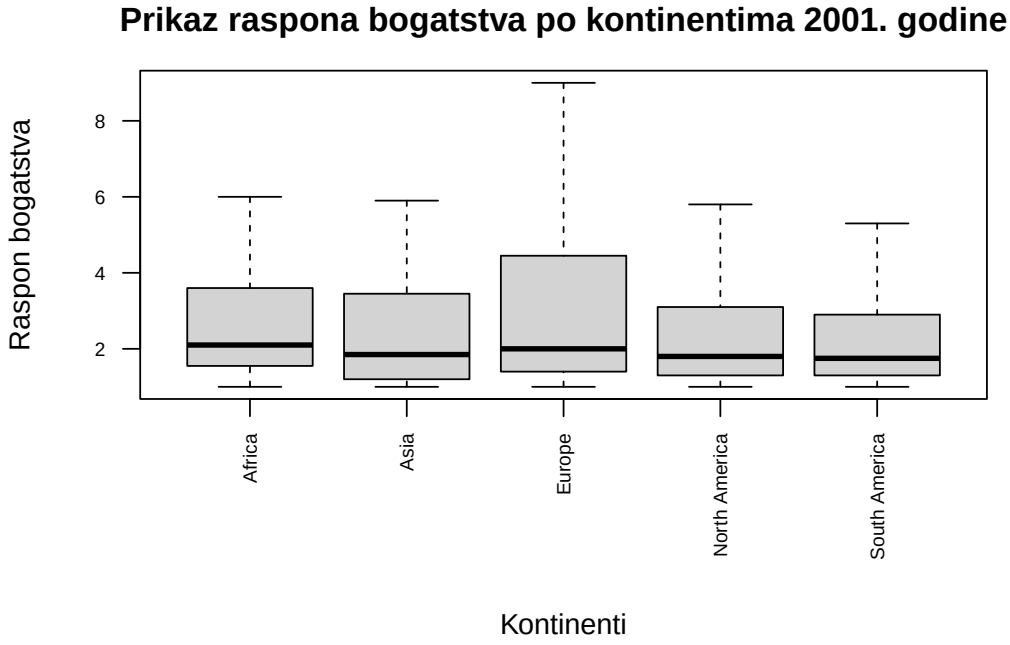

Prikaz raspona bogatstva po kontinentima 1996. godine
Raspon bogatstva
15
10
5
Africa
Asia
Europe
North America
South America
Kontinenti

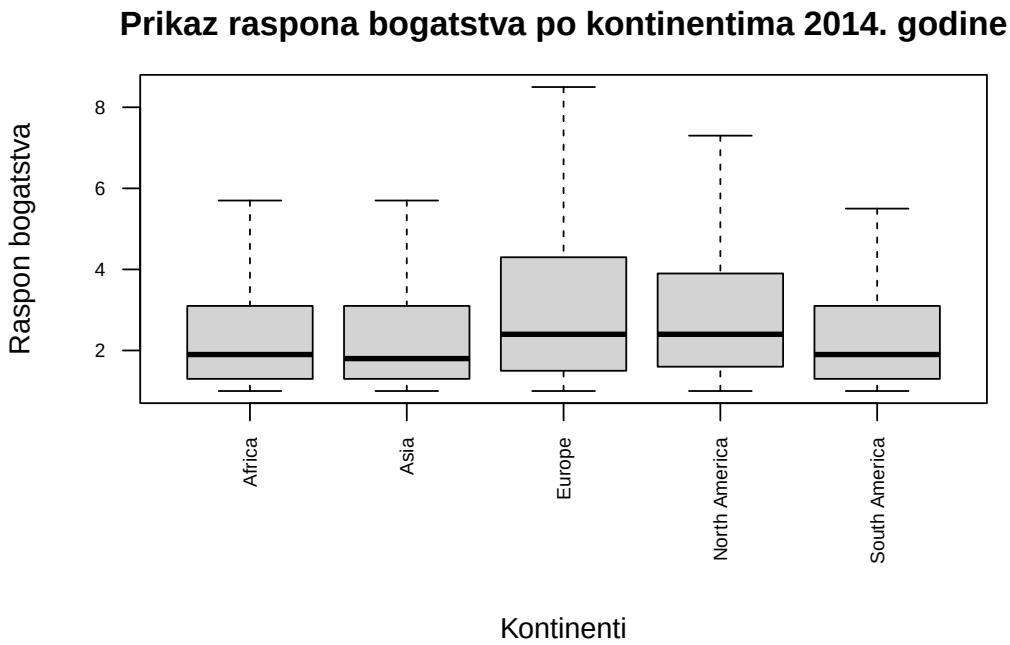

Prikaz raspona bogatstva po kontinentima 2001. godine
Raspon bogatstva
60
50
40
30
20
10
0
Africa
Asia
Europe
North America
South America
Kontinenti



Na prvi pogled, uočavamo da je sredina (medijan) bogatstva veoma blizu za sve kontinente, ali da određeni kontinenti (Europa, Sjeverna Amerika, Azija) imaju i velik broj stršećih vrijednosti - vrijednosti izvan gornjeg "brka" boxplota. Stršeće vrijednosti veoma velikih vrijednosti spljošćuju prikaz pa je zbog toga teško doći do pravilnog zaključka. Zbog toga ćemo prikazati boxplotove i bez njih.







Na prikazu bez stršećih vrijednosti možemo vidjeti da za 1996. godinu Afrika, Azija i Europa imaju viši medijan od Sjeverne i Južne Amerike, a pritom daleko najniži medijan ima Južna Amerika. Za 2001. godinu medijani svih kontinenata su podjednaki, a za 2014. godinu izgleda da je medijan Europe i Sjeverne Amerike viši od medijana preostala tri kontinenta. Važno je istaknuti i da za 2001. te 2014. godinu Europa ima veoma

velik IQR što upućuje na veliku varijabilnost podataka za taj kontinent.

Kako bismo ove zaključke poduprli testiranjem, za sve tri godine pokušat ćemo provesti testiranje jednakosti sredina jednofaktorskom analizom varijance pri čemu će faktor biti parametar location.region. Da bismo mogli provoditi Anovu, trebale bi vrijediti sljedeće pretpostavke: populacije trebaju biti nezavisne i imati normalnu distribuciju, a varijance za svaku grupu trebale bi biti homogene. Ako ove pretpostavke nisu zadovoljene, moguće je koristiti i neparametarsku verziju testa, Kruskal-Wallisov test.

Kako bismo testirali normalnost podataka, provest ćemo Lillieforsovu inačicu KS testa za sve grupe: u slučaju da je p-vrijednost manja od 0.05, odbacujemo nultu hipotezu da podaci potječu iz normalne distribucije. Kako bismo testirali homogenost varijance, provest ćemo Bartlettov test: u slučaju da je p-vrijednost manja od 0.05, odbacujemo nultu hipotezu da su varijance svih grupa podataka jednake.

```
## Loading required package: nortest
```

```
## Test normalnosti podataka za 1996. godinu
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  billionaires.1996$wealth.worth.in.billions
## D = 0.23464, p-value < 2.2e-16
```

```
## Test za provjeru jednakosti varijanci za 1996. godinu
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  billionaires.1996$wealth.worth.in.billions by billionaires.1996$location.region
## Bartlett's K-squared = 34.665, df = 4, p-value = 5.443e-07
```

Zbog sažetosti prikaza, nije prikazano provođenje prikladnih testova za godine 2001. i 2014., a za godinu 1996. prikazan je rezultat provođenja Lillieforsove inačice KS testa nad cijelom populacijom (provedeni su i testovi za zasebne kontinente, ali je njihovo izvođenje zbog sažetosti zakomentirano), kao i rezultat provođenja Bartlettovog testa za homogenost varijanci. P-vrijednosti za sve testove manje su od 0.05.

Na razini značajnosti 0.05 zbog veoma niske p-vrijednosti odbacujemo nultu hipotezu Lillieforseove inačice KS testa za cijelu populaciju, ali i za pojedinačne grupe (za sve tri godine) - ne možemo pretpostaviti normalnost podataka. Isto tako, na istoj razini značajnosti odbacujemo i nultu hipotezu Bartlettovog testa (za sve tri godine) - ne možemo pretpostaviti homogenost varijanci. Na temelju ovih rezultata, možemo pretpostaviti da rezultat provođenja testova na temelju jednofaktorske Anove nije veoma vjerodostojan.

