STAT 435 HW 3

Pat McCornack

2022-11-03

Q1 - Question 1 of Chapter 4 of textbook

$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$

$1 - p(x) = 1 - \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$

$= \frac{1+e^{\beta_0+\beta_1 x}-e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$

$= \frac{1}{1+e^{\beta_0+\beta_1 x}}$

$\frac{p(x)}{1-p(x)} = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} * \frac{1+e^{\beta_0+\beta_1 x}}{1}$

$\frac{p(x)}{1-p(x)} = e^{\beta_0+\beta_1 x}$

Q2 - Question 2 of Chapter 4 of textbook

Q3 - Question 5 of Chapter 4 of textbook

a.

We would expect LDA to perform better on both the training set and test set if the Bayes decision boundary is linear.

b.

For a non-linear Bayes decision boundary we would expect QDA to perform better on the training and test sets

c.

We would expect the predication accuracy of QDA relative to LDA to remain unchanged as the sample size n increases because the suitability of QDA over LDA is dependent on the nature of the decision boundary and not the sample size.

d.

False. While we could achieve a superior error rate using QDA on the training data because of its flexibility this wouldn't necessarily translate to a better error rate on the test set. The higher flexibility makes it more likely that QDA would pick up on anomalies in the training set and lead to over-fitting.

Q4 - Question 6 of Chapter 4 of textbook

a.

The probability that a student that studied 40 hours and had a GPA of 3.5 receives an A in the class is **37.75%**.

```
B0 <- -6
B1 <- 0.05
B2 <- 1
hours <- 40
GPA <- 3.5

prob.A <- (exp(B0 + B1*hours + B2*GPA)) / (1 + exp(B0 + B1*hours + B2*GPA))
prob.A
```

```
## [1] 0.3775407
```

b.

The same student would need to study 50 hours to have a 50% probability of getting an A in the class.

```
(log((.5/(1-.5))) - (B0 + B2*GPA)) / .05
```

```
## [1] 50
```

Q5 - Question 7 of Chapter 4 of textbook

There is a **75.2%** chance that the company issues a dividend.

```
# x is last year's % profit
div_mean <- 10
no_div_mean <- 0
var <- 36
x <- 4

# 80% of companies issued dividends

# Density }

f1_density <- 1/(sqrt(2*pi*var))*exp((-(x-div_mean)^2)/(2*var))

f2_density <- 1/(sqrt(2*pi*var))*exp((-(x-no_div_mean)^2)/(2*var))

prob <- (.8 * f1_density)/(.8 * f1_density + .2 * f2_density)
prob
```

```
## [1] 0.7518525
```

Q6 - Question 10 of Chapter 4 (All parts but (g))

Q7 - Question 13 of Chapter 4 (Use LDA, QDA, logistic regression, regularized logistic regression)

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

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

```
library(glmnet)
```

```
## Loading required package: Matrix
```

```
## Loaded glmnet 4.1-4
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
```

```
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.2
## v purrr   0.3.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x tidyr::expand() masks Matrix::expand()
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x tidyr::pack()   masks Matrix::pack()
## x tidyr::unpack() masks Matrix::unpack()
```

```
data(Weekly)
names(Weekly)
```

```
## [1] "Year"      "Lag1"      "Lag2"      "Lag3"      "Lag4"      "Lag5"
## [7] "Volume"    "Today"     "Direction"
```

a.

First I look at the summary statistics and scatterplot matrix for the data. No patterns become immediately apparent through the summary statistics, but it's worth noting that there are more weeks with positive returns than negative.

When looking at the scatterplot matrix most pairs of variables yield uncorrelated clouds and have matching very low correlation coefficients. The only notable exception is the Year vs. Volume data where we see an upward trend. This plot is blown up for further inspection below. I also plotted the Volume histogram because of its skewed distribution.

