LDA,QDA,Normality

November 24, 2020

[1]:
```
set.seed(72007)
```

[2]:
```
#Drawing ninety six percent random sample from Iris data
set.seed(72007)
index<-sample(1:nrow(iris), nrow(iris)*0.96)
index
sdata<-iris[index,]
```

[2]:
1. 11 2. 138 3. 67 4. 121 5. 85 6. 147 7. 10 8. 130 9. 124 10. 117 11. 69 12. 5 13. 145 14. 99 15. 37 16. 26 17. 149 18. 93 19. 42 20. 43 21. 16 22. 119 23. 50 24. 97 25. 23 26. 144 27. 122 28. 126 29. 148 30. 127 31. 88 32. 80 33. 29 34. 64 35. 78 36. 105 37. 19 38. 47 39. 62 40. 7 41. 15 42. 87 43. 112 44. 9 45. 100 46. 92 47. 150 48. 35 49. 52 50. 71 51. 59 52. 72 53. 34 54. 133 55. 111 56. 86 57. 68 58. 27 59. 8 60. 49 61. 89 62. 128 63. 91 64. 6 65. 81 66. 136 67. 63 68. 143 69. 70 70. 53 71. 140 72. 98 73. 94 74. 120 75. 83 76. 3 77. 131 78. 77 79. 57 80. 28 81. 45 82. 101 83. 4 84. 118 85. 75 86. 56 87. 142 88. 39 89. 132 90. 13 91. 109 92. 107 93. 114 94. 12 95. 51 96. 95 97. 110 98. 96 99. 137 100. 82 101. 84 102. 2 103. 41 104. 1 105. 65 106. 76 107. 58 108. 36 109. 134 110. 115 111. 135 112. 21 113. 22 114. 79 115. 90 116. 74 117. 60 118. 24 119. 103 120. 108 121. 14 122. 102 123. 40 124. 139 125. 106 126. 141 127. 38 128. 31 129. 32 130. 129 131. 33 132. 113 133. 17 134. 73 135. 123 136. 116 137. 18 138. 30 139. 44 140. 54 141. 48 142. 20 143. 146 144. 66

