MODEL COMPARISON FOR THE PREDICTION OF STOCK PRICES IN THE NYSE

Victoria Switlyk

A Thesis

Submitted to the Graduate College of Bowling Green
State University in partial fulfillment of
the requirements for the degree of

MATER OF SCIENCE

August 2018

Committee:

Junfeng Shang, Advisor

John Chen

Wei Ning

ABSTRACT

Junfeng Shang, Advisor

The stock market is an integral part of investments as well as the economy as a whole. The prediction of stock prices is a exciting and challenging problem that has been considered by many due to the complexity and noise within the market as well as the potential profit that can be yielded from accurate predictions.

The purpose of this study is to construct and compare models used for the prediction of weekly closing prices for some of the top stocks in the NYSE as well as to discuss the relationship between stock prices and the predictor variables. Relationships considered in the study include that with macroeconomic variables such as the Federal Funds Rate and the M1 money supply as well as market indexes such as the CBOE Volatility Index, the Wilshire 5000 Total Market Full Cap Index, the CBOE interest rate for 10-year T-notes and bonds, and NYSE commodity indexes including XOI and HUI.

Models are built using methods of regression analysis and time series analysis. Models are analyzed and compared with one another by considering their predictive ability, accuracy, fit to the underlying model assumptions, and usefulness in application. The final models considered are a pooled regression model considering the median weekly closing price across all stocks, a varying intercept model considering the weekly closing price for each individual stock, and an ARIMA time series model that predicts the median weekly closing stock price based on past prices.

ACKNOWLEDGMENTS

I would like to acknowledge and thank my advisor Junfeng Shang for all of her help, advice, and mentoring throughout my studies. She is an amazing person to work with, and I cannot thank her enough. I would also like to thank my remaining committee, John Chen and Wei Ning, for their willingness to provide their assistance and expertise and for the dedication of their time.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1 DATA DESCRIPTION

1.1 Introduction to Data

When it comes to creating an investment portfolio to build up held assets, there are several different asset classes to choose from which include bonds, cash equivalents, and equities. Equities, or stocks, are the most volatile type of these asset classes but also have the ability create a large profit. The key is to know when to buy and when to sell. This is why understanding patterns in equity prices is vital to those who wish to invest in the stock market.

The purpose of this study is to estimate the values of time series data within each selected stock of the New York Stock Exchange (NYSE). The study focuses on various stocks within a certain exchange rather than in a certain index such as the S&P 500, because we want to keep the values estimated to be consistent within the same market. Stocks within the S&P 500 are sold in different exchanges which can affect the estimated price of that stock. Because of this, the model provides insight exclusively for participation within the NYSE. Out of all possible exchanges, the NYSE was chosen in this model since it is the largest stock exchange in the world.

Stock data is time series since it contains prices over time where each time point is related to the previous time point. It is possible to look at stock prices over different time intervals such as hourly, daily, weekly and monthly. This study looks at weekly time intervals. The reason is that when looking at daily stock data, there are some days that are missing. Excluding these data points would make the intervals between time points unequal, which would not be appropriate for time series analysis. A solution to this problem would be to fill in the missing data. However, this study chooses to instead use weekly time points since there can be a great amount of variation among stock prices per day. This variation could be largely due to white noise which we are not necessarily interested in tracking. This study is interested in looking at trends over longer periods of time.

There are approximately 2,800 companies that have equities listed in the NYSE. Instead of

trying to fit a model that encompasses all of these stocks, this model focuses on a subset of stocks within the NYSE. There are two options for doing this. The first is to take a random sample of stocks and create a model looking at the stocks within that sample. The second is to take a selection of the most common or most popular stocks within the exchange. This study follows the latter option. Using a model to create predictions for more popular stocks makes the study more relevant in application for those who participate within the NYSE. On the other hand, a random selection could include more obscure stocks, although which could provide a more diverse look at equites, would be less practical for application. The selection of common stocks that are modeled in this study are those within the NYSE 100.

1.2 New York Stock Exchange 100

The New York Stock Exchange 100, or NYSE 100, composes of a list of very promising, high achieving, and popular stocks. A stock must be well established to become a part of the NYSE 100. Since this group consists of successful and popular stocks, it provides a sample of stocks for this model that is relevant in application. This study looks at weekly closing stock prices in the NYSE 100 for each Friday from January 1, 2000 through December 23, 2017. This model includes 85 stocks from the NYSE 100. There are some stocks within the NYSE 100 that are missing from this study because they did not have the full range of data between these two dates. With this data for individual stocks, this model includes three different variables.

The first variable related to the NYSE 100 is the variable which this model is a predictor for. For each of the 85 stocks, this study contains data for the closing price for each Friday from January 1, 2000 through December 23, 2017. The goal of this model is to be able to predict the closing price for a stock at a certain time point. For each stock, there are 939 data points which were retrieved from Yahoo! Finance. This study defines the closing price as Y_{it} for each stock $i = 1, .., 85$ for each week $t = 1, ..., 939$. The other two variables describing each of these stocks are used as variables to predict the closing price.

One variable that is used as a predictor is the volume of stock sold each week. For each Friday, this study looks at the data for the volume of each stock sold at each time point. The volume sold

in the previous week is used to predict the volume sold in the current week. The data was retrieved from Yahoo! Finance. This study defines the volume sold as V_{it} for each stock $i = 1, ..., 85$ for each week $t = 1, ..., 939$.

