HW5

2025-03-25

Setup

We will begin by examining some numerical and graphical summaries of the Weekly data, which is part of the ISLR2 library. This data set contains 1,089 weekly returns for the S&P 500 stock index for 21 years, from the beginning of 1990 to the end of 2010. For each date, we have recorded the percentage returns for each of the five previous trading days, Lag1 through Lag5. We have also recorded Volume (the number of shares traded on the previous day, in billions), Today (the percentage return on the date in question), and Direction (whether the market was Up or Down on this date). Our goal is to predict Direction (a qualitative response) using the other features.

Note: Direction is Up if Today is positive, otherwise, it is Down. So, Today must not be used to model Direction.

Questions 1

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

Answer

Nemrical summary

```
library(ISLR2)
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
```

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

Graphical summary

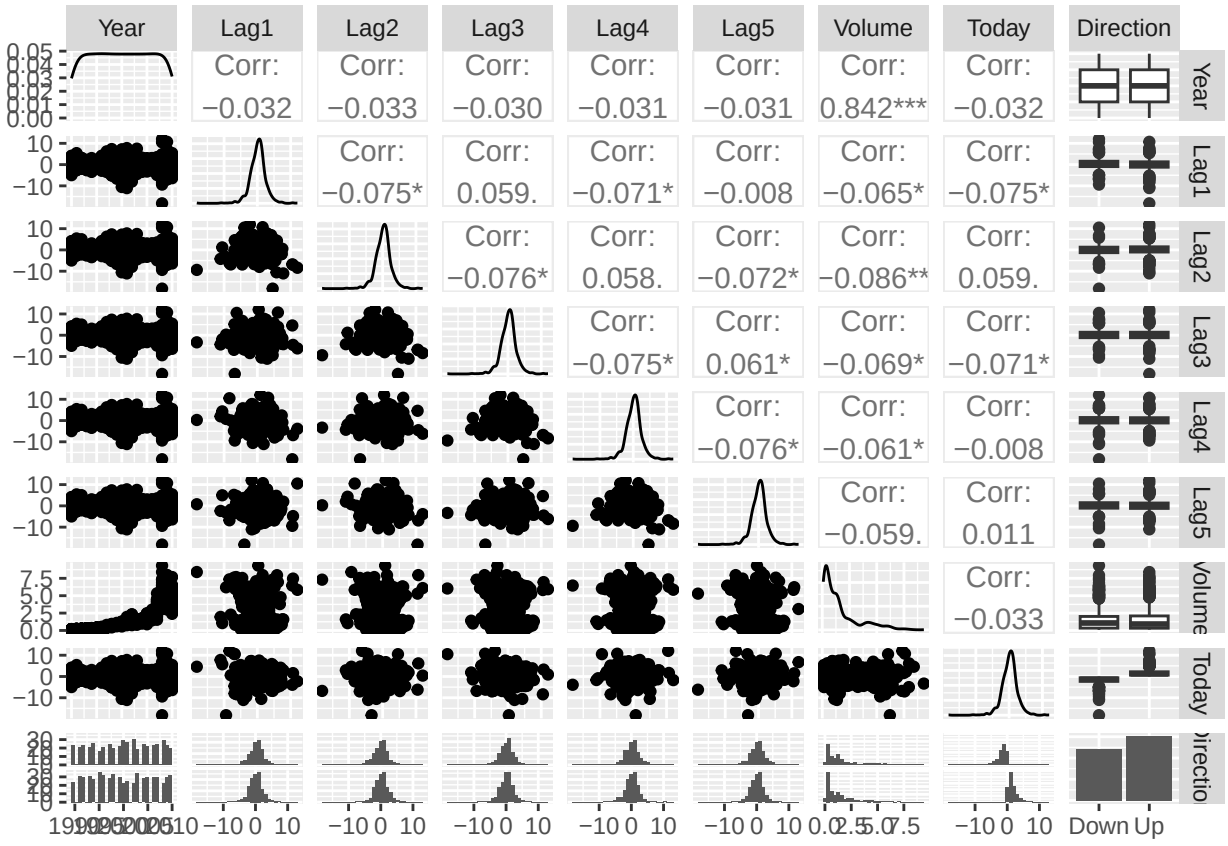
```
library(ggplot2)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
ggpairs(Weekly)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



There is little correlation among the variables, except between Volume and Year. Volume increases over time (Year).

Question 2

Use the data set to perform a logistic regression with Direction as the response and Lag1 as a predictor. Use the summary function to print the results.

Answer