Ipak, provest ćemo jednofaktorsku Anovu, oslanjajući se pritom na robustnost testa jednakosti sredina s nultom hipotezom da su srednje vrijednosti svih uzoraka jednake. Uz to, provest ćemo i Kruskal-Wallisov test čija je nulta hipoteza da su medijani svih uzoraka jednaki. Pritom treba paziti na uvjet za primjenjivost Kruskal-Wallisovog testa: veličina svakog uzorka mora biti barem 5.

```
## Testovi za 1996. godinu (frekvencije po kontinentima:  19 119 104 141 40 )
```

```
##                  Df Sum Sq Mean Sq F value Pr(>F)
## location.region   4   42.9  10.732   2.601 0.0357 *
## Residuals       418 1724.8   4.126
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  wealth.worth.in.billions by location.region
## Kruskal-Wallis chi-squared = 23.831, df = 4, p-value = 8.636e-05
```

```
## Testovi za 2001. godinu (frekvencije po kontinentima:  24 72 124 289 28 )
##                   Df Sum Sq Mean Sq F value Pr(>F)
## location.region    4     43   10.65   0.539  0.707
## Residuals        532  10512   19.76
##
##  Kruskal-Wallis rank sum test
##
## data:  wealth.worth.in.billions by location.region
## Kruskal-Wallis chi-squared = 5.2313, df = 4, p-value = 0.2644
## Testovi za 2014. godinu (frekvencije po kontinentima:  94 413 470 562 114 )
##                    Df Sum Sq Mean Sq F value   Pr(>F)
## location.region     4    630  157.39   4.807 0.000742 ***
## Residuals        1648  53962   32.74
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Kruskal-Wallis rank sum test
##
## data:  wealth.worth.in.billions by location.region
## Kruskal-Wallis chi-squared = 41.766, df = 4, p-value = 1.865e-08
```

Na razini značajnosti 0.05, za godine 1996. i 2014. odbacujemo H0 i za jednofaktorsku Anovu (srednje vrijednosti svih uzoraka su jednake), ali i za Kruskal-Wallisov test (medijani svih uzoraka su jednaki). Za 2001. godinu ne možemo odbaciti H0 ni za jedan od navedenih testova. Naravno, testovima ne možemo potvrditi H0 tj. reći da ona sigurno vrijedi, ali možemo reći da je velika vjerojatnost da su 2001. godine srednje vrijednosti i medijani bili otprilike jednaki, dok za preostale dvije godine nisu (1996. i 2014. godine postoje kontinenti koji imaju statistički značajno veću sredinu milijardi od ostalih). Uočavamo da Kruskal-Wallisov test kao rezultat ima manje p-vrijednosti od Anove.

Kako bismo potvrdili jesu li naše pretpostavke napravljene na temelju boxplota ispravne, dodatno bi mogli provesti i usporedbe s jednim stupnjem slobode.

2. zadatak

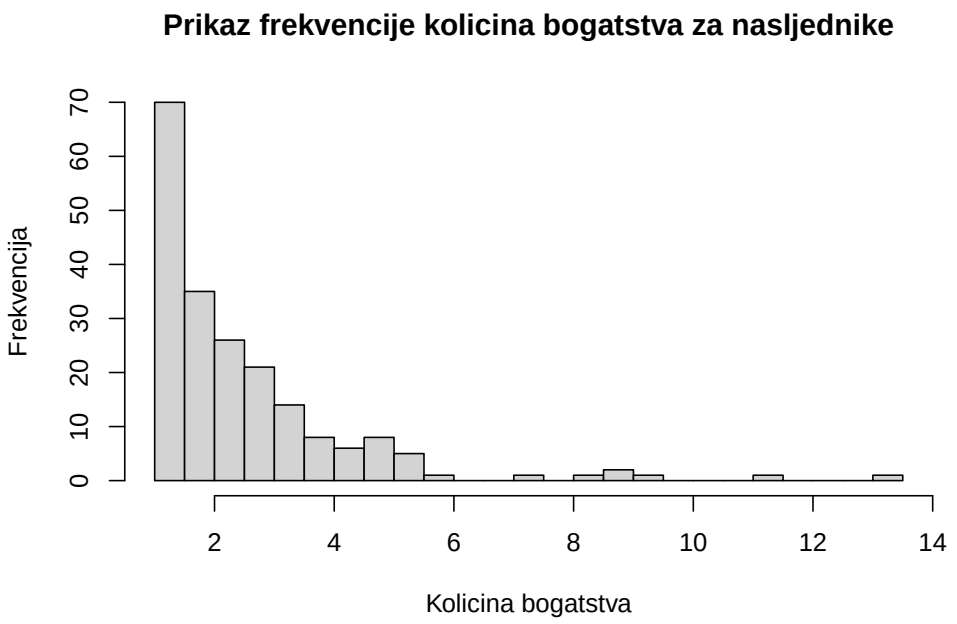
Kako bismo odredili postoji li značajna razlika u bogatstvu između ljudi koji su naslijedili bogatstvo i onih koji nisu, ponovno ćemo koristiti tablice koje predstavljaju broj milijardera ovisno o godini (na početku smo utvrdili da svi milijarderi imaju podatak o tome jesu li naslijedili bogatstvo, kao i o godini koje su bili na određenom ranku pa stoga nemamo potrebu za dodatnom eliminacijom podataka iz tablice).

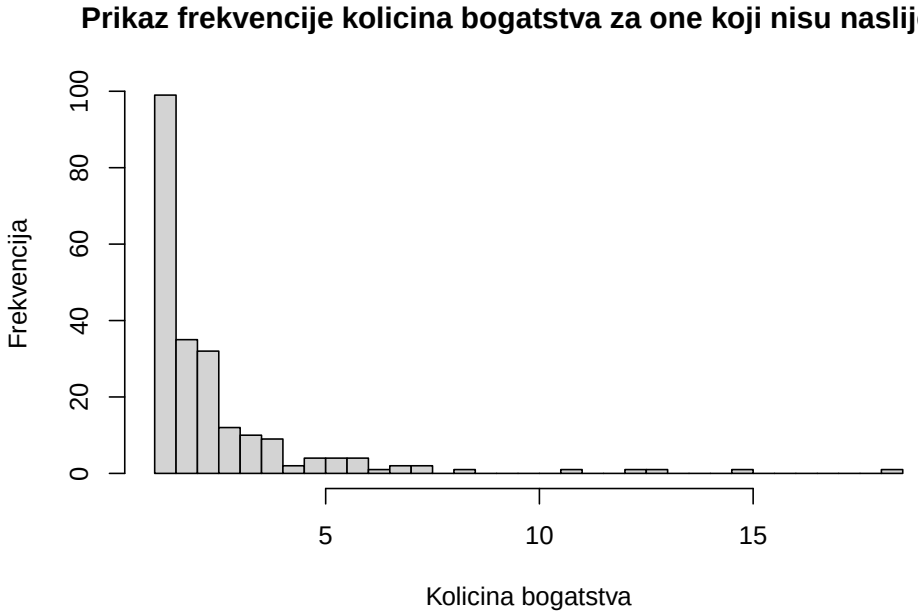
Za svaku od te tri godine, stvorit ćemo po 2 tablice - jedna će sadržavati podatke o onima koji su naslijedili bogatstvo, a druga o onima koji nisu. Zbog jednostavnijeg prikaza kasnije, vrijednosti atributa wealth.how.inherited zamijenit ćemo vrijednostima "inherited" te "not inherited".

```
billionaires.1996 <- billionaires[billionaires["year"] == 1996,]
billionaires.2001 <- billionaires[billionaires["year"] == 2001,]
billionaires.2014 <- billionaires[billionaires["year"] == 2014,]

for (i in 1:length(billionaires.1996$wealth.how.inherited)) {
  if (billionaires.1996$wealth.how.inherited[i] == "not inherited") {
    next
  } else {
    billionaires.1996$wealth.how.inherited[i] <- "inherited"
  }
}
```

