Analiza podataka iz edukacijske domene

Tin Ivan Križ

1. Sadržaj

2. Uvod

U ovom radu ću se baviti eksploratornom analizom dobivenih podataka iz edukacijske domene. Dobiveni podatkovni skup sadrži podatke o riješenosti svih ispita iz predmeta "Osnove elektrotehnike" u 2012./2013. akademskoj godini.

Cilj eksploratorne analize je:

1. Dobiti bolji uvid u dobivene podatke
2. Uočiti anomalije
3. Uočiti bitne varijable
4. Stvoriti i testirati pretpostavke o podatcima

Rad je organiziran tako da slijedi i prikazuje potrebne korake analize u programskom jeziku R. Za svaki korak će biti objašnjena motivacija, implementacija i rezultat koji je najčešće popraćen opažanjima i zaključcima. Treba naglasiti da cilj ovog rada nije stvaranje modela koji bi mogli opisati testirane relacije, već dobiti što bolji uvid u podatke i uočiti koje su varijable bitne i naslutiti kako one koreliraju. Također, analiza se odnosi na ispite samo jedne akademske godine. U skladu s time, svaki zaključak nema dovoljno čvrstu matematičku podlogu, već više govori o nećemo što bi se dalo naslutiti na temelju dosadašnje analize.

3. Analiza

3.1 Učitavanje i prilagodba podataka

Najprije moram učitati podatke iz *CSV* datoteke u računalnu strukturu data frame (*hrv.* okvir podataka) kako bih njima mogao raspolagati. Ovi podatci su relativno jednostavni pošto ne sadrže nepoznate vrijednosti. Jedino je potrebno pretvoriti stupac "datum" u odgovarajući tip i podatci su spremni za analizu.

```
#Učitavanje podataka
odgovori <- read.csv("studenti_odgovori_ispiti_OE.csv", fileEncoding = "UTF-8
" , stringsAsFactors = F, na.strings = c("NULL","NA"))
odgovori$datum <- as.Date(odgovori$datum, "%d.%m.%Y")
odgovori <- tbl_df(odgovori)

# Početni uvid u podatke i tipove podataka stupaca
head(odgovori)

## # A tibble: 6 x 6
##   studentID IDpitanja odgovor      datum  nivo naziv.ispita
##       <int>     <int>   <int>     <date> <int>        <chr>
## 1      7694      2737       1 2012-11-22     3           MI
## 2      7416      2737       1 2012-11-22     3           MI
## 3      7529      2737       0 2012-11-22     3           MI
## 4      7527      2737       0 2012-11-22     3           MI
```

```
## 5       7030       2737        0 2012-11-22      3            MI
## 6       7376       2737        1 2012-11-22      3            MI
```

3.2 Identifikacija anomalija

Pri analizi ne želim imati neočekivane anomalije u podatcima. Zato, prije analize, pokušavam pronaći i ukloniti sve potencijalne anomalije.

3.2.1 Ima li svaki student za isti ispit jednak broj zadataka?

```
# Broj studenata koji ima neispravan ukupan broj bodova
brojLosihZapisa <- group_by(odgovori, studentID, naziv.ispita) %>% summarise(
UkupanBrojZadataka =  n()) %>% arrange(desc(UkupanBrojZadataka)) %>% filter(U
kupanBrojZadataka!=10 & naziv.ispita %in% c("MI", "ZI") ||  UkupanBrojZadatak
a !=20 & naziv.ispita %in% c("Ljetni", "Jesenski", "Zimski", "Dekanski"))

nrow(brojLosihZapisa) # 0 zapisa
```

```
## [1] 0
```

Opažanja

- Svaki student ima valjani broj zadataka u skladu s pristupljenim ispitom.

3.2.2 Postoje li NA (nepostojeći) zapisi?

```
anyNA(odgovori)
```

```
## [1] FALSE
```

Opažanja

- Ne postoje NA vrijednosti u podatcima.

3.3 Početni uvid u podatke

Uvjeren u konzistentnost podataka, mogu krenuti s analizom podataka. Najprije pokušavam dobiti osnovne informacije o podatcima u nadi da ću primijetiti kauzalne odnose među varijablama.

3.3.1 Izlaznost na ispite

Zanima me koliko studenata je izašlo na pojedinu vrstu ispita. Također me zanima kako se odnose broj studenata koji su izašli na MI a nisu na ZI ili obrnuto ili koji nisu uopće izlazili kontinuirano.

```
pristupiloIspitu <- group_by(odgovori, naziv.ispita) %>% distinct(naziv.ispit
a, studentID) %>% summarise(Pristupilo = n())
pristupiloIspitu
```

```
## # A tibble: 6 x 2
##   naziv.ispita Pristupilo
##          <chr>      <int>
## 1     Dekanski         51
## 2     Jesenski        177
## 3       Ljetni         57
## 4           MI        721
## 5           ZI        543
## 6       Zimski        189
```

Opažanja

- Različit broj osoba je izašao na MI od ZI.

```
# Uvid u uspješnost studenata na kontinuiranoj nastavi (MI+ZI)
kontinuirano <- filter(odgovori, naziv.ispita=="MI" | naziv.ispita == "ZI") %
>% mutate(naziv.ispita="Kontinuirano") %>% group_by(studentID, naziv.ispita)
%>% summarise(Ukupno = sum(odgovor))
#head(kontinuirano)
rokovi <- filter(odgovori, naziv.ispita!="MI" & naziv.ispita != "ZI") %>%grou
p_by(studentID, naziv.ispita) %>% summarise(Ukupno = sum(odgovor))
#head(rokovi)

# Spajam u jedinstvenu tablicu
spojeniIspiti <- rbind(rokovi, kontinuirano) %>% arrange(desc(Ukupno))
# sample_n(ungroup(spojeniIspiti), 6)   # MEĐUREZULTAT - dodatak

# Stvaram novu tablicu iz koje mogu čitati koliko je studenata pristupilo koj
em tipu ispitivanja
pristupiloIspituProsireno <- group_by(spojeniIspiti, naziv.ispita) %>% summar
ise(Pristupilo = n())
head(pristupiloIspituProsireno)
```

```
## # A tibble: 5 x 2
##   naziv.ispita Pristupilo
##          <chr>      <int>
## 1     Dekanski         51
## 2     Jesenski        177
## 3 Kontinuirano        722
## 4       Ljetni         57
## 5       Zimski        189
```

Opažanja

- Jedan student je izašao samo na ZI.

```
# Priprema podataka - "Raširit" ću podatke kako bi imali bolji pregled uspjeh
a pojedinog studenta

spojeniSpread <- mutate(spojeniIspiti, Ukupno = Ukupno/20) %>% spread(naziv.i
spita,  Ukupno,fill = NA, convert = FALSE, drop = TRUE, sep = NULL)
sample_n(ungroup(spojeniSpread),5) # MEĐUREZULTAT
```

```
spojeniSpreadSviIspiti <- group_by(odgovori, studentID, naziv.ispita) %>% sum
marise(Rezultat=mean(odgovor)) %>% spread(naziv.ispita,  Rezultat,fill = NA,
convert = FALSE, drop = TRUE, sep = NULL)
#head(spojeniSpreadSviIspiti) # MEĐUREZULTAT
```

```
## # A tibble: 5 x 6
##   studentID Dekanski Jesenski Kontinuirano Ljetni Zimski
##       <int>    <dbl>    <dbl>        <dbl>  <dbl>  <dbl>
## 1      7198       NA       NA         0.30   0.30   0.25
## 2      7579       NA      0.1         0.05     NA     NA
## 3      7287       NA       NA         0.40   0.65   0.30
## 4      7651       NA       NA         0.65     NA     NA
## 5      7696      0.6      0.3         0.25     NA   0.35
```

Opažanja

- Ovaj oblik podatkovnog skupa može biti jako koristan za usporedbe uspjeha/izlaznosti studenata na različitim vrstama ispita.

```
# Identifikacija studenata koji nisu uopće izašli kontinuirano
nrow(filter(spojeniSpread, is.na(Kontinuirano))) # Ima 10 studenata
```

```
## [1] 10
```

Opažanja

- 10 studenata nije uopće pristupilo MI ili ZI.

