

## R – OBUKA MEŠOVITI MODELI

18-20.04.2019.

Filozofski fakultet, Univerzitet u Banja Luci

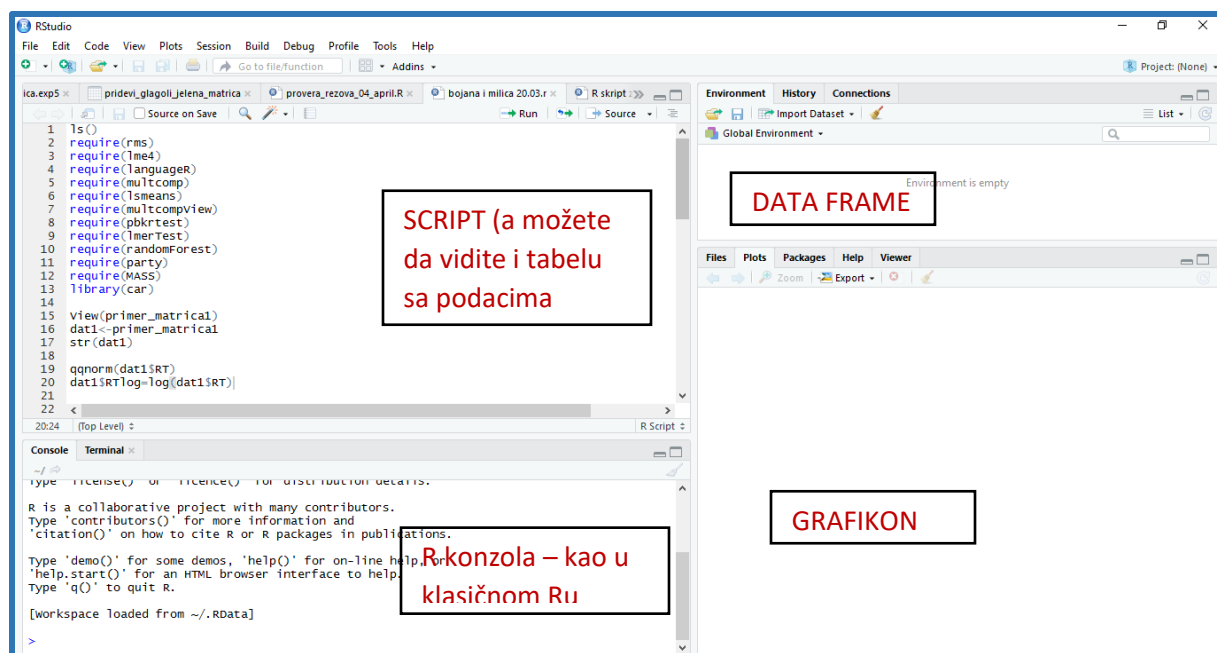
Milica Popović Stijačić

Bojana Dinić

### **I UPOZNAVANJE SA R-om<sup>1</sup>- KRATKI KRATKI (kratki) INTRO u R**

<sup>1</sup>Ko zna ovaj basic, samo neka pređe na sledeće poglavlje ☺

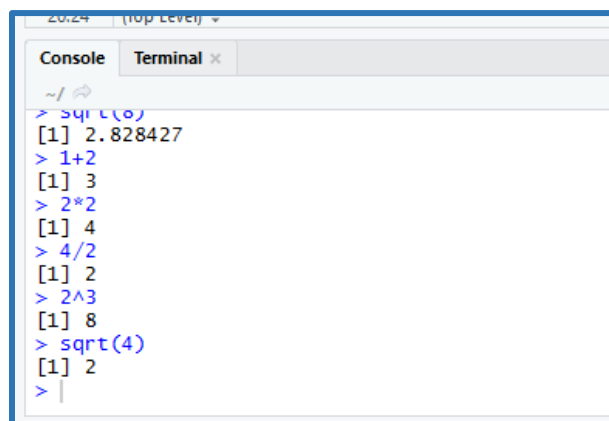
R programu možete pristupiti preko klasične r konzole, ili preko nešto više „user friendly“ varijante – R studija (R studio). Na slici 1 prikazan je „print screen“ R studija. Može se videti da postoje četiri kvadranta, dva gornja i dva donja. U gornjem levom kvadrantu možete da pišete svoj skript, koji sačuvate i kasnije pokrenete, i koji će imati ekstenziju „.Rscript“. U ovom kvadrantu, takođe, možete videti i svoje podatke, tako što kliknete na odgovarajući jezičak sa podacima (kao kad prelazite iz jednog sheet-a u drugi, u ekselu). Opet, klikom na odgovarajući jezičak, vraćate se na svoj skript. U gornjem desnom kvadrantu nalazi se tzv. „environment“, i tu će vam se prikazati deo sa vašim dataframe-om., koji je sada na slici 1 prazan. Takođe, u tom kvadrantu, klikom na „import dataset“ birate vaše fajlove sa podacima. U donjem levom kvadrantu se nalazi R konzola, i to je praktično klasičan R; drugim rečima, klasičan R program ima samo konzolu, gde ukucavate komandu i stisnete „ENTER“, a ako biste hteli da čuvate history i skript, to biste posebno pravili u nekom notepad-u. U R studiju, najbolja strategija je da komande kucate u skriptu (gornji levi kvadrant), a zatim, da biste izvršili komandu, selektujete šta želite da se izvrši i kliknete „Ctrl+ENTER“. Ovo je zgodno, pogotovo zato što u skriptu možete da pišete komentare, tako što pre komentara ukucate „#“ (tarabu) pa upišete komentar (na primer: #Pocinjem da ucim R ☺). Na ovaj način, R-u saopštavate da preskoči sve ono što počinje sa tarabom. I, na kraju, u donjem desnom kvadrantu nalazi se deo ekrana u kom će vam se iscrtavati grafikoni, koje posle možete sačuvati u jpeg ili png formatu. Tople preporuke za upotrebu R studija ☺.



**Slika 1.** Print screen interfejsa od R studija

## 1) R kao kalkulator – kratki primer

Samo ukucate račun koji želite u konzoli i pritisnete ENTER (slika 2).

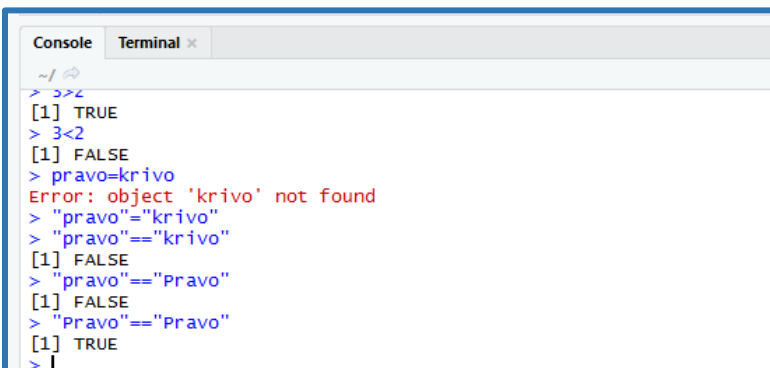


**Slika 2.** Print screen R konzole pri primeni osnovnih računskih operacija

## 2) Logički operatori -> TRUE & FALSE

Na slici 3 prikazana je logička provera istinitosti tvrdnji. Ono što je bitno, to je da samo jedan znak „=“ ne znači jednako, već da biste proverili jednakost, morate da koristite dva znaka „==“. Dodatno, za stringove, to jest reci, morate koristiti navodnike, nebitno da li jednostruke ili

dvostruke. Pored toga, vodite računa o malim i velikim slovima, jer ih R razlikuje, pa „pravo“ nije isto što i „Pravo“ ☺

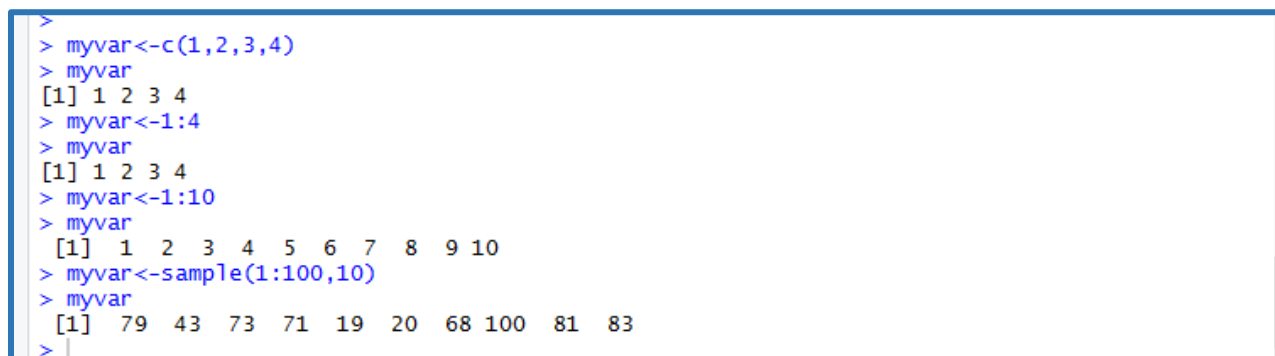


```
Console Terminal x
~/
> >>
[1] TRUE
> 3<2
[1] FALSE
> pravo=krivo
Error: object 'krivo' not found
> "pravo"="krivo"
> "pravo"=="krivo"
[1] FALSE
> "pravo"=="Pravo"
[1] FALSE
> "Pravo"=="Pravo"
[1] TRUE
>
```

**Slika 3.** Provera logičnosti iskaza – mala velika slova, znak jednakosti

### 3) Kreiranje varijabli i objekata

U R u možete da sami pravite svoje data frame ove, i evo nekoliko načina da generišete varijable. Na slici 4 dato je nekoliko različitih načina da generišete varijable koje sadrže cele brojeve – „integers“. U poslednjem redu na slici 4, pomoću funkcije `sample`, možete slučajno generisati određen skup brojeva iz zadatog opsega.

