Domácí úkol 4

Diskriminační analýza

Marie Melínová

Zadání úkolu č. 4 je jednoduché. V datovém souboru máme obvody břicha, předloktí a kolene u mužů a žen. Úkolem je pak tedy zjistit, jestli na základě těchto rozměrů lze určit, které z nezařazených osob jsou ženy a které muži.

Data si klasicky načteme funkcí `read.spss()` a pro přehlednost přejmenujeme jednotlivé proměnné. Pak si vytvoříme tři samostnatné datasety - **jeden**, kde jsou údaje pouze za muže, **druhý**, kde jsou údaje jen za ženy a **třetí**, který obsahuje 10 nezařazených lidí, které se budeme snažit zařadit.

```
library(foreign)
data = read.spss("du4_9.sav", to.data.frame = TRUE)

colnames(data) <- c("ID", "oBRICHO", "oPREDLOKTI", "oKOLENE", "POHLAVI")
levels(data$POHLAVI) <- c("2", "0", "1")  #0 -> muž, 1 -> žena, 2 -> nezařazeno

muzi <- data[data$POHLAVI == 0, 2:4]
zeny <- data[data$POHLAVI == 1, 2:4]
nez <- data[data$POHLAVI == 2, 2:4]
```

Průzkum dat z hlediska předpokladů a použitelnosti diskriminační analýzy

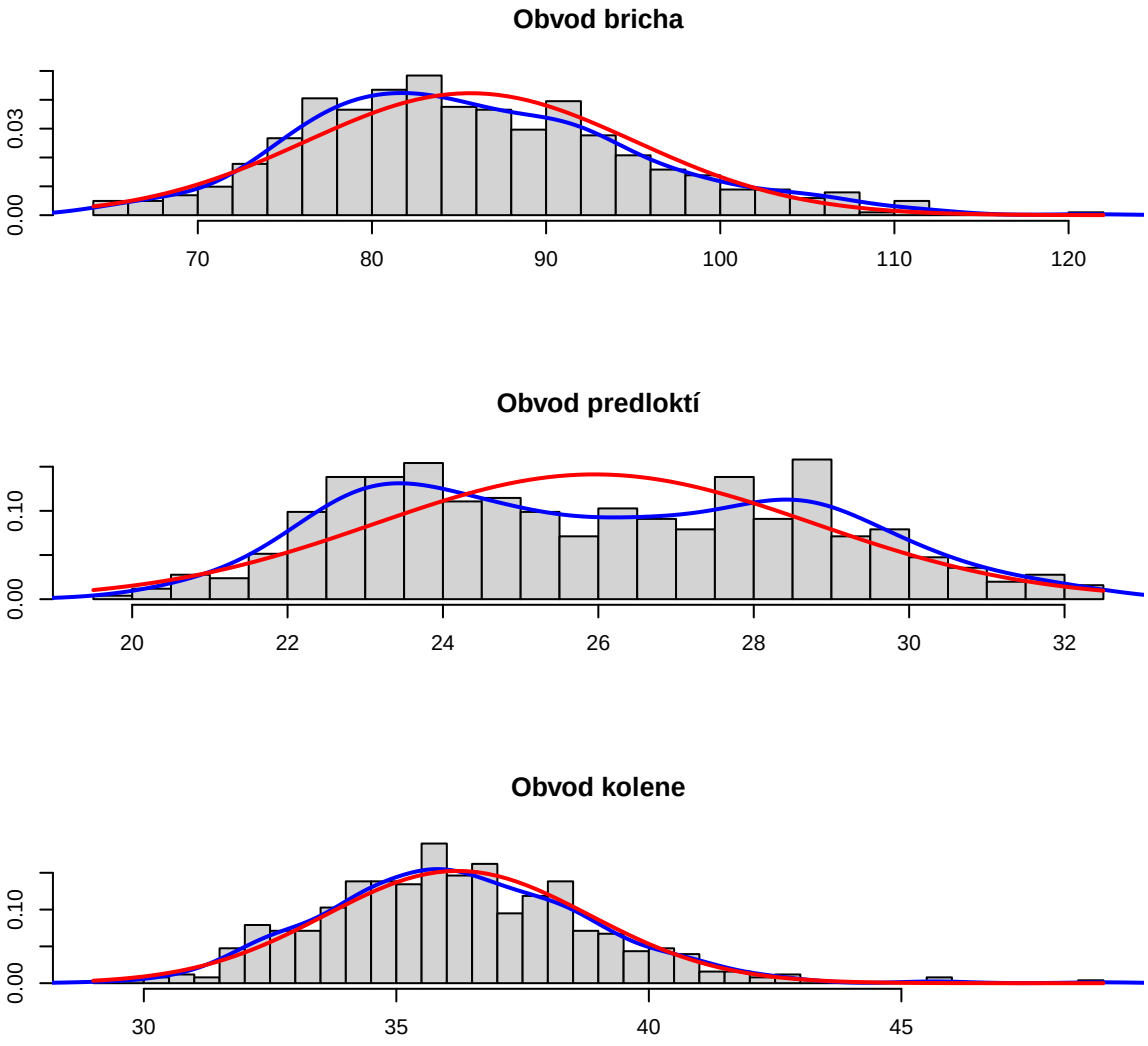
Před použitím diskriminační analýzy se musíme podívat, zda-li jsou data pro použití této metody vhodná. Jako první se ujistíme, že všechny řádky matice jsou "plné" - čili, že v datasetech nejsou chybějící hodnoty. Následně se alespoň okrajově ujistíme, že hodnoty u každé kategorie pochází z normálního rozdělení.

```
# kontrola, jestli se v jednotlivých datasetech nenachází chybějící hodnoty
c(sum(is.na(muzi)), sum(is.na(zeny)), sum(is.na(nez)))
```

```
## [1] 0 0 0
```

