A3

Elliot Gavrin

2023-01-23

https://github.com/19eag3/BIOL432_Assignment3

Part I:

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

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

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

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

```r
MyData <-read.csv("https://colauttilab.github.io/Data/ColauttiBarrett2013Data.csv")
```

```r
dim(MyData)
```

```
## [1] 432  23
```

```r
head(MyData)
```

```
##   Ind  Site Row Pos Mat Pop Region Flwr07 FVeg07 InfMass07 Fruits07 Flwr08
## 1   1 1_BEF   1   1 C13   C  NORTH  39285   41.0       9.7      625  39613
## 2   2 1_BEF   1   2 C11   C  NORTH  39293   40.0       6.5      329  39625
## 3   3 1_BEF   1   3 A16   A  NORTH  39286   32.5      16.0     1029  39610
## 4   4 1_BEF   1   4 C20   C  NORTH  39305   45.0       1.5       46  39638
## 5   5 1_BEF   1   5  T3   T  SOUTH  39293   47.5      17.4     1121  39616
## 6   6 1_BEF   1   6 C18   C  NORTH  39291   55.0      13.1      504  39625
##   FVeg08 HVeg08 InfMass08 Flwr09 FVeg09 HVeg09 InfMass09 Flwr10 FVeg10 HVeg10
## 1   79.0   71.0       9.7  39980     64   82.0      11.7  40345   68.7   66.0
## 2  117.0  108.0      36.1  39993     70   65.5      10.1  40350   83.4   96.0
## 3   70.0   62.5      28.9     NA     NA     NA        NA     NA     NA     NA
## 4   94.5   87.5      18.2  40000     84   72.0       9.6  40369   69.9   57.5
## 5   97.0   90.0      47.9  39986     66   65.5      38.1  40353   70.1   65.5
## 6  128.5  124.0      30.7  39995     91   82.0      16.7  40357   97.0   77.0
##   InfMass10
## 1       1.9
## 2       7.3
## 3        NA
## 4       5.7
## 5      27.9
## 6       4.9
```

```r
tail(MyData)
```

```
##     Ind      Site Row Pos Mat Pop Region Flwr07 FVeg07 InfMass07 Fruits07
## 427 427 3_Timmins   6  19 E16   E    MID     NA     NA        NA        0
## 428 428 3_Timmins   6  20 C30   C  NORTH     NA     NA        NA        0
## 429 429 3_Timmins   6  21 A18   A  NORTH     NA     NA        NA        0
## 430 430 3_Timmins   6  22  T7   T  SOUTH     NA     NA        NA        0
## 431 431 3_Timmins   6  23 S10   S  SOUTH     NA     NA        NA        0
## 432 432 3_Timmins   6  24  T2   T  SOUTH     NA     NA        NA        0
##     Flwr08 FVeg08 HVeg08 InfMass08 Flwr09 FVeg09 HVeg09 InfMass09 Flwr10 FVeg10
## 427     NA     NA     NA        NA     NA     NA     NA        NA     NA     NA
## 428  39664  101.5     96      13.7  40035   90.5   83.5      10.3  40387     79
## 429  39657   63.5     52       7.1  40018   68.5   61.0       6.5  40373     71
## 430     NA     NA     NA        NA     NA     NA     NA        NA     NA     NA
## 431     NA     NA     NA        NA     NA     NA     NA        NA     NA     NA
## 432     NA     NA     NA        NA     NA     NA     NA        NA     NA     NA
##     HVeg10 InfMass10
## 427     NA        NA
## 428   75.8       6.9
## 429   67.2       4.5
## 430     NA        NA
## 431     NA        NA
## 432     NA        NA
```

