Home Work Data Analysis 1

Md Shamsuzzaman

11.08.2017

1 Executive Summary

The Weekly data from the ISLR package is the data regarding the weekly percentage returns for the S&P of 500 stock index in between 1990 and 2010. This data set, containing 1089 variables, has been analysed by logistic regression along with the linear and quadratic discriminant analysis. At the beginning, the logistic regression has been applied on the full data set by using direction as the response with the possible outputs up and down. Lag1 to lag5 and volume have been treated as the predictors, and no assumption has not been used in this analysis. From the summary result, The AIC value has been found is 1500.4 and the lag2 has been accepted as the predictor due to the p-value (0.0296) which is less than 0.05. As a result, the null hypothesis has been rejected by concluding that the information of lag2 is important. Along with that , no strong relationship has not been found from the correlation matrix. It has been stated that 54 data set are labeled down and 557 are labeled as up according to the confusion matrix. It also can be stated that by this logistic model 56.10% data is classified correctly while 43.89% is misclassified. After getting this information, only lag2 data has been used as the predictor and the data period from 1990 to 2008 as the training data. The resulted AIC value (1282.5), which is less than the full data set model, indicates that the second model is better than the full model. Additionally, the confusion matrix from this logistic regression model suggests that 7 data are correctly labeled down and 79 data are correctly labeled up from 156 data. Finally, It has been found from the QDA confusion matrix that 6 data are correctly classified in down and 79 data are correctly classified as up. Along with that we can say, 54% data is correctly assigned while 45% is misclassified which is higher than that in LDA. So, according to the classification it can be expressed that the LDA model is better than the QDA model.

2 Introduction

The Weekly data from the ISLR package is the data regarding the weekly percentage returns for the S&P of 500 stock index in between 1990 and 2010. This study was conducted to get the information from the lag1 to lag5 and volume could be used as to predict the directions (up and down). This data set has been analysed through a quantitative probability model. The goal of this data analysis was to get the misclassification rate to distinguish the models. The quantitive model is following below.

```
P(X)=P(Y=1|X)= Ao + Ax
The null hypothesis: Ai is equal to 0 . (x is not important or not related to Y).
Alternative hypothesis: Ai is not 0. ( Here x is important).
```

3 Data Collection

This data set has been collected on the basis of weekly percentage returns for the S&P 500 stock index between 1990 and 2010. Where , Year - The year that the observation was recorded, Lag1 - Percentage return for previous week, Lag2 - Percentage return for 2 weeks previous, Lag3 - Percentage return for 3 weeks previous, Lag4 - Percentage return for 4 weeks previous, Lag5 - Percentage return for 5 weeks previous, Volume - Volume of shares traded (average number of daily shares traded in billions), Today - Percentage return for this week, and Direction - A factor with levels Down and Up indicating whether the market had a positive or negative return on a given week.

4 Summary Details

Different visual graphs can express the linearity information.