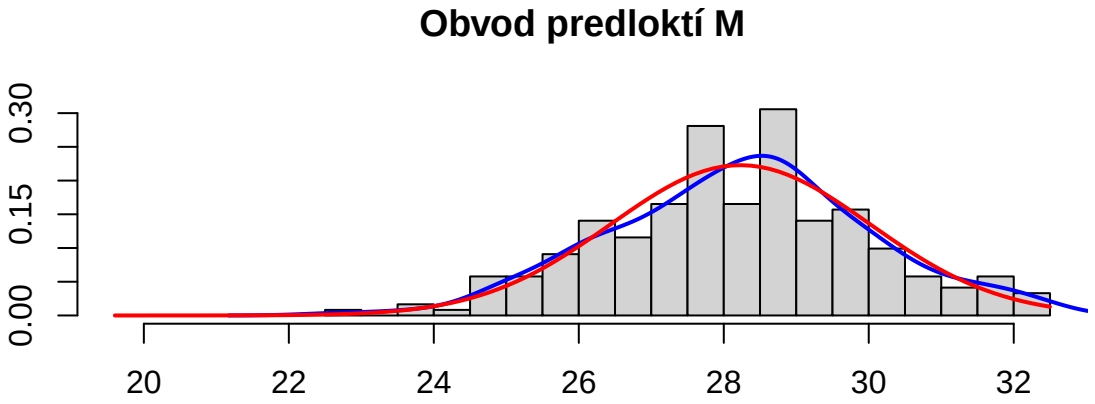
Díky této kontrole si nyní můžeme být jistí, že pro diskriminační analýzu můžeme využít všechny řádky, protože ani v jednom není žádná chybějící hodnota.

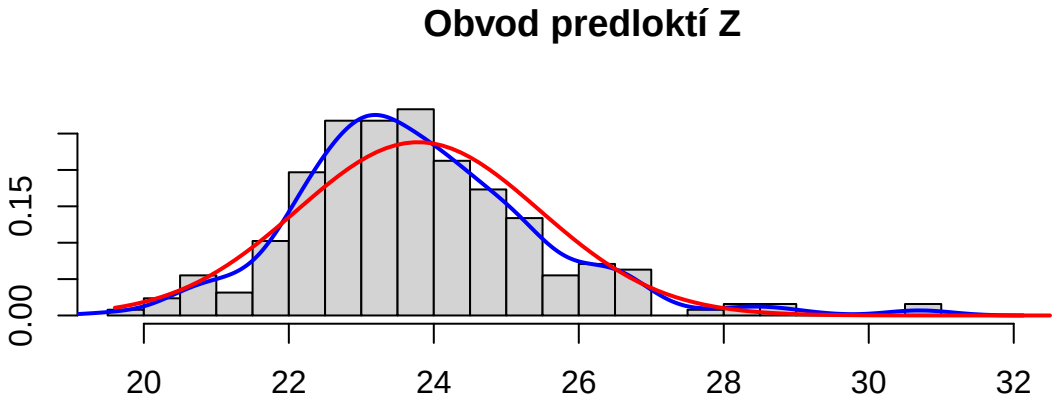
```
par(mfrow = c(3, 1))
hist(data$oBRICHO, breaks = 30, freq = F, main = "Obvod bricha", ylab = "", xlab = "")
lines(density(data$oBRICHO), col = "blue", lwd = 2)
curve(dnorm(x, mean = mean(data$oBRICHO), sd = sd(data$oBRICHO)), add = T, col = "red",
    lwd = 2)

hist(data$oPREDLOKTI, breaks = 30, freq = F, main = "Obvod predloktí", ylab = "",
    xlab = "")
lines(density(data$oPREDLOKTI), col = "blue", lwd = 2)
curve(dnorm(x, mean = mean(data$oPREDLOKTI), sd = sd(data$oPREDLOKTI)), add = T,
    col = "red", lwd = 2)

hist(data$oKOLENE, breaks = 30, freq = F, main = "Obvod kolene", ylab = "", xlab = "")
lines(density(data$oKOLENE), col = "blue", lwd = 2)
curve(dnorm(x, mean = mean(data$oKOLENE), sd = sd(data$oKOLENE)), add = T, col = "red",
    lwd = 2)
```



