MVA Project - LDA

Chun-Jung Chen & Akshay Arora

Question 1: Deos Family conditions affect students' final grade in Math?

To fit the Linear Discriminant Analysis, I transformed depedent variables from continuous to factor variable.

Convert Final math and subsetting training and testing dataset

The dependent variable is the final math scores, I split it out into four groups include first, second, third, and fourth quantiles.

```
(q = quantile(d3.q1$FinalMath))
```

```
##   0%  25%  50%  75% 100%
##    0    8   11   14   20
```

Separating the original dataset to training and testing by 75/25 rules

```
set.seed(20201109)
d3.q1.train = d3.q1[sample(nrow(d3.q1), nrow(d3.q1) * 0.75),
    ]
d3.q1.test = d3.q1[-sample(nrow(d3.q1), nrow(d3.q1) * 0.75),
    ]
```

Linear Discriminant Analysis

According to the Linear Discriminant Analysis report, the grop mean in each quantile group does not contain large difference. The scattor plot can illustrate this phenomenon.

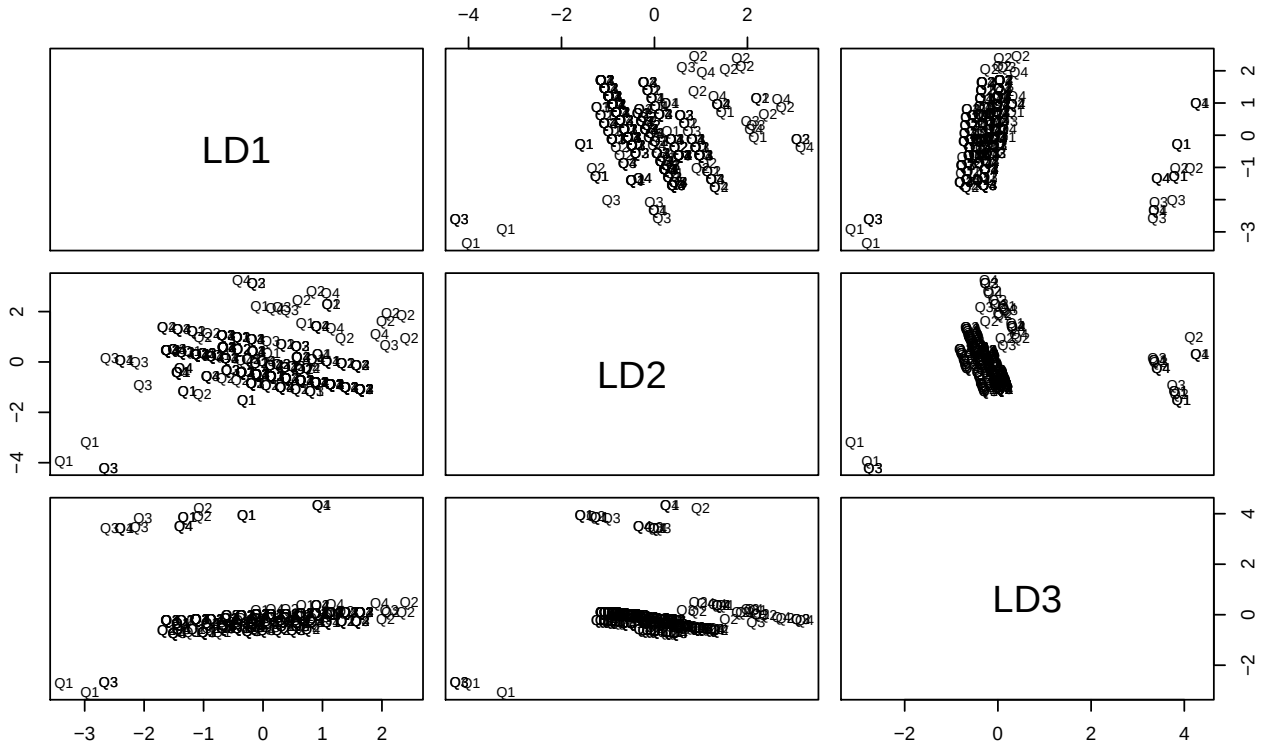
```
d3.q1.lda = lda(formula = d3.q1.train$FinalMath ~ ., data = d3.q1.train)
d3.q1.lda
```

```
## Call:
## lda(d3.q1.train$FinalMath ~ ., data = d3.q1.train)
##
## Prior probabilities of groups:
##        Q1        Q2        Q3        Q4
## 0.2635379 0.3068592 0.2274368 0.2021661
##
## Group means:
##     famsize  Pstatus0    famrel2   famrel3   famrel4   famrel5   famsup1
## Q1 5.342466 0.9452055 0.08219178 0.1780822 0.4657534 0.2465753 0.6849315
## Q2 6.023529 0.8705882 0.02352941 0.1647059 0.5764706 0.2352941 0.6470588
## Q3 4.714286 0.9365079 0.04761905 0.1904762 0.3968254 0.3333333 0.6031746
## Q4 4.767857 0.8571429 0.07142857 0.1607143 0.4107143 0.3571429 0.5535714
##
## Coefficients of linear discriminants:
##                  LD1         LD2         LD3
## famsize   0.24941351 -0.08317644  0.01617145
## Pstatus0 -1.25748843 -1.83899016 -0.39822249
## famrel2   1.57779424  2.93643231  6.56783398
## famrel3   2.71693904  3.28834929  2.42648496
## famrel4   3.55161186  3.38167124  2.74908117
## famrel5   2.53914981  4.16829620  2.55361469
## famsup1   0.06210534 -0.92110198  0.37750095
##
## Proportion of trace:
##    LD1    LD2    LD3
## 0.6650 0.2755 0.0595
```

