MAT 443: HW 6

Name: Eric Agyemang

**Question 1**

To prove: Let be (4.2) and be (4.3) are equal to each other, we can get this by using property of ratio and proportions. Subtract numerator from denominator on both sides of (4.2) and (4.3)

Divide both sides

Which is equivalent to

**Question 5**

(a) If the Bayes decision boundary is linear, we expect **QDA** to perform better on the training set due to its flexibility and **LDA** on the test set.

(b) If the Bayes decision boundary is non-linear, we expect **QDA** to perform better on the training set as well as on the test set.

(c) In general, QDA fits better when the sample size n increases as the variance is no longer a concern. It is a concern for QDA when sample size is small and variance for different classes is not same

(d)FALSE. Because with high variance for more flexibility approach of **QDA** we may over fit the data for small number of observations and get inferior test error rate

**Question 10**

[1] 1089 9

> head(Weekly)

Year Lag1 Lag2 Lag3 Lag4 Lag5

1 1990 0.816 1.572 -3.936 -0.229 -3.484

2 1990 -0.270 0.816 1.572 -3.936 -0.229

3 1990 -2.576 -0.270 0.816 1.572 -3.936

4 1990 3.514 -2.576 -0.270 0.816 1.572

5 1990 0.712 3.514 -2.576 -0.270 0.816

6 1990 1.178 0.712 3.514 -2.576 -0.270

Volume Today Direction

1 0.1549760 -0.270 Down

2 0.1485740 -2.576 Down

3 0.1598375 3.514 Up

4 0.1616300 0.712 Up

5 0.1537280 1.178 Up

6 0.1544440 -1.372 Down

> str(Weekly)

'data.frame': 1089 obs. of 9 variables:

$ Year : num 1990 1990 1990 1990 1990 1990 1990 1990 1990 1990 ...

$ Lag1 : num 0.816 -0.27 -2.576 3.514 0.712 ...

$ Lag2 : num 1.572 0.816 -0.27 -2.576 3.514 ...

$ Lag3 : num -3.936 1.572 0.816 -0.27 -2.576 ...

$ Lag4 : num -0.229 -3.936 1.572 0.816 -0.27 ...

$ Lag5 : num -3.484 -0.229 -3.936 1.572 0.816 ...

$ Volume : num 0.155 0.149 0.16 0.162 0.154 ...

$ Today : num -0.27 -2.576 3.514 0.712 1.178 ...

$ Direction: Factor w/ 2 levels "Down","Up": 1 1 2 2 2 1 2 2 2 1 ...

> summary(Weekly)

Year Lag1

Min. :1990 Min. :-18.1950

1st Qu.:1995 1st Qu.: -1.1540

Median :2000 Median : 0.2410

Mean :2000 Mean : 0.1506

3rd Qu.:2005 3rd Qu.: 1.4050

Max. :2010 Max. : 12.0260

Lag2 Lag3

Min. :-18.1950 Min. :-18.1950

1st Qu.: -1.1540 1st Qu.: -1.1580

Median : 0.2410 Median : 0.2410

Mean : 0.1511 Mean : 0.1472

3rd Qu.: 1.4090 3rd Qu.: 1.4090

Max. : 12.0260 Max. : 12.0260

Lag4 Lag5

Min. :-18.1950 Min. :-18.1950

1st Qu.: -1.1580 1st Qu.: -1.1660

Median : 0.2380 Median : 0.2340

Mean : 0.1458 Mean : 0.1399

3rd Qu.: 1.4090 3rd Qu.: 1.4050

Max. : 12.0260 Max. : 12.0260

Volume Today

Min. :0.08747 Min. :-18.1950

1st Qu.:0.33202 1st Qu.: -1.1540

Median :1.00268 Median : 0.2410

Mean :1.57462 Mean : 0.1499

3rd Qu.:2.05373 3rd Qu.: 1.4050

Max. :9.32821 Max. : 12.0260

Direction

Down:484

Up :605

> correlation<-cor(Weekly[-9])

