

R

μ

Vežbe 8

Povezanost veličina



- Kada se posmatra povezanost dve veličine, često se uočava da promena jedne od njih dovodi do promene druge
- Ova povezanost može da se razlikuje po:
 - smeru
 - jačini
 - obliku povezanosti
- Pozitivan smer povezanosti – povećanje jedne od veličina dovodi do povećanja druge veličine
- Negativan smer povezanosti – povećanje jedne od veličina dovodi do smanjenja druge veličine

Povezanost veličina



- Jačina povezanosti – koliko promena jedne veličine utiče na drugu veličinu
- Funkcionalna veza je najjača – svakoj vrednosti jedne veličine odgovara **tačno jedna** vrednost druge veličine
- Veze koje su slabije od funkcionalne se zovu **stohastičke** ili **korelativne** i podložne su manjim ili većim odstupanjima
- Postoje različiti oblici povezanosti: linearan, kvadratni, eksponencijalni...

Linearna povezanost



- Linearni oblik povezanosti predstavlja linearnu vezu između dve veličine
- Za ovakav oblik povezanosti se korišćenjem **linearne regresije** nalazi **regresiona prava**
- Ova prava (oblika $y = a + bx$) je ona prava koja ispunjava kriterijum da je zbir kvadrata „vertikalnih rastojanja“ tačaka koje predstavljaju zabeležene vrednosti od te prave što manji



Linearna regresija

```
library(UsingR)

data(home)
attach(home)

plot(old, new)

#funkcija lm vraca koeficijente a i b regresione prave
koeficijenti = lm(new ~ old)

#funkcija abline ce na osnovu dobijenih koeficijenata da iscrta liniju
na grafiku
abline(koeficijenti)

detach(home)
```



Linearna regresija

```
#funkcija simple.lm ce odraditi izracunati parametre regresione prave i  
iscrtati odgovarajuci grafik  
simple.lm(old, new)  
  
#direktan pristup koeficijentima tegresione prave  
lm = simple.lm(old, new)  
coef(lm)  
  
#ovde ce, osim prikaza grafika vrednosti sa regresionom pravom biti  
prikazani i dodatni grafici koji  
#predstavljaju analizu "ostataka" tj. razlike stvarnih vrednosti od  
onih koje prikazuje regresiona prava  
simple.lm(old, new, show.residuals = TRUE)  
  
#funkcija resid ce da izdvoji samo razlike stvarnih vrednosti od onih  
koje su dobijene na regresionoj pravoj  
ostaci = resid(lm)  
plot(ostaci)
```

Testiranje jačine regresione veze



- Vrlo je bitno znati koliko je neka regresiona veza jaka – koliki je stepen korelacije (povezanosti) između dve veličine
- **Pearson-ov** test korelacije koristi se kada bar jedna od veličina koje se posmatraju ima normalnu raspodelu
- **Spearman-ov** test korelacije koristi se kada veličine nemaju normalnu raspodelu, ili kada je neka od veličina koje se posmatraju data preko skupa rangiranih podataka
- Oba ova testa daju veličine između -1 i 1 – koeficijent korelacije
- Što je apsolutna vrednost koeficijenta korelacije bliža 1, to je regresiona veza jača

Testiranje jačine regresione veze



```
#pearsonov koeficijent korelacije
cor(old, new)
summary(lm(new ~ old))

#spearmanov koeficijent korelacije
#funkcija rank ce da vrati rang liste vrednosti koja joj je prosledjena
x = rank(old)
y = rank(new)
cor(x, y)
```




- Netipične vrednosti mogu imati veliki uticaj na linearnu regresiju, i poželjno je da one pre izračunavanja koeficijenata regresione prave budu izbačene

```
#preko funkcije identify moze se pronaci indeks neke tacke klikom na
grafik
#ovo je korisno pri otkrivanju outlier-a
plot(old, new)
identify(old, new, n=1)

sa_netipicnim = lm(new ~ old)
bez_netipicnih = lm(new[-9] ~ old[-9])

plot(old, new)
abline(sa_netipicnim, col = "blue", lty = 1)
abline(bez_netipicnih, col = "red", lty = 2)
```



```
attach(florida)
names(florida)

sa_netipicnim = simple.lm(BUSH, BUCHANAN)
identify(BUSH, BUCHANAN, n=2)

bush = BUSH[-50]
bush = bush[-13]
buchanan = BUCHANAN[-50]
buchanan = buchanan[-13]

bez_netipicnih = simple.lm(bush, buchanan)

koeficijenti_netipicni = coef(sa_netipicnim)
koeficijenti = coef(bez_netipicnih)

plot(BUSH, BUCHANAN)
abline(koeficijenti_netipicni, col = "blue")
abline(koeficijenti, col = "red", lty = 2)

#vrednost za 250000
verdnost1 = koeficijenti_netipicni[1] + koeficijenti_netipicni[2] *
250000
vrednost2 = koeficijenti[1] + koeficijenti[2] * 250000
```

Zadatak 1 - postavka



- Data je biblioteka **MASS**, i u njoj data frame **USCereal**. U ovom data frame-u istražiti veze između:
 - proizvođača i police
 - police i masti
 - ugljenih hidrata i šećera
- Za analizu je dozvoljeno koristiti tabele, bar plot-ove i scatter plot-ove.

Zadatak 1 - rešenje



```
library(MASS)
attach(UScereal)

#analiza proizvođača i police
tabela = table(mfr, shelf)
tabela
prop.table(tabela, 2)
barplot(tabela, beside = T, legend.text = T, args.legend = c(x =
"topright"))

#analiza masti i police
tabela = table(shelf, fat)
tabela
prop.table(tabela, 1)
barplot(tabela, beside = T, legend.text = T, args.legend = c(x =
"topright"))

#analiza ugljenih hidrata i šećera
tabela = table(carbo, sugars)
tabela

#iz ovog grafika se vidi da između ove dve veličine ne postoji veza
plot(carbo, sugars)
detach(UScereal)
```

Zadatak 2



- Iz skupa podataka za cene kuća (homedata) izvući cene i istražiti vezu između podataka **old** i **new**. Da li postoji linearna veza? Da li postoje netipične vrednosti? Naći ih, i proceniti novu vrednost kuće koja je 1970. godine vredela 75000\$.

```
attach(homedata)
names(homedata)

plot(old, new)
identify(old, new, n=1)

sa_netipicnim = simple.lm(old, new)
bez_netipicnog = simple.lm(old[-9], new[-9])

koeficijenti = coef(bez_netipicnog)
vrednost = koeficijenti[1] + 75000 * koeficijenti[2]
vrednost

detach(homedata)
```

Zadatak 3



- Iz skupa podataka **emissions** iscrtati vezu između BDP-a i promenljive CO2. Naći netipičnu vrednost i iscrtati regresionu pravu sa njom i bez nje.

```
attach(emissions)
names(emissions)

plot(GDP, CO2)
identify(GDP, CO2, n = 1)

sa_netipicnim = simple.lm(GDP, CO2)
bez_netipicnog = simple.lm(GDP[-1], CO2[-1])

plot(GDP, CO2)
abline(sa_netipicnim, col = "blue")
abline(bez_netipicnog, col = "red", lty=2)

detach(emissions)
```