```r
str(MyData)
```

```
## 'data.frame':    432 obs. of  23 variables:
##  $ Ind      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Site     : chr  "1_BEF" "1_BEF" "1_BEF" "1_BEF" ...
##  $ Row      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Pos      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Mat      : chr  "C13" "C11" "A16" "C20" ...
##  $ Pop      : chr  "C" "C" "A" "C" ...
##  $ Region   : chr  "NORTH" "NORTH" "NORTH" "NORTH" ...
##  $ Flwr07   : int  39285 39293 39286 39305 39293 39291 39287 39293 39287 39303 ...
##  $ FVeg07   : num  41 40 32.5 45 47.5 55 51.5 38 51 37.5 ...
##  $ InfMass07: num  9.7 6.5 16 1.5 17.4 13.1 3 0.6 0 12.2 ...
##  $ Fruits07 : int  625 329 1029 46 1121 504 122 15 0 896 ...
##  $ Flwr08   : int  39613 39625 39610 39638 39616 39625 39614 39621 39617 39638 ...
##  $ FVeg08   : num  79 117 70 94.5 97 ...
##  $ HVeg08   : num  71 108 62.5 87.5 90 124 85 76.5 91 96.5 ...
##  $ InfMass08: num  9.7 36.1 28.9 18.2 47.9 30.7 23.3 8.5 28.5 16.6 ...
##  $ Flwr09   : int  39980 39993 NA 40000 39986 39995 39990 39993 39996 39997 ...
##  $ FVeg09   : num  64 70 NA 84 66 91 80 62 58 99 ...
##  $ HVeg09   : num  82 65.5 NA 72 65.5 82 76 60.5 56.5 93.5 ...
##  $ InfMass09: num  11.7 10.1 NA 9.6 38.1 16.7 11.1 2.6 1.7 18.8 ...
##  $ Flwr10   : int  40345 40350 NA 40369 40353 40357 40350 NA 40358 40357 ...
##  $ FVeg10   : num  68.7 83.4 NA 69.9 70.1 97 105 NA 50.8 85.1 ...
##  $ HVeg10   : num  66 96 NA 57.5 65.5 77 98 57 49.5 82 ...
##  $ InfMass10: num  1.9 7.3 NA 5.7 27.9 4.9 5.3 5.5 1.5 15.3 ...
```

```r
summary(MyData)
```

```
##       Ind            Site                Row             Pos
##  Min.   :  1.0   Length:432         Min.   :1.000   Min.   : 1.00
##  1st Qu.:108.8   Class :character   1st Qu.:2.000   1st Qu.: 6.75
##  Median :216.5   Mode  :character   Median :3.500   Median :12.50
##  Mean   :216.5                      Mean   :3.495   Mean   :12.53
##  3rd Qu.:324.2                      3rd Qu.:5.000   3rd Qu.:18.25
##  Max.   :432.0                      Max.   :6.000   Max.   :25.00
##
##      Mat                Pop               Region             Flwr07
##  Length:432         Length:432         Length:432         Min.   :39275
##  Class :character   Class :character   Class :character   1st Qu.:39294
##  Mode  :character   Mode  :character   Mode  :character   Median :39306
##                                                           Mean   :39310
##                                                           3rd Qu.:39317
##                                                           Max.   :39368
##                                                           NA's   :170
##      FVeg07        InfMass07        Fruits07          Flwr08
##  Min.   :18.50   Min.   : 0.00   Min.   :   0.0   Min.   :39608
##  1st Qu.:35.12   1st Qu.: 2.50   1st Qu.:   0.0   1st Qu.:39632
##  Median :45.00   Median : 8.80   Median :   0.0   Median :39640
##  Mean   :45.30   Mean   :12.07   Mean   : 251.0   Mean   :39645
##  3rd Qu.:54.88   3rd Qu.:17.10   3rd Qu.: 310.2   3rd Qu.:39657
##  Max.   :92.00   Max.   :70.10   Max.   :3211.0   Max.   :39709
##  NA's   :170     NA's   :193                      NA's   :65
##      FVeg08           HVeg08         InfMass08         Flwr09
##  Min.   : 42.00   Min.   : 35.00   Min.   : 0.00   Min.   :39980
##  1st Qu.: 73.00   1st Qu.: 66.50   1st Qu.: 8.20   1st Qu.:40002
##  Median : 88.50   Median : 82.00   Median :19.30   Median :40021
##  Mean   : 90.96   Mean   : 84.42   Mean   :23.65   Mean   :40019
##  3rd Qu.:107.12   3rd Qu.:100.50   3rd Qu.:33.75   3rd Qu.:40035
##  Max.   :164.00   Max.   :159.50   Max.   :98.50   Max.   :40072
##  NA's   :64       NA's   :47       NA's   :50      NA's   :97
##      FVeg09           HVeg09         InfMass09          Flwr10
##  Min.   : 23.00   Min.   : 27.00   Min.   :  0.00   Min.   :40332
##  1st Qu.: 74.00   1st Qu.: 68.50   1st Qu.:  3.20   1st Qu.:40358
##  Median : 92.00   Median : 87.00   Median : 11.70   Median :40366
##  Mean   : 92.87   Mean   : 86.16   Mean   : 17.13   Mean   :40371
##  3rd Qu.:112.00   3rd Qu.:102.50   3rd Qu.: 25.60   3rd Qu.:40385
##  Max.   :151.00   Max.   :156.00   Max.   :113.40   Max.   :40422
##  NA's   :93       NA's   :75       NA's   :75       NA's   :112
##      FVeg10           HVeg10         InfMass10
##  Min.   : 28.10   Min.   : 26.10   Min.   :  0.00
##  1st Qu.: 67.00   1st Qu.: 63.92   1st Qu.:  2.00
##  Median : 81.00   Median : 77.25   Median :  7.65
##  Mean   : 82.29   Mean   : 76.26   Mean   : 18.06
##  3rd Qu.: 96.00   3rd Qu.: 87.97   3rd Qu.: 24.68
##  Max.   :142.00   Max.   :142.00   Max.   :135.40
##  NA's   :112      NA's   :86       NA's   :86
```

```r
respDat <- MyData %>% #Responses Dataset
  dplyr::select(1:7)
features <- MyData %>% # Features Dataset
  dplyr::select(-c(1:7))

scaled <-features %>% # scaled variables dataset
  mutate_all(scale)

scaled %>% #Determine which variables have NA
  select_if(function(x) any(is.na(x))) %>%
  names()
```

```
##  [1] "Flwr07"    "FVeg07"    "InfMass07" "Flwr08"    "FVeg08"    "HVeg08"
##  [7] "InfMass08" "Flwr09"    "FVeg09"    "HVeg09"    "InfMass09" "Flwr10"
## [13] "FVeg10"    "HVeg10"    "InfMass10"
```