The second predictor variable is how stocks within the NYSE are categorized by sector. This is a way that stocks are grouped with other stocks that cater to the same sections of the economy and market. The 85 stocks that are modeled in this study belong to 7 distinct sectors. 20 stocks belong to the Consumer Goods (CG) sector. This sector includes both cyclical and non-cyclical goods and services. Industry groups within this sector include automobile manufacturers, home construction, leisure goods and services, textiles and apparel, entertainment, broadcasting, retail, food, consumer services, cosmetics, and household products. 18 stocks belong to the Basic Materials (BSC) sector. Industry groups within this sector include chemicals, mining, metals, forest products, and paper. 16 stocks belong to the Financial (FIN) sector. Industry groups within this sector include banks, financial services, and insurance. 12 stocks belong to the Industrial (IDU) sector. Industry groups within this sector include construction, and industrial goods and services such as building materials, industrial equipment, aerospace, electrical components, and industrial transportation. 10 stocks belong to the Healthcare (HCR) sector. Industry groups within this sector include biotechnology, healthcare providers, medical products, and pharmaceuticals. 5 stocks belong to the Technology (TEC) sector. Industry groups within this sector include communications technology, technology services, technology hardware and equipment, and software. The last 4 stocks belong to the Utilities (UT) sector which includes electric, gas, and water utilities. Sectors will be treated as categorical variables in the model and are discussed later in this study.

1.3 Macroeconomic Variables

Changes in stock prices are connected to several aspects of the economy, including those at the highest levels. Macroeconomic variables are those features of a national or international economy that describe the state of the market as a whole. These variables tend to be recorded monthly or annually rather than weekly. Because of this, macroeconomic variables are more useful for observing trends over long periods of time rather than accounting for variation in the short run.

Within this modeling process, we will consider several of these factors to see any possible relations between the changes in equities and the economy at a macroeconomic level.

The Federal Funds Rate (DFF) is the rate of interest in which institutions exchange funds held at Federal Reserve Banks. Institutions may lend portions of balances and funds to other institutions. This interest rate is influenced by the Federal Reserve, and decisions on the rate are determined by the state of the market. Changes in the interest rate influences spending. If the funds rate is high, exchange and spending is deterred resulting in decreased stock prices. On the other hand, if the funds rate is low, the cost of exchanging funds and spending becomes cheaper. This encourages borrowing and spending, leading to increased stock prices. The data obtained for the funds rate is monthly, however there are cases where consecutive months have the same interest rate when the rate is not changed. The data was originally released by the Board of Governors of the Federal Reserve System and was retrieved from the Federal Reserve Bank of St. Louis (FRED). This study defines DFF as DFF_t for each week $t = 1, ..., 939$. (Online, a)

M1 is the entire supply of physical money in the United States and is composed of federal notes and coins as well as some accounts such as demand deposits. M1 is also called narrow money because it includes only physical money and liquid assets that can be easily converted to physical money. M1 will always increase over time due to inflation. The data obtained is weekly, was originally released by the Board of Governors of the Federal Reserve System, and was retrieved from FRED. The units for this data is in billions of U.S. dollars and is seasonally adjusted. This study defines M1 as $M1_t$ for each week $t = 1, .., 939$. (Online, b)

1.4 Market Indexes

This model will look at different indexes that are used to illustrate different aspects of the equity market. Unlike the macroeconomic factors which give a broad look on the state of the economy, these indexes can be used to look at the state of the stock market specifically.

The Chicago Board Options Exchange (CBOE) created the Volatility Index (VIX) to measure market expectations of volatility in stock index prices. The VIX serves as a way to measure market risk. A low VIX index indicates low expected volatility meaning that stock prices are not expected

to change quickly. A high VIX index indicates a high expected volatility meaning that stock prices are expected to change quickly. A higher amount of volatility also indicates increased uncertainty and risk in the market, which can deter investments and spending. The data obtained is weekly. The VIX data was originally released by the CBOE and was retrieved from FRED. This study defines VIX as VIX_t for each week $t = 1, ..., 939$. (Online, c)

The CBOE also has an index that measures the interest rate for 10-year T-notes and bonds. TNX is the ticker symbol for this index. Equities or stocks are a type of asset class along with bonds. Since both equities and bonds are used in financial portfolios, it is possible that changes in bond rates can affect whether or not a person decides to invest in stocks or bonds. The data obtained is weekly and was retrieved from Yahoo! Finance. This study defines TNX as TNX_t for each week $t = 1, ..., 939$.

The Wilshire 5000 Total Market Full Cap Index is one of several Wilshire Indexes. The Total Market Index is known as being a comprehensive measure of equity in the U.S. market by including the average price of nearly 5000 different stocks from various exchanges. The data obtained is weekly, was originally published by Wilshire Associates, and was retrieved from FRED. This study defines the Wilshire 5000 as WIL_t for each week $t = 1, ..939$. (Online, e)

Stock prices can also be related to prices of major commodities within the United States. The NYSE provides current prices of various commodity groups alongside the prices of their equities for reference. The NYSE and looks at three different commodity groups. The first is softs and includes goods such as coffee, cocoa, sugar, and cotton. The NYSE does not have an index for summarizing the prices of softs, so instead this model looks to the other two commodity groups. The second commodity group is energy and includes fuels such as gas and oil. In this study, the changes in prices of fuels are modeled using the NYSE ARCA Oil and Gas Index (XOI). The index provides the average prices of major oil and gas components within the market. The data is weekly and was retrieved from Yahoo! Finance. This study defines XOI as XOI_t for each week $t = 1, ..., 939$. The third and final commodity group is precious metals and includes rates for gold, silver, and platinum. In this study, the changes in prices of precious metals are modeled

by the NYSE ARCA Gold Bugs Index (HUI). The index provides the average prices of stocks in companies within the gold mining industry. The data obtained is weekly and was retrieved from Yahoo! Finance. This study defines HUI as HUI_t for each week $t = 1, ..., 939$. (Online, d)

CHAPTER 2 INTRODUCTION TO THE STUDY

2.1 Objective of the Study

The prediction of stock prices is an interesting and challenging endeavor that has been considered by economists, financial analysts, statisticians and computer scientists alike. The stock market has intrigued many due to the complexity and uncertainty of the market as well as the potential financial gain that can come from accurate predictions. A classic dream is to "make it big" on the stock market and data modelers have considered many different techniques to achieve this end.

