# Uvod u linearne modele

## Predgovor

maj 2012. godine

# Sadržaj

# Organizacija podataka u R-u

- Podaci imaju standardni tabelarni format u kome su kolone promenljive, a redovi entiteti
- Ovaj tip organizacije se u R-u naziva data frames
- Promenljive mogu biti:
  - numeričke (*numeric*), a to su merenja ili prebrojive vrednosti
  - kategoričke (*categorial*), koje se uobičajeno nazivaju faktori (*factors*)
  - uređeni faktori (*ordered factors*)
- Navedeni tipovi se često nazivaju i klase ili tipovi promenljivih
- Postoje i drugi tipovi podataka, kao što su logički (Boolean) i sl.



# Organizacija podataka u R-u

- Funkcija `str()` prikazuje sažeti opis strukture tabele (ali i drugih tipova objekata u R-u)
- Funkcija `summary()` sumira svaku promenljivu, već prema tome kojem tipu pripada
- Kada ukucate samo naziv tabele, izlistaće vam se cela. Zato je bolje koristiti prigodne funkcije `head()` i/ili `tail()`
- Osnovni oblik tabele u R-u je tzv. "dugački" (*long*), dok se u programima kao što su SPSS ili Statistica koristi "široko" prikazivanje (*wide*)
- R ima mogućnost konverzije iz jednog u drugi format

# "Dugački" i "široki" formati tabela

```
> state.x77 <- as.data.frame(state.x77)
> long <- reshape(state.x77, idvar='state', ids=row.names(state.x77),
+  times=names(state.x77), timevar='Characteristic',
+  varying=list(names(state.x77)), direction='long')
> head(long)
```

```
                     Characteristic Population      state
Alabama.Population       Population       3615    Alabama
Alaska.Population        Population        365     Alaska
Arizona.Population       Population       2212    Arizona
Arkansas.Population      Population       2110   Arkansas
California.Population    Population      21198 California
Colorado.Population      Population       2541   Colorado
```

```
> wide <- reshape(long, direction = 'wide',
+  new.row.names = unique(long$state))
> wide[1:6,1:6]
```

```
               state Population Income Illiteracy Life Exp Murder
Alabama      Alabama       3615   3624        2.1    69.05   15.1
Alaska        Alaska        365   6315        1.5    69.31   11.3
Arizona      Arizona       2212   4530        1.8    70.55    7.8
Arkansas    Arkansas       2110   3378        1.9    70.66   10.1
California California      21198   5114        1.1    71.71   10.3
Colorado     Colorado       2541   4884        0.7    72.06    6.8
```



# R paketi

- Paketi obuhvataju funkcije, podatke i dokumentaciju
- Mi ćemo ovde koristiti veći broj paketa
- Oni se mogu instalirati ili putem odgovarajućih menija ili direktno, pozivanjem funkcije:

```
> install.packages('rms')
```

- Paketi se instaliraju samo jednom, a najnovije verzije se preuzimaju pomoću funkcije:

```
> update.packages()
```

# Korišćenje R paketa

- U toku R *sesije*, paket se poziva na dva načina:

```
> require(rms)
```

```
> library(rms)
```

- Ovaj drugi način održava široko raširenu konfuziju i raspravu o terminologiji “paket” ili “biblioteka”



# Pristupanje dokumentaciji

- Da bi paket dospeo na CRAN, on prolazi niz provera kvaliteta
- Pre svega, svaka funkcija i tabela sa podacima mora biti dokumentovana
- Funkcija `data()` prikazuje imena i kratke opise podataka u nekom određenom paketu:

```
> data(package='lme4')

Data sets in package 'lme4':

Dyestuff                Yield of dyestuff by batch
Dyestuff2               Yield of dyestuff by batch
Pastes                  Paste strength by batch and cask
Penicillin              Variation in penicillin testing
cake                    Breakage angle of chocolate cakes
cbpp                    Contagious bovine pleuropneumonia
sleepstudy              Reaction times in a sleep deprivation study
```

# Pristupanje dokumentaciji

- Funkcija `help()` (ili samo `?`) sa argumentom koji je naziv funkcije ili tabele sa podacima, prikazaće odgovarajuću dokumentaciju
- Ova funkcija za pomoć ima više varijanti:
  - `help()` (`?`) $\rightarrow$ dokumentacija
  - `help.search()` (`??`) $\rightarrow$ pretraživanje sistema pomoći i dokumentacije
  - `RSiteSearch()` $\rightarrow$ pretražuje ključnu reč ili frazu (pod navodnim znacima) na `R-help` listi, internet stranama i sl., koristeći `http://search.r-project.org`, a rezultate pretrage prikazuje u brauzeru



# Sadržaj

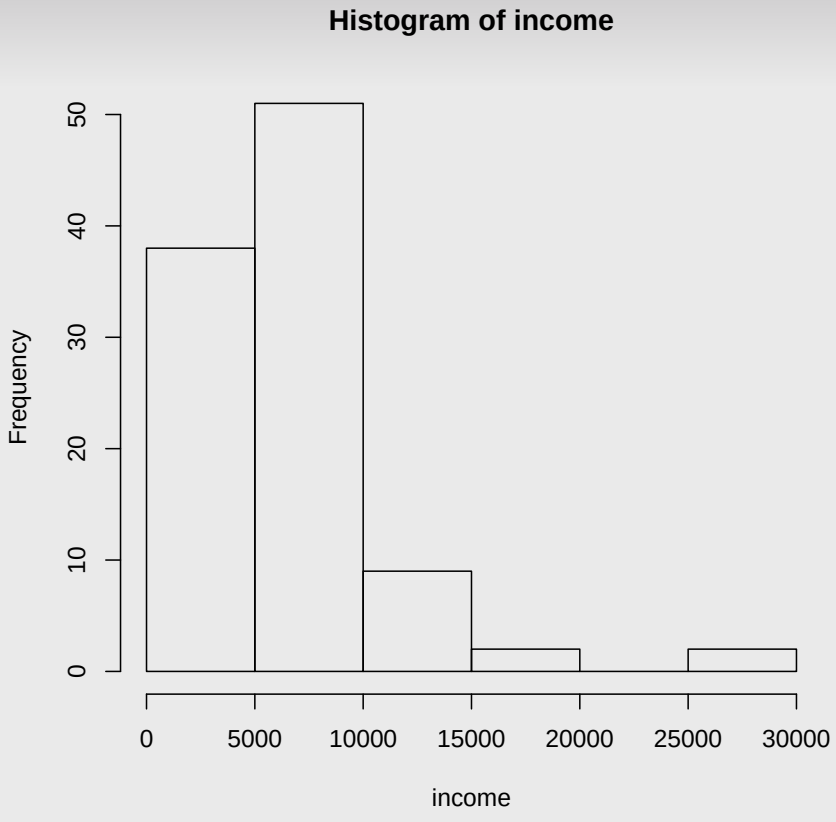
# ISPITIVANJE PODATAKA

- Oblik distribucije podataka uobičajeno se ispituje dijagramom stubaca – histogramom

```
> require(car)
> data(Prestige)
> attach(Prestige)

The following object(s) are masked from 'package:datasets':
    women

> head(Prestige)

                    education income women prestige census type
gov.administrators      13.11  12351 11.16     68.8   1113 prof
general.managers        12.26  25879  4.02     69.1   1130 prof
accountants             12.77   9271 15.70     63.4   1171 prof
purchasing.officers     11.42   8865  9.11     56.8   1175 prof
chemists                14.62   8403 11.68     73.5   2111 prof
physicists              15.64  11030  5.13     77.6   2113 prof

> hist(income)
```



