574_HW5

Question 1

a

```
library(MASS)
library(ggplot2)

root <- read.table("Data/rootstock.txt", header = TRUE, stringsAsFactors = FALSE)

root$RootStock <- factor(root$RootStock)
str(root)
```

```
## 'data.frame':    48 obs. of  5 variables:
##  $ RootStock: Factor w/ 6 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 2 2 ...
##  $ Y1       : num  1.11 1.19 1.09 1.25 1.11 1.08 1.11 1.16 1.05 1.17 ...
##  $ Y2       : num  2.57 2.93 2.87 3.84 3.03 ...
##  $ Y3       : num  3.58 3.75 3.93 3.94 3.6 3.51 3.98 3.62 4.09 4.06 ...
##  $ Y4       : num  0.76 0.821 0.928 1.009 0.766 ...
```

```
lda_fit <- lda(RootStock ~ Y1 + Y2 + Y3 + Y4, data = root)

lda_fit$scaling
```

```
##           LD1        LD2       LD3       LD4
## Y1  3.0479952  -1.140083 -1.002448 23.419063
## Y2 -1.7025953  -1.215888  1.672714 -3.076804
## Y3  4.2332645   7.166403  3.045553 -2.011416
## Y4 -0.4785144 -11.520302 -5.506192  3.101660
```

b

```
prop_var <- lda_fit$svd^2 / sum(lda_fit$svd^2)
round(prop_var, 3)
```

```
## [1] 0.642 0.271 0.078 0.009
```

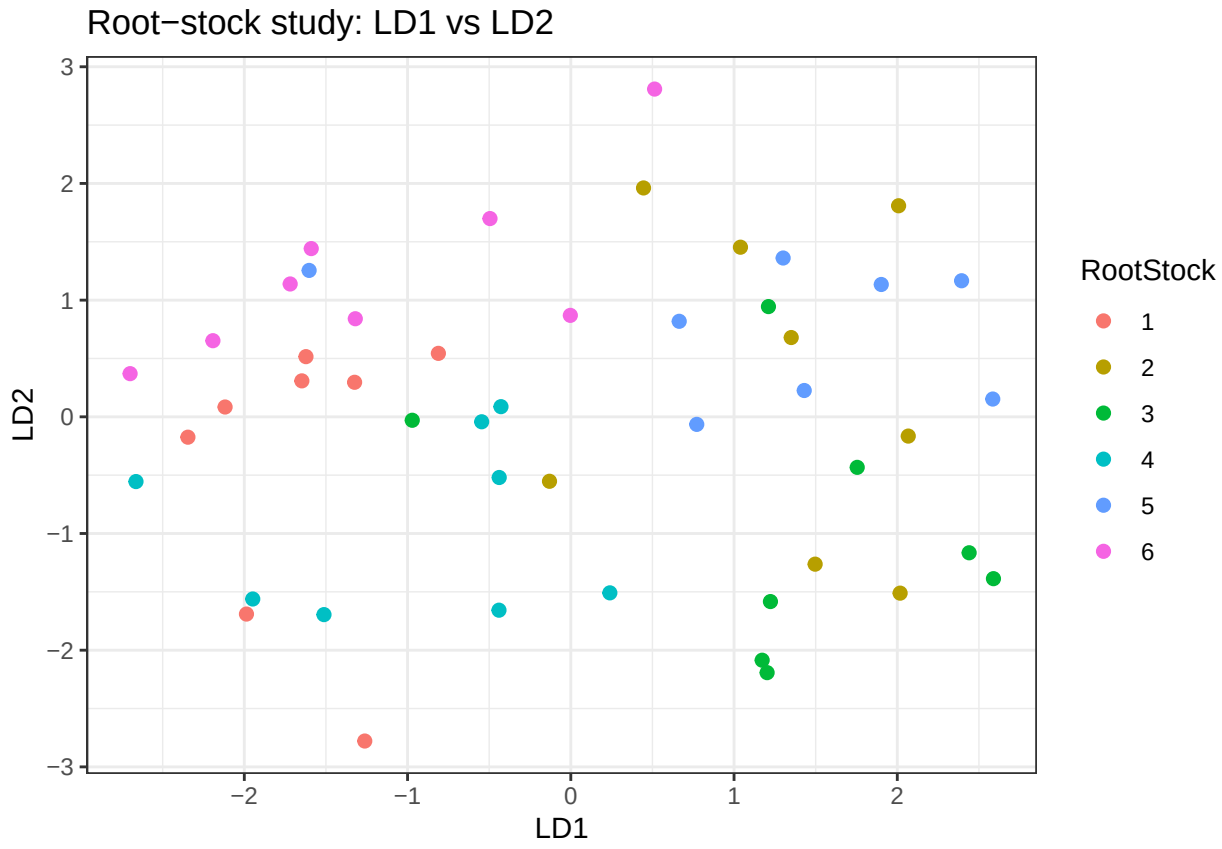
```
cumsum(prop_var)
```

```
## [1] 0.6420523 0.9127112 0.9911160 1.0000000
```

