RPS
Analiza podatkov

A. Blejec

21. november 2013

Kazalo

Povzetek

Primer analize podatkov

O rasti in velikosti ljudi imamo nekaj mnenj, ki jih lahko izrazimo v obliki raziskovalnih vprašanj. Najprej si zastavimo vprašanja.

Vprašanja

Nekaj vprašanj, na katere bi radi odgovorili je:

- Ali so fantje večji od deklet?
- Ali so fantje težji od deklet?
- Ali sta razpon rok in višina približno enaka?
- Ali drži Galtonovo opažanje glede višine otrok in staršev?
- ...

Zbrali smo nekaj podatkov o študentih, s katerimi si bomo lahko poskusili odgovoriti. Nato zberemo podatke, s katerimi bomo poskusili odgovoriti na vprašanja. Ker predvidevamo, da nas bo zanimalo še kaj, zberemo podatke o še nekaj spremenljivkah.

```
lfn <- "Podatki2012.txt"
```

Podatki

Podaki so o študentih 3. letnika biologije v letu 2012/13 so v datoteki lfn in na http://bit.ly/16oBVpR

```
fpath <- "http://bit.ly/16oBVpR"
data <- read.table(fpath, header = TRUE, sep = "\t")
names(data)
```

```
 [1] "starost" "mesec"   "spol"    "masa"    "visina" 
 [6] "roke"    "cevelj"  "lasje"   "oci"     "mati"   
[11] "oce"     "majica" 
```

Opisna statistika

```
summary(data[, 1:6])
```

```
    starost          mesec         spol        masa      
 Min.   :20.00   Min.   : 0.000   F:33   Min.   :50.00  
 1st Qu.:21.00   1st Qu.: 5.000   M:10   1st Qu.:55.50  
 Median :21.00   Median : 7.000          Median :61.00  
 Mean   :22.07   Mean   : 6.814          Mean   :63.42  
 3rd Qu.:22.00   3rd Qu.: 9.500          3rd Qu.:70.00  
 Max.   :59.00   Max.   :11.000          Max.   :91.00  
                                                        
     visina           roke      
 Min.   :156.0   Min.   :154.0  
 1st Qu.:164.0   1st Qu.:163.2  
 Median :170.0   Median :167.8  
 Mean   :169.9   Mean   :169.3  
 3rd Qu.:173.5   3rd Qu.:172.5  
 Max.   :189.0   Max.   :193.0  
                 NA's   :5      
```

Ali pri podatkih kaj opazite?

Nenavadni podatki

Kaj storiti s tistim, ki je napisal, da je rojen v mesecu 0?
Eden pa je star 59 let??

Popravljanje podatkov

Odstranimo ta starga

```
data <- data[data$starost < 30, ]
```

Podatke o mesecu 0 spremenimo v NA

```
data[data$mesec == 0, "mesec"] <- NA
table(data$mesec)
```

```
 1  2  3  4  5  6  7  8  9 10 11 
 1  3  2  3  4  3  7  5  2  5  6 
```

Nadaljevanje opisa

```
summary(data[, 7:dim(data)[2]])
```

```
    cevelj         lasje   oci         mati
Min.   :36.00   S:19    S:23    Min.   :157.0
1st Qu.:38.00   T:23    T:19    1st Qu.:160.0
Median :39.00                   Median :165.0
Mean   :39.93                   Mean   :165.6
3rd Qu.:41.00                   3rd Qu.:168.0
Max.   :48.00                   Max.   :180.0
                                NA's   :5
     oce          majica
Min.   :170.0   L : 4
1st Qu.:174.0   M :19
Median :179.0   S :16
Mean   :179.1   XL: 1
3rd Qu.:182.0   XS: 2
Max.   :190.0
NA's   :5
```

1 Višina in spol

Primerjajte razpone vrednosti višin študentov in staršev.

Višina po spolu

Povzetek višin glede na spol

```
  summary(data$mati)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
  157.0   160.0   165.0   165.6   168.0   180.0       5 
```

```
  by(data$visina, data$spol, summary)
```

```
data$spol: F
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  156.0   163.0   168.0   166.8   170.0   178.0 
---------------------------------------------
data$spol: M
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  171.0   180.0   180.0   180.2   183.0   189.0 
```

```
  summary(data$oce)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
  170.0   174.0   179.0   179.1   182.0   190.0       5 
```

Doseg spremenljivk v objektu `data.frame`

Poglejte kakšne so vrednosti spremenljivke `visina`! Ali je v delovnem prostoru (workspace)? Do spremenljivk lahko pridem posredno na več načinov

- `data$visina`
- `data[, 'visina']`
- `data[,5]`

Neposreden dostop

Neposreden dostop do spremenljivk omogoči

```
  attach(data)
  length(visina)
```

```
[1] 42
```