Plots for Linear Discriminant variables

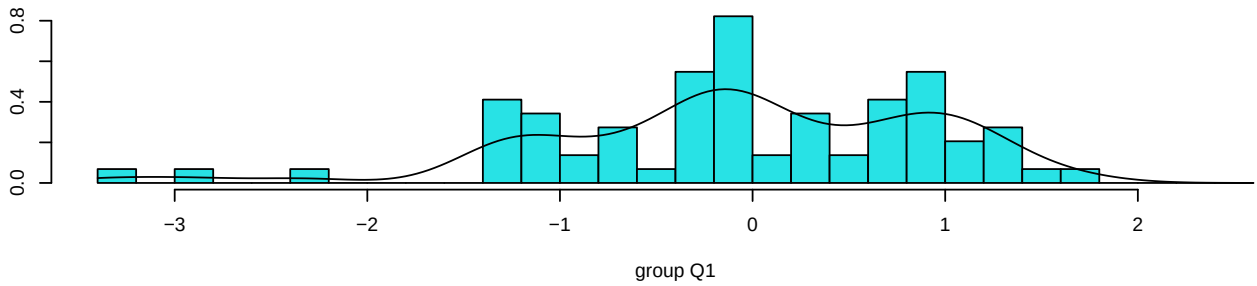
The LD variables do not clearly split each quantile group