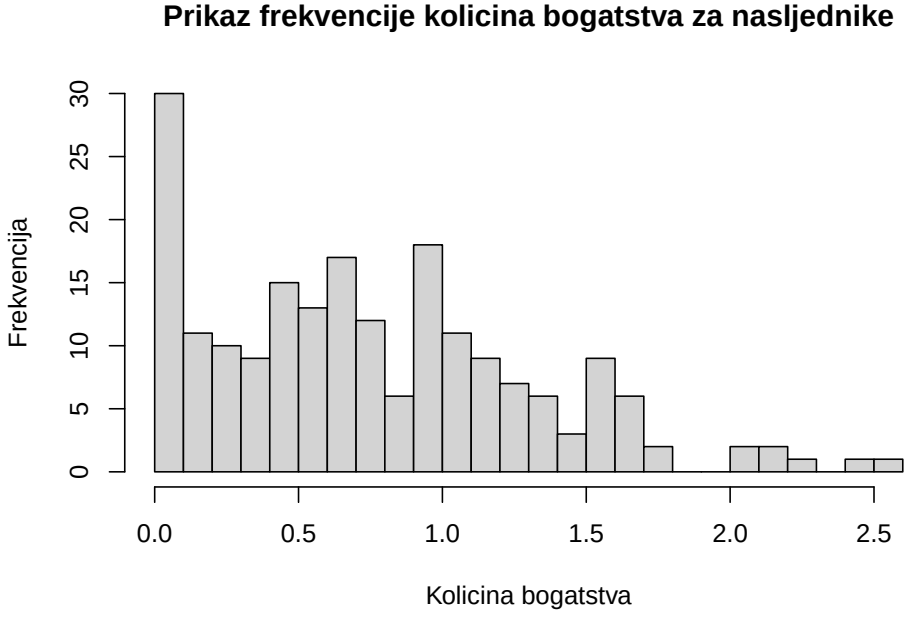
```r
for (i in 1:length(billionaires.2001$wealth.how.inherited)) {
  if (billionaires.2001$wealth.how.inherited[i] == "not inherited") {
    next
  } else {
    billionaires.2001$wealth.how.inherited[i] <- "inherited"
  }
}

for (i in 1:length(billionaires.2014$wealth.how.inherited)) {
  if (billionaires.2014$wealth.how.inherited[i] == "not inherited") {
    next
  } else {
    billionaires.2014$wealth.how.inherited[i] <- "inherited"
  }
}

billionaires.1996.inherited <- billionaires.1996[billionaires.1996["wealth.how.inherited"] != "not inhe
billionaires.1996.not.inherited <- billionaires.1996[billionaires.1996 ["wealth.how.inherited"] == "not
billionaires.2001.inherited <- billionaires.2001[billionaires.2001["wealth.how.inherited"] != "not inhe
billionaires.2001.not.inherited <- billionaires.2001[billionaires.2001["wealth.how.inherited"] == "not
billionaires.2014.inherited <- billionaires.2014[billionaires.2014["wealth.how.inherited"] != "not inhe
billionaires.2014.not.inherited <- billionaires.2014[billionaires.2014["wealth.how.inherited"] == "not
```

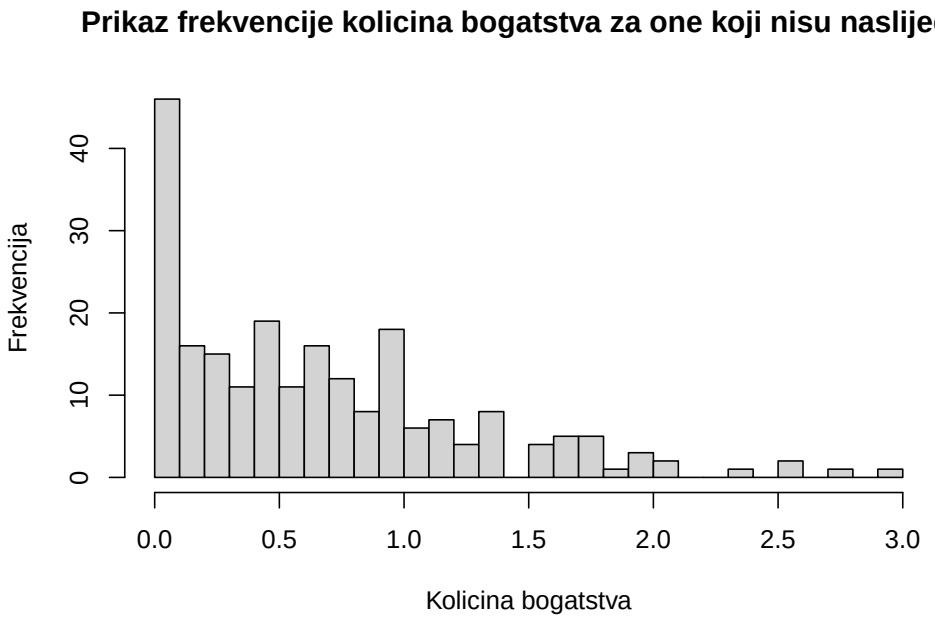
Količinu bogatstva pojedinih milijardera prikazat ćemo histogramima.



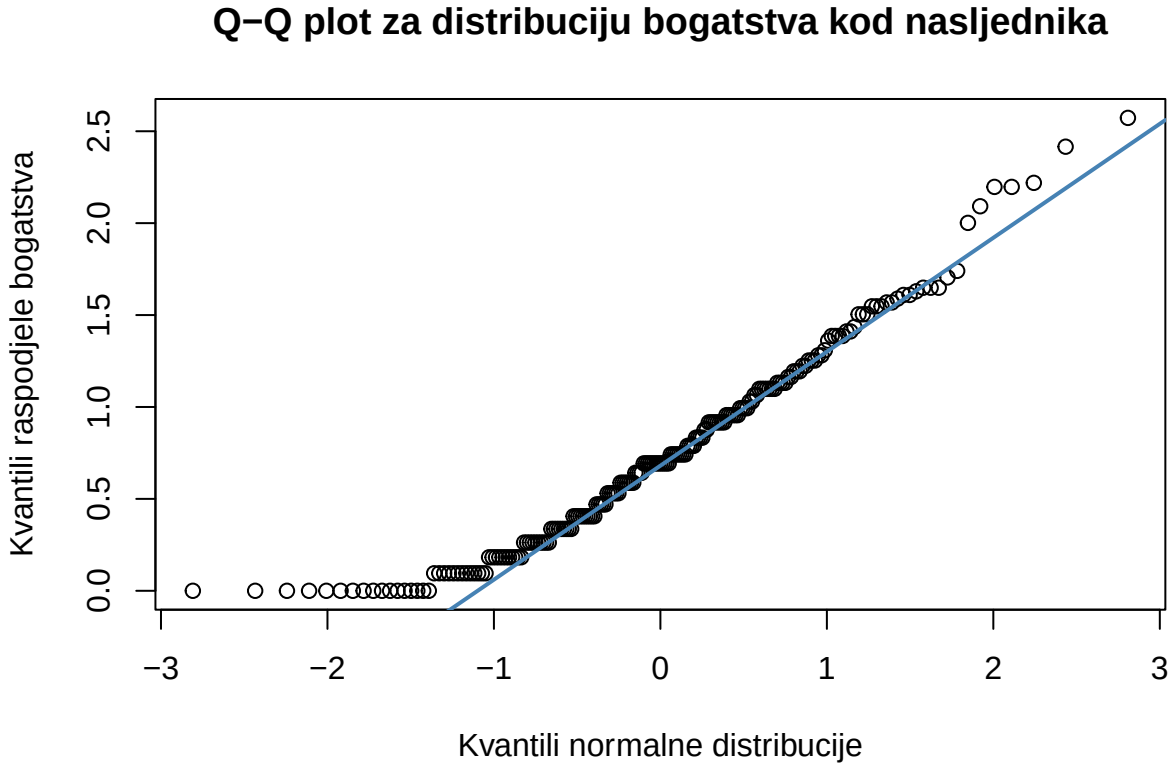


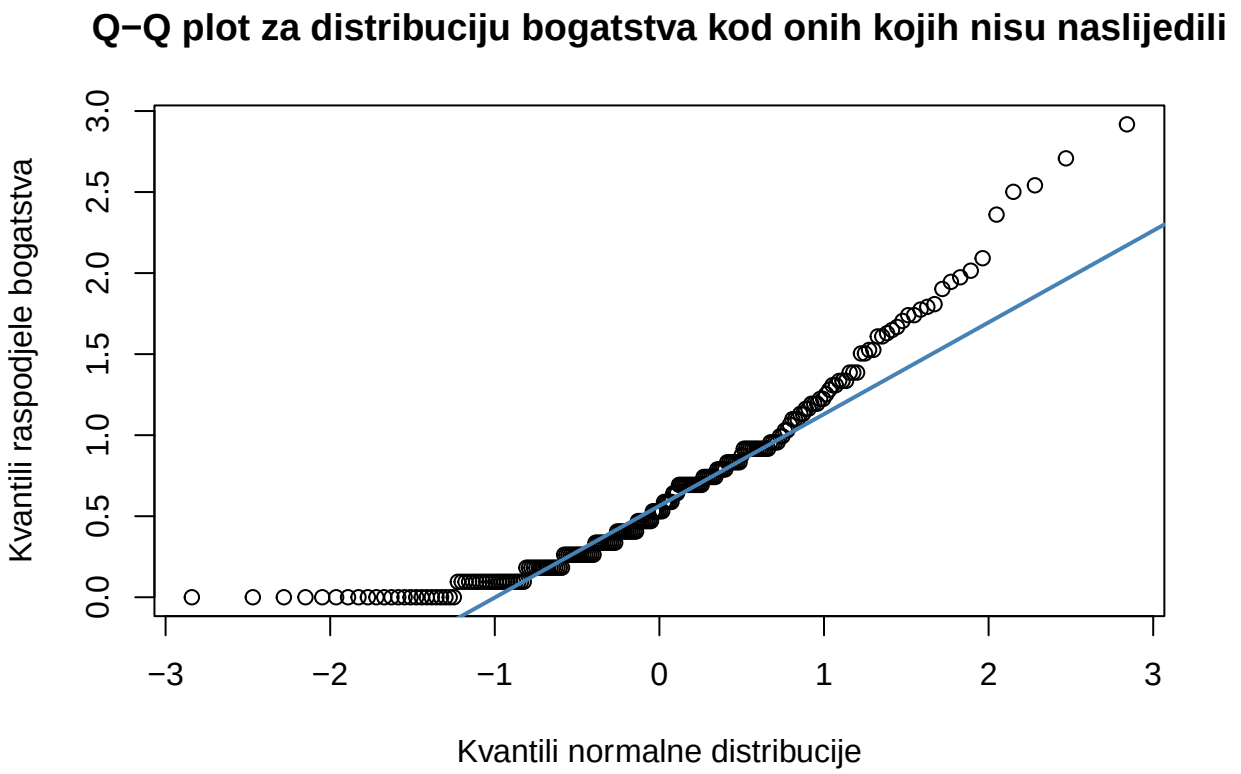
Prikaz stavljamo samo za jednu od godina zbog sažetosti - histogrami za preostale godine izgledaju veoma slično ovima. Za sve histograme uočavamo veoma izraženi desni rep. Kako bismo mogli provoditi F-test, kao i t-test, potrebna je barem približna normalnost podataka. Zbog toga ćemo podatke o iznosu bogatstva transformirati koristeći funkciju log.





