Domácí úkol 1

Průzkumová analýza dat a metoda hlavních komponent

Marie Melínová

Než začnu s plněním samotného domácího úkolu, musím si načíst data. Dle zadání mám pořadové číslo 9 a tím pádem pro další prácí nahrávám soubor s příslušným pořadovým číslem.

Datový soubor je podobný, jako s kterým jsme pracovali na cvičení č.3, nyní však budeme pracovat s daty o ženách. Data mají 254 pozorování a 7 proměnných, ze kterých slouží proměnná "id" jako identifikátor.

```
library(foreign)
data = read.spss("du1_9.sav",to.data.frame=TRUE)
```

Než budeme dále pokračovat, přejmenujeme si jednotlivé proměnné tak, aby se nám s nimi dále lépe pracovalo. Dále si také míry uložíme do nového datového souboru, který nazveme *miry*.

```
prom <- c("ID", "RAMENA", "HRUDNIK", "BOKY", "PREDLOKTI", "KOLENA", "ZAPESTI")
colnames(data) <- prom
miry <- data[,c(-1)]
head(miry)
```

```
##   RAMENA HRUDNIK  BOKY PREDLOKTI KOLENA ZAPESTI
## 1  100.1    81.1  90.4      22.3   34.2    14.5
## 2  111.4    94.5 110.2      26.8   40.5    16.6
## 3  100.3    85.3  95.9      23.9   35.6    15.2
## 4  107.2    98.2  94.9      25.0   37.7    16.2
## 5  101.5    91.0 103.0      26.5   38.2    16.0
## 6   98.5    80.8  85.8      21.5   33.0    14.0
```