3.3.2 Uvid u rezultate pojedinog pitanja

Nastojim odgovoriti na sljedeća pitanja:

- Postoje li bolje i lošije riješena pitanja?
- Koji tip ispita ima najbolje ili najlošije riješeno pitanje?

```
# Prosječna riješenost jednog pitanja
paste("Prosjecna rijesenost pitanja je " , 100 * round(mean(odgovori$odgovor)
,4), "%") # 44,07 %
```

```
## [1] "Prosjecna rijesenost pitanja je  44.07 %"
```

```
# Grupiram po pitanjima i ispitima
odgovoriGrupiraniPitanja <- group_by(odgovori, IDpitanja, naziv.ispita) %>% s
ummarise(ProsjekRjesenosti = mean(odgovor)) %>% arrange(desc(ProsjekRjesenost
i))

# Najbolje riješena pitanja
head(odgovoriGrupiraniPitanja, 5)
```

```
## # A tibble: 5 x 3
## # Groups:   IDpitanja [5]
##   IDpitanja naziv.ispita ProsjekRjesenosti
```

```
##       <int>        <chr>            <dbl>
## 1      2896       Ljetni        0.9649123
## 2      2946     Dekanski        0.9215686
## 3      2895       Ljetni        0.8947368
## 4      2941     Dekanski        0.8627451
## 5      2900       Ljetni        0.8421053
```

```
# Najgore riješena pitanja
tail(odgovoriGrupiraniPitanja, 5)
```

```
## # A tibble: 5 x 3
## # Groups:   IDpitanja [5]
##   IDpitanja naziv.ispita ProsjekRjesenosti
##       <int>        <chr>             <dbl>
## 1      2908       Ljetni        0.10526316
## 2      2883       Zimski        0.09523810
## 3      2940     Jesenski        0.08474576
## 4      2746           MI        0.05686546
## 5      2890       Zimski        0.04761905
```

```
# Koliko ima loše riješenih, dobro riješenih i odlično riješenih?
odgovoriGrupiraniPitanja$Rijesenost <- cut(odgovoriGrupiraniPitanja$ProsjekRj
esenosti, breaks = c(0, 0.33, 0.66, 1), labels = c("Losa (<33%)", "Dobra (33%
-66%)", "Odlicna (>66%)"))

group_by(odgovoriGrupiraniPitanja, Rijesenost) %>% summarise(n())
```

```
## # A tibble: 3 x 2
##        Rijesenost `n()`
##            <fctr> <int>
## 1     Losa (<33%)    37
## 2 Dobra (33%-66%)    50
## 3  Odlicna (>66%)    13
```

Opažanja

- Pitanja su najčešće riješena oko 44%, ali su skoro 3 puta češća pitanja koja su loše riješena (<33%) od onih koja su odlično riješena (>66%).

3.3.3 Kako prolaznost na ispitu ovisi o riješenosti najtežeg/najlakšeg pitanja

Na početku sam dobio uvid u velike razlike koje nastaju u težini zadataka na ispitu. Zanima me s kojom sigurnošću mogu tvrditi da će osoba koja je riješila najteži zadatak proći na ispitu ili da će osoba koja nije riješila najlakši zadatak pasti na ispitu.

```
# Odmah stvaram listu studenata koji su prošli na ispitu
studentiProsli <- odgovori %>% group_by(studentID, naziv.ispita)%>% summarise
(Rjesenost=mean(odgovor)) %>% mutate(Prosao=Rjesenost>=0.5) #%>% distinct()

# Znam koji je student prošao koji ispit
studentiProsli$Rjesenost <- NULL
#head(studentiProsli) # MEĐUREZULTAT
```

```
mean(studentiProsli$Prosao)
```

```
## [1] 0.4666283
```

Opažanja

- Prosječna prolaznost studenata je 46,62%. Ovaj podatak mogu, u nastavku, usporediti s prolaznošću studenata koji su riješili najteži ili nisu riješili najlakši zadatak.

```
# ----------------------------
# Pripremam tablice
# ----------------------------

# Najtezih zadataka
najtezaPitanja <- odgovoriGrupiraniPitanja %>% ungroup() %>% group_by(naziv.i
spita) %>% filter(ProsjekRjesenosti == min(ProsjekRjesenosti))
# najtezaPitanja # MEĐUREZULTAT

# Najlaksih zadataka
najlaksaPitanja <- odgovoriGrupiraniPitanja %>% ungroup() %>% group_by(naziv.
ispita) %>% filter(ProsjekRjesenosti == max(ProsjekRjesenosti))
# najlaksaPitanja # MEĐUREZULTAT
```

3.3.3.1 Uspjeh studenata koji su riješili najteže pitanje

```
rijesioNajtezi <- group_by(odgovori, studentID, naziv.ispita) %>%filter( IDpi
tanja%in%  najtezaPitanja$IDpitanja) %>% mutate(RijesioNajtezi = odgovor==1)
%>% select(studentID, naziv.ispita, RijesioNajtezi)
rijesioNajtezi %>% inner_join(studentiProsli,by = c("studentID", "naziv.ispit
a"))%>% group_by(RijesioNajtezi) %>% summarise(Prolaznost = mean(Prosao))
```

```
## # A tibble: 2 x 2
##   RijesioNajtezi Prolaznost
##            <lgl>      <dbl>
## 1          FALSE  0.4195719
## 2           TRUE  0.7788018
```

Opažanja

- Ako znamo da je student točno riješio najteži zadatak, njegova vjerojatnost prolaska poraste s 46,6% na 77,9%.
- Ako znamo da student nije točno riješio najteži zadatak, njegova vjerojatnost prolaska pada s 46,6% na 42,0%.

3.3.3.2 Uspjeh studenata koji su riješili najlakše pitanje

```
rijesioNajlaksi <- group_by(odgovori, studentID, naziv.ispita) %>%filter( IDp
itanja%in%  najlaksaPitanja$IDpitanja) %>% mutate(RijesioNajlaksi = odgovor==
1) %>% select(studentID, naziv.ispita, RijesioNajlaksi)
rijesioNajlaksi %>% inner_join(studentiProsli,by = c("studentID", "naziv.ispi
ta"))%>% group_by(RijesioNajlaksi) %>% summarise(Prolaznost = mean(Prosao))
```

```
## # A tibble: 2 x 2
##   RijesioNajlaksi Prolaznost
##             <lgl>      <dbl>
## 1           FALSE  0.1407186
## 2            TRUE  0.5441595
```

Opažanja

- Ako znamo da je student točno riješio najlakši zadatak, njegova vjerojatnost prolaska poraste s 46,6% na 54,4%.
- Ako znamo da student nije točno riješio najlakši zadatak, njegova vjerojatnost prolaska pada s 46,6% na 14,1%.

Zaključak

Činjenica da je student točno riješio najteži zadatak mu povećava vjerojatnost prolaska za 31,3% što je približno jednako vrijednosti za koju se umanji vjerojatnost prolaska studenta ako znamo da nije točno riješio najlakše pitanje(32,5%).

3.3.4 Uvid u rezultate pojedine vrste ispita

U ovom odlomku promatram razlike u riješenosti pojedinih tipova ispita. Redom sam analizirao riješenosti ispita i razmišljao o objašnjenju istih. Odgovorimo najprije na pitanje: koji su najbolje i najlošije riješeni ispiti?

```
odgovoriGrupiraniIspit <- group_by(odgovori, naziv.ispita)
summarise(odgovoriGrupiraniIspit, ProsjekRjesenosti = mean(odgovor)) %>% arra
nge(desc(ProsjekRjesenosti))
```

```
## # A tibble: 6 x 2
##   naziv.ispita ProsjekRjesenosti
##          <chr>             <dbl>
## 1     Dekanski         0.4970588
## 2           ZI         0.4896869
## 3           MI         0.4796117
## 4       Ljetni         0.4429825
## 5     Jesenski         0.4062147
## 6       Zimski         0.3126984
```

Opažanja

- Postoji velika razlika u rezultatima ispita. Objektivno je dekanski rok najbolje, a zimski rok najgore riješeni ispit.