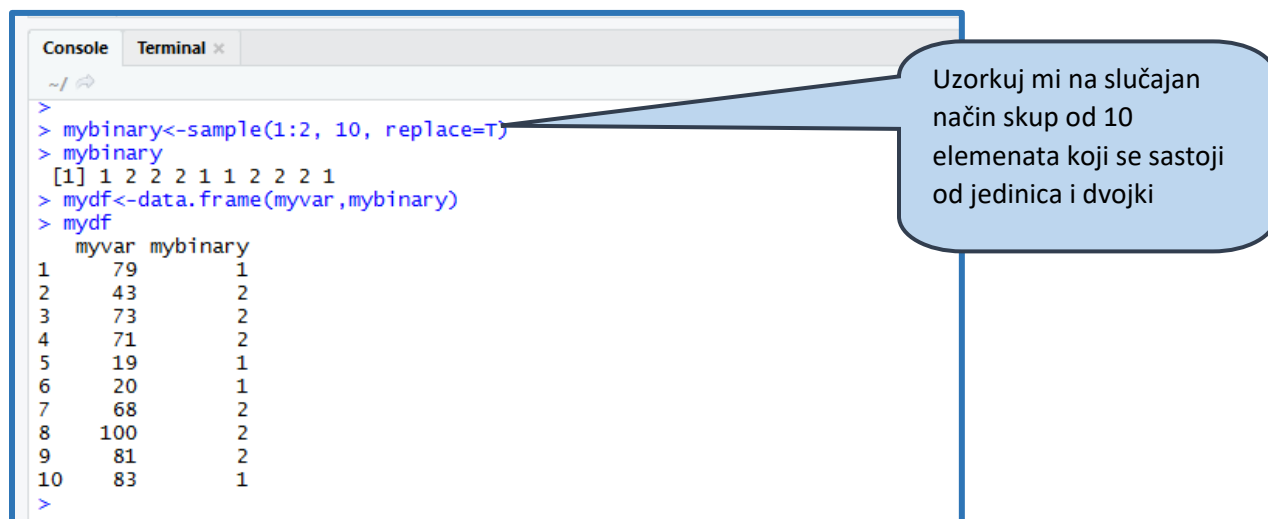


```
>
> myvar<-c(1,2,3,4)
> myvar
[1] 1 2 3 4
> myvar<-1:4
> myvar
[1] 1 2 3 4
> myvar<-1:10
> myvar
[1] 1 2 3 4 5 6 7 8 9 10
> myvar<-sample(1:100,10)
> myvar
[1] 79 43 73 71 19 20 68 100 81 83
>
```

**Slika 4.** Dodeljivanje brojeva varijabli – nekoliko načina

Možete generisati i binarnu varijablu (slika 5), pomoću iste funkcije `sample`, samo što mu još u zagradi dodate „`replace=T`“, što znači da se uzorkovanje radi sa vraćanjem, što je u ovom slučaju obavezno naglasiti.

Na kraju, napravite svoj prvi „data frame“ (slika 5), tako što ćete da spojite *myvar* i *mybinary*, pomoću funkcije „data.frame“. Pomoću funkcije „colnames“ možete da promenite nazive svojih varijabli, u, na primer „tezina“ i „pol“...



```
> mybinary<-sample(1:2, 10, replace=T)
> mybinary
[1] 1 2 2 2 1 1 2 2 2 1
> mydf<-data.frame(myvar,mybinary)
> mydf
  myvar mybinary
1    79         1
2    43         2
3    73         2
4    71         2
5    19         1
6    20         1
7    68         2
8   100         2
9    81         2
10   83         1
```

Uzorkuj mi na slučajan način skup od 10 elemenata koji se sastoji od jedinica i dvojki

Slika 5. Pravljenje data frame-a

```
> colnames(mydf)<-c("tezina","pol")
> mydf
  tezina pol
1     79   1
2     43   2
3     73   2
4     71   2
5     19   1
6     20   1
7     68   2
8    100   2
9     81   2
10    83   1
```

Slika 6. Dodeljivanje imena kolonama

#### 4) Tipovi podataka

R u osnovi razlikuje više tipova podataka, ali najosnovniji i najbitniji za osnovnu statistiku su (Slika 7):

- Integer (ceo broj: 1, 2, 3, 4)
- Numeric (realan broj: 2.3, 3.6...)
- Factor
- Character („musko“, „zensko“, „Zelim da naucim mesovite modele“)

- Logical (Boolean, TRUE, FALSE, NA)

```
> class(2)
[1] "numeric"
> class(4.55)
[1] "numeric"
> class(2L)
[1] "integer"
> class("musko")
[1] "character"
> class(TRUE)
[1] "logical"
>
>
> str(mydf)
'data.frame': 10 obs. of 2 variables:
 $ tezina: int 79 43 73 71 19 20 68 100 81 83
 $ pol : int 1 2 2 2 1 1 2 2 2 1
```

**Slika 7.** Ispitivanje vrste podataka pomoću funkcije „class“ i „str“

Ukoliko želite da proverite koji tip podataka je neka varijabla, koristite funkciju *class*, odnosno *str* za čitav data frame. U ovom našem primeru, naš data frame *mydf* ima dve varijable, *tezina* i *pol* i R ih je obe kodirao kao *int*edžere. U stvari, *pol* bi trebalo da se tretira kao faktor u kasnijim analizama. Pomoću funkcije *as.factor* ili *as.numeric*, ili *as.character* možemo da menjamo vrstu podataka u matrici, to jest data frame-u (slika 8).

```
> mydf$pol<-as.factor(mydf$pol)
> class(mydf$pol)
[1] "factor"
>
```

**Slika 8.** Promena vrste podataka pomoću funkcije *as.factor*

## 5) Osnovne funkcije za pokazatelje deskriptivne statistike

U R u postoje generičke funkcije za ispitivanje osnovnih deskriptivnih pokazatelja (slika 9).

```
>
>
> mean(tezina)
[1] 56.4
> median(tezina)
[1] 59.5
> sd(tezina)
[1] 27.18333
> hist(tezina)
```

**Slika 9.** Računanje osnovnih deskriptivnih pokazatelja

A iste te pokazatelje pomoću funkcija *tapply* možemo prikazati po nivoima faktora – slika 10. A možemo ukombinovati i *with*, pomoću kojeg kažemo da nam iz određenog data frame-a izračuna pomoću *tapply*, aritmetičku sredinu, ili šta već želimo.

```
>
> tapply(mydf$tezina, mydf$pol, mean)
 1      2
55.50 57.75
>
> with(mydf, tapply(tezina, pol, mean))
 1      2
55.50 57.75
```

**Slika 10.** Pomoću funkcije *tapply*, možemo prikazati deskriptivne pokazatelje po nivoima faktora

I na kraju, možemo izračunati t test ☺ (slika 11), pomoću, gle čuda, funkcije *t.test*.

```
> t.test(tezina~pol, data=mydf)

welch Two Sample t-test

data:  tezina by pol
t = -0.11525, df = 5.5253, p-value = 0.9123
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -51.03154  46.53154
sample estimates:
mean in group 1 mean in group 2
      55.50      57.75
```

**Slika 11.** Računanje t – testa

Data frame možemo sačuvati na disku (van R-a) na više načina – slika 12:

```
write.table(mydf, file = "Podaci_vezba1.txt", append = FALSE, sep = "\t", eol=
"\n", col.names=TRUE, row.names=FALSE)

ili

write.csv(mydf, "podaci vezba1.csv")
```

**Slika 12.** Upisivanje data frame-a u .txt i .csv fajl

## 6) Učitavanje podataka

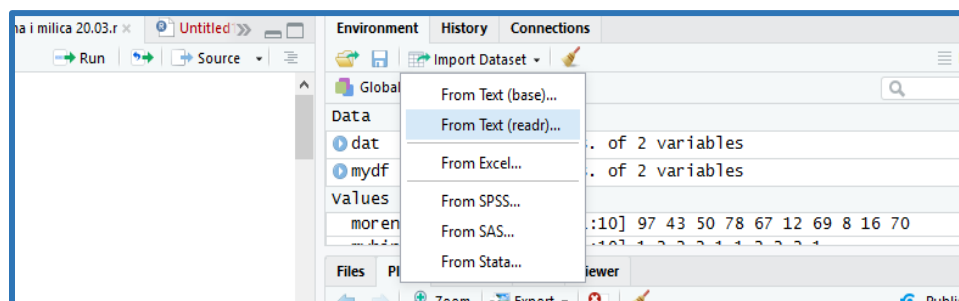
U realnosti, retko kad imamo potrebu da pravimo data frame u R-u, već učitavamo sopstvene realne podatke, smeštene u .txt, .csv ili .xls formatu. Ranije, dok se češće koristila R

konsola, podaci su se učitavali pomoću komande `read.table`, a još bolje je da se odmah dodeli ime tom data frame-u, poput „dat“, „df“, „mydata“ i sl.

```
> read.table("podaci_vezba.txt", sep="\t", header=T)
  tezina pol
1     79   2
2     86   2
3     58   1
4     67   1
5     14   2
6     38   1
7     16   1
8     61   1
9     52   2
10    93   1
> dat<-read.table("podaci_vezba.txt", sep="\t", header=T)
> |
```

**Slika 13.** Učitavanje matrice podataka preko komande `read.table`.

Druga verzija je još lakša, jer se ide preko, na početku spomenutog, gornjeg desnog kvadranta, gdje se u jeziku „environment“ ide na „import data set“, a zatim se odabere format u kojem je fajl sa podacima, koji se potom pojavljuju u gornjem levom kvadrantu ako su lepo učitani (slika 14).



**Slika 14.** Učitavanje podataka preko R studia

## II UVOD U MEŠOVITE LINEARNE MODELE – KRATKI, KRATKI (kratki) INTRO

### 1) Mali osvrt na klasične linearne modele

Osnovna funkcija linearnih modela je sledeća:

$$Y = a + bX + \varepsilon,$$

pri čemu je  $Y$  naša zavisna varijabla koju predviđamo,  $a$  je intercept, ili odsečak na y osi,  $b$  je nagib regresione prave,  $X$  je naša nezavisna varijabla, to jest prediktor, a  $\varepsilon$  je greška merenja, šum u podacima, ono što nismo sistematski varirali, to jest reziduali. Prikazani linearni model je sa jednim prediktorom, ali ako ih imamo više, kao što je najčešće slučaj, onda model zapisujemo ovako:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_i X_i + \varepsilon.$$

U ovom drugom modelu, umesto  $a$  imamo  $\beta_0$ , a umesto  $b$  imamo  $\beta_i$ , to jest onoliko koeficijenata za nagib, koliko imamo prediktora  $X_i$ , i na kraju, opet imamo rezidualne, ili grešku, to jest, neobjašnjeni deo modela  $\varepsilon$ . Drugi model predstavlja, u stvari, model višestruke linearne regresije.