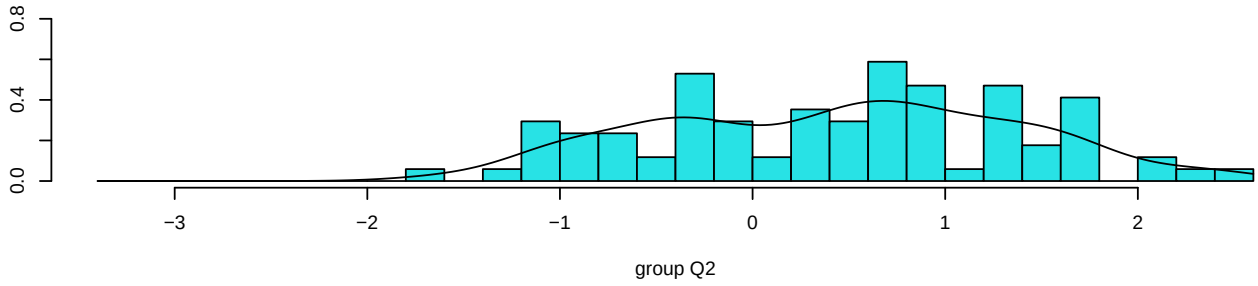
```
plot(d3.q1.lda)
```

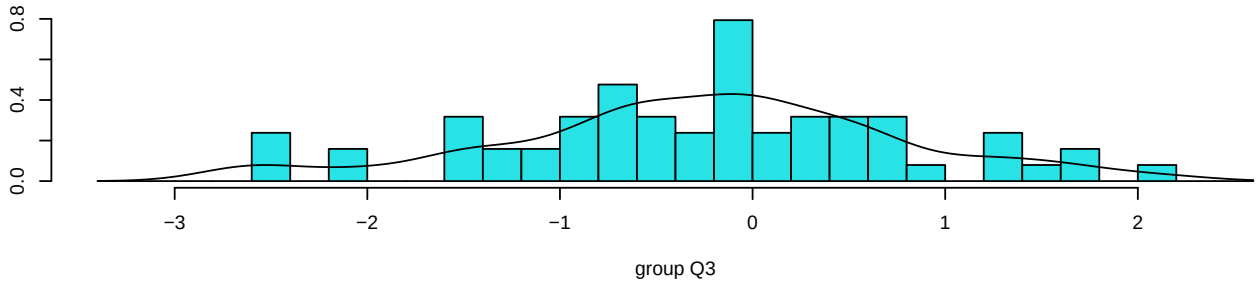


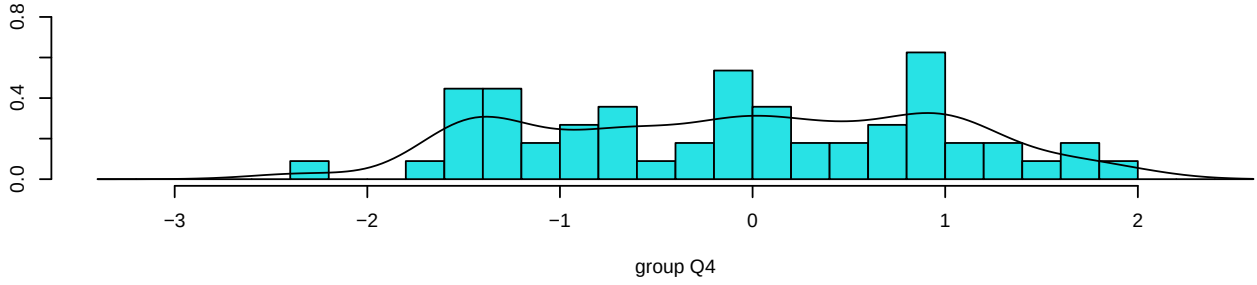
The plot shows that the center of each group still located around 0, there is no clear split for each group to obtan.

```
plot(d3.q1.lda, dimen = 1, type = "b")
```









Testing Accuracy

As a result of the table, the LDA does not offer an accurate prediction for the final math score separated by quantile. However, LDA on Q2 has the best prediction since the majority of the Q2 are correctly forecasted.

```
lda.train.finalmath = predict(d3.q1.lda)
d3.q1.train$lda = lda.train.finalmath$class
table(d3.q1.train$lda, d3.q1.train$FinalMath)
```

```
##
##      Q1 Q2 Q3 Q4
##   Q1 15 11 14  8
##   Q2 36 54 24 29
##   Q3 18 16 21 17
##   Q4  4  4  4  2
```

Same results occur in the testing subset, LDA has the most accurate prediction records on Q2.

```
lda.test.finalmath = predict(d3.q1.lda, d3.q1.test)
d3.q1.test$lda = lda.test.finalmath$class
table(d3.q1.test$lda, d3.q1.test$FinalMath)
```

```
##
##      Q1 Q2 Q3 Q4
##   Q1  4  5  6  5
##   Q2 10 20  8  8
##   Q3  5  9  4  5
##   Q4  0  1  2  1
```

Question 2: Does parents' jobs and education level influence students' first period of grade in Math?

To fit the Linear Discriminant Analysis, I transformed depedent variables from continuous to factor variable.

Convert First math and subsetting training and testing dataset

The dependent variable is the first math scores, I split it out into four groups include first, second, third, and fourth quantiles.

```
(q = quantile(d3.q2$FirstMath))
```

```
##   0%  25%  50%  75% 100%
##    3    8   11   13   19
```

Separating the original dataset to training and testing by 75/25 rules

```
set.seed(20201109)
d3.q2.train = d3.q2[sample(nrow(d3.q2), nrow(d3.q2) * 0.75),
    ]
d3.q2.test = d3.q2[-sample(nrow(d3.q2), nrow(d3.q2) * 0.75),
    ]
```

Linear Discriminant Analysis

According to the Linear Discriminant Analysis report, we should consider these three LD since they present almost the same proportion of the trace.