In literature, perhaps the most common method for stock prediction is through machine learning and neural networks due to its versatile nature in using many predictor variables and its lenient model assumptions. The most common neural network is the Artificial Neural Network (ANN) see in studies such as one by Moghaddama, Moghaddamb, and Esfandyari (2016), who consider the prediction of daily NASDAQ rates using the day of the week and historical prices as inputs to produce accurate predictions. The drawback with neural networks is that they act as a sort of "black box" building relations between the stock prices and the predictors meaning that it is difficult to interpret the model and the relationships therein. So although neural networks and machine learning tend to be the most popular choice for stock predictions, this study turns to more classical methods including multiple linear regression and time series analysis to provide more meaningful interpretations.

Chang, Wang, and Zhou (2012) who study the daily stock trends using another type of neural network, the evolving partially connected neural network (EPCNN) explain that "mining stock market trend is a challenging task due to its high volatility and noisy environment". Stock prices can be very volatile especially in the short run, which is why unlike many other studies, this study will consider a longer time interval using weekly data rather than daily data to account for this noisy short term environment. Chang, Wang, and Zhou (2012) also express the strong relationship between stock trends and other outside factors which is why in this study, we consider many other

economic variables as described in Chapter 1.

This study takes advantage of the ability of classical models to provide insight in the relationship between stock prices and other predictor variables. We also consider larger time periods to account for the access noise in the short term.

Previous studies such as those by Al-Tamimi, Alwan, and Rahman (2011) and Sharif, Purohit, and Pillai (2015) consider regression analysis for the prediction of stock prices using other predictor variables even though it is understood that there is a dependent relationship within the data due to its time series nature. However, as this study will show, despite failing to meet the underlying assumption of independence, regression can still be used to statistically show the relationships between stocks and other variables that are known from an economic standpoint.

2.2 Outline of the Study

In this section we briefly discuss what you can expect to see in this study. As stated previously, the goal of this study is to compare various models used for the prediction of the weekly closing stock price for selected stocks in the NYSE.

In Chapter 1, you become familiar with the the data that was collected and is used for the modeling process as well as for testing the fit of the models. In Chapter 2, we have discussed the motivation of the study and literature related to the prediction of stock prices. Chapter 3 is where we begin to look at each model separately. This chapter is dedicated to the pooled multiple linear regression model which is based on the median weekly stock price over all indexes. Chapter 4 discusses the time series model which takes advantage of the time series nature of the data. This model also considers the pooled weekly median stock price over all indexes. Chapter 5 analyzes the third model which is also a multiple linear regression model. However, instead of considering the pooled price, this model considers each of the 85 selected stocks in the NYSE individually. This study is concluded in Chapter 6 which gives comparisons of the three models as well as discusses further considerations for modeling.

CHAPTER 3 POOLED REGRESSION MODEL

3.1 Multiple Linear Regression

Multiple linear regression is a model used create predictions based on information that is known of other variables. This study uses regression models to show how the variation of stock prices are related to our predictor variables and how these variables can be used to explain variation in prices. Since the data are time series, we consider time as one of these predictor variables. In this study we consider several regression models, however, this chapter focuses on a pooled model.

How the data are currently, there are multiple values of Y_{it} for each time period since we are considering the closing price for each individual stock for each time period. When considering each stock type, we are essentially considering a varying intercept model, or a model that considers the changes or variation among each stock type. For our first model, we will look at a pooled model which examines the top stocks of the NYSE as a whole.

To create the pooled model, we consider predicting Y_t which represents the median closing stock price for each week t. When pooling the closing price over all stock indexes, we consider the median rather than the average. The average merges the variability within the data over time while the median will retain the patterns of variation.

Furthermore, the distribution of prices at time t tends to be right skewed. This is because there are some larger and more popular companies within the NYSE 100 that have much higher stock prices than other companies. Figure 3.1 illustrates the difference in the distribution of closing prices when looking at the price for each stock for each week as opposed to considering the median price for each week. For the original price data, we see that the distribution is highly right skewed. For the median price, we see that the distribution is no longer skewed, but it does appear to be bimodal. This could be due to fluctuations in the economy over time causing prices to sit at different levels over different periods. Considering the median is a different perspective on pooling index prices compared to the calculations usually used for stock market indexes, such as the S&P

500 which considers the weighted average price which is dependent on the number of stock shares for that company.

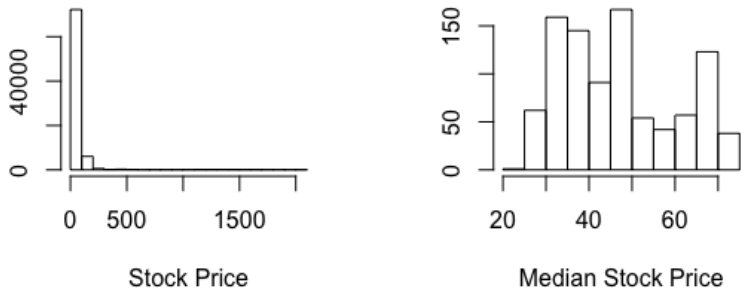

Figure 3.1: Distribution of Price

We also remember that one of our predictor variables, the volume of stock i sold at time t is also a variable that depends of the stock index. Similarly with the closing price, in this model we consider the median volume sold at time t over each stock i. In other words, we are looking at V_t rather than V_{it}.

