```
# Classification and Discriminant Analysis
# Exercise 4

library(ElemStatLearn)
data(SAheart)
# str(SAheart)
heart <- SAheart[,-5] # exclude the variable famhist
x.heart = heart[,1:8] # everything except chd
y.heart = heart[,9] # chd


################################################################################


# 1. Plot the (scaled) data into the space of principal components
# (function princomp) to distinguish the groups given by the variable chd.

library(ElemStatLearn)
data(SAheart)
# str(SAheart)
heart <- SAheart[,-5]
x.heart = heart[,1:8]
y.heart = heart[,9]

x.heart.scaled <- scale(x.heart, center=TRUE, scale=TRUE)
heart.pc <- princomp(x.heart.scaled)
# heart.pc <- princomp(x.heart, cor=TRUE)
# heart.pc <- princomp(~., data=x.heart, cor=TRUE)
# heart.pc <- prcomp(x.heart, scale=TRUE)
plot(heart.pc$scores[,1],heart.pc$scores[,2],pch=y.heart, col=y.heart+1)
title(main='South African Hearth Disease Data\nPrincipal Components')
legend('topright',pch=c(0,1), col=c(1,2),
legend=c('coronary heart disease = 0', 'coronary heart disease = 1'), cex=0.8)


################################################################################

# 2. Linear regression with the indicator matrix (LS):

library(ElemStatLearn)
data(SAheart)
# str(SAheart)
heart <- SAheart[,-5]
heart.ind <- heart

# (a) Construct the indicator matrix consisting of two columns for the variable
# chd.

heart.indy = array(data=0, dim = c(nrow(heart), 2))

for (i in 1:nrow(heart)) {
        if (heart[i, 9] == 0) {
                        heart.indy[i,1] = 1
                } else {
                        heart.indy[i,2] = 1
                }
}

heart.ind[,9] <- heart.indy
```

```
# Select the training set of 300 observations (set the random seed!) and
# apply the LS-regression with the indicator matrix.

set.seed(123)
obs <- 300
train <- sample(1:nrow(heart),size=obs, replace=FALSE)

mod.ind <- lm(chd ~ ., data=heart.ind[train, ])
mod.ind

        # Call:
        # lm(formula = chd ~ ., data = heart.ind[train, ])

        # Coefficients:
        #              [,1]        [,2]
        # (Intercept)   1.7798606  -0.7798606
        # sbp          -0.0022913   0.0022913
        # tobacco      -0.0082423   0.0082423
        # ldl          -0.0319983   0.0319983
        # adiposity     0.0028711  -0.0028711
        # typea        -0.0089255   0.0089255
        # obesity       0.0076676  -0.0076676
        # alcohol       0.0000511  -0.0000511
        # age          -0.0100738   0.0100738

# Predict the group membership for the test data and compute the rate of
# misclassification.

predict.ind <- as.numeric(predict(mod.ind, newdata = heart.ind[-train, ])[, 1] <
predict(mod.ind, newdata = heart.ind[-train, ])[, 2])
TAB.ind <- table(heart$chd[-train], predict.ind)
mklrate.ind <- 1 - sum(diag(TAB.ind))/sum(TAB.ind)
mklrate.ind
        # [1] 0.2901235

# (b) Repeat the procedure 100 times (without seed) and plot the rates of
# misclassication using a boxplot.

set.seed(seed=NULL)
obs <- 300
proc <- 100
mklrate.ind.proc <- rep(0,proc)

for (i in 1:proc){
        train <- sample(1:nrow(heart),size=obs, replace=FALSE)
        mod.ind <- lm(chd ~ ., data=heart.ind[train, ])
        predictind <- as.numeric(predict(mod.ind, newdata = heart.ind[-train, ])[, 1]
        < predict(mod.ind, newdata = heart.ind[-train, ])[, 2])
        TAB <- table(heart$chd[-train], predictind)
        mklrate.ind.proc[i] <- 1 - sum(diag(TAB))/sum(TAB)
}

boxplot(mklrate.ind.proc, notch=TRUE, ylab='Rate of misclassification')
title(main='South African Hearth Disease Data
Linear regression with indicator matrix (LS)')


##############################################################################
```

```
# 3. Linear Discrimant Analysis (LDA): function lda from library(MASS)

library(ElemStatLearn)
data(SAheart)
# str(SAheart)
heart <- SAheart[,-5]