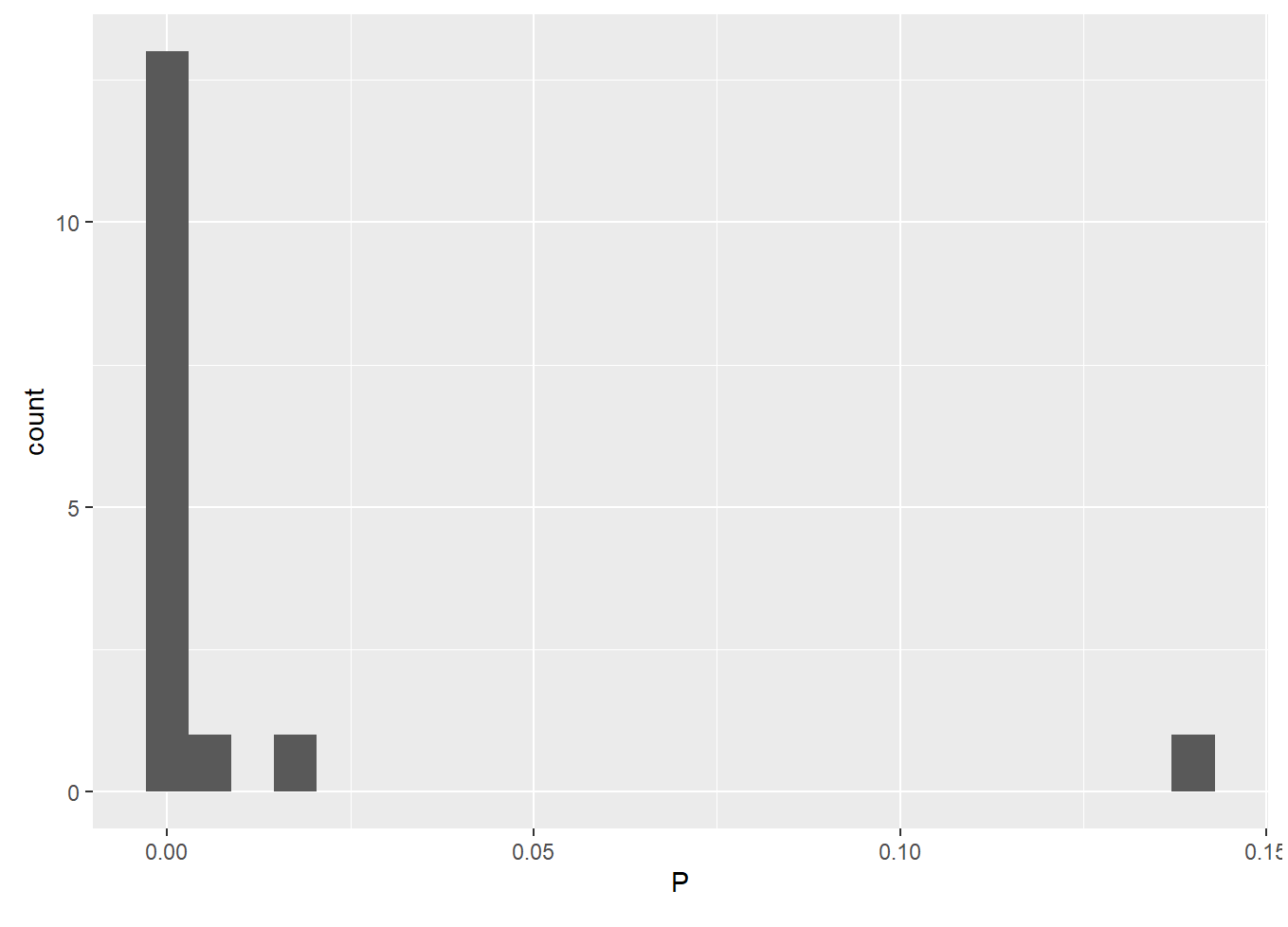
```r
ScalComp<-scaled %>% #replace NA with 0 to maintain the constant mean
  mutate(Flwr07 = ifelse(is.na(Flwr07),0,Flwr07),
         Flwr08 = ifelse(is.na(Flwr08),0,Flwr08),
         Flwr09 = ifelse(is.na(Flwr09),0,Flwr09),
         Flwr10 = ifelse(is.na(Flwr10),0,Flwr10),
         FVeg07 = ifelse(is.na(FVeg07),0,FVeg07),
         FVeg08 = ifelse(is.na(FVeg08),0,FVeg08),
         FVeg09 = ifelse(is.na(FVeg09),0,FVeg09),
         FVeg10 = ifelse(is.na(FVeg10),0,FVeg10),
         HVeg08 = ifelse(is.na(HVeg08),0,HVeg08),
         HVeg09 = ifelse(is.na(HVeg09),0,HVeg09),
         HVeg10 = ifelse(is.na(HVeg10),0,HVeg10),
         InfMass07 = ifelse(is.na(InfMass07),0,InfMass07),
         InfMass08 = ifelse(is.na(InfMass08),0,InfMass08),
         InfMass09 = ifelse(is.na(InfMass09),0,InfMass09),
         InfMass10 = ifelse(is.na(InfMass10),0,InfMass10),
         )
ScalComp %>% #double checking for any NAs
  select_if(function(x) any(is.na(x))) %>%
  names()
```

```
## character(0)
```

```r
#Dimention Reducing
library(tidyr)
FeatureSel<-ScalComp %>%
  mutate(Pop=MyData$Pop) %>%
  pivot_longer(cols=-Pop,
               names_to="Trait",
               values_to="Conc")
FeatureSel %>%
  group_by(Trait) %>%
  summarise(MeanConc=mean(Conc),
            sd=sd(Conc),
            max=max(Conc),
            min=min(Conc))
```

```
