Homework #2

Dustin Leatherman

April 13, 2019

```
knitr::opts_chunk$set(echo = TRUE)
library(kableExtra)
library(knitr)
library(car)
```

```
## Loading required package: carData
```

```
library(tidyverse)
```

```
## -- Attaching packages ----------------------------------------------------------------------------
```

```
## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------------------------------- tidy
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x dplyr::recode() masks car::recode()
## x purrr::some()   masks car::some()
```

```
library(Hotelling)
```

```
## Loading required package: corpcor
```

```
library(ggplot2)
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
# executes a wilks lambda test to determine the significance of an linear discriminant function
lda.sig <- function(model.lda, model.lm) {
  require(MASS)
  require(car)
  require(dplyr)

  m <- model.lda
```

```
  m.lm <- model.lm
  m.man <- Manova(m.lm)
  groupName <- m$terms[[2]]
  E <- m.man$SSPE
  H <- m.man$SSP[[groupName]]

  n <- m$N
  p <- m.lm$coefficients[1,] %>% length
  k <- m$lev %>% length
  s <- min(k - 1, p)

  # calculate eigenvalues of E^-1H
  lambda <-
    solve(E)%*%H %>%
    eigen %>%
    .$values %>%
    Re

  # subset based on number of possible linear discrim functions
  lambda <- lambda[1:s]

  # inline function to calculate Wilk's Lambda for testing signifance of discriminant functions
  V <- function(m,l) (n - 1 - 0.5 * (p + k)) * sum(log(1 + lambda[m:l]))

  for(i in 1:(k - 1)) {
    V.i <- V(i, k - 1)
    V.p <- 1 - pchisq(V.i, (p - i + 1) * (k - i))
    print(paste("EigenValues: ", lambda))
    print(paste("Test Statistic for Eigenvalue ", i, ": ", V.i))
    print(paste("P-value for Eigenvalue ", i, ": ", V.p))
  }
}
```

1

Four measurements were made on two species of flea beetles. The variables were:

y1 = distance of transverse groove from posterior border of prothorax (µm) y2 = length of elytra (in 0.01 mm) y3 = length of second antennal joint (in µm) y4 = length of third antennal joint (in µm)

The data can be found in the data set beetles.csv in the Week One course content on D2L.

(a) Find the discriminant function coefficient vector.

```
beetles <- read.csv("~/Downloads/beetles.csv")
beetles.lda <- lda(Species ~ y1 + y2 + y3 + y4, data = beetles)
beetles.lda$scaling %>% kable
```

Table 1: Standardized Coefficients

	x
y1	1.1158383
y2	-0.6728884
y3	-0.3159640
y4	-0.5562256

	LD1
y1	0.0931292
y2	-0.0350869
y3	-0.0290482
y4	-0.0383234

(b) Find the standardized coefficients.

$z_{ij} = a_1^* \frac{y_{ij1} - \bar{y}}{s_1} + a_2^* \frac{y_{ij2} - \bar{y}}{s_2} + \ldots + a_p^* \frac{y_{ijp} - \bar{y}}{s_p}$

$a^* = (\sqrt{diag S_p l})a$

```
beetle.lm <- lm(cbind(y1, y2, y3, y4) ~ Species, data = beetles)

# run manova to retrieve the Error Matrix (E)
beetle.manova <- Manova(beetle.lm)
E <- beetle.manova$SSPE

# calculate pooled Standard variance. This formula works for J groups
n <- beetles %>% count %>% pull
j <- beetles %>% dplyr::select(Species) %>%  unique %>% count %>% pull
S.pl <- E / (n - j)

# calculate and display standardized coefficients
a.star <- sqrt(diag(S.pl)) * beetles.lda$scaling[,1]
a.star %>% kable(
  caption = "Standardized Coefficients"
)
```

(c) Calculate t-tests for individual variables.

```
beetle.hot <- hotelling.test(cbind(y1,y2,y3,y4) ~ Species, data = beetles)
beetle.hot
```

```
## Test stat:  30.581
## Numerator df:  4
## Denominator df:  34
## P-value:  7.802e-11
```

There is convincing evidence that at least one of the coefficents of the linear discriminant function is non-zero (Hotelling's T^2. p-value = 7.802e-11).