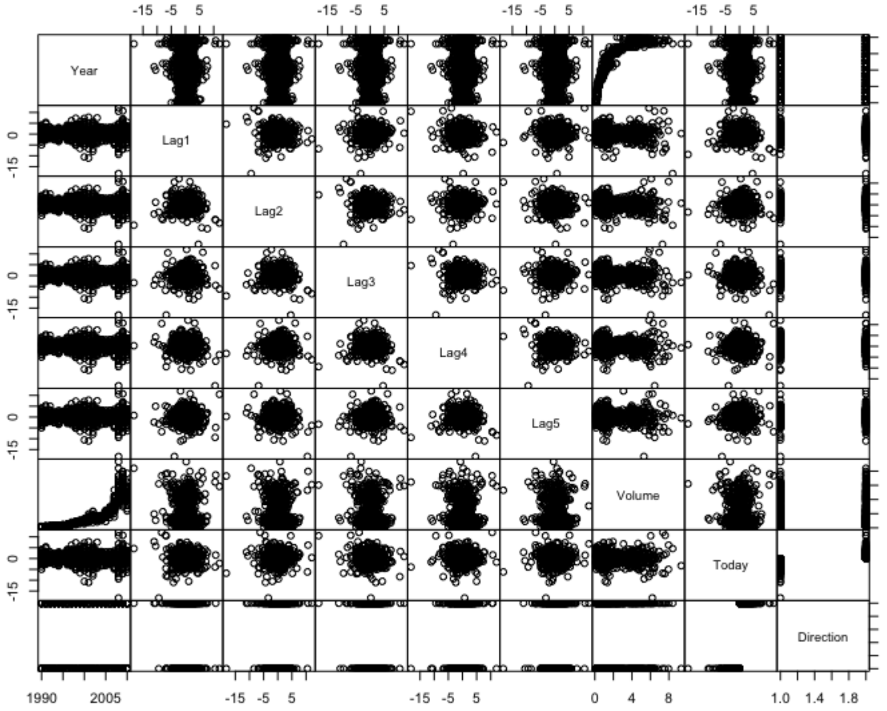

Fig 1. Scattered plot of Weekly data set

The scattered plot does not reveal any strong linear relationship though there is a linearity between the year and volume variable where most of them have no correlation.

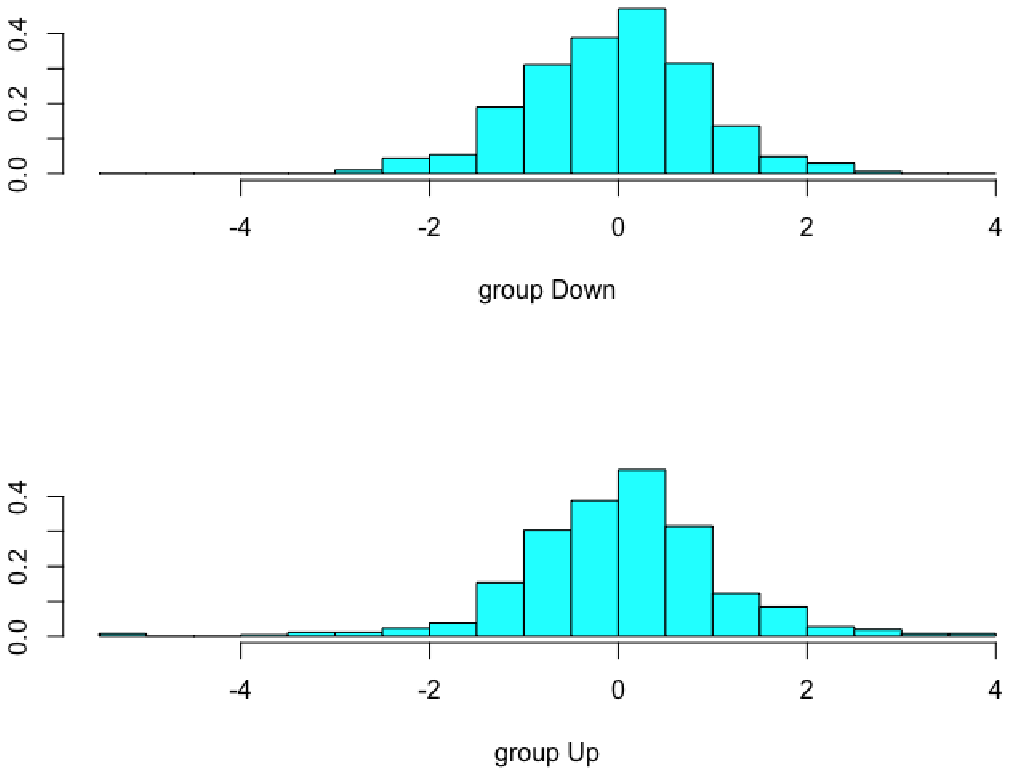

Fig 2. Fitted plot of LDA

From the fitted LDA plot, it seems the data classified in the group up and down are normal.

5 Analysis

```
> names(Weekly)
[1] "Year"      "Lag1"      "Lag2"      "Lag3"      "Lag4"      "Lag5"      "Volume"
[8] "Today"     "Direction"
> summary(Weekly)
      Year           Lag1               Lag2               Lag3
 Min.   :1990   Min.   :-18.1950   Min.   :-18.1950   Min.   :-18.1950
 1st Qu.:1995   1st Qu.: -1.1540   1st Qu.: -1.1540   1st Qu.: -1.1580
 Median :2000   Median :  0.2410   Median :  0.2410   Median :  0.2410
 Mean   :2000   Mean   :  0.1506   Mean   :  0.1511   Mean   :  0.1472
 3rd Qu.:2005   3rd Qu.:  1.4050   3rd Qu.:  1.4090   3rd Qu.:  1.4090
 Max.   :2010   Max.   : 12.0260   Max.   : 12.0260   Max.   : 12.0260
      Lag4               Lag5              Volume            Today          Direction
 Min.   :-18.1950   Min.   :-18.1950   Min.   :0.08747   Min.   :-18.1950
 1st Qu.: -1.1580   1st Qu.: -1.1660   1st Qu.:0.33202   1st Qu.: -1.1540
 Median :  0.2380   Median :  0.2340   Median :1.00268   Median :  0.2410
 Mean   :  0.1458   Mean   :  0.1399   Mean   :1.57462   Mean   :  0.1499
 3rd Qu.:  1.4090   3rd Qu.:  1.4050   3rd Qu.:2.05373   3rd Qu.:  1.4050
 Max.   : 12.0260   Max.   : 12.0260   Max.   :9.32821   Max.   : 12.0260
 Down:484
 Up  :605
```