## # A tibble: 16 × 5
##    Trait      MeanConc    sd   max    min
##    <chr>         <dbl> <dbl> <dbl>  <dbl>
##  1 Flwr07    -5.05e-14 0.778  2.82 -1.68
##  2 Flwr08    -6.89e-14 0.922  3.11 -1.82
##  3 Flwr09    -4.78e-14 0.880  2.80 -2.04
##  4 Flwr10     6.10e-14 0.860  2.86 -2.23
##  5 Fruits07   3.00e-17 1      6.09 -0.516
##  6 FVeg07    -1.54e-16 0.778  3.57 -2.05
##  7 FVeg08     1.80e-16 0.923  3.08 -2.07
##  8 FVeg09     2.11e-16 0.886  2.32 -2.78
##  9 FVeg10     2.24e-16 0.860  2.87 -2.61
## 10 HVeg08     8.76e-17 0.944  3.21 -2.11
## 11 HVeg09    -1.62e-16 0.909  2.95 -2.50
## 12 HVeg10    -2.10e-16 0.895  3.49 -2.66
## 13 InfMass07  1.59e-17 0.743  4.56 -0.949
## 14 InfMass08  1.03e-17 0.940  3.68 -1.16
## 15 InfMass09  1.85e-17 0.909  5.25 -0.934
## 16 InfMass10  3.54e-17 0.895  4.80 -0.739
```

```r
PVals <-FeatureSel %>%
  group_by(Trait) %>%
  summarise(P=anova(lm(Conc ~ Pop))[1,"Pr(>F)"]) %>%
  dplyr::select(Trait,P)
ggplot(aes(x=P), data=PVals)+
  geom_histogram(bins=25)
```




```r
Keep<-PVals %>% #Keeping features that are less then 0.05 p value
  filter(PVals$P <0.05)
Keep<-paste(Keep$Trait)

ScaledSub<-ScalComp %>%
  dplyr::select(all_of(Keep))
names(ScaledSub) #Scaled Features Data set
```

```
##  [1] "Flwr08"    "Flwr09"    "Flwr10"    "Fruits07"  "FVeg07"    "FVeg08"
##  [7] "FVeg09"    "FVeg10"    "HVeg08"    "HVeg09"    "HVeg10"    "InfMass07"
## [13] "InfMass08" "InfMass09" "InfMass10"
```