U terminima mešovityh efekata,  $\beta_i$  koeficijenti predstavljaju FIKSNE (fixed) efekte, a  $\varepsilon$  predstavlja SLUČAJNE (random) efekte. U klasičnim linearnim modelima, deo koji se odnosi na slučajne efekte ne procenjujemo, on prosto „visi“, i jedino što možemo da uradimo jeste da vidimo kako se reziduali distribuiraju, kako bismo proverili da li su zadovoljene osnovne pretpostavke linearnih modela:

1. Linearan odnos između ZV i NV
2. Odsustvo multikolinearnosti – kod višestruke regresije, što se proverava ili preko jednostavnih korelacija svakog prediktora sa svakim, ili preko VIF-a (*variance inflation factor*, koji treba da je ispod 8, dok vrednost preko 10 govori o ozbiljnoj multikolinearnosti).
3. Normalna distribucija reziduala – drugim rečima, greške se distribuiraju normalno, to znači da su slučajne i da nisu posledica neke sistematske varijacije. Ovo se proverava



preko QQ plota ili testiranjem normalnosti distribucije rezidula pomoću Kolmogorov-Smirnov testa.

4. Odsustvo uticajnih tačaka: autlejeri (štrčci) i ekstremne vrednosti
5. Homoskedastičnost varijanse – govori o jednakosti varijanse podataka duž fitovanih vrednosti (procenjenih vrednosti). Za proveru stavljamo u odnos fitovane vrednosti i rezidualne, skater dijagram treba da prikaže „jaje“, a ne „peščani sat“, „trougao“ i slično ☺. Ako postoji, trebalo bi primeniti ili kvadratnu transformaciju ili primeniti neki model iz porodice Generalizovanih linearnih modela (logistička regresija, mešoviti logit modeli)
6. Odsustvo autokorelacija greški – nekorelisanost reziduala, to jest, greške treba da su nezavisne: Durbin Watson statistik (traba da je između 1.5 i 2.5, a inače se kreće od 0-4, vrednosti bliske krajnjim granicama govore o korelisanosti grešaka)

Naravno, poenta klasičnih linearnih modela je da se pomoću što manje prediktora „pokupe“ to jest sa što manje manipulacija objasni fenomen, a da pri tom, neobjašnjen deo bude što manji. Inače, matematički je nemoguće odsustvo greške – ona potiče makar od nesavršenost mernog instrumenta ☺, što je u društvenim naukama standardan slučaj. I evo, tu dolazimo do veze između linearnih modela i ANOVE, to jest analize varijanse, gde se pomoću F testa stavljaju u odnos objašnjena i neobjašnjena varijansa (to jest varijansa reziduala). U tom smislu, samo malo podsećanje, da je F test uvek pozitivan i da, što je veći, do je bolji, jer to znači da je varijansa koja je objašnjena nezavisnim varijablama veća od varijanse reziduala, to jest greške.

Ovo je bio kratki, kratki osvrt na klasične linearne modele, neophodan za objašnjenje mešoviti modela.

\*\*\*\*\* Vežba 1 (pogledati tutorijal 1 od Bodo Wintera (2013))\*\*\*\*\*

Zamislamo da želimo da ispitamo kako se muškarci i žene razlikuju po visini glasa. Treba nam jedna varijabla sa prosečnom frekvencijom glasova žena i jedna varijabla koja se odnosi na pol.

```
#Pravljenje data frame
```

```
pol <- c(rep("musko",3),rep("zensko",3))
```

```
#sa rep, kažemo R-u da 3 puta ponovi musko, to jest 3 puta  
zensko
```

```
vis_glasa <- c(233,204,242,130,112,142)
```

```
my.df <- data.frame(pol, vis_glasa)#napravi mi data frame
```

```
#Ajde da vidimo kako izgleda naš data frame
```

```
my.df
```

|   | pol    | vis_glasa |
|---|--------|-----------|
| 1 | musko  | 233       |
| 2 | musko  | 204       |
| 3 | musko  | 242       |
| 4 | zensko | 130       |
| 5 | zensko | 112       |
| 6 | zensko | 142       |

```
# Sada ćemo da napravimo prvi linearan model
```

```
lm1<- lm(vis_glasa ~ pol, data = my.df)
```

```
summary(lm1)
```

Call:

```
lm(formula = vis_glasa ~ pol, data = my.df)
```

Residuals:

|  | 1     | 2       | 3      | 4     | 5       | 6      |
|--|-------|---------|--------|-------|---------|--------|
|  | 6.667 | -22.333 | 15.667 | 2.000 | -16.000 | 14.000 |

**Coefficients:**

|             | Estimate | Std. Error | t value | Pr(> t )     |
|-------------|----------|------------|---------|--------------|
| (Intercept) | 226.33   | 10.18      | 22.224  | 2.43e-05 *** |
| polzensko   | -98.33   | 14.40      | -6.827  | 0.00241 **   |

---

**Signif. codes:** 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**Residual standard error:** 17.64 on 4 degrees of freedom

**Multiple R-squared:** 0.921, **Adjusted R-squared:** 0.9012

**F-statistic:** 46.61 on 1 and 4 DF, **p-value:** 0.002407

# Dakle, model je značajan, i vidimo standardne pokazatelje, kao što su  $R^2$ ,  $\beta$  koeficijente, i njihovu značajnost, te F test čitavog modela i procenu značajnosti modela na osnovu F testa.

# Kada bi bilo više prediktora u modelu, p vrednosti koeficijenata i p vrednost F testa bi se razlikovala, jer F test procenjuje značajnost modela sa svim fiksnim efektima, a p vrednost koeficijenta procenjuje značajnost samo za jedan prediktor.

Na ovom mestu dolazimo do prve bitne stavke u interpretaciji koeficijenata fiksnih efekata, bilo da su u pitanju standardni linearni ili mešoviti linearni modeli:

1) **INTERPRETACIJA KOEFICIJENATA FIKSNIH EFEKATA – PRAVILO 1**

Kada imamo KATEGORIJALNI PREDIKTOR, onda intercept predstavlja aritmetičku sredinu referentne kategorije, u ovom slučaju to je „musko“, jer R automatski ređa kategorije uzlazno po brojevima ili po abecedi.

Ovo se može proveriti:

```
with (my.df, tapply (vis_glasa, pol, mean))
```

```
musko    zensko
```

```
226.3333 128.0000
```

Kao što se može videti, intercept iz našeg modela ima istu vrednost kao i aritmetička sredina visine glasa za muski pol.

Šta je onda drugi koeficijent?

„Polzensko“ koeficijent je negativan broj, i ako se oduzme od intercepta, dobija se 128 što je aritmetička sredina za zenski pol. Drugim rečima, nagib u jednostavnoj regresiji govori o tome za koliko se promeni visina glasa (to jest zavisna varijabla) kada se nezavisna pomeri za jednu jedinicu na x osi. U slučaju sa kategorijalnim prediktorom, to znači, za koliko se promeni zavisna varijabla kada se pređe sa referentne na sledeću kategoriju. Konkretno, u ovom primeru, koeficijent polmusko je ocena razlike u visini glasa između muškog i ženskog pola.

### \*\*\*\*Vežba 2 – slučaj jednostavnog LM sa kontinuiranim prediktorom\*\*\*\*

Zamislimo da želimo da ispitamo kako se visina glasa menja sa porastom godina. Možemo da modifikujemo data frame i dodamo varijablu godine.

```
#Pravljenje data framea
```

```
godine <- c(14,23,35,48,52,67)
```

```
vis_glasa <- c(252,244,240,233,212,204)
```

```
my.df <- data.frame(godine, vis_glasa) #napravi mi data frame
```

```
#Ajde da vidimo kako izgleda naš data frame
```

```
my.df
```

|   | godine | vis_glasa |
|---|--------|-----------|
| 1 | 14     | 252       |
| 2 | 23     | 244       |
| 3 | 35     | 240       |

|   |    |     |
|---|----|-----|
| 4 | 48 | 233 |
| 5 | 52 | 212 |
| 6 | 67 | 204 |

# Sada ćemo da napravimo drugi linearni model

```
lm2<- lm(vis_glasa ~ godine, data = my.df)
```

```
summary(lm2)
```

Call:

```
lm(formula = vis_glasa ~ godine, data = my.df)
```

Residuals:

|  | 1      | 2      | 3     | 4     | 5      | 6      |
|--|--------|--------|-------|-------|--------|--------|
|  | -2.338 | -2.149 | 4.769 | 9.597 | -7.763 | -2.115 |

Coefficients:

|             | Estimate | Std. Error | t value | Pr(> t )     |
|-------------|----------|------------|---------|--------------|
| (Intercept) | 267.0765 | 6.8522     | 38.98   | 2.59e-06 *** |
| godine      | -0.9099  | 0.1569     | -5.80   | 0.00439 **   |

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.886 on 4 degrees of freedom

Multiple R-squared: 0.8937, Adjusted R-squared: 0.8672

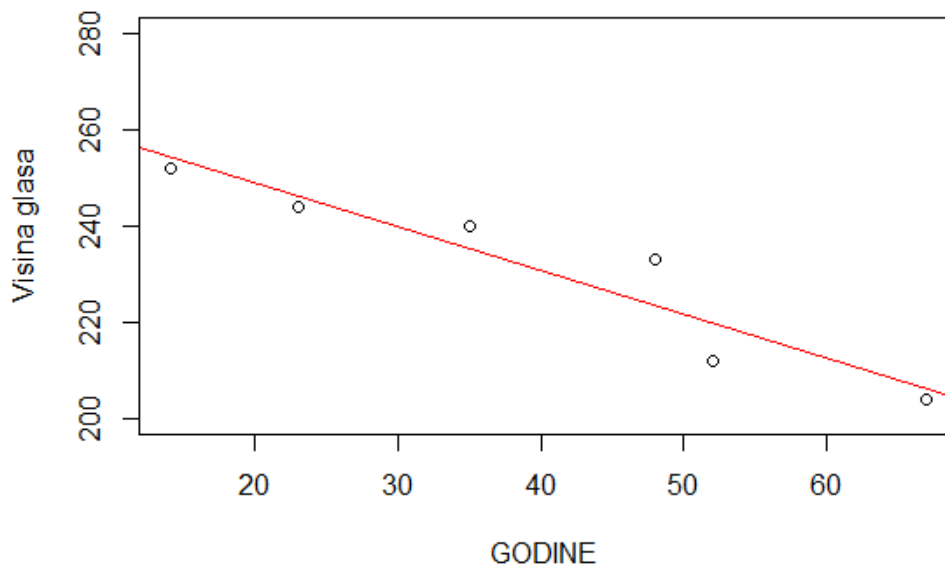
F-statistic: 33.64 on 1 and 4 DF, p-value: 0.004395

## 2) INTERPRETACIJA KOEFICIJENATA FIKSNIH EFEKATA – PRAVILO 2

Kada imamo KONTINUIRANI PREDIKTOR, onda intercept predstavlja aritmetičku sredinu kriterijuma kad je prediktor 0, u ovom slučaju to znači kolika bi bila visina glasa nerođenog deteta (pri 0 god), što nema mnogo smisla.