> correlation

Year Lag1 Lag2

Year 1.00000000 -0.032289274 -0.03339001

Lag1 -0.03228927 1.000000000 -0.07485305

Lag2 -0.03339001 -0.074853051 1.00000000

Lag3 -0.03000649 0.058635682 -0.07572091

Lag4 -0.03112792 -0.071273876 0.05838153

Lag5 -0.03051910 -0.008183096 -0.07249948

Volume 0.84194162 -0.064951313 -0.08551314

Today -0.03245989 -0.075031842 0.05916672

Lag3 Lag4 Lag5

Year -0.03000649 -0.031127923 -0.030519101

Lag1 0.05863568 -0.071273876 -0.008183096

Lag2 -0.07572091 0.058381535 -0.072499482

Lag3 1.00000000 -0.075395865 0.060657175

Lag4 -0.07539587 1.000000000 -0.075675027

Lag5 0.06065717 -0.075675027 1.000000000

Volume -0.06928771 -0.061074617 -0.058517414

Today -0.07124364 -0.007825873 0.011012698

Volume Today

Year 0.84194162 -0.032459894

Lag1 -0.06495131 -0.075031842

Lag2 -0.08551314 0.059166717

Lag3 -0.06928771 -0.071243639

Lag4 -0.06107462 -0.007825873

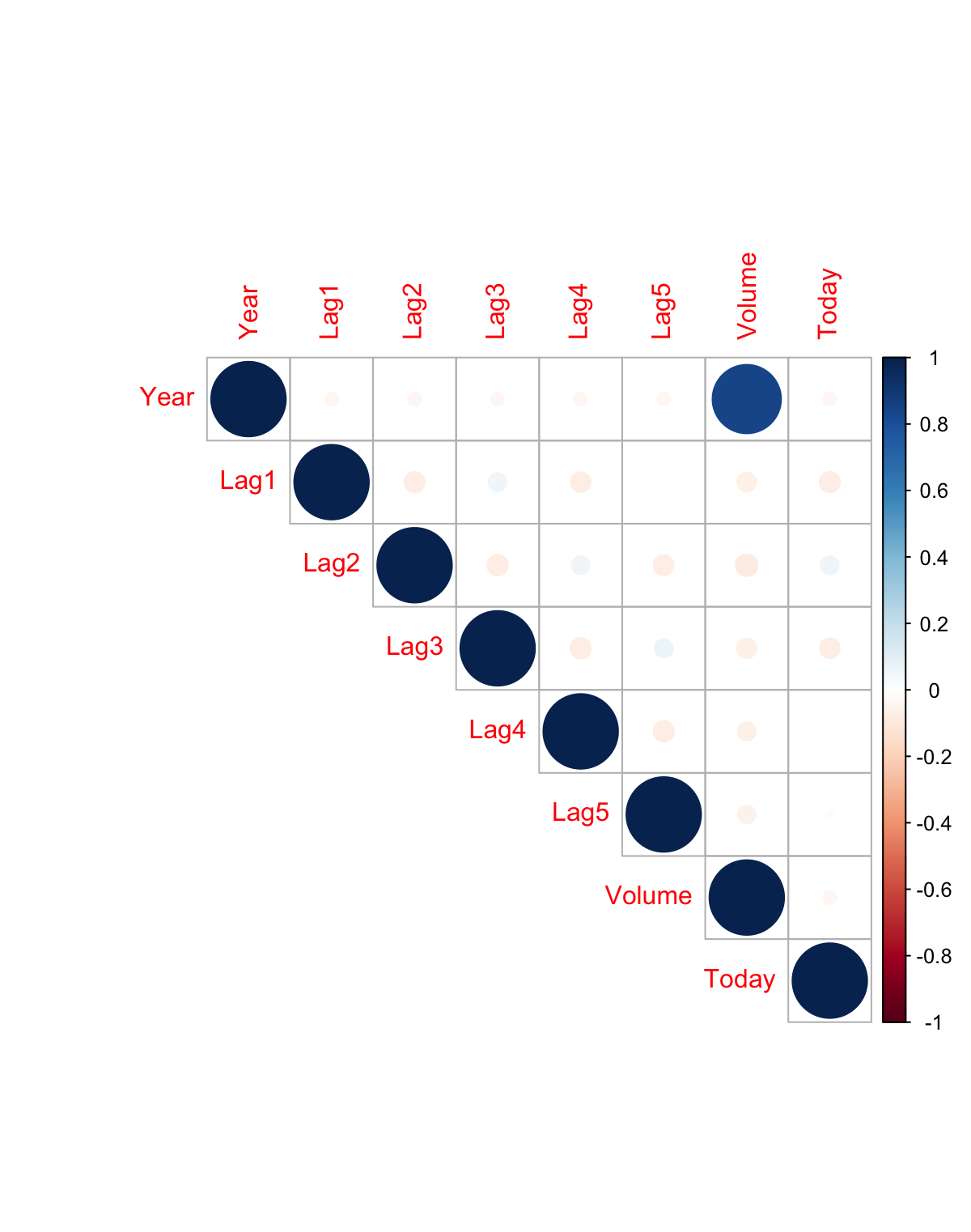
Lag5 -0.05851741 0.011012698

Volume 1.00000000 -0.033077783

Today -0.03307778 1.000000000

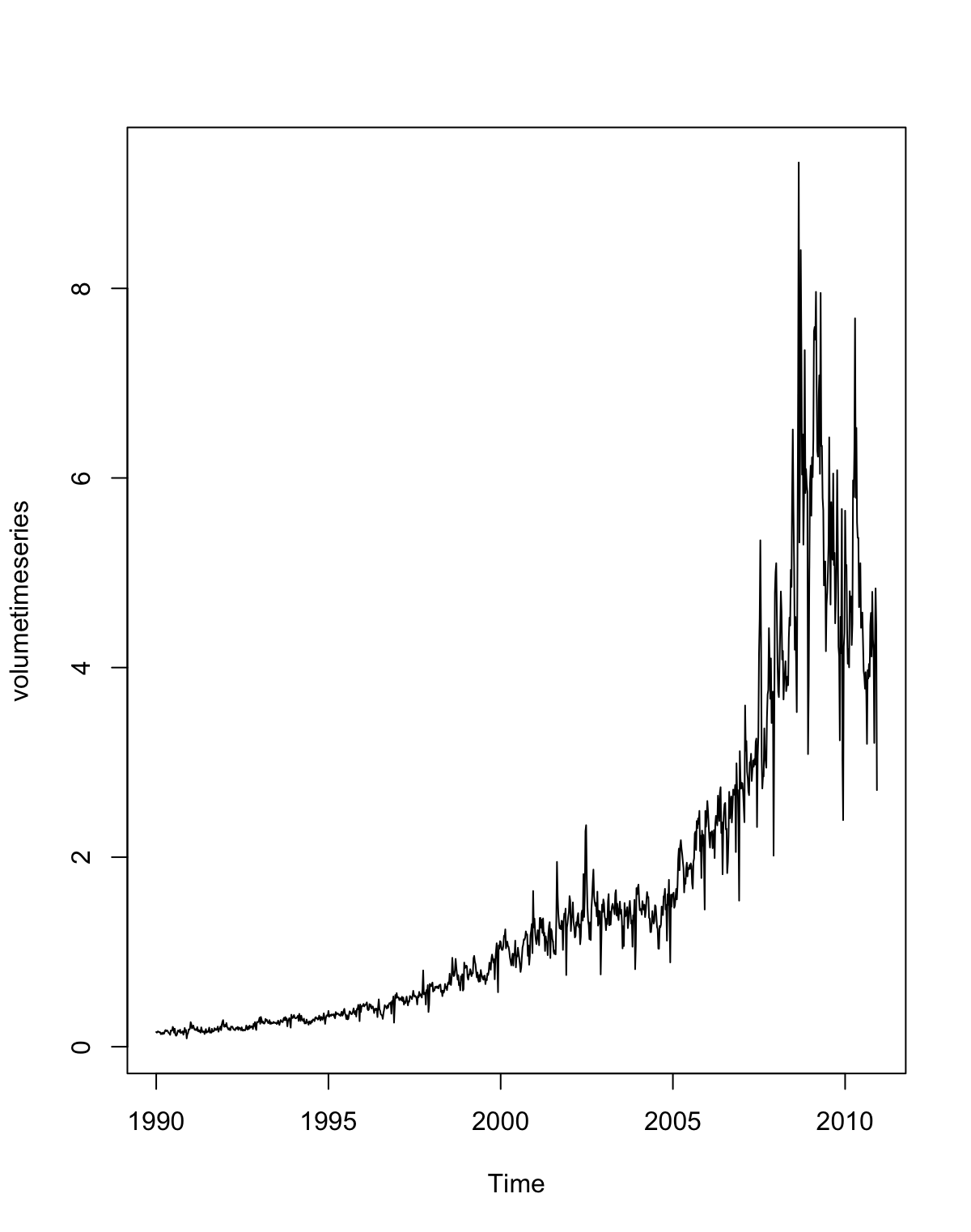
>library(ggplot2)

