# Predicting Weekly Returns with Logistic Regression, LDA, and QDA

Data Analysis Assignment Lesson 7

David Chen

March 3rd, 2020

# Introduction

In this report, we will be using the `Weekly` dataset to predict if the S&P 500 markets had positive or negative returns for the week. Since we are trying to predict a binary response ('Up' or 'Down' for weekly returns), we will be using the following models: logistic regression, linear discriminant analysis (LDA), and quadratic discriminant analysis (QDA). We ultimately observe that `Lag2`, the percentage return from two weeks ago, is the only significant variable, and that a logistic regression / LDA model with just that predictor performs the best.

# Data

For this analysis, we will be analyzing the `Weekly` dataset. As listed from the online documentation, this dataset provides the S&P 500 weekly percentage returns from 1990 to 2010. The following variables are included:

1. `Year` - Year the data was observed.
2. `Lag1` - Percentage return for the previous week
3. `Lag2` - Percentage return from two weeks ago
4. `Lag3` - Percentage return from three weeks ago
5. `Lag4` - Percentage return from four weeks ago
6. `Lag5` - Percentage return from five weeks ago
7. `Volume` - Volume of shares traded
8. `Today` - Percentage return for the week.
9. `Direction` - Whether the market had positive ('Up') or negative ('Down') returns for the week.

## Exploratory Data Analysis

We begin by examining a summary of all the variables. From the summary statistics, we observe that all the `Lag` variables are fairly consistent with each other. We can also note the distribution of `Direction`, how 55.6% of the data is `Up`.

```
summary(Weekly)
```

```
##       Year           Lag1               Lag2               Lag3         
##  Min.   :1990   Min.   :-18.20   Min.   :-18.20   Min.   :-18.20  
##  1st Qu.:1995   1st Qu.: -1.15   1st Qu.: -1.15   1st Qu.: -1.16  
##  Median :2000   Median :  0.24   Median :  0.24   Median :  0.24  
##  Mean   :2000   Mean   :  0.15   Mean   :  0.15   Mean   :  0.15  
##  3rd Qu.:2005   3rd Qu.:  1.40   3rd Qu.:  1.41   3rd Qu.:  1.41  
##  Max.   :2010   Max.   : 12.03   Max.   : 12.03   Max.   : 12.03  
##       Lag4               Lag5             Volume          Today        
##  Min.   :-18.20   Min.   :-18.20   Min.   :0.09   Min.   :-18.20  
##  1st Qu.: -1.16   1st Qu.: -1.17   1st Qu.:0.33   1st Qu.: -1.15  
##  Median :  0.24   Median :  0.23   Median :1.00   Median :  0.24  
##  Mean   :  0.15   Mean   :  0.14   Mean   :1.57   Mean   :  0.15  
##  3rd Qu.:  1.41   3rd Qu.:  1.40   3rd Qu.:2.05   3rd Qu.:  1.40  
##  Max.   : 12.03   Max.   : 12.03   Max.   :9.33   Max.   : 12.03  
##  Direction
##  Down:484
##  Up  :605
```

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

Next, we examine a correlation plot between all the variables except Distribution (given that it is binary). From Figure 1, we can clearly see that only `Volume` and `Year` has a large correlation coefficient, the rest of the values seemingly below 0.1. Investigating further, we see in Figure 2 that `Volume` and `Year` appear to have an exponential relationship, and thus a log transformation may be beneficial. We examine this relationship in Figure 3, applying a log transformation to Volume and comparing it to `Direction`, the variable we will attempt to predict. This transformation does seem to be beneficial.