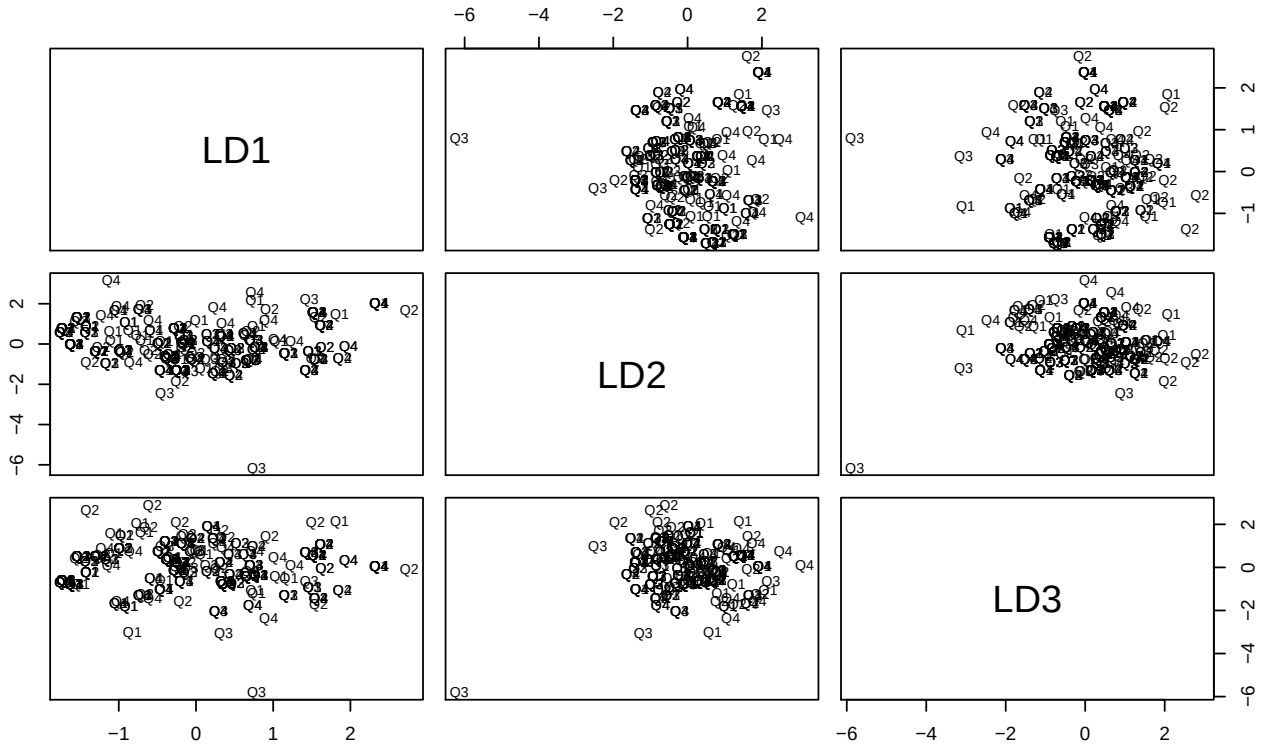
```
d3.q2.lda = lda(formula = d3.q2.train$FirstMath ~ ., data = d3.q2.train)
d3.q2.lda
```

```
## Call:
## lda(d3.q2.train$FirstMath ~ ., data = d3.q2.train)
##
## Prior probabilities of groups:
##        Q1        Q2        Q3        Q4
## 0.2671480 0.3104693 0.1696751 0.2527076
##
## Group means:
##         Medu1     Medu2     Medu3     Medu4     Fedu1     Fedu2     Fedu3
## Q1 0.20270270 0.2702703 0.3243243 0.1891892 0.3648649 0.2162162 0.2837838
## Q2 0.11627907 0.2674419 0.2441860 0.3720930 0.1279070 0.3255814 0.3023256
## Q3 0.08510638 0.2978723 0.2553191 0.3404255 0.1489362 0.3191489 0.1914894
## Q4 0.10000000 0.1857143 0.2285714 0.4714286 0.1428571 0.2857143 0.2000000
##         Fedu4      Mjob2     Mjob3     Mjob4     Mjob5      Fjob2     Fjob3
## Q1 0.1351351 0.04054054 0.4459459 0.2432432 0.1081081 0.02702703 0.6351351
## Q2 0.2441860 0.09302326 0.4069767 0.1744186 0.1627907 0.09302326 0.5348837
## Q3 0.3191489 0.12765957 0.4255319 0.2553191 0.1276596 0.06382979 0.6382979
## Q4 0.3714286 0.14285714 0.2285714 0.3428571 0.1714286 0.04285714 0.4285714
##        Fjob4      Fjob5
## Q1 0.2567568 0.04054054
## Q2 0.2674419 0.08139535
## Q3 0.2553191 0.02127660
## Q4 0.2857143 0.18571429
##
## Coefficients of linear discriminants:
##                LD1         LD2        LD3
## Medu1 -0.66487483  0.68108495  3.5139037
## Medu2 -0.55808449 -0.10776372  3.3120853
## Medu3 -0.69092867  0.48494406  3.4300140
## Medu4  0.21322534  0.24495178  4.1616404
## Fedu1 -2.47107185  6.93600230  5.2010306
## Fedu2 -1.09511120  5.63067360  6.2541053
## Fedu3 -2.01315954  5.96676537  6.3901024
## Fedu4 -1.17615366  5.04341813  4.8525642
## Mjob2  0.90276715  0.21889349 -2.2290746
## Mjob3 -0.19601016 -0.56058444 -1.1034386
## Mjob4  0.51277898  0.51990442 -2.0958013
## Mjob5 -0.28997539  0.07688812 -1.5811940
## Fjob2 -0.51185616 -2.32477440  2.5484886
## Fjob3 -0.38994825 -1.79428891  0.4200917
## Fjob4 -0.08088151 -1.71775513  0.7860098
## Fjob5  0.78394706  0.66601722  1.7820150
##
## Proportion of trace:
##    LD1    LD2    LD3
## 0.4460 0.3009 0.2531
```

