RPS
Analiza podatkov

A. Blejec

12. marec 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 <- file.path("../data", lfn)
> 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 kaj opazite?

Nenavadni podatki

Kaj storiti s tistim, ki je napisal, da je rojen v mesecu 0?

Eden pa je star 59 let??

Nadaljevanje opisa

```
> summary(data[, 7:dim(data)[2]])
     cevelj      lasje  oci         mati
 Min.   :36.00   S:19   S:24   Min.   :155.0
 1st Qu.:38.00   T:24   T:19   1st Qu.:160.0
 Median :39.00                 Median :165.0
 Mean   :40.02                 Mean   :165.4
 3rd Qu.:41.50                 3rd Qu.:168.0
 Max.   :48.00                 Max.   :180.0
```

```
                                NA's   :5
     oce         majica
Min.   :170.0   L : 5
1st Qu.:174.2   M :19
Median :179.5   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$mati)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  155.0   160.0   165.0   165.4   168.0   180.0       5
> by(data$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$spol: M
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  171.0   178.5   180.0   180.0   182.5   189.0
> summary(data$oce)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  170.0   174.2   179.5   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[,v̈isina]̈`
- `data[,5]`

Neposreden dostop

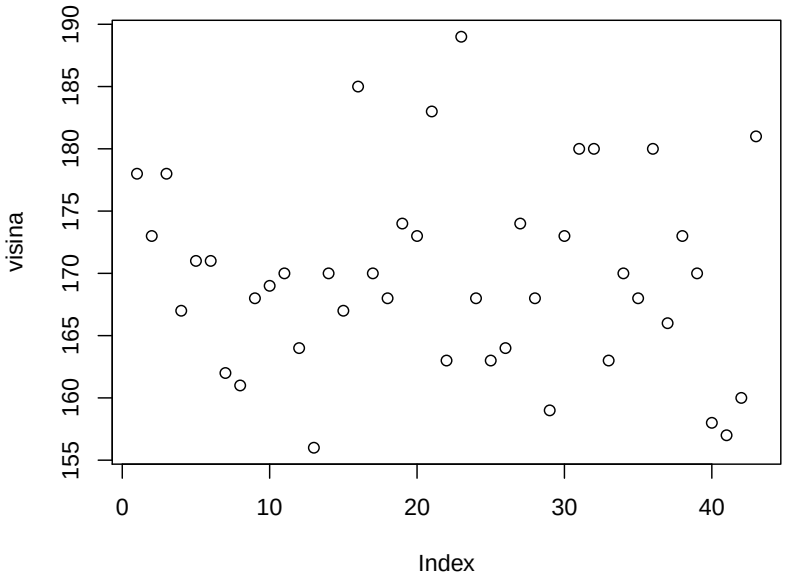
Neposreden dostop do spremenljivk omogoči

```
> attach(data)
> length(visina)
[1] 43
> visina[1:5]
[1] 178 173 178 167 171
```

Grafični prikazi

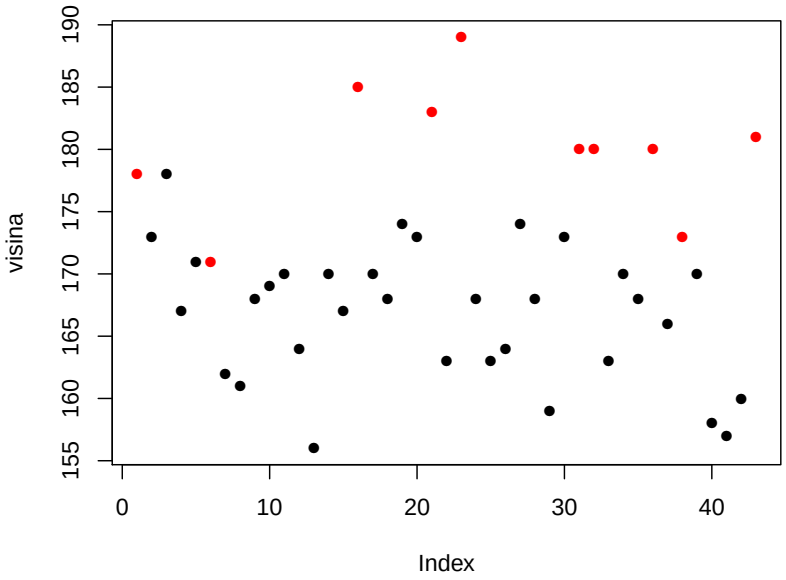
Grafični prikaz podatkov