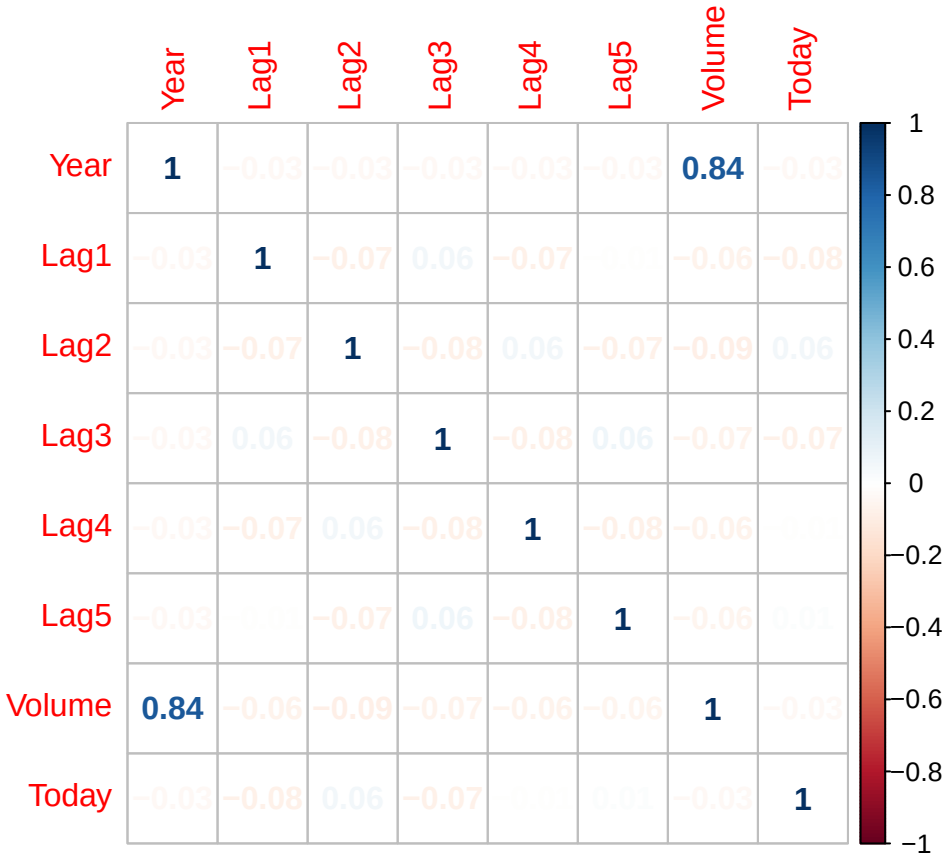


Figure 1: Weekly Dataset Correlations

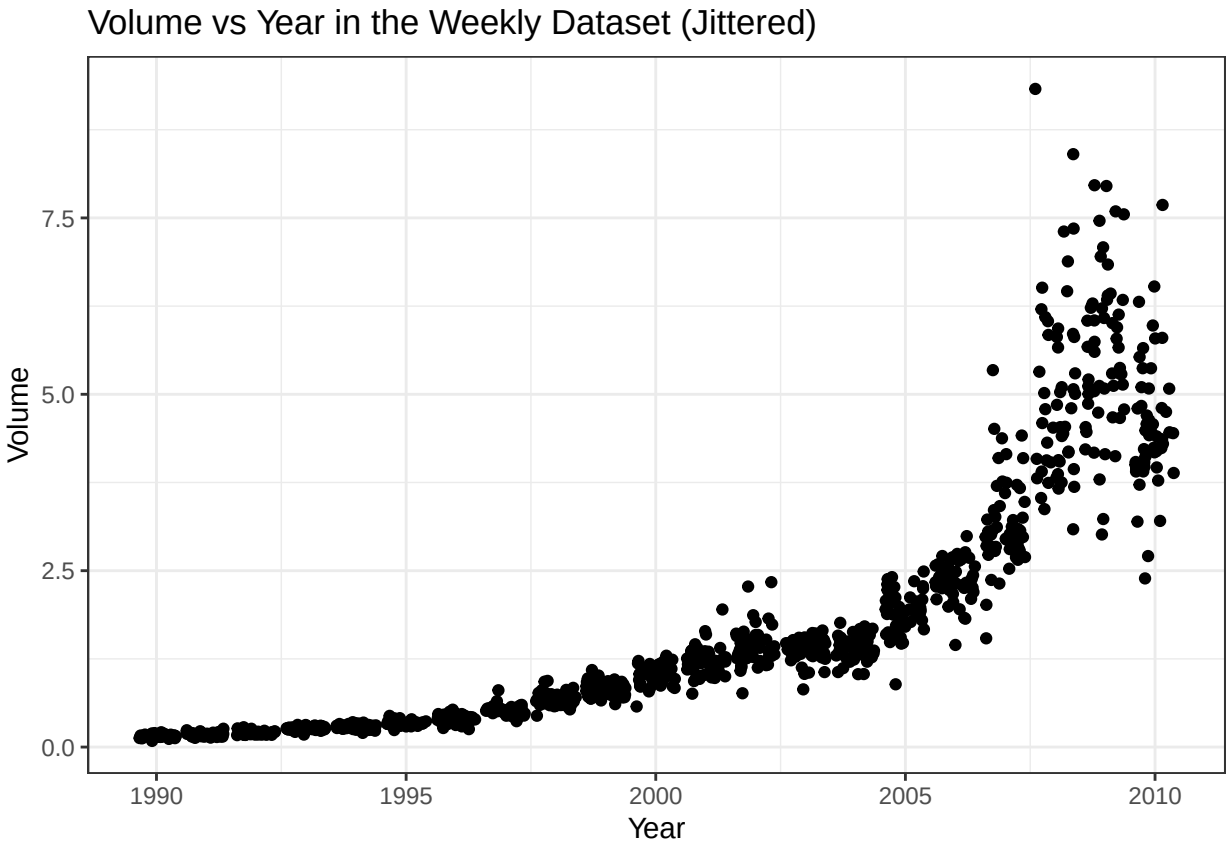


Figure 2: Volume vs Year Scatterplot

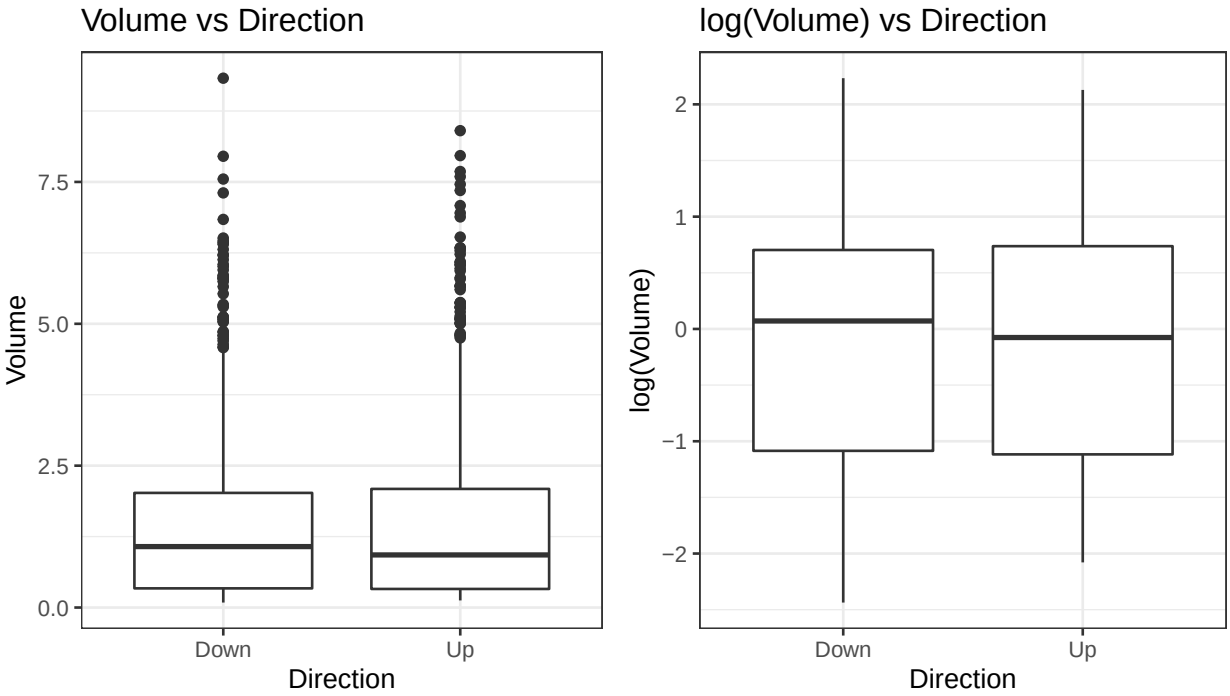


Figure 3: Volume vs Direction boxplots

# Analysis

*Post-submission note: Specificity, sensitivity, and additional details are described in Data Analysis Assignment 9 - Classification: Nonparametric Methods*

## Logistic Regression on the Full Dataset

We begin by conducting a logistic regression for `Direction` using the full dataset. We use the predictors `Lag1` through `Lag5`, and `Volume`, applying a log transformation to `Volume` (as explained previously). From this model. we observe in the summary statistics below that only `Lag2` appears to be significant at a 95% level of significance.