Možemo grafički prikazati ovu regresiju (slika 15):

```
plot (my.df$godine, my.df$vis_glasa, xlab="GODINE", ylab="Visina glasa", ylim= c (200, 280), abline (lm (vis_glasa ~ godine), col="red"))
```



**Slika15. Regresioni model godina i visine glasa**

Grafikon smo iscrtali kako bismo imali bolji uvid u interpretaciju regresionih koeficijenata.

Šta znači nagib u slučaju KONTINUIRANOG prediktora? Iz ispisa modela, može se videti da je koeficijent kojim se ocenjuje uticaj godina na variranje visine glasa značajan i iznosi -0.9099, što znači da se za svaku godinu, frekvencija glasa snižava za 0.90 Hz.

### 3) INTERPRETACIJA KOEFICIJENATA FIKSNIH EFEKATA – PRAVILO 3

Da bi intercept mogao smisljeno da se interpretira, varijable je potrebno centrirati, to jest, od svakog podatka jedne varijable oduzme se njegova aritmetička sredina.

```
#centriranje podataka
```

```
my.df$godine.c = my.df$godine-mean(my.df$godine)
```

```
my.df #proverimo kakav nam je data frame
```

|   | godine | vis_glasa | godine_c   |
|---|--------|-----------|------------|
| 1 | 14     | 252       | -25.833333 |
| 2 | 23     | 244       | -16.833333 |
| 3 | 35     | 240       | -4.833333  |
| 4 | 48     | 233       | 8.166667   |
| 5 | 52     | 212       | 12.166667  |
| 6 | 67     | 204       | 27.166667  |

```
#ponovo pokrenemo model
```

```
lm2 = lm(vis_glasa ~ godine_c, my.df)
```

```
summary(lm2)
```

Call:

```
lm(formula = vis_glasa ~ godine_c, data = my.df)
```

Residuals:

| 1      | 2      | 3     | 4     | 5      | 6      |
|--------|--------|-------|-------|--------|--------|
| -2.338 | -2.149 | 4.769 | 9.597 | -7.763 | -2.115 |

Coefficients:

|             | Estimate | Std. Error | t value | Pr(> t )     |
|-------------|----------|------------|---------|--------------|
| (Intercept) | 230.8333 | 2.8113     | 82.11   | 1.32e-07 *** |
| godine_c    | -0.9099  | 0.1569     | -5.80   | 0.00439 **   |

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.886 on 4 degrees of freedom

Multiple R-squared: 0.8937, Adjusted R-squared: 0.8672

F-statistic: 33.64 on 1 and 4 DF, p-value: 0.004395

Može se videti na osnovu nove vrednosti intercepta, a i na osnovu grafika da sada intercept predstavlja vrednost Y za prosečnu vrednost Xa, konkretno, on sada iznosi 230, što jeste vrednost Y varijable kada X ima prosečnu vrednost.

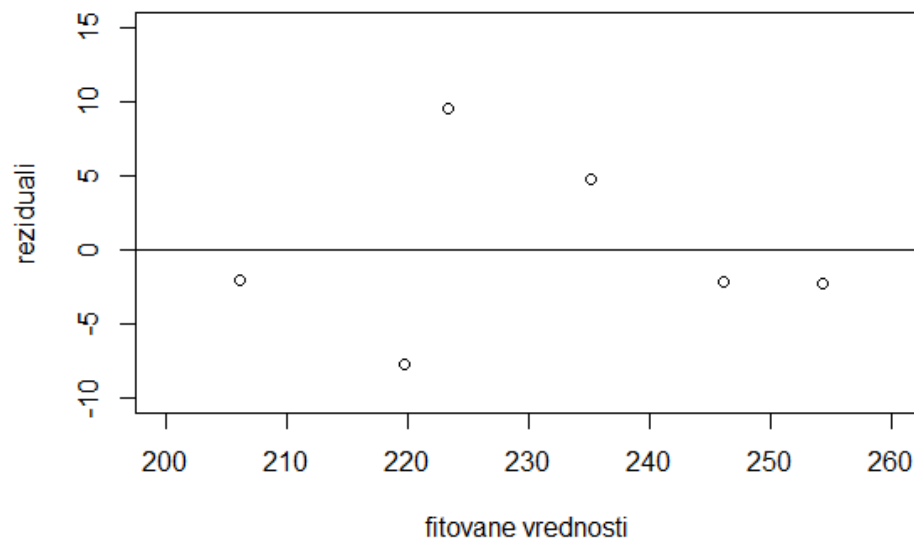
### \*\*\*\*Vežba 3 – provera klasičnih pretpostavki linearnih modela\*\*\*\*

#### # Provera linearnosti ZV i prediktora

Za proveru linearnosti model potrebno je napraviti „residual plot“ (slika 16), gde zapravo stavljamo u odnos fitovane vrednosti na x osi (predviđene sredine na osnovu regresione jednačine), i rezidualne na y osi.

#Kod za rezidualni plot

```
plot(fitted(lm2), residuals(lm2), xlab="reziduali", ylab="fitovane  
vrednosti", xlim=c(200,260), ylim=c(-10,15), abline(0,0))
```



**Slika 16. Rezidualni plot-provera linearnosti ZV i prediktora**

Ukoliko se primeti „U“ oblik, ili neko veliko odstupanje, potrebno bi bilo primeniti ili log transformaciju ZV, ili primeniti čak kvadratnu transformaciju fiksnog efekta – na primer godine su često u kvadratnom odnosu sa drugim varijablama. Takođe, moguće je da se dodavanjem fiksnog efekta to jest njegove interakcije sa već postojećim fiksnim efektom, ispravi ovaj problem (Winter, 2013).



### #Provera kolinearnosti fiksnih efekata

Ovo se odnosi na proveru ineterkorelacija među fiksnim efektima.

```
# Provera međusobnih korelacija prediktora
```

```
mydf.cor = cor (my.df)
```

```
mydf.cor
```

```
          godine vis_glasa  godine_c
godine    1.000000 -0.945371  1.000000
vis_glasa -0.945371  1.000000 -0.945371
godine_c   1.000000 -0.945371  1.000000
```

Što se tiče „leka“ za multiokolinearnost – posebne debate se vode. Najčešće se primenjuje analiza glavnih komponenti. Zatim, ako svi fiksni efekti mere sličnu stvar, onda zadržati najsmisleniji i najrelevantniji za istraživanje

### #Provera homoskedastičnosti

Već je rečeno na početku, treba napraviti kao i za linearnost rezidualni plot, to jest staviti u odnos rezidualne i fitovane vrednosti ZV. Oblik raspršenja reziduala treba da bude „jajast“, okrugao, u svakom slučaju podjednako raspoređen oko linije. Ukoliko nije pravilnog oblika, već ima oblik trougla, trapezoida i sl, znači da model ne fituje jednako dobro sve vrednosti fiksnih efekata. Obično je dovoljno primeniti na primer log transformaciju.

### #Provera normalnosti distribucije reziduala

Prema Winteru (2013) ovo je najmanje bitan uslov LM jer su oni vrlo robustni na odstupanje podataka od normalnosti. Ipak, proverava se tako što se napravi QQ plot (slika 17) reziduala linearnog modela ili histogram.

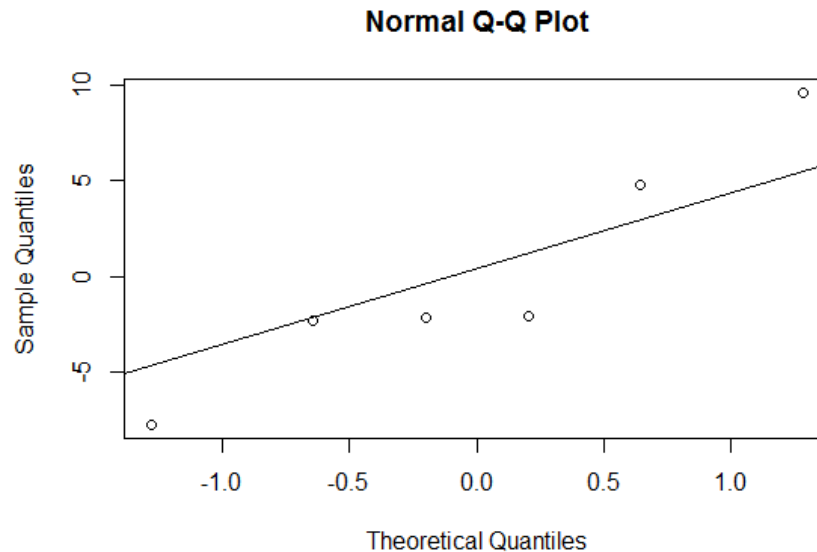
```
# kod za histogram
```

```
hist(residuals(lm2))
```

```
# kod za QQ plot – provera normalnosti reziduala
```

```
qqnorm(residuals(lm2))
```

```
qqline(residuals(lm2)) #dodaje liniju na qq plot
```



**Slika 17. QQ plot reziduala LM**

#### # Odsustvo uticajnih tačaka

Nisu svi autlajeri uticajne tačke (Baayen & Milin, 2010). Uticajne su samo one koje značajno menjaju značenje modela – u smislu da smanjuju ili povećavaju vrednost  $R^2$  ili da menjaju značenje prediktora – promena smeru korelacije. Kako znamo da li je outlier uticajan? Tako što možemo proveriti model sa i bez autlajera, ili možemo da računamo  $DFbeta$  pokazatelje koji govore o uticajnosti pojedinih podataka na model. Često se nazivaju i dijagnostikom „odstrani jednog“, to jest govore o tome kako se menja koeficijent ukoliko se jedan podatak ukloni.