# DIJAGRAM STUBACA ZA PROMENLJIVU 'PRIHODI'



# Problem broja razreda (stubaca)

- Vrlo često, broj razreda koji funkcija `hist()` određuje automatski, nije optimalan
- Zato je dobro utvrditi optimalni broj razreda pomoću Fridman-Diakonis algoritma (Freedman-Diaconis):

$$\lceil \frac{(max-min)}{h} \rceil, \text{ gde je } h = 2 \times IQR \times n^{-1/3}$$

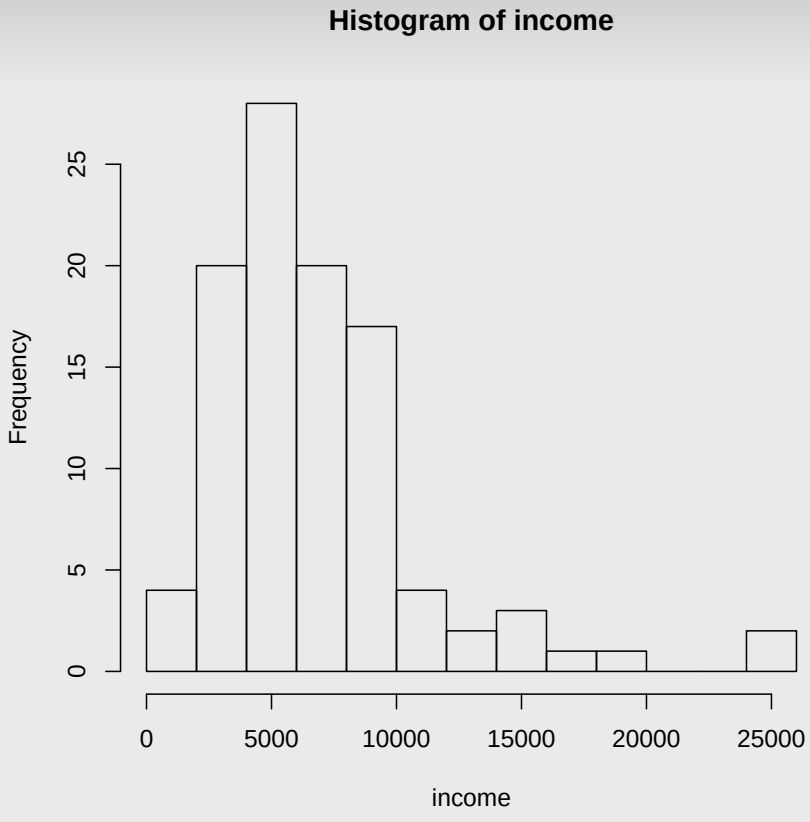
odnosno, u obliku podesnijem za direktno izračunavanje:

$$\lceil \frac{n^{1/3}(min-max)}{2(Q_3-Q_1)} \rceil$$

```
> nclass.FD(income)
[1] 15
> hist(income, breaks='FD')
```



# Korigovani dijagram stubaca



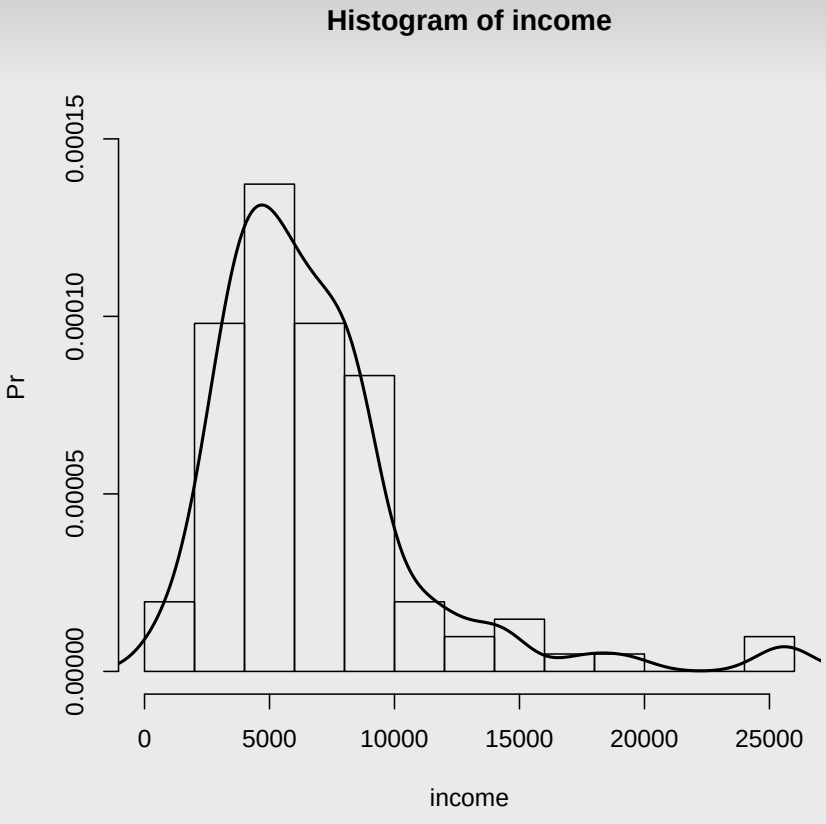
# Gustina raspodele

- Na postojeći dijagram, može se naneti i gustina raspodele
- U R-u koristimo neparametarsku ocenu gustine raspodele

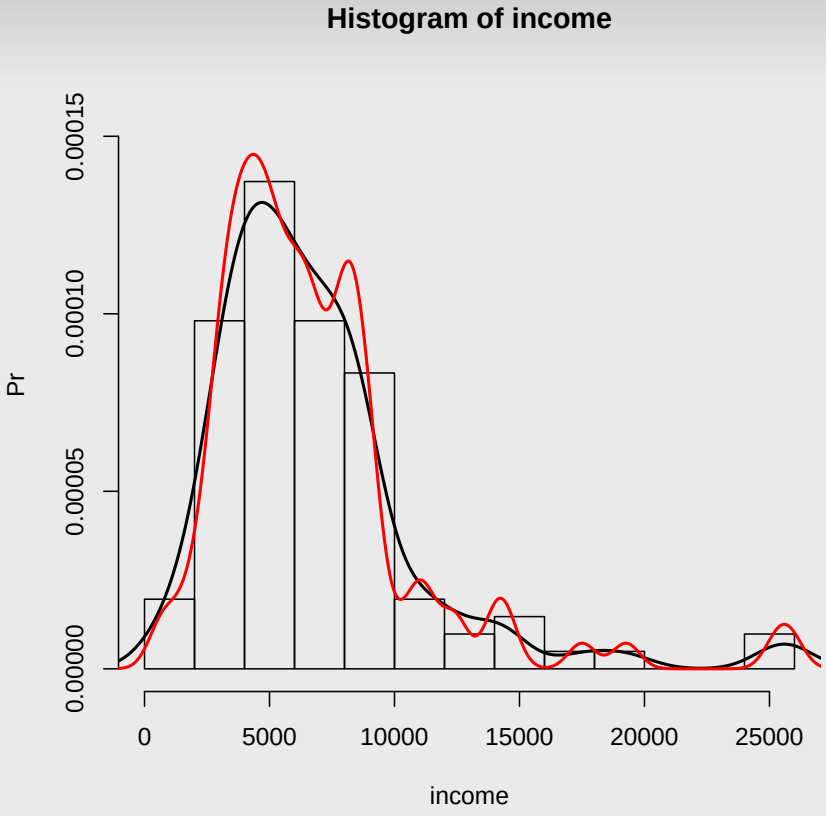
```
> hist(income, breaks='FD', probability=TRUE, ylab='Pr', ylim=c(0,0.00016))
> lines(density(income), lwd=2)
> lines(density(income, adjust=0.5), lwd=2, col='red')
```