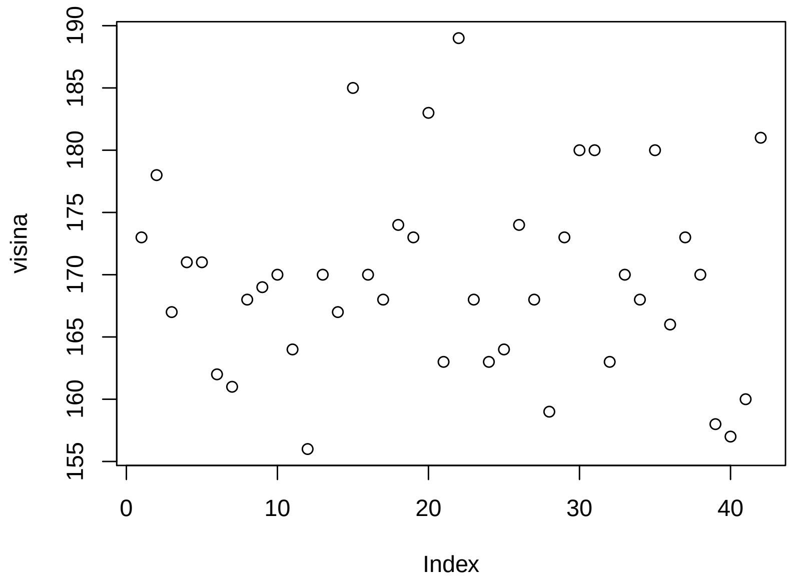
```
  visina[1:5]
```

```
[1] 173 178 167 171 171
```

Grafični prikazi

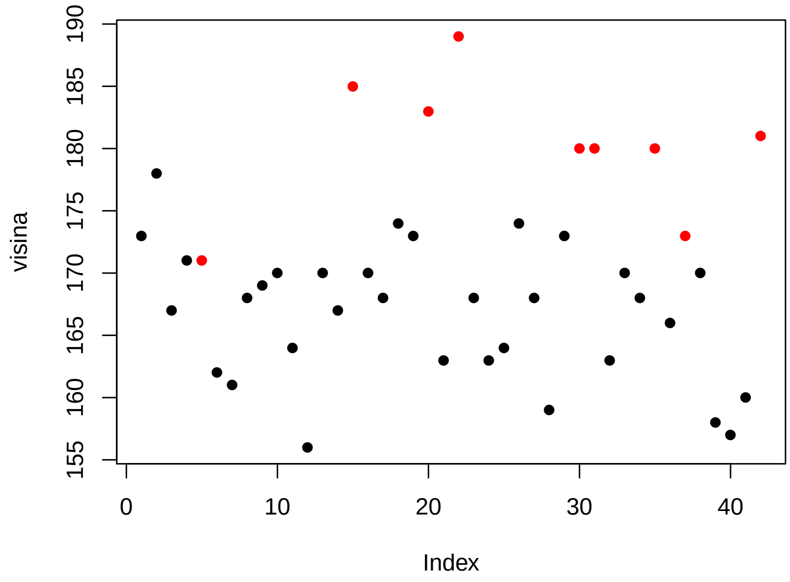
Grafični prikaz podatkov

```
plot(visina)
```



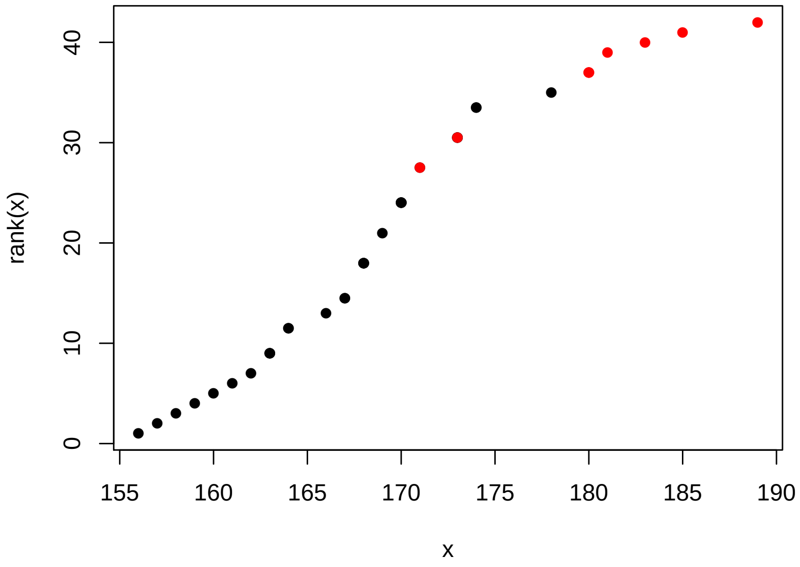
Grafični prikaz podatkov

```
plot(visina, pch = 16, col = spol)
```



