

11 Lineární diskriminační analýza (LDA)

Příklad 1. V souboru `head.txt` máme k dispozici antropometrické údaje mladých dospělých lidí (převážně studentů vysokých škol z Brna a Ostravy). Známe také pohlaví zaznamenaných jedinců (proměnná `sex`). Pomocí lineární diskriminační analýzy sestrojte funkci, která bude na základě na základě tělesné výšky (proměnná `body.H`), délky hlavy (proměnná `head.L`), šířky hlavy (proměnná `head.W`), šířky dolní čelisti (proměnná `bigo.W`) a šířky obličeje (proměnná `bizyg.W`) rozlišovat muže a ženy. Všechny rozměry byly měřeny v milimetrech.

Načteme datový soubor a zkontrolujeme, že R pracuje s proměnnou pohlaví jako s faktorem. Pokud by byla v datovém souboru kódována například pomocí 0 a 1, tak by s ní R pracovalo jako s numerickou proměnnou, nikoli kategoriální. V takovém případě bychom ji museli změnit na kategoriální pomocí funkce `factor()`.

```
head <- read.table('DATA/head.txt', header=T)
is.factor(head$sex)

## [1] TRUE
```

Zjistíme počet pozorování a odhady vektoru středních hodnot a varianční matice zvlášť pro muže a pro ženy.

```
table(head$sex)

##
##   f   m
## 100  75

colMeans(head[head$sex=='f', 2:6])

##   body.H   head.L   head.W   bigo.W   bizyg.W
## 1667.33  185.01  146.92  100.57  133.46

cov(head[head$sex=='f', 2:6])

##           body.H      head.L      head.W      bigo.W      bizyg.W
## body.H  4516.93040  121.855253  85.976162  42.789798  88.028485
## head.L   121.85525   42.838283   5.526061   6.943737   5.278182
## head.W    85.97616    5.526061  28.478384  10.056162  22.724040
## bigo.W    42.78980    6.943737  10.056162  22.085960  13.735152
## bizyg.W   88.02848    5.278182  22.724040  13.735152  37.341818

colMeans(head[head$sex=='m', 2:6])

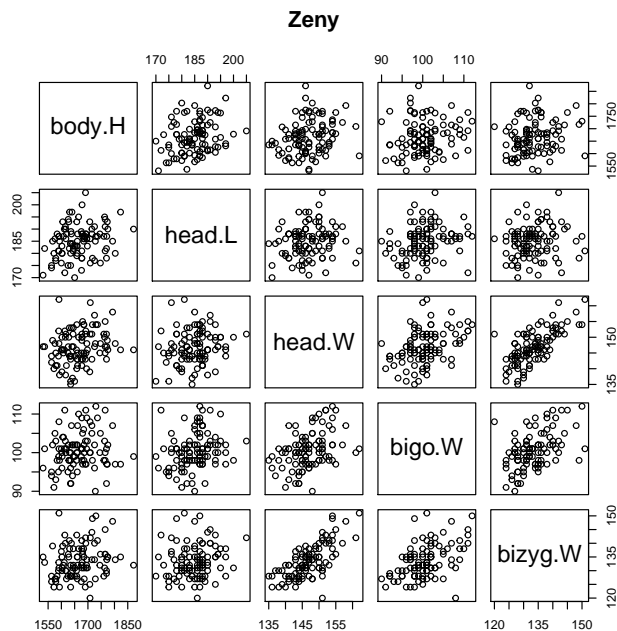
##   body.H   head.L   head.W   bigo.W   bizyg.W
## 1789.7200  195.9467  155.6533  107.8133  140.2933

cov(head[head$sex=='m', 2:6])

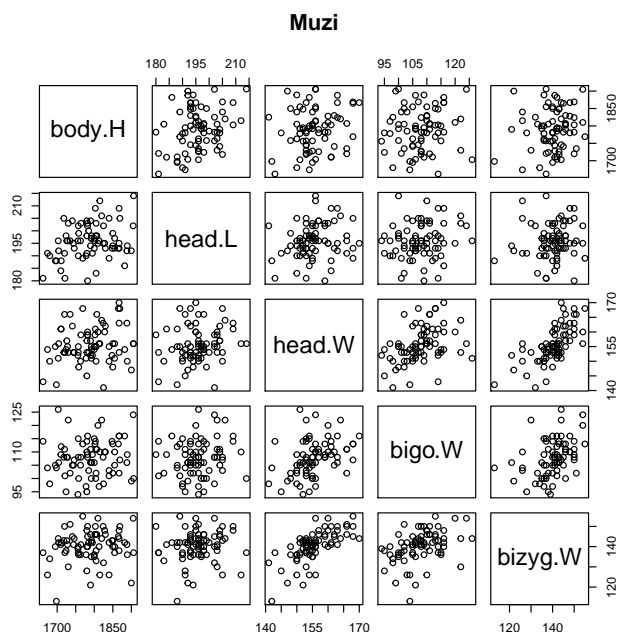
##           body.H      head.L      head.W      bigo.W      bizyg.W
## body.H  3564.85297  123.295676  64.671892  37.717297  61.826486
## head.L   123.29568   48.591712   8.008288   7.300721   5.623964
## head.W    64.67189    8.008288  36.986306  15.880360  31.832793
## bigo.W    37.71730    7.300721  15.880360  47.234955  21.285225
## bizyg.W   61.82649    5.623964  31.832793  21.285225  59.507387
```

