HW3#5

(a)Produce some numerical and graphical summaries of the Weekly data. Do there appear to be any patterns?

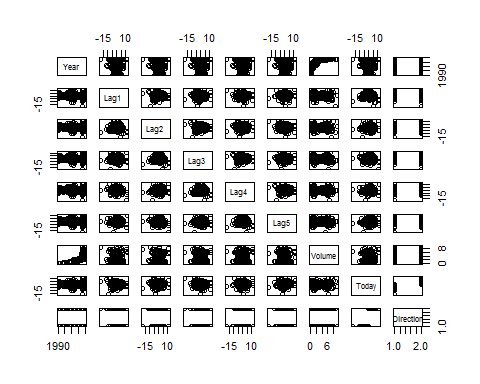
library(ISLR)

## Warning: package 'ISLR' was built under R version 3.2.5

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   
## Min. :-18.1950 Min. :-18.1950 Min. :0.08747   
## 1st Qu.: -1.1580 1st Qu.: -1.1660 1st Qu.:0.33202   
## Median : 0.2380 Median : 0.2340 Median :1.00268   
## Mean : 0.1458 Mean : 0.1399 Mean :1.57462   
## 3rd Qu.: 1.4090 3rd Qu.: 1.4050 3rd Qu.:2.05373   
## Max. : 12.0260 Max. : 12.0260 Max. :9.32821   
## Today Direction   
## Min. :-18.1950 Down:484   
## 1st Qu.: -1.1540 Up :605   
## Median : 0.2410   
## Mean : 0.1499   
## 3rd Qu.: 1.4050   
## Max. : 12.0260

pairs(Weekly)

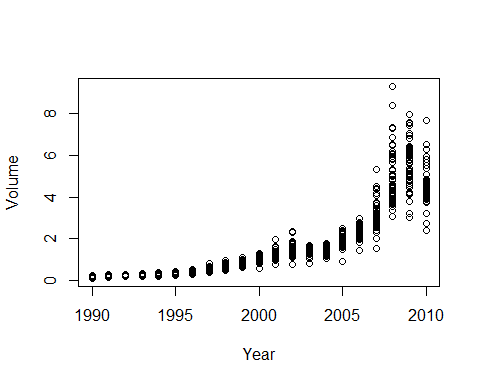


cor(Weekly[,-9])

## Year Lag1 Lag2 Lag3 Lag4  
## Year 1.00000000 -0.032289274 -0.03339001 -0.03000649 -0.031127923  
## Lag1 -0.03228927 1.000000000 -0.07485305 0.05863568 -0.071273876  
## Lag2 -0.03339001 -0.074853051 1.00000000 -0.07572091 0.058381535  
## Lag3 -0.03000649 0.058635682 -0.07572091 1.00000000 -0.075395865  
## Lag4 -0.03112792 -0.071273876 0.05838153 -0.07539587 1.000000000  
## Lag5 -0.03051910 -0.008183096 -0.07249948 0.06065717 -0.075675027  
## Volume 0.84194162 -0.064951313 -0.08551314 -0.06928771 -0.061074617  
## Today -0.03245989 -0.075031842 0.05916672 -0.07124364 -0.007825873  
## Lag5 Volume Today  
## Year -0.030519101 0.84194162 -0.032459894  
## Lag1 -0.008183096 -0.06495131 -0.075031842  
## Lag2 -0.072499482 -0.08551314 0.059166717  
## Lag3 0.060657175 -0.06928771 -0.071243639  
## Lag4 -0.075675027 -0.06107462 -0.007825873  
## Lag5 1.000000000 -0.05851741 0.011012698  
## Volume -0.058517414 1.00000000 -0.033077783  
## Today 0.011012698 -0.03307778 1.000000000

Thera is high correlation between Year and Volume, lets plot bivariate plot for these variables

attach(Weekly)  
plot(Year,Volume)

 Median volume is increasing each year

(b)Use the full data set to perform a logistic regression with Direction as the response and the five lag variables plus Volume as predictors.

glm\_model <- glm(Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + Volume, data = Weekly, family = binomial)  
summary(glm\_model)

##   
## 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

## findings

(c)Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression.

ods <- predict(glm\_model, type = "response")  
glm\_pred <- rep("Down", length(ods))  
glm\_pred[ods > 0.5] <- "Up"  
table(glm\_pred, Direction)

## Direction  
## glm\_pred Down Up  
## Down 54 48  
## Up 430 557

## findings

(d)Now fit the logistic regression model using a training data period from 1990 to 2008, with "Lag2" as the only predictor. Compute the confusion matrix and the overall fraction of correct predictions for the held out data (that is, the data from 2009 to 2010).

Weekly\_Pre2009 <- Weekly[Year<2009, ]  
Weekly\_Post2009 <- Weekly[Year>2008, ]  
glm\_model <- glm(Direction ~ Lag2, data = Weekly\_Pre2009, family = binomial)  
summary(glm\_model)

##   
## Call:  
## glm(formula = Direction ~ Lag2, family = binomial, data = Weekly\_Pre2009)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.536 -1.264 1.021 1.091 1.368   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.20326 0.06428 3.162 0.00157 \*\*  
## Lag2 0.05810 0.02870 2.024 0.04298 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1354.7 on 984 degrees of freedom  
## Residual deviance: 1350.5 on 983 degrees of freedom  
## AIC: 1354.5  
##   
## Number of Fisher Scoring iterations: 4

## findings

ods <- predict(glm\_model, Weekly\_Post2009, type = "response")  
glm\_pred <- rep("Down", length(ods))  
glm\_pred[ods > 0.5] <- "Up"  
table(glm\_pred, Weekly\_Post2009$Direction)