#### # Računanje dfbeta

```
dfbeta(lm2)
```

|   | (Intercept) | godine_c    |
|---|-------------|-------------|
| 1 | -0.8002664  | 0.06437573  |
| 2 | -0.5220150  | 0.02736278  |
| 3 | 0.9678744   | -0.01456709 |
| 4 | 2.0026352   | 0.05092767  |
| 5 | -1.7103247  | -0.06479736 |

**6   -0.7828787   -0.06622744**

Kako se tumače ovi koeficijenti? Tako što ukoliko bi se prvi podatak otklonio, onda bi koeficijent za nagib iznosio  $-0.9099 - 0.064 = -0.97$ . Ako je koeficijent nagiba negativan, onda se Dfbeta vrednost oduzima, a ako je pozitivan onda se sabira (Winter, 2013).

Ne postoji pravilo koja DF vrednost je velika a koja je mala, najbolje je razmišljati u tom pravcu da ako se menja predznak koeficijenta, onda je taj podatak sigurno uticajan. Tek posle toga možemo gledati apsolutne vrednosti DF beta pokazatelja, koje ako iznose upola koliko i nagib, onda su upozoravajuće – na primer, ako je nagib 2, a Dfbeta 1 ili -1, ili ako je nagib 4, a DF beta 2 ili -2.

Jedini podaci koji mogu da se isključe bez razmišljanja su teoretski nemoguće vrednosti (Bayen & Milin, 2010; Winter, 2013) – na primer u leksičkoj odluci, RT-ovi ispod 5 ms u leksičkoj odluci, ili predugački i besmisleni, a da je populacija „normalna“ – RTovi preko 5 s.

### **# Nezavisnost merenja!!!**

Najvažnija pretpostavka svih statističkih testova!

Šta je nezavisnost merenja? Najprostiji primer jeste slučaj bacanja novčića, pri čemu ishod svakog narednog bacanja ne zavisi od prethodnog. U društvenim naukama to znači da ne potiču svi podaci to jest sva merenja sa jednog stimulusa od jednog ispitanika, to jest, sva pojedinačna merenja (data points) trebalo bi da potiču of različitih ispitanika.

Zašto je ovo toliko važno? Zato što zavisna merenja, to jest ona koja potiču od istog ispitanika povećavaju šansu pronalaženja spurioznog efekta, to jest, postoji opasnost da se spuriozni efekat proglasi značajnim!

Međutim, u psihološkim istraživanjima često imamo više pojedinačnih podataka od istog ispitanika bilo po istom nivou faktora, bilo po istom stimulusu.

Tu na svetlo izlaze razne statističke analize, koje bismo spram nacrtu istraživanja. Tako, ako nam je neki faktor ponovljen po ispitanicima radimo ANOVU za ponovljena merenja, ili t-test za zavisne uzorke i slično. Ili često primenjujemo split-plot ANOVU, ili kombinovanu ANOVU gde koristimo ANOVU za zavisne i nezavisne uzorke.

Tradicionalno, istraživači su radili uprosečavanja po ispitanicima i po stimulusima i koristili dve vrste analiza (Baayen & Milin, 2010; Jaeger, 2011; Popović Stijačić, Mihić i Filipović Đurđević, 2018), a zatim računali  $F1$  i  $F2$  test. Da bismo efekat proglasili značajnim, morala su oba test da budu značajna. Računanje  $F$  testa po stimulusima je uveo Klark (Clark, 1971), smatrajući da „individualne“ razlike mogu postojati i u populaciji stimulusa, što se naročito odnosi na populaciju reči, važnu za domen psiholingvistike. Međutim, Klark (Clark, 1971) je smatrao da su  $F1$   $F2$  test nedovoljni, i objedinio ih je u jedan pokazatelj - kvazi  $F_{min}$ , smatrajući da tek značajnost ovog pokazatelja nedvosmisleno ukazuje na značajan efekat. Ipak, retko ko je koristio ovaj pokazatelj, i svi su se oslanjali samo na  $F1$  i  $F2$  testove, pri čemu značajnost jednog ukazuje da se efekat može generalizovati na populaciju ispitanika, a drugi da se efekat može generalizovati na populaciju stimulusa.

MODELI MEŠOVITIH EFEKATA rešavaju problem nezavisnosti merenja, i skraćuju postupak računanja  $F1$  i  $F2$  testa. Pored toga, oni na sofisticiraniji način modeluju grešku, dakle, nema potrebe za uprosečavanjima, jer se koristi takozvani „long data“ format.

## 2) Mali osvrt na mešovite linearne modele

Ako se osvrnemo na prethodni primer gde smo modelovali jednostavnim linearnim modelom visinu glasa i pol, te visinu glasa i godine:

$$\text{visina glasa} \sim \text{pol} + \epsilon,$$

**pol** je fiksni, determinisani, sistematski efekat, a  $\epsilon$  je slučajni, probabilistički efekat modela. Kod mešovitih linearnih modela, fiksni deo je potpuno isti, to jest, tumači se na isti način kao i kod klasičnih linearnih modela, dok se slučajni deo modela procenjuje tako što se aritmetička sredina mapira na 0, a ocenjuje se varijansa slučajnog efekta. Slučajni efekat može da potiče od ispitanika, ali i od stimulusa.

#### \*\*\*\*\*Vežba 4. Mešoviti modeli – teorijske osnove\*\*\*\*\*

```
# Pravljenje novog data frame a kako bismo kreirali novi LM:

visina_tona ~ pol + uctivost + e

#Imamo 2 fiksna efekta: pol i uctivost, pri čemu se uctivost ponavlja po ispitanicima

subj<-c(rep(1:6,8))
pol<-c(rep("m",24),rep("z",24))
uctivost<-c(rep(c(rep(c("uctiv","neuctiv"),times=c(6,6))), 4))
vis_glasa<-sample(160:260,48, replace=T)
mydf3<-data.frame(subj, pol, uctivost, vis_glasa)

mydf3
```

|    | subj | pol | uctivost | vis_glasa |
|----|------|-----|----------|-----------|
| 1  | 1    | m   | uctiv    | 243       |
| 2  | 2    | m   | uctiv    | 175       |
| 3  | 3    | m   | uctiv    | 258       |
| 4  | 4    | m   | uctiv    | 256       |
| 5  | 5    | m   | uctiv    | 172       |
| 6  | 6    | m   | uctiv    | 182       |
| 7  | 1    | m   | neuctiv  | 210       |
| 8  | 2    | m   | neuctiv  | 239       |
| 9  | 3    | m   | neuctiv  | 255       |
| 10 | 4    | m   | neuctiv  | 180       |
| 11 | 5    | m   | neuctiv  | 190       |
| 12 | 6    | m   | neuctiv  | 165       |

```
### KORISNO ### Brz uvid u strukturu podataka ###

#funkcija dim - ispitivanje dimenzionalnosti matrice
#funkcije str - uvid u podatke
#funkcija head - prvih 6 redova matrice
#funkcija tail - poslednjih 6 redova matrice

dim(mydf3)

[1] 48 4
```

```
str(mydf3)
```

```
'data.frame':      48 obs. of  4 variables:
 $ subj      : int   1 2 3 4 5 6 1 2 3 4 ...
 $ pol       : Factor w/ 2 levels "m","z": 1 1 1 1 1 1 1 1 1 1 ...
 $ uctivost  : Factor w/ 2 levels "neuctiv","uctiv": 2 2 2 2 2 2 1 1 1 1 ...
 $ vis_glasa: int   243 175 258 256 172 182 210 239 255 180 ...
```

```
head(mydf3)
```

```
subj pol uctivost vis_glasa
1    1  m    uctiv      243
2    2  m    uctiv      175
3    3  m    uctiv      258
4    4  m    uctiv      256
5    5  m    uctiv      172
6    6  m    uctiv      182
```

```
tail(mydf3)
```

```
      subj pol uctivost vis_glasa
43     1   z  neuctiv      181
44     2   z  neuctiv      204
45     3   z  neuctiv      171
46     4   z  neuctiv      211
47     5   z  neuctiv      165
48     6   z  neuctiv      191
```

```
# Potrebno je da subjekte tretiramo kao faktor, kako bismo mogli da prikazemo variranja po ispitanicima
```

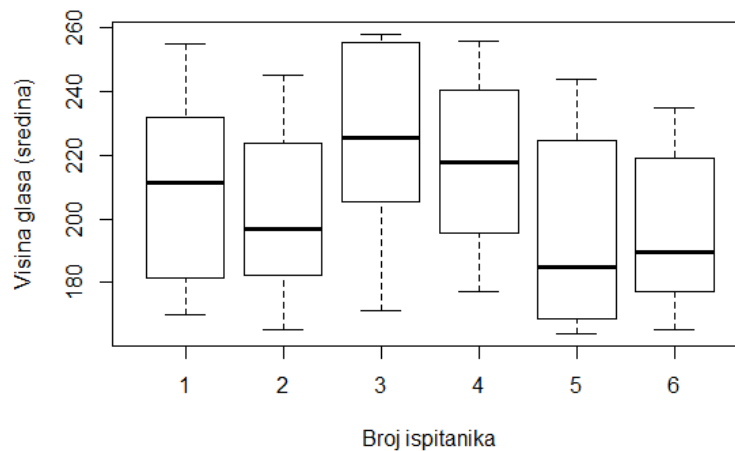
```
mydf3$subj<-as.factor(mydf3$subj)
```

```
is.factor(mydf3$subj)
```

```
[1] TRUE
```

```
# Hajde da vidimo šta se dešava sa variranjima slučajnih efekata (slika 17).
```

```
# Plotovanje variranja aritmetičkih sredina visine glasa među ispitanicima
plot(mydf3$subj,mydf3$vis_glasa, xlab="Broj ispitanika",
ylab="Visina glasa (sredina)")
```