Now let's test each variable using two-sample t-tests

```r
# test for equal variance assumption
bartlett.test(y1 ~ Species, data = beetles)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  y1 by Species
## Bartlett's K-squared = 1.6628, df = 1, p-value = 0.1972
```

```r
bartlett.test(y2 ~ Species, data = beetles)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  y2 by Species
## Bartlett's K-squared = 0.063557, df = 1, p-value = 0.801
```

```r
bartlett.test(y3 ~ Species, data = beetles)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  y3 by Species
## Bartlett's K-squared = 3.6999, df = 1, p-value = 0.05442
```

```r
bartlett.test(y4 ~ Species, data = beetles)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  y4 by Species
## Bartlett's K-squared = 0.4635, df = 1, p-value = 0.496
```

```r
# Test for normality assumptions
shapiro.test(beetles$y1)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  beetles$y1
## W = 0.97097, p-value = 0.4021
```

```r
shapiro.test(beetles$y2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  beetles$y2
## W = 0.95948, p-value = 0.1714
```

```r
shapiro.test(beetles$y3)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  beetles$y3
## W = 0.96192, p-value = 0.2066
```

```r
shapiro.test(beetles$y4)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  beetles$y4
## W = 0.98331, p-value = 0.8199
```

```r
# run two sample t-tests
y1.res <- t.test(y1 ~ Species, data = beetles, var.equal = TRUE)
y1.res
```

```
##
##  Two Sample t-test
##
## data:  y1 by Species
## t = -3.8879, df = 37, p-value = 0.0004049
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -22.701122  -7.146246
## sample estimates:
## mean in group Carduorum  mean in group Oleracea
##                179.5500                194.4737
```

```r
y1.res.est <- unname(y1.res$estimate)

# Estimated average difference between Oleracea and Carduorum
y1.res.est[2] - y1.res.est[1]
```

```
## [1] 14.92368
```

```r
y2.res <- t.test(y2 ~ Species, data = beetles, var.equal = TRUE)
y2.res
```

```
##
##  Two Sample t-test
##
## data:  y2 by Species
## t = 3.8652, df = 37, p-value = 0.0004326
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  11.29879 36.19595
```

```
## sample estimates:
## mean in group Carduorum  mean in group Oleracea
##                290.8000                267.0526
```

```
y2.res.est <- unname(y2.res$estimate)

# Estimated average difference between Oleracea and Carduorum
y2.res.est[2] - y2.res.est[1]
```

```
## [1] -23.74737
```

```
y3.res <- t.test(y3 ~ Species, data = beetles, var.equal = TRUE)
y3.res
```

```
##
##  Two Sample t-test
##
## data:  y3 by Species
## t = 5.6911, df = 37, p-value = 1.645e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  12.77101 26.89214
## sample estimates:
## mean in group Carduorum  mean in group Oleracea
##                157.2000                137.3684
```

```
y3.res.est <- unname(y3.res$estimate)

# Estimated average difference between Oleracea and Carduorum
y3.res.est[2] - y3.res.est[1]
```

```
## [1] -19.83158
```

```
y4.res <- t.test(y4 ~ Species, data = beetles, var.equal = TRUE)
y4.res
```

```
##
##  Two Sample t-test
##
## data:  y4 by Species
## t = 5.0343, df = 37, p-value = 1.269e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  13.98665 32.82914
## sample estimates:
## mean in group Carduorum  mean in group Oleracea
##                209.2500                185.8421
```

```
y4.res.est <- unname(y4.res$estimate)

# Estimated average difference between Oleracea and Carduorum
y4.res.est[2] - y4.res.est[1]
```

```
## [1] -23.40789
```

There is no evidence that y1, y2, or y4 have differing variances between the two species of beetles. There is suggestive but inconclusive evidence that y3 has differing variances between the two species of beetles (Bartlett Test. p-value=0.05442).

There is no evidence that y1, y2, y3, or y4 are not normal via the Shapiro Wilk Test.

There is convincing evidence that the average value of y1 is non-zero between the two species of beetles (two-sample T-test. p-value=0.0004049). It is estimated that y1 for Oleracea has an average value of 14.92368 more than Carduorum. With 95% confidence, the average value of y1 for Oleracea is between 7.146246 and 22.701122 more than Carduorum.

There is convincing evidence that the average value of y2 is non-zero between the two species of beetles (two-sample T-test. p-value=0.0004326). It is estimated that y2 for Oleracea has an average value of 23.74737 less than Carduorum. With 95% confidence, the average value of y2 for Oleracea is between 11.29879 and 36.19595 less than Carduorum.

