DA410_Project3_MattGraham

In project 3, we are working through Linear Discriminate Analysis (LDA) on newborns who were victims of Sudden Infant Death Syndrome (SIDS). Group 1 is our control group, Group 2 are victims of SIDS.

```
library(nnspat)  # used for dist2full()
library("dplyr")  # used to select numeric datatypes
library("ggplot2")
library(reshape)  # used for melting matricies
library("klaR")
library(MASS)
library(readxl)
library(BSDA)
```

Part 1: Creating and understanding our LDA

Import data

```
sids <- read_excel("C:/mattgraham93.github.io/school/22_3_DA410/data/SIDS.xlsx")

sids$HR <- as.numeric(sids$HR)
sids$BW <- as.numeric(sids$BW)
sids$Factor68 <- as.numeric(sids$Factor68)
sids$Gesage <- as.numeric(sids$Gesage)


sids
```

Group <dbl>	HR <dbl>	BW <dbl>	Factor68 <dbl>	Gesage <dbl>
1	115.6	3060	0.291	39
1	108.2	3570	0.277	40
1	114.2	3950	0.390	41
1	118.8	3480	0.339	40
1	76.9	3370	0.248	39
1	132.6	3260	0.342	40
1	107.7	4420	0.310	42
1	118.2	3560	0.220	40
1	126.6	3290	0.233	38
1	138.0	3010	0.309	40

1-10 of 65 rows Previous **1** 2 3 4 5 6 7 Next

Create LDA

```
sids.lda <- lda(Group ~ HR + BW + Factor68 + Gesage, data=sids, prior=c(0.5, 0.5))
```

Group means

```
sids.lda$means
```

```
##         HR       BW  Factor68 Gesage
## 1 129.1816 3429.898 0.3104082  40.00
## 2 132.9500 2964.688 0.4018125  39.25
```

Coefficients of discriminates

```
sids.lda$scaling
```

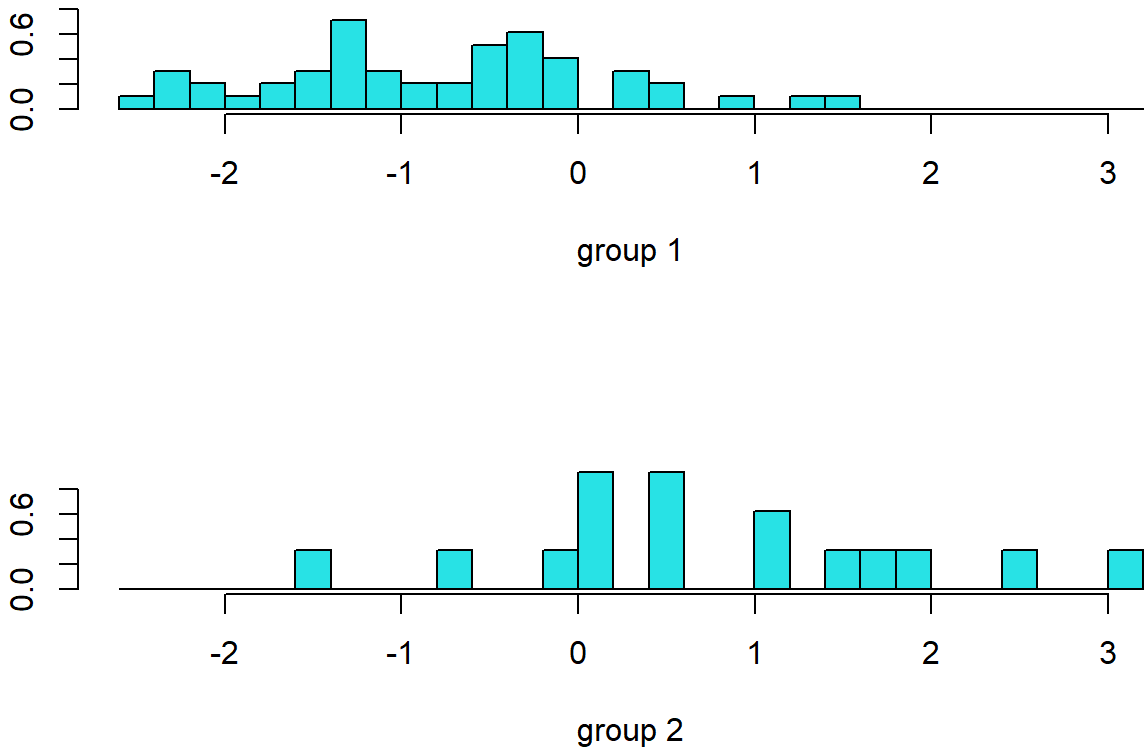
```
##                   LD1
## HR         0.001365159
## BW        -0.001115406
## Factor68 10.007618822
## Gesage   -0.157165685
```

Interpretation

When looking at our coefficients of linear discriminates, it is evident that 2 variables have the greatest impact on determining whether a baby has SIDS or not: Factor68 seems to greatly increase the likelihood while Gesage slighly reduces the risk.

Plot LDA

```
plot(sids.lda)
```



When observing the above plot, we can note how abnormal the spreads are between our groups. Group 2 displays how being ~5% beyond the expected mean can lead to a higher likelihood of a newborn developing SIDS.

Part 2: Predictions and summary

```
newinf <- rbind( c(100, 3000, 0.3, 40) ) # classification of newborn
dimnames(newinf) <- list(NULL, c('HR', 'BW', 'Factor68','Gesage'))
newinf <- data.frame(newinf)
predict(sids.lda, newdata = newinf)
```

```
## $class
## [1] 1
## Levels: 1 2
##
## $posterior
##           1         2
## 1 0.6658141 0.3341859
##
## $x
##          LD1
## 1 -0.4428163
```

Our output is fascinating. I thought of limiting it to just the class to keep it simple. Personally, seeing the full summary makes things interesting.

To start, we can conclude that our baby is likely to not be affected by SIDS. We can see this through the posterior probability returning ~0.666 and our predicted value ~-.443. When looking at our plot above, we can start to see why/where this is helpful.

Z test

```
group1.pred <- rbind( c(sids.lda$means[1,]) )
names(group1.pred) <- NULL
dimnames(group1.pred) <- list(NULL, c('HR', 'BW', 'Factor68','Gesage'))

group1.pred <- data.frame(group1.pred)

newinf.pred.val <- predict(sids.lda, newdata = newinf)$x
group1.pred.val <- predict(sids.lda, newdata=group1.pred)$x

z.stat <- (newinf.pred.val - group1.pred.val) / (sids.lda$svd / sqrt(sids.lda$N))
z.stat
```

```
##         LD1
## 1 0.4310675
```

Observing our z.stat above, we can conclude that we are far within rejection region at nearly any percentile.

This is exactly how we can prove to our new mothers how confident we are that their pregancies are within a range of normality.