Slika 17. Variranje aritmetičkih sredina visine glasa ispitanika – demonstracija slučajnih efekata ispitanika

Može se videti da aritmetičke sredine visine glasova variraju po ispitanicima. Dakle, višestruki odgovori od istog ispitanika ne mogu se tretirati kao nezavisni jedan od drugog. Da bi se rešila zavisnost merenja, uvodi se SLUČAJNI EFEKAT (*random effect*) za ISPITANIKE. Na ovaj način dozvoljava se da postoji različiti „baseline“ za svakog ispitanika. Kako se to modeluju slučajni efekti koji dozvoljavaju različit „baseline“ – tako što se pretostavi SLUČAJNI INTERCEPT za svakog ispitanika, pri čemu mešoviti modeli rade procenu intercepta za svakog ispitanika. Zato se ovi modeli nazivaju mešovitim, jer uključuju i fiksne i slučajne efekte, dok klasični LM su modeli koji uključuju samo fiksne efekte. Dodavajući slučajni efekat LM modelu, mi pokušavamo da modelujemo taj šum, to jest neobjašnjeni deo varijanse, koji potiče od individualnih razlika. Novi model sa slučajnim efektom ispitanika, to jest sa slučajnim interceptom ispitanika izgleda ovako:

$$\text{visina glasa} \sim \text{uctivost} + \text{pol} + (1|\text{ispitanik}) + \epsilon,$$

1 označava intercept, i na ovaj način kažemo modelu da može da očekuje da će da bude više odgovora po ispitaniku - na ovaj način se rešava problem zavisnosti

Primetite da greška i dalje postoji u modelu!

```
# Dodajemo u data frame i variranja koja potiču od različitih ajtema, kojih u ovom izmišljenom primeru ima 12, po 6 za svaku situaciju učtivosti.
```

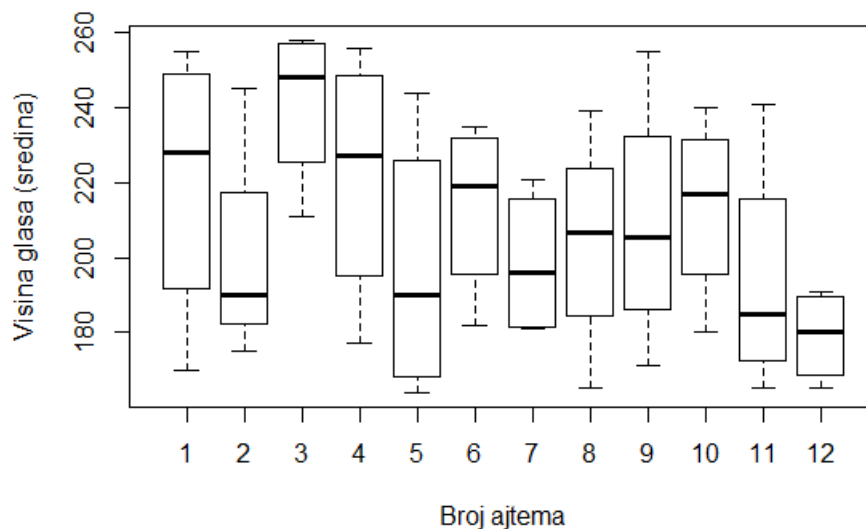
```
item<-c(rep(1:12,4))
```

```
mydf4<-data.frame(subj, pol,uctivost, item, vis_glasa)
```

```
mydf4
```

```
# Plotujemo variranja sredina visine glasa po ajtemima (slika 18).
```

```
plot(mydf4$item,mydf4$vis_glasa, xlab="Broj ajtema",  
ylab="Visina glasa (sredina)")
```



Slika 18. Variranja sredina visine glasa po ajtemima – demonstracija slučajnih efekata ajtema

Kao što se na slici 18 može videti, postoje slučajna variranja koja potiču od samih ajtema, i koja nisu sistematski obuhvaćena fiksnim faktorima pola i učtivosti. Drugim rečima, odgovori različitih ispitanika mogu biti pod uticajem slučajnog faktora koji potiče od varijacije stimulusa, odnosno, različiti odgovori na jedan stimulus ne mogu biti posmatrani kao nezavisni. To znači da u naš model možemo da dodamo još jedan slučajni efekat koji potiče od stimulusa:

$$\text{visina glasa} \sim \text{uctivost} + \text{pol} + (1|\text{ispitanik}) + (1|\text{ajtem}) + \epsilon.$$



Treba primetiti da i dalje postoji greška u modelu, dakle, modelovanjem dve vrste slučajnih efekata ne iscrpljujemo nesistematska slučajna variranja!

U čemu je lepota ovog modela? Već smo spomenuli da kod mešovitih modela ne radimo nikakva uprosečavanja – imamo matricu u dugačkom formatu (sa sirovim podacima), a modelu kažemo koja to merenja nisu nezavisna, to jest koja se merenja ponavljaju bilo po ispitanicima (1|ispitanik) bilo po stimulusima (1|ajtem), pri čemu ne gubimo informaciju o individualnim variranjima unutar ispitanika i unutar ajtema (što se inače uprosečavanjem dešava).

Ovaj model se često naziva RANDOM INTERCEPT MODEL, i znači da smo za svakog ispitanika i za svaki stimulus prepostavili drugačiji „baseline“ zavisne varijable, u ovom slučaju, visine glasa. Istovremeno, ovaj model kaže da su fiksni efekti, ma kavi bili isti za sve ispitanike i sve stimulse. Ali da li je ovo ispravna pretpostavka?

Ovde dolazimo do *slučajnog variranja nagiba*, odnosno, možemo očekivati da fiksni efekat ima drugačiji nagib u zavisnosti od stimulusa ili ispitanika, što je često i slučaj. Uvođenjem ovih efekata dobijamo RANDOM SLOPE MODEL, gde se dozvoljava ne samo variranje intercepta, već i variranje nagiba fiksnog efekta za svakog ispitanika i za svaki stimulus. Kako ćemo reći programu da hoćemo da nam varira nagib?

$$\text{visina glasa} \sim \text{uativost} + \text{pol} + (1 + \text{uativost} | \text{ispitanik}) + (1 | \text{ajtem}) + \varepsilon.$$

1 označava intercept, i na ovaj način kažemo modelu da može da očekuje da ima različit baseline visine glasa za svakog ispitanika ali i to da će odgovor varirati u odnosu na glavni faktor, to jest fiksni efekat, u ovom slučaju to je uativost.

U ovom slučaju nije baš smisleno uključiti variranja nagiba koja potiču od stimulusa, jer stimulusi su ugnježdjeni u faktor uativosti – ne mogu istovremeno pripadati i uativom i neuativom nivou.

Baayen smatra (Baayen et al, 2013) da nije dovoljno u model uključiti samo variranje intercepta, već da je potrebno uključiti maksimalnu random strukturu, koja je opravdana nacrtom, a ne podacima (o ovome će biti kasnije reči; „data driven“ pristup zastupa Baayen, 2017). Možda najjednostavniji podsetnik jeste da u manipulacijama između grupa, kada su u ispitanici ugnježdjeni u dve

eksperimentalne situacije (svaki ispitanik pripada samo jednoj od eksperimentalnih grupa), što znači da ispitanik i tretman, to jest grupa ne mogu da interreaguju (ne znamo kako bi ispitanik reagovao na drugu manipulaciju), odnosno, drugim rečima, random slope po ispitanicima ne bi bilo moguće definisati – usled samog dizajna. U globalu, pravilo je sledeće: unutar grupne manipulacije zahtevaju i variranja po interceptu i variranja nagiba, dok među grupna variranja zahtevaju definisanje samo random intercept slučajnih efekata.

Postoje situacije kada model takođe nije moguće definisati, ali zato što ne postoji dovoljan broj jedinica posmatranja, pa model ne može da konvergira, to nema dovoljan broj podataka da bi se ocenili parametri modela. Ipak, Baar i saradnici (Baar et al, 2013) smatra da uvek treba primeniti maksimalnu random strukturu, a to je ona maksimalna random struktura koja je dozvoljena eksperimentalnim dizajnom. Na taj način se može generalizovati na sve buduće ispitanike i stimulse. Kako Baar navodi (Baar et al, 2013), model koji ne uključuje random intercepte ima smanjenu snagu, dok model koji ne uključuje random nagibe imaju povećanu verovatnoću greške tipa I. Na ovaj način, on zaključuje da su modeli sa maksimalnom strukturom idealni, jer imaju i dovoljnu snagu, i smanjenu verovatnoću greške tipa I. Baar (Baar et al, 2013) predlaže da se u model svakako uključi random nagib, i da se nikako ne završi samo na random intercept modelu, zato smatra da istraživači prvo probaju da isključe interakciju nagiba i intercepta:

$$\text{visina glasa} \sim \text{uctivost} + \text{pol} + (1|\text{ispitanik}) + (0 + \text{uctivost}|\text{ispitanik}) + (1|\text{ajtem}) + \epsilon.$$

Ukoliko bi ispitanici sa višom frekvencijom glasa inače bili manje osetljivi na variranja procedure – to jest učtivosti, ovaj model to ne bi mogao da uhvati, jer mu je isključena interakcija intercepta i nagiba.

Sledeća redukcija modela išla bi u pravcu da se isključi jedan random intercept, ali da se ipak zadrži nagib:

$$\text{visina glasa} \sim \text{uctivost} + \text{pol} + (0 + \text{uctivost}|\text{ispitanik}) + (1|\text{ajtem}) + \epsilon.$$

U ovom modelu, isključeno je variranje intercepta po ispitanicima. Ali, ukoliko je više opravdano dizajnom, moguće je isključiti i variranja po ajtemima. Uglavnom, tu na scenu može da stupi „data driven“ pristup, pa da se porede likelihood modela sa jednim odnosno sa drugim random interceptom.