There is convincing evidence that the average value of y3 is non-zero between the two species of beetles (two-sample T-test. p-value=1.645e-06). It is estimated that y3 for Oleracea has an average value of 19.83158 less than Carduorum. With 95% confidence, the average value of y3 for Oleracea is between 12.77101 and 26.89214 less than Carduorum.

There is convincing evidence that the average value of y4 is non-zero between the two species of beetles (two-sample T-test. p-value=1.269e-05). It is estimated that y4 for Oleracea has an average value of 23.40789 less than Carduorum. With 95% confidence, the average value of y4 for Oleracea is between 13.98665 and 32.82914 less than Carduorum.

(d) Compare the results of (b) and (c) as to the contribution of each variable to separation of the groups

The direction of the comparison is the same for both the standardized coefficients and the estimated value for the univariate t-test. For example, the standardized coefficient for y1 is positive and the estimated value from the t-test is positive, while the other 3 variables are negative for both the standardized coefficient and the estimated value. The standardized coefficient for y1 has largest magnitude but has the smallest average difference between the two species of beetle. The Std Coefficient for y2 and y4 are similar in magnitude as well as estimated value from the t-test. The coefficient for y3 is the smallest in magnitude but has an estimated value that falls between the estimated values y2,y3 and y1 indicating that there may be a non-linear relationship between the estimates from a t-test and the standardized coefficient.

2

Baten, Tack, and Baeder (1958) compared judges' scores on fish prepared by three methods. Twelve fish were cooked by each method, and several judges tasted fish samples and rated each on four variables: aroma, flavor, texture, and moisture. The data can be found in the data set fish.csv in the Week Two course content on D2L.

(a)

Carry out tests of significance for the discriminant functions and find the relative importance of each as in $\frac{\lambda_i}{\sum \lambda_i}$

where λ_i are the eigenvalues. Do these two procedures agree as to the number of important discriminant functions?

Since there is more than two groups, Wilk's Lambda will be used to test for significance.

```
fish <- read.csv("~/Downloads/fish.csv")
fish$MethodFactor <- factor(fish$Method)

fish.lda <- lda(MethodFactor ~ Aroma + Flavor + Texture + Moisture, data = fish)
fish.lm <- lm(cbind(Aroma, Flavor, Texture, Moisture) ~ MethodFactor, data = fish)

lda.sig(fish.lda, fish.lm)
```

```
## [1] "EigenValues:  3.03556432554679"  "EigenValues:  0.126104161122293"
## [1] "Test Statistic for Eigenvalue  1 :  47.6881706842222"
## [1] "P-value for Eigenvalue  1 :  1.13324864714492e-07"
## [1] "EigenValues:  3.03556432554679"  "EigenValues:  0.126104161122293"
## [1] "Test Statistic for Eigenvalue  2 :  3.74106697260471"
## [1] "P-value for Eigenvalue  2 :  0.290815848532028"
```

There is convincing evidence that at least one of the eigenvalues are non-zero (p-value = 1.1332e-07). There is no evidence that the second eigenvalue is non-zero (p-value = 0.2908). There is convincing evidence that the first discriminant function is significant while the second one is not.

```
3.03556432554679 / (3.03556432554679 + 0.126104161122293) = 0.9501147
```

The first discriminant function has a total variance explained of 0.95. This agrees with the significance proven by a Wilk's Lambda Test.

(b) Find the standardized coefficients and comment on the contribution of the variables to separation of groups

```
# run manova to retrieve the Error Matrix (E)
fish.manova <- Manova(fish.lm)
E <- fish.manova$SSPE

# calculate pooled Standard variance. This formula works for J groups
n <- fish %>% count %>% pull
j <- fish %>% dplyr::select(Method) %>%  unique %>% count %>% pull
S.pl <- E / (n - j)

# calculate and display standardized coefficients
a.star <- sqrt(diag(S.pl)) * fish.lda$scaling[,1]
a.star %>% kable(
  caption = "Standardized Coefficients"
)
```

Flavor followed by Texture are the most dominant variables for determining the cooking method.