# Dijagram stubaca i gustina raspodele



# Dijagram stubaca i gustina raspodele



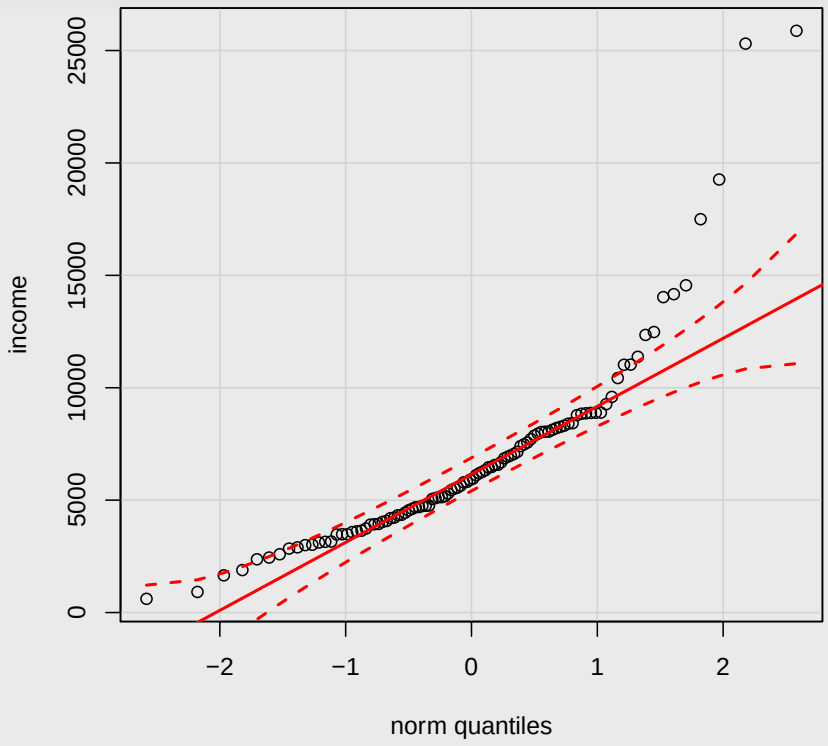



# Kvantil-kvantil dijagram

- Za mene, lično, mnogo je informativniji dijagram koji rastežu očekivana normalna raspodela (teorijska) i opažena distribucija podataka

```
> qqPlot(income)
```

# KVANTIL-KVANTIL DIJAGRAM






# KVANTIL-KVANTIL DIJAGRAM

- pored funkcije `qqPlot()` koja se nalazi u paketu CAR postoje i druge, slicne, vrlo korisne funkcije:
  - STATS → `qqnorm()`
  - LATTICE → `qqmath()`
- Funkcije `qqPlot()` i `qqmath()` dopustaju definisanje očekivane raspodele:

| raspodela | gustina | raspodela | kvantili | slučajni brojevi |
|---|---|---|---|---|
| normalna (Gausova) | dnorm | pnorm | qnorm | rnorm |
| hi-kvadrat | dchisq | pchisq | qchisq | rchisq |
| F | df | pf | qf | rf |
| t | dt | pt | qt | rt |
| binomna | dbinom | pbinom | qbinom | rbinom |
| uniformna | dunif | punif | qunif | runif |

# Preduslovi za linearno modelovanje

- Linearni odnos između promenljivih
- Više podataka za svakog ispitanika (za svaku promenljivu)
- Normalna raspodela (greške)
- Nezavisnost ispitanika
- Odsustvo perfektne multikolinearnosti

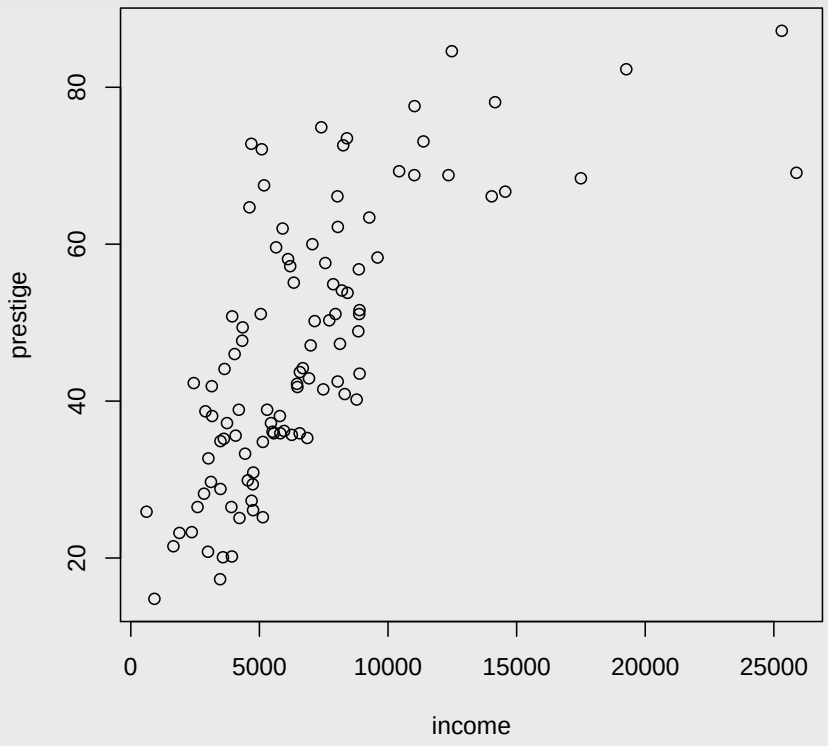


# Dijagram raspršenja (skater-dijagram)

- Za ispitivanje odnosa između dve kvantitativne promenljive najčešće se koristi dijagram raspršenja

```
> plot(income, prestige)
```

# Dijagram raspršenja






# Dijagram raspršenja (skater-dijagram)

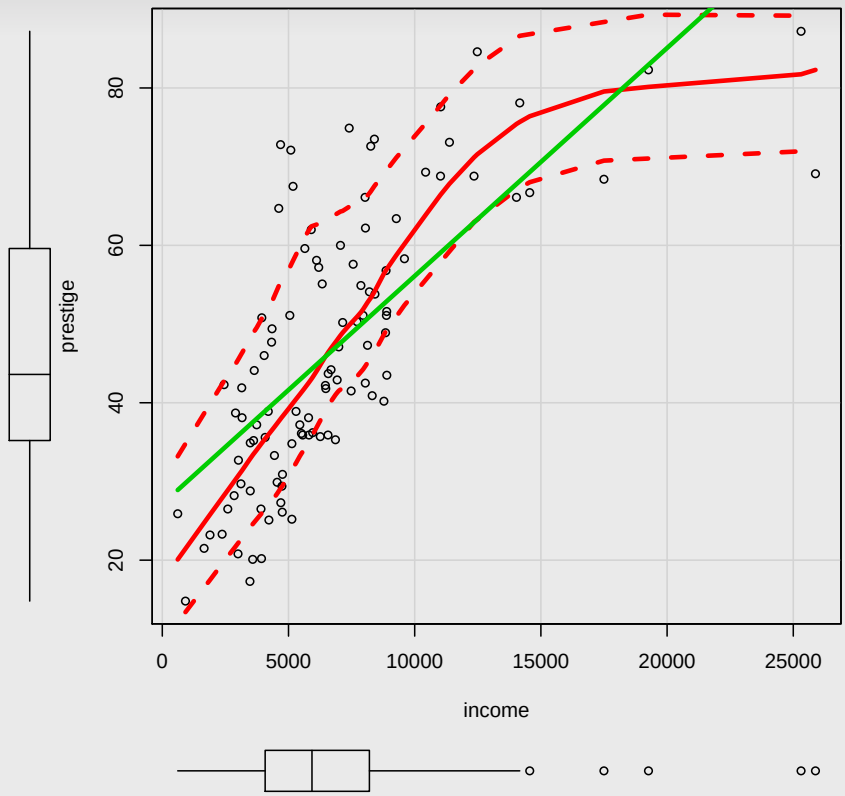
- Za ispitivanje odnosa između dve kvantitativne promenljive najčešće se koristi dijagram raspršenja

```
> plot(income, prestige)
```