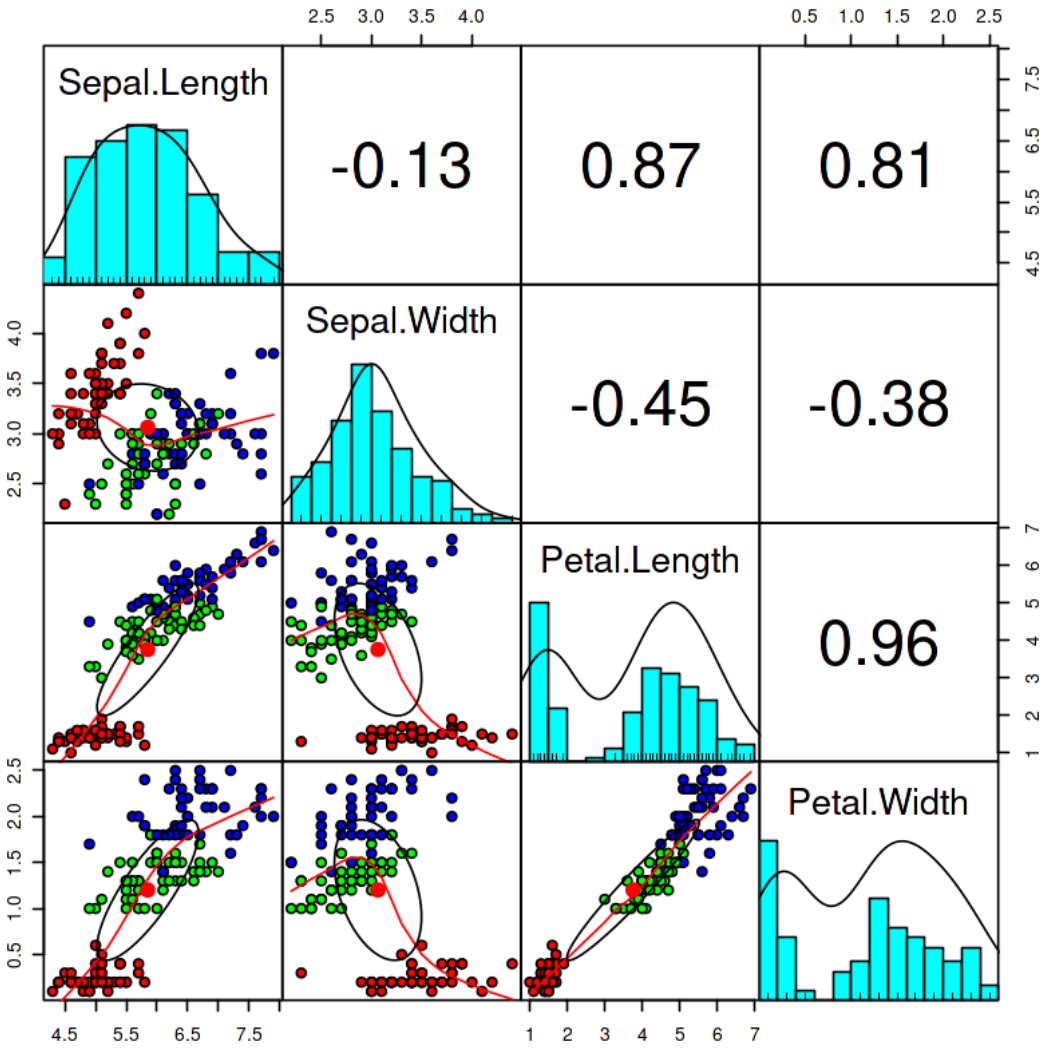
[3]:
```
head(sdata)
```

[3]:
A data.frame: 6 × 5

	Sepal.Length <dbl>	Sepal.Width <dbl>	Petal.Length <dbl>	Petal.Width <dbl>	Species <fct>
11	5.4	3.7	1.5	0.2	setosa
138	6.4	3.1	5.5	1.8	virginica
67	5.6	3.0	4.5	1.5	versicolor
121	6.9	3.2	5.7	2.3	virginica
85	5.4	3.0	4.5	1.5	versicolor
147	6.3	2.5	5.0	1.9	virginica

[4]:
```
#constructing scatter plots
library(psych)
pairs.panels(sdata[1:4],
             gap=0,
             bg=c("red","green","blue")[sdata$Species],
             pch=21
             )
```

[4]:

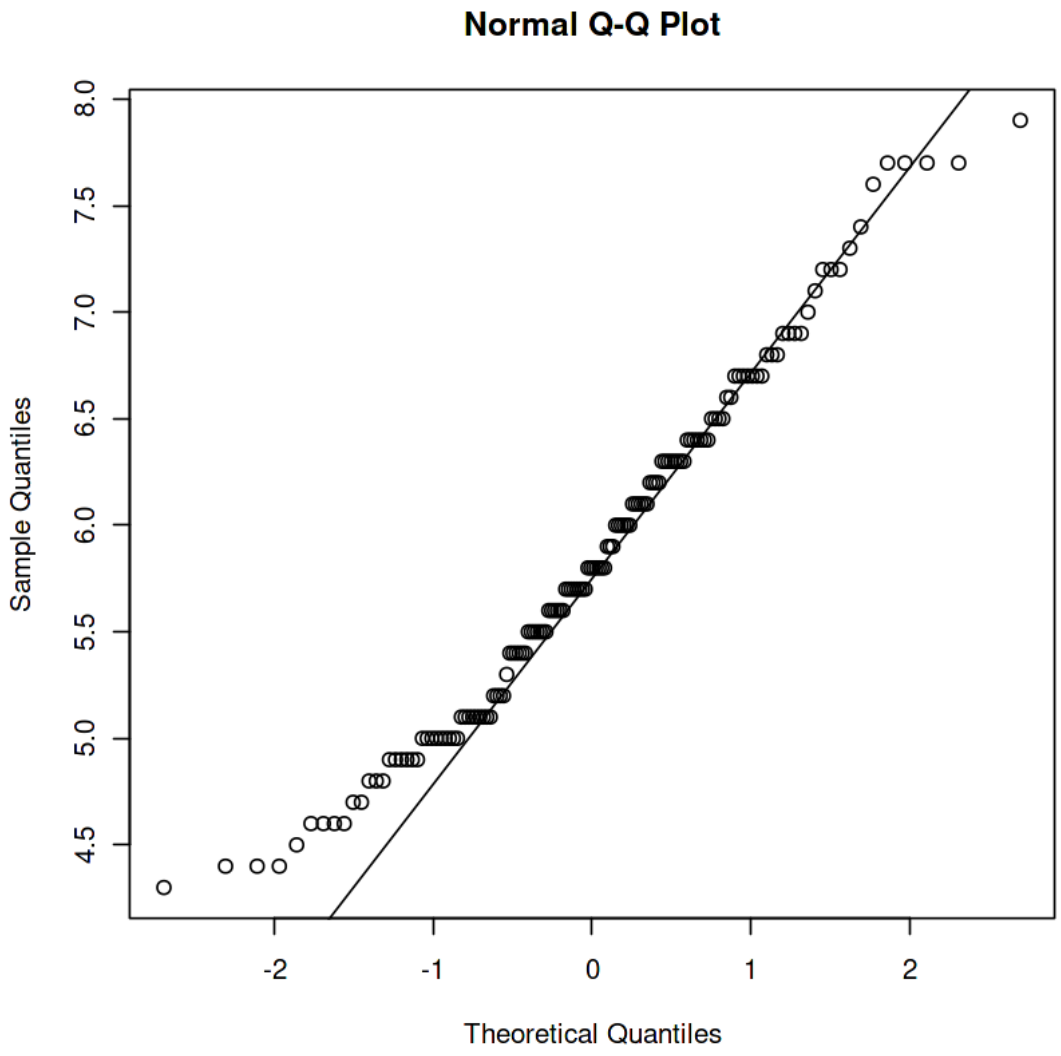


From scatter plot we see that sepal.length is negatively skewed,sepal.width is positively skewed,petal.length and petal.width are not normally distributed.