Opravdanost ovakvih preporuka Baar (Baar et al, 2013) vidi u tome što overfitovanje podataka ne dovodi do dramatičnog povećanja greške tipa I, što je pokazao simulacijama, ali da nedovoljan fit eksperimentalnog dizajna (*overfitting the data vs underfitting the design*) ne dovodi do poželjene generalizacije efekata na populaciju ispitanika i stimulusa.

### **\*\*\*Vežba 5: Mešoviti linearni modeli: priprema podataka, građenje modela, provera pretpostavki linearnih modela\*\*\***

```
##### Kod prilagođen za radionicu u Banja Luci, april, 2019 #####
#####LEPmler2018_cas3: Uvod u analizu mesovitih efekata### ovo je naziv originalnog skripta
# Za pocetak, ucitacemo potrebne pakete

library(lme4) # da bismo pravili modele
library(lmerTest) # da nam budu prikazane p vrednosti
library(ggplot2) # za grafikone
library(gridExtra) # za uredjivanje visestrukih grafikona
library(languageR) # za razne stvari
library(lattice) # za grafikone

# I ucitacemo jedan dataframe

#dat.im=read.table("vezba 5.txt",sep="\t",T) #ovo je tradicionalno preko konzole
vezba 5 <- read.delim("C:/Users/lenovo/Desktop/LMERDUSICA OBUKA/LEPlmer2018-
master/Dusica.imenice.2004.txt")

View (vezba 5)

dat.im <- vezbe 5

# proverimo dimenzije

dim(dat.im)

# mali uvid u podatke

head(dat.im)
```

```

# napravimo uvid u strukturu podataka
str(dat.im)

# zadržimo u data frame-u samo reci
dat.im = dat.im[dat.im$Leksikalnost == "word",]

# zadržimo u data frame-u samo tacne odgovore
dat.im = dat.im[dat.im$error_code == "C",]

# proverimo da li je ZV normalno distribuirana – zatarabljen kod iz ggplota, samo se odarabi i
radi
# g1 = ggplot(dat.im, aes(RT)) + geom_density()
# g2 = ggplot(dat.im, aes(sample=RT)) +
#   stat_qq() + stat_qq_line()
qqnorm(dat.im$RT)

# za sada cemo primeniti log transformaciju RT
dat.im$RT = log(dat.im$RT)

# opet proverimo distribuciju
qqnorm(dat.im$RT)

# Transformisemo frekvenciju reci, jer znamo da stoji u log odnosu sa RT (SETITE SE
#USLOVA O LINEARNOSTI)
dat.im$frekv = log(dat.im$Frekvencija)
qqnorm(dat.im$Frekvencija)
qqnorm(dat.im$frekv)

# Pored toga, kontinuirane prediktore treba centrirati na nulu zbog smislenosti intercepta, a jos je
# bolje normalizovati vrednosti:
dat.im$frekv.sirovo = dat.im$frekv #da bismo sacuvali sirove
                                #frekvencije u data frameu

dat.im$frekv = scale(dat.im$frekv) #normalizovanje

# da vidimo sta smo uradili sa frekv:
qqnorm(dat.im$frekv.sirovo)

```

```

qqnorm(dat.im$frekv) #sveli na (0,1)
mean(dat.im$frekv.sirovo)
exp(mean(dat.im$frekv.sirovo))
round(mean(dat.im$frekv),5)
# skaliramo i broj znacenja
dat.im$NoS = scale(dat.im$NoS)

#### GRAĐENJE MODELA SA SLUČAJNIM EFEKTIMA ####

#Počinjemo sa random interceptom i variranjem ispitanika, jer da ništa ne znamo o nacrtu,
#znamo da se ispitanici razliku međusobno po većini svojstava

# Slučajni efekti ispitanika

lmer1 = lmer( RT ~ 1 + (1|Subject), data = dat.im)

```

Pored toga sto ocekujemo da se ispitanici razlikuju po brzini,ocekujemo i da vreme reagovanja nece biti isto za sve reci napravimo model koji informisemo o tome da ocekujemo razlicit intercept za svaku rec, tj. ocekujemo da se reci razlikuju medjusobno po brzini kojom se reaguje na njih posto nismo prikazali sve reci srpskog jezika, a dodatno, zelimo da svoje nalaze generalizujemo na citavu populaciju reci naseg jezika ni reci ne mozemo tretirati kao fiksne efekte, vec kao slucajne.

```

#Slučajni efekat stimulusa

lmer2 = lmer( RT ~ 1 + (1|Rec), data = dat.im)

# Mozemo da napravimo model koji istovremeno informisemo da ocekujemo i razlike izmedju
# ispitanika i razlike izmedje reci

lmer3 = lmer( RT ~ 1 + (1|Subject) + (1|Rec), data = dat.im)
summary(lmer3)

```

Mozemo i da proverimo da li je opravdano ukljuciti svaki od ova dva slucajna efekta. Setite se - ovo je "data driven" pristup. Uporedimo model koji sadrzi samo ispitanike i model koji sadrzi i ispitanike i stimulse, da bismo proverili da li je opravdano ukljuciti stimulse kao random efekat

```
# Provera opravdanosti slučajnih efekata – poređenje modela

anova(lmer1, lmer3) # jeste

# Uporedimo model koji sadrzi samo stimulse i model koji sadrzi i ispitanike i stimulse
# da bismo proverili da li je opravdano ukljuciti ispitanike kao random efekat

anova(lmer2, lmer3) # jeste

# Dakle, model sa oba izvora slučajnih efekata je opravdan i dizajnom (dva izvora zavisnosti
# merenja: ispitanici i reci) i podacima.

##### KAKO DODAJEMO FIKSNE EFEKTE? #####

# Na isti nacin kao uobicnim linearnim modelima:

lmer4 = lmer(RT ~ frekv + (1|Subject) + (1|Rec), data = dat.im)
```

Da li dodavanje frekvencije kao fiksnog efekta cini da model bolje opisuje podatke? Da li je opravdan podacima ili nepotrebno usloznjava model? Nekontrolisano dodavanje prediktora moze da dovede do tzv. overfitting-a.

```
anova(lmer3, lmer4)

# Vidimo da je opravdano ukljuciti frekvenciju, jer model koji nju sadrzi ima manji AIC, manji
# BIC i veci loglikelihood. Tek kad utvrdimo da dodavanje prediktora čini model opravdano
# boljim gledamo koeficijente iz modela.

summary(lmer4)
```

```
##### ŠTA DOBIJAMO KAD PRIKAŽEMO REZIME MODELA?#####
#####

# Prve linije daju osnovne podatke o algoritmu, formuli koju smo primenili i podacima
Linear mixed model fit by REML. t-tests use Satterthwaite's
  method ['lmerModLmerTest']
Formula: RT ~ frekv + (1 | Subject) + (1 | Rec)
Data: dat.im

# Potom dobijamo REML (Restricted Maximum Likelihood) kriterijum konvergiranja
# (koji moze da posluži kao indeks za goodness of fit te i za poredjenje modela)

REML criterion at convergence: -1591.8

# Dobijamo osnovne podatke o distribuciji reziduala
# (za sada se čini da je simetrična, kasnije ćemo to dalje proveravati)

Scaled residuals:
    Min       1Q   Median       3Q      Max
-4.6145 -0.6549 -0.1456  0.5323  5.5305
```

Dolazimo do dela ispisa u kom su prikazani parametri za slučajne efekte. Rekli smo da se za njih procenjuje varijansa/standardna devijacija. Vidimo procenu za slučajni intercept za reci, procenu za slučajni intercept za ispitanike i rezidual. Rezidual je ono što smo u običnom linearnom modelu označavali kao gresku (ono čiju strukturu ne razumemo). Možemo da kažemo i da smo gresku iz lm razdvojili na deo čiju strukturu razumemo (različite prosečne brzine ispitanika i reci) i deo čiju strukturu ne razumemo (gresku).

```
Random effects:
 Groups   Name                Variance Std.Dev.
 Rec      (Intercept)  0.003043  0.05516
 Subject  (Intercept)  0.012521  0.11190
 Residual                    0.024610  0.15688
Number of obs: 2095, groups:  Rec, 90; Subject, 24
```

Na kraju, prikazani su koeficijenti za FIKSNE EFEKTE. Mi imamo jedan kontinuirani prediktor. To znaci da nam intercept kaze koju vrednost ima ZV kada je vrednost NV jednaka nuli. Da bi ovo bilo smisljeno, centrirali smo prediktor na nulu, što znaci da nula sada oznacava prosečnu frekvencu, te dobijamo podatak o vrednosti ZV (tj. RT) za prosečnu vrednost NV (tj. frekvence). Drugi koeficijent odnosi se na prediktor i govori nam za koliko se promeni vrednost ZV, kada se vrednost NV poveca za jedan. Vidimo da je povecanje frekvence za jedno mesto na skali praceno skracenjem vremena reakcije za 0.029, kao i da je ova promena statisticki znacajna.

```
Fixed effects:
              Estimate Std. Error      df t value Pr(>|t|)
(Intercept)  6.457019   0.023819 25.960000 271.092 < 2e-16 ***
frekv       -0.024173   0.006766 84.730000  -3.573 0.000586 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Correlation of Fixed Effects:

(Intr)

frekv 0.002

#####

# parcijalni fiksni efekat prediktora mozemo ovako da dobijemo:

```
plotLMER.fnc(lmer4)
```

Medjutim, pored toga sto informisemo model o tome da ocekuje razlicita prosečna vremena reagovanja od razlicitih ispitanika i za razlicite reci mozemo da se zapitamo i da li je neki fiksni efekat bas isti za sve ispitanike.

```
# Da vidimo kako izgleda efekat frekvencije odvojeno za svakog ispitanika
```

```
ggplot(dat.im, aes(x=frekv, y=RT)) +
  geom_point() +
  geom_smooth(method = "lm", se = TRUE) +
  facet_wrap(~Subject)
```



```
# Mozemo da informisemo model o tome da ocekujemo razlicit efekat frekvencije za razlicite
# ispitanike

lmer5 = lmer( RT ~ frekv + (1 + frekv|Subject) + (1|Rec), data =
  dat.im)

# proverimo da li je ovo opravdano podacima

anova(lmer4, lmer5) # zapravo nije potreban slucajni nagib...