Orientačně ověříme linearitu vztahů mezi proměnnými u obou pohlaví.

```
plot(head[head$sex=='f', 2:6], main='Zeny')
```

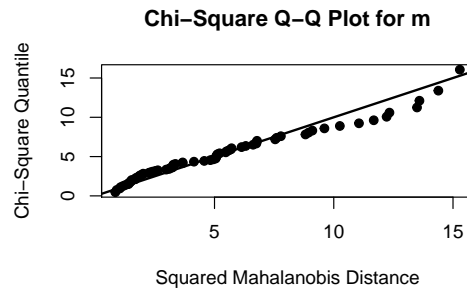
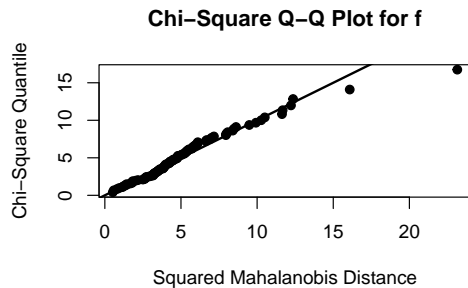


```
plot(head[head$sex=='m', 2:6], main='Muži')
```



Jedním z předpokladů lineární diskriminační analýzy je to, že pozorování ve všech skupinách pocházejí z vícerozměrného normálního rozdělení. Ověříme tedy tento předpoklad.

```
library(MVN)
par(mfrow=c(1,2))
mvn(head, subset='sex', mvnTest = 'mardia', multivariatePlot = 'qq')$multivariateNormality
```



```
## $f
##           Test      Statistic      p value Result
## 1 Mardia Skewness  50.5746767007763 0.042931208765896    NO
## 2 Mardia Kurtosis  0.862869966797547 0.38820896499104    YES
## 3           MVN           <NA>           <NA>       NO
##
## $m
##           Test      Statistic      p value Result
## 1 Mardia Skewness  54.6432331943984 0.0183339920325426    NO
## 2 Mardia Kurtosis  1.6376808454306 0.101488288386985    YES
## 3           MVN           <NA>           <NA>       NO

mvn(head, subset='sex', mvnTest = 'hz')$multivariateNormality
```

```
## $f
##           Test      HZ      p value MVN
## 1 Henze-Zirkler 0.9067157 0.2858417 YES
##
## $m
##           Test      HZ      p value MVN
## 1 Henze-Zirkler 1.081892 0.002485063 NO
```

Ženy:

Mardiův test pro šikmost:

Hodnota testovací statistiky

p-hodnota

Mardiův test pro špičatost:

Hodnota testovací statistiky

p-hodnota

Závěr

Henzeův-Zirklerův test:

Hodnota testovací statistiky

p-hodnota

Závěr

Muži:

Mardiův test pro šikmost:

Hodnota testovací statistiky
 p -hodnota
 Mardiov test pro špičatost:
 Hodnota testovací statistiky
 p -hodnota
 Závěr

Henzeův-Zirklerův test:
 Hodnota testovací statistiky
 p -hodnota
 Závěr

```
## ---
## biotools version 3.1
```