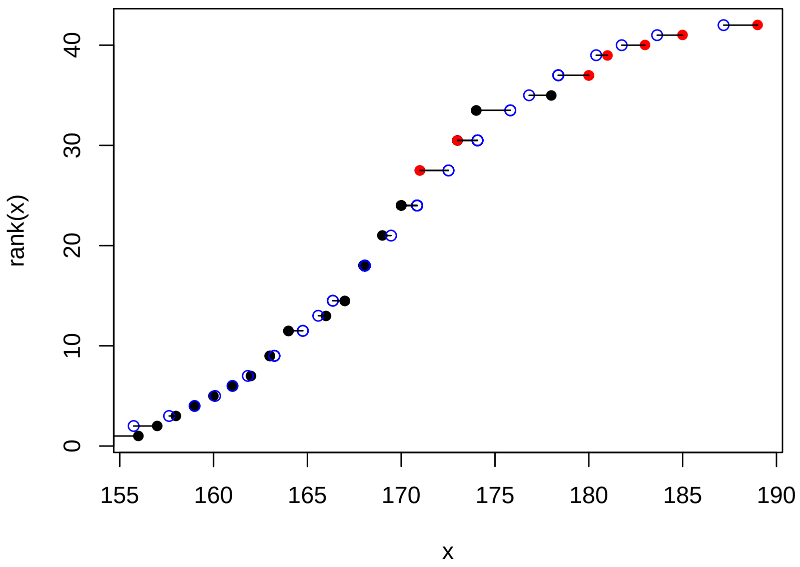
Kumulativa

```
  x <- visina
  plot(x, rank(x), pch = 16, col = spol)
```



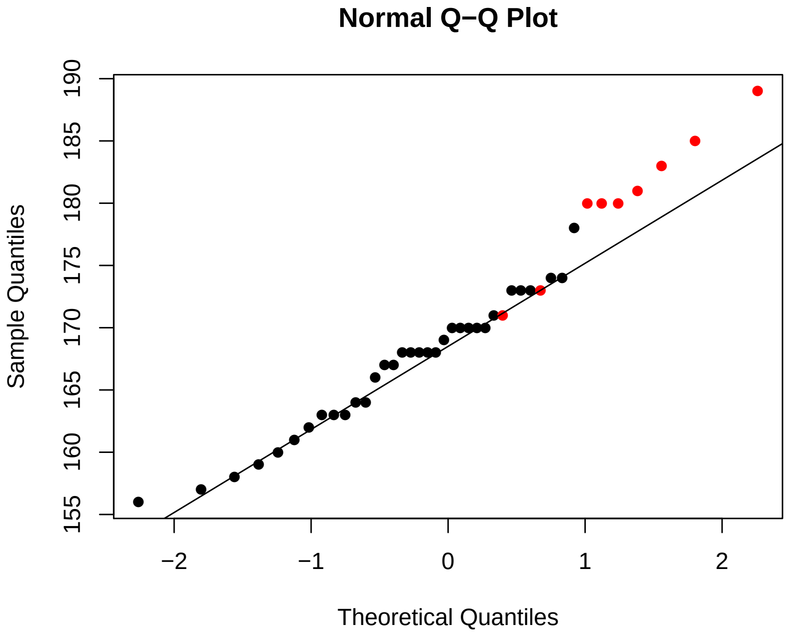
Kumulativa in normalna aproksimacija

```
  x <- visina
  plot(x, rank(x), pch = 16, col = spol)
  q <- qnorm((rank(x) - 0.5)/length(x), mean(x),
+     sd(x))
  points(q, rank(x), col = 4)
  segments(x, rank(x), q, rank(x))
```



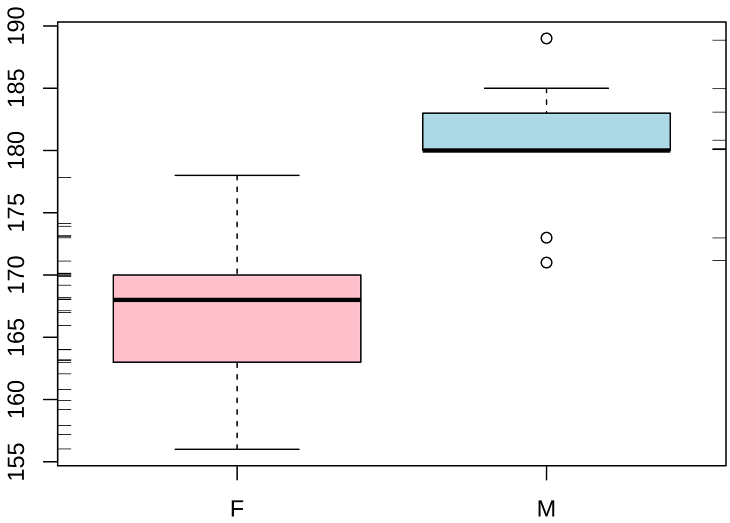
Slika kvantilov

```
qqnorm(visina, col = spol, pch = 16)
qqline(visina)
```



Boxplot

```
boxplot(visina ~ spol, col = c("pink", "lightblue"))
rug(jitter(visina[spol == "F"]), side = 2)
rug(jitter(visina[spol == "M"]), side = 4)
```



Dorišite točke za mediane. Pomagajte si s `str()`, `locator()`.

2 Testiranje višin

Student t-test

```
t.test(visina ~ spol)
```

```
	Welch Two Sample t-test

data:  visina by spol
t = -6.4643, df = 12.502, p-value = 2.55e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -17.901862  -8.906219
sample estimates:
mean in group F mean in group M
       166.8182        180.2222
```

Lahko tudi tako:

```
t.test(visina[spol == "F"], visina[spol == "M"])
```

Oglejte si, kaj vrne funkcija `t.test()`. Dorišite točki povprečij.