```
log_reg = glm(Direction ~ Lag1, family = 'binomial', data = Weekly)
summary(log_reg)
```

```
##
## Call:
## glm(formula = Direction ~ Lag1, family = "binomial", data = Weekly)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.456  -1.263   1.041   1.087   1.277
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)  0.23024    0.06124   3.760  0.00017 ***
## Lag1        -0.04313    0.02622  -1.645  0.10001
## ---
## 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: 1493.5  on 1087  degrees of freedom
## AIC: 1497.5
##
## Number of Fisher Scoring iterations: 4
```

Question 3

Direction has two levels - Up and Down. The glm function in R uses a dummy variable taking 0 and 1, and then it models $\log\left(\frac{P(Y=1|X=x)}{P(Y=0|X=x)}\right)$. Check how the levels are coded. Hint: contrasts(Weekly$Direction). By default, R follows the alphabetical order during the conversion.

Answer

```
contrasts(Weekly$Direction)
```

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

Up is denoted by 1.

Question 4

Interpret the fitted regression coefficient.

Answer

If Lag1 is increased by one percent, the log odds decreases by 0.04313 %. When Lag1 = 0 log odds is 0.23024.

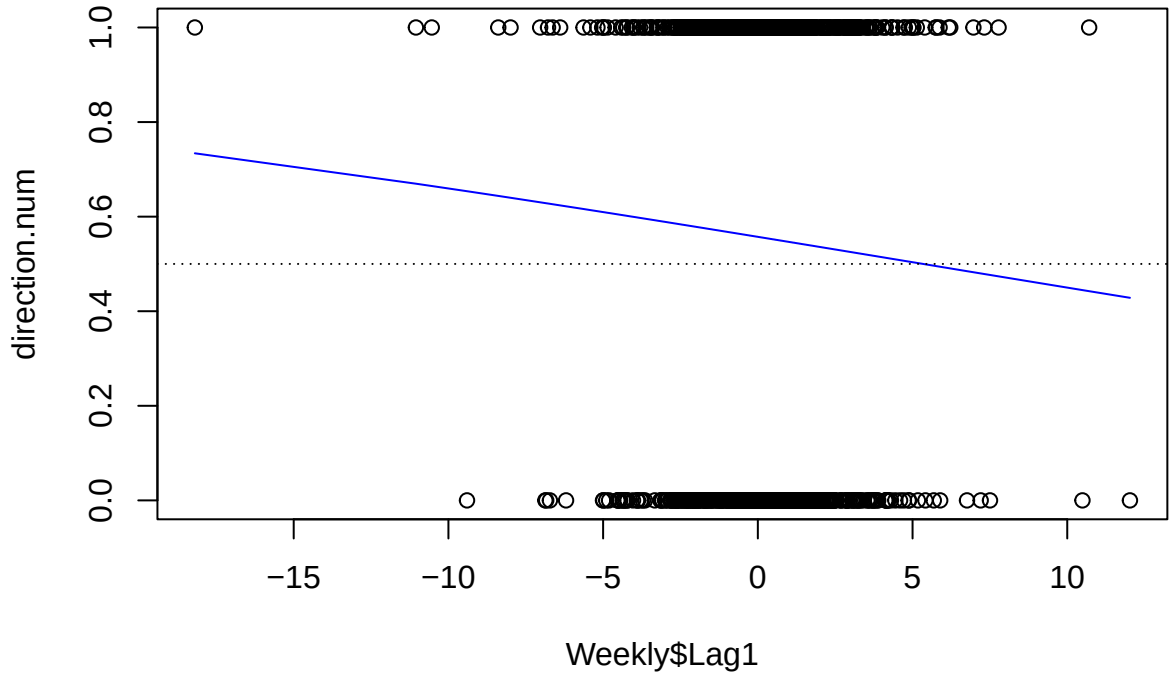
Question 5

Calculate the fitted probabilities for the entire training data (Do not print all values; print only the first few). Draw a scatter plot of the data and add the fitted probability curve. Hint: use type = "response" in the predict() function.