Kao i prethodni put, zbog sažetosti su prikazani histogrami samo za jednu godinu. Distribucija podataka sada je puno bliže normalnoj, ali je desni rep i dalje veoma izražen - teško je pretpostaviti normalnost podataka. Ovo ćemo provjeriti prikazivanjem Q-Q grafa za 1996. godinu.





Na temelju prikazanih Q-Q grafova, vidimo da je distribucija čak i logaritmiranih iznosa bogatstva daleko od normalne. Q-Q grafovi izgledaju veoma slično i za preostale dvije godine, a zbog sažetosti ovdje nisu prikazani. Kako bismo dodatno provjerili možemo li provoditi daljnje testove s pretpostavkom normalnosti podataka, provest ćemo Lillieforsovu inačicu Kolmogorov-Smirnov testa.

```
## Lilliefors test za one koji su naslijedili bogatstvo, 1996. godina

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  billionaires.1996.inherited.log$wealth.worth.in.billions
## D = 0.086995, p-value = 0.0008251

## Lilliefors test za one koji nisu naslijedili bogatstvo, 1996. godina

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  billionaires.1996.not.inherited.log$wealth.worth.in.billions
## D = 0.13035, p-value = 8.286e-10

## Lilliefors test za one koji su naslijedili bogatstvo, 2001. godina

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  billionaires.2001.inherited.log$wealth.worth.in.billions
## D = 0.10622, p-value = 2.097e-06

## Lilliefors test za one koji nisu naslijedili bogatstvo, 1996. godina
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  billionaires.2001.not.inherited.log$wealth.worth.in.billions
## D = 0.15784, p-value < 2.2e-16
```

```
## Lilliefors test za one koji su naslijedili bogatstvo, 2014. godina
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  billionaires.2014.inherited.log$wealth.worth.in.billions
## D = 0.095729, p-value = 6.882e-12
```

```
## Lilliefors test za one koji nisu naslijedili bogatstvo, 2014. godina
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  billionaires.2014.not.inherited.log$wealth.worth.in.billions
## D = 0.12199, p-value < 2.2e-16
```

Na temelju rezultata izvođenja Lillieforsove inačice KS testa, zbog veoma male p-vrijednosti (manje od 0.05) odbacujemo H0: podaci dolaze iz normalne razdiobe (za obje skupine za sve tri godine).

U daljnjem testiranju oslanjat ćemo se na robustnost t-testa, ali ćemo nakon provođenja parametarskih testova provesti i neparametarske inačice. Uz dodatnu pretpostavku nezavisnosti podataka, sada ćemo provesti Fisherov test kako bismo odredili možemo li pri t-testu koristiti varijantu s jednakim varijancama.

```
## F-test za provjeru jednakosti varijanci, 1996. godina
```

```
##
##  F test to compare two variances
##
## data:  billionaires.1996.inherited.log$wealth.worth.in.billions and billionaires.1996.not.inherited.l
## F = 0.86996, num df = 200, denom df = 221, p-value = 0.3153
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.6638944 1.1421064
## sample estimates:
## ratio of variances
##          0.8699597
```

```
## F-test za provjeru jednakosti varijanci, 2001. godina
```

```
##
##  F test to compare two variances
##
## data:  billionaires.2001.inherited.log$wealth.worth.in.billions and billionaires.2001.not.inherited.l
## F = 1.0444, num df = 222, denom df = 314, p-value = 0.7206
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8207317 1.3359471
## sample estimates:
## ratio of variances
##           1.044419
```

```
## F-test za provjeru jednakosti varijanci, 2014. godina
```

```
##
```

```
##  F test to compare two variances
##
## data:  billionaires.2014.inherited.log$wealth.worth.in.billions and billionaires.2014.not.inherited.l
## F = 1.0829, num df = 501, denom df = 1150, p-value = 0.2858
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.9355341 1.2590879
## sample estimates:
## ratio of variances
##           1.082939
```

Za sva tri testa dobili smo p-vrijednost značajno veću od 0.05. Zbog toga ćemo pri provođenju t-testa ići s pretpostavkom da su varijance populacija jednake (ne odbacujemo H0 F-testa). Hipoteze pri provođenju t-testa nad dvije populacije biti će:

H0: razlika srednjih vrijednosti između dviju populacija jednaka je 0
H1: razlika srednjih vrijednosti između dvije populacije veća je od 0

pri čemu je prva populacija skupina milijardera koji su naslijedili bogatstvo, a druga populacija skupina milijardera koji nisu naslijedili bogatstvo.

```
## t-test za jednakost srednjih vrijednosti dvije populacije, 1996. godina

##
##  Two Sample t-test
##
## data:  billionaires.1996.inherited.log$wealth.worth.in.billions and billionaires.1996.not.inherited.l
## t = 1.5191, df = 421, p-value = 0.06474
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.007264524          Inf
## sample estimates:
## mean of x mean of y
## 0.7549957 0.6696943

## t-test za jednakost srednjih vrijednosti dvije populacije, 2001. godina

##
##  Two Sample t-test
##
## data:  billionaires.2001.inherited.log$wealth.worth.in.billions and billionaires.2001.not.inherited.l
## t = 2.662, df = 536, p-value = 0.004001
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.06376485        Inf
## sample estimates:
## mean of x mean of y
## 0.9070956 0.7397417

## t-test za jednakost srednjih vrijednosti dvije populacije, 2014. godina

##
##  Two Sample t-test
##
## data:  billionaires.2014.inherited.log$wealth.worth.in.billions and billionaires.2014.not.inherited.l
## t = 4.7793, df = 1651, p-value = 9.574e-07
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
```

```
##  0.1296578       Inf
## sample estimates:
## mean of x mean of y
## 1.0879316 0.8901764
```

Rezultati provođenja t-testa nad dvije populacije s razinom značajnosti 0.05 upućuju na sljedeći zaključak:

-za godinu 1996 ne možemo odbaciti H0 - ne postoji značajna razlika između srednjih vrijednosti bogatstva onih koji su bogatstvo naslijedili i onih koji bogatstvo nisu naslijedili

-za godinu 2001 odbacujemo H0 - postoji značajna razlika između srednjih vrijednosti bogatstva onih koji su bogatstvo naslijedili i onih koji bogatstvo nisu naslijedili (srednja vrijednost bogatstva onih koji su naslijedili bogatstvo vrlo je vjerojatno veća od srednje vrijednosti bogatstva onih koji nisu naslijedili)

-za godinu 2014 odbacujemo H0 - postoji značajna razlika između srednjih vrijednosti bogatstva onih koji su bogatstvo naslijedili i onih koji bogatstvo nisu naslijedili (srednja vrijednost bogatstva onih koji su naslijedili bogatstvo vrlo je vjerojatno veća od srednje vrijednosti bogatstva onih koji nisu naslijedili)

Uz to, uočavamo da je p-vrijednost za godinu 1996 ipak veoma blizu razini značajnosti pa je za tu godinu teško dati pouzdani zaključak - lako je moguće da bi s malo drugačijim datasetom rezultat bio odbacivanje H0. Provest ćemo i neparametarsku alternativu t-testu: Mann-Whitney-Wilcoxonov test sa sljedećim hipotezama:

H0: medijani obje populacije jednaki su
H1: medijan prve populacije veći je od medijana druge populacije

pri čemu je prva populacija skupina milijardera koji su bogatstvo naslijedili, a druga populacija skupina milijardera koji bogatstvo nisu naslijedili. Pretpostavka testa je da su uzorci iz istih distribucija (do na translaciju).

```
## MWW test, 1996. godina
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  billionaires.1996.inherited$wealth.worth.in.billions and billionaires.1996.not.inherited$weal
## W = 24879, p-value = 0.0203
## alternative hypothesis: true location shift is greater than 0
```

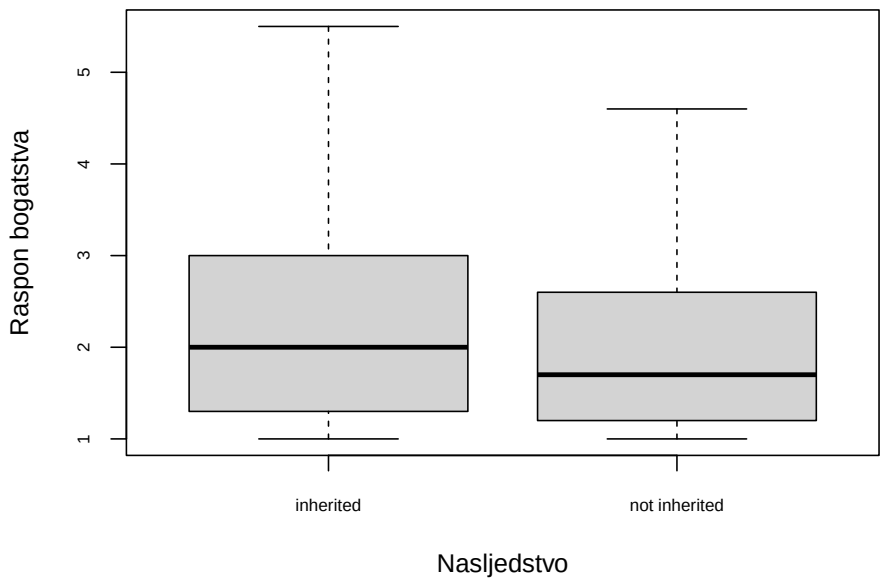
```
## MWW test, 2001. godina
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  billionaires.2001.inherited$wealth.worth.in.billions and billionaires.2001.not.inherited$weal
## W = 40563, p-value = 0.001086
## alternative hypothesis: true location shift is greater than 0
```

```
## MWW test, 2014. godina
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  billionaires.2014.inherited$wealth.worth.in.billions and billionaires.2014.not.inherited$weal
## W = 336316, p-value = 5.302e-08
## alternative hypothesis: true location shift is greater than 0
```

Na razini značajnosti 0.05, odbacujemo H0 za sve tri godine - ne možemo pretpostaviti da je medijan bogatstva za one koji su naslijedili bogatstvo i one koji nisu naslijedili bogatstvo jednak. Budući da je H1 bio da je medijan onih koji su bogatstvo naslijedili veći od medijana onih koji bogatstvo nisu naslijedili, zaključujemo da je vrlo vjerojatno da su oni koji su naslijedili bogatstvo u prosjeku bogatiji od onih koji nisu. Ovo ćemo provjeriti koristeći prikaz boxplotom za zasebne godine, pritom prikazujući boxplot s i bez stršećih vrijednosti.



Zbog sažetosti prikaza, prikazani su boxplotovi samo za 1996. godinu. Boxplotovi za preostale dvije godine veoma su slični ovima. Možemo uočiti da je medijan bogatstva kod onih koji su naslijedili bogatstvo viši od medijana onih koji bogatstvo nisu naslijedili. Uz to, možemo uočiti da u grupi onih koji bogatstvo nisu naslijedili ima puno više stršećih vrijednosti. Drugim riječima, u grupi onih koji bogatstvo nisu naslijedili ima velik broj milijardera čija količina bogatstva veoma odudara od prosjeka.

3. zadatak

Kako bismo provjerili koreliranost između određenih parametara i količine bogatstva, prvo ćemo provesti linearnu regresiju na zasebnim parametrima.

U podacima smo uočili prisutnost 3 boolean varijable koje za svaki unos imaju vrijednost True. Zbog toga te varijable nećemo razmatrati pri provođenju linearne regresije. Uz to, kao regresor nećemo razmatrati rank jer je on ovisan o količini bogatstva, a ne obrnuto. Isto tako, nećemo razmatrati ni ime milijardera ni ime tvrtke - u pitanju su kategoričke varijable s veoma velikim brojem mogućih vrijednosti.

Regresiju ćemo provoditi na setu podataka pročišćenom od svim neunesenih podataka (negativni brojevi, 0, prazan string) - neuneseni podaci u tablici će biti zamijenjeni vrijednošću NA. Uz to, na taj dataset primijenit ćemo transformaciju tako da uklonimo sve whitespace znakove s početka ili kraja stringova, kao i promijenimo sva slova u lowercase. Time će za određene kategorijske varijable biti ukupno manje kategorija (npr. "finance" i "Finance" bit će jedna kategorija).

```
require(tidyverse)

clean_string <- function(x) {
  if (x == '' | (is.numeric(x) & x <= 0) | (is.character(x) & x == '0')) {
    return (NA)
  } else if (!is.numeric(x)) {
    return (str_squish(tolower(x)))
  } else {
    return (x)
  }
}

billionaires.unique <- billionaires

for (i in 1:nrow(billionaires)) {
  for (j in 1:length(billionaires)) {
    billionaires.unique[i, j] <- clean_string(billionaires.unique[i, j])
  }
}
```

Nad pročišćenim podacima provest ćemo linearnu regresiju - prvo na zasebnim parametrima da odredimo parametre koji su značajni i sami po sebi, a nakon toga na kombinacijama parametara, pritom pokušavajući model ostaviti što jednostavnijim (što manje parametara).

Zbog sažetosti, provođenje linearne regresije na po 1 regresoru nije prikazano. Provođenjem linearne regresije na po 1 regresoru, bilo numeričkom ili kategoričkom, dolazimo do zaključka da su samostalno statistički značajni parametri year, company.type, demographics.age, location.region, wealth.type, wealth.how.industry, wealth.how.inherited.

Kombinacijom regresora pokušat ćemo postići objašnjenje većeg postotka varijance modelom, pritom koristeći R^2 i prilagođeni R^2 kao glavnu metriku.

```
#### Kombinirano

# svi parametri
# fit.combined = lm(log(wealth.worth.in.billions)~year + company.founded + company.relationship + compa
# summary(fit.combined)
# Multiple R-squared:  0.4624,  Adjusted R-squared:  0.1595
# F-statistic: 1.527 on 218 and 387 DF,  p-value: 0.0001597
# vecina unosa deleted zbog NA kod location.gdp
```

```
# minimalni model

# fit.combined = lm(log(wealth.worth.in.billions)~year + company.relationship + company.sector + demogr
# summary(fit.combined)
# Multiple R-squared:  0.3505,  Adjusted R-squared:  0.1546
# F-statistic: 1.789 on 515 and 1708 DF,  p-value: < 2.2e-16
```

Zbog sažetosti i preglednosti, kod za provođenje linearne regresije i prikaz sažetka o modelima zakomentiran je. Kombiniranjem različitih regresora pri provođenju linearne regresije moguće je objasniti veći postotak ukupne varijance. Pritom je potrebno pokušavati postići što jednostavniji model kod kojeg bi regresori međusobno trebali biti neovisni. Kako bismo modelom objasnili što veći udio varijance, transformirali smo izlaz tj. zavisnu varijablu log-transformacijom.

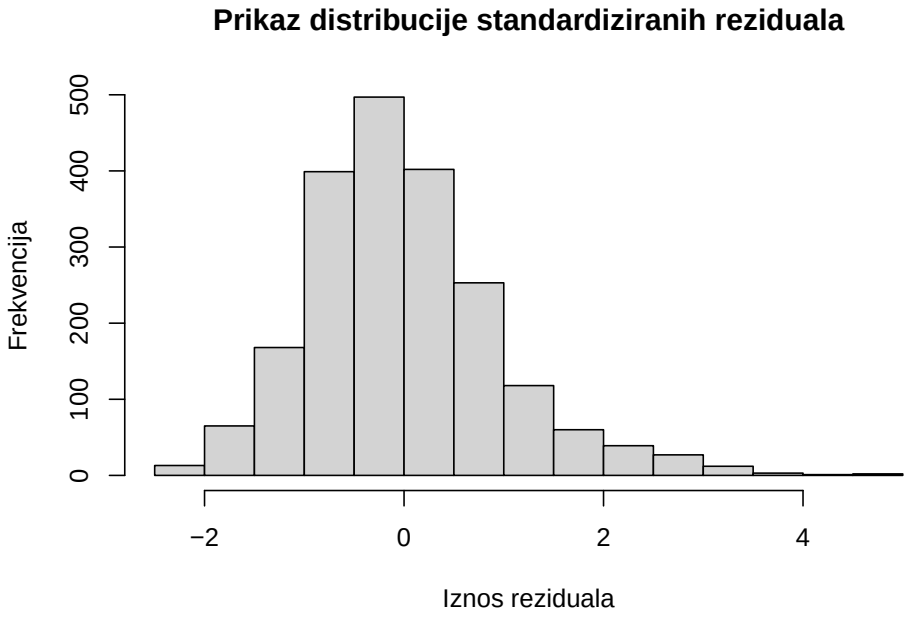
Kombinacijom parametara year, company.relationship, company.sector i demographics.age kao regresora dobivamo iznos R^2 = 0.3505 i iznos prilagođenog R^2 = 0.1546. Prilagođeni R^2 značajno je niži zbog veoma visokog broja razina (engl. levels) kategorijske varijable company.sector od kojih nisu sve jednako značajne za model.