Dále je potřeba ověřit předpoklad shodných variančních matic. K tomu použijeme Boxův M test. Pozn.: Pokud někomu nešlo nainstalovat balíček biotools, lze použít funkci z balíčku heplots.

```
library('biotools')
boxM(head[,2:6], grouping=head$sex)

##
## Box's M-test for Homogeneity of Covariance Matrices
##
## data: head[, 2:6]
## Chi-Sq (approx.) = 24.749, df = 15, p-value = 0.05342
```

```
library('heplots')
boxM(head[,2:6], group=head$sex)

##
## Box's M-test for Homogeneity of Covariance Matrices
##
## data: head[, 2:6]
## Chi-Sq (approx.) = 24.749, df = 15, p-value = 0.05342
```

Hodnota testovací statistiky
 p -hodnota
 Závěr

Dále otestujeme hypotézu o shodnosti vektorů středních hodnot mezi skupinami. Pokud bychom hypotézu nezamítli, pak vybrané proměnné nezpůsobují rozdíly mezi skupinami, takže bychom nesestavili účinné pravidlo, které by nám je pomohlo třídit.

```
library('ICSNP')
HotellingsT2(head[head$sex=="f",2:6], head[head$sex=="m",2:6])

##
## Hotelling's two sample T2-test
##
## data: head[head$sex == "f", 2:6] and head[head$sex == "m", 2:6]
## T.2 = 53.421, df1 = 5, df2 = 169, p-value < 2.2e-16
## alternative hypothesis: true location difference is not equal to c(0,0,0,0,0)
```

Hodnota testovací statistiky
p-hodnota
Závěr

Na základě pozorování nyní sestavíme funkci pro rozlišení mužů a žen.

```
library('MASS')
head.lda <- lda(sex ~ body.H + head.L + head.W + bigo.W + bizyg.W, data=head)
head.lda

## Call:
## lda(sex ~ body.H + head.L + head.W + bigo.W + bizyg.W, data = head)
##
## Prior probabilities of groups:
##      f      m
## 0.5714286 0.4285714
##
## Group means:
##   body.H  head.L  head.W  bigo.W  bizyg.W
## f 1667.33 185.0100 146.9200 100.5700 133.4600
## m 1789.72 195.9467 155.6533 107.8133 140.2933
##
## Coefficients of linear discriminants:
##                LD1
## body.H    0.008650545
## head.L    0.056486003
## head.W    0.076367221
## bigo.W    0.047399638
## bizyg.W  -0.023550688
```

Ve výstupu vidíme apriorní pravděpodobnosti (tj. pravděpodobnosti odhadnuté z původních hodnot), dále vektory skupinových průměrů a koeficienty lineární diskriminační funkce pro jednotlivé proměnné.

Podívejme, jak dobře funkce zařazuje muže a ženy. K tomu použijeme funkci `predict()`.

```
fitted <- predict(head.lda)
```

Výstup (nyní uložený pod jménem `fitted`) poskytuje zařazení `class`, aposteriorní pravděpodobnosti `posterior` a hodnotu diskriminační funkce `x`. Zařazení je provedeno na základě aposteriorní pravděpodobnosti - pozorování je přiřazeno do skupiny, pro niž má vyšší aposteriorní pravděpodobnost. Podívejme se na klasifikační tabulku správně a mylně zařazených objektů.

```
(tab <- table(fitted$class, head$sex))

##
##      f  m
## f 91  9
## m  9 66
```

Vypočítáme podíl správně a mylně zařazených pozorování:

```
#spravna klasifikace
sum(diag(tab))/sum(tab)

## [1] 0.8971429

#mylna klasifikace
(tab[1,2] + tab[2,1])/sum(tab)

## [1] 0.1028571
```

Podíl správně zařazených objektů
Podíl mylně zařazených objektů

Podíl mylně zařazených objektů můžeme srovnat s náhodnou klasifikací, kdy bychom vzali v potaz pouze apriorní pravděpodobnosti, nikoli naměřené veličiny:

```
#mylna klasifikace pri nahodnem zarazovani
p <- head.lda$prior
2*p[1]*p[2]

##          f
## 0.4897959
```

Lineární diskriminační analýza snížila podíl špatného zařazení z na

Na základě sestavené funkce se pokusíme zařadit dva nové případy, první má naměřené hodnoty 1820, 190, 165, 110, 152 a druhý 1700, 185, 154, 99, 130 (hodnoty jsou v pořadí výška, délka hlavy, šířka hlavy, šířka dolní čelisti, šířka obličeje).

```
predict(head.lda, newdata=list(body.H=c(1820, 1700), head.L=c(190, 185), head.W=c(165, 154),
                               bigo.W=c(110, 99), bizyg.W=c(152, 130)))

## $class
## [1] m f
## Levels: f m
##
## $posterior
##          f          m
## 1 0.01656476 0.9834352
## 2 0.79926415 0.2007358
##
## $x
##          LD1
## 1  1.9111046
## 2 -0.2527112
```