When conducting regression modeling, we first partition the data into two parts: training and testing data. The training data is used in the process of creating the model. Once created, we use the chosen model to make predictions for the testing data in order to check the fit of our model and make sure that there is not an overfitting of the model to the training data. Since the pooled data contains one data point or observation for each week, t, the data set for this model contains 939 data points. The data are randomly partitioned into the two groups with a 70%-30% split. In other words, there are 657 observations in the training data and 282 observation in the testing data. For the purpose of creating the model, use only the training data. We will look at the testing data later in our analysis.

Before creating the model, we look at the mathematical qualities that go into the creation of a multiple linear regression model. For this model, we assume that the relationship between the median closing stock price at week t and our selected set of predictors variables, $X_{t1}, ..., X_{tk}$,

roughly follows the linear regression model

$$Y_t = \beta_0 + \beta_1 X_{t1} + ... + \beta_k X_{tk} + \epsilon_t,$$

where ϵ_t is a random variable that represents the error. We suppose $E(\epsilon_t) = 0$ such that the expected median closing stock price,

$$E(Y_t) = \beta_0 + \beta_1 X_{t1} + ... + \beta_k X_{tk},$$

is a function of our predictor variables. β_0 is the intercept and $\beta_1, ..., \beta_k$ are the slopes. All of the beta coefficients are defined as fixed and unknown parameters. The regression model uses the least squares estimates of $\beta_1, ..., \beta_k$ which are the values that minimize the residual sum of squares (RSS),

$$RSS = \sum_{t=1}^{n} \epsilon_t^2 = \sum_{t=1}^{n} (Y_t - \beta_0 - \sum_{j=1}^{k} \beta_j X_{tj})^2.$$

The notation for a multiple regression model can be simplified by writing the model in terms of vectors and matrices. We set Y as the vector of median closing prices Y_t from $t = 1, ..., 657$. X is set as the $(k+1)X657$ matrix of all K predictor variables X_{tj} from $t = 1, ..., 657$ and $j = 1, ..., k$. The first column of the matrix is a vector of all ones corresponding to the intercept. We then have β as the vector of the unknown parameters $\beta_0, ..., \beta_k$ and ϵ as the vector of error terms from $t = 1, ..., 657$. Defined with matrices, we can rewrite the model as

$$Y = X\beta + \epsilon.$$

We can also rewrite the residual sum of squares as

$$RSS = \epsilon'\epsilon = (Y - X\beta)'(Y - X\beta).$$

Finally, we will define the least squares estimates of the models beta parameters as

$$\hat{\beta} = (X'X)^{-1}X'Y.$$

The k predictor variables in X come from the group of original predictor variables discussed in the first chapter. We decide which predictor variables to include in the model based on variable selection methods.

3.2 Stepwise Variable Selection

Classically, there are three popular methods of variable selection which include forward selection, backwards elimination, and stepwise selection. Each of these methods determine which predictor variables should be included in the model based on some selection criterion. We consider variable selection with the goal of minimizing the Akaike Information Criterion (AIC), which is defined as

$$AIC = 2k - 2log(L(\beta)),$$

where k is the number of parameters and $L(\beta)$ is the likelihood function. The median closing prices are assumed to be normally distributed such that $Y \sim N(X\beta, \sigma^2)$. This means that the likelihood function of the beta coefficients given some point, can be defined as

$$L(\beta) = f(Y) = \prod_{i=1}^{n} f(Y_t).$$

AIC is a criterion used for model comparison where the ideal model is the one with the smallest AIC. The criterion considers the fit of the model to the data using the likelihood function. Subtracting the likelihood function means that the AIC will decrease with an increased likelihood function. We can choose the maximum likelihood estimate, $\hat{\beta}$ of the model parameters at which the likelihood function reaches its maximum possible value. In other words, the maximum likelihood estimate of the beta parameters is the value that maximizes the occurrence of the data that the model is built on. The added $2k$ indicates that a model with a greater number of parameters, while

possibly increasing the likelihood, decreases the AIC. In other words, when considering AIC as our selection criterion, a less complex model is preferred to avoid an overfitting of the model to the training data.

Now that we have established AIC as the selection criterion, we consider the three classical methods of variable selection. The method of forward selection begins with a null model where $Y = 1$. Then for each step or iteration in the process, a variable is added until either the AIC is minimized or there are no more predictor variables to introduce into the model. Backward elimination begins with a full model where $Y = X\beta + \epsilon$. In the full model, X contains all of the predictor variables that we are considering for the model. Then for each step or iteration in the process, a variable is removed from the model until the AIC reaches a minimum value. This study uses the method of stepwise selection since it is a combination of the previous methods. Stepwise selection begins with the null model where $Y = 1$. For each iteration in the process, a variable is either introduced to or removed from the model. The process ends once the AIC is minimized.