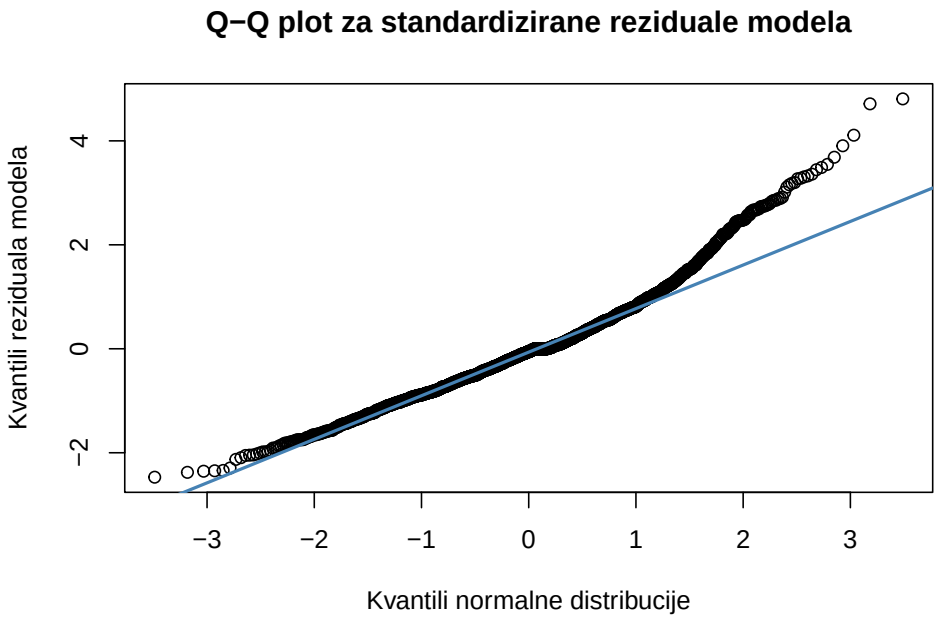
U slučaju da koristimo zakomentirani model s 13 parametara, iznos R^2 bio bi jednak 0.4624, ali korištenjem tog modela uvodimo prevelik broj parametara - cilj nam je na kraju koristiti što jednostavniji model (jedan od razloga za ovo je želja za izbjegavanjem pojave da model dobro predviđa samo za ovaj konkretni skup podataka - unosom novih podataka model bi mogao imati puno lošija previđanja). U modelu s 4 gore navedenih parametara, sva 4 parametra veoma utječu na iznos R^2 i iznos prilagođenog R^2. Izostavljanjem pojedinih parametara, posebice parametara company.relationship i company.sector, udio varijance objašnjene modelom značajno pada. Značajnost pojedinih parametara provjerit ćemo korištenjem funkcije anova:

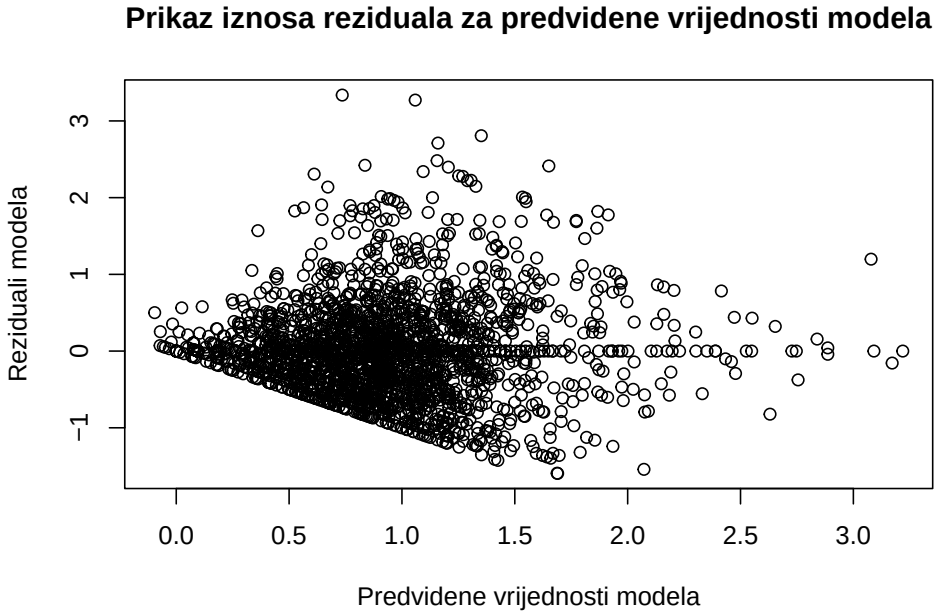
```
## Analysis of Variance Table
##
## Response: log(wealth.worth.in.billions)
##                         Df Sum Sq Mean Sq F value    Pr(>F)
## year                     1   7.48  7.4808 15.2770 9.647e-05 ***
## company.relationship    67  81.89  1.2222  2.4960 6.536e-10 ***
## company.sector         446 341.84  0.7665  1.5652 2.427e-10 ***
## demographics.age         1  20.04 20.0422 40.9294 2.033e-10 ***
## Residuals             1708 836.37  0.4897
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Na temelju rezultata dobivenih korištenjem funkcije anova, vidimo da su na razini značajnosti od 0.05 sva četiri parametra veoma značajna u modelu, a da najveću količinu varijance objašnjavaju parametri company.relationship i company.sector. Naravno, treba razmišljati o tome da Anova kao pretpostavku ima normalnost podataka iz svih populacija, kao i homoskedastičnost varijanci. Za trenutni skup podataka, već znamo da je normalnost narušena pa rezultati provođenja Anova postupka nisu nužno vjerodostojni.

Za provođenje dodatnih testova koristeći naš model, trebali bismo provjeriti svojstva reziduala - očekujemo da su reziduali homogeni, nezavisni i normalno distribuirani. U slučaju da ove pretpostavke ne vrijede, narušena je vjerodostojnost dodatnih testova provedenih koristeći naš model. Kako bismo provjerili svojstva reziduala našeg modela, prikazat ćemo standardizirane reziduale pomoću histograma, Q-Q grafa, a provest ćemo i Lillieforsovu inačicu KS testa.


Prikaz distribucije standardiziranih reziduala
Frekvencija
0
100
200
300
400
500
−2
0
2
4
Iznos reziduala


Q−Q plot za standardizirane reziduale modela
Kvantili reziduala modela
−2
0
2
4
−3
−2
−1
0
1
2
3
Kvantili normalne distribucije




