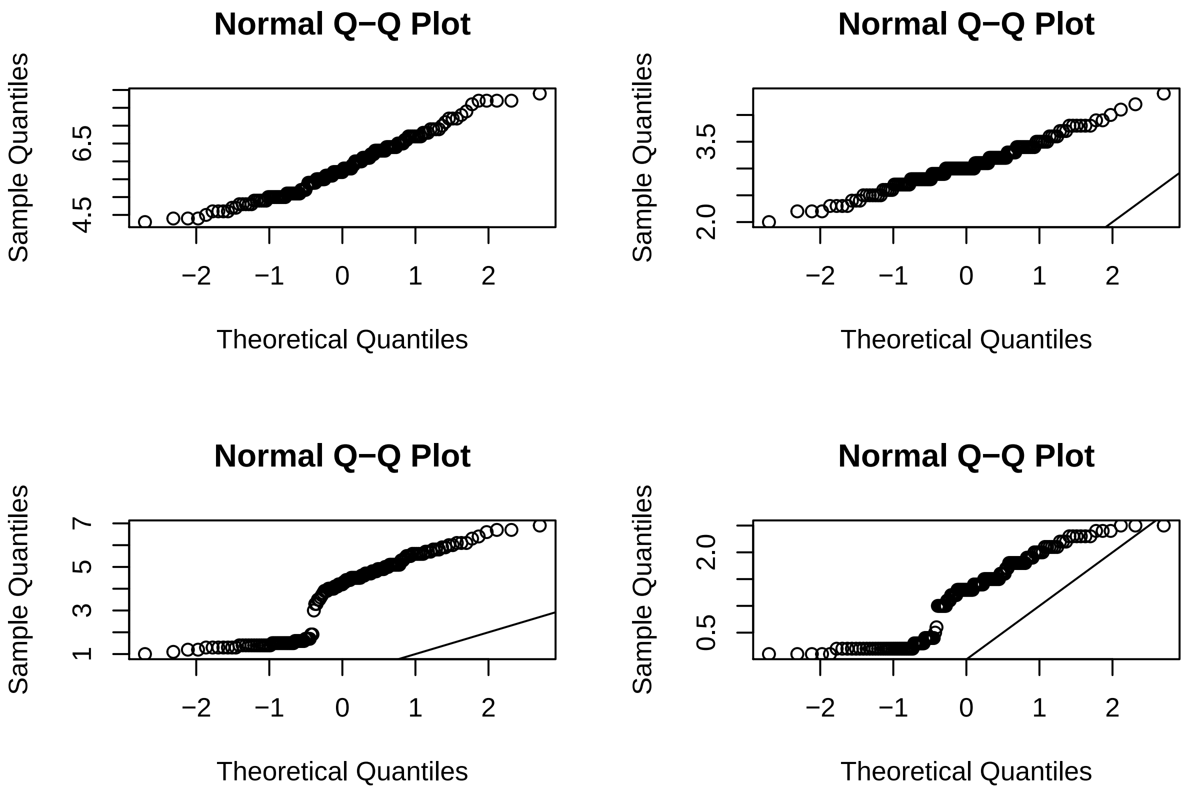
M445 HW 6

Erick Castillo

3/26/2021

Problem 2: Recall that the LDA method requires that the data is normally distributed. This can be checked with QQ-plots.



The data does not appear to align with the quantiles of the normal distribution. Therefore, the data must be transformed. This is done in the following block of code. The first line of code simply standardizes the data to be more inline with the quantiles of $N(0, 1)$ by taking the difference of a value by the mean of the row (using the center method), then dividing by the standard deviation of the row (using the scale method). LDA is also implemented and used to make predictions:

```
#preprocessing method must be declared to transform the data.
preproc.parameter <- train %>%
  preProcess(method = c("center", "scale"))

#here data is transformed from multivariate to univariate observations.
```

```
traintrans <- preproc.parameter %>% predict(train)
testtrans <- preproc.parameter %>% predict(test)

#the model is then declared. predictions are derived from such a model.
lda.mod = lda((factor(Species)~.), data = traintrans)
lda.pred <- lda.mod %>% predict(testtrans)
lda.pred
```

```
## $class
## [1] virginica virginica virginica virginica virginica
## Levels: setosa versicolor virginica
##
## $posterior
##           setosa   versicolor virginica
## 146 4.009305e-38 1.233670e-04 0.9998766
## 147 2.995978e-35 8.502040e-03 0.9914980
## 148 2.968680e-34 4.272486e-03 0.9957275
## 149 1.158559e-39 1.856019e-05 0.9999814
## 150 1.307814e-32 2.140862e-02 0.9785914
##
## $x
##           LD1        LD2
## 146 -5.740301  1.6490850
## 147 -5.293216 -0.3719992
## 148 -5.091086  0.8550232
## 149 -5.987756  2.3959450
## 150 -4.816665  0.4137047
```

```
data.frame(orig_data = test$Species, pred_vals= lda.pred$class)
```

```
##   orig_data pred_vals
## 1 virginica virginica
## 2 virginica virginica
## 3 virginica virginica
## 4 virginica virginica
## 5 virginica virginica
```

Notice that the predictions made on the test set of the last five variables show that there is a very high chance that these values are virginicas. This is most certainly the case, so this model worked very well.

The following code uses the QDA method of classification utilizing the same train and test split from the code above. This method does not require the preprocessing method used before.

```
qda.mod = qda((Species~.), data = train)
qda.pred <- predict(qda.mod, test)
qda.pred
```

```
## $class
## [1] virginica virginica virginica virginica virginica
## Levels: setosa versicolor virginica
##
```

```
## $posterior
##            setosa   versicolor virginica
## 146 2.072178e-150 4.998132e-09 1.0000000
## 147 2.692510e-124 2.486478e-04 0.9997514
## 148 1.002502e-133 1.339368e-03 0.9986606
## 149 2.358170e-155 1.523325e-06 0.9999985
## 150 8.519591e-119 7.250993e-02 0.9274901
```

```
data.frame(orig_data = test$Species, pred_vals= qda.pred$class)
```

```
##   orig_data pred_vals
## 1 virginica virginica
## 2 virginica virginica
## 3 virginica virginica
## 4 virginica virginica
## 5 virginica virginica
```

This method was able to classify all the test values accurately. Do note that the percentages in this table are slightly smaller than the percentages seen in the LDA method.