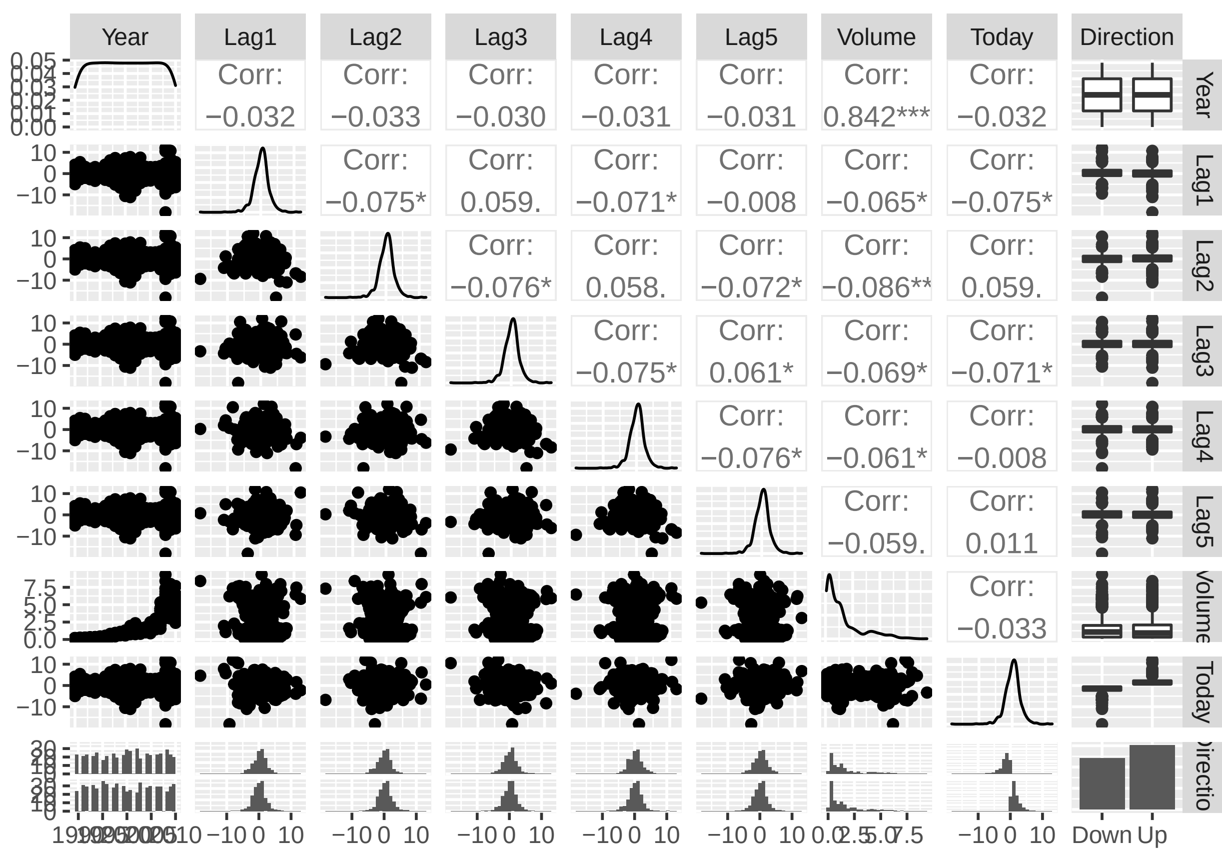
```
summary(Weekly)
```

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

```
ggpairs(Weekly)
```

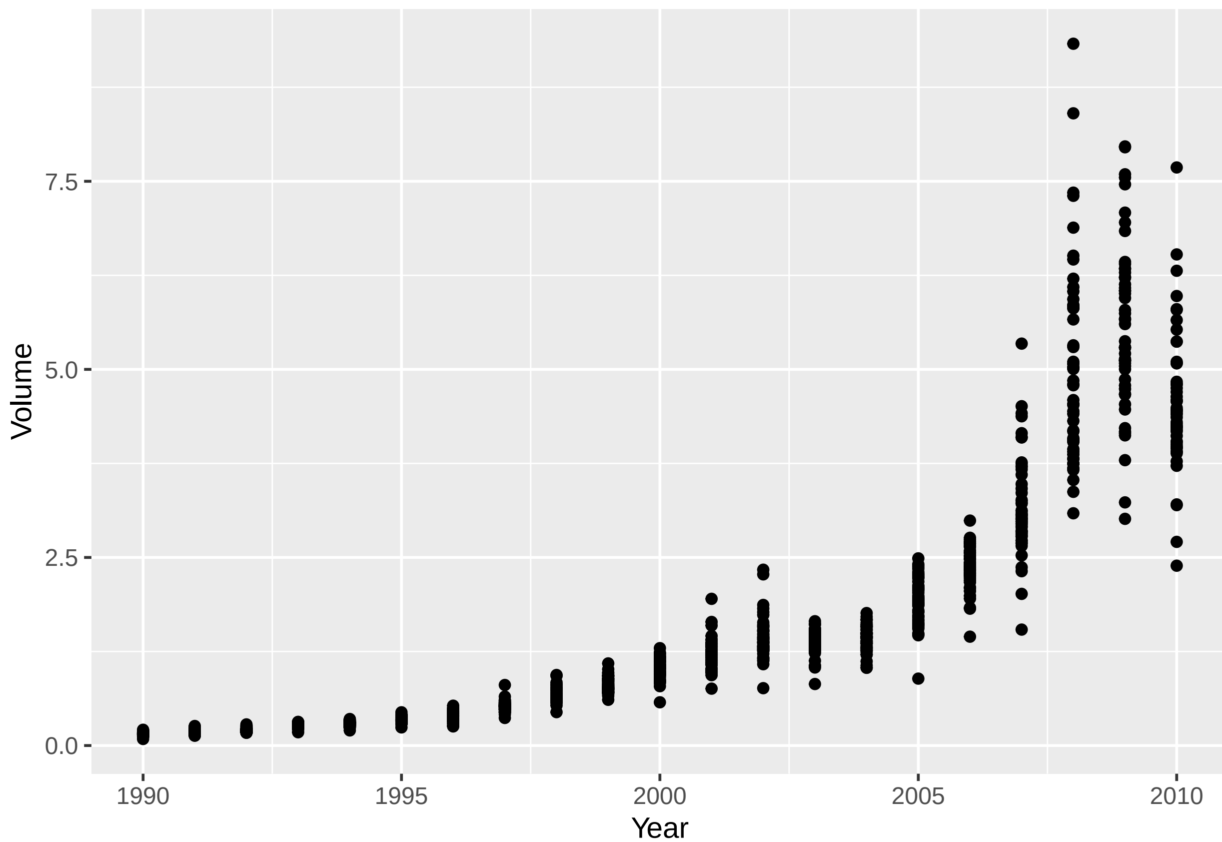
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using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



Notable points about the Year vs. Volume scatterplot are the overall upward trend and the increase in spread over time. We can also see that the drop in volume after 2008 corresponds to the recession at that time.