3 Teža

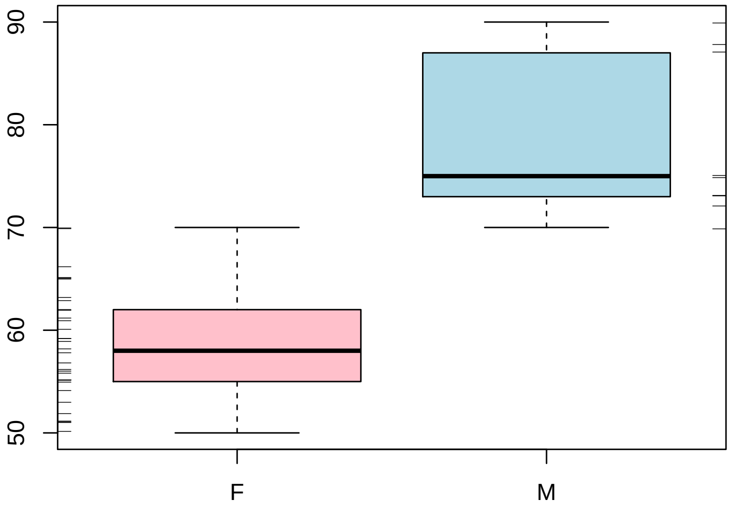
Teža in spol

Izberite si nekaj prejšnjih prikazov in

- Raziščite kako je s težo pri dekletih in fantih.
- Izračunajte novo spremenljivko $BMI = masa/visina^2$
- Kaj pa velja za BMI?

```
  t.test(masa ~ spol)

        Welch Two Sample t-test

data:  masa by spol
t = -7.0271, df = 10.154, p-value = 3.324e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -25.71686 -13.35385
sample estimates:
mean in group F mean in group M
       58.57576        78.11111
```

```
  boxplot(masa ~ spol, col = c("pink", "lightblue"))
  rug(jitter(masa[spol == "F"]), side = 2)
  rug(jitter(masa[spol == "M"]), side = 4)
```



4 Galton in višina otrok in staršev

Velikost staršev in potomcev

Galton je ugotavljal korelacijo med velikostjo staršev in potomcev.

Uvedel je pojem regresija, ki izvira iz ugotovitve, da so velikost staršev in potomcev v posebnem razmerju, ki zagotavlja 'regesijo' k povprečju.

Fantje

```
  with(data, plot(oce, visina, col = spol, pch = 16,
+     xlim = range(visina)))
  abline(c(0, 1), col = "blue")
  abline(lm(visina ~ oce, data = data), col = 3,
+     lwd = 3)
  abline(lm(visina ~ oce, data = data[data$spol ==
+     "M", ]), col = "red", lwd = 3)
  abline(lm(visina ~ oce, data = data[data$spol ==
+     "F", ]), lwd = 3)
```

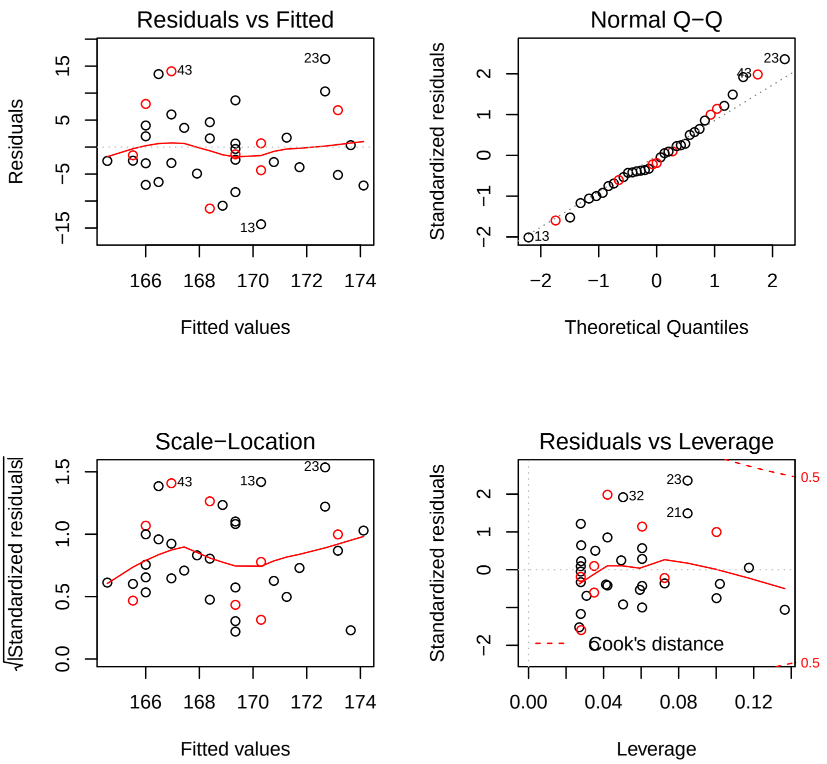
Koeficienti

```
  fit <- lm(visina ~ oce, data = data)
  summary(fit)
```