Plots for Linear Discriminant variables

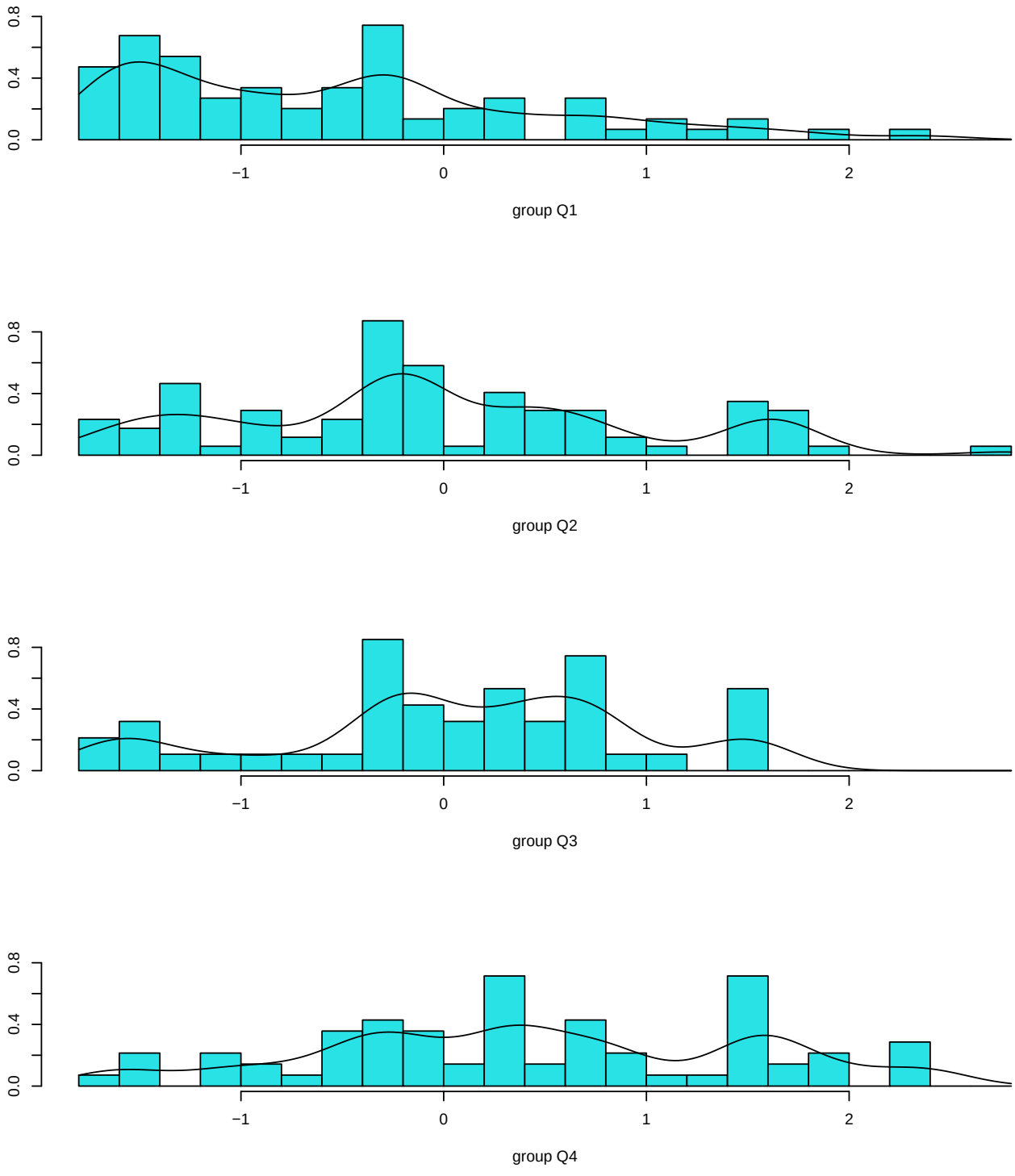
The LD variables do not clearly split each quantile group

```
plot(d3.q2.lda)
```