# Da pogledamo kako izgledaju brojke

summary(lmer5)
```

Mozemo da primetimo dve stvari:

- 1) variranje za nagib frekvence je mnogo manje nego variranje za intercept za ispitanike ili reci
- 2) korelacija izmedju intercepta za ispitanike i nagiba efekta frekvence jednaka je -1, što znaci da su ispitanici koji su bili brzi bili istovremeno i osetljiviji na frekvenciju. Medjutim, to sto je korelacija ovako visoka je cesto znak da smo ukljucili nepotrebne parametre u model, što nam, uostalom, poredjenje dva modela i sugerise.

```
# da vidimo korelaciju izmedju intercepta za ispitanika i nagiba po ispitaniku:

plot(ranef(lmer5)$Subject)
```

Medjutim, postoji glediste po kom variranje nagiba frekvence po ispitaniku treba ostaviti u modelu, jer je opravdano nacrtom (cak i ako nije opravdano podacima), te doprinosi razresavanju problema zavisnih merenja.

```
# Ako se odlucimo da ostavimo ovu tzv. "slucajnu interakciju", mozemo da pokusamo
# da iskljucimo koralaciju izmedju intercepta za ispitanika i nagiba po ispitaniku:

lmer6 = lmer( RT ~ frekv + (1 + frekv||Subject) + (1|Rec), data
  = dat.im)

# to smo postigli sa dve vertikalne linije, a mogli smo isto to i ovako:

lmer6 = lmer( RT ~ frekv + (1 |Subject) + (0 + frekv|Subject) +
  (1|Rec), data = dat.im)

anova(lmer5, lmer6)
```

```
# vidimo da ni ovo nije opravdano podacima

# a kad pogledamo ispis, vidimo da je variarnje nagiba zaista blisko nuli
summary(lmer6)

# za fiksne efekte treba da proverimo i da li postoji nelinearna komponenta
lmer6n = lmer( RT ~ poly(frekv,degree=2,raw=T) +
              (1 + frekv||Subject) + (1|Rec), data = dat.im)
anova(lmer6, lmer6n)

# vidimo da ni ovo nije opravdano podacima
```

Model sa nelinearnim efektom čak ima nešto lošiji fit. Pogledacemo ispis, tek da vidimo kako se izlazi na kraj sa nelinearnostima u linearnom modelu primetite da za prediktor frekv sada postoje dva koeficijenta: prvi se odnosi na linearnu komponentu a drugi se odnosi na kvadratnu komponentu.

```
summary(lmer6n)
```

Na slican nacin na koji smo se pitali da li postoji variranje nagiba efekta frekvencije po ispitanicima, mozemo da se zapitamo i da li postoji analogno variranje po recima. Medjutim, frekvencija nije ponovljena po recima, tj. jedna rec je uvek iste frekvencije pa bi ovo pitanje bilo besmisleno, tj. ne bi bilo opravdano nacrtom. To mozemo da ucinimo za neki prediktor koji je ponovljen po recima. U ovom slucaju, u te svrhe moze da nam posluzi varijabla (koju sam napravila za potrebe demonstracije) koja se zove Brzina.ispitanika. To je kategorijalna varijabla koja je napravljena tako sto su ispitanici podeljeni u dve grupe (brzi, spori) na osnovu medijane varijable SubjSpeed što je prosečno vreme reakcije ispitanika u eksperimentu.

```
# da pogledamo prvo da li su brzi ispitanici brzi na svim recima, kao i da li su podjednako brzi
na

# razlicitim recima:

# ggplot(dat.im, aes(x=Brzina.ispitanika, y=RT)) +
# geom_point() +
# geom_smooth(method = "lm", se = TRUE) +
# facet_wrap(~Rec)
```

```
#Napravimo model u koji unesemo informaciju o tome da očekujemo razlike u odnosima između  
#brzih i sporih ispitanika za različite reči. Ovo je dozvoljeno, pošto su svaku reč videli i brzi i  
#spori ispitanici, odnosno, faktor Brzina.ispitanika je ponovljen po rečima.
```

```
lmer7 = lmer( RT ~ frekv + (1 + frekv||Subject) +  
              (1 + Brzina.ispitanika|Rec), data = dat.im)  
anova(lmer6, lmer7)
```

```
# ponovo, vidimo da ovo nije opravdano podacima
```

```
# Kad pogledamo rezime modela, vidimo i da je variranje vrlo nisko, a korelacija ponovo veoma  
# visoka:
```

```
summary(lmer7)
```

```
# ponovo isključimo korelaciju:
```

```
lmer8 = lmer( RT ~ frekv + (1 + frekv||Subject) +  
              (1 + Brzina.ispitanika||Rec), data = dat.im)  
anova(lmer6, lmer8)
```

```
# ponovo, vidimo da ovo nije opravdano podacima
```

```
# a kad pogledamo rezime modela, vidimo i da je variranje vrlo nisko
```

```
summary(lmer8)
```

```
# ovo je svakako bilo u svrhu ilustracije, pa se vraćamo na model lmer6
```

Na ovaj način možemo da dodajemo nove prediktore za njih testiramo nelinearnosti, fiksne i slučajne interakcije opravdanost podacima. Međutim, interpretacija fiksnih efekata u lmer-u je ista kao interpretacija u LM-u, kojom smo se bavili na početku, stoga nećemo uključivati nove prediktore u ovu analizu.

```
# Za koeficijente iz modela možemo da procenimo 95% intervale poverenja
```

```
confint(lmer6, method="Wald")
```

```
# Vidimo da se nasi efekti uvek nalaze sa iste strane nule
```

## ##### PROVERA PRETPOSTAVKI MODELA #####

Sada cemo da proverimo da li su prekršeni neki od preduslova za primenu linernog modela:

```
# napravimo kolonu sa predviđenim vrednostima ZV:
```

```
dat.im$RT.fitted = predict(lmer6)
```

```
# ako nas zanima procenat objasnjene varijanse:
```

```
cor(dat.im$RT, dat.im$RT.fitted)^2
```

```
# napravimo kolonu sa rezidualima:
```

```
dat.im$RT.res = residuals(lmer6)
```

### #Provera homoskedastičnosti

Plotujemo korelaciju između fitovanih vrednosti i reziduala da proverimo da li postoji homogenost varijanse - ovo treba da bude jedno lepo "jaje".

```
# klasičan prikaz
```

```
plot(predict(lmer6), residuals(lmer6),  
      xlab="reziduali", ylab="fitovane vrednosti", abline(0,0))
```

```
#prikaz u ggplotu
```

```
ggplot(dat.im, aes(x=RT.fitted, y=RT.res)) +  
  geom_point() +  
  geom_smooth(method = "lm", se = TRUE)
```

**# Proverimo da li se reziduali normalno distribuiraju** - ovo treba da bude što sličnije ravnoj liniji:

```
#klasičan prikaz
```

```
qqnorm(residuals(lmer6))
```

```
qqline(residuals(lmer6)) #dodaje liniju na qq plot
```

```
# isto to može i ovako, pa prikaze i skater i normalnost :)
```

```
par(mfcol=c(1,2))
```

```
qqnorm(resid(lmer6))
```

```
plot(fitted(lmer6), resid(lmer6))  
par(mfcol=c(1,1))
```

### # Provera uticajnosti tačaka

Sad cemo da izbacimo tacke sa velikim rezidualima da proverimo da li uticu previse na model.

```
# refitujemo model na podskupu tacaka  
lmer6t = lmer( RT ~ frekv + (1 + frekv||Subject) + (1|Rec),  
              data = dat.im,  
              subset=abs(scale(resid(lmer6)))<2.5)  
summary(lmer6t)  
# Provera distribucije i homogenosti nakon izbacivanja outliera, i to onih čiji se reziduali nalaze  
# izvan opsega  $\pm 2.5$  SD  
par(mfcol=c(1,2))  
qqnorm(resid(lmer6t))  
plot(fitted(lmer6t), resid(lmer6t))  
par(mfcol=c(1,1))  
#Vidimo da sad reziduali izgledaju mnogo bolje. Efekti su opstali i kad smo se otarasili strcaka
```

## DOBRA PRAKSA 1: OPIS REZULTATA MEŠOVITIH MODELA

- 2) Mali osvrt na logističku regresiju
- 3) Mali osvrt na mešovite logit modele

## REFERENCE

Navedeni kodovi su sinteza nekoliko domaćih i stranih izvora:

1. Baar, D., Levy, R., Scheepers, C. & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3). doi:10.1016/j.jml.2012.11.001.
2. Baayen, H. & Milin, P. (2010). Analyzing Reaction Times. *International Journal of Psychological Research*, 3(2), 12-28.
3. Baayen, R. H. (2012). Mixed-effects models. In Cohn, A. C., Fougerson, C., and Huffman, M.K. (Eds.) *Handbook of Laboratory Phonology*, 668 – 677. Oxford: Oxford University Press. [pdf](#)
4. Clark, H. (1973). The Language-as-Fixed-Effect Fallacy: A Critique of Language Statistics in Psychological Research. *Journal Of Verbal Learning And Verbal Behavior* 12, 335-359.
5. Dušica Filipović Đurđević - LEPlmer2018, koristili smo deo materijala koji je prof. Dušica koristila na svojoj radionici koju je sprovedla na Filozofskom fakultetu u Beogradu tokom novembra 2018. Ukoliko želite materijal sa njene radionice, možete mu pristupiti preko GitHub-a, na sledećem linku: <https://github.com/dfdurdevic/LEPlmer2018>.
6. Matuschek, H., Kliegl, R., Vasishth, S., Baayen, R. H., and Bates, D. (2017). Balancing Type I Error and Power in Linear Mixed Models. *Journal of Memory and Language*, 94, 305 - 315. [http pdf](#)
7. Popović Stijačić, M., Mihić, L., & Filipović Đurđević, D. [2018]. Analyzing data from memory tasks: Comparison of ANOVA, logistic regression and mixed logit model. *Psihologija*, 51(4), 469-488.
8. Radanović, J. i Vaci, N. (2013). Analiza vremena reakcije modelovanjem linearnih mešovityh efekata. *Primenjena psihologija*, 6(3), 311-332.
9. Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications. arXiv:1308.5499. [<http://arxiv.org/pdf/1308.5499.pdf>]