První případ má aposteriorní pravděpodobnosti a, proto byl zařazen do skupiny Druhý případ má aposteriorní pravděpodobnosti a, proto byl zařazen do skupiny

Pro výběr proměnných můžeme použít dopřednou krokovou metodu. K tomu slouží funkce `greedy.wilks()` z balíčku `klaR`. Funkce postupně vybírá proměnné, které vedou ke snížení hodnoty Wilskova Λ . Pokud přidání žádné další proměnné nevede ke snížení, algoritmus končí.

```

library('klaR')
greedy.wilks(sex ~ body.H + head.L + head.W + bigo.W + bizyg.W, data=head)

## Formula containing included variables:
##
## sex ~ body.H + head.W + head.L + bigo.W
## <environment: 0x00000000205befc0>
##
##
## Values calculated in each step of the selection procedure:
##
##      vars Wilks.lambda F.statistics.overall p.value.overall
## 1 body.H      0.5255017           156.20924    5.903695e-26
## 2 head.W      0.4467734           106.49132    8.081779e-31
## 3 head.L      0.4025040            84.61351    1.312298e-33
## 4 bigo.W      0.3905933            66.30883    1.045292e-33
##      F.statistics.diff p.value.diff
## 1           156.209239 5.903695e-26
## 2           30.309019 1.311737e-07
## 3           18.807447 2.457533e-05
## 4            5.183937 2.403590e-02

```

V tomto případě algoritmus vybral čtyři proměnné z původních pěti. Šířka obličeje tedy dále nepřispívá k lepší klasifikaci. Sestavme tedy funkci bez ní a podívejme se na podíl správně zařazených objektů.

```

head.lda2 <- lda(sex ~ body.H + head.L + head.W + bigo.W, data=head)
head.lda2

## Call:
## lda(sex ~ body.H + head.L + head.W + bigo.W, data = head)
##
## Prior probabilities of groups:
##      f      m
## 0.5714286 0.4285714
##
## Group means:
##      body.H  head.L  head.W  bigo.W
## f 1667.33 185.0100 146.9200 100.5700
## m 1789.72 195.9467 155.6533 107.8133
##
## Coefficients of linear discriminants:
##              LD1
## body.H 0.008620688
## head.L 0.057662299
## head.W 0.059469511
## bigo.W 0.042039058

fitted2 <- predict(head.lda2)

(tab2 <- table(fitted2$class, head$sex))

##
##      f  m
## f 91  9
## m  9 66

```

```
sum(diag(tab2))/sum(tab2)
## [1] 0.8971429
```

Vidíme, že vynecháním proměnné šířka obličej se podíl správně zařazených objektů nezhoršil.

Příklad 2. V souboru `Howell.csv` máme k dispozici kranio-metrické údaje z různých populací. Nás zajímají muži (kategorie `M` proměnné `Sex`) ze 3 populací (proměnná `Population`) - ZULU, BUSHMAN a AUSTRALI. Konkrétně máme tyto kranio-metrické rozměry (vše v milimetrech):

- ZYB - bizygomatická šířka,
- ZMB - zygomaticomaxilární šířka,
- BPL - délka obličejové části lebky,
- NPH - výška horní části obličejového skeletu,
- NLH - výška nosu,
- OBH - výška očníce levé strany,
- WCB - minimální šířka lebky.

Načteme datový soubor. Protože v databázi jsou chybějící pozorování kódovány jako 0, je potřeba při načítání zadat, aby se 0 braly jako `NA`. Vybereme pozorování a proměnné, které nás zajímají, a zbavíme se nyní prázdných kategorií proměnné `Population`.

```
cranio <- read.csv('DATA/Howell.csv',header=T, na.strings='0')
howells.data <- cranio[cranio$Sex == 'M' & cranio$Population %in% c('ZULU', 'BUSHMAN', 'AUSTRALI'),
  c('Population', 'ZYB', 'ZMB', 'BPL', 'NPH', 'NLH', 'OBH', 'WCB')]
howells.data$Population <- factor(howells.data$Population)
```