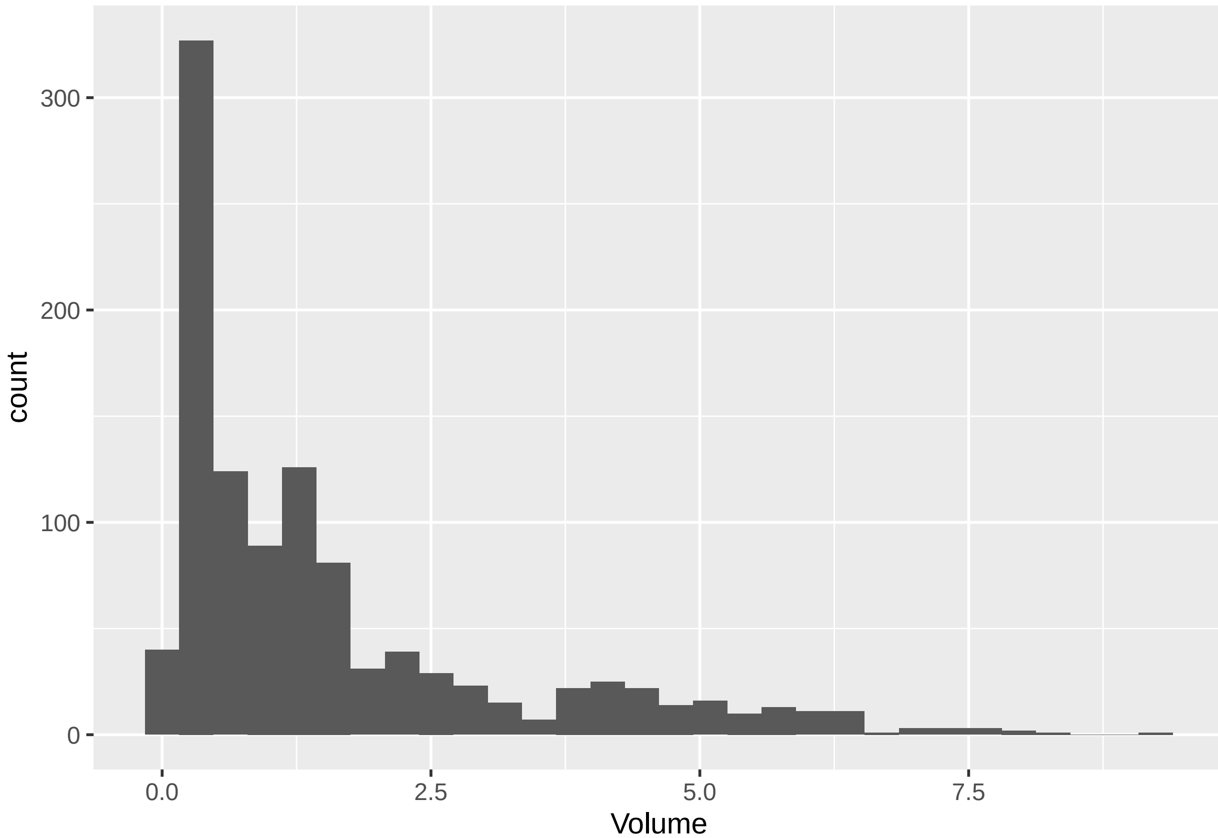
The volume histogram shows a right-tailed distribution. This makes sense since very high volume weeks would be an exception with lower volume weeks being more typical.

```
ggplot(data = Weekly, aes(x=Year, y=Volume)) +
  geom_point()
```




```
ggplot(data = Weekly, aes(x=Volume)) +
  geom_histogram()

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



b.

According to the model the p-value of Lag2 is less than .05 which indicates it may be statistically significant.