```
> plot(visina)
```



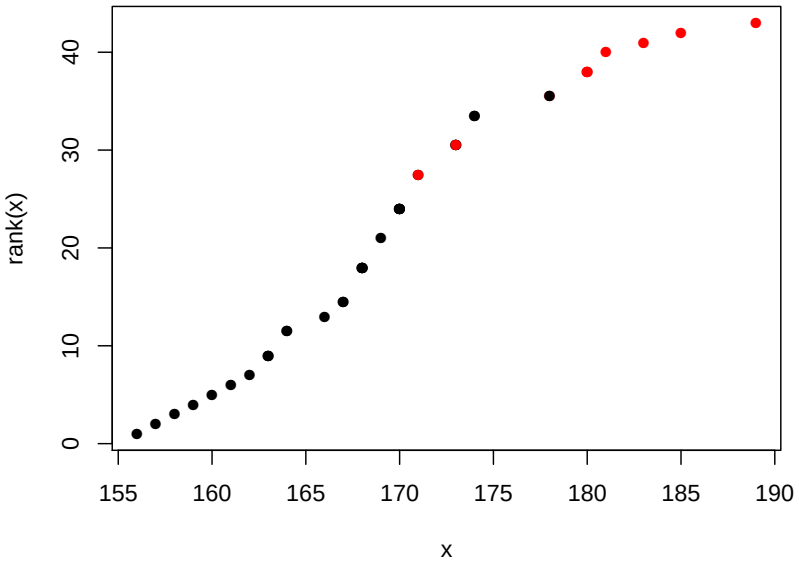
Grafični prikaz podatkov

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



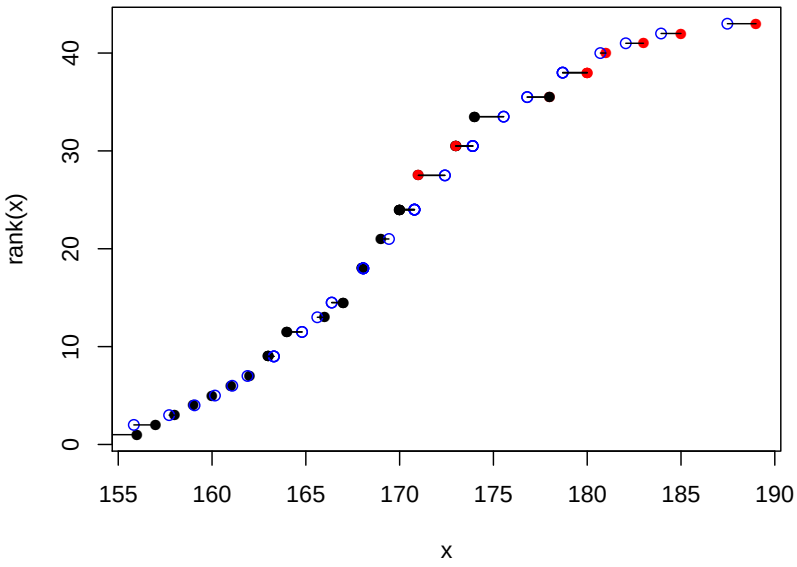
Kumulativa

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



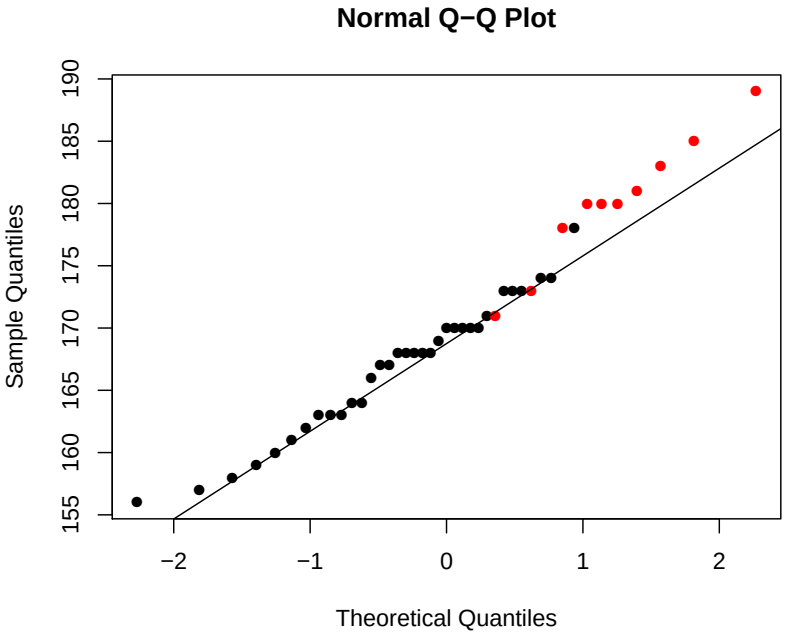
Kumulativa in normalna aproksimacija

