Chapter 4 Problem 10

Andira Putri

September 19, 2018

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

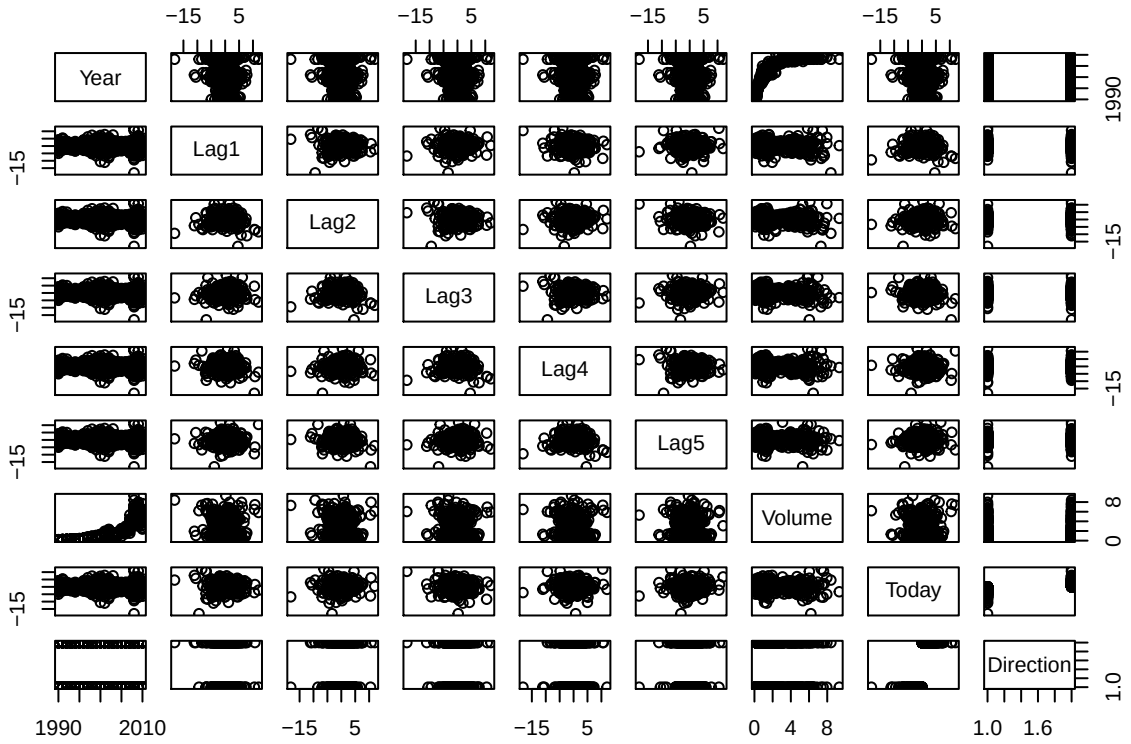
```
library(ISLR)
```

```
## Warning: package 'ISLR' was built under R version 3.5.1
```

```
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)
```



Volume increases as year increases (positively correlated).

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. Print the results. Which predictors are statistically significant?

```
glm.fits=glm(Direction~Lag1+Lag2+Lag3+Lag4+Lag5+Volume,data=Weekly,family=binomial)
#glm fits a series of generalized linear models
#'family=binomial' tells R to perform a logistic regression
summary(glm.fits)
```

```
##
## 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
```

Lag2 appears to be the only statistically significant predictor.

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.

```
#outputs probabilities in the form P(Y=1|X)
glm.probs=predict(glm.fits,type="response")

#generates dummy variables for Up and Down
contrasts(Weekly$Direction)
```

```
##      Up
## Down  0
## Up    1
```

```
#create vector of 1089 "Down"" elements
glm.pred=rep("Down",1089)

#change "Down" to "Up" if probability exceeds 0.5
glm.pred[glm.probs>0.5]="Up"

#generates confusion matrix
table(glm.pred,Weekly$Direction)
```

```
##
## glm.pred Down  Up
##     Down   54  48
##     Up    430 557
```

There are 611 correct classifications out of a total 1089 observations, so the accuracy is $611/1089 = 0.561$. In particular, the Up accuracy, with 557 correct predictions on 605 total Up observations, is $557/605 = 0.92$. The Down accuracy, with 54 correct predictions on 484 total Down observations, is $54/484 = 0.11$, which is pretty bad. Perhaps changing the probability threshold could improve the accuracy.

d.) 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 (2009-2010).

```
#split the Weekly data set and perform logistic regression
training<-Weekly[Weekly$Year<=2008,]
test<-Weekly[Weekly$Year>2008,]
```

```
logreg=glm(Direction~Lag2,data=training,family=binomial)
summary(logreg)
```

```
##
## Call:
## glm(formula = Direction ~ Lag2, family = binomial, data = training)
##
## 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
```

```
#Make predictions on our test data with our classifier
glm.probs2=predict(logreg,test,type="response")
contrasts(test$Direction)
```

```
##      Up
## Down  0
## Up    1
```

```
glm.pred=rep("Down",nrow(test))
glm.pred[glm.probs2>0.5]="Up"
table(glm.pred,test$Direction)
```

```
##
## glm.pred Down Up
##     Down    9  5
##     Up     34 56
```

1. Total accuracy: $(9+56)/104 = 0.625$
2. Up accuracy: $56/(56+5) = 0.918$
3. Down accuracy: $9/(34+9) = 0.209$

e.) Repeat (d) using LDA.

```
library(MASS)
lda.fit=lda(Direction~Lag2,data=training)
lda.fit
```

```
## Call:
## lda(Direction ~ Lag2, data = training)
##
```

```
## 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.fit,test)
names(lda.pred)
```

```
## [1] "class"     "posterior" "x"
```

```
lda.class=lda.pred$class
table(lda.class,test$Direction)
```

```
##
## lda.class Down Up
##      Down    9  5
##      Up     34 56
```

The confusion matrix is the exact same as the one given by logistic regression.

f.) Repeat (d) using QDA.

```
library(MASS)
qda.fit=qda(Direction~Lag2,data=training)
qda.fit
```

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

```
qda.pred=predict(qda.fit,test)
names(qda.pred)
```

```
## [1] "class"     "posterior"
```

```
qda.class=qda.pred$class
table(qda.class,test$Direction)
```

```
##
## qda.class Down Up
##      Down    0  0
##      Up     43 61
```

QDA correctly predicted all Up directions, but incorrectly predicted all Down directions. The total accuracy is 0.5.

g.) Repeat (d) using KNN, K=1.

```
library(class)
```

```
## Warning: package 'class' was built under R version 3.5.1
```

```
train.X <- as.matrix(training$Lag2)
test.X <- as.matrix(test$Lag2)
set.seed(1)
knn.pred<-knn(train.X,test.X,training$Direction,k=1)
table(knn.pred,test$Direction)
```

```
##
## knn.pred Down Up
##     Down   21 30
##     Up     22 31
```

The total accuracy is (21+31)/104 = 0.5.

h.) Which of these methods provide the best results on this data?

Logistic regression and LDA prove to be equally the best methods.