Zjistíme počet pozorování a odhady vektoru středních hodnot a varianční matice zvlášť pro každou populaci.

```
table(howells.data$Population)
##
## AUSTRALI BUSHMAN ZULU
##      52      41      55

colMeans(howells.data[howells.data$Population=='AUSTRALI',-1])
##      ZYB      ZMB      BPL      NPH      NLH      OBH      WCB
## 136.76923  98.34615 105.50000  64.76923  49.69231  33.46154  71.30769

cov(howells.data[howells.data$Population=='AUSTRALI',-1])
##      ZYB      ZMB      BPL      NPH      NLH      OBH      WCB
## ZYB 17.396682  7.297134  5.2156863  2.926094  3.045249  1.108597  6.0723982
## ZMB  7.297134 16.505279  5.1764706  3.238311  3.343891  1.562594  3.6757164
## BPL  5.215686  5.176471 19.9803922  8.215686  3.333333  1.843137  0.6666667
## NPH  2.926094  3.238311  8.2156863 17.318250  8.162896  3.147813  3.5037707
## NLH  3.045249  3.343891  3.3333333  8.162896  7.197587  1.791855  3.0573152
## OBH  1.108597  1.562594  1.8431373  3.147813  1.791855  3.665158  0.9336350
## WCB  6.072398  3.675716  0.6666667  3.503771  3.057315  0.933635  9.2368024
```



```
colMeans(howells.data[howells.data$Population=='BUSHMAN',-1])

##          ZYB          ZMB          BPL          NPH          NLH          OBH          WCB
## 123.56098  92.19512  93.65854  57.51220  43.75610  30.82927  70.00000

cov(howells.data[howells.data$Population=='BUSHMAN',-1])

##          ZYB          ZMB          BPL          NPH          NLH          OBH          WCB
## ZYB 22.002439 10.787805 15.421341 15.055488  6.815244  5.523171  9.550
## ZMB 10.787805 23.460976 10.943293  9.822561  5.423780  2.959146  5.600
## BPL 15.421341 10.943293 27.930488 15.954268  7.039634  2.090244  4.275
## NPH 15.055488  9.822561 15.954268 28.256098 12.378049  5.264634  3.825
## NLH  6.815244  5.423780  7.039634 12.378049  8.639024  3.857317  1.525
## OBH  5.523171  2.959146  2.090244  5.264634  3.857317  5.795122  2.000
## WCB  9.550000  5.600000  4.275000  3.825000  1.525000  2.000000  9.400

colMeans(howells.data[howells.data$Population=='ZULU',-1])

##          ZYB          ZMB          BPL          NPH          NLH          OBH          WCB
## 129.94545  95.87273 102.38182  67.32727  50.00000  33.76364  71.98182

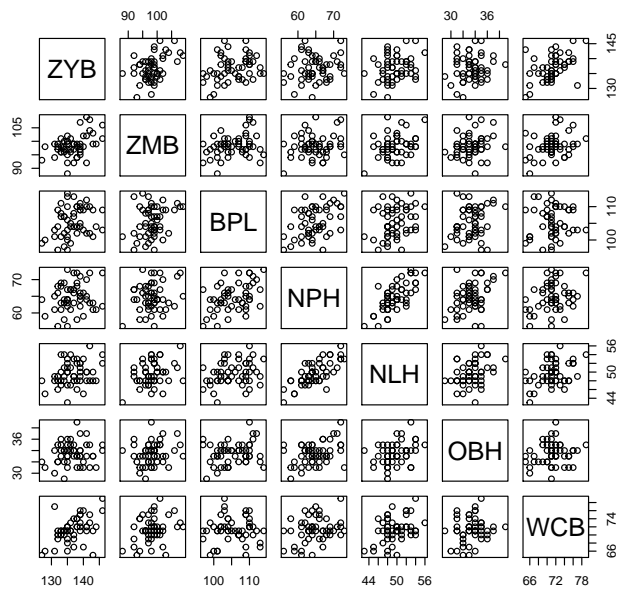
cov(howells.data[howells.data$Population=='ZULU',-1])

##          ZYB          ZMB          BPL          NPH          NLH          OBH
## ZYB 16.645118 11.400337  3.8175084  4.592256  5.0555556  1.8942761
## ZMB 11.400337 23.483502 10.4754209  7.672054  4.6296296  1.6730640
## BPL  3.817508 10.475421 37.2404040  3.780135 -0.1481481 -0.5191919
## NPH  4.592256  7.672054  3.7801347 16.557576  7.5000000  3.6343434
## NLH  5.055556  4.629630 -0.1481481  7.500000  6.5555556  2.0000000
## OBH  1.894276  1.673064 -0.5191919  3.634343  2.0000000  3.1097643
## WCB  8.554545  7.997643 -5.4373737  2.969024  3.4259259  1.0882155
##          WCB
## ZYB  8.554545
## ZMB  7.997643
## BPL -5.437374
## NPH  2.969024
## NLH  3.425926
## OBH  1.088215
## WCB 18.129293
```

