Project #3

Milica Cudina

2024-03-31

Problem #1 (5+10+5+10+10+10+10+10=70 points)

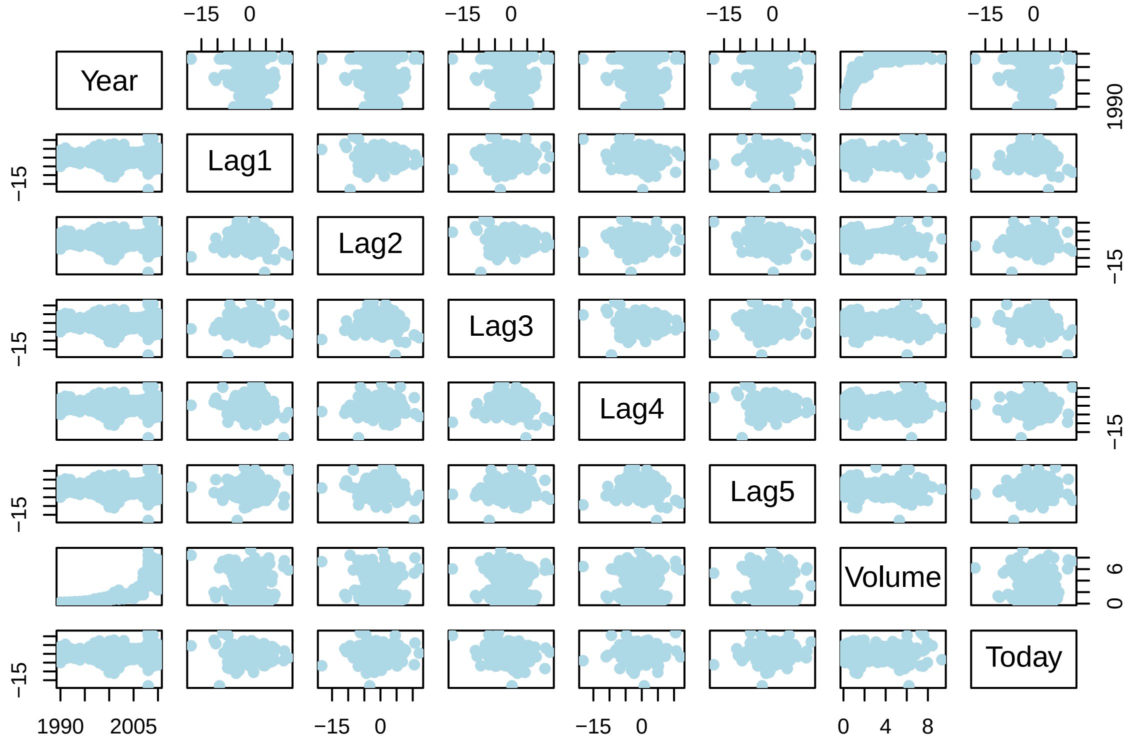
Solve **Problem 4.8.13** (pp. 192-193) from the textbook.

Hint: Here is a list of libraries you will need:

```
library(MASS)
library(ISLR2)
##
## Attaching package: 'ISLR2'
## The following object is masked from 'package:MASS':
##
##     Boston
library(e1071)
```

Solution: First, here is some exploratory data analysis.

```
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
##  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
##  Direction
##  Down:484
##  Up  :605
##
##
##
##
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
plot(Weekly[, -9], pch=19, col="lightblue")
```



As time goes by, there is more and more trading. So, there is a nice correlation between `Year` and `Volume`. Other than that, I cannot discern a pattern.

```
mlr.fit <- glm(
  Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + Volume,
  data = Weekly,
  family = binomial
)
summary(mlr.fit)
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
##     Volume, family = binomial, data = Weekly)
##
```

```
## 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 is the only significant one.

Now, it's time for the **confusion matrix**.

```
probs <- predict(mlr.fit, type = "response")
glm.pred=rep("Down", length(probs))
glm.pred[probs>0.5]<-"Up"
tab <- table(glm.pred, Weekly$Direction)
tab
##         
## glm.pred Down  Up
##     Down   54  48
##     Up    430 557
sum(diag(tab)) / sum(tab)
## [1] 0.5610652
mean(Weekly$Direction=="Up")
## [1] 0.5555556
```

The prediction is correct a bit over 56% of the time. However, the proportion of the realized "Up"s was just under 56%. So, constantly saying "Up" would work almost as well as our logistic regression.

Now, for training and testing.

```
train <- Weekly$Year < 2009

fit <- glm(Direction ~ Lag2, data = Weekly[train, ], family = binomial)
pred <- predict(fit, Weekly[!train, ], type = "response") > 0.5
(t <- table(ifelse(pred, "Up (pred)", "Down (pred)"), Weekly[!train, ]$Direction))
##              
##               Down Up
##   Down (pred)    9  5
##   Up (pred)     34 56
sum(diag(t)) / sum(t)
## [1] 0.625
```

```r
attach(Weekly)
train <- (Year< 2009)
test=Weekly[!train,]
dim(test)
## [1] 104   9
dim(test)
## [1] 104   9
fit.tr <- glm(Direction ~ Lag2, data = test, family = binomial)

probs <- predict(fit.tr, data=Weekly[!train, ], type = "response")
length(probs)
## [1] 104
glm.pred=rep("Down", length(probs))
glm.pred[probs>0.5]<-"Up"
length(glm.pred)
## [1] 104
length(test$Direction)
## [1] 104
tab <- table(glm.pred, test$Direction)
tab
##
## glm.pred Down Up
##     Down    8  4
##     Up     35 57
sum(diag(tab)) / sum(tab)
## [1] 0.625
```