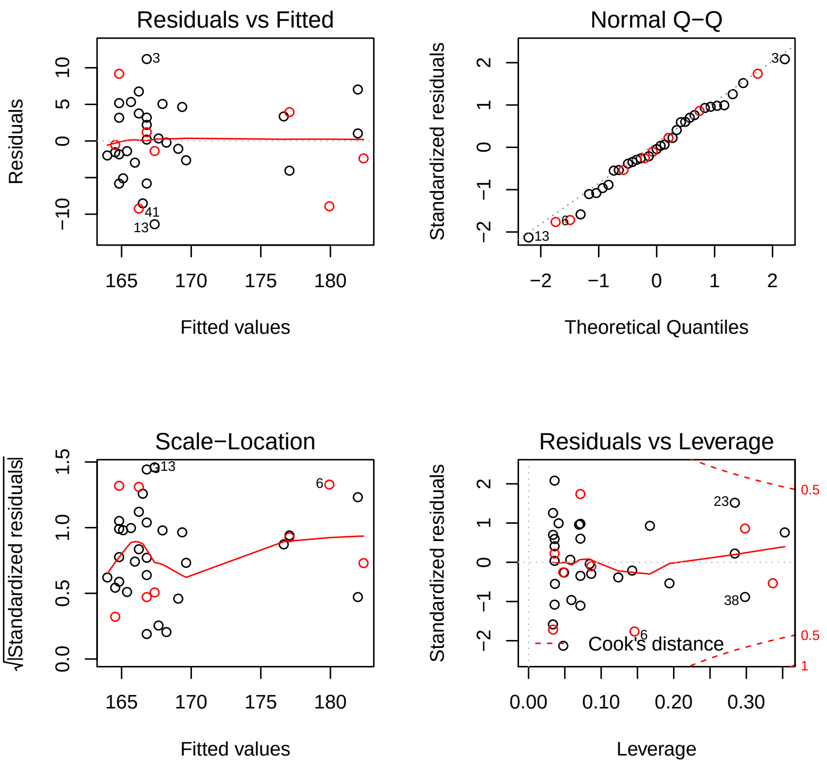
```
Call:
lm(formula = visina ~ oce, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-14.298  -4.298  -1.343   3.998  16.315 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)  83.4128    39.1670   2.130   0.0403 *
oce           0.4774     0.2186   2.183   0.0358 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.231 on 35 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared: 0.1199,	Adjusted R-squared: 0.09474 
F-statistic: 4.767 on 1 and 35 DF,  p-value: 0.0358 
```

Grafična analiza

```
  par(mfrow = c(2, 2))
  plot(fit, col = spol)
```

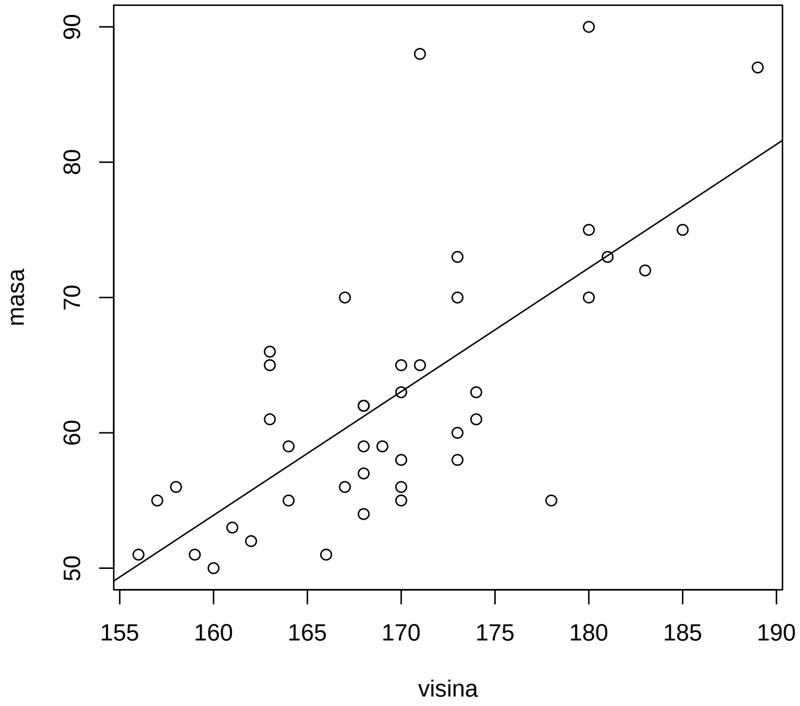


```
  fit <- lm(visina ~ spol * oce, data = data)
  summary(fit)
Call:
lm(formula = visina ~ spol * oce, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-11.3717  -2.6359  -0.2208   3.7604  11.1943 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept) 115.8618    34.2922   3.379  0.00188 **
spolM       -10.4524    72.3167  -0.145  0.88596   
oce           0.2830     0.1920   1.474  0.14986   
spolM:oce     0.1264     0.4003   0.316  0.75420   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.475 on 33 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared: 0.5244,	Adjusted R-squared: 0.4812 
F-statistic: 12.13 on 3 and 33 DF,  p-value: 1.645e-05

  par(mfrow = c(2, 2))
  plot(fit, col = spol)
```