```
## Lilliefors test za normalnost standardiziranih reziduala
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rstandard(model)
## D = 0.07887, p-value < 2.2e-16
```

Iz histograma, Q-Q grafa i Lilliforsove inačice KS testa vidljivo je da reziduali nisu normalno distribuirani (izražen dugačak desni rep distribucije reziduala) - u slučaju provođenja dodatnih testova s modelom, vjerodostojnost testova bi bila veoma narušena.

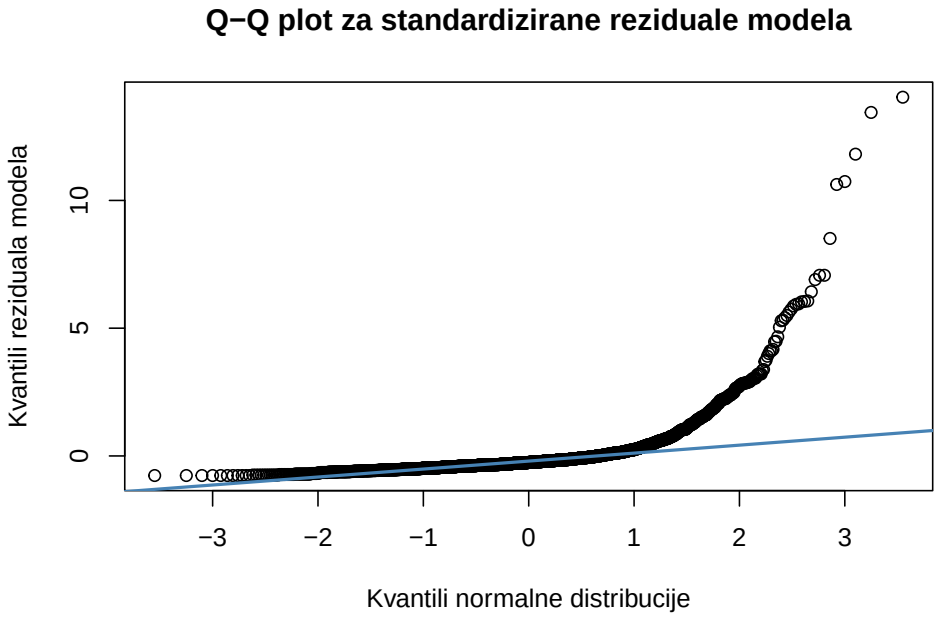
Na temelju iznosa R^2 = 0.3505 vidimo da naš konačni model s 4 regresora uspješno objašnjava 35.05% varijance u iznosu logaritmirane zavisne varijable wealth.worth.in.billions. Drugim riječima, našim modelom mogao bi se pokušati predvidjeti logaritmiran iznos bogatstva određenog milijardera na temelju vrijednosti regresora, ali ta vrijednost vjerojatno ne bi bila veoma precizna - što višu količinu varijance model objašnjava, veća je preciznost samih predikcija. Nažalost, veoma je malen broj parametara s brojčanom vrijednošću pa je stoga model teško poboljšati koristeći transformacije ulaznih podataka (npr. kvadriranje ulaznih vrijednosti nekog parametra), kao i koristeći interakcijske članove (npr. umnožak vrijednosti dva parametra).

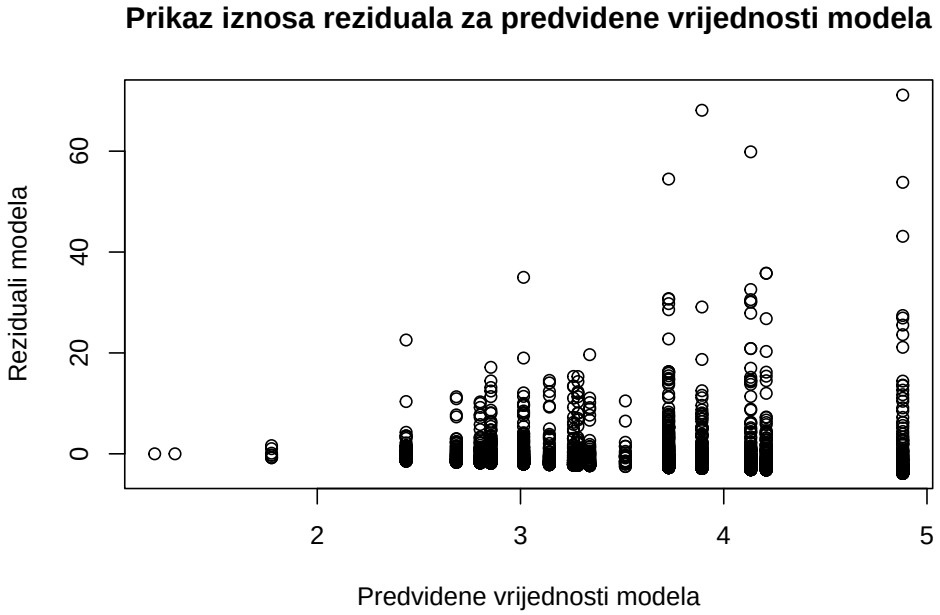
4. zadatak

Kako bismo mogli odabrati karijeru po kriteriju da se obogatimo, provest ćemo linearnu regresiju s regresorom wealth.how.industry i zavisnom varijablom wealth.worth.in.billions. Pritom ćemo kao dataset koristiti podatke koje smo pripremili za provođenje linearne regresije čišćenjem stringova i zamjenom praznih stringova, negativnih vrijednosti i vrijednosti nula vrijednošću NA.

```
##
## Call:
## lm(formula = wealth.worth.in.billions ~ wealth.how.industry,
##     data = billionaires.unique)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.881 -2.083 -1.293  0.046 71.119
##
## Coefficients:
##                                                 Estimate Std. Error t value
## (Intercept)                                        1.300      5.078   0.256
## wealth.how.industryconstrution                     1.137      5.104   0.223
## wealth.how.industryconsumer                        2.429      5.083   0.478
## wealth.how.industrydiversified financial           2.909      5.093   0.571
## wealth.how.industryenergy                          1.983      5.097   0.389
## wealth.how.industryhedge funds                     2.040      5.116   0.399
## wealth.how.industrymedia                           2.593      5.089   0.509
## wealth.how.industrymining and metals               1.842      5.106   0.361
## wealth.how.industrymoney management                1.554      5.088   0.305
## wealth.how.industrynon-consumer industrial         1.961      5.102   0.384
## wealth.how.industryother                           1.384      5.108   0.271
## wealth.how.industryprivate equity/leveraged buyout 2.216      5.178   0.428
## wealth.how.industryreal estate                     1.715      5.087   0.337
## wealth.how.industryretail, restaurant              2.833      5.087   0.557
## wealth.how.industryservices                       -0.100      7.181  -0.014
## wealth.how.industrytechnology-computer             3.581      5.090   0.703
## wealth.how.industrytechnology-medical              1.502      5.101   0.294
## wealth.how.industryventure capital                 0.475      5.386   0.088
##                                                 Pr(>|t|)
## (Intercept)                                        0.798
## wealth.how.industryconstrution                     0.824
## wealth.how.industryconsumer                        0.633
## wealth.how.industrydiversified financial           0.568
## wealth.how.industryenergy                          0.697
## wealth.how.industryhedge funds                     0.690
## wealth.how.industrymedia                           0.610
## wealth.how.industrymining and metals               0.718
## wealth.how.industrymoney management                0.760
## wealth.how.industrynon-consumer industrial         0.701
## wealth.how.industryother                           0.786
## wealth.how.industryprivate equity/leveraged buyout 0.669
## wealth.how.industryreal estate                     0.736
## wealth.how.industryretail, restaurant              0.578
## wealth.how.industryservices                        0.989
## wealth.how.industrytechnology-computer             0.482
## wealth.how.industrytechnology-medical              0.768
## wealth.how.industryventure capital                 0.930
##
## Residual standard error: 5.078 on 2579 degrees of freedom
##   (17 observations deleted due to missingness)
## Multiple R-squared:  0.01619,    Adjusted R-squared:  0.009705
## F-statistic: 2.497 on 17 and 2579 DF,  p-value: 0.0006239
```

Provjerit ćemo izgled reziduala za dobiveni model na isti način kao i u prethodnom zadatku.



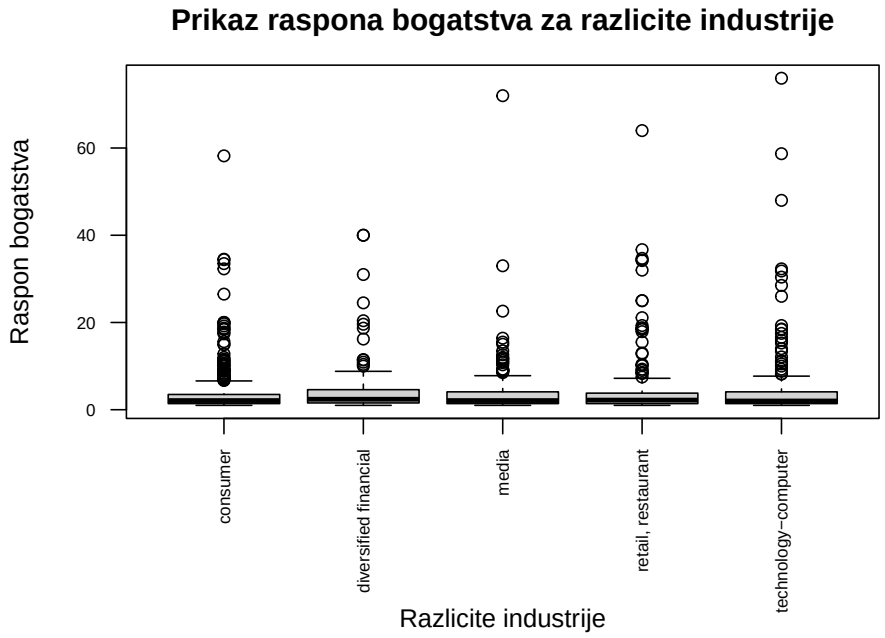


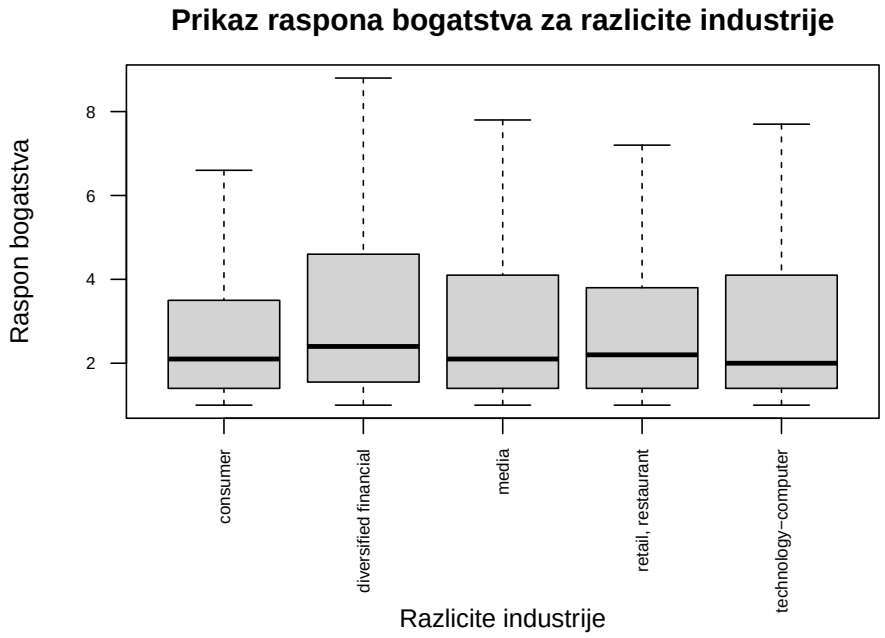