In the Discriminant Analysis Tutorial we went through the process of writing linear models to select appropriate features. Briefly explain why that is not necessary for this data set.

The purpose of writing linear models to select appropriate features was to reduce the amount of dimensions to only the dimensions with a p value of less than 0.05. This is not necessary for this data set because there are not many dimensions and the majority have a p-value under 0.05.

1. Use the lda() function in the MASS package to run one or more LDA model(s) that distinguish genetic populations and regions.

```r
LDAPop <-lda(x=ScaledSub, grouping=MyData$Pop)
LDAReg <-lda(x=ScaledSub, grouping=MyData$Region)
```

2. Explain how many LD axes you need to distinguish among the three sites, and among the six populations.

LDA generates axises to create distinctions between categories by minimizing scatter. Therefore, there will be 15 LD axises.

3. Explore the objects in your LD models. What does the $scaling slice show you? How does this relate to the LD eigenvectors? Briefly explain the difference between the PC axes of a PCA and the LD axes of an LDA.

The scaling is normalized for the groups. The scaling slice shows the loading of each variable. The largest loadings (positive or negative) are the variables that contribute most to the discriminant function. This relates to the LD eigenvectors… The differences between PCA and LDA is that the axes of an LDA show how well the categories are distinguished from one another. The PC axes show the variation between variables in the data set.

```r
head(LDAPop$scaling)
```

```
##                   LD1         LD2        LD3        LD4        LD5
## Flwr08    0.308090159  0.78793184 -0.3237967  0.5445341 -0.3056926
## Flwr09   -0.234172385 -0.35100353  0.4160291  0.6408027  1.4526164
## Flwr10   -0.005226207  0.01881148 -0.3502045 -0.9001104 -0.3140239
## Fruits07  0.291674638  0.14646070 -0.3194333  0.2347204  0.2262379
## FVeg07    0.423301694  0.36298383  0.1788792  0.7694274 -0.7727573
## FVeg08    0.168407935 -0.92885661  1.1829844  0.4644879  1.1858798
```

```r
dim(LDAPop$scaling)
```

```
## [1] 15  5
```

```r
head(LDAReg$scaling)
```

```
##                   LD1          LD2
## Flwr08    0.648033702 -0.439700212
## Flwr09   -0.184880907 -0.008613114
## Flwr10   -0.009696155  0.168442930
## Fruits07  0.328461445 -0.370684749
## FVeg07    0.362555829 -0.213020913
## FVeg08   -0.371419545  0.614938817
```

```r
dim(LDAReg$scaling)
```

```
## [1] 15  2
```

```r
round(LDAPop$scaling, 2)
```

```
##             LD1   LD2   LD3   LD4   LD5
## Flwr08     0.31  0.79 -0.32  0.54 -0.31
## Flwr09    -0.23 -0.35  0.42  0.64  1.45
## Flwr10    -0.01  0.02 -0.35 -0.90 -0.31
## Fruits07   0.29  0.15 -0.32  0.23  0.23
## FVeg07     0.42  0.36  0.18  0.77 -0.77
## FVeg08     0.17 -0.93  1.18  0.46  1.19
## FVeg09     0.18  0.03  0.42 -0.20 -0.82
## FVeg10     0.57  0.37 -0.41  0.24 -0.03
## HVeg08     0.44  0.23 -1.17 -0.58  0.05
## HVeg09     0.22  0.08 -0.17 -0.83  0.42
## HVeg10    -0.15 -0.64 -0.30  0.64 -0.77
## InfMass07  0.14  0.00  0.46 -0.99 -0.41
## InfMass08 -0.39 -0.76 -0.56  0.15  0.08
## InfMass09  0.05 -0.42  1.18  0.11  0.07
## InfMass10  0.21  0.94  0.03 -0.15  0.49
```

```r
round(LDAReg$scaling, 2)
```

```
##             LD1   LD2
## Flwr08     0.65 -0.44
## Flwr09    -0.18 -0.01
## Flwr10    -0.01  0.17
## Fruits07   0.33 -0.37
## FVeg07     0.36 -0.21
## FVeg08    -0.37  0.61
## FVeg09    -0.01  0.44
## FVeg10     0.60 -0.43
## HVeg08     0.51 -0.66
## HVeg09     0.26  0.30
## HVeg10    -0.56 -0.69
## InfMass07  0.00  0.88
## InfMass08 -0.65 -0.65
## InfMass09 -0.30  0.86
## InfMass10  0.76  0.25
```