```
> 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))
```



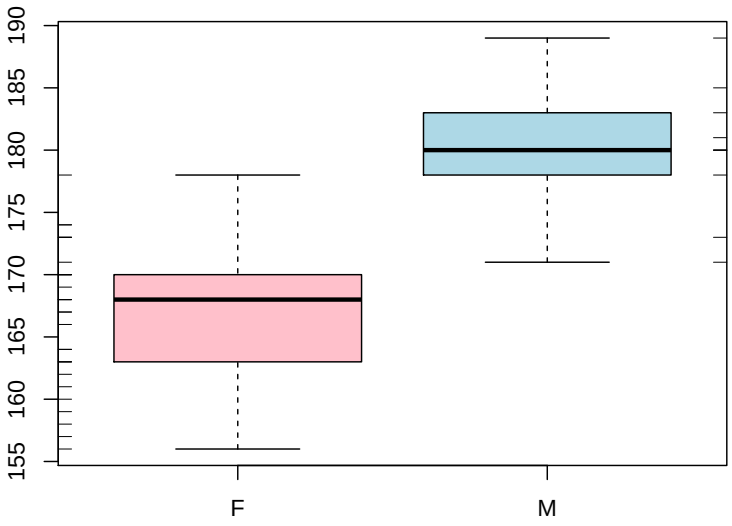
Slika kvantilov

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



Boxplot

```
> boxplot(visina ~ spol, col = c("pink", "lightblue"))
> rug(visina[spol == "F"], side = 2)
> rug(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 = -6.8838, df = 15.244, p-value = 4.77e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -17.257671  -9.105965
sample estimates:
mean in group F mean in group M
       166.8182        180.0000
```

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.
- Kaj pa velja za BMI ($BMI = masa/visina^2$)

4 Galton in višina otrok in staršev

Velikost staršev in potomcev

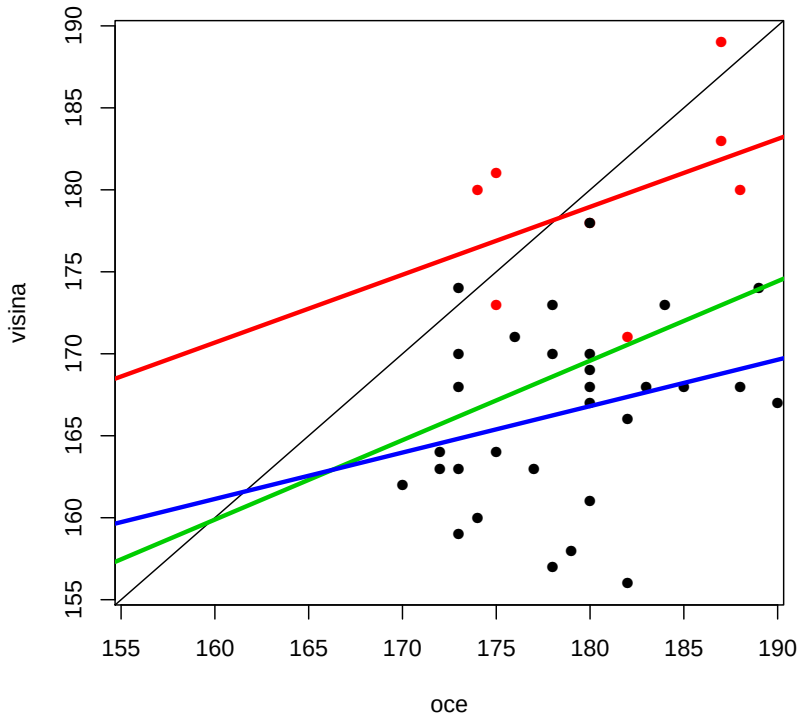
Galton je ugotavljal korelacijo med velikostjo staršev in potomcev.

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))
> 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", ]), col = "blue", lwd = 3)
```