[5]:
```
##====Q-Q Plot===##
par(mfrow=c(2,2))
```

[6]:
```
qqnorm(sdata[,1])
qqline(sdata[,1])
```

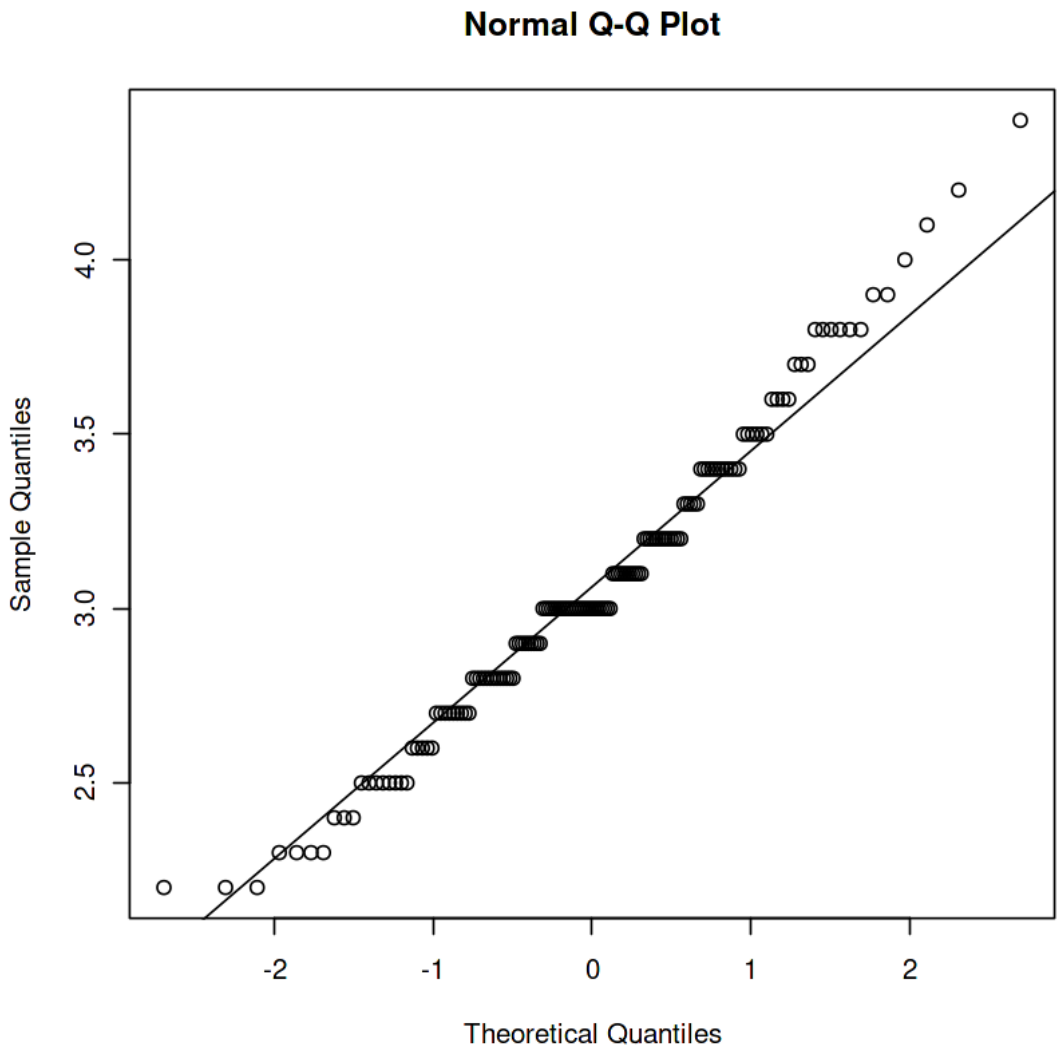
[6]:



sepal.length is negatively skewed

[7]:
```
qqnorm(sdata[,2])
qqline(sdata[,2])
```

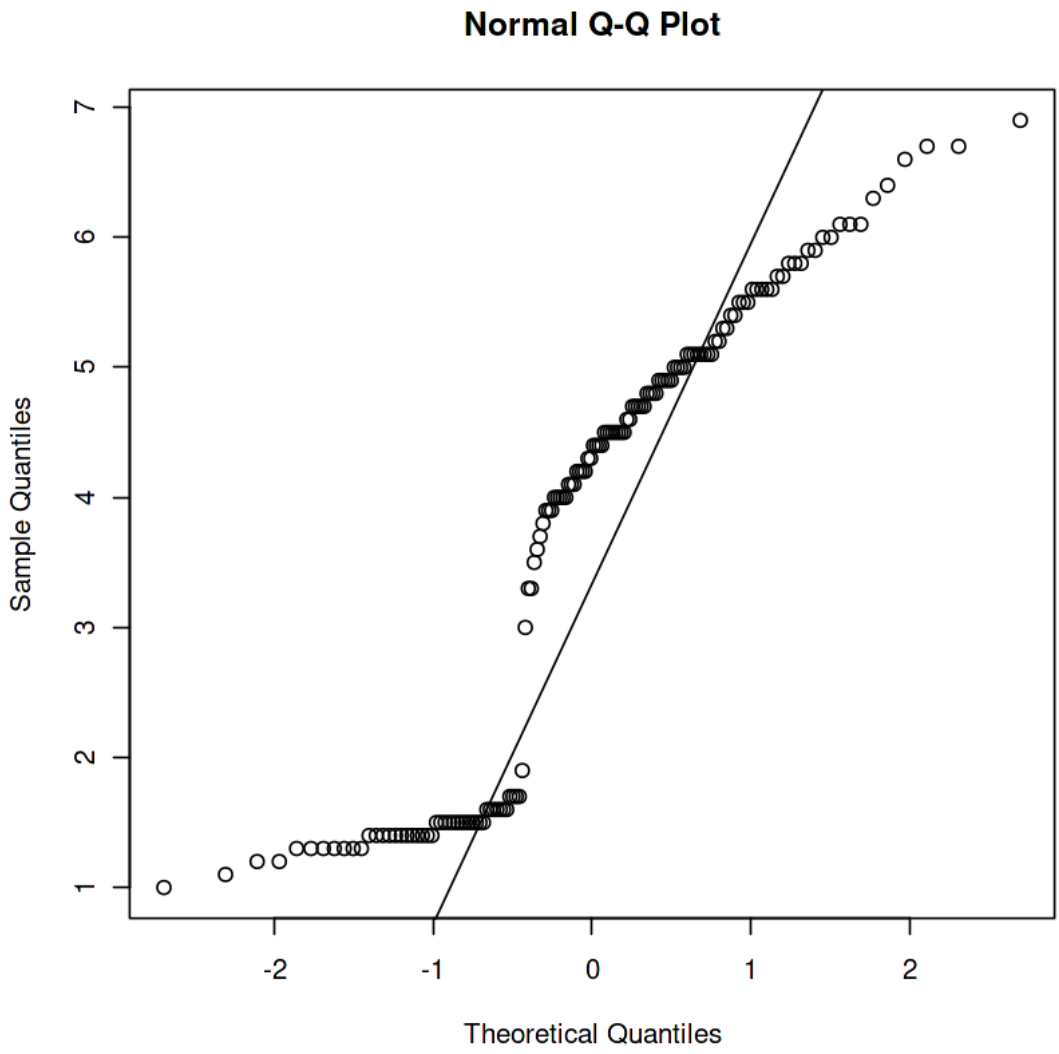
[7]:



sepal.width is positively skewed

[8]:
```
qqnorm(sdata[,3])
qqline(sdata[,3])
```

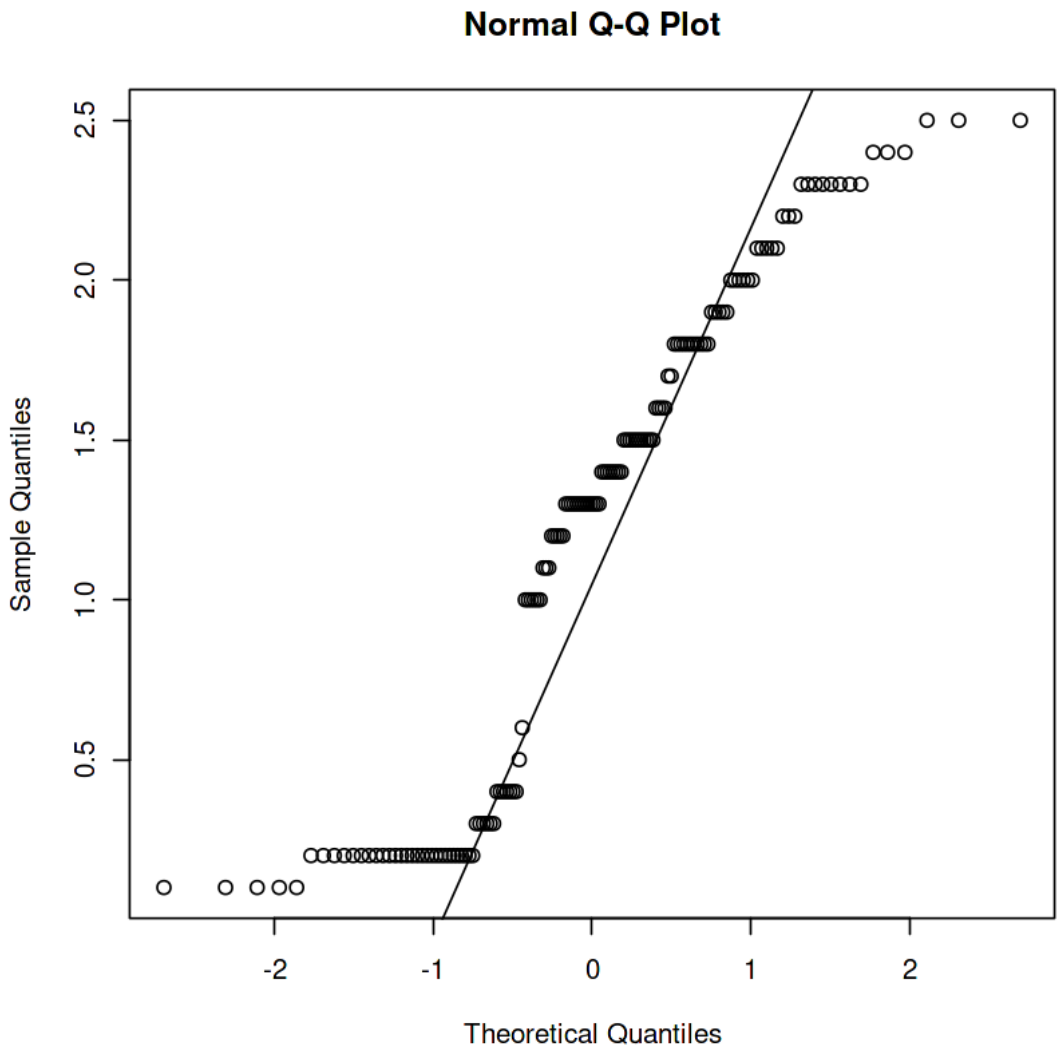
[8]:



petal legth is not normally distributed

[10]:
```
qqnorm(sdata[,4])
qqline(sdata[,4])
```

[10]:



petal width is not normally distributed

[11]:
```
#Data Partition with 75% training data
ind <- sample(2,nrow(sdata),
              replace=TRUE,
              prob=c(0.75,0.25))
```

[12]:
```
training <- sdata[ind==1,]
testing <- sdata[ind==2,]
```

[13]:
```
str(training)
```

```
'data.frame':   115 obs. of  5 variables:
 $ Sepal.Length: num  5.4 6.4 5.6 6.9 5.4 6.3 4.9 6.5 5 5.5 …
```

```
 $ Sepal.Width : num  3.7 3.1 3 3.2 3 2.5 3.1 3 3.6 3.5 …
 $ Petal.Length: num  1.5 5.5 4.5 5.7 4.5 5 1.5 5.5 1.4 1.3 …
 $ Petal.Width : num  0.2 1.8 1.5 2.3 1.5 1.9 0.1 1.8 0.2 0.2 …
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 3 2 3 2 3 1 3 1
1 …
```

[14]:
```
str(testing)
```

```
'data.frame':   29 obs. of  5 variables:
 $ Sepal.Length: num  7.2 6.3 6.2 6.7 5.1 5 6.2 5.7 5.7 5.8 …
 $ Sepal.Width : num  3 2.7 2.2 3.3 2.5 3 2.8 2.6 3.8 4 …
 $ Petal.Length: num  5.8 4.9 4.5 5.7 3 1.6 4.8 3.5 1.7 1.2 …
 $ Petal.Width : num  1.6 1.8 1.5 2.5 1.1 0.2 1.8 1 0.3 0.2 …
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 3 3 2 3 2 1 3 2 1
1 …
```

For training we have 115 observations and for testing we have 29 observations

[15]:
```
#Linear Discriminant Analysis
library(MASS)
linear <- lda(Species~.,training)
linear
```

[15]:
```
Call:
lda(Species ~ ., data = training)

Prior probabilities of groups:
    setosa versicolor  virginica
 0.3130435  0.3304348  0.3565217

Group means:
           Sepal.Length Sepal.Width Petal.Length Petal.Width
setosa         4.958333    3.422222     1.430556   0.2277778
versicolor     5.992105    2.818421     4.342105   1.3526316
virginica      6.600000    2.970732     5.553659   2.0365854

Coefficients of linear discriminants:
                   LD1       LD2
Sepal.Length  1.099628  0.287000
Sepal.Width   1.336225  1.637013
Petal.Length -2.411562 -1.435176
Petal.Width  -2.838229  3.551885

Proportion of trace:
   LD1    LD2
0.9931 0.0069
```

From the Prior probabilities of the groups we see that 31.30% belongs to setosa,33.04% belongs to versicolor and 35.65% belongs to virginica.Then there are group mean values.First Discriminant

Function(LD1) is the linear combination of the four variables.