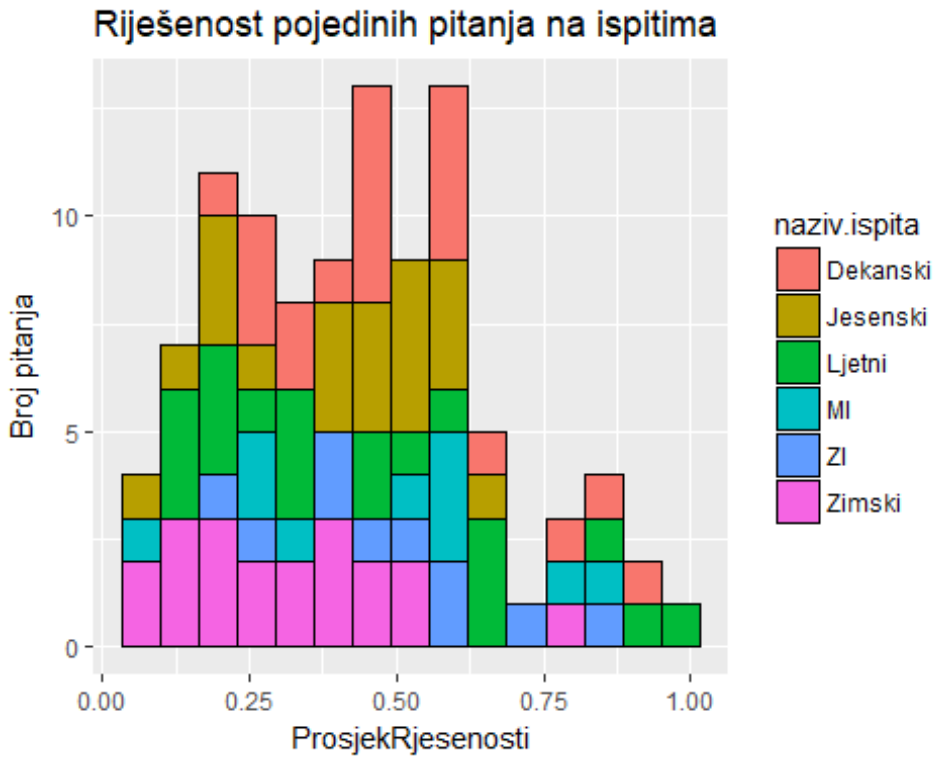
3.4 Grafički uvid u podatke

U ovom odlomku grafički prikazujem podatke kako bih dobio bolji uvid u njih. Zapisujem samo opažanja i pokušavam uočiti bitne varijable i moguće zanimljive kauzalne odnose među njima.

3.4.1 Zastupljenost težih i lakših pitanja

Pokušavam dobiti bolji uvid u zastupljenost težih i lakših pitanja po ispitima gdje lakoću pitanja cijenim kao visoku riješenost.

```
# Prikaz balansiranosti težine pitanja sa svih ispita
ggplot(odgovoriGrupiraniPitanja ,aes(x= ProsjekRjesenosti, fill=naziv.ispita)
) + geom_histogram(bins=15, color = "black") + labs(y ="Broj pitanja", title
="Riješenost pojedinih pitanja na ispitima")
```



Opažanja

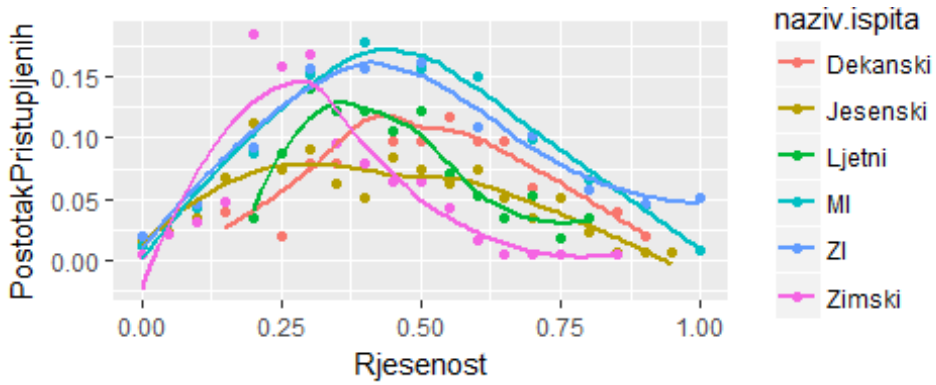
- Ima puno manje lakših pitanja od onih težih.
- Zimski rok ima veliku zastupljenost kod težih zadataka, a ljetni i dekanski imaju više lakših zadataka.

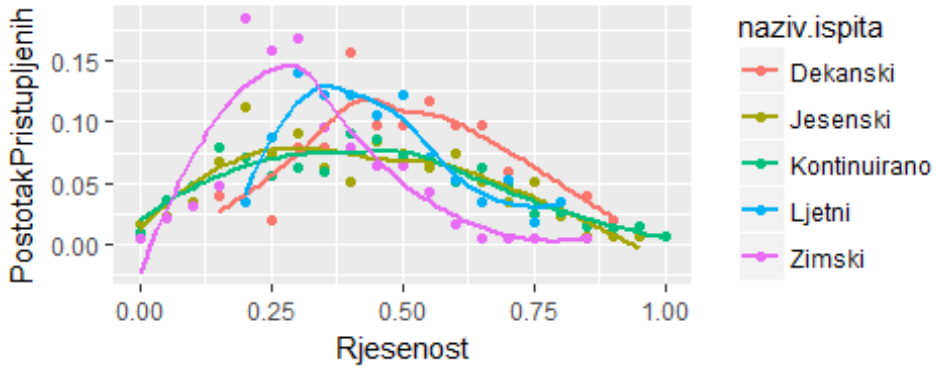
3.4.2 Usporedba razdioba riješenosti ispita

Cilj mi je napraviti linijski graf koji prikazuje riješenosti pojedinih ispita relativno s brojem pristupljenih studenata. Tako mogu usporediti razdiobe riješenosti ispita.

```
# ----------------------------
# Stvaranje tablica koje govore o rezultatima studenata na ispitima
# ----------------------------

# Svi ispiti
sviIspiti <- group_by(odgovori, studentID, naziv.ispita) %>% summarise(Rjesen
```

```
ost = mean(odgovor)) %>% group_by(Rjesenost, naziv.ispita) %>% summarise(Broj
Ucenika= n())

# Dodajem stupac koji mi govori koliko je sveukupno studenata pristupilo ispi
tu...
uspjehSvi <- inner_join(sviIspiti, pristupiloIspitu, "naziv.ispita")
# ... i dijelim dva stupca kako bih dobio relativnu vrijednost riješenosti
uspjehSvi$PostotakPristupljenih <- uspjehSvi$BrojUcenika / uspjehSvi$Pristupi
lo
#head(sviIspiti) # MEĐUREZULTAT



# Svi ispiti uz: Kontinuirano = MI + ZI
# Tablica spojenih ispita već postoji - spojeniIspiti
uspjehSpojeni <- mutate(spojeniIspiti, Rjesenost=Ukupno/20) %>% group_by(nazi
v.ispita, Rjesenost) %>% summarise(BrojUcenika = n())

# Dodajem stupac koji mi govori koliko je sveukupno studenata pristupilo ispi
tu...
uspjehSpojeni <- inner_join(uspjehSpojeni, pristupiloIspituProsireno, "naziv.
ispita")
# ... i dijelim dva stupca kako bih dobio relativnu vrijednost riješenosti
uspjehSpojeni$PostotakPristupljenih <- uspjehSpojeni$BrojUcenika / uspjehSpoj
eni$Pristupilo
# head(uspjehSpojeni) # MEĐUREZULTAT

# ----------------------------
# Stvaranje grafova razdioba
# ----------------------------

gRazdiobeSvi <- ggplot(uspjehSvi, aes(x=Rjesenost, y = PostotakPristupljenih,
color=naziv.ispita)) + geom_point() + stat_smooth( method="loess", aes(group=
naziv.ispita), se = FALSE)
gRazdiobeSpojeni <- ggplot(uspjehSpojeni, aes(x=Rjesenost, y = PostotakPristu
pljenih, color=naziv.ispita)) + geom_point() + stat_smooth( method="loess", a
es(group=naziv.ispita), se = FALSE)

grid.arrange(gRazdiobeSvi, gRazdiobeSpojeni, nrow=2, ncol=1)
```





Opažanja

- Zimski rok ima najizraženiji vrh grafa i to pri niskoj riješenosti.
- S druge strane, jesenski rok ima najujednačenije vrijednosti riješenosti.
- Najsličnija Gaussovoj krivulji je krivulja za MI, a nakon nje i za ZI, dok kontinuirana nastava (kao zbroj MI i ZI) ima vrlo ujednačenu razdiobu riješenosti

3.5 Teze

Sad već imam dobar uvid u podatke i pokušavam identificirati kauzalne odnose među bitnim varijablama i dokazati njihovu uvjerljivost u obliku teza. Na početku teze objašnjavam što i kako ju testiram, a na kraju dolazim do zaključka o valjanosti. Ponovo naglašavam da su zaključci više slutnje na temelju opažanja na podatcima jedne akademske godine kojoj fali prava matematička analiza povjerljivosti.