# (a) Select the training set of 300 observations (set the random seed!) and
# apply the LDA.

set.seed(123)
obs <- 300
train <- sample(1:nrow(heart),size=obs, replace=FALSE)

library(MASS)
mod.lda <- lda(chd ~ ., data=heart[train, ])
mod.lda

        # Call:
        # lda(chd ~ ., data = heart[train, ])

        # Prior probabilities of groups:
        #         0         1
        # 0.6533333 0.3466667

        # Group means:
        #        sbp  tobacco      ldl adiposity    typea  obesity  alcohol      age
        # 0 135.2959 2.714541 4.341531  23.79709 51.92347 25.67709 16.20526 38.49490
        # 1 145.0865 4.925192 5.272308  27.54202 55.23077 26.30298 20.51087 50.19231

        # Coefficients of linear discriminants:
        #                       LD1
        # sbp        0.0119510752
        # tobacco    0.0429908792
        # ldl        0.1668999888
        # adiposity -0.0149752631
        # typea      0.0465544631
        # obesity   -0.0399936205
        # alcohol   -0.0002665466
        # age        0.0525436865

# Predict the group membership for the test data and compute the rate of
# misclassification.

predict.lda <- predict(mod.lda,newdata=heart[-train,])$class
TAB.lda <- table(heart$chd[-train], predict.lda)
mklrate.lda <- 1 - sum(diag(TAB.lda))/sum(TAB.lda)
mklrate.lda
        # [1] 0.2962963


# (b) Repeat the procedure 100 times (without seed) and plot the rates of
# misclassification using a boxplot.

set.seed(seed=NULL)
obs <- 300
proc <- 100
mklrate.lda.proc <- rep(0,proc)
```

```
for (i in 1:proc){
        train <- sample(1:nrow(heart),size=obs, replace=FALSE)
        mod.lda <- lda(chd ~ ., data=heart[train, ])
        predict.lda <- predict(mod.lda,newdata=heart[-train,])$class
        TAB.lda <- table(heart$chd[-train], predict.lda)
        mklrate.lda.proc[i] <- 1 - sum(diag(TAB.lda))/sum(TAB.lda)
}

boxplot(mklrate.lda.proc, notch=TRUE, ylab='Rate of misclassification')
title(main='South African Hearth Disease Data
Linear Discriminant Analysis (LDA)')


# Compare the boxplot with the previous one.
# Which method works better in this case?

boxplot(mklrate.ind.proc, mklrate.lda.proc, notch=TRUE, names=c('LS','LDA'),
ylab='Rate of misclassification')
title(main='South African Hearth Disease Data
Linear regression with indicator matrix (LS)
Linear Discriminant Analysis (LDA)')

summary(mklrate.ind.proc)
        #    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
        #  0.2160  0.2778  0.3025  0.2996  0.3164  0.3827
summary(mklrate.lda.proc)
        #    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
        #  0.1975  0.2778  0.3025  0.3009  0.3210  0.3704

sd(mklrate.ind.proc)
        # [1] 0.03448022
sd(mklrate.lda.proc)
        #       [1] 0.03309229

IQR(mklrate.ind.proc)
        # [1] 0.03858025
IQR(mklrate.lda.proc)
        # [1] 0.04320988

# In this case, probably IND/LS.
# IND/LS's and LDA's median are approximately the same.
# IND/LS's whiskers are closer together.
# IND/LS has smaller IQR, but slightly larger sd.
# However, IND/LS shows the highest mklrate of both (max.).
```