Již na první pohled si můžeme všimnout, že proměnná `Obvod předloktí` má bimodální rozdělení, proto si tuto proměnnou necháme vykreslit zvlášť pro ženy a zvlášť pro muže.

```
par(mfrow = c(2, 1))
hist(muzi$oPREDLOKTI, breaks = 30, freq = F, main = "Obvod predloktí M", ylab = "",
    xlab = "", xlim = c(19.6, 32.5))
lines(density(muzi$oPREDLOKTI), col = "blue", lwd = 2)
curve(dnorm(x, mean = mean(muzi$oPREDLOKTI), sd = sd(muzi$oPREDLOKTI)), add = T,
    col = "red", lwd = 2)

hist(zeny$oPREDLOKTI, breaks = 30, freq = F, main = "Obvod predloktí Z", ylab = "",
    xlab = "", xlim = c(19.6, 32.5))
lines(density(zeny$oPREDLOKTI), col = "blue", lwd = 2)
curve(dnorm(x, mean = mean(zeny$oPREDLOKTI), sd = sd(zeny$oPREDLOKTI)), add = T,
    col = "red", lwd = 2)
```





Na základě vykreslení této proměnné by se dalo říci, že rozměr předloktí odlišuje ženy a muže nejlépe. Dle grafické analýzy bychom o ostatních proměnných mohli tvrdit, že ačkoli jsou lehce sešikmené, přibližně se řídí normálním rozdělením.

Model diskriminační analýzy

Než přejdeme k vytvoření modelu pro zařazení pohlaví nezařazeným osobám, provedeme na modelu křížovou validaci. Následně pak vytvoříme predikci jednotlivých nezařazených osob.

```
library(MASS)

krizova_validace <- function(data, K = 10) {
    n <- nrow(data)
    folds <- cut(seq(1, n), breaks = K, labels = F)
    errors <- numeric(K)

    for (i in 1:K) {
        test_indices <- which(folds == 1)
        train_data <- data[-test_indices, ]
        test_data <- data[test_indices, ]

        modelLDA <- lda(POHLAVI ~ ., data = train_data)
        predicted <- predict(modelLDA, newdata = test_data)
        errors[i] <- mean(predicted$class != test_data$POHLAVI)
    }

    return(mean(errors))
}

data_train <- data[data$POHLAVI != 2, 2:5]
data_train$POHLAVI <- (as.numeric(data_train$POHLAVI) - 2)

kv <- krizova_validace(data_train)
kv
```

```
## [1] 0.16
```

Hodnota křížové validace je 0.16 a znamená, že průměrná chyba klasifikace modelu při použití křížové validace je 16 %. Tato hodnota představuje průměrný podíl špatných klasifikací v rámci jednotlivých testovacích sad.

Nyní přejdeme k definovaní samotného modelu a predikce.

```
data_test <- data[data$POHLAVI == 2, 2:4]

model <- lda(POHLAVI ~ ., data = data_train)
predictions <- model |>
    predict(data_test)
predictions$class
```

```
##  [1] 0 0 1 0 1 1 0 1 1 0
## Levels: 0 1
```

Model diskriminační analýzy zařadí nezařazené osoby **č. 1, 2, 4, 7, 10** jako muže a nezařazené osoby **č. 3, 5, 6, 8, 9** jako ženy.

Toto rozřazení si ještě zkusíme vykreslit pomocí grafu. Na hodnotě `y = 0` se nachází muži z trénovací množiny (modrá barva), na hodnotě `y = 1` se nachází ženy z trénovací množiny (červená barva) a na hodnotě `y = 2` se nachází muži a ženy z testovací množiny, neboli lidé, u kterých jsme zjišťovali zařazení pomocí modelu `lda` (rozřazeni podle barev na muže a ženy).

```
ldaData_train <- as.data.frame(matrix(ncol = 0, nrow = 496))
ldaData_train$POHLAVI <- c(data_train$POHLAVI)
ldaData_train$LDA <- c(predict(model)$x)

ldaData_test <- as.data.frame(matrix(ncol = 0, nrow = 10))
ldaData_test$POHLAVI <- (as.numeric(predictions$class) - 1)
ldaData_test$LDA <- predictions$x

plot(x = ldaData_train[ldaData_train$POHLAVI == 0, ]$LDA, y = rep(0,
    nrow(ldaData_train[ldaData_train$POHLAVI == 0, ])), ylim = c(-5,
    8), xlim = c(min(ldaData_train$LDA), max(ldaData_train$LDA)),
    col = "blue", ylab = "", xlab = "")
points(x = ldaData_train[ldaData_train$POHLAVI == 1, ]$LDA, y = rep(1,
    nrow(ldaData_train[ldaData_train$POHLAVI == 1, ])), col = "red")
points(x = ldaData_test[ldaData_test$POHLAVI == 0, ]$LDA, y = rep(2,
    nrow(ldaData_test[ldaData_test$POHLAVI == 0, ])), col = "blue")
points(x = ldaData_test[ldaData_test$POHLAVI == 1, ]$LDA, y = rep(2,
    nrow(ldaData_test[ldaData_test$POHLAVI == 1, ])), col = "red")
```