- Tipično se prikazuju i regresione linije (najmanjih kvadrata i/ili neparametarske), što nam omogućava funkcija iz paketa car

```
> scatterplot(income, prestige, span=.6, lwd=3,
+   labels=rownames(Prestige))
```

# Dijagram raspršenja (skater-dijagram)



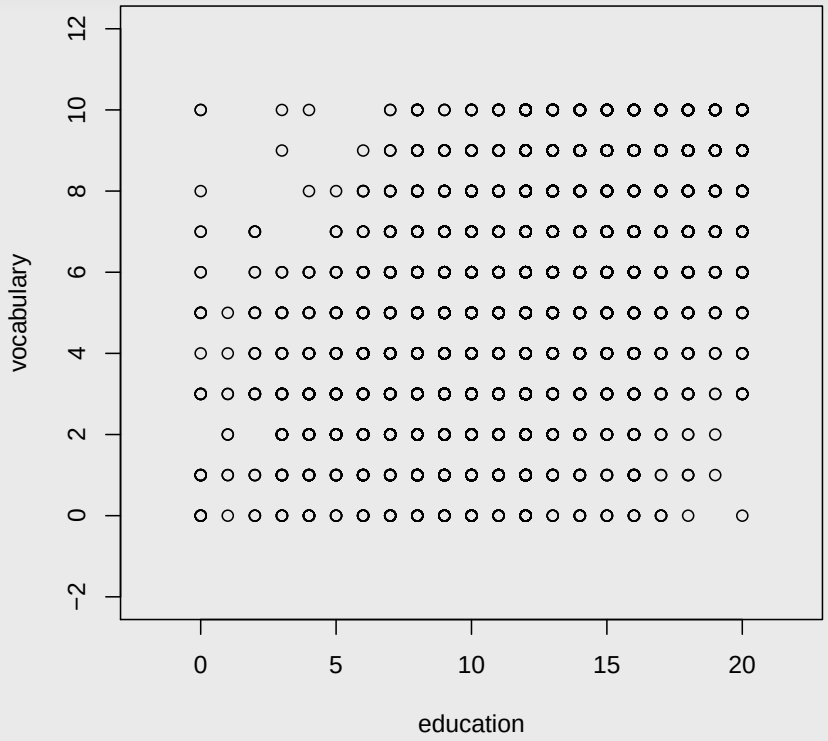



# Dijagram raspršenja (skater-dijagram)

- Za ispitivanje odnosa između diskretnih kvantitativnih promenljivih dijagram raspršenja nije baš najinformativniji

```
> detach(Prestige)
> data(Vocab)
> attach(Vocab)
> plot(education, vocabulary, xlim=c(-2,22), ylim=c(-2,12))
```

# Dijagram raspršenja (skater-dijagram)






# Dijagram raspršenja (skater-dijagram)

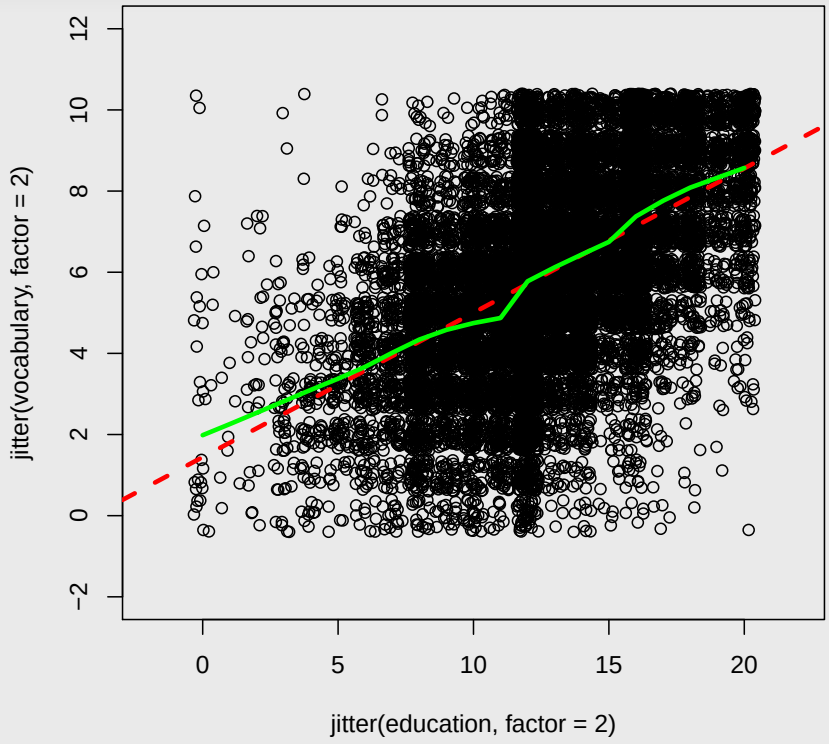
- Za ispitivanje odnosa između diskretnih kvantitativnih promenljivih dijagram raspršenja nije baš najinformativniji

```
> detach(Prestige)
> data(Vocab)
> attach(Vocab)
> plot(education, vocabulary, xlim=c(-2,22), ylim=c(-2,12))
```

- Zato je, obično, korisno i dovoljno dodati malo “šuma” – jitter() je funkcija koja dodaje neku malu, normalno distribuiranu vrednost

```
> plot(jitter(education, factor=2), jitter(vocabulary, factor=2),
+   xlim=c(-2,22), ylim=c(-2,12))
> abline(lm(vocabulary ~ education), lwd=3, lty=2, col='red')
> lines(lowess(education, vocabulary, f=.2), lwd=3, col='green')
```

# Dijagram raspršenja (skater-dijagram)



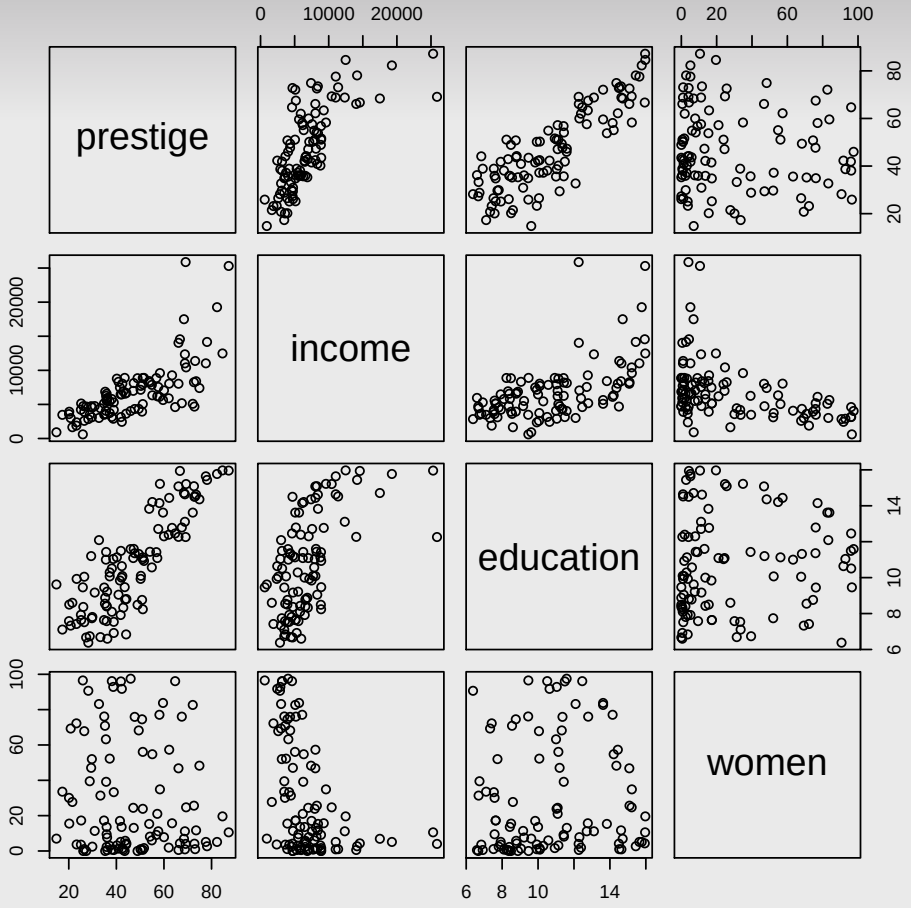



# Dijagram raspršenja za više promenljivih

- Čak i kada imamo veći broj promenljivih, što je tipično za višestruku regresiju, postoji podesan način za vizuelnu analizu odnosa
- Matrični dijagram raspršenja prikazuje odnose *po parovima* (marginalne; *pairwise; marginal*)

```
> detach(Vocab)
> attach(Prestige)

The following object(s) are masked from 'package:datasets':
    women

> pairs(cbind(prestige, income, education, women))
```