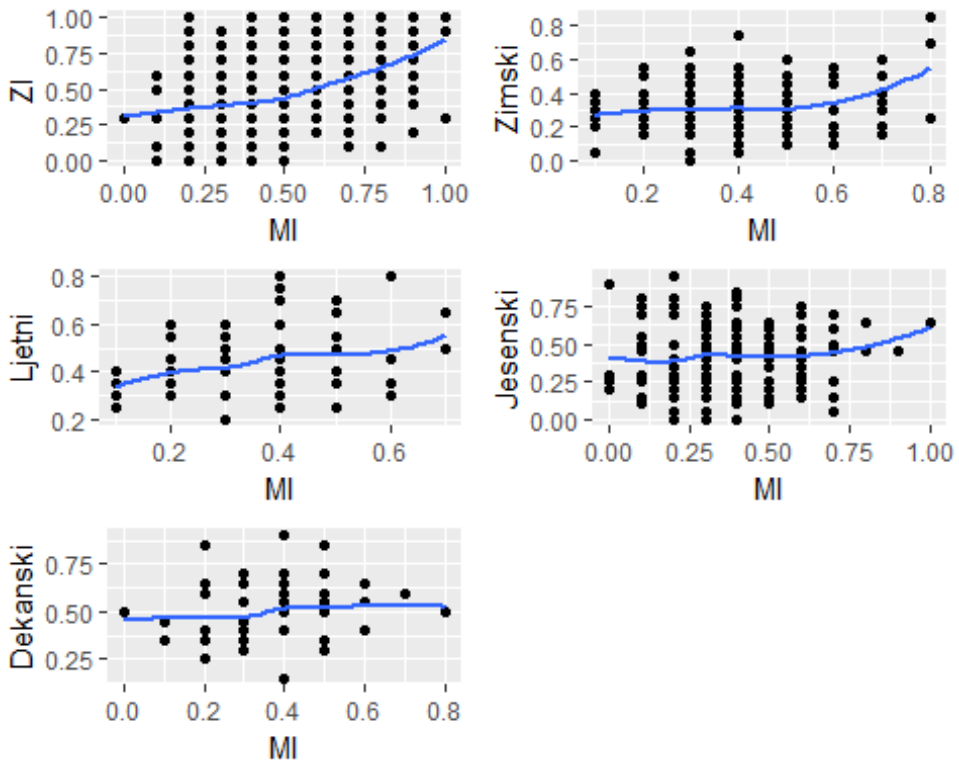
Bitne varijable koje dovodim u kauzalne odnose:

1. Rezultat na ispitu
2. Pristupanje ispitu
3. Riješenost najtežeg/ najlakšeg pitanja

3.5.1 Rezultati ispita ovise o rezultatima na MI

Testiram ovisi li rezultat pojedine vrste ispita o rezultatima na MI, ali pritom promatram skup studenata koji su pristupili oba ispita.

```
ovisnostMIZI <- spojeniSpreadSviIspiti %>% filter(!is.na(ZI) & !is.na(MI)) %>
% select(studentID, MI,ZI)
ovisnostMIZR <- spojeniSpreadSviIspiti %>% filter(!is.na(Zimski) & !is.na(MI)
) %>% select(studentID, MI,Zimski)
ovisnostMILJR <- spojeniSpreadSviIspiti %>% filter(!is.na(Ljetni) & !is.na(MI
)) %>% select(studentID, MI,Ljetni)
ovisnostMIJR <- spojeniSpreadSviIspiti %>% filter(!is.na(Jesenski) & !is.na(M
I)) %>% select(studentID, MI,Jesenski)
ovisnostMIDR <- spojeniSpreadSviIspiti %>% filter(!is.na(Dekanski) & !is.na(M
I)) %>% select(studentID, MI,Dekanski)

grafMIZI <- ggplot(ovisnostMIZI, aes(x=MI, y = ZI)) + geom_point() + stat_smo
oth( method="loess", se = FALSE)
grafMIZR <- ggplot(ovisnostMIZR, aes(x=MI, y = Zimski)) + geom_point() + stat
_smooth( method="loess", se = FALSE)
grafMILJR <- ggplot(ovisnostMILJR, aes(x=MI, y = Ljetni)) + geom_point() + st
at_smooth( method="loess", se = FALSE)
grafMIJR <- ggplot(ovisnostMIJR, aes(x=MI, y = Jesenski)) + geom_point() + st
at_smooth( method="loess", se = FALSE)
grafMIDR <- ggplot(ovisnostMIDR, aes(x=MI, y = Dekanski)) + geom_point() + st
at_smooth( method="loess", se = FALSE)

grid.arrange(grafMIZI, grafMIZR, grafMILJR, grafMIJR, grafMIDR, nrow=3, ncol=
2)
```



Opažanja

- Očekivano je ovisnost rezultata na rokovima o rezultatu MI sve manja što je ispit dalje od MI. Dekanski rok nema gotovo nikakvu pozitivnu korelaciju s rezultatima MI.

Zaključak

Na temelju boljih rezultata MI možemo pretpostaviti da će student bolje proći na vremenski bližim ispitima. S druge strane, za jesenski i dekanski rok, koji su vremenski daleko, nam informacija o rezultatu na MI gotovo ništa ne znači.

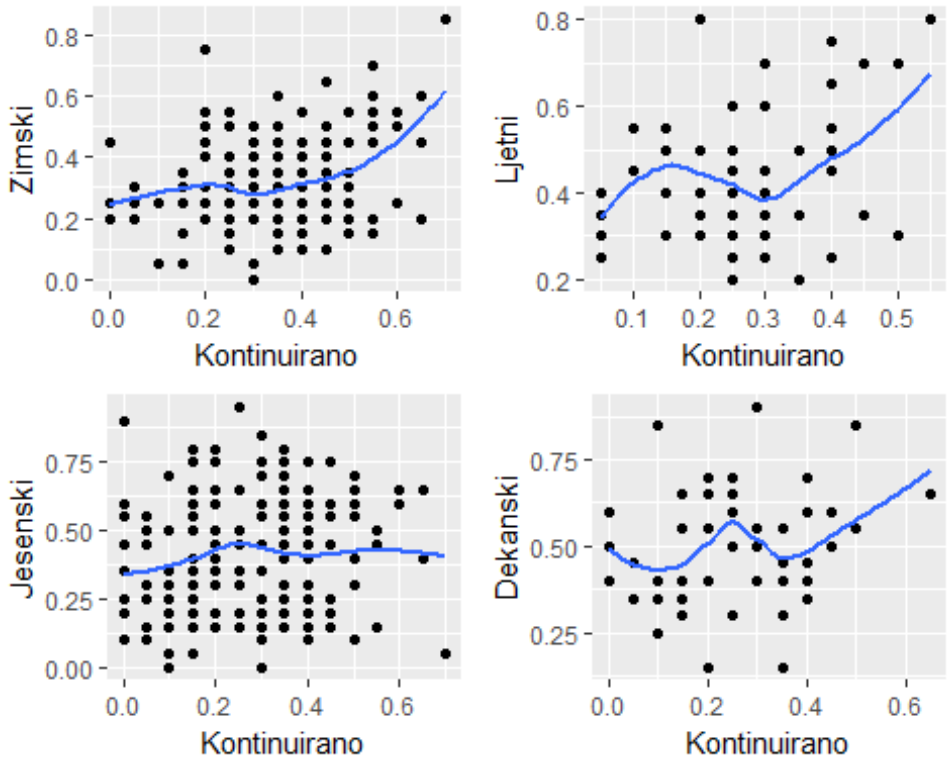
Dakle, teza je točna, ali gubi na značaju za vremenski udaljenije ispite.

3.5.2 Rezultati ispita ovise o rezultatima u kontinuiranoj nastavi

Već smo utvrdili da rezultati na MI pozitivno koreliraju s rezultatima na rokovima, iako je veza slabija što je rok vremenski dalji. U ovom poglavlju ću pokušati istražiti odnos rezultata rokova s rezultatima na kontinuiranoj nastavi(MI + ZI). Pri početnom uvidu u podatke, utvrdili smo da postoji veliki broj studenata koji nisu pristupili jednoj od ove dvije provjere. Studente koji nisu pristupili ispitu ću tretirati kao da su postigli 0 bodova na istom.