Regresija

```
plot(visina, masa)
abline(lm(masa ~ visina))
```



Regresija

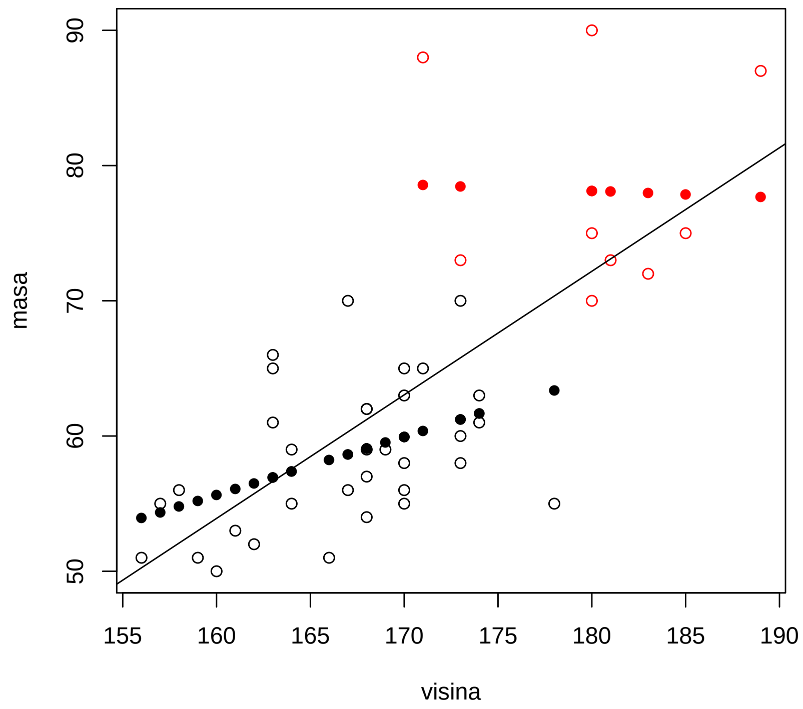
```
  cor(masa, visina)
[1] 0.7049331
  fit <- lm(masa ~ visina)
  summary(fit)
Call:
lm(formula = masa ~ visina)

Residuals:
    Min      1Q  Median      3Q     Max
-15.354  -4.140  -1.786   3.579  24.042

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -92.2700    24.6887  -3.737 0.000581 ***
visina        0.9136     0.1453   6.286 1.87e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.202 on 40 degrees of freedom
Multiple R-squared: 0.4969,     Adjusted R-squared: 0.4844
F-statistic: 39.51 on 1 and 40 DF,  p-value: 1.874e-07
```

Regresija

```
  fit <- lm(masa ~ visina * spol)
  summary(fit)
```

```
Call:
lm(formula = masa ~ visina * spol)

Residuals:
   Min     1Q Median     3Q    Max 
-8.381 -4.148 -1.588  3.022 11.878 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)  
(Intercept)  -13.1070    31.2526  -0.419   0.6773  
visina         0.4297     0.1872   2.295   0.0274 *
spolM        100.1884    73.0398   1.372   0.1782  
visina:spolM  -0.4795     0.4113  -1.166   0.2509  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.738 on 38 degrees of freedom
Multiple R-squared: 0.6966,	Adjusted R-squared: 0.6727 
F-statistic: 29.09 on 3 and 38 DF,  p-value: 6.079e-10
```

```
  plot(visina, masa, col = spol)
  abline(lm(masa ~ visina))
  points(visina, predict(fit), pch = 16, col = spol)
```



Analiza variance

```
  fvis <- cut(visina, breaks = c(155, 165, 175,
+     200), labels = c("M", "S", "V"))
  table(fvis)
```

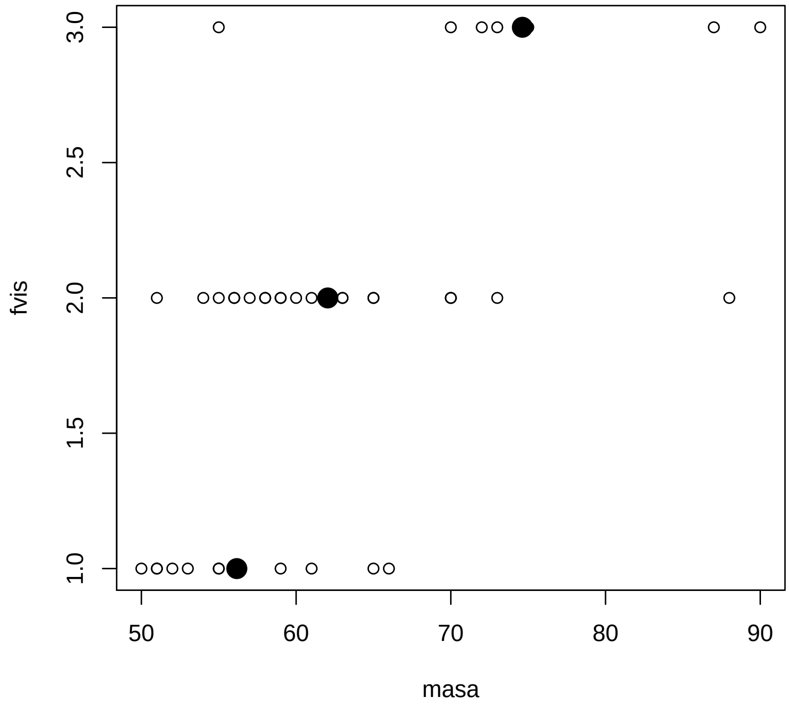
```
fvis
 M  S  V
12 22  8
```

```
   (m <- by(masa, fvis, mean))
```

```
fvis: M
[1] 56.16667
---------------------------------------------
fvis: S
[1] 62.04545
---------------------------------------------
fvis: V
[1] 74.625
```

