Assignment 3

Brian Morales

September 17, 2022

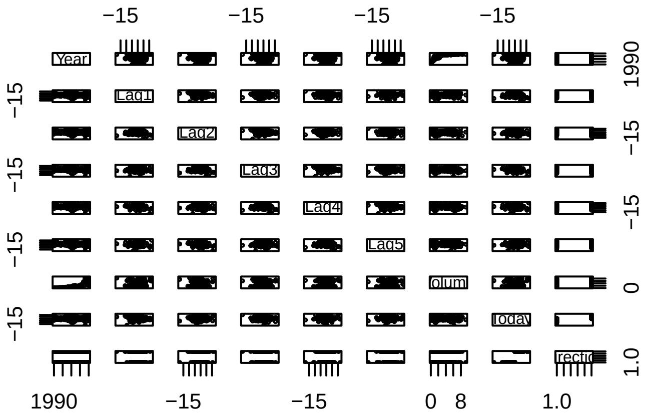
13a.

```
library(readr)
Weekly <- read_csv("~/Desktop/Fall-2022/Stats-Learning/ALL-CSV-FILES/Weekly.csv", show_col_types = FALS
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
##  Length:1089
##  Class :character
##  Mode  :character
##
##
##
```

```
Weekly$Direction = as.factor(Weekly$Direction)
```

```
plot(Weekly, cex = 0.3)
```



There appears to be an exponential pattern with `Volume` and `Year` and every other variable seems random. We can see that `Volume` is increasing over time, meaning the the average number shares increased from 1990 t0 2010.

13b.

```
glm.fits <- glm(
    Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + Volume,
    data = Weekly, family = binomial
  )
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
```

Yes, only one predictor seems to be statistically significant, `Lag2`. `Lag2` has a significant code of 0.01 so its not very significant either. The positive coefficient in `Lag2` is implying that if the weekly market had a positive return yesterday, than it is likely to go up today.

13c

```
glm.probs <- predict(glm.fits, type = "response")

glm.pred <- rep("Down", nrow(Weekly))
glm.pred[glm.probs > .5] = "Up"
table(glm.pred, Weekly$Direction)
```

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

```
(583 + 23)/1089
```

```
## [1] 0.5564738
```

The confusion matrix tells us what we predicted correctly and what we predicted incorrectly. The diagonals of the matrix shows us the number of correct predictions and the off-diagonals indicates what we incorrectly predicted. Here our model correctly predicted that our market would go down 51 days and up for 555, total of 606 correct predictions. We incorrectly predicted up when the market was actually down 433 days and down when is was actually up 50 days a total of 483 incorrect predictions. Our model is working a little better than random guessing, however, this can be deceptive because we are training and testing on the same dataset. Lets train on part of the data and test on the remaining held out data.

13d

```
train <- (Weekly$Year < 2009)
Weekly.2008 <- Weekly[!train, ]
Direction.2010 <- Weekly$Direction[!train]

glm.fits <- glm(
    Direction ~ Lag2,
    data = Weekly, family = binomial, subset = train
  )
glm.probs <- predict(glm.fits, Weekly.2008,
    type = "response")
```

```
glm.pred <- rep("Down", nrow(Weekly.2008))
glm.pred[glm.probs > .5] <- "Up"
table(glm.pred, Direction.2010)
```

```
##          Direction.2010
## glm.pred Down Up
##     Down    9  5
##     Up     34 56
```

```
65/104
```

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

```
39/104
```

```
## [1] 0.375
```

As before the diagonals of the matrix displays what our model predicted correct and the off-diagonals are what it predicted incorrectly. Notice that our accuracy is 62.5 which is better than our previous model. This suggest that our model is predicting better than guessing at random.