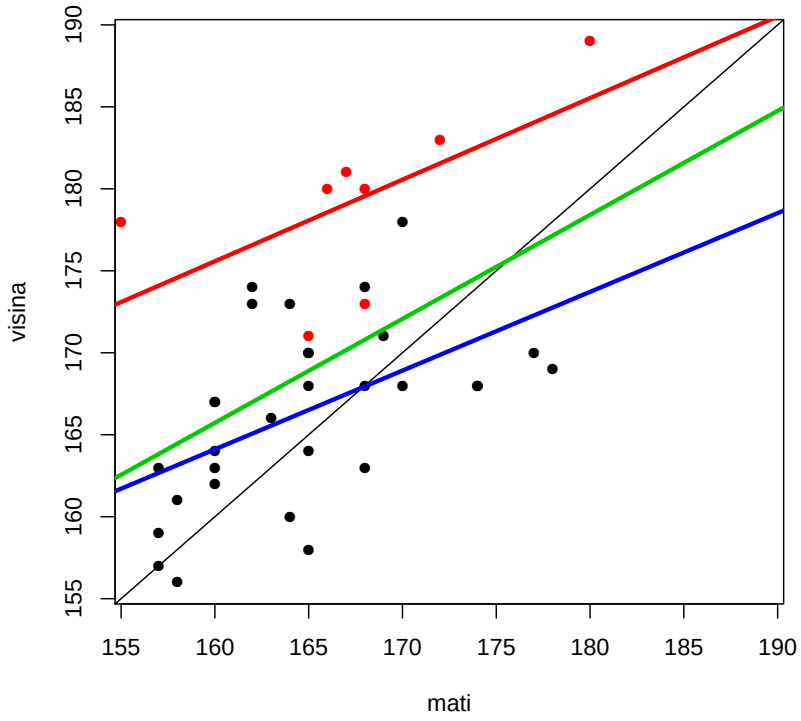
Fantje



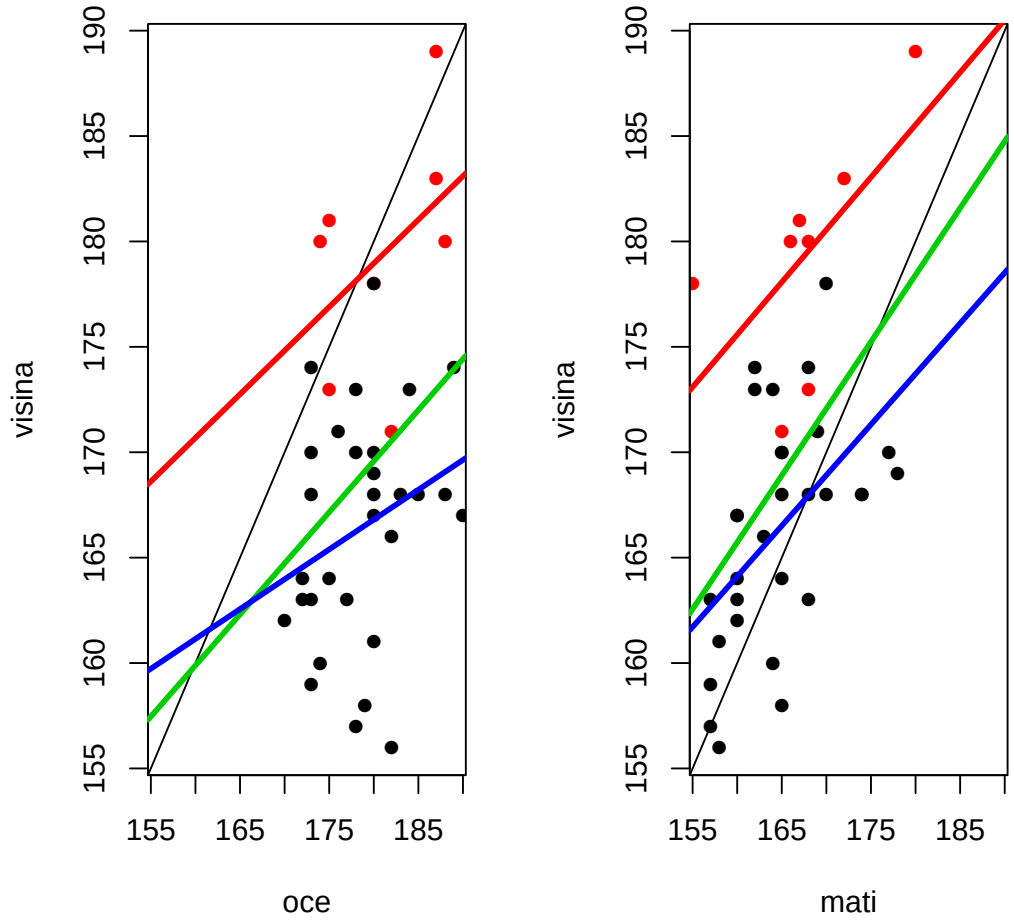
Dekleta

```
> with(data, plot(mati, visina, col = spol, pch = 16,
+     xlim = range(visina)))
> abline(c(0, 1))
> 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", ]), col = "blue", 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 = ".")))
```