```
prob = predict(log_reg, newdata = Weekly, type = "response")
head(prob)
```

```
##         1         2         3         4         5         6
## 0.5486092 0.5601786 0.5845145 0.5196650 0.5497196 0.5447404
```

```
direction.num = ifelse(Weekly$Direction == "Up", 1, 0)
par(mfrow = c(1,1))
plot(Weekly$Lag1, direction.num)
sorted.Lag1 = sort(Weekly$Lag1, index.return = TRUE)
lines(sorted.Lag1$x, prob[sorted.Lag1$ix], col = "blue")
abline(h = 0.5, lty = "dotted")
```



Question 6

Use a 0.5 threshold of the probability to calculate the fitted Direction (Up or Down). Determine how many observations were correctly or incorrectly classified. What percentage of market movement is correctly predicted? Hint: get the confusion matrix using table(fitted_Direction, Obs_Direction).

Answer

```
fitted_direction_prob = ifelse(prob > 0.5, "Up", "Down")
table(fitted_direction_prob, Weekly$Direction)
```

```
##
## fitted_direction_prob Down  Up
##                  Down    8  10
##                  Up    476 595
```

```
(correct_percent = mean(fitted_direction_prob == Weekly$Direction)*100)
```

```
## [1] 55.3719
```

Question 7

Use the data set to perform a linear discriminant analysis (LDA) with Direction as the response and Lag1 as a predictor. Use the summary function to print the results.

Answer

```
library(MASS)
```

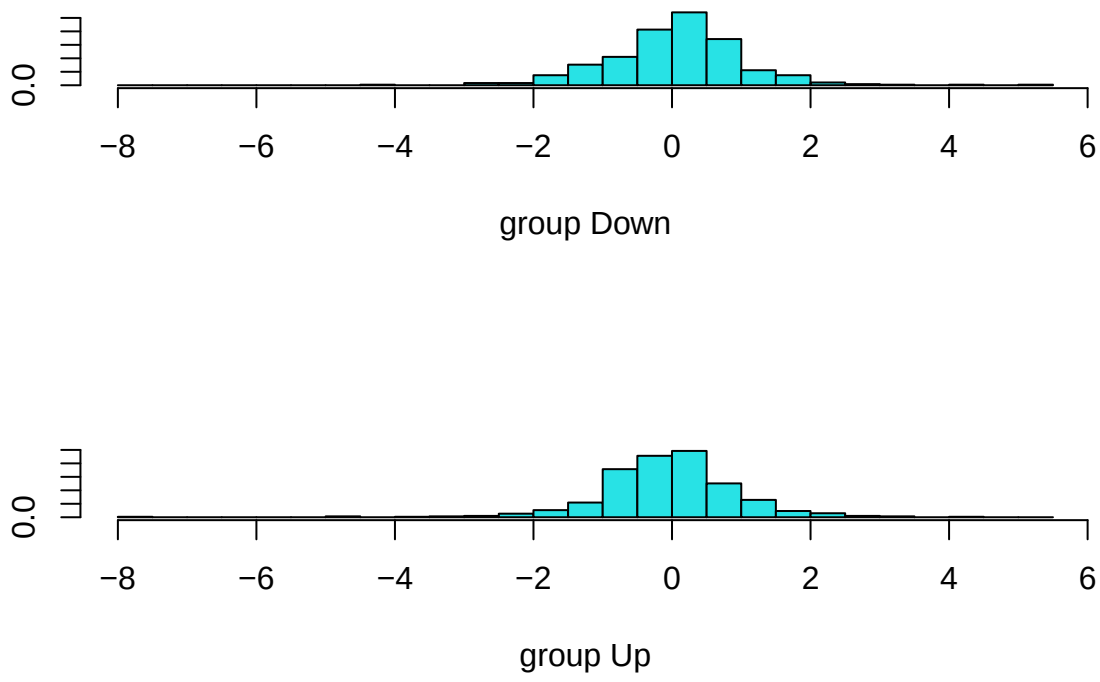
```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:ISLR2':
##
##     Boston
```