Orientačně ověříme linearitu vztahů mezi proměnnými u všech populací.

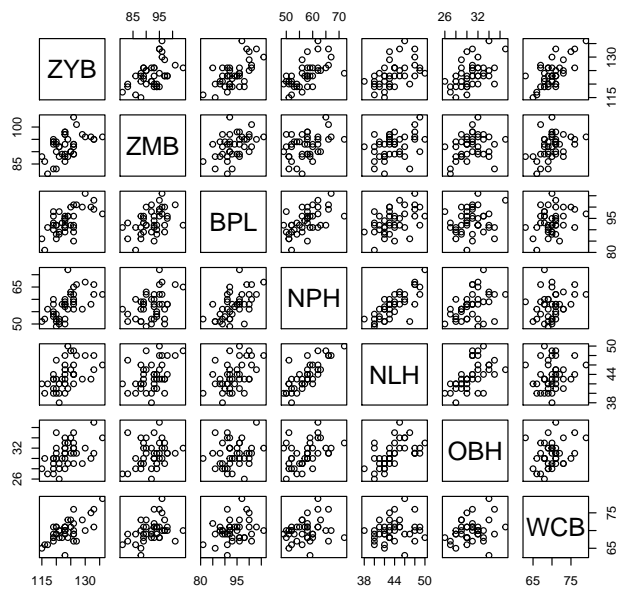
```
plot(howells.data[howells.data$Population=='AUSTRALI',-1], main='AUSTRALI')
```

AUSTRALI

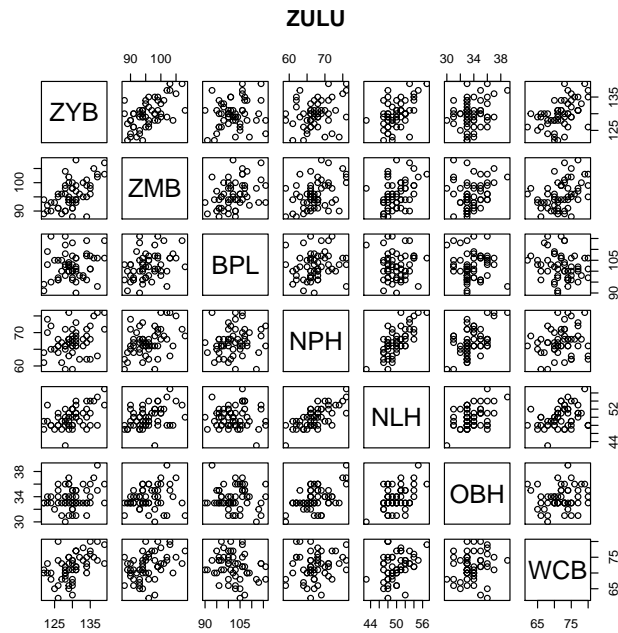


```
plot(howells.data[howells.data$Population=='BUSHMAN',-1], main='BUSHMAN')
```

BUSHMAN

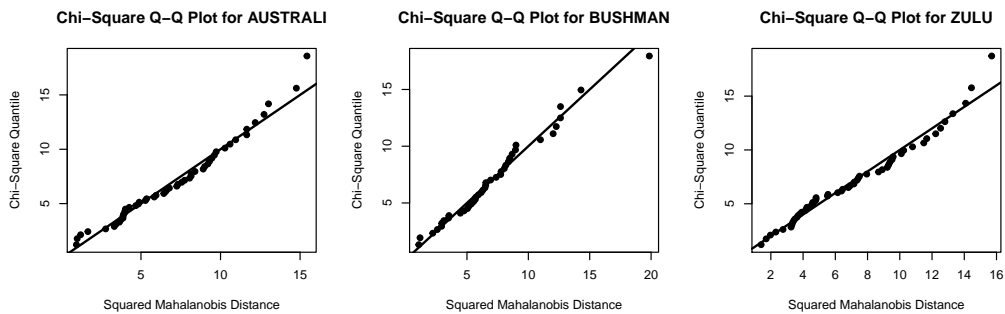


```
plot(howells.data[howells.data$Population=='ZULU',-1], main='ZULU')
```