1.099628* Sepal.Length+1.336225* Sepal.Width-2.411562* Petal.Length-2.838229* Petal.Width

Percentage separation achieved by the first discrimant function is 99.31% and by the 2nd discriminant function is 0.69%

[16]:
```
linear$counts
```

[16]:
setosa 36 **versicolor** 38 **virginica** 41

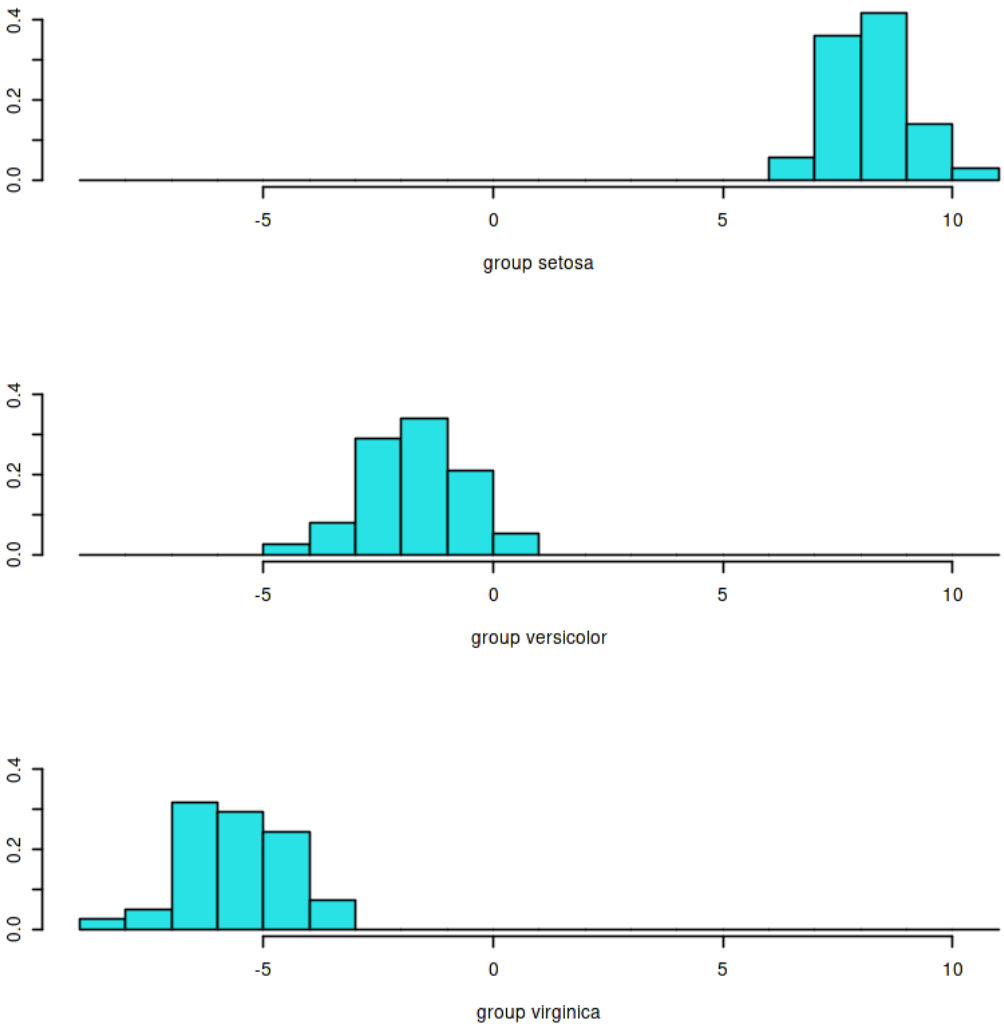
[17]:
```
attributes(linear)
```

[17]:
$names 1. 'prior' 2. 'counts' 3. 'means' 4. 'scaling' 5. 'lev' 6. 'svd' 7. 'N' 8. 'call' 9. 'terms' 10. 'xlevels'

$class 'lda'

[18]:
```
#Histogram
p <- predict(linear,training)
ldahist(data=p$x[,1], g=training$Species)
```