Table 3.1 shows each of the iterations of the stepwise process for the pooled training data. In this variable selection procedure, all of the possible predictor variables were included in the final model. This means that each variable increased the maximum likelihood function so as to outweigh the cost of adding an additional variable. In the end, the procedure leaves us with the full model. Looking at some of the specifics of the stepwise selection, we see that there are 9 iterations, and for each iteration a variable was introduced to the model. We also see that for each iteration, the AIC decreases at a slower rate. This indicates that there are diminishing returns for the reduction of AIC due to the cost of adding an additional variable. Finally, we note that the variables introduced into the model first are the variables that increase the likelihood function the most.

Even though the stepwise process gives a full model, this is not the final model. To select the final model, we must consider the significance of the predictor variables and possible multicollinearity between predictor variables.

Table 3.1: Iterations of the stepwise process and the corresponding AIC

Iteration	Add/Remove	AIC
0	-	3422.66
1	$+WIL$	1869.37
2	$+XOI$	1388.94
3	$+VIX$	1290.56
4	$+HUI$	1243.28
5	$+TNX$	1238.86
6	$+DFF$	1205.88
7	$+M1$	1201.46
8	$+TIME$	1193.40
9	$+V$	1188.19

3.3 Model Assumptions and Diagnosis

In this section, we select the final model for predicting the median stock price in the NYSE by considering multicollinearity as well as the significance of the predictor variables. After obtaining the final model, we check to see if our model satisfies the underlying assumptions of a multiple linear regression model.

We see from the previous section that the stepwise process gives the full model as the model with the lowest AIC. A reason for this outcome could be possible multicollinearity. Multicollinearity in a model occurs when predictor variables are highly correlated with one another. This means that each predictor variable can be used to explain the variation in our response variable, however, the information that is explained by each variable will be the same. In other words, highly correlated predictor variables tell us the same information about our response. Multicollinearity is undesirable because it adds unnecessary complexity to the model.

One way to check for multicollinearity is by assessing the correlation between each of the predictor variables. Figure 3.2 gives the correlation between each of the variables in the model rounded to the nearest tenth. Notice that some of the strongest correlations involving predictor variables include time, the Wilshire 5000 index, and M1. For example, there is a strong positive correlation between M1 and time (nearly 90%). This relationship is due to the effect of inflation over time. Inflation causes prices to increase over time relating to an increase of the money supply.

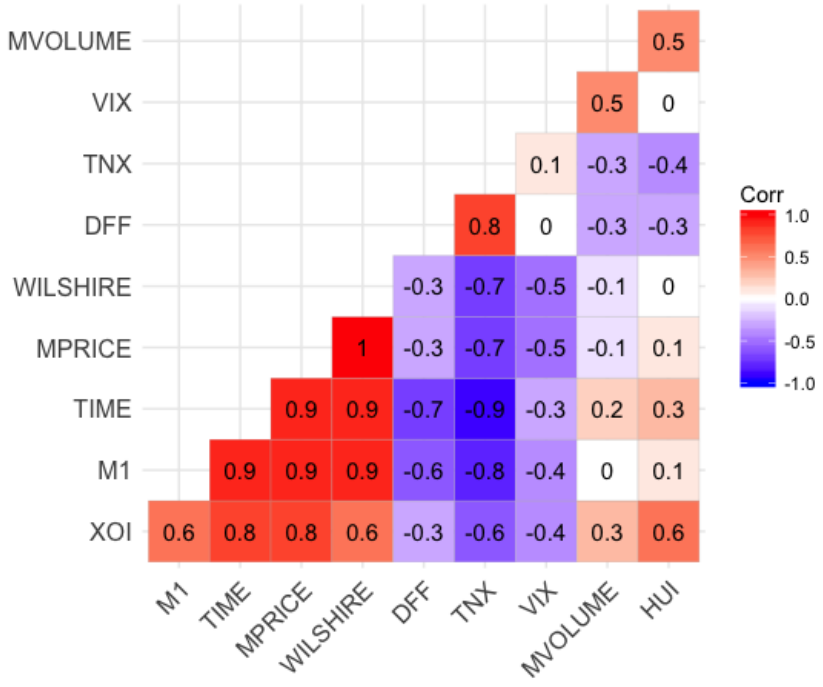

Figure 3.2: Correlation Matrix

The result of inflation over time can be an underlying effect for correlation between our predictors and time.

To deal with multicollinearity, it is best to remove predictor variables that are highly correlated with other predictors. The Variance Inflation Factor (VIF) for each variable can be used to determine which predictors should be removed from the model and is defined as

$$VIF_j = \frac{1}{1 - R_j^2}.$$

If we consider the $X_1, ..., X_k$ predictor variables, then R_j^2 is the correlation coefficient for the fit of X_j on the remaining $k-1$ variables. The correlation coefficient will be discussed later in further detail, but it represents the amount of variation in X_j explained by the remaining predictors. If X_j is highly correlated with the other predictors, R_j^2 will be closer to 1 meaning that VIF_j will be larger. A general rule of thumb is that a VIF greater than 10 indicates a problem of multicollinearity.

Table 3.2 gives the VIF for each X_j in the full model. M1 is the most highly correlated with the other predictor variables with a VIF of 133.05 followed by time with a VIF of 93.76. When

Table 3.2: VIF of predictor variables

Variable	VIF	VIF
WIL	53.12	5.79
XOI	7.52	4.35
VIX	2.76	2.69
HUI	4.75	3.28
TNX	12.82	7.34
DFF	5.22	3.69
$M1$	133.05	-
$TIME$	93.76	-
V	3.79	2.60

analyzing the correlation matrix, we can see that these two variables are indeed highly correlated with the other predictors. We now consider a reduced model where M1 and time are removed. In this model, all VIF's are less than 10 indicating that this solved the problem of multicollinearity in the model.