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
##     log(Volume), family = binomial, data = Weekly)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.692  -1.260   0.993   1.085   1.466
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.2256     0.0622    3.62  0.00029 ***
## Lag1         -0.0413     0.0264   -1.57  0.11758
## Lag2          0.0583     0.0268    2.18  0.02943 *
## Lag3         -0.0161     0.0266   -0.60  0.54621
## Lag4         -0.0279     0.0264   -1.06  0.29122
## Lag5         -0.0146     0.0264   -0.55  0.58043
## log(Volume)  -0.0513     0.0561   -0.92  0.35999
## ---
## 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: 1485.9  on 1082  degrees of freedom
## AIC: 1500
##
## Number of Fisher Scoring iterations: 4
```

Using this full model, we calculate the confusion matrix and the overall fraction of correct predictors. This confusion matrix indicates to us that the model predominately predicts 'Up' for every observation, leading to many false positives (predicting 'Up' when the truth is 'Down'). We also observe that the fraction of correct predictions is **0.561**. This is only slightly better than the naive model of predicting 'Up' for every observation (0.556 correct).

Table 1: Full Data Logistic Regression Confusion Matrix

| | True Down | True Up |
|---|---|---|
| Down | 59 | 52 |
| Up | 425 | 553 |

**Splitting the Data Into Training and Testing Sets**

Although including the full dataset allows for the most training data, we want to check how our models respond to new observations. Thus, moving forward, we train the data solely on the years 1990-2008, leaving 2009 and 2010 as the testing set.

## Logistic Regression with Training / Testing Split

Since `Lag2` was the only significant variable previously, we begin by attempting to build a model with `Lag2` as the only predictor. Applying the model to the testing set, we observe that the confusion matrix behaves similarly to the previous model, where 'Up' predictions are over represented.

We can also note that the overall fraction of correct predictions is **0.625**, higher than the naive model (predict 'Up' every time) result of 0.587.

Table 2: Logistic Regression Confusion Matrix

| | True Down | True Up |
|---|---|---|
| Down | 9 | 5 |
| Up | 34 | 56 |

## Linear Discriminant Analysis (LDA)

Next, we perform a linear discriminant analysis (`Lag2` as the only predictor). As expected, the confusion matrix and proportion of correct predictions is identical to that of the logistic regression. Thus, the proportion of accurate predictions is the same at **0.625**.

Table 3: LDA Confusion Matrix

| | True Down | True Up |
|---|---|---|
| Down | 9 | 5 |
| Up | 34 | 56 |

## Quadratic Discriminant Analysis (QDA)

Next we proceed with a QDA. Unlike LDA, a QDA allows for unique covariance matrices and creates a quadratic relationship between the predictor and response.

Again using `Lag2` as the only predictor, we observe that the QDA model follows the naive approach of simply predicting 'Up' for every observation. This yields an accuracy of **0.587**, worse than the previous two methods.

Table 4: QDA Confusion Matrix

| | True Down | True Up |
|---|---|---|
| Down | 0 | 0 |
| Up | 43 | 61 |

**Comparison**

Comparing the logistic regression, LDA, and QDA for accuracy while using only `Lag2` to predict `Direction`, we observe that the logistic regression and LDA models performed the best. Both models predicted the correct response in the 2009-2010 test data 62.5% of the time, compared to the QDA which had a value of 58.7%. We also note that the QDA model followed the naive approach of predicting 'Up' for every observation, and thus does not seem to be very reliable.

### Variable Selection, Interactions, and Transformations

In the previous models, we did not include predictors besides `Lag2`. Thus, in order to check if there will be a model with more accurate predictions, we will increase the number of variables, include interactions, and attempt to perform transformations.

To begin, we start with the logistic regression model. We include interactions between Lags that are next to each other (i.e. `Lag1` and `Lag2`, `Lag2` and `Lag3`, etc.), but no log transformations. However, this full model only accurately predicts half of the results, so we attempt to include a backwards stepwise selection (with AIC). While only 5 variables remained, the number of correct predictions barely changed (up to 0.51). However, conducting a log transformation on `Volume` resulted in a prediction accuracy of 0.606. This ultimately supports our idea that the log transformation is beneficial, but fails to reach the previous results with only `Lag2`.

Attempting to include the full interaction terms and the log transformation to the LDA and QDA model, we observe similarly poor results. The LDA model predicts 0.558 of the observations while the QDA somehow dips to 0.49.

Ultimately, the inclusion of additional interaction terms, predictors, and transformations failed to improve the prediction accuracy of the models. We maintain that the logistic regression with just `Lag2` as the predictor is the most accurate among these models.

## Conclusion

In conclusion, we observe that the logistic regression and the LDA model with just `Lag2` as the predictor is the most accurate (0.625). However, it still fails to achieve results substantially greater than the naive model (0.587), and thus one must exercise caution before attempting to apply this model. Ultimately, we would conclude that the `Lag` variables, along with `Volume`, have very little association with the weekly S&P 500 returns. Moving forward we should investigate other variables to create a more accurate model.