Posouzení normality jednotlivých proměnných

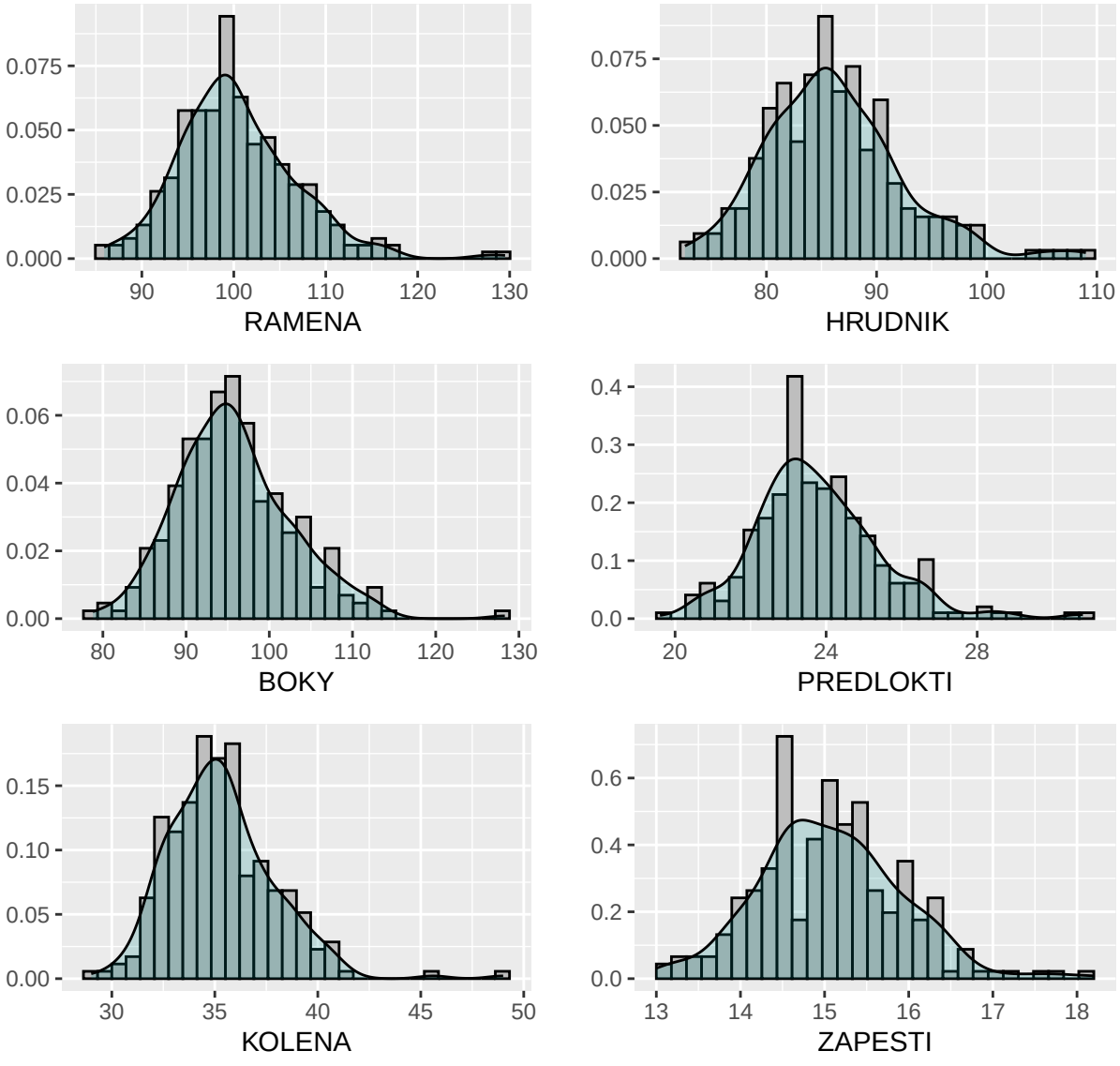
Ze všeho nejdřív se podíváme na základní statistiky, které se našeho souboru týkají. Můžeme si všimnout, že šikmost žádné proměnné nepřesahuje 1 a nejvyšší špičatost má proměnná *KOLENA*.

```
library(psych)
describe(miry)[,-c(1,2,7, 10)]
```

```
##             mean   sd median trimmed  min   max skew kurtosis   se
## RAMENA    100.36 6.50  99.50  100.01 85.9 129.5 0.82     1.92 0.41
## HRUDNIK    86.14 6.21  85.50   85.75 72.6 109.0 0.73     1.16 0.39
## BOKY       95.79 6.94  95.05   95.45 78.8 128.3 0.65     1.31 0.44
## PREDLOKTI  23.79 1.68  23.60   23.68 19.6  30.8 0.85     1.90 0.11
## KOLENA     35.29 2.59  35.10   35.14 29.0  49.0 0.97     2.96 0.16
## ZAPESTI    15.07 0.85  15.00   15.05 13.0  18.2 0.36     0.57 0.05
```