Očekujem da će studenti s boljim uspjehom u kontinuiranoj nastavi bolje prolaziti na rokovima jer su pokazali više truda i volje kroz semestar.

```
# Priprema podataka - studentima koji nisu pristupili se pridodaje 0 bodova
spojeniSpread <- mutate(spojeniSpread, Kontinuirano = ifelse(is.na(Kontinuira
no), 0, Kontinuirano))

ovisnostKoZR <- spojeniSpread %>% filter(!is.na(Zimski) & !is.na(Kontinuirano
)) %>% select(studentID, Kontinuirano,Zimski)
ovisnostKoLJR <- spojeniSpread %>% filter(!is.na(Ljetni) & !is.na(Kontinuiran
o)) %>% select(studentID, Kontinuirano,Ljetni)
ovisnostKoJR <- spojeniSpread %>% filter(!is.na(Jesenski) & !is.na(Kontinuira
no)) %>% select(studentID, Kontinuirano,Jesenski)
ovisnostKoDR <- spojeniSpread %>% filter(!is.na(Dekanski) & !is.na(Kontinuira
no)) %>% select(studentID, Kontinuirano,Dekanski)

# Priprema grafova
grafKoZR <- ggplot(ovisnostKoZR, aes(x=Kontinuirano, y = Zimski)) + geom_poin
t() + stat_smooth( method="loess", se = FALSE)
grafKoLJR <- ggplot(ovisnostKoLJR, aes(x=Kontinuirano, y = Ljetni)) + geom_po
int() + stat_smooth( method="loess", se = FALSE)
grafKoJR <- ggplot(ovisnostKoJR, aes(x=Kontinuirano, y = Jesenski)) + geom_po
int() + stat_smooth( method="loess", se = FALSE)
grafKoDR <- ggplot(ovisnostKoDR, aes(x=Kontinuirano, y = Dekanski)) + geom_po
int() + stat_smooth( method="loess", se = FALSE)

# Završni prikaz
grid.arrange(grafKoZR, grafKoLJR, grafKoJR, grafKoDR, nrow=2, ncol=2)
```



Opažanja

1. Zimski rok

- Studenti koji su kontinuirano skupili manje ili jednako od 15% na ispitima, imaju ravno 0% prolaznosti na zimskom roku. Čak i oni koji nisu uopće izlazili kontinuirano konzistentno padaju na zimskom roku.
- Iznenađujuće su uspješni studenti koji su ostvarili 20-25% na kontinuiranoj nastavi.
- Također, iznenađujuće pada prolaznost studenata koji su u skoro najvišoj kategoriji uspješnosti u kontinuiranoj nastavi. Radi se o studentima s uspjehom od oko 60%. Moguće da su to su studenti koji nisu bili zadovoljni ocjenom pa su se malo precijenili.

2. Ljetni rok

- Pristupili su samo studenti koji su imali više od 0 bodova kontinuirano.
- Opet se javlja neočekivana povećana uspješnost, ali ovaj put na 15% kontinuiranih bodova. Ispada da su osobe s 15% kontinuiranih bodova riješile uspješnije ispit od osoba koje su imale 20-35% kontinuirano.

3. Jesenski rok

- Krivulja je približno ujednačena.
- Nastaje nagli pad prolaska na jesenskom roku među studentima s najviše ostvarenim brojem bodova kontinuirano(>=65%).
- Studenti koji su imali prolaz na kontinuiranoj nastavi su jako loše riješili ispit.

4. Dekanski rok

- Neočekivano dobro prolaze studenti koji su kontinuirano postigli oko 25% bodova.
- Prolaze svi studenti koji su imali strogo iznad 40% kontinuirano skupljenih bodova.
- Uočavam pad prolaznosti kod studenata kategorije 30-40% bodova kontinuirano.

Zaključak

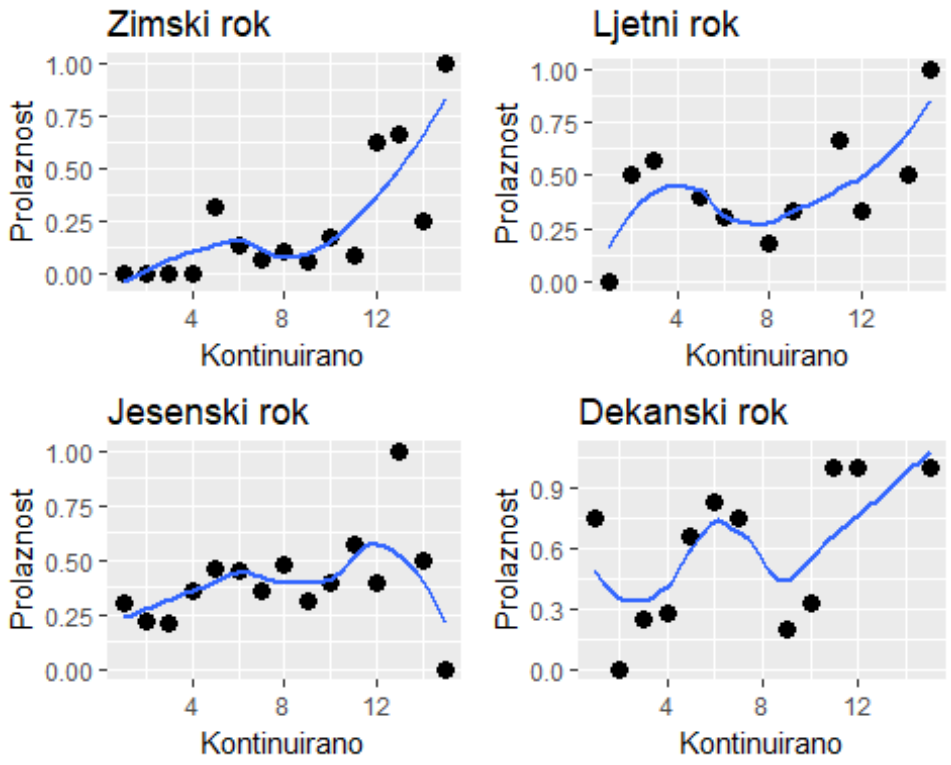
Krivulja ovisnosti rezultata ispita o rezultatima u kontinuiranoj nastavi poprima specifičan oblik za svaki ispit.
Krivulju karakterizira lošiji rezultati studenata koji su osvojili 30-35% kontinuiranih bodova. S druge strane, ispite su neočekivano dobro riješili studenti koji su osvoji 15-25% kontinuiranih bodova. Od uzorka najviše odskače krivulja jesenskog roka. Zbog ujednačenosti krivulje jesenskog roka, zaključujem da broj bodova na jesenskom roku gotovo uopće ne ovisi o broju bodova na kontinuiranoj nastavi.

3.5.3 Ovisnost prolaznosti rokova o uspjehu na kontinuiranoj nastavi

Prolaznost i riješenost ispita često nisu u potpunom skladu zbog visokog postotka studenata koji imaju riješenost ispita netom ispod prolaza(40-45%). U ovom poglavlju pokušavam vidjeti hoće li se prolaznost ispita ponašati jednako kao riješenost ispita u ovisnosti o uspjehu na kontinuiranoj nastavi. Da bih bolje prikazao prolaznost na ispitima, podijelio sam studente u kategorije prema njihovom uspjehu na kontinuiranoj nastavi.

```
# ----------------------------
# Priprema kategorija
# ----------------------------
spojeniSpreadZR <- spojeniSpread %>% select(studentID, Kontinuirano, Zimski)
%>% filter(!is.na(Zimski))
spojeniSpreadZR$KategorijaKontinuiranog <- cut(spojeniSpreadZR$Kontinuirano,
15)
spojeniSpreadZR$ProsliRok <- spojeniSpreadZR$Zimski >= 0.5
kategorijeZR <- group_by(spojeniSpreadZR, KategorijaKontinuiranog) %>% summar
ise(UspjehZR = mean(Zimski), ProlaznostZR = mean(ProsliRok))


spojeniSpreadLJR <- spojeniSpread %>% select(studentID, Kontinuirano, Ljetni)
%>% filter(!is.na(Ljetni))
spojeniSpreadLJR$KategorijaKontinuiranog <- cut(spojeniSpreadLJR$Kontinuirano
, 15)
spojeniSpreadLJR$ProsliRok <- spojeniSpreadLJR$Ljetni >= 0.5
kategorijeLJR <- group_by(spojeniSpreadLJR, KategorijaKontinuiranog) %>% summ
arise(UspjehLJR = mean(Ljetni), ProlaznostLJR = mean(ProsliRok))

spojeniSpreadJR <- spojeniSpread %>% select(studentID, Kontinuirano, Jesenski
) %>% filter(!is.na(Jesenski))
spojeniSpreadJR$KategorijaKontinuiranog <- cut(spojeniSpreadJR$Kontinuirano,
15)
spojeniSpreadJR$ProsliRok <- spojeniSpreadJR$Jesenski >= 0.5
kategorijeJR <- group_by(spojeniSpreadJR, KategorijaKontinuiranog) %>% summar
```

```
ise(UspjehJR = mean(Jesenski), ProlaznostJR = mean(ProsliRok))


spojeniSpreadDR <- spojeniSpread %>% select(studentID, Kontinuirano, Dekanski
) %>% filter(!is.na(Dekanski))
spojeniSpreadDR$KategorijaKontinuiranog <- cut(spojeniSpreadDR$Kontinuirano,
15)
spojeniSpreadDR$ProsliRok <- spojeniSpreadDR$Dekanski >= 0.5
kategorijeDR <- group_by(spojeniSpreadDR, KategorijaKontinuiranog) %>% summar
ise(UspjehDR = mean(Dekanski), ProlaznostDR = mean(ProsliRok))

# ----------------------------
# Priprema grafova
# ----------------------------
gProlaznostKoZR <- ggplot(kategorijeZR, aes(x=as.numeric(KategorijaKontinuira
nog), y = ProlaznostZR)) + geom_point(size=3) + stat_smooth( method="loess",
se = FALSE) + labs(x="Kontinuirano", title ="Zimski rok", y="Prolaznost")

gProlaznostKoLJR <- ggplot(kategorijeLJR, aes(x=as.numeric(KategorijaKontinui
ranog), y = ProlaznostLJR)) + geom_point(size=3) + stat_smooth( method="loess
", se = FALSE) + labs(x="Kontinuirano", title ="Ljetni rok", y="Prolaznost")

gProlaznostKoJR <- ggplot(kategorijeJR, aes(x=as.numeric(KategorijaKontinuira
nog), y=ProlaznostJR)) + geom_point(shape=16, size=3) + labs(x="Kontinuirano"
,y ="Prolaznost", title ="Jesenski rok") + stat_smooth(method="loess", se = F
ALSE)

gProlaznostKoDR <- ggplot(kategorijeDR, aes(x=as.numeric(KategorijaKontinuira
nog), y=ProlaznostDR)) + geom_point(shape=16, size=3) + labs(x="Kontinuirano"
, y ="Prolaznost", title="Dekanski rok") + stat_smooth( method="loess", se =
FALSE)


# Konačan prikaz
grid.arrange(gProlaznostKoZR, gProlaznostKoLJR, gProlaznostKoJR, gProlaznostK
oDR, nrow=2, ncol=2)
```