From the summary of the Weekly data set, 484 data are labeled as down while 605 are labeled as up among 1089 observations. Now, the fitted model of the full data set is necessary to see which information appears to be important. Then after that we can perform the confusion table to get the misclassification rate.

```
> cor(Weekly[,-c(1,8,9)])
               Lag1        Lag2        Lag3        Lag4         Lag5      Volume
Lag1    1.000000000 -0.07485305  0.05863568 -0.07127388 -0.008183096 -0.06495131
Lag2   -0.074853051  1.00000000 -0.07572091  0.05838153 -0.072499482 -0.08551314
Lag3    0.058635682 -0.07572091  1.00000000 -0.07539587  0.060657175 -0.06928771
Lag4   -0.071273876  0.05838153 -0.07539587  1.00000000 -0.075675027 -0.06107462
Lag5   -0.008183096 -0.07249948  0.06065717 -0.07567503  1.000000000 -0.05851741
Volume -0.064951313 -0.08551314 -0.06928771 -0.06107462 -0.058517414  1.00000000
> 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 Pr(>|z|)
(Intercept)  0.26686    0.08593   3.106   0.0019 **
Lag1        -0.04127    0.02641  -1.563   0.1181
```

```
Lag2         0.05844    0.02686   2.175   0.0296 *
Lag3        -0.01606    0.02666  -0.602   0.5469
Lag4        -0.02779    0.02646  -1.050   0.2937
Lag5        -0.01447    0.02638  -0.549   0.5833
Volume      -0.02274    0.03690  -0.616   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

> coef(glm.fit)
(Intercept)        Lag1        Lag2        Lag3        Lag4        Lag5      Volume
 0.26686414 -0.04126894  0.05844168 -0.01606114 -0.02779021 -0.01447206 -0.02274153
> summary(glm.fit)$coef[,4]
(Intercept)        Lag1        Lag2        Lag3        Lag4        Lag5      Volume
0.001898848 0.118144368 0.029601361 0.546923890 0.293653342 0.583348244 0.537674762
> glm.probs=predict(glm.fit, type="response")
> glm.probs[1:10]
        1         2         3         4         5         6         7         8
0.6086249 0.6010314 0.5875699 0.4816416 0.6169013 0.5684190 0.5786097 0.5151972
        9        10
0.5715200 0.5554287
> contrasts(Direction)
     Up
Down  0
Up    1
> glm.pred=rep("Down",1089)
> glm.pred[glm.probs>.50]="up"
> table(glm.pred, Direction)
        Direction
glm.pred Down  Up
    Down   54  48
    up    430 557
```

The logistic regression of the full data set without year and today, the information appears to be significant is lag2 with the p-value 0.0296. It indicates that null hypothesis is rejected. So, the information regarding lag2 is important. Along with that , no strong relationship has not been found from the correlation matrix. It has been stated that 54 data set are labeled down and 557 are labeled as up according to the confusion matrix. It also can be stated that by this logistic model 56.10% data is classified correctly while 43.89% is misclassified. Now, we can use only lag2 data as the predictor and the data period from 1990 to 2008 as the training data because the lag2 information is important.

```
> train=(Year < 2008)
> Weekly.2008= Weekly[!train,]
> dim(Weekly.2008)
[1] 156   9
> Direction.2008=Direction[!train]
> glm.fit2=glm(Direction~Lag2, data=Weekly, family=binomial, subset=train)
> summary(glm.fit2)

Call:
glm(formula = Direction ~ Lag2, family = binomial, data = Weekly,
    subset = train)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.395  -1.274   1.028   1.082   1.305

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.22658    0.06621   3.422 0.000621 ***
Lag2         0.04716    0.03230   1.460 0.144293
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1280.6  on 932  degrees of freedom
Residual deviance: 1278.5  on 931  degrees of freedom
AIC: 1282.5

Number of Fisher Scoring iterations: 4

> coef(glm.fit2)
```

```
(Intercept)        Lag2
 0.22658454  0.04715526
> summary(glm.fit2)$coef
              Estimate Std. Error  z value     Pr(>|z|)
(Intercept) 0.22658454 0.06621167 3.422124 0.0006213398
Lag2        0.04715526 0.03229838 1.459988 0.1442932457
> summary(glm.fit2)$coef[,4]
 (Intercept)         Lag2
0.0006213398 0.1442932457
> glm.probs=predict (glm.fit2, Weekly.2008, type="response")
> glm.pred2=rep("Down",156)
> glm.pred2[glm.probs>.50]="up"
> table(glm.pred2, Direction.2008)
         Direction.2008
glm.pred2 Down Up
     Down    7  5
     up     65 79
```