The plot shows that the center of each group still located around 0, there is no clear split for each group to obtan.

```
plot(d3.q2.lda, dimen = 1, type = "b")
```



Testing Accuracy

As a result of the table, the LDA contains an accurate prediction for the final math score separated by quantile as the elements at the diagonal of the table are the largest number in their columns or rows.

```
lda.train.firstmath = predict(d3.q2.lda)
d3.q2.train$lda = lda.train.firstmath$class
table(d3.q2.train$lda, d3.q2.train$FirstMath)
```

```
##
##      Q1 Q2 Q3 Q4
##   Q1 36 20 10 13
##   Q2 23 43 19 17
##   Q3  1  4  7 10
##   Q4 14 19 11 30
```

For the application of the test dataset, the result is similar but not as good as the trainning dataset.

```
lda.test.firstmath = predict(d3.q2.lda, d3.q2.test)
d3.q2.test$lda = lda.test.firstmath$class
table(d3.q2.test$lda, d3.q2.test$FirstMath)
```

```
##
##      Q1 Q2 Q3 Q4
##   Q1 12  4  5  6
##   Q2  9 16  4  7
##   Q3  1  1  4  3
##   Q4  4  6  3  8
```

Question 3: Does student's learning conditions really impact students' final grade math score and Portuguese scores in average?

To fit the Linear Discriminant Analysis, I transformed depedent variables from continuous to factor variable.

Convert final average math score and subsetting training and testing dataset

The dependent variable is the final average math scores, I split it out into four groups include first, second, third, and fourth quantiles.

```
(q = quantile(d3.q3$FinalAvg))
```

```
##   0%  25%  50%  75% 100%
##  0.0  9.5 11.5 13.5 18.5
```

Separating the original dataset to training and testing by 75/25 rules

```
set.seed(20201109)
d3.q3.train = d3.q3[sample(nrow(d3.q3), nrow(d3.q3) * 0.75),
    ]
d3.q3.test = d3.q3[-sample(nrow(d3.q3), nrow(d3.q3) * 0.75),
    ]
```

Linear Discriminant Analysis

According to the Linear Discriminant Analysis report, the LD3 can be eliminated since LD1 and LD2 have already entailed over 95% of data