> library(corrplot)



> volumetimeseries <- ts(Volume, frequency=52, start=c(1990,1))

> plot.ts(volumetimeseries)



The correlation between the Lag Variables and Today’s returns are close to zero. This can be noticed as volume is increasing over the years which is also shown in the time series plot.

(b)

> glm.fit=glm(Direction~Lag1+Lag2+Lag3+Lag4+Lag5+Volume, data=Weekly,family = binomial)

> summary(glm.fit)

Call:

glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +

Volume, family = binomial, data = Weekly)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.6949 -1.2565 0.9913 1.0849 1.4579

Coefficients:

Estimate Std. Error z value

(Intercept) 0.26686 0.08593 3.106

Lag1 -0.04127 0.02641 -1.563

Lag2 0.05844 0.02686 2.175

Lag3 -0.01606 0.02666 -0.602

Lag4 -0.02779 0.02646 -1.050

Lag5 -0.01447 0.02638 -0.549

Volume -0.02274 0.03690 -0.616

Pr(>|z|)

(Intercept) 0.0019 \*\*

Lag1 0.1181

Lag2 0.0296 \*

Lag3 0.5469

Lag4 0.2937

Lag5 0.5833

Volume 0.5377

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1496.2 on 1088 degrees of freedom

Residual deviance: 1486.4 on 1082 degrees of freedom

AIC: 1500.4

Number of Fisher Scoring iterations: 4

Lag2 is the onlt predictor that appears to be statistically significant as its p-value is less than 0.05

(c)

glm.prods<-predict(glm.fit,type = "response")

glm.pred<- rep("Down",Length(glm.probs))

glm.pred[glm.probs>0.5]<- "up"

table(glm.pred, Direction)

Direction

Glm2.pred Down Up

Down 54 48

Up 430 557

In this particular scenario, we conclude that percentage of correct predictions on the training data is (54+557)/1089 which is equal to 56.1065%. This can be also be interpreted as 43.8935% is the training error rate, which is often overly optimistic. The overall accuracy of the model is 56,11%.The sensitivity is 56.43% which explains that we are able to perform better than the baseline

(d)

> train<-(Year<2009)

> Weekly.0910<-Weekly[!train,]

> dim(Weekly.0910)

[1] 104 9

> glm.pred=rep("Down",104)

> glm.pred[glm.probs>0.5]="Up"

> table(Direction.0910,glm.pred,dnn = c("Actual Direction", "Predicted Direction"))

Predicted Direction

Actual Direction Down Up

Down 9 34

Up 5 56

In this particular scenario, we may conclude that the percentage of correct predictions on the test data is 62.5%.In other words, using the most significant variable, Lag 2, in our logistic regression model we find the overall accuracy of the model to be (9+56)/104 which is equal to 62.5%

(e)

> lda.fit<-lda(Direction~Lag2,data=Weekly,subset=train)

> lda.predict=predict(lda.fit,Weekly.0910)

> lda.class=lda.predict$class

> table(Direction.0910,lda.class,dnn = c("Actual Direction", "Predicted Direction"))

Predicted Direction

Actual Direction Down Up

Down 9 34

Up 5 56

Using Linear Discriminant Analysis we got overall accuracy to be 62.5% which is the same as Logistic Regression.

