Zixiao Wu 515491-Individual Assignment 3

Zixiao Wu

2022-09-23

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com (http://rmarkdown.rstudio.com).

4.7 Exercises Problem 10

```
library(ISLR)
library(MASS)
library(class)
head(Weekly)
```

```
##   Year   Lag1   Lag2   Lag3   Lag4   Lag5    Volume  Today Direction
## 1 1990  0.816  1.572 -3.936 -0.229 -3.484 0.1549760 -0.270      Down
## 2 1990 -0.270  0.816  1.572 -3.936 -0.229 0.1485740 -2.576      Down
## 3 1990 -2.576 -0.270  0.816  1.572 -3.936 0.1598375  3.514        Up
## 4 1990  3.514 -2.576 -0.270  0.816  1.572 0.1616300  0.712        Up
## 5 1990  0.712  3.514 -2.576 -0.270  0.816 0.1537280  1.178        Up
## 6 1990  1.178  0.712  3.514 -2.576 -0.270 0.1544440 -1.372      Down
```

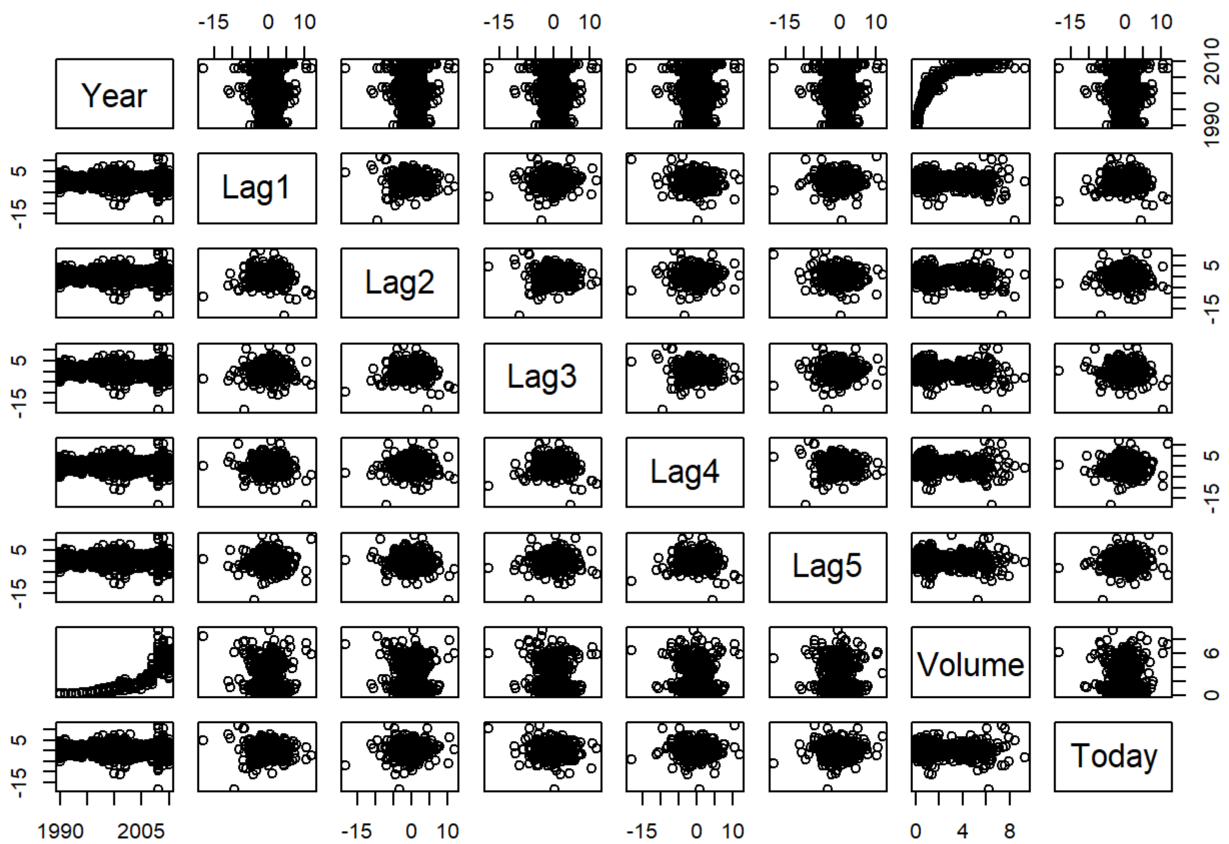
```
attach(Weekly)
```

#(a)

```
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
##
##
##
##
```

```
pairs(Weekly[,1:8])
```




```
cor(Weekly[,1:8])
```

```
##               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
```

```
From the numerical and graphical summaries above, we can know that there may be a positive rela
tionship between 'Year' and 'Volume' only. Also, four lag observations are almost the same.
```