```
log.mod <- glm(Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + Volume, family = 'binomial', data = Weekl

summary(log.mod)
```

```
##
## 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
```

c.

The model correctly predicts 56% of the observations. This is only marginally better than a random classifier. It makes more false-positive predictions where it predicts the direction will be up when the true value is down than false-negative predictions.

```
probs <- predict(log.mod, Weekly, type = "response")
n <- dim(Weekly)[1]
pred <- rep("Down", n)
pred[probs > 0.5] <- "Up"

table(pred, Weekly$Direction)
```

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

```
correct <- (557 + 54) / (557 + 54 + 48 + 430)
correct
```

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

d.

The logistic model correctly predicted 62.5% of observations from the test dataset.

```
train <- filter(Weekly, Year <= 2008)
test <- filter(Weekly, Year > 2008)

lag2.log.mod <- glm(Direction ~ Lag2, data = Weekly, family = 'binomial')
prob <- predict(lag2.log.mod, test, type = 'response')
n <- dim(test)[1]
pred <- rep("Down", n)
pred[prob > 0.5] <- "Up"

table(pred, test$Direction)
```

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

```
correct <- (9 + 56) / (9 + 5 + 34 + 56)
correct
```

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

e.

The prediction accuracy using the test dataset and LDA is 62.5%.

```r
library(MASS)
```

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

## The following object is masked from 'package:dplyr':
##
##     select
```

```r
# Prediction using Linear Discriminant Analysis
lda.mod <- lda(Direction ~ Lag2, data = train)
lda.pred <- predict(lda.mod, test)

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

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

```r
(9 + 56) / (9 + 5 + 34 + 56)
```

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

f.

The QDA model predicted 58.7% of observations correctly.

```r
qda.mod <- qda(Direction ~ Lag2, data = train)
qda.pred <- predict(qda.mod, test)

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

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

```r
61 / (43 + 61)
```

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

g.

knn prediction accuracy of the test dataset is 50%.

```r
library(class)
```

```
## Warning: package 'class' was built under R version 4.2.2
```

```r
train.predictors <- dplyr::select(train, Lag2)
train.response <- dplyr::select(train, Direction)

test.predictors <- dplyr::select(test, Lag2)
test.response <- dplyr::select(test, Direction)
```

```r
knn.pred <- knn(train.predictors, test.predictors, train.response[,1], k=1)


table(knn.pred, test.response[,1])
```

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

```r
(21 + 31)/(21 + 30 + 22 + 31)
```

```
## [1] 0.5
```

Q8

Read in data from csv

```r
dat <- read.csv("./Hw3data.csv")
```

Implement LDA and QDA using data

```r
lda.fit <- lda(response ~ ., data = dat)
lda.pred <- predict(lda.fit, dat)
table(lda.pred$class, dat$response)
```

```
##
##      0  1
##   0 29 25
##   1 21 25
```

```r
qda.fit <- qda(response ~ ., data = dat)
qda.pred <- predict(qda.fit, dat)
table(qda.pred$class, dat$response)
```

```
##
##      0  1
##   0 47  4
##   1  3 46
```