# Dijagram raspršenja za više promenljivih






# Dijagram raspršenja za više promenljivih

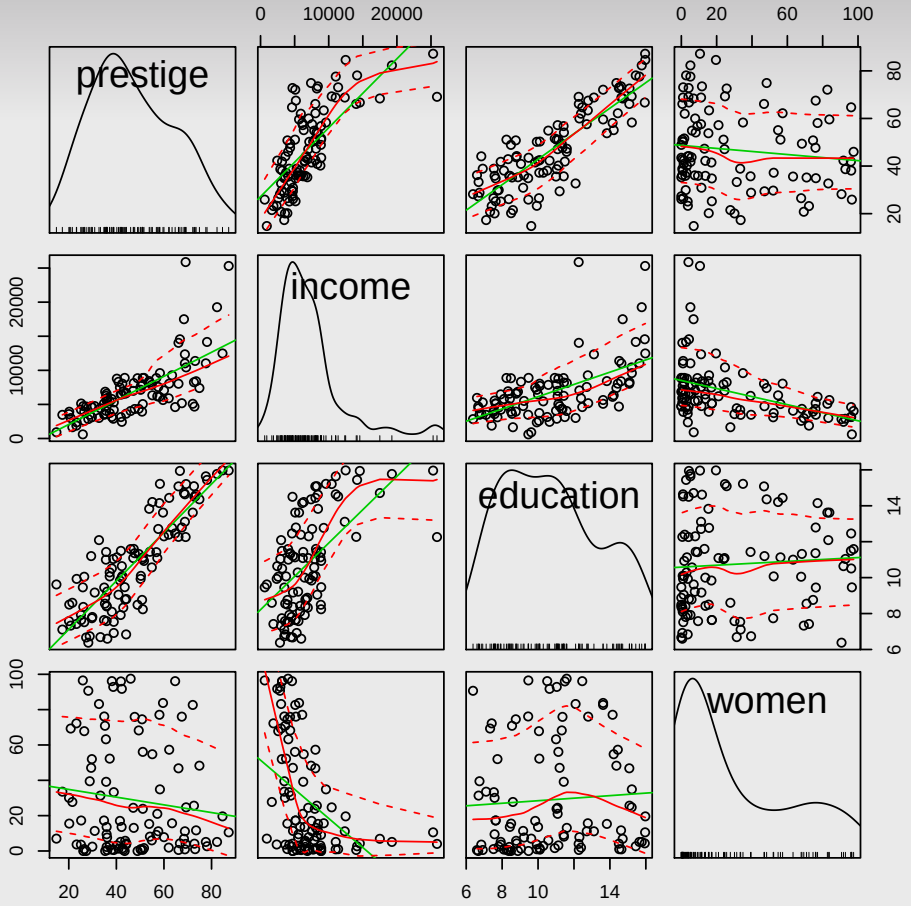
- Čak i kada imamo veći broj promenljivih, što je tipično za višestruku regresiju, postoji podesan način za vizuelnu analizu odnosa
- Matrični dijagram raspršenja prikazuje odnose *po parovima* (marginalne; *pairwise; marginal*)

```
> detach(Vocab)
> attach(Prestige)
> pairs(cbind(prestige, income, education, women))
```

- Ponovo nam paket CAR daje prigodnu funkciju sa dopunskim informacijama

```
> scatterplotMatrix(cbind(prestige, income, education, women),
+     diag='density', span=.75)
```

# Dijagram raspršenja za više promenljivih






# Sadržaj

# Čemu služe transformacije podataka?

- "popravljanje" normalnosti i/ili simetričnosti raspodele
- stabilizovanje raspršenosti (varijabiliteta)
- obezbeđivanje linearnog odnosa među promenljivama (preduslova za linearno modelovanje)
- Porodica stepenih (i korenskih) funkcija se najčešće koristi u ove svrhe
- Neko $x$ zamenjuje se sa $x' = x^{\lambda}$:

$$\begin{array}{ll} x' = x^2 & \text{za } \lambda = 2 \\ x' = \sqrt{x} & \text{za } \lambda = 1/2 \\ x' = 1/x & \text{za } \lambda = -1 \end{array}$$



# Boks-Koks transformacija (*Box-Cox*)

- Boks-Koks transformacije su, u suštini, veoma slične stepenim i korenskim transormacijama

$$x' = x^{(\lambda)} = \begin{cases} \frac{x^{\lambda}-1}{\lambda} & \text{ako } \lambda \neq 0 \\ log_e x & \text{ako } \lambda = 0 \end{cases}$$

- Fox (2002) konstatuje da je log-transformacija neka vrsta transformacije na "nulti" stepen
- Boks-Koks transformacija ima smisla samo za pozitivne vrednosti
- A to se može postići dodavanje dovoljno velike konstante (*start*)

```
> bcPower(1:5, 2)
[1]  0.0  1.5  4.0  7.5 12.0
```

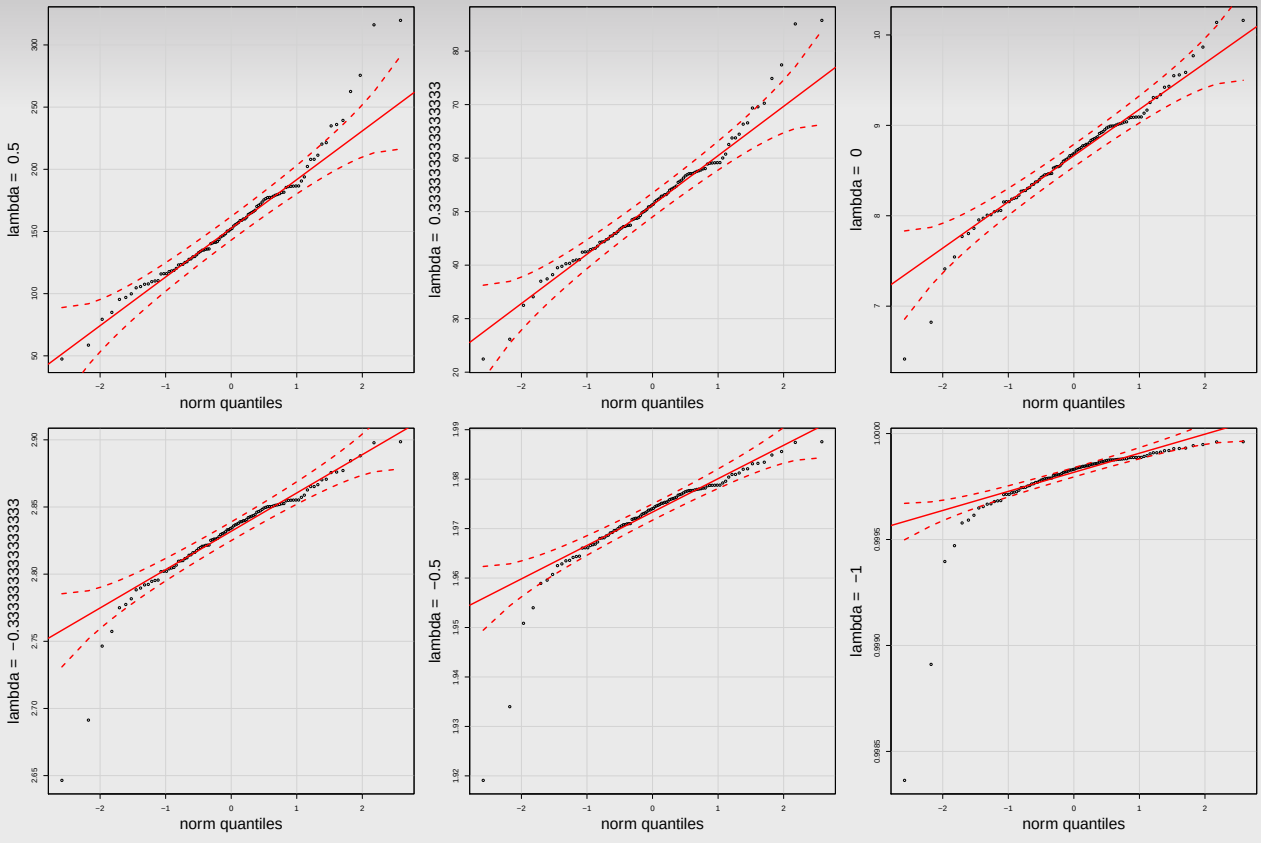
# Transformacija za normalnost i simetriju