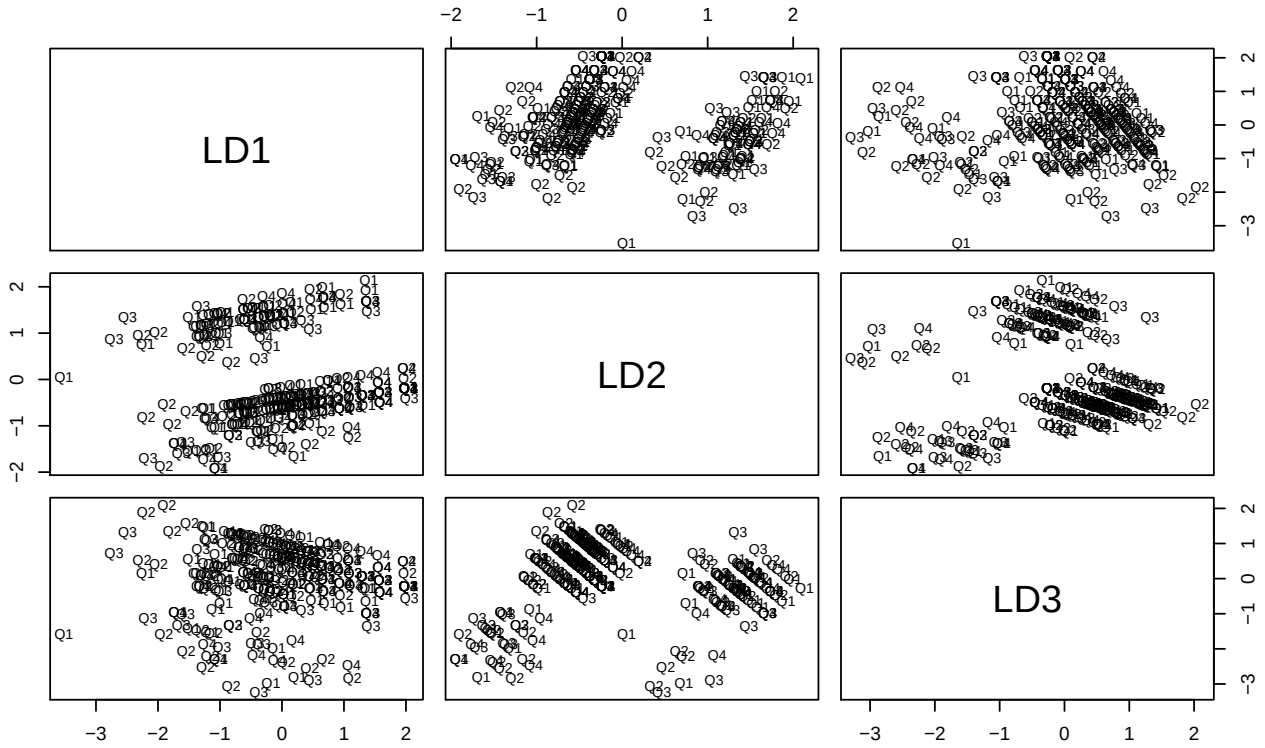
```
d3.q3.lda = lda(formula = d3.q3.train$FinalAvg ~ ., data = d3.q3.train)
d3.q3.lda
```

```
## Call:
## lda(d3.q3.train$FinalAvg ~ ., data = d3.q3.train)
##
## Prior probabilities of groups:
##        Q1        Q2        Q3        Q4
## 0.2490975 0.2635379 0.2166065 0.2707581
##
## Group means:
##    internet1 romantic1 freetime normtraveltime
## Q1 0.8695652 0.4202899 3.260870       1.720526
## Q2 0.8219178 0.2876712 3.178082       1.762760
## Q3 0.8500000 0.3000000 3.183333       1.641179
## Q4 0.8933333 0.2800000 3.226667       1.601564
##
## Coefficients of linear discriminants:
##                        LD1        LD2        LD3
## internet1       0.9168843  0.8416167  2.2612413
## romantic1      -0.6138136  1.8952246 -0.7814600
## freetime       -0.0118888  0.2166360  0.3525651
## normtraveltime -2.5106062 -0.5029440  1.0340816
##
## Proportion of trace:
##    LD1    LD2    LD3
## 0.7133 0.2680 0.0188
```

Plots for Linear Discriminant variables

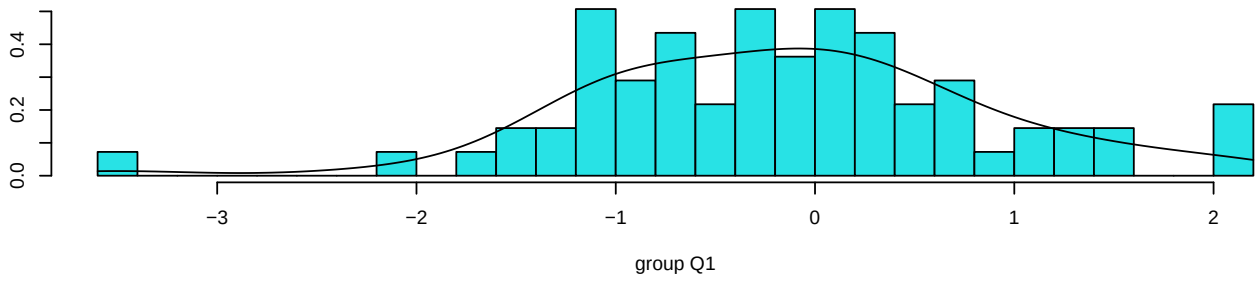
The LD variables do not clearly split each quantile group