Before we choose the final model, we first check to see if each of the variables within the model are significant. If a predictor variable is significant, then the variation in the predictor variable can be used to explain the variation in the median closing price. In terms of the model, the slope coefficient for the variable will be significantly greater than zero. When testing for the significance of a variable we can test the null hypothesis that the beta coefficient for that variable is equal to zero versus the alternative hypothesis that the beta coefficient is not equal to zero. We can also think of the test as

$$H_0 : E(Y) = \beta_0 + \beta_1 X_1 + ... + \beta_7 X_7$$

$$H_a : E(Y) = \beta_0 + \beta_1 X_1 + ... + \beta_6 X_6.$$

Here, the null hypothesis is that the expected value of the median closing price follows a reduced model where one variable is removed leaving 6 predictor variables. The alternate hypothesis is that the expected value follows the full model instead where the 7 predictor variables are included. The 7 variables in the model for the alternative hypothesis are the 7 remaining variables

after $M1$ and $TIME$ were removed due to multicollinearity. The test statistic is

$$F = \frac{SSR_F - SSR_R}{SSE_F/(n-k-1)},$$

where SSR_F and SSR_R are the sum of squares for the regression model for the full and reduced models. These represent the amount of variation explained by the model. SSE_F is the error sum of squares for the model or the amount of variation that is not explained by the model. $n-k-1$ is the degrees of freedom under the full model where $n = 657$ is the number of observations in the training data and $k = 7$ is the number of predictor variables that we are considering for the full model. We are interested to see if the gain in the regression sum of squares is significant enough to justify keeping that additional variable in the model. The corresponding p-value for the test is

$$p - value = P(F_{1,n-k-1} \geq F),$$

where 1 represents the additional variable that the full model has over the reduced model. Thus $F_{1,n-k-1}$ represents an F-distribution with 1 and $n-k-1$ degrees of freedom, and the p-value is the probability that we obtain a value of F from this distribution that is greater than our test statistic. Table 3.3 gives the F-statistics and corresponding p-values for each of the 7 variables that we are considering.

Table 3.3: Significance of predictor variables

Variable	F-Statistic	p-value
WIL	17493.01	<0.0001
XOI	941.14	<.0001
VIX	123.02	<.0001
HUI	53.85	<.0001
TNX	6.72	.0098
DFF	35.50	<.0001
V	.1194	.7298

For most of the predictor variables, the p-values for the test statistics are close to zero indicating that for each of those tests, there is sufficient evidence to reject the null hypothesis and conclude

that the predictor variable is significant within the model. The only insignificant variable in the model is the median volume sold in week t which has an F-statistic of 0.1194 and a corresponding p-value of 0.7298. Since the p-value is much greater than 0, there is not sufficient evidence to reject the null hypothesis, and we conclude that the median volume is not a significant variable for predicting the median closing price when also considering our other predictor variables.

After removing volume, we can define the final pooled model as

$$\hat{Y}_t = \hat{\beta}_0 + \hat{\beta}_1 WIL_t + \hat{\beta}_2 XOI_t + \hat{\beta}_3 VIX_t + \hat{\beta}_4 HUI_t + \hat{\beta}_5 TNX_t + \hat{\beta}_6 DFF_t,$$

where $\hat{Y}_t$ is the predicted median closing price at week t, $\hat{\beta}_0, ..., \hat{\beta}_6$ are the parameter estimates for the y-intercept and slope coefficients. The chosen predictor variables include the Wilshire 5000 Index (WIL), the Oil and Gas Index (XOI), the Volatility Index (VIX), Gold Bugs Index (HUI), the interest rate for 10-year T-notes and bonds (TNX), and the Federal Funds Rate (DFF).

Before considering the specifics of the model such as interpretations and applications, we first discuss the diagnostics of the model. We wish to see whether the selected model satisfies the underlying assumptions for a linear regression model. The four main assumptions of the model as shown by Dielman (2005), are linearity between the closing price and our predictor variables, independence of the error terms or residuals, constant variance or homoscedasticity of the residuals, and normality for the distribution of the residuals. We analyze each assumption individually.

The chosen pooled model is a linear regression model meaning that it is assumed that there is a linear relationship between the weekly median closing stock price and the predictor variables. This assumption can be visualized by looking at a plot of the residuals and fitted values. If there is a linear relationship, then the residuals should be distributed centered around a straight line across all values to. The plot gives a red line that represents the center of the residuals for each set of values. It appears that the relationship is linear except for large fitted values where it appears to be slightly curved.

Next we wish to test whether the residuals are independent of each other. This means that we

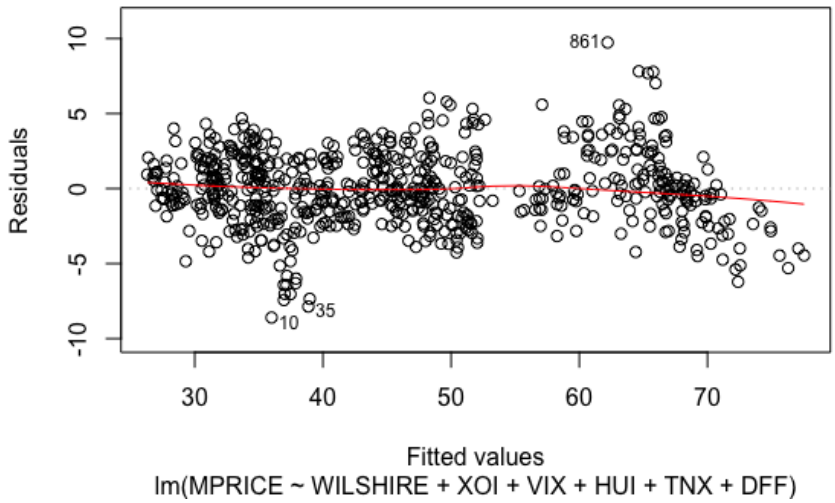

Figure 3.3: Residuals vs Fitted Values

are interested in the correlation between the residuals. The correlation can be visualized by looking at a plot of the autocorrelation function for the residuals.