Because DF1 + DF2 explain 91 % of the between-group variance, I retain two discriminant functions.

c

```
scores <- as.data.frame(predict(lda_fit)$x)
scores$RootStock <- root$RootStock
ggplot(scores, aes(LD1, LD2, colour = RootStock)) +
  geom_point(size = 2) +
  labs(title = "Root-stock study: LD1 vs LD2") +
  theme_bw()
```



Stocks 1 and 6 cluster on the left of LD1 and around the midpoint of LD2. Stocks 2 and 3 cluster on the right of LD1 and spread across most of LD2. These two seem well separated. Stocks 4 and 5 are more mixed in with the rest and more difficult to separate.

d

```
new_tree <- data.frame(Y1 = 1.09, Y2 = 2.69, Y3 = 4.38, Y4 = 1.29)

pred <- predict(lda_fit, new_tree)
pred$class
```

```
## [1] 3
## Levels: 1 2 3 4 5 6
```

```
pred$posterior
```

```
##            1         2         3          4         5          6
## 1 0.00621337 0.2534963 0.3665987 0.03435888 0.3287843 0.01054838
```

Rootstock 3 has the largest posterior probability, but the model is not super confident between stocks 3 and 5 both having a probability around .3

Question 2

Basic Data

```
a <- c(0.32, 0.05, 0.19, -2.13, -0.14)

bird1 <- c(156, 245, 31.60, 18.50, 20.50)

mean_surv <- c(157.3, 241.0, 31.4, 18.5, 20.8)
mean_died <- c(158.4, 241.5, 31.4, 18.4, 20.8)

pi <- c(surv = 0.43, died = 0.57)
```

a

```
Z  <- sum(a * bird1)
cat("Bird 1 Z score is", Z)
```

```
## Bird 1 Z score is 25.899
```

b

```
Z_surv <- sum(a * mean_surv)
Z_died <- sum(a * mean_died)

cat("Survived mean is", Z_surv, " \nDied mean is
    ", Z_died)
```

```
## Survived mean is 26.035
## Died mean is
##      26.625
```

c

```
midpt   <- 0.5 * (Z_surv + Z_died)          # 26.33
class_eq <- ifelse(Z < midpt, "survived", "died")

cat("The Midpoint is", midpt, "\nGiven this data, Bird 1 ", class_eq)
```

```
## The Midpoint is 26.33
## Given this data, Bird 1  survived
```

d

The Probaibliteis provided classify the bird as dead, but differs from the midpoint rule where the bird survives. This is because predict uses a larger prior for "died" which shifts the decision boundary.

e

We assume that the sample rates accurately represent the true population class probabilities for sparrows that strike windows

f

No. LDA uses the pooled covariance matrix and effectively rescales each variable naturally.