Ověříme předpoklady lineární diskriminační analýzy.

```
library(MVN)
par(mfrow=c(1,3))
mvn(howells.data, subset='Population', mvnTest = 'mardia',
    multivariatePlot = 'qq')$multivariateNormality
```



```
## $AUSTRALI
##           Test           Statistic           p value Result
## 1 Mardia Skewness  61.7824580535975 0.967207294007091   YES
## 2 Mardia Kurtosis -0.527521156668937 0.597831728443825   YES
## 3           MVN                <NA>                <NA>   YES
##
## $BUSHMAN
##           Test           Statistic           p value Result
## 1 Mardia Skewness  98.3414107802788 0.135693050518559   YES
## 2 Mardia Kurtosis  0.0599958258921211 0.952158959159005   YES
## 3           MVN                <NA>                <NA>   YES
##
## $ZULU
```

```

##          Test          Statistic          p value Result
## 1 Mardia Skewness    93.9448325630975 0.214796045720732    YES
## 2 Mardia Kurtosis  -0.295992327049272 0.767235941123309    YES
## 3          MVN              <NA>              <NA>        YES

mvn(howells.data, subset='Population', mvnTest = 'hz')$multivariateNormality

## $AUSTRALI
##          Test          HZ          p value MVN
## 1 Henze-Zirkler 0.9066086 0.5696042 YES
##
## $BUSHMAN
##          Test          HZ          p value MVN
## 1 Henze-Zirkler 0.967527 0.06989468 YES
##
## $ZULU
##          Test          HZ          p value MVN
## 1 Henze-Zirkler 0.9760937 0.06599569 YES

```

Populace australských domorodců:

Mardiův test pro šikmost:

Hodnota testovací statistiky

p-hodnota

Mardiův test pro špičatost:

Hodnota testovací statistiky

p-hodnota

Závěr

Henzeův-Zirklerův test:

Hodnota testovací statistiky

p-hodnota

Závěr

Populace Křováků:

Mardiův test pro šikmost:

Hodnota testovací statistiky

p-hodnota

Mardiův test pro špičatost:

Hodnota testovací statistiky

p-hodnota

Závěr

Henzeův-Zirklerův test:

Hodnota testovací statistiky

p-hodnota

Závěr

Populace Zulu:

Mardiův test pro šikmost:

Hodnota testovací statistiky

p-hodnota

Mardiův test pro špičatost:

Hodnota testovací statistiky

p-hodnota

Závěr

Henzeův-Zirklerův test:

Hodnota testovací statistiky

p-hodnota

Závěr

```
library('biotools')
boxM(howells.data[,-1], grouping=howells.data$Population)

##
## Box's M-test for Homogeneity of Covariance Matrices
##
## data:  howells.data[, -1]
## Chi-Sq (approx.) = 70.37, df = 56, p-value = 0.09371
```

```
library('heplots')
boxM(howells.data[,-1], group=howells.data$Population)

##
## Box's M-test for Homogeneity of Covariance Matrices
##
## data:  howells.data[, -1]
## Chi-Sq (approx.) = 70.37, df = 56, p-value = 0.09371
```

Hodnota testovací statistiky

p-hodnota

Závěr

Dále otestujeme hypotézu o shodnosti vektorů středních hodnot mezi skupinami.

```
model <- manova(as.matrix(howells.data[,-1]) ~ howells.data$Population)
summary(model, test='Wilks')

##
##              Df  Wilks approx F num Df den Df    Pr(>F)
## howells.data$Population  2 0.18552  26.245    14   278 < 2.2e-16 ***
## Residuals              145
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hodnota Wilksovy testovací statistiky

p-hodnota

Závěr

Na základě pozorování nyní sestavíme funkci pro rozlišení populací.