- Postavlja se pitanje kako utvrditi vrednost eksponenta tako da dobijemo normalnu ili, barem, simetričnu raspodelu?!
- Jedan način je da tu vrednost tražimo putem "pokušaja i pogrešaka" (napipavanjem); na primer:

```
> par(mfrow=c(2,3), mar=c(5, 5, 1, 1), cex.lab=2)
> for (p in c(1/2, 1/3, 0, -1/3, -1/2, -1))
+  qqPlot(bcPower(income, lambda=p),
+  ylab=paste('lambda = ', p))
> par(mfrow=c(1,1))
```



# Transformacija za normalnost i simetriju




Univerzitetski centar za primenjenu statistiku

# Transformacija za normalnost i simetriju

- Postoje i komforniji načini za određivanje optimalne vrednosti za stepeni parametar
- Paket car nam pruža prigodnu funkciju `powerTransform()`:

```
> summary(powerTransform(income))

bcPower Transformation to Normality
       Est.Power Std.Err. Wald Lower Bound Wald Upper Bound
income    0.1793   0.1108          -0.0379           0.3965

Likelihood ratio tests about transformation parameters
                            LRT df         pval
LR test, lambda = (0)  2.710304  1 9.970200e-02
LR test, lambda = (1) 47.261001  1 6.213585e-12
```



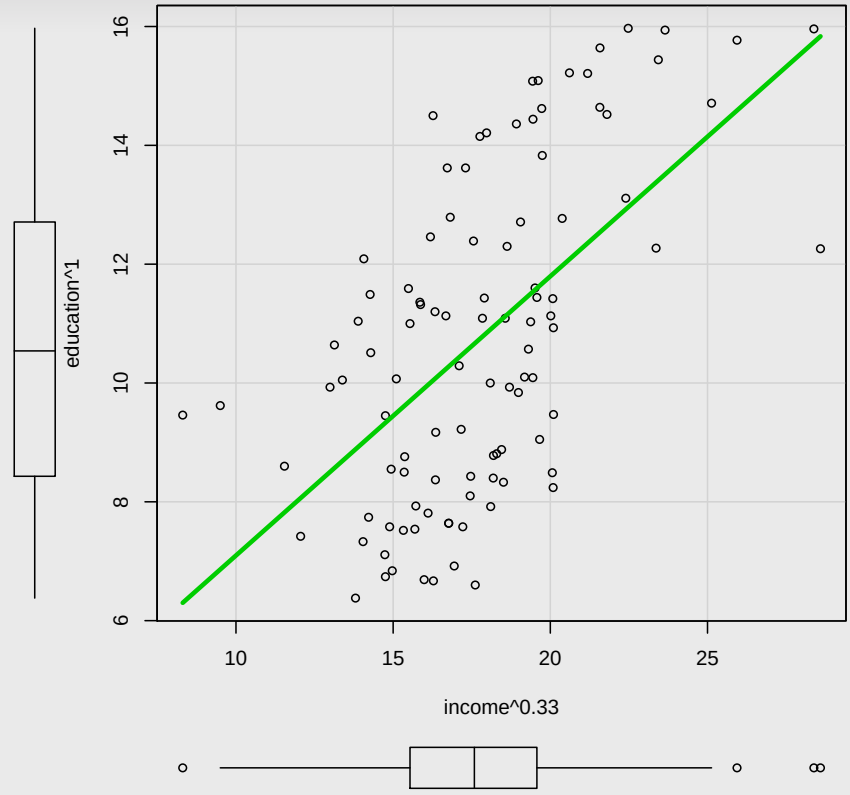
# Transformacija za normalnost i simetriju

- Takođe, moguće je utvrditi transformaciju za dve promenljive, da bi se dobila bivarijatna normalna:

```
> summary(powerTransform(cbind(income, education)))

bcPower Transformations to Multinormality
          Est.Power Std.Err. Wald Lower Bound Wald Upper Bound
income       0.2617   0.1014           0.0629           0.4604
education    0.4242   0.4033          -0.3663           1.2146

Likelihood ratio tests about transformation parameters
                                LRT df         pval
LR test, lambda = (0 0)     7.694014  2 2.134352e-02
LR test, lambda = (1 1)    48.872743  2 2.440159e-11
LR test, lambda = (0.33 1)  2.406223  2 3.002585e-01

> scatterplot(income^.33 , education^1, lwd=3, smooth=FALSE)
```