#(b)

```
log = glm(Direction ~ Lag1+Lag2+Lag3+Lag4+Lag5+Volume, data=Weekly, family=binomial)
summary(log)
```

```
##
## 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
```

```
We can see that Lag2 is significant in 99% confidencial level.
```

#(c)

```
log_prob = predict(log, type="response")
log_pred = rep("down", 1089)
log_pred[log_prob > 0.5] = "up"
table(log_pred, Direction)
```

```
##         Direction
## log_pred Down  Up
##     down   54  48
##     up    430 557
```

```
#mean(log_pred==Direction)
(54+557)/(54+557+48+430)
```

```
## [1] 0.5610652
```

```
557/(430+557)
```

```
## [1] 0.5643364
```

```
48/(54+48)
```

```
## [1] 0.4705882
```

```
From the confusion matrix we can know that the correct rate is 56.1%. When predicting a up mark
et, the sensitivity is 56.4%. Also we can see the Type 1 error is 47.1%.
```

#(d)

```
train = (Year<2009)
Test = Weekly[!train ,]
Test_Direction= Direction[!train]

log2 = glm(Direction ~ Lag2, data=Weekly, family=binomial, subset=train)

log_prob2 = predict(log2,Test, type="response")
log_pred2 = rep("Down", nrow(Test))
log_pred2[log_prob2>0.5] = "Up"
table(log_pred2,Test_Direction)
```

```
##          Test_Direction
## log_pred2 Down Up
##      Down    9  5
##      Up     34 56
```

```
mean(log_pred2==Test_Direction)
```

```
## [1] 0.625
```

```
We can see the model has a correct rate of 62.5%.
```

#(e)

```
lda = lda(Direction ~ Lag2, data=Weekly, subset=train)
lda_pred = predict(lda,Test)
table(lda_pred$class, Test_Direction)
```

```
##       Test_Direction
##        Down Up
##   Down    9  5
##   Up     34 56
```

```
mean(lda_pred$class==Test_Direction)
```

```
## [1] 0.625
```

```
We can see the LDA model has the same corret rate as Logistic model.
```

#(g)

```
set.seed(114514)
train.x = Lag2[train]
test.x = Lag2[!train]
train_direction = Direction[train]
dim(train.x) = c(985,1)
dim(test.x) = c(104,1)

knn_pred = knn(train.x, test.x, train_direction, k=1)
table(knn_pred, Test_Direction)
```

```
##         Test_Direction
## knn_pred Down Up
##     Down   21 29
##     Up     22 32
```

```
mean(knn_pred==Test_Direction)
```

```
## [1] 0.5096154
```

```
We can see that the correct rate is 50%, lower than the models above.
```

#(h)

```
The logistic regression and the LDA. They are the methods with a higher correct rate, sensitivi
ty and precision.
```

#(i)

```
#knn, k = 5
knn_pred2 = knn(train.x, test.x, train_direction, k=5)
table(knn_pred2, Test_Direction)
```

```
##          Test_Direction
## knn_pred2 Down Up
##      Down   15 22
##      Up     28 39
```

```
mean(knn_pred2==Test_Direction)
```

```
## [1] 0.5192308
```

```
#knn, k = 10
knn_pred3 = knn(train.x, test.x, train_direction, k=10)
table(knn_pred3, Test_Direction)
```

```
##           Test_Direction
## knn_pred3 Down Up
##      Down   19 19
##      Up     24 42
```

```
mean(knn_pred3==Test_Direction)
```

```
## [1] 0.5865385
```

```
#logistic
log2 = glm(Direction ~ Lag2 + Volume, data=Weekly, family=binomial, subset=train)
log_prob2 = predict(log2, Test, type="response")
log_pred2 = rep("Down", 104)
log_pred2[log_prob2>0.5] = "Up"
table(log_pred2,Test_Direction)
```

```
##           Test_Direction
## log_pred2 Down Up
##      Down   20 25
##      Up     23 36
```

```
mean(log_pred2==Test_Direction)
```

```
## [1] 0.5384615
```

```
#lda
lda2 = lda(Direction ~ Lag2 + Volume, data=Weekly, subset=train)
lda_pred2 = predict(lda2,Test)
table(lda_pred2$class, Test_Direction)
```

```
##       Test_Direction
##         Down Up
##   Down   20 25
##   Up     23 36
```

```
mean(lda_pred2$class==Test_Direction)
```

```
## [1] 0.5384615
```

```
As the experiment above, the model which appears to provide the best results is knn(k = 10) mod
el, with a correct rate of 58%. However, it is still lower than the initial lda model.
```