```
library('MASS')
how.lda <- lda(Population ~ ZYB + ZMB + BPL + NPH + NLH + OBH + WCB, data=howells.data)
how.lda

## Call:
## lda(Population ~ ZYB + ZMB + BPL + NPH + NLH + OBH + WCB, data = howells.data)
```

```
##
## Prior probabilities of groups:
## AUSTRALI BUSHMAN ZULU
## 0.3513514 0.2770270 0.3716216
##
## Group means:
##          ZYB      ZMB      BPL      NPH      NLH      OBH      WCB
## AUSTRALI 136.7692 98.34615 105.50000 64.76923 49.69231 33.46154 71.30769
## BUSHMAN  123.5610 92.19512  93.65854 57.51220 43.75610 30.82927 70.00000
## ZULU     129.9455 95.87273 102.38182 67.32727 50.00000 33.76364 71.98182
##
## Coefficients of linear discriminants:
##          LD1      LD2
## ZYB -0.19925338 -0.18258632
## ZMB  0.03386777 -0.03969644
## BPL -0.06286477  0.04818895
## NPH  0.05502890  0.13314518
## NLH -0.17897848  0.11188320
## OBH -0.02518774  0.10227486
## WCB  0.10632779  0.12385882
##
## Proportion of trace:
## LD1 LD2
## 0.77 0.23
```

Ve výstupu vidíme apriorní pravděpodobnosti (tj. pravděpodobnosti odhadnuté z původních hodnot), dále vektory skupinových průměrů a koeficienty obou lineárních diskriminačních funkcí pro jednotlivé proměnné.

Přiřazení do skupin je opět na základě nejvyšší hodnoty aposteriorní pravděpodobnosti.

```
fit <- predict(how.lda)

(tab.h <- table(fit$class, howells.data$Population))

##
##          AUSTRALI BUSHMAN ZULU
## AUSTRALI      45      1      6
## BUSHMAN       1     33      1
## ZULU          6      7     48

sum(diag(tab.h)) / sum(tab.h)

## [1] 0.8513514
```

Podíl správně zařazených objektů

Zkusme zařadit neznámé pozorování s hodnotami ZYB 130, ZMB 98, BPL 100, NPH 68, NLH 51, OBH 34 a WCB 70.

```
predict(how.lda, newdata=list(ZYB=130, ZMB=98, BPL=100, NPH=68, NLH=51, OBH=34, WCB=70))

## $class
## [1] ZULU
```

```
## Levels: AUSTRALI BUSHMAN ZULU
##
## $posterior
##   AUSTRALI   BUSHMAN   ZULU
## 1 0.101297 0.01802436 0.8806786
##
## $x
##      LD1      LD2
## 1 -0.169831 0.8221662
```

Pozorování má aposteriorní pravděpodobnosti (pro AUSTRALI), (pro BUSHMAN) a (pro ZULU), proto bylo zařazeno k populaci

Můžeme zkusit vybrat proměnné pomocí dopředné krokové metody.

```
library('klaR')
greedy.wilks(Population ~ ZYB + ZMB + BPL + NPH + NLH + OBH + WCB, data=howells.data)

## Formula containing included variables:
##
## Population ~ ZYB + NPH + WCB + NLH + BPL
## <environment: 0x00000000226a6148>
##
##
## Values calculated in each step of the selection procedure:
##
##   vars Wilks.lambda F.statistics.overall p.value.overall F.statistics.diff
## 1  ZYB   0.3979242          109.69551    9.672195e-30         109.695506
## 2  NPH   0.2571663           69.97943    2.469448e-41          39.408603
## 3  WCB   0.2185048           54.30612    1.770872e-44          12.650969
## 4  NLH   0.2034279           43.20876    6.610294e-45           5.262102
## 5  BPL   0.1919026           36.17372    5.160336e-45           4.234092
##   p.value.diff
## 1 9.672195e-30
## 2 2.153833e-14
## 3 8.678415e-06
## 4 6.235646e-03
## 5 1.636441e-02
```

Funkce vybrala 5 z našich původních 7 proměnných. Sestavíme diskriminační funkce pouze z vybraných proměnných a podíváme se na podíl správně zařazených objektů.

```
how.lda2 <- lda(Population ~ ZYB + NPH + WCB + NLH + BPL, data=howells.data)
fit2 <- predict(how.lda2)
(tab.h2 <- table(fit2$class, howells.data$Population))

##
##           AUSTRALI BUSHMAN ZULU
## AUSTRALI         45         1     5
## BUSHMAN           0         32     1
## ZULU              7         8    49

sum(diag(tab.h2)) / sum(tab.h2)

## [1] 0.8513514
```