Opažanja

- Na krivuljama prolaza se još jasnije vidi prethodno opisani uzorak kojeg karakterizira lošiji uspjeh kategorija oko 30-35% uspjeha kontinuirano i boljeg uspjeha kategorija koje sadrže studente koji su postigli oko 15-25% kontinuirano.
- Također se jasnije vidi odstupanje od uzorka na jesenskom roku u području studenata koji su već imali osigurani prolaz kontinuirano.

Zaključak

Promatrajući ovisnost rezultata ispita o rezultatima na kontinuiranoj nastavi, ne mogu doći do jedinstvenog zaključka kao što sam mogao u slučaju promatranja samo rezultata MI. Korelacija nije jedinstveno pozitivna. Ipak, primjećujem drugačiji uzorak. Svaki ispit ima povećanu prolaznost oko studenata koji imaju 15-25% i smanjenu prolaznost oko studenata koji imaju 30-35% osvojenih bodova na kontinuiranoj nastavi.

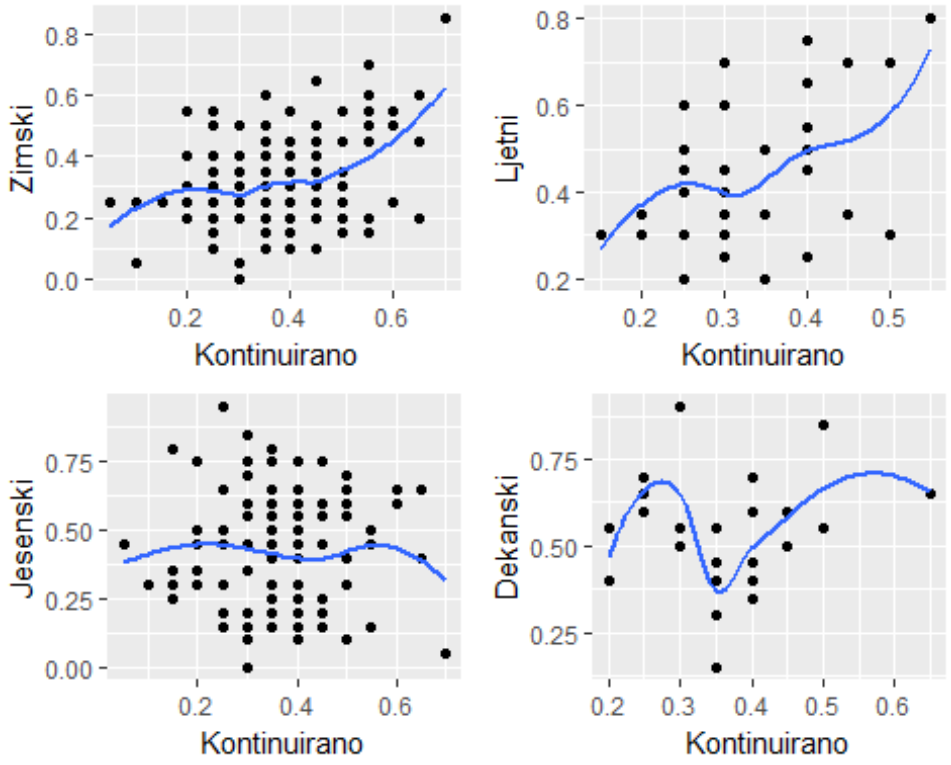
3.5.4 Prvi maksimum krivulje nastaje zbog izlaznosti na ZI

Razmišljao sam o mogućim razlozima nastanka prvog maksimuma i zanima me: Je li upravo to kategorija studenata koji nisu izlazili na ZI? Možda su to studenti koji su dobro riješili MI, ali nisu stigli naučiti za ZI pa su se iskazali na roku? U ovom poglavlju me zanimaju samo rezultati oko spomenutog prvom maksimuma. Tezu mogu testirati tako da provjerim:

2. Nastaje li prvi maksimum ako uklonim studente koji nisu pristupili obje provjere
3. Postoji li razlika u rezultatima studenata koji jesu i nisu pristupili ZI

3.5.4.1 Uspješnost samo studenata koji su pristupili obje provjere

```
# Što ako promatram samo studente koji su izasli na oba ispita kontinuirano?
cista <- spojeniSpreadSviIspiti %>% filter(!is.na(ZI) & !is.na(MI)) %>% mutat
e(Kontinuirano = (MI+ZI)/2)
ovisnostKoZR <- cista %>% filter(!is.na(Zimski) & !is.na(Kontinuirano)) %>% s
elect(studentID, Kontinuirano,Zimski)
ovisnostKoLJR <- cista %>% filter(!is.na(Ljetni) & !is.na(Kontinuirano)) %>%
select(studentID, Kontinuirano,Ljetni)
ovisnostKoJR <- cista %>% filter(!is.na(Jesenski) & !is.na(Kontinuirano)) %>%
select(studentID, Kontinuirano,Jesenski)
ovisnostKoDR <- cista %>% filter(!is.na(Dekanski) & !is.na(Kontinuirano)) %>%
select(studentID, Kontinuirano,Dekanski)

grafKoZR <- ggplot(ovisnostKoZR, aes(x=Kontinuirano, y = Zimski)) + geom_poin
t() + stat_smooth( method="loess", se = FALSE)
grafKoLJR <- ggplot(ovisnostKoLJR, aes(x=Kontinuirano, y = Ljetni)) + geom_po
int() + stat_smooth( method="loess", se = FALSE)
grafKoJR <- ggplot(ovisnostKoJR, aes(x=Kontinuirano, y = Jesenski)) + geom_po
int() + stat_smooth( method="loess", se = FALSE)
grafKoDR <- ggplot(ovisnostKoDR, aes(x=Kontinuirano, y = Dekanski)) + geom_po
int() + stat_smooth( method="loess", se = FALSE)

grid.arrange(grafKoZR, grafKoLJR, grafKoJR, grafKoDR, nrow=2, ncol=2)
```



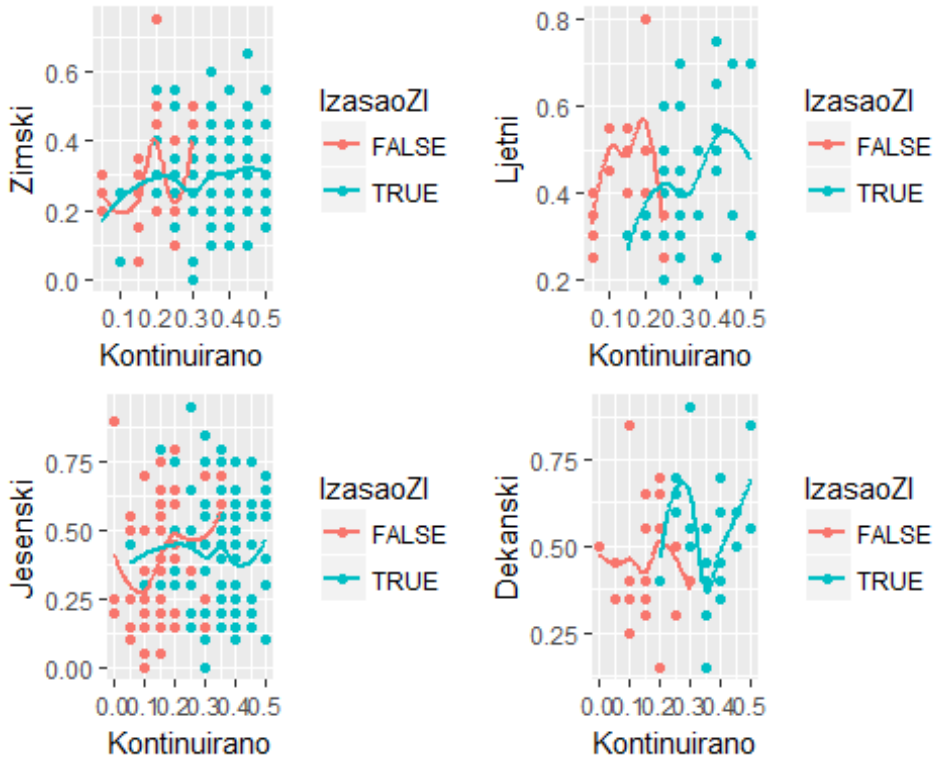
Opažanja

- Svejedno nastaje prvi maksimum oko studenata s 20-25% uspjeha kontinuirano.