Here, according to this model the AIC value is 1282.5, which is less than the full model.So, it is better than the full model. Additionally, the confusion matrix from this logistic regression model suggests that 7 data are correctly labeled down and 79 data are correctly labeled up from 156 data. The misclassification rate is 44.87%. Now, we can go for the discriminate analysis or LDA.

```
> lda.fit=lda(Direction~Lag2, data=Weekly, subset=train)######fitting LDA
> lda.fit
Call:
lda(Direction ~ Lag2, data = Weekly, subset = train)

Prior probabilities of groups:
     Down        Up
0.4415863 0.5584137

Group means:
           Lag2
Down 0.07329612
Up   0.27110173

Coefficients of linear discriminants:
           LD1
Lag2 0.4876669
> plot(lda.fit)
> lda.pred=predict(lda.fit, Weekly.2008)
> names(lda.pred)
[1] "class"     "posterior" "x"
> lda.class=lda.pred$class
> table(lda.class, Direction.2008)
          Direction.2008
lda.class Down Up
     Down    6  5
     Up     66 79
> mean(lda.class==Direction.2008)
[1] 0.5448718
> mean(lda.class!=Direction.2008)
[1] 0.4551282
```

It has been found from the QDA confusion matrix that 6 data are correctly classified in down and 79 data are correctly classified as up. Along with that we can say, 54% data is correctly assigned while 45% is misclassified which is higher than that in LDA.

6 Conclusion

Lag2 appears to be significant due to the p-value of it which is less than predefined value 0.05, and this significance has been found from the fitted logistic regression model of the full data set excluding year and today variables. The logistic regression of the full data set without year and today, the information appears to be significant is lag2 with the p-value 0.0296. It indicates that null hypothesis is rejected. So, the information regarding lag2 is important. Along with that , no strong relationship has been found from the correlation matrix. It has been stated that 54 data set are labeled down and 557 are labeled as up according to the confusion matrix. It also can be stated that by this logistic model 56.10% data is classified correctly while 43.89% is misclassified. Only lag2 data as the predictor and the data period from 1990 to 2008 as the training data because the lag2 information is important. Due to the lower AIC value (1282.5), which is less than the full model reveals that this model is better than the full model. Finally, LDA and QDA were run for the reduces model and found that the misclassification rate of LDA is lower than the QDA. As a result, it can be concluded that the LDA model can be use for the prediction of direction of average response with lag2 as the predictor in the period from 1990 to 2008.