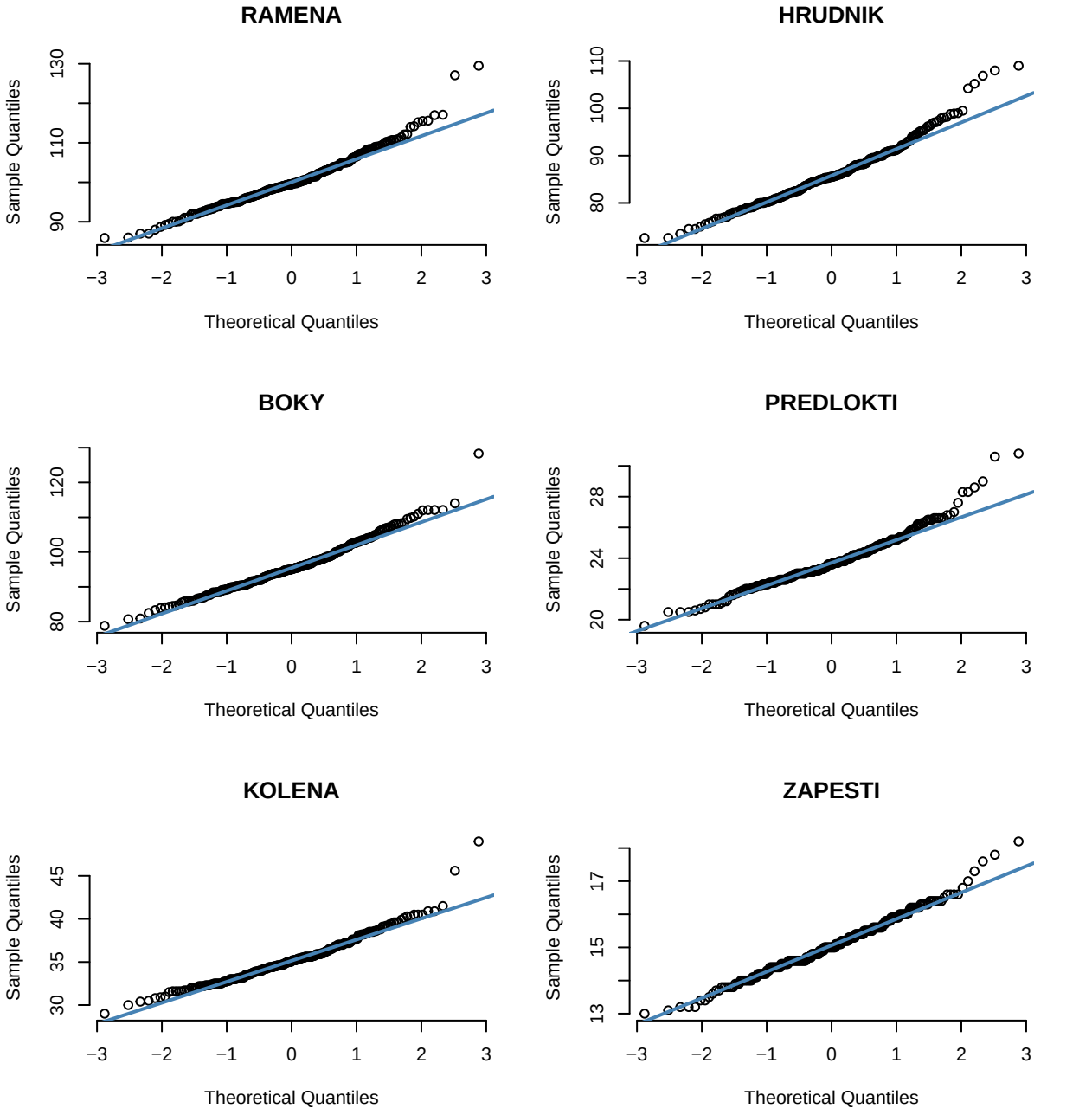
Nyní se na rozdělení jednotlivých proměnným podíváme i graficky.

```
library(ggplot2)
library(gridExtra)

plot <- list()

for (i in 1:ncol(miry)) {
  plot[[i]] <- ggplot(data = miry, aes_string(x = miry[[i]])) +
    geom_histogram(aes(y = after_stat(density)), colour = "black", fill = "grey") +
    geom_density(alpha=0.2, fill="darkcyan") + labs(x = prom[i+1],y="")
}

do.call(grid.arrange,plot)
```



Dále se pokusím sestrojit Q-Q ploty, které nám pomohou lépe posoudit porušení (či neporušení) normality jednotlivých proměnných.

```
par(mfrow=c(3,2))

for (i in 1:ncol(miry)){
  qqnorm(miry[[i]], pch = 1, frame = FALSE, main = prom[i+1])
  qqline(miry[[i]], col = "steelblue", lwd = 2)
}
```