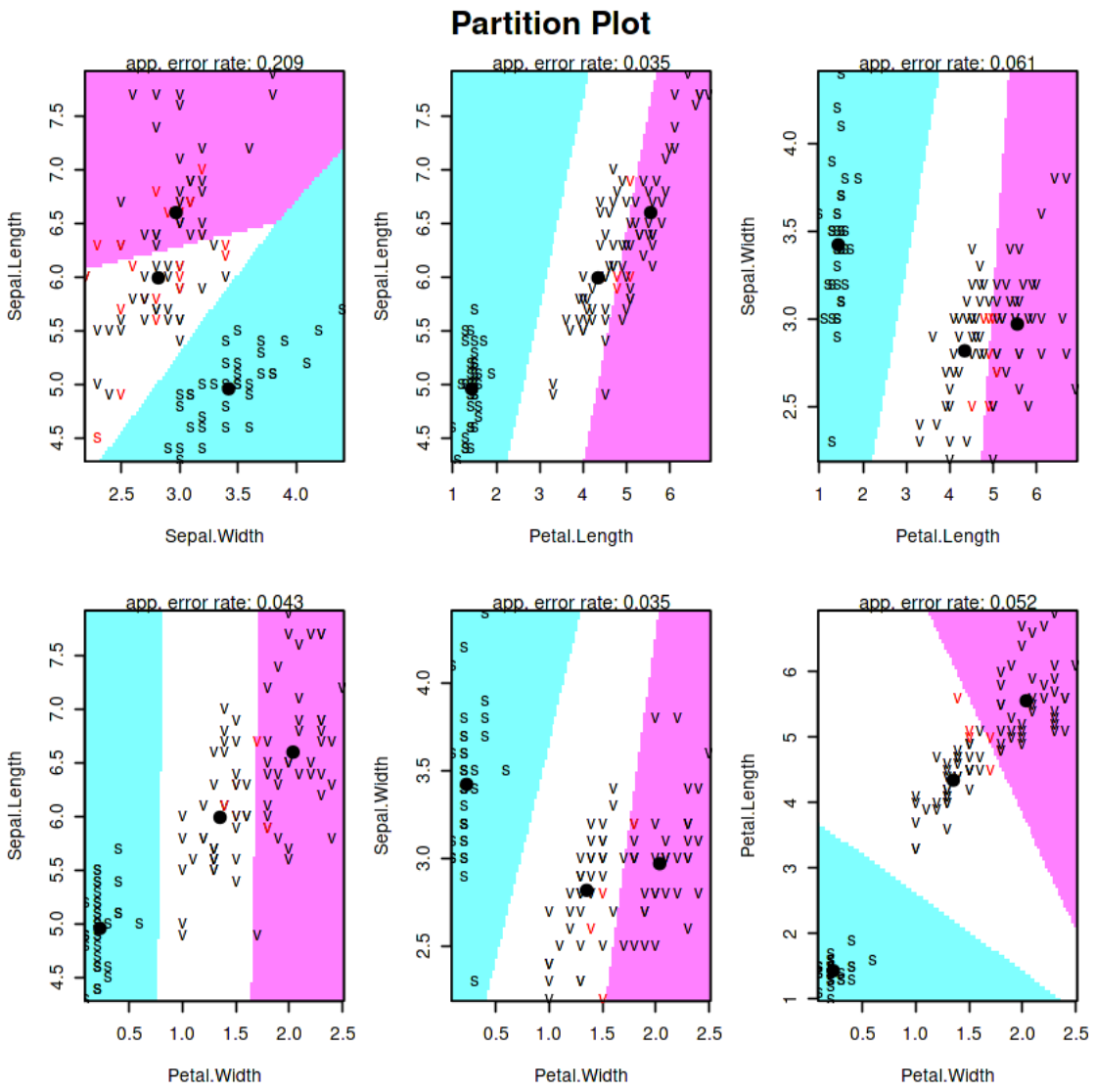
[18]:



These histograms are based on LD1. From this histograms we see that separation between 1st and 2nd species, separation between 1st and 3rd species are clear. There are some overlap between 2nd and 3rd species.

[19]:
```
# Partition Plot
library(klaR)
partimat(Species~.,data=training,method="lda")
```

[19]:



[20]:
```
# Confusion Matrix and accuracy- Training data
p1 <- predict(linear,training)$class
tab <- table(Predicted=p1, Actual=training$Species)
tab
```

[20]:
```
            Actual
Predicted    setosa versicolor virginica
  setosa         36          0         0
  versicolor      0         36         1
  virginica       0          2        40
```

setosa is 100% accurately classified, there is 2 misclassification in versicolor, 1 misclassification in virginica.

[21]:
```
accuracy <- sum(diag(tab))/sum(tab)
accuracy
```

[21]:
0.973913043478261

The accuracy is 97.39% on training data

[22]:
```
# Confusion Matrix and accuracy- Testing data
p2 <- predict(linear,testing)$class
tab1 <- table(Predicted=p2, Actual=testing$Species)
tab1
```

[22]:
```
            Actual
Predicted    setosa versicolor virginica
  setosa         12          0         0
  versicolor      0         10         0
  virginica       0          0         7
```

3 classes are 100% accurately classified.

[23]:
```
accuracy1 <- sum(diag(tab1))/sum(tab1)
accuracy1
```

[23]:
1

The accuracy on testing data is 100%

Quadratic Discriminant Analysis

[24]:
```
#Data Partition with 75% training data
set.seed(72007)
ind <- sample(2,nrow(sdata),
              replace=TRUE,
              prob=c(0.75,0.25))
```

[25]:
```
training <- iris[ind==1,]
testing <- iris[ind==2,]
```

[26]:
```
#Quadratic Discriminant Analysis
library(MASS)
quad <- qda(Species~.,training)
quad
```

[26]:
```
Call:
qda(Species ~ ., data = training)

Prior probabilities of groups:
    setosa versicolor  virginica
 0.3391304  0.3304348  0.3304348
```

```
Group means:
           Sepal.Length Sepal.Width Petal.Length Petal.Width
setosa         5.010256    3.438462     1.471795   0.2538462
versicolor     5.939474    2.757895     4.244737   1.3184211
virginica      6.647368    3.000000     5.600000   2.0368421
```

[28]:
```
attributes(quad)
```

[28]:
$names 1. 'prior' 2. 'counts' 3. 'means' 4. 'scaling' 5. 'ldet' 6. 'lev' 7. 'N' 8. 'call' 9. 'terms' 10. 'xlevels'