7 Appendix

7.1 R Code

```
library(ISLR)
library(MASS)
attach(Weekly)
names(Weekly)
summary(Weekly)
par(mex=0.5)
pairs(Weekly, gap=0, cex.labels = 0.9)
cor(Weekly[,-c(1,8,9)])
glm.fit=glm(Direction~Lag1+Lag2+Lag3+Lag4+Lag5+Volume,data=Weekly, family=binomial)
summary(glm.fit)
coef(glm.fit)
summary(glm.fit)$coef[,4]
glm.probs=predict(glm.fit, type="response")
glm.probs[1:10]
contrasts(Direction)
glm.pred=rep("Down",1089)
glm.pred[glm.probs>.50]="up"
table(glm.pred, Direction)
train=(Year < 2008)
Weekly.2008= Weekly[!train,]
dim(Weekly.2008)
Direction.2008=Direction[!train]
glm.fit2=glm(Direction~Lag2, data=Weekly, family=binomial, subset=train)
summary(glm.fit2)
coef(glm.fit2)
summary(glm.fit2)$coef
summary(glm.fit2)$coef[,4]
glm.probs=predict (glm.fit2, Weekly.2008, type="response")
glm.pred2=rep("Down",156)
glm.pred2[glm.probs>.50]="up"
table(glm.pred2, Direction.2008)
mean(glm.pred2== Direction.2008)
mean(glm.pred2!=Direction.2008)
library(MASS)
lda.fit=lda(Direction~Lag2, data=Weekly, subset=train)
lda.fit
plot(lda.fit)
lda.pred=predict(lda.fit, Weekly.2008)
names(lda.pred)
lda.class=lda.pred$class
table(lda.class, Direction.2008)
mean(lda.class==Direction.2008)
mean(lda.class!=Direction.2008)
qda.fit=qda(Direction~Lag2, data=Weekly, subset=train)
qda.fit
qda.class=predict(qda.fit, Weekly.2008)$class
table(qda.class, Direction.2008)
mean(qda.class==Direction.2008)
mean(qda.class!=Direction.2008)
```