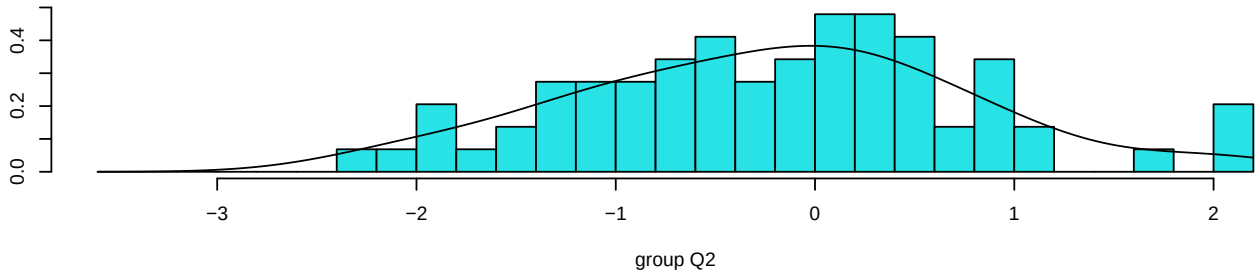
```
plot(d3.q3.lda)
```

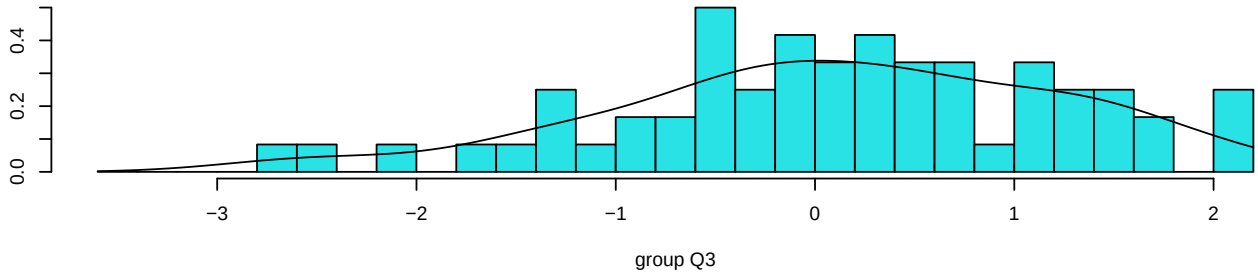


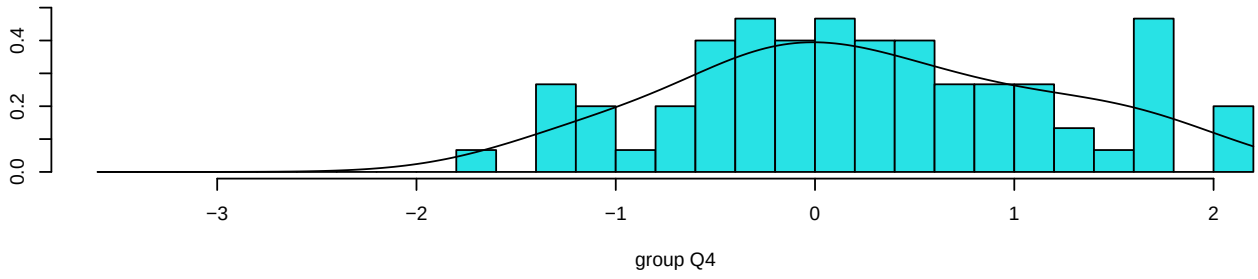
The plot shows that the center of each group still located around 0, there is no clear split for each group to obtan.

```
plot(d3.q3.lda, dimen = 1, type = "b")
```









Testing Accuracy

As a result of the table, the LDA does not offer an accurate prediction for the final math score separated by quantile since the diagonal elements are not the significant largest number of their columns elements.

```
lda.train.FinalAvg = predict(d3.q3.lda)
d3.q3.train$lda = lda.train.FinalAvg$class
table(d3.q3.train$lda, d3.q3.train$FinalAvg)
```

```
##
##      Q1 Q2 Q3 Q4
##   Q1 24 20 14 18
##   Q2 22 29 17 21
##   Q3  0  0  0  0
##   Q4 23 24 29 36
```

Same results occur in the testing subset, LDA does not effectively work for question 3

```
lda.test.FinalAvg = predict(d3.q3.lda, d3.q3.test)
d3.q3.test$lda = lda.test.FinalAvg$class
table(d3.q3.test$lda, d3.q3.test$FinalAvg)
```

```
##
##      Q1 Q2 Q3 Q4
##   Q1  8  4  3  8
##   Q2  8 11  6  5
##   Q3  0  0  0  0
##   Q4  4 11 14 11
```