Univerzitetski centar za primenjenu statistiku

# Transformacija za normalnost i simetriju



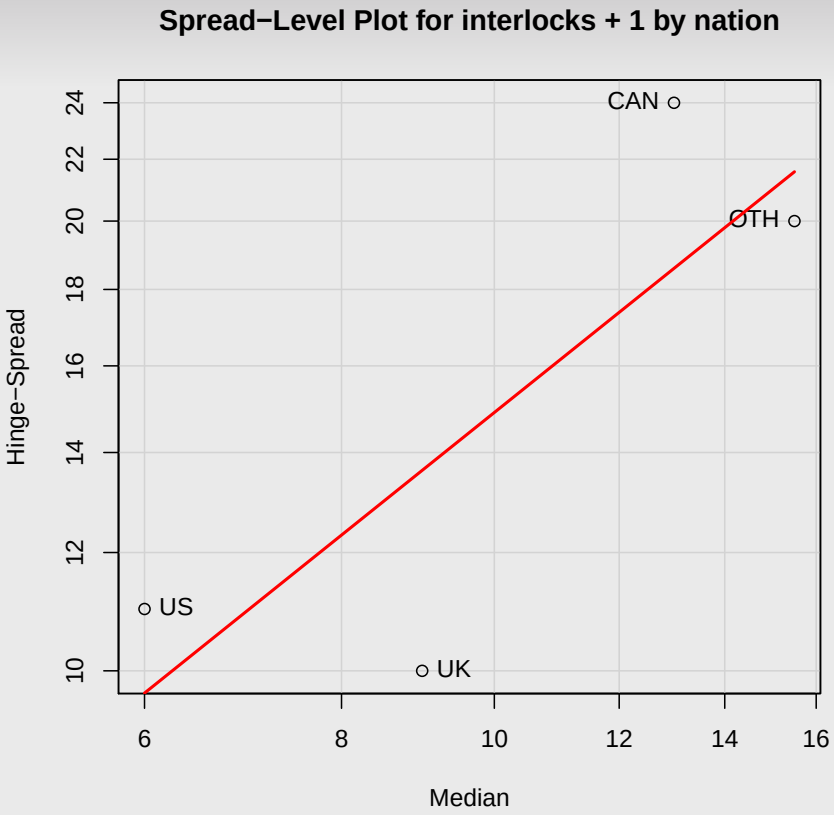



# Transformacije za ujednačavanje raspršenja

- Čest je slučaj da raspršenje (variranje) različitih promenljivih nije ujednačeno
- S tim u vezi, nije retko da su parametri pozicije (*location*) i skale (*scale*) u korelaciji
- Skala-raspršenje dijagram (*sprad-level plot*; Tukey, 1977), podesan je za ispitivanje ovog odnosa
- Paket car ima funkciju koja prikazuje pomenute odnose:

```
> detach(Prestige)
> attach(Ornstein)
> spreadLevelPlot(interlocks + 1 ~ nation)

    LowerHinge Median UpperHinge Hinge-Spread
US           2    6.0         13           11
UK           4    9.0         14           10
CAN          6   13.0         30           24
OTH          4   15.5         24           20
Suggested power transformation:  0.1534487
```

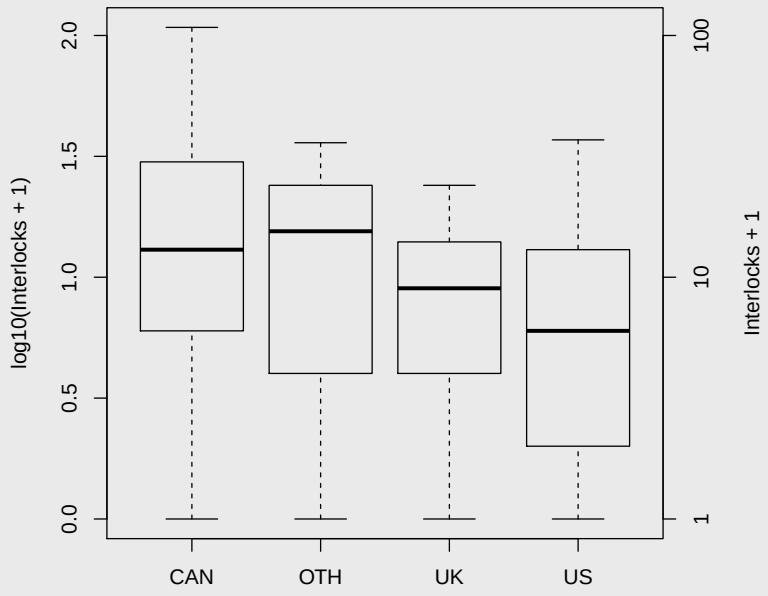
# Transformacija za normalnost i simetriju






# Transformacije za ujednačavanje raspršenja

- U prethodnom slučaju, predložena vrednost za $\lambda$ je 0.15, a to je blizu proste log-transformacije



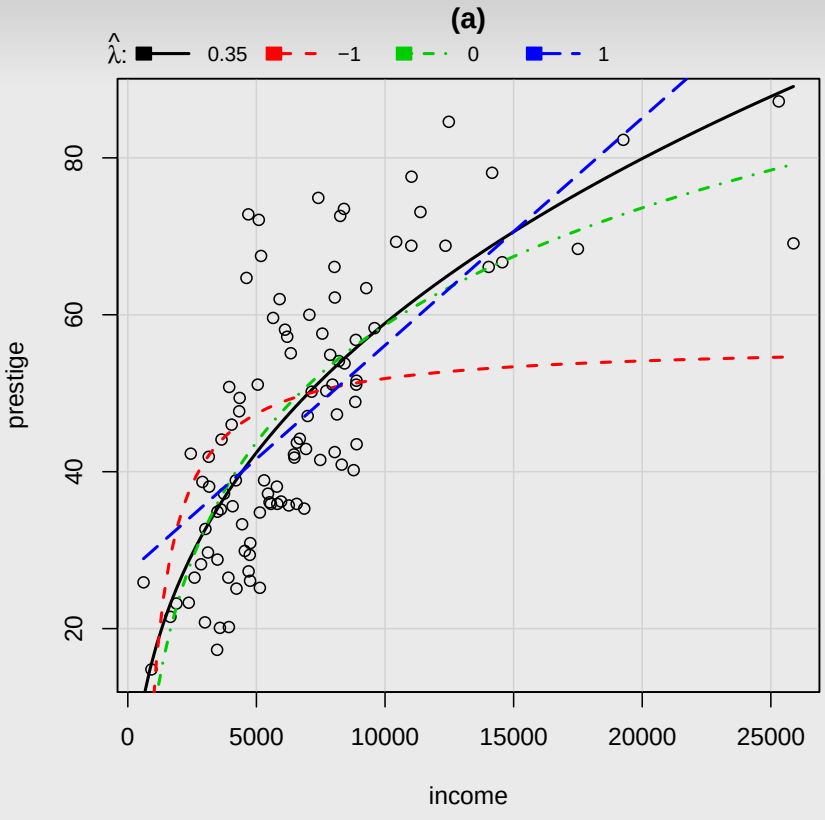
# Transformacije za linearnost odnosa

- Već smo prethodno istakli da osnovna funkcija `powerTransform()` iz paketa CAR može da sugeriše kako da postignemo linearizaciju odnosa između dve promenljive
- Međutim, ovaj paket nudi i još neke zanimljive mogućnosti za puno "razumevanje" podataka i njihovih odnosa

```
> detach(Ornstein)
> attach(Prestige)

The following object(s) are masked from 'package:datasets':
    women

> invTranPlot(prestige ~ income, lwd=2, xlab='income', main='(a)')

      lambda      RSS
1  0.3491486 12723.73
2 -1.0000000 22166.44
3  0.0000000 13477.93
4  1.0000000 14616.17
```



# Transformacija za linearnost odnosa



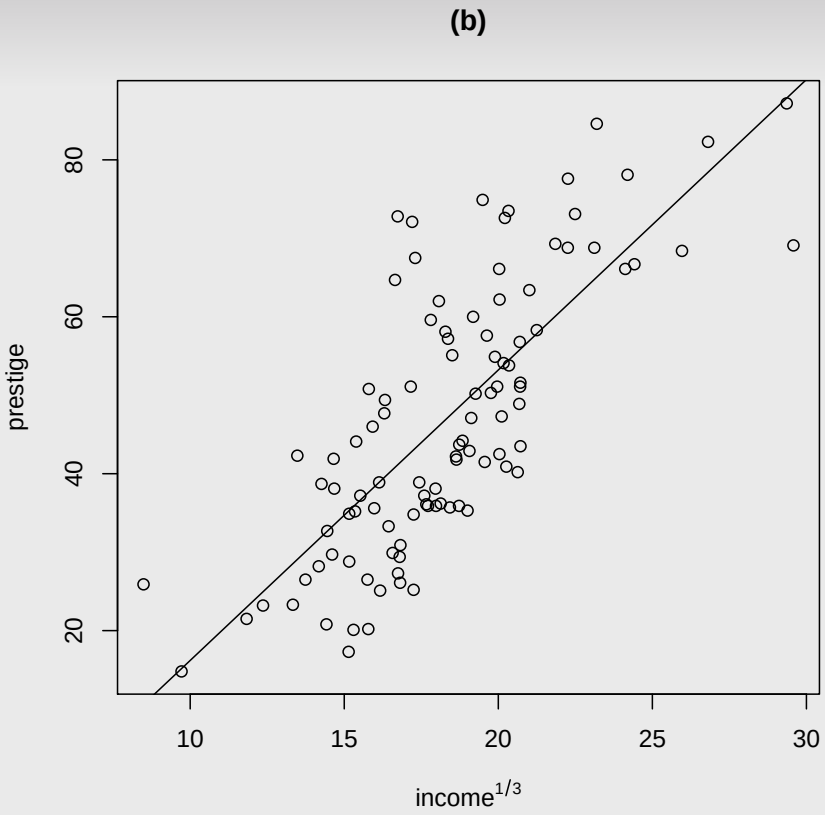
# Transformacije za linearnost odnosa

- Već smo prethodno istakli da osnovna funkcija `powerTransform()` iz paketa CAR može da sugeriše kako da postignemo linearizaciju odnosa između dve promenljive
- Međutim, ovaj paket nudi i još neke zanimljive mogućnosti za puno "razumevanje" podataka i njihovih odnosa

```
> detach(Ornstein)
> attach(Prestige)
> invTranPlot(prestige ~ income, lwd=2, xlab='income', main='(a)')

> plot(prestige ~ I(income^(1/3)), xlab=expression(income^{1/3}),
+   main='(b)')
> abline(lm(prestige ~ I(income^(1/3))))
```