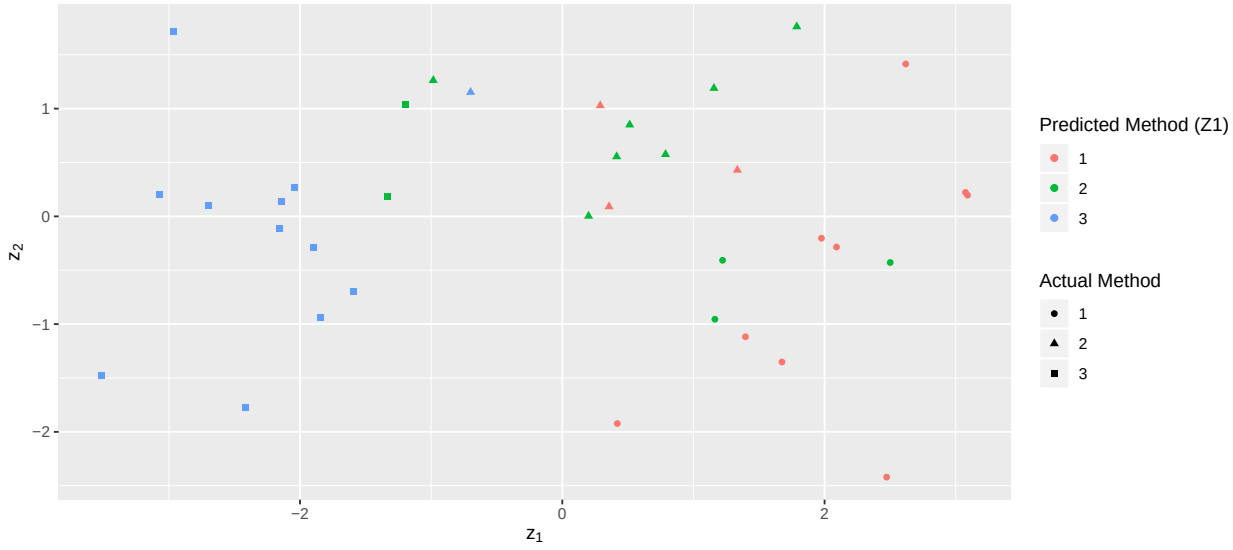
Table 2: Standardized Coefficients

	x
Aroma	-0.0268346
Flavor	1.6374967
Texture	-1.1244356
Moisture	-0.5127268

(c) Plot the first two discriminant functions for each observation and for the mean vectors

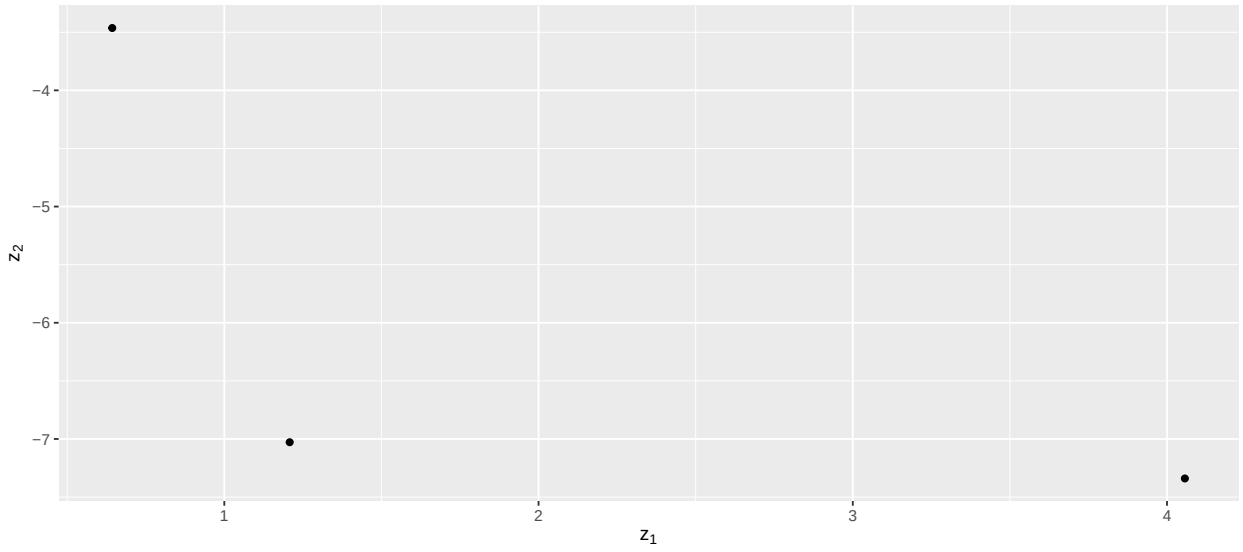
```
pred.z1 <- predict(fish.lda)$x[,1]
pred.z2 <- predict(fish.lda)$x[,2]

qplot(pred.z1, pred.z2, color=factor(fish$Method), shape = predict(fish.lda)$class) +
  scale_color_discrete(name="Predicted Method (Z1)") +
  scale_shape_discrete(name="Actual Method") +
  xlab(expression(z[1])) +
  ylab(expression(z[2]))
```




```
# Not sure if this is what is meant by plot the discrimant functions for the mean vectors
pred.z1.mean <- rowSums(fish.lda$scaling[,1] * fish.lda$means)
pred.z2.mean <- rowSums(fish.lda$scaling[,2] * fish.lda$means)

qplot(pred.z1.mean, pred.z2.mean) +
  scale_color_discrete(name="Predicted Method (Z1)") +
  scale_shape_discrete(name="Actual Method") +
  xlab(expression(z[1])) +
  ylab(expression(z[2]))
```