4.

```r
LDAPop_pred<-predict(LDAPop)
str(LDAPop_pred)
```

```
## List of 3
##  $ class    : Factor w/ 6 levels "A","C","E","J",..: 3 2 2 2 2 4 2 3 2 5 ...
##  $ posterior: num [1:432, 1:6] 0.2183 0.0276 0.0704 0.1765 0.0607 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : NULL
##   .. ..$ : chr [1:6] "A" "C" "E" "J" ...
##  $ x        : num [1:432, 1:5] -0.689 -0.37 -1.094 -0.444 -0.188 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : NULL
##   .. ..$ : chr [1:5] "LD1" "LD2" "LD3" "LD4" ...
```

```r
head(LDAPop_pred$posterior)
```

```
##                A         C           E          J           S          T
## [1,] 0.218251311 0.2935224 0.321705024 0.06398673 0.027081100 0.07545344
## [2,] 0.027641572 0.9213695 0.011303922 0.02186120 0.006297469 0.01152632
## [3,] 0.070374803 0.4910398 0.368436274 0.03106982 0.006305244 0.03277408
## [4,] 0.176467821 0.2902972 0.246966848 0.07490051 0.052524414 0.15884316
## [5,] 0.060668260 0.5203577 0.097616183 0.22187954 0.024550974 0.07492734
## [6,] 0.004810821 0.1784533 0.002478227 0.48038752 0.305866909 0.02800322
```

```r
dim(LDAPop_pred$x)
```

```
## [1] 432   5
```

```r
LDAReg_pred<-predict(LDAReg)
str(LDAReg_pred)
```

```
## List of 3
##  $ class    : Factor w/ 3 levels "MID","NORTH",..: 1 2 1 1 1 2 2 2 2 1 ...
##  $ posterior: num [1:432, 1:3] 0.5 0.101 0.505 0.444 0.5 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : NULL
##   .. ..$ : chr [1:3] "MID" "NORTH" "SOUTH"
##  $ x        : num [1:432, 1:2] -0.686 -1.75 -1.26 -0.215 -0.971 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : NULL
##   .. ..$ : chr [1:2] "LD1" "LD2"
```

```r
head(LDAReg_pred$posterior)
```

```
##            MID     NORTH      SOUTH
## [1,] 0.4996583 0.3822577 0.11808402
## [2,] 0.1011240 0.8667695 0.03210652
## [3,] 0.5052991 0.4390376 0.05566333
## [4,] 0.4443613 0.3415585 0.21408027
## [5,] 0.4999942 0.4175263 0.08247955
## [6,] 0.3223039 0.3647091 0.31298703
```

```r
dim(LDAReg_pred$x)
```

```
## [1] 432   2
```

```r
xtabs(~MyData$Pop+LDAPop_pred$class)
```

```
##           LDAPop_pred$class
## MyData$Pop  A  C  E  J  S  T
##          A 11 16 15  0  0  3
##          C  3 57 15  3  3  5
##          E  5 15 62  1  2  1
##          J  1 11  4 38 21  7
##          S  0 10  1 18 46  5
##          T  2 16  8  4  9 14
```

```r
xtabs(~MyData$Region+LDAReg_pred$class)
```

```
##              LDAReg_pred$class
## MyData$Region MID NORTH SOUTH
##         MID   103    34    31
##         NORTH  48    73    10
##         SOUTH  48     8    77
```

5. Write some text to explain what you learned about the Lythrum data from your LDA models. Compare results to the PCA results and projection of loadings in the PCA Tutorial. Which traits distinguish genetic populations and regions best, respectively? Formulate biological hypotheses to explain the LDA results. If you need a refresher on this experiment, recall that we also used it in the GAM Chapter in R Stats Crash Course book on Perusall.

I learned that there are certain traits that distinguish Lythrum from each other. These can be examined further by comparing the scaling factors in the LDA for regions and population to determine which traits contribute most towards distinguishing populations from each other. PCA only looks for the groupings with the most variation, therefore there is a lot of overlap in the PCA results compared to the LDA. LDA is efficient at separating data into distinct categories.

In the populations, FVeg07 abd FVeg10 have a strong positive contribution and InfMass08 and Flwr09 have a strong negative contribution to LD1. Flwr08 and InfMass10 have a strong positive contribution and FVeg08, HVeg10, InfMass08, and InfMass09 have a strong negative contribution in LD2. In the regions, Flwr08, Fruits07, and InfMass10 have a strong positive contribution and HVeg10, InfMass08 have a strong negative contribution to LD1. FVeg08 and InfMass09 have a strong positive contribution and HVeg8, HVeg10, InfMass08, and Fveg10 have a strong negative contribution in LD2.

Possible explanations could be a genetic difference in the gene pool between regions could cause distict categorizations of region and genetic population. There could be unique environmental pressures in each reagion that cause these differences in genetic traits.