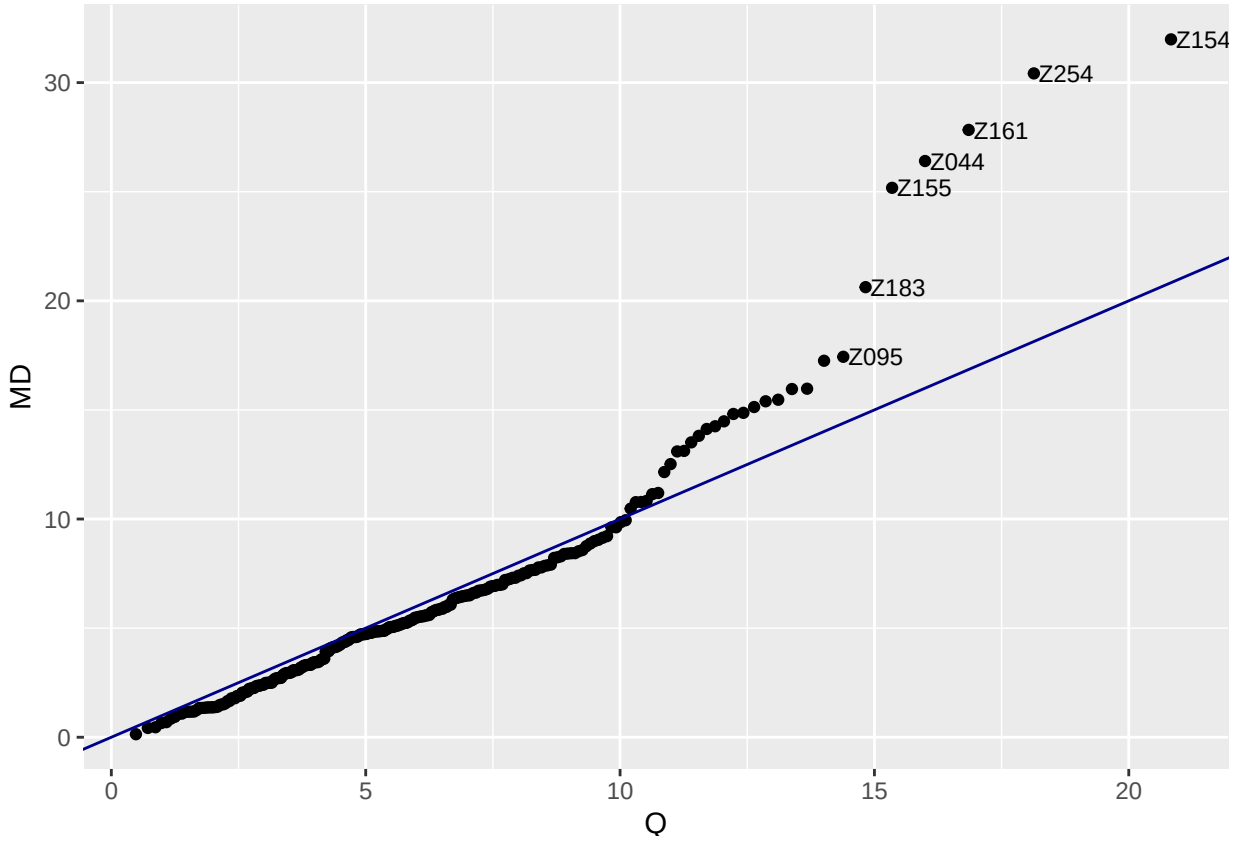


U všech Q-Q plotů si můžeme všimnout levostranného sešikmení, které se projevují odlehlými pozorováními nad naší pomocnou modrou čarou.

Posouzení vícerozměrné normality

K posouzení vícerozměrné normality využijeme χ^2 diagram, který využívá Mahalnobisobu vzdálenost. V grafu snadno nalezneme pozorování, která příliš narušují vícerozměrnou normalitu - na pravé straně rozdělení máme velký počet odlehlých pozorování.

```
data$MD <- mahalanobis(miry, colMeans(miry), cov(miry))
data$rank <- rank(data$MD)
data$p <- (data$rank - 0.5)/nrow(data)
data$Q <- qchisq(data$p, 6, ncp = 0, lower.tail = TRUE)

ggplot(data, aes(x = Q,y = MD, label = ifelse(nrow(data) - rank < 7, as.character(ID),''))) +
  geom_point() +
  geom_text(hjust = 0, nudge_x = 0.1, size = 3) +
  geom_abline(aes(intercept = 0,slope = 1), color = "darkblue")
```