3

Do a classification analysis on the beetle data in beetle.csv.

(a) Find the classification function

What is the cutoff point:

$z = \frac{1}{2}(\bar{z}_1 + \bar{z}_2)$

```
zbar.o <- sum(beetles.lda$means[2,] * beetles.lda$scaling)
zbar.c <- sum(beetles.lda$means[1,] * beetles.lda$scaling)

z1 <- max(zbar.c, zbar.o)
z2 <- min(zbar.c, zbar.o)

cutoff <- 0.5 * (zbar.c + zbar.o)

#beetles$.class <- ifelse(
  #rowSums(beetles.lda$scaling * beetles[,-1]) > cutoff, "Oleracea", "Carduorum"
#)
```

Linear Discriminant Function

$z = 0.0931y_1$ - $0.0351y_2$ - $0.029y_3$ - $0.0383y_4$

Linear Classification Function

$$z(x) = \begin{cases} Oleracea & x > -4.2194 \\ Carduorum & else \end{cases}$$

(b) Find the classification table using the linear classification function.

```r
table(beetles$Species, predict(beetles.lda)$class)
```

```
##
##             Carduorum Oleracea
##   Carduorum        19        1
##   Oleracea          0       19
```

4

Do a classification analysis on the fish data in fish.csv.

(a) Find the linear classification functions.

$L_1(y) = \bar{y}_1^T S_{pl}^{-}1y - \frac{1}{2}\bar{y}_1^T S_{pl}^{-}1\bar{y}_1$ $L_2(y) = \bar{y}_2^T S_{pl}^{-}1y - \frac{1}{2}\bar{y}_2^T S_{pl}^{-}1\bar{y}_2$ $L_3(y) = \bar{y}_3^T S_{pl}^{-}1y - \frac{1}{2}\bar{y}_3^T S_{pl}^{-}1\bar{y}_3$

```r
# I left this in here because I spent a good amount of time on it. I didn't need to go this far to calc

E <- fish.manova$SSPE

# calculate pooled Standard variance. This formula works for J groups
n <- fish %>% count %>% pull
j <- fish %>% dplyr::select(Method) %>%  unique %>% count %>% pull
S.pl <- E / (n - j)

ybar <- mean(fish.lda$means)

# Linear Classification Function
L <- function(y) {
  sapply(1:3, function(i){
    ybar_i <- fish.lda$means[i,]
    ((ybar_i %*% solve(S.pl) * y) - 0.5 * (t(ybar_i) %*% solve(S.pl) * ybar_i)) %>% sum
  })
}

li.y <- apply(fish[,2:5], 1, L)

li <- apply(t(li.y), 1, function(x) {
  which(x==max(x))
})

# comparing my calculation of L_i against the original. Mine was a little worse than the lda function w
table(li, fish$Method)
```

```
##
## li   1  2  3
##   1  9  3  0
##   2  3  7  1
##   3  0  2 11
```

(b) Find the classification table using the linear classification functions in (a) (assuming population covariance matrices are equal).

```
table(fish$Method, predict(fish.lda)$class)
```

```
##
##      1  2  3
##   1  9  3  0
##   2  3  7  2
##   3  0  1 11
```

(c) Find the classification table using quadratic classification functions (assuming population covariance matrices are not equal)

```
fish.qda <- qda(MethodFactor ~ Aroma + Flavor + Texture + Moisture, data = fish)
table(fish$Method, predict(fish.qda)$class)
```

```
##
##      1  2  3
##   1 11  1  0
##   2  3  7  2
##   3  0  1 11
```