- Krivulja jesenskog roka je sada još ujednačenija.

3.5.4.2 Izravna usporedba rezultata studenata po pristupanju ZI

```
spojenoSpreadSve <- spojeniSpreadSviIspiti %>% mutate(Kontinuirano = (ifelse(
is.na(MI), 0, MI) + ifelse(is.na(ZI), 0, ZI)) /2)
ovisnostMIZR <- spojenoSpreadSve %>% filter(!is.na(Zimski) & !is.na(MI) & Kon
tinuirano <= 0.5) %>% select(studentID, Kontinuirano,Zimski, ZI) %>% mutate(I
zasaoZI=!is.na(ZI))
ovisnostMILJR <- spojenoSpreadSve %>% filter(!is.na(Ljetni) & !is.na(MI) & Ko
ntinuirano <= 0.5) %>% select(studentID, Kontinuirano,Ljetni, ZI)%>% mutate(I
zasaoZI=!is.na(ZI))
ovisnostMIJR <- spojenoSpreadSve %>% filter(!is.na(Jesenski) & !is.na(MI) & K
ontinuirano <= 0.5) %>%select(studentID, Kontinuirano,Jesenski,ZI)%>% mutate(
IzasaoZI=!is.na(ZI))
ovisnostMIDR <- spojenoSpreadSve %>% filter(!is.na(Dekanski) & !is.na(MI) & K
ontinuirano <= 0.5) %>%select(studentID, Kontinuirano,Dekanski,ZI)%>% mutate(
IzasaoZI=!is.na(ZI))

grafMIZR <- ggplot(ovisnostMIZR, aes(x=Kontinuirano, y = Zimski, color=Izasao
ZI)) + geom_point() + stat_smooth( method="loess", se = FALSE, aes(group=Izas
aoZI))
grafMILJR <- ggplot(ovisnostMILJR, aes(x=Kontinuirano, y = Ljetni, color=Izas
aoZI)) + geom_point() + stat_smooth( method="loess", se = FALSE, aes(group=Iz
asaoZI))
grafMIJR <- ggplot(ovisnostMIJR, aes(x=Kontinuirano, y = Jesenski, color=Izas
aoZI)) + geom_point() + stat_smooth( method="loess", se = FALSE, aes(group=Iz
asaoZI))
grafMIDR <- ggplot(ovisnostMIDR, aes(x=Kontinuirano, y = Dekanski, color=Izas
aoZI)) + geom_point() + stat_smooth( method="loess", se = FALSE, aes(group=Iz
asaoZI))

grid.arrange(grafMIZR, grafMILJR, grafMIJR, grafMIDR, nrow=2, ncol=2)
```



Opažanja

- Na ljetnom roku bolje prolaze studenti koji nisu izlazili na ZI, no, sveukupno gledajući, nema korelacije između pristupanja završnom ispitu i uspjeha na roku.

Zaključak

Činjenica o pristupanju studenta na ZI ne utječe na njegove bolje ili lošije rezultate na roku. Dakle, teza nije valjana.

3.5.5 Na dekanskom bolje prolaze studenti koji su pristupili jesenskom roku

Prikaz rezultata dekanskog roka me zainteresirao za pitanje: prolaze li bolje na dekanskom roku oni studenti koji su odlučili izlaziti na ljetni ili na jesenski rok? Naime, moglo bi se pomisliti da je studentima koji su pristupali jesenskom roku još uvijek svježe gradivo jer su ga učili tek tjedan dana ranije. Testirajmo ovu tezu.

```
izasliJR <- filter(odgovori, naziv.ispita == "Jesenski")
izasliLJR <- filter(odgovori, naziv.ispita == "Ljetni")

spojeniSpreadDR$PristupioJesenskom <- spojeniSpreadDR$studentID %in% izasliJR
$studentID
spojeniSpreadDR$PristupioLjetnom <- spojeniSpreadDR$studentID %in% izasliLJR$
studentID

spojeniSpreadDR$TakoderPristupio <- if_else(spojeniSpreadDR$PristupioJesensko
m, "Jesenskom", ifelse(spojeniSpreadDR$PristupioLjetnom, "Ljetnom", "Ni JR ni
```

```
LJR"))
# head(spojeniSpreadDR) # MEĐUREZULTAT

group_by(spojeniSpreadDR, TakoderPristupio) %>% summarise(ProlaznostDR=mean(P
rosliRok))

## # A tibble: 3 x 2
##   TakoderPristupio ProlaznostDR
##              <chr>        <dbl>
## 1        Jesenskom    0.5789474
## 2          Ljetnom    0.4000000
## 3     Ni JR ni LJR    0.3333333
```

Opažanja

- Vjerojatnost prolaska na dekanskom roku je značajno veća za studente koji su izašli na jesenskom roku u usporedbi sa studentima koji su odlučili pristupiti ljetnom roku
- Studenti koji nisu pristupili ni jednom roku imaju najmanju vjerojatnost za prolaz.

Zaključak

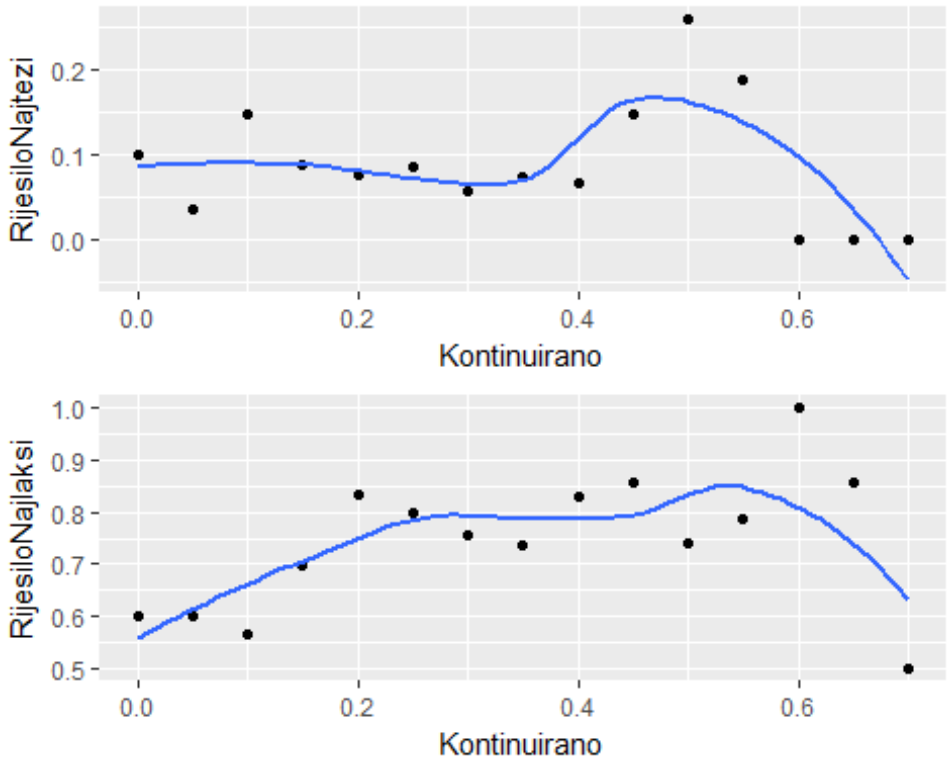
Teza je valjana. Ne mogu provjeriti je li valjana baš zato što su studenti koji su pristupali jesenskom roku još uvijek zagrijani za predmet, ali rezultati daju zaključiti da su na dekanskom roku uspješniji oni studenti koji su prethodno pristupili jesenskom roku.