AOV

```
plot(masa, fvis)
points(m, 1:3, pch = 16, cex = 2)
```

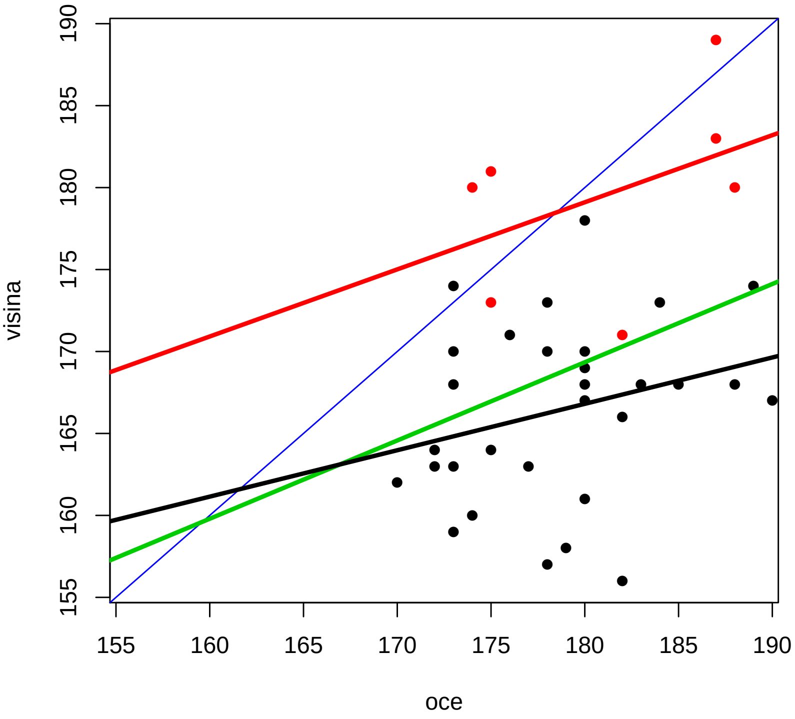


AOV

```
fit <- lm(masa ~ 0 + fvis)
summary(fit)
```

```
Call:
lm(formula = masa ~ 0 + fvis)

Residuals:
    Min      1Q  Median      3Q     Max 
-19.625  -4.510  -1.167   2.924  25.954 

Coefficients:
      Estimate Std. Error t value Pr(>|t|)    
fvisM   56.167      2.295   24.48   <2e-16 ***
fvisS   62.045      1.695   36.61   <2e-16 ***
fvisV   74.625      2.811   26.55   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.949 on 39 degrees of freedom
Multiple R-squared: 0.9855,	Adjusted R-squared: 0.9843 
F-statistic: 881.4 on 3 and 39 DF,  p-value: < 2.2e-16
```