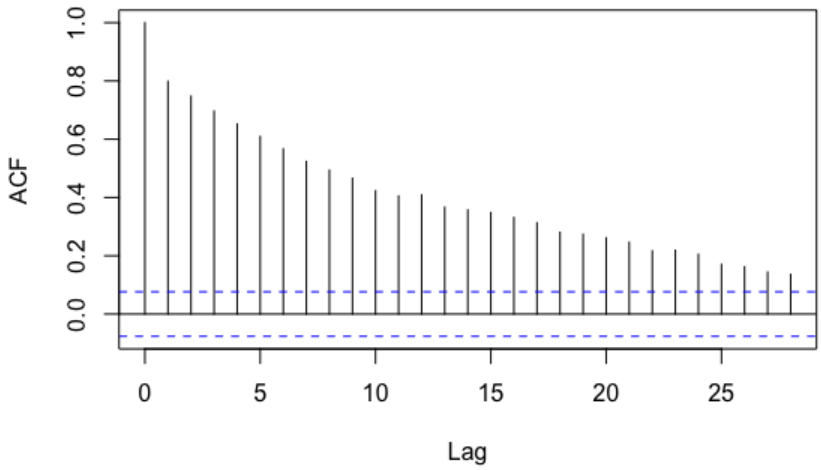

Figure 3.4: Autocorrelation of the Residuals

Based on the plot, we notice that there are significantly high correlations among the residuals. We can also check for the independence of the residuals with formal testing using the Durbin-Watson test where we test the null hypothesis that the autocorrelation among the residuals is zero

against the alternative which suggests that the autocorrelation is greater than zero, or

$$H_0 : \rho = 0$$

$$H_a : \rho > 0.$$

The test statistic is defined as

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2},$$

where we consider the ratio of the sum of squares of the difference between each residual and the previous residual and the sum of squares of the errors. The value of the test statistic obtain from the residuals of the pooled model is $t = 0.39664$ with a corresponding p-value of nearly zero. These results indicate that there is significant evidence to reject the null hypothesis and conclude that the autocorrelation of the residuals is greater than zero. In other words, we conclude that the residuals are not independence and that the model assumption of independence is violated. One possible reason that we obtain this result is because the data are time series data. For time series data, there is generally a correlation among previous values which could indicate a correlation in the residuals. For example, the closing price for one week will be correlated to the closing price of the previous week since in general, there is not going to be a major shift in the economy that will drastically change the median stock price in one week as opposed to a longer time period.

The third model assumption that we must acknowledge is that of constant variance of the residuals or homoscedasticity. We assume that the residuals have variances that are equal and unknown. This can be checked visually by looking at a plot of the residuals versus the fitted values. If the variances are equal, then we should expect the residuals to appear randomly and equally spread among all values. We note from the plot that the residuals across all values tend to be centered around zero showing that the model tends to provide predicted prices centered around the actual prices across all values. We also notice that there are a few values that have extremely high or low residuals that could be due to the presence of outliers where the model was not as

accurate in its predictions. Finally, we can see that for larger values such as predicted prices over $70, we see that the variance appears to be smaller than for other values. When the closing median stock prices is predicted to be a high value such as above $70, the corresponding residuals tend to be smaller. This means that these high predictions tend to be higher than the actual closing price. In conclusion, the variance appears to be constant except for some extremely large fitted values.

Finally, we consider the assumption of normality for the distribution of the errors or residuals which we assume in the creation of the model. In other words, $\epsilon_t \sim N(0, \sigma^2)$ for some variance. For the model to be correct, it is assumed that $E(\epsilon_t) = 0$. We must furthermore assume that the errors also follow a normal distribution when testing for the significance of the model parameters as well as forecasting with the model. Normality can be checked both graphically as well as with the use of formal testing. Visually, we turn to the normal Q-Q plot of the standardized residuals versus the theoretical quantities where the residuals follow the normal distribution. If the assumption of normality is met, then the standardized residuals and the theoretical values should be very similar meaning that the normal Q-Q plot shows values that do not stray far from a straight line.

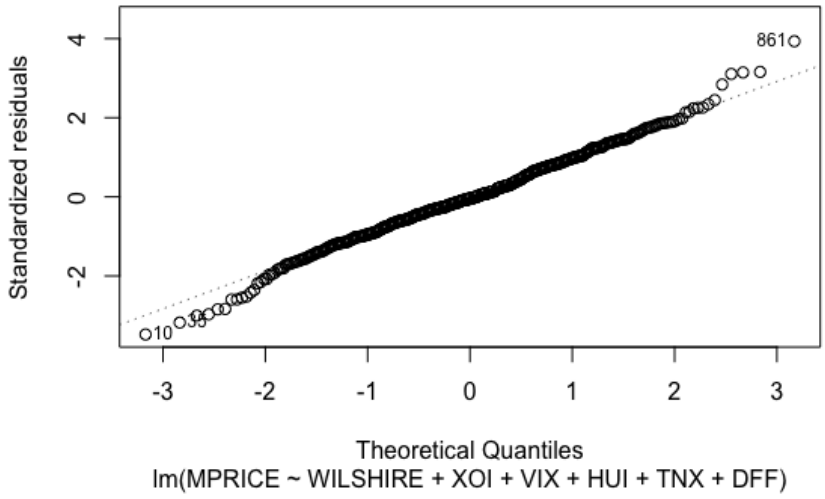

Figure 3.5: Normal Q-Q Plot

Based on the plot, points near zero tend to follow a straight line while points farther out tend to stray farther from the line created a curved shape on the end points. This indicates that values closer to the zero are closer to the theoretical values following the normal. This observation is intuitive