```
lda_fit = lda(Direction ~ Lag1, data = Weekly)
lda_fit
```

```
## Call:
## lda(Direction ~ Lag1, data = Weekly)
##
## Prior probabilities of groups:
##      Down        Up
## 0.4444444 0.5555556
##
## Group means:
##            Lag1
## Down 0.28229545
## Up   0.04521653
##
## Coefficients of linear discriminants:
##           LD1
## Lag1 0.424602
```

```
plot(lda_fit)
```



Question 8

Similar to the logistic regression, determine how many observations were correctly or incorrectly classified. What percentage of market movement is correctly predicted? Hint: use the predict() function to the LDA fit. The output is a list, where "class" gives the predicted Y using a 0.5 threshold.

Answer

```
lda_pred = predict(lda_fit)
table(lda_pred$class, Weekly$Direction)
```

```
##
##         Down  Up
##   Down    8  10
##   Up    476 595
```

```
(mean(lda_pred$class == Weekly$Direction)*100)
```

```
## [1] 55.3719
```

Question 9

Use the full data set to perform a logistic regression with Direction as the response and the five lag variables plus Volume as predictors. Use the summary function to print the results. Do any of the predictors appear to be statistically significant? If so, which ones? Check the performance as before.

Answer

```
log_reg_2 = glm(Direction ~ . - Year - Today, data = Weekly, family = "binomial")
summary(log_reg_2)
```

```
##
## Call:
## glm(formula = Direction ~ . - Year - Today, 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
```

Yes, Lag2 is statistically significant.

```
log_reg_2_pob = predict(log_reg_2, type = "response")
log_reg_2_pred = ifelse(log_reg_2_pob > 0.5, "Up", "Down")
log_reg_2_pred = as.factor(log_reg_2_pred)
table(log_reg_2_pred, Weekly$Direction)
```

```
##
## log_reg_2_pred Down  Up
##           Down   54  48
##           Up    430 557
```

```
(mean(log_reg_2_pred == Weekly$Direction)*100)
```

```
## [1] 56.10652
```

Question 10

Use the full data set to perform an LDA with Direction as the response and the five lag variables plus Volume as predictors. Check the performance of the LDA.

Answer

```
lda_fit_2 = lda(Direction ~ . - Year - Today, data = Weekly)
lda_fit_2
```

```
## Call:
## lda(Direction ~ . - Year - Today, data = Weekly)
##
## Prior probabilities of groups:
##      Down        Up
## 0.4444444 0.5555556
##
## Group means:
##            Lag1        Lag2       Lag3      Lag4      Lag5   Volume
## Down 0.28229545 -0.04042355 0.20764669 0.2000207 0.1878347 1.608536
## Up   0.04521653  0.30428099 0.09885124 0.1024562 0.1015388 1.547483
##
## Coefficients of linear discriminants:
##                LD1
## Lag1   -0.21451867
## Lag2    0.30090869
## Lag3   -0.08015487
## Lag4   -0.14217986
## Lag5   -0.07271067
## Volume -0.12269898
```

```
lda_fit_2_pred = predict(lda_fit_2)
table(lda_fit_2_pred$class, Weekly$Direction)
```

```
##
##        Down  Up
##   Down   52  46
##   Up    432 559
```

```
(mean(lda_fit_2_pred$class == Weekly$Direction)*100)
```

```
## [1] 56.10652
```