Fantje

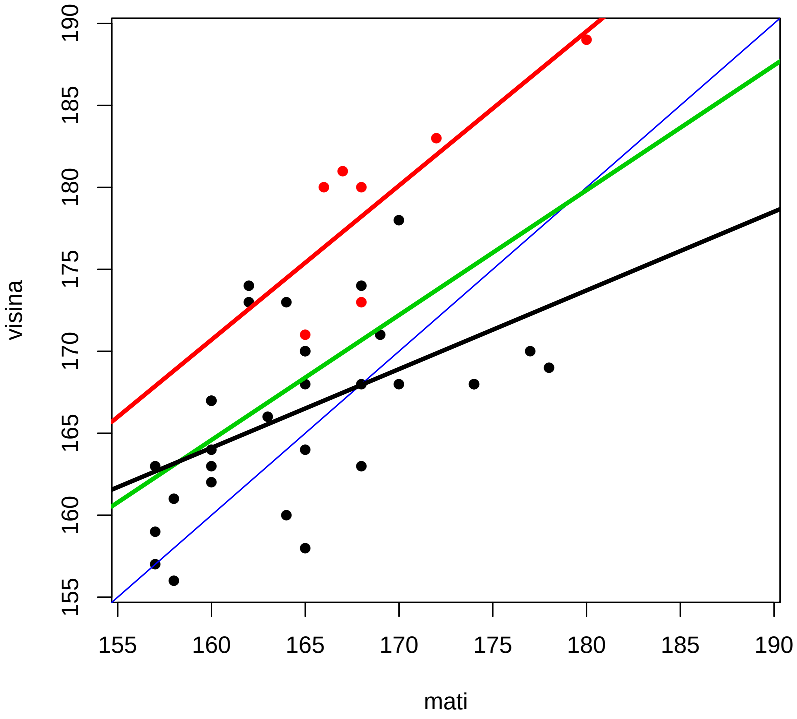


visina
oce
155
160
165
170
175
180
185
190
155
160
165
170
175
180
185
190

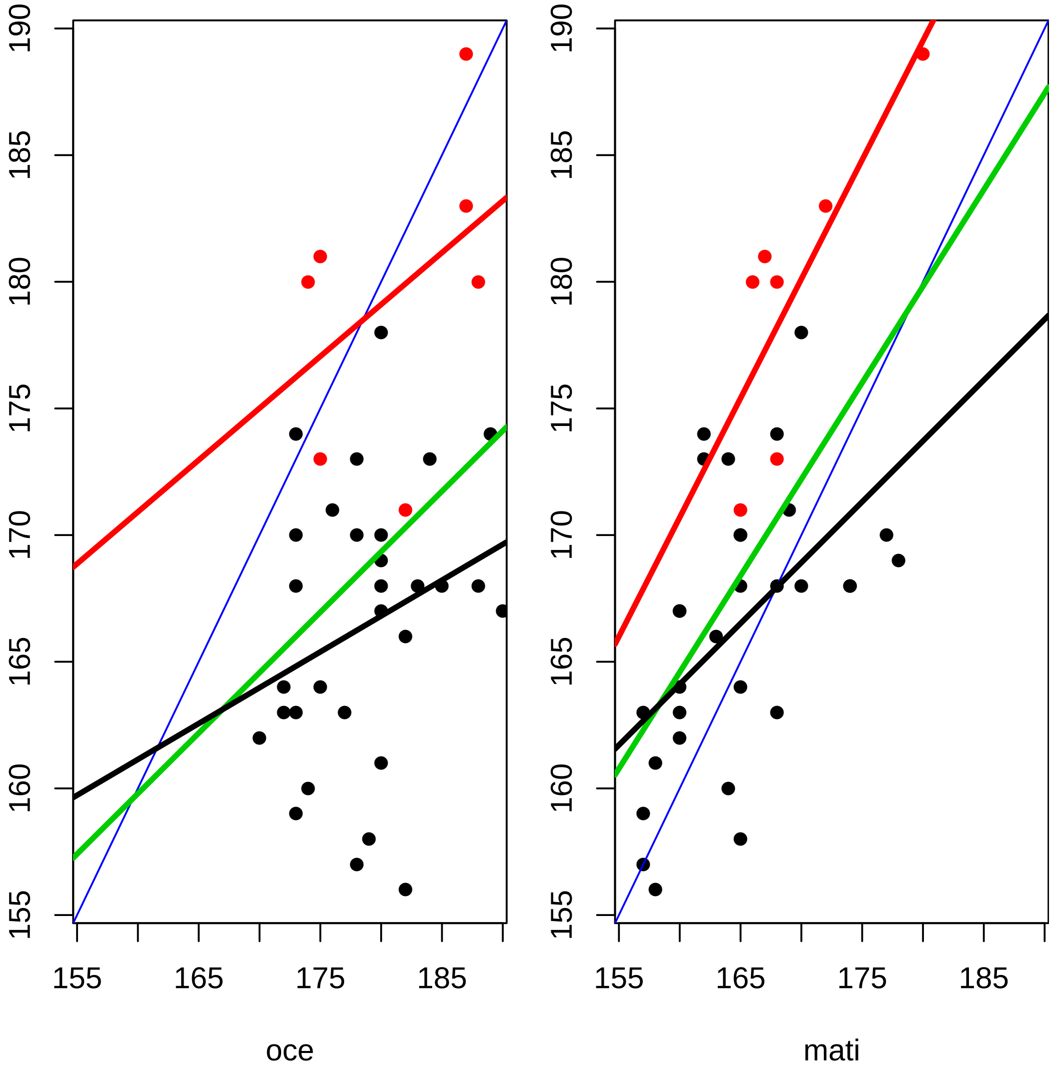
Dekleta

```
  with(data, plot(mati, visina, col = spol, pch = 16,
+      xlim = range(visina)))
  abline(c(0, 1), col = "blue")
  abline(lm(visina ~ mati, data = data), col = 3,
+      lwd = 3)
  abline(lm(visina ~ mati, data = data[data$spol ==
+      "M", ]), col = "red", lwd = 3)
  abline(lm(visina ~ mati, data = data[data$spol ==
+      "F", ]), lwd = 3)
```

Dekleta



Fantje in dekleta



SessionInfo

Windows 7 x64 (build 7601) Service Pack 1

- R version 2.15.1 (2012-06-22), x86_64-pc-mingw32
- Locale: LC_COLLATE=Slovenian_Slovenia.1250, LC_CTYPE=Slovenian_Slovenia.1250, LC_MONETARY=Slovenian_Slovenia.1250, LC_NUMERIC=C, LC_TIME=Slovenian_Slovenia.1250
- Base packages: base, datasets, graphics, grDevices, stats, utils
- Other packages: patchDVI 1.9
- Loaded via a namespace (and not attached): tools 2.15.1

Project path: D:/_Y/R/rps
Main file : ../doc/Opisna.Rnw

View as vignette

Project files can be viewed by pasting this code to R console:

```
> projectName <-"rps";  mainFile <-"Opisna"

  commandArgs()
  library(tkWidgets)
  openPDF(file.path(dirname(getwd()), "doc", paste(mainFile,
+     "PDF", sep = ".")))
  viewVignette("viewVignette", projectName, file.path("../doc",
+     paste(mainFile, "Rnw", sep = ".")))
```