# Transformacija za linearnost odnosa



# Transformacije proporcija

- Stepena transformacija proporcija (ili procenata) obično ne daje zadovoljavajuće rezultate
- U stvari, nema dobrog rešenja kada su vrednosti ekstremne; blizu teorijskog minimuma i maksimuma, tj. $(0 i 1)$
- Jedna od najčešće korišćenih transformacija proporcija jeste logit transformacija – logaritam racia šansi (*log-odds ratio*):

$$logit(p) = log_e \frac{p}{1-p}$$



# Transformacije proporcija

- Funkcija `logit()` iz paketa CAR za nas automatski računa logit-vrednosti
- Takođe, problem ekstrema rešava tako što remapira interval $(0, 1)$ na $(.025, .975)$

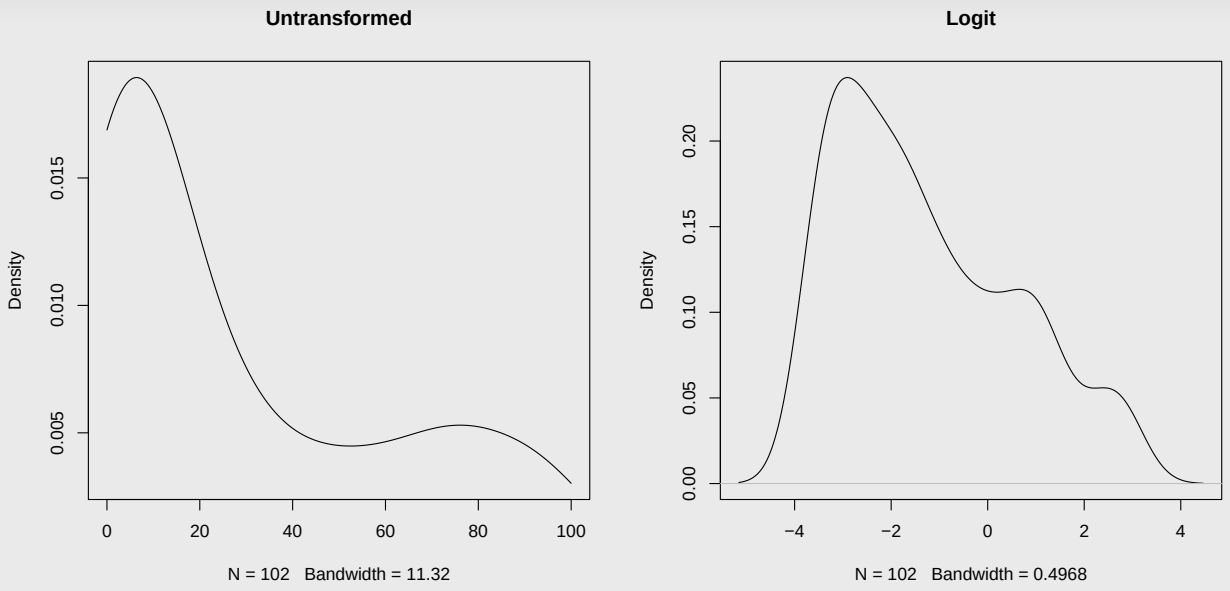
```
> logit(seq(0.1, 0.9, 0.1))

[1] -2.1972246 -1.3862944 -0.8472979 -0.4054651  0.0000000  0.4054651  0.8472979
[8]  1.3862944  2.1972246

> logit(seq(0, 1, 0.1))

 [1] -3.6635616 -1.9924302 -1.2950457 -0.8001193 -0.3846743  0.0000000
 [7]  0.3846743  0.8001193  1.2950457  1.9924302  3.6635616

Warning message:
In logit(seq(0, 1, 0.1)) : proportions remapped to (0.025, 0.975)
```

# Transformacije proporcija

```
> par(mfrow=c(1,2))
> plot(density(women, from=0, to=100), main='Untransformed')
> plot(density(logit(women), adjust=0.75), main='Logit')
> par(mfrow=c(1,1))
```



# Transformacije proporcija




```
Warning in logit(women) : proportions remapped to (0.025, 0.975)
```

Univerzitetski centar za primenjenu statistiku

# Sadržaj





# Rad sa podacima u R-u

- Ponovite osnovne komande za pregledanje podataka u R-u
  - Pregledajte koje tabele podataka sadrži paket car
  - Utvrdite strukturu jedne od tih tabela
  - Utvrdite deskriptivne pokazatelje za tu tabelu
  - Pogledajte početak i kraj tabele
  - Dajte komandu R-u da je ta tabela aktivna za rad
  - Deaktivirajte tabelu

# ISPITIVANJE RASPODELA

- Koje osnovne načine prikazivanja raspodela u R-u poznajete?
  - Odaberite neku od tabela, aktivirajte je za rad u R-u i prikažite njenu raspodelu pomoću dijagrama stubaca
  - Napravite novi dijagram stubaca, ali sada sa optimalnim brojem razreda
  - Ponovite tu komandu i tražite da vrednosti budu iskazane kao ocenjene verovatnoće
  - Zatim, nanesite gustinu raspodele
  - Prikažite istu raspodelu pomoću kvantil-kvantil dijagrama



# ISPITIVANJE ODNOSA

- Kakvi odnosi između promenljivih moraju biti da biste mogli primeniti neki od postupaka linearnog modelovanja?
  - Prikažite odnos promenljive koju ste prethodno odabrali i još jedne promenljive
  - Da li je njihov odnos (približno) linearan
  - Ako nije, kako biste prevazišli ovaj problem
  - Pokušajte da eksperimentišete sa nekom od prigodnih funkcija iz paketa CAR

# Transformacije

- Navedite postupke za transformisanje podataka koje ste upamtili sa predavanja
- Da li znate za još neke?
- Hajde da počistimo "radni prostor" (*workspace*) u R-u. Jedan način da to uradite jeste:

```
> search()
```

  a zatim primenite funkciju `detach()`

- Zadajte R-u sledeći niz komandi i razmislite o tome šta vam one omogućavaju:

```
> attach(UN)
> summary(powerTransform(UN))
> with(UN, summary(powerTransform(cbind(infant.mortality, gdp))))
> summary(powerTransform(UN[ , c('infant.mortality', 'gdp')]))
> detach(UN)
```



# Transformacije: naprednije mogućnosti

- Upoznajte se sa još nekim mogućnostima:

```
> summary(p1 <- with(Prestige, powerTransform(cbind(income, education))))
> testTransform(p1, lambda=c(0.33, 1))
> testTransform(p1, lambda=c(0, 1))

> summary(p2 <- with(Prestige,
+  powerTransform(cbind(income, education) ~ type)))
> testTransform(p2, c(0, 1))
```

K R A J