7.2 Log File

```
> install.packages("ISLR", dependencies = FALSE)
> library("ISLR", lib.loc="/Library/Frameworks/R.framework/Versions/3.3/Resources/library")
> library(ISLR)
> library(MASS)
> attach(Weekly)
> names(Weekly)
[1] "Year"      "Lag1"      "Lag2"      "Lag3"      "Lag4"      "Lag5"      "Volume"
[8] "Today"     "Direction"
> summary(Weekly)
      Year           Lag1               Lag2               Lag3
 Min.   :1990   Min.   :-18.1950   Min.   :-18.1950   Min.   :-18.1950
 1st Qu.:1995   1st Qu.: -1.1540   1st Qu.: -1.1540   1st Qu.: -1.1580
 Median :2000   Median :  0.2410   Median :  0.2410   Median :  0.2410
 Mean   :2000   Mean   :  0.1506   Mean   :  0.1511   Mean   :  0.1472
 3rd Qu.:2005   3rd Qu.:  1.4050   3rd Qu.:  1.4090   3rd Qu.:  1.4090
 Max.   :2010   Max.   : 12.0260   Max.   : 12.0260   Max.   : 12.0260
      Lag4               Lag5              Volume            Today          Direction
 Min.   :-18.1950   Min.   :-18.1950   Min.   :0.08747   Min.   :-18.1950   Down:484
 1st Qu.: -1.1580   1st Qu.: -1.1660   1st Qu.:0.33202   1st Qu.: -1.1540   Up  :605
 Median :  0.2380   Median :  0.2340   Median :1.00268   Median :  0.2410
 Mean   :  0.1458   Mean   :  0.1399   Mean   :1.57462   Mean   :  0.1499
 3rd Qu.:  1.4090   3rd Qu.:  1.4050   3rd Qu.:2.05373   3rd Qu.:  1.4050
 Max.   : 12.0260   Max.   : 12.0260   Max.   :9.32821   Max.   : 12.0260
> par(mex=0.5)
> pairs(Weekly, gap=0, cex.labels = 0.9)
> cor(Weekly[,-c(1,8,9)])
               Lag1        Lag2        Lag3        Lag4         Lag5      Volume
Lag1    1.000000000 -0.07485305  0.05863568 -0.07127388 -0.008183096 -0.06495131
Lag2   -0.074853051  1.00000000 -0.07572091  0.05838153 -0.072499482 -0.08551314
Lag3    0.058635682 -0.07572091  1.00000000 -0.07539587  0.060657175 -0.06928771
Lag4   -0.071273876  0.05838153 -0.07539587  1.00000000 -0.075675027 -0.06107462
```

```
Lag5   -0.008183096 -0.07249948  0.06065717 -0.07567503  1.000000000 -0.05851741
Volume -0.064951313 -0.08551314 -0.06928771 -0.06107462 -0.058517414  1.00000000
> 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 Pr(>|z|)
(Intercept)  0.26686    0.08593   3.106   0.0019 **
Lag1        -0.04127    0.02641  -1.563   0.1181
Lag2         0.05844    0.02686   2.175   0.0296 *
Lag3        -0.01606    0.02666  -0.602   0.5469
Lag4        -0.02779    0.02646  -1.050   0.2937
Lag5        -0.01447    0.02638  -0.549   0.5833
Volume      -0.02274    0.03690  -0.616   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

> coef(glm.fit)
(Intercept)        Lag1        Lag2        Lag3        Lag4        Lag5      Volume
 0.26686414 -0.04126894  0.05844168 -0.01606114 -0.02779021 -0.01447206 -0.02274153
> summary(glm.fit)$coef[,4]
(Intercept)        Lag1        Lag2        Lag3        Lag4        Lag5      Volume
0.001898848 0.118144368 0.029601361 0.546923890 0.293653342 0.583348244 0.537674762
> glm.probs=predict(glm.fit, type="response")
> glm.probs[1:10]
```

```
        1         2         3         4         5         6         7         8 
0.6086249 0.6010314 0.5875699 0.4816416 0.6169013 0.5684190 0.5786097 0.5151972 
        9        10 
0.5715200 0.5554287 
> contrasts(Direction)
     Up
Down  0
Up    1
> glm.pred=rep("Down",1089)
> glm.pred[glm.probs>.50]="up"
> table(glm.pred, Direction)
        Direction
glm.pred Down  Up
    Down   54  48
    up    430 557
> train=(Year < 2008)
> Weekly.2008= Weekly[!train,]
> dim(Weekly.2008)
[1] 156   9
> Direction.2008=Direction[!train]
> glm.fit2=glm(Direction~Lag2, data=Weekly, family=binomial, subset=train)
> summary(glm.fit2)

Call:
glm(formula = Direction ~ Lag2, family = binomial, data = Weekly, 
    subset = train)

Deviance Residuals: 
   Min      1Q  Median      3Q     Max  
-1.395  -1.274   1.028   1.082   1.305  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)  0.22658    0.06621   3.422 0.000621 ***
Lag2         0.04716    0.03230   1.460 0.144293    
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1 

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1280.6  on 932  degrees of freedom
```

```
Residual deviance: 1278.5  on 931  degrees of freedom
AIC: 1282.5

Number of Fisher Scoring iterations: 4

> coef(glm.fit2)
(Intercept)        Lag2
 0.22658454  0.04715526
> summary(glm.fit2)$coef
              Estimate Std. Error  z value     Pr(>|z|)
(Intercept) 0.22658454 0.06621167 3.422124 0.0006213398
Lag2        0.04715526 0.03229838 1.459988 0.1442932457
> summary(glm.fit2)$coef[,4]
 (Intercept)         Lag2
0.0006213398 0.1442932457
> glm.probs=predict (glm.fit2, Weekly.2008, type="response")
> glm.pred2=rep("Down",156)
> glm.pred2[glm.probs>.50]="up"
> table(glm.pred2, Direction.2008)
         Direction.2008
glm.pred2 Down Up
     Down    7  5
     up     65 79
> mean(glm.pred2== Direction.2008)
[1] 0.04487179
> mean(glm.pred2!=Direction.2008)
[1] 0.9551282
> library(MASS)
> lda.fit=lda(Direction~Lag2, data=Weekly, subset=train)######fitting LDA
> lda.fit
Call:
lda(Direction ~ Lag2, data = Weekly, subset = train)

Prior probabilities of groups:
     Down        Up
0.4415863 0.5584137

Group means:
           Lag2
Down 0.07329612
Up   0.27110173
```

```
Coefficients of linear discriminants:
           LD1
Lag2 0.4876669
> plot(lda.fit)
> lda.pred=predict(lda.fit, Weekly.2008)
> names(lda.pred)
[1] "class"     "posterior" "x"
> lda.class=lda.pred$class
> table(lda.class, Direction.2008)
         Direction.2008
lda.class Down Up
     Down    6  5
     Up     66 79
> mean(lda.class==Direction.2008)
[1] 0.5448718
> mean(lda.class!=Direction.2008)
[1] 0.4551282
> qda.fit=qda(Direction~Lag2, data=Weekly, subset=train)######fitting QDA
> qda.fit
Call:
qda(Direction ~ Lag2, data = Weekly, subset = train)

Prior probabilities of groups:
     Down        Up
0.4415863 0.5584137

Group means:
           Lag2
Down 0.07329612
Up   0.27110173
> qda.class=predict(qda.fit, Weekly.2008)$class
> table(qda.class, Direction.2008)
         Direction.2008
qda.class Down Up
     Down    0  0
     Up     72 84
> mean(qda.class==Direction.2008)
[1] 0.5384615
> mean(qda.class!=Direction.2008)
[1] 0.4615385
```