Identifikace odlehlých pozorování

Můžeme vidět, že každá proměnná má určité odlehlé pozorování, díky kterým jsou rozdělení jednotlivých proměnných sešikmené. Nyní se pokusíme o jejich identifikaci.

```
outliers <- list()

for (i in 1:ncol(miry)) {
  outliers[[i]] <- data.frame(
    promenna = colnames(miry)[i],
    pořadí = which(miry[[i]] %in% boxplot.stats(miry[[i]])$out),
    hodnota = miry[which(miry[[i]] %in% boxplot.stats(miry[[i]])$out),1])
}

outliers <- do.call(rbind, outliers)
table(outliers$promenna)
```

```
##
##      BOKY   HRUDNIK    KOLENA PREDLOKTI    RAMENA   ZAPESTI
##         2         5         2         7         4         4
```

Řekla bych, že vzhledem k ženské stavbě těla je vysvětlení odlehlých pozorování u proměnné *BOKY* a *HRUDNIK* celkem jasná. Ač může být daná žena v ostatních rozměrech považována za "normální", rozměry hrudníku (či boků) se mohou značně odlišovat.

Co se týče dalších proměnných, řekla bych, že se bude jednat o ženy, které se vymykají všem proměnným najednou z důvodu vyšší tělesné hmotnosti.

Určení domenzionality dat, interpretace komponent

Nejdříve nalezneme jednotlivé komponenty v našich date, a kolik variability vysvětlují. K tomu využijeme funkci PCA, která je přímo v základních funkcích R.

Vzhledem k tomu, že máme všechny proměnné ve stejných jednotkách, můžeme s parametrem *scale* analyzovat jak kovarianční, tak korelační matici. Já jsem za tento parametr zvolila hodnotu *TRUE*, tedy budu analyzovat korelační matici.

Dále si také znázorníme matici komponentních zátěží.

```
PCA <- prcomp(miry, scale = TRUE)
round(t(PCA$sdev*t(PCA$rotation)), 8)
```

```
##                 PC1         PC2         PC3         PC4         PC5         PC6
## RAMENA    0.8720417  0.34814538 -0.11052808  0.26985023 -0.07155099  0.16787747
## HRUDNIK   0.8818038  0.38122714 -0.00653289 -0.09916311  0.02935885 -0.25758494
## BOKY      0.8698540 -0.01312037  0.40227727 -0.21522851 -0.14716429  0.11564731
## PREDLOKTI 0.9220452 -0.08628673 -0.13440990 -0.13950892  0.30026574  0.12123966
## KOLENA    0.8542773 -0.34537354  0.24137759  0.27420126  0.06199216 -0.11676858
## ZAPESTI   0.8424004 -0.30121899 -0.39179477 -0.07866957 -0.19622371 -0.03785463
```

Na základě matice komponentních zátěží si můžeme všimnout, že první komponenta silně koreluje se všemi proměnnými.

```
summary(PCA)
```

```
## Importance of components:
##                           PC1    PC2     PC3    PC4     PC5     PC6
## Standard deviation     2.1411 0.6958 0.63555 0.4794 0.40018 0.37104
## Proportion of Variance 0.7641 0.0807 0.06732 0.0383 0.02669 0.02295
## Cumulative Proportion  0.7641 0.8447 0.91206 0.9504 0.97705 1.00000
```

Z výpisu funkce PCA můžeme vidět, že první komponenta nám vysvětluje 76.41 % celkové variability. Další komponenty už nepřesahují více než 10 % vysvětlené variability. Můžeme tedy říct, že prostorová výraznost žen v našem vzorku se dá vysvětlit jednou silnou dimenzí a nekolika slabšími.

Pro lepší názornost si ještě znázorníme tzv. SCREE plot, ve kterém můžeme spatřit postupné klesání vlastních čísel hlavních komponent. I zde můžeme vidět, že velký zlom nastává po první komponentě.

```
screeplot(PCA, type="lines")
```