```
## Lilliefors test za normalnost standardiziranih reziduala
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rstandard(fit.industry)
## D = 0.25569, p-value < 2.2e-16
```

Vidimo da kao niti u prethodnome zadatku, reziduali nisu normalno distribuirani. Uz to, vidljiva je i razlika u varijabilnosti reziduala ovisno o konkretnoj industriji. Zbog nezadovoljenih pretpostavki o rezidualima, moramo razmišljati o tome da dodatni testovi provedeni koristeći ovaj model nisu nužno vjerodostojni.

Kako bismo odabrali industriju kojom bi se bavili s ciljem da se obogatimo, proučavat ćemo konkretne industrije koje su u regresiji bile što značajnije (što manji iznos p-vrijednosti t-testa s "H0: regresijski koeficijent za određenu industriju je jednak 0"), a istovremeno imaju i visoki iznos procjene regresijskog koeficijenta. Konkretno, to su sljedeće industrije: "technology-computer", "retail, restaurant", "diversified financial", "media" te "consumer" . Za iste ćemo prikazati raspon bogatstva koristeći boxplot.





Na temelju prikazanih boxplotova, mogli bismo zaključiti da je medijan za ovih 5 industrija otprilike jednak. Kako bismo provjerili ovu tvrdnju, provest ćemo testiranje jednakosti sredina jednofaktorskom analizom varijance pri čemu će faktor biti parametar wealth.how.industry, ali samo s 5 razina - 5 odabranih industrija. Da bismo mogli provoditi Anovu, trebale bi vrijediti već navedene pretpostavke: populacije trebaju biti nezavisne i imati normalnu distribuciju, a varijance za svaku grupu trebale bi biti homogene. Ako ove pretpostavke nisu zadovoljene, test gubi na vjerodostojnosti - u tom slučaju moguće je koristiti i neparametarsku verziju testa, Kruskal-Wallisov test.

Kako bismo testirali normalnost podataka, provest ćemo Lillieforsovu inačicu KS testa za sve grupe: u slučaju da je p-vrijednost manja od 0.05, odbacujemo nultu hipotezu da podaci potječu iz normalne distribucije.

Kako bismo testirali homogenost varijance, provest ćemo Bartlettov test: u slučaju da je p-vrijednost manja od 0.05, odbacujemo nultu hipotezu da su varijance svih grupa podataka jednake.

```
## Lilliefors test za normalnost podataka iz industrije: technology-computer

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  billionaires.best.industry$wealth.worth.in.billions[billionaires.best.industry$wealth.how.indu
## D = 0.32934, p-value < 2.2e-16

## Lilliefors test za normalnost podataka iz industrije: retail, restaurant

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  billionaires.best.industry$wealth.worth.in.billions[billionaires.best.industry$wealth.how.indu
## D = 0.31793, p-value < 2.2e-16

## Lilliefors test za normalnost podataka iz industrije: diversified financial

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  billionaires.best.industry$wealth.worth.in.billions[billionaires.best.industry$wealth.how.indu
## D = 0.2902, p-value < 2.2e-16

## Lilliefors test za normalnost podataka iz industrije: media

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  billionaires.best.industry$wealth.worth.in.billions[billionaires.best.industry$wealth.how.indu
## D = 0.31561, p-value < 2.2e-16

## Lilliefors test za normalnost podataka iz industrije: consumer

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  billionaires.best.industry$wealth.worth.in.billions[billionaires.best.industry$wealth.how.indu
## D = 0.29757, p-value < 2.2e-16

## Bartlett test za provjeru jednakosti varijanci

##
##  Bartlett test of homogeneity of variances
##
## data:  billionaires.best.industry$wealth.worth.in.billions by billionaires.best.industry$wealth.how.i
## Bartlett's K-squared = 96.271, df = 4, p-value < 2.2e-16
```

Rezultat provođenja Lillieforsove inačice KS testa za svih 5 razina daje veoma sličan rezultat - p-vrijednost je za sve razine značajno manja od 0.05. Na razini značajnosti 0.05, odbacujemo nultu hipotezu Lillieforsove inačice KS testa - ne možemo pretpostaviti da podaci dolaze iz normalne distribucije.

Isto tako, odbacujemo i nultu hipotezu Bartlettovog testa - ne možemo pretpostaviti jednakost varijanci. Zbog navedenih kršenja pretpostavke, provest ćemo Kruskal-Wallisov test s nultom hipotezom jednakosti medijana za svih 5 uzoraka (pritom pazeći na uvjet veličine uzoraka od barem 5). Kako bismo rezultat provođenja Kruskal-Wallisovog testa mogli usporediti, provest ćemo i jednofaktorsku Anovu, pritom se oslanjajući na robustnost testa.

```
## Frekvencije po industrijama:  471 167 219 281 208
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  wealth.worth.in.billions by wealth.how.industry
## Kruskal-Wallis chi-squared = 1.8427, df = 4, p-value = 0.7647
```

```
## ANOVA
```

```
##                        Df Sum Sq Mean Sq F value Pr(>F)
## wealth.how.industry     4    203   50.67   1.255  0.286
## Residuals            1341  54138   40.37
```

Na razini značajnosti 0.05, ne možemo odbaciti nultu hipotezu Kruskal-Wallisovog testa, kao ni nultu hipotezu testa jednakosti sredina jednofaktorskom analizom varijance. Drugim riječima, zaključujemo da su vrlo vjerojatno medijani iznosa bogatstva odabranih 5 industrija otprilike jednaki. Zbog toga, svejedno je koju ćemo od navedenih industrija odabrati, ali je bitno da odaberemo jednu od tih 5 industrija, a ne neku od ostalih. Kako bismo to dokazali, za kraj ćemo provesti i Kruskal-Wallisov test za sve industrije.

```
## Frekvencije pojedinih industrija:  1 97 471 167 132 67 219 90 249 107 83 25 280 281 1 208 111 8
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  wealth.worth.in.billions by wealth.how.industry
## Kruskal-Wallis chi-squared = 39.908, df = 17, p-value = 0.001333
```

Na razini značajnosti 0.05, odbacujemo nultu hipotezu Kruskal-Wallisovog testa - ne možemo zaključiti da je medijan iznosa bogatstva jednak za sve industrije. U obzir treba uzeti i činjenicu da dvije grupe imaju frekvenciju manju od 5 - takve grupe bi se mogle izbaciti kako bi test bio pouzdaniji, no zbog sažetosti prikaza provođenje testa nad takvoj, pročišćenoj tablici, nije prikazan (p-vrijednost ispada neznatno drugačija: 0.001048 pa stoga zaključak nakon provođenja testa ostaje isti).

Oslanjajući se na iznos regresijskog koeficijenta za industriju "technology-computer", p-vrijednost provođenja t-testa, kao i prikaz na boxplotu, odabrali bismo karijeru u industriji "technology-computer". Kao drugi izbor na temelju regresijskog koeficijenta, p-vrijednosti provođenja t-testa i prikaza na boxplotu, odabrali bismo karijeru u industriji "diversified financial".