##   
## glm\_pred Down Up  
## Down 9 5  
## Up 34 56

## findings

(e)Repeat (d) using LDA.

library(MASS)  
lda\_model <- lda(Direction ~ Lag2, data = Weekly\_Pre2009)  
lda\_model

## Call:  
## lda(Direction ~ Lag2, data = Weekly\_Pre2009)  
##   
## Prior probabilities of groups:  
## Down Up   
## 0.4477157 0.5522843   
##   
## Group means:  
## Lag2  
## Down -0.03568254  
## Up 0.26036581  
##   
## Coefficients of linear discriminants:  
## LD1  
## Lag2 0.4414162

lda\_pred <- predict(lda\_model, Weekly\_Post2009)  
table(lda\_pred$class, Weekly\_Post2009$Direction)

##   
## Down Up  
## Down 9 5  
## Up 34 56

## findings

1. Repeat (d) using QDA.

qda\_model <- qda(Direction ~ Lag2, data = Weekly\_Pre2009)  
qda\_model

## Call:  
## qda(Direction ~ Lag2, data = Weekly\_Pre2009)  
##   
## Prior probabilities of groups:  
## Down Up   
## 0.4477157 0.5522843   
##   
## Group means:  
## Lag2  
## Down -0.03568254  
## Up 0.26036581

qda\_pred <- predict(qda\_model, Weekly\_Post2009)  
table(qda\_pred$class, Weekly\_Post2009$Direction)

##   
## Down Up  
## Down 0 0  
## Up 43 61

## findings

1. Repeat (d) using KNN with K=1

set.seed(12)  
library(class)

## Warning: package 'class' was built under R version 3.2.5

## with K =5  
knn\_pred <- knn(as.matrix(Weekly\_Pre2009[,c("Lag2")]), as.matrix(Weekly\_Post2009[,c("Lag2")]), Weekly\_Pre2009$Direction, k = 1)  
table(knn\_pred, Weekly\_Post2009$Direction)

##   
## knn\_pred Down Up  
## Down 21 29  
## Up 22 32

## findings

1. examine whether it is worth to include interactions via a forward selection scheme for LDA, which greedily minimizes the test error as it adds variables to the model one at a time.

library(MASS)  
## step 1  
lda\_model <- lda(Direction ~ Lag2, data = Weekly\_Pre2009)  
lda\_model

## Call:  
## lda(Direction ~ Lag2, data = Weekly\_Pre2009)  
##   
## Prior probabilities of groups:  
## Down Up   
## 0.4477157 0.5522843   
##   
## Group means:  
## Lag2  
## Down -0.03568254  
## Up 0.26036581  
##   
## Coefficients of linear discriminants:  
## LD1  
## Lag2 0.4414162

lda\_pred <- predict(lda\_model, Weekly\_Post2009)  
table(lda\_pred$class, Weekly\_Post2009$Direction)

##   
## Down Up  
## Down 9 5  
## Up 34 56

mean(lda\_pred$class !=Weekly\_Post2009$Direction)

## [1] 0.375

library(MASS)  
## step 2  
lda\_model <- lda(Direction ~ Lag2:Lag3, data = Weekly\_Pre2009)  
lda\_model

## Call:  
## lda(Direction ~ Lag2:Lag3, data = Weekly\_Pre2009)  
##   
## Prior probabilities of groups:  
## Down Up   
## 0.4477157 0.5522843   
##   
## Group means:  
## Lag2:Lag3  
## Down -0.1937158  
## Up -0.6405132  
##   
## Coefficients of linear discriminants:  
## LD1  
## Lag2:Lag3 0.1012928

lda\_pred <- predict(lda\_model, Weekly\_Post2009)  
table(lda\_pred$class, Weekly\_Post2009$Direction)

##   
## Down Up  
## Down 0 0  
## Up 43 61

mean(lda\_pred$class !=Weekly\_Post2009$Direction)

## [1] 0.4134615

library(MASS)  
## step 3  
lda\_model <- lda(Direction ~ Lag2:Lag4, data = Weekly\_Pre2009)  
lda\_model

## Call:  
## lda(Direction ~ Lag2:Lag4, data = Weekly\_Pre2009)  
##   
## Prior probabilities of groups:  
## Down Up   
## 0.4477157 0.5522843   
##   
## Group means:  
## Lag2:Lag4  
## Down 0.78824608  
## Up 0.04407141  
##   
## Coefficients of linear discriminants:  
## LD1  
## Lag2:Lag4 0.1287072

lda\_pred <- predict(lda\_model, Weekly\_Post2009)  
table(lda\_pred$class, Weekly\_Post2009$Direction)

##   
## Down Up  
## Down 1 4  
## Up 42 57

mean(lda\_pred$class !=Weekly\_Post2009$Direction)

## [1] 0.4423077

library(MASS)  
## step 4  
lda\_model <- lda(Direction ~ Lag2:Lag5, data = Weekly\_Pre2009)  
lda\_model

## Call:  
## lda(Direction ~ Lag2:Lag5, data = Weekly\_Pre2009)  
##   
## Prior probabilities of groups:  
## Down Up   
## 0.4477157 0.5522843   
##   
## Group means:  
## Lag2:Lag5  
## Down -0.3132494  
## Up -0.3497535  
##   
## Coefficients of linear discriminants:  
## LD1  
## Lag2:Lag5 0.1105356

lda\_pred <- predict(lda\_model, Weekly\_Post2009)  
table(lda\_pred$class, Weekly\_Post2009$Direction)

##   
## Down Up  
## Down 0 0  
## Up 43 61

mean(lda\_pred$class !=Weekly\_Post2009$Direction)

## [1] 0.4134615