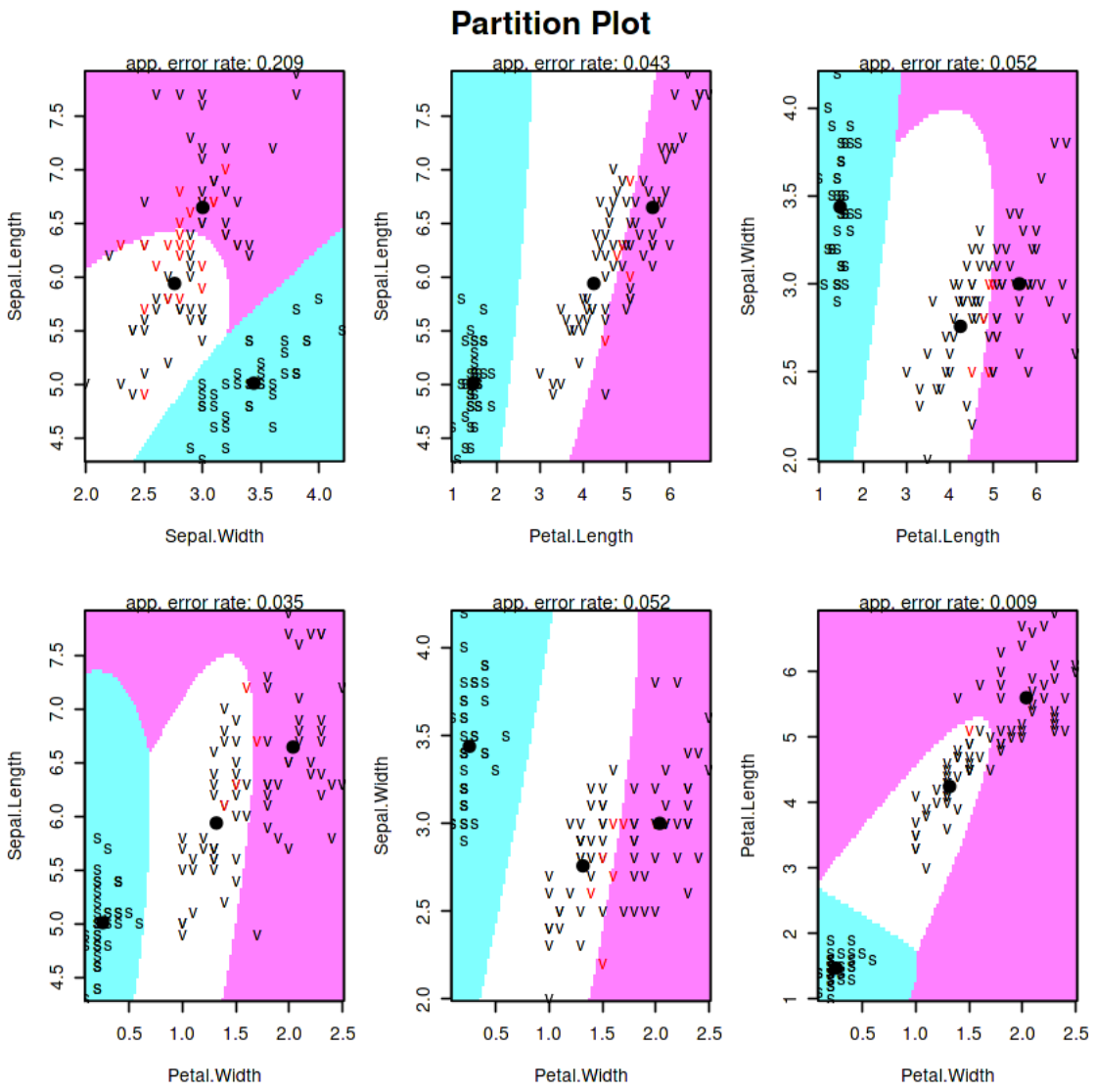
$class 'qda'

[29]:
```
quad$counts
```

[29]:
setosa 39 **versicolor** 38 **virginica** 38

[31]:
```
# Partition Plot
library(klaR)
partimat(Species~.,data=training,method="qda")
```

[31]:



[32]:
```
# Confusion Matrix and accuracy- Training data
p3 <- predict(quad,training)$class
tab3 <- table(Predicted=p3, Actual=training$Species)
tab3
```

[32]:
```
            Actual
Predicted    setosa versicolor virginica
  setosa         39          0         0
  versicolor      0         37         1
  virginica       0          1        37
```

[35]:
```
accuracy3 <- sum(diag(tab3))/sum(tab3)
accuracy3
```

[35]:
0.982608695652174

[33]:
```
# Confusion Matrix and accuracy- Testing data
p4 <- predict(quad,testing)$class
tab4 <- table(Predicted=p4, Actual=testing$Species)
tab4
```

[33]:
```
            Actual
Predicted    setosa versicolor virginica
  setosa         11          0         0
  versicolor      0         11         0
  virginica       0          1        12
```

[34]:
```
accuracy4 <- sum(diag(tab4))/sum(tab4)
accuracy4
```

[34]:
0.971428571428571