(f)

> qda.fit<-qda(Direction~Lag2,data=Weekly,subset=train)

> qda.predict=predict(qda.fit,Weekly.0910)

> qda.class=qda.predict$class

> table(Direction.0910,qda.class,dnn = c("Actual Direction", "Predicted Direction"))

Predicted Direction

Actual Direction Down Up

Down 0 43

Up 0 61

Using Quadratic Discriminant Analysis we get overall accuracy to be 58.65%.QDA predicts that direction will always be up.

(g)

> test.x<-as.matrix(Lag2[!train])

> train.Direction<-Direction[train]

> set.seed(1)

> knn.pred<-knn(train.x,test.x,train.Direction,k=1)

> table(Direction.0910,knn.pred,dnn = c("Actual Direction", "Predicted Direction"))

Predicted Direction

Actual Direction Down Up

Down 21 22

Up 30 31

Using KNN method we get accuracy (21+31)/(21+31+22+30) which is 50%.We are using only Lag2 as a variable in the matrix to find nearest neighbor.

(h)

If we use overall accuracy of the model as our judging criteria we get Logistic Regression and LDA giving the best results and output

(i)

> glm.fit=glm(Direction~Lag1\*Lag2, data=Weekly,family = binomial, subset=train)

> glm.probs=predict(glm.fit,Weekly.0910,type="response")

> glm.pred=rep("Down",104)

> glm.pred[glm.probs>0.5]="Up"

> table(Direction.0910,glm.pred,dnn = c("Actual Direction", "Predicted Direction"))

Predicted Direction

Actual Direction Down Up

Down 7 36

Up 8 53

[1] 0.576931

> lda.fit<-lda(Direction~Lag2\*Lag1,data=Weekly,subset=train)

> lda.predict=predict(lda.fit,Weekly.0910)

> lda.class=lda.predict$class

> table(Direction.0910,lda.class,dnn = c("Actual Direction", "Predicted Direction"))

Predicted Direction

Actual Direction Down Up

Down 7 36

Up 8 53

[1] 0.5769231

> qda.fit<-qda(Direction~Lag2\*Lag1,data=Weekly,subset=train)

> qda.predict=predict(qda.fit,Weekly.0910)

> qda.class=qda.predict$class

> table(Direction.0910,qda.class,dnn = c("Actual Direction", "Predicted Direction"))

Predicted Direction

Actual Direction Down Up

Down 23 20

Up 36 25

[1] 0.4615385

> Weeklymod<-Weekly[,c(2:6)]

> standardized.x<-scale(Weeklymod)

> test<-986:1089

> train.x<-standardized.x[-test,]

> test.x<-standardized.x[test,]

> train.y=Direction[-test]

> test.y=Direction[test]

> set.seed(1)

> knn.pred<-knn(train.x,test.x,train.y,k=10)

> table(test.y,knn.pred,dnn = c("Actual Direction", "Predicted Direction"))

Predicted Direction

Actual Direction Down Up

Down 17 26

Up 20 41

[1] 0.5576923

## Question 5

p= c(0.1, 0.15, 0.2, 0.2, 0.55, 0.6, 0.6, 0.65, 0.7, 0.75)

There are two commmon ways to combine these results together into a single class prediction.One is amajority approach and the second is average approach.

## Majority Approach

sum(p >= 0.5) > sum(p < 0.5)

## [1] TRUE

The number of red predictions is greater than the number of green predictions based on a 50% threshold,thus RED

## Average Approach

mean(p)

## [1] 0.45

The average of the probabilities is less than the 50% threshol,thus GREEN