3.5.6 Riješenost najtežeg/najlakšeg pitanja na roku ovisi o rezultatima na kontinuiranoj nastavi

Najteža pitanja su najčešća ona pitanja koja se posebno obrađuju na nastavi, a studenti koji su aktivno pratili nastavu najčešće imaju više ostvarenih bodova kontinuirano. Testirajmo ovu logiku.

```
# Relacija koja mi govori koji su studenti riješili najteža pitanja na rokovi
ma
uspjehKontinuirano <- spojeniSpread %>% select(studentID, Kontinuirano) %>% m
utate(Kontinuirano=ifelse(is.na(Kontinuirano), 0, Kontinuirano))

rijesioNajteziRok <- filter(rijesioNajtezi, naziv.ispita != "MI" & naziv.ispi
ta!="ZI")
# rijesioNajteziRok  # MEĐUREZULTAT

rijesioNajlaksiRok <- filter(rijesioNajlaksi, naziv.ispita != "MI" & naziv.is
pita!="ZI")
# rijesioNajlaksiRok # MEĐUREZULTAT

rijesioNajteziKo <- inner_join(rijesioNajteziRok, uspjehKontinuirano, by="stu
dentID")
rijesioNajteziKo <- rijesioNajteziKo %>% group_by(Kontinuirano) %>% summarise
(RijesiloNajtezi = mean(RijesioNajtezi))
#rijesioNajteziKo # MEĐUREZULTAT
```

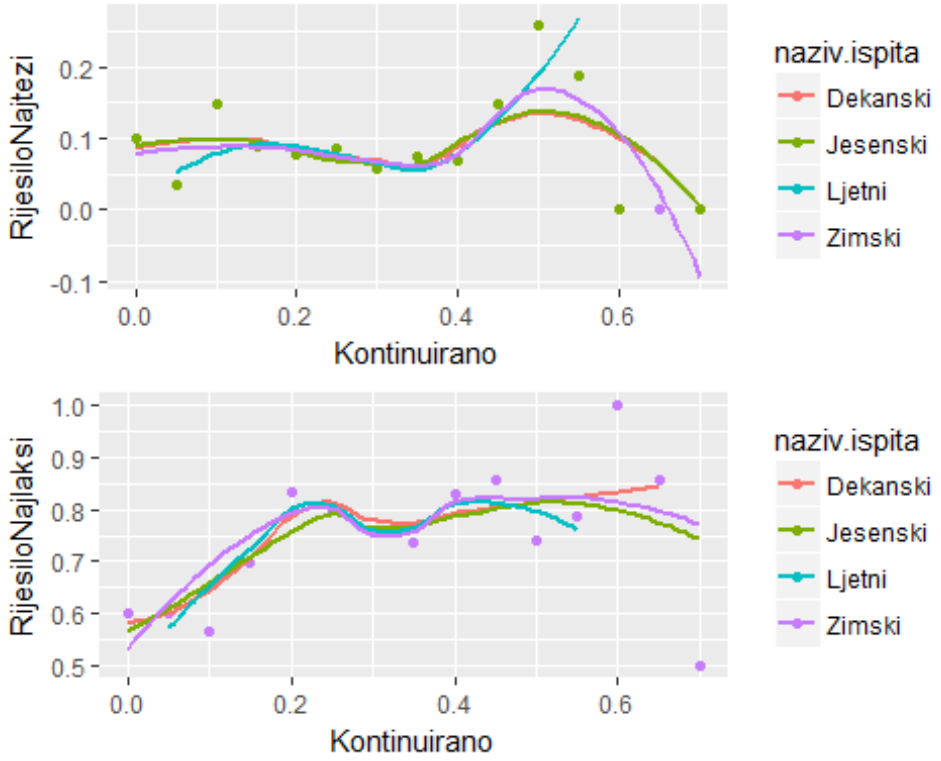
```
rijesioNajlaksiKo <- inner_join(rijesioNajlaksiRok, uspjehKontinuirano, by="studentID")
rijesioNajlaksiKo <- rijesioNajlaksiKo %>% group_by(Kontinuirano) %>% summarise(RijesiloNajlaksi = mean(RijesioNajlaksi))
# rijesioNajlaksiKo # MEĐUREZULTAT

# Grafovi
gNajteziKo <- ggplot(rijesioNajteziKo, aes(x=Kontinuirano, y = RijesiloNajtezi)) + geom_point() + stat_smooth( method="loess", se = FALSE)
gNajlaksiKo <- ggplot(rijesioNajlaksiKo, aes(x=Kontinuirano, y = RijesiloNajlaksi)) + geom_point() + stat_smooth( method="loess", se = FALSE)

# Konačan prikaz
grid.arrange(gNajteziKo, gNajlaksiKo, nrow=2, ncol=1)
```



Opažanja

- Ne mogu potvrditi striktno pozitivnu korelaciju.
- Kod obje vrste pitanja neočekivano loše prolaze studenti s najviše postignutih bodova kontinuirano.
- Svaka kategorija kontinuiranog uspjeha je riješila preko 50% najlakše pitanje.
- Nijedna kategorija nije riješila više od 26% najteže pitanje.

```r
# Relacija koja mi govori koji su studenti riješili najlakša pitanja na rokov
ima
# uspjehKontinuirano <- spojeniSpread %>% select(studentID, Kontinuirano) %>%
mutate(Kontinuirano=ifelse(is.na(Kontinuirano), 0, Kontinuirano))

#rijesioNajteziRok <- filter(rijesioNajtezi, naziv.ispita != "MI" & naziv.isp
ita!="ZI")
# rijesioNajteziRok  # MEĐUREZULTAT

# rijesioNajlaksiRok <- filter(rijesioNajlaksi, naziv.ispita != "MI" & naziv.
ispita!="ZI")
# rijesioNajlaksiRok # MEĐUREZULTAT

rijesioNajteziKoGrupiran <- inner_join(rijesioNajteziRok, uspjehKontinuirano,
by="studentID")
rijesioNajteziKoGrupiran <- rijesioNajteziKoGrupiran %>% group_by(Kontinuiran
o) %>% mutate(RijesiloNajtezi = mean(RijesioNajtezi))
# rijesioNajteziKoGrupiran # MEĐUREZULTAT

rijesioNajlaksiKoGrupiran <- inner_join(rijesioNajlaksiRok, uspjehKontinuiran
o, by="studentID")
rijesioNajlaksiKoGrupiran <- rijesioNajlaksiKoGrupiran %>% group_by(Kontinuir
ano) %>% mutate(RijesiloNajlaksi = mean(RijesioNajlaksi))
# rijesioNajlaksiKoGrupiran # MEĐUREZULTAT

# Grafovi
gNajteziKoGrupirani <- ggplot(rijesioNajteziKoGrupiran, aes(x=Kontinuirano, y
= RijesiloNajtezi, color=naziv.ispita)) + geom_point() + stat_smooth( method=
"loess", se = FALSE, aes(group=naziv.ispita))
gNajlaksiKoGrupiran <- ggplot(rijesioNajlaksiKoGrupiran, aes(x=Kontinuirano,
y = RijesiloNajlaksi,color=naziv.ispita)) + geom_point() + stat_smooth( metho
d="loess", se = FALSE,aes(group=naziv.ispita))

# Konačan prikaz
grid.arrange(gNajteziKoGrupirani, gNajlaksiKoGrupiran, nrow=2, ncol=1)
```



Opažanja

- Svaki ispit ima jednaku distribuciju riješenosti najtežeg/najlakšeg pitanja u ovisnosti o rezultatima na kontinuiranoj nastavi.

Zaključak

Uspješnost studenata na najlakšim/najtežim pitanjima ispita ne ovisi o njihovom uspjehu u kontinuiranoj nastavi. Zaključak je jednak za sve vrste ispita.

4. Zaključak

Kroz analizu podataka o rezultatima ispita na predmetu "Osnove elektrotehnike" iz 2012./2013. akademske godine, uočio sam sljedeće bitne varijable: rezultati na pojedinom ispitu, izlazak na pojedini ispit i riješenost najtežeg/najlakšeg zadatka.

Proučavajući njihove međusobne odnose, došao sam do više zanimljivih opažanja od kojih ističem:

- Bolji rezultat studenta na MI uglavnom znači i bolji rezultat na roku. Relacija slabi za ispite koji su vremenski udaljeniji pa je tako relacija za dekanski rok gotovo nepostojeća.
- Studenti koji su ostvarili 15-25% bodova na kontinuiranoj nastavi konzistentno ostvaruju bolje rezultate na ispitima od studenata koji su ostvarili 30-35%.

- Bolje rezultate na dekanskom roku ostvaruju oni studenti koji su prethodno izašli na jesenski od onih koji su izašli na ljetni rok.
- Rokove jednako dobro rješavaju studenti koji jesu i nisu izlazili na ZI.

Analizu bih dalje mogao produbiti kroz izradu konkretnih matematičkih modela za naslućene relacije i kroz korištenje podataka za druge akademske godine.