since theoretically the mean is places at zero. Values on the end points of the graph indicate values that differ greatly from the normal model. These values can indicate a skewed distribution for the residuals as well as outliers. Unusual values will always be present as very large or very small values and will stray from the normal. Since the data we are analyzing for the pooled model is the median closing stock price, this will exclude the effects of outliers present across individual stock indexes. However, outliers can still be present in the form of extremely large or small median prices. The unusual prices can be due to external effects or anomalies in the market or across the economy as a whole. For example, we see that there is a curve of many smaller values that stray from the line. These could indicate unusually lower prices due to some economic downturn such as the Great Recession which was brought on by a crash in the housing market.

The normality assumption can also be checked with formal testing using the Anderson-Darling test for normality. This test is versatile because it can be used to check if a sample distribution fits any probability distribution. This means that this test can be applied specifically for the testing of normality. The null hypothesis that the residuals come from a normal distribution is tested against the alternative hypothesis that the errors are not normally distributed.

$$H_0 : \epsilon_t \text{ are normally distributed}$$

$$H_a : \epsilon_t \text{ are not normally distributed.}$$

The Anderson-Darling test considers the distance between the sample distribution of the observed residuals and the hypothesized normal distribution. The test statistic A^2 is used to quantify the discrepancy rather than simply looking at the plot. The test also places a higher weight on the endpoint values which is where there tends to be the greatest difference from the normal. The test statistic is defined as

$$A^2 = -n \sum_{i=1}^{n} \frac{2i-1}{n} [log(F(e_t)) + log(1 - F(e_{n+1-t}))],$$

where the function $F(*)$ is the cumulative normal distribution. Therefore, if the standardized resid-

ual distribution closely follows the normal, then it is expected to have a smaller test statistic. The value of test statistic for the residuals based on the pooled model is $A^2 = 0.96234$. The corresponding p-value for the test statistic is 0.01516. If we are basing our decision with a standard 95% confidence level, we can conclude that there is sufficient evidence to reject the null hypothesis and conclude that the residuals do not follow a normal distribution. Holding a 95% level of confidence indicates that we allow the type I error, or the probability of falsely rejecting the null hypothesis, to be up to 5%. In other words, the p-value tells us that the probability that we obtain a test statistic of 0.96234 given that the residuals come from a normal distribution is approximately 2%. In conclusion, the test results indicate that the normality assumption is not satisfied. This result could possibly be related to a problem with linearity in the model.

3.4 Interpretation and Fit of the Model

In this final section, we consider the fit of the model by considering the error and analyzing the testing data. We also discuss interpretations of the model coefficients and what the model tells us about how the median stock prices relate to the predictor variables.

Perhaps the most common way to assess the fit or the predictive capabilities of the model is with the coefficient of determination which is a value that represents the percentage of the variability in the response value that can be explained by the variability in the predictor variables. The coefficient of determination is defined as

$$R^2 = 1 - \frac{SSR}{SST},$$

where SSR is the sum of the squares of the residuals, or $\sum_t e_t^2$, and SST is the total sum of squares, or $\sum_t (y_t - \bar{y})^2$. The residuals represent the amount of error caused by the discrepancies between the estimated values and the actual values. Therefore the ratio of the residual sum of squares and the total sum of squares gives the percentage of variation related to the residuals. This represents the unexplained variance which is not accounted for by the model. Therefore R^2 then gives the variation of the median stock price that is accounted for by the model. Models that have a higher coefficient of determination tend to be a better fit and give better predictions because the

model is able to explain more of the variation in the stock price. The coefficient of determination for the pooled model is $R^2 = 0.9664$. This indicates that 96.64% of the total variance in the weekly median stock price for the top stocks in the NYSE 100 is linearly associated with the variance in the Wilshire 5000 Index (WIL), the Oil and Gas Index (XOI), the Volatility Index (VIX), Gold Bugs Index (HUI), the interest rate for 10-year T-notes and bonds (TNX), and the Federal Funds Rate (DFF). The percentage of explained variation is very high indicating that the model provides a good fit for the median stock price for the top stocks in the NYSE.

The problem with considering R^2 as a measurement of the fit for the model is that it will always increase when more predictor variables are added to the model. This means that based solely on the R^2 a better model would be a model with more predictor variables. However, this is not true. As discussed previously, when creating a model, the goal is to have a well fit model that is as simple as possible. Additional variables can bring additional explanation to the variation of the response variable, however, there is a point where adding an additional variable is not worth the increasing the complexity of the model. An example would be predictor variables that are highly correlated with each other. If we have too many variables, it is likely we will see a high correlation among predictors. Highly correlated predictors give the information about the variability in the response, so it would be unnecessary to include all of them into the model. To account for the complexity of the model when considering the fit, we look instead to the adjusted coefficient of determination which is defined as

$$R^2_{adj} = 1 - \frac{SSR/(n-K-1)}{SST/(n-1)}.$$

Looking at the equation, the adjusted value is similar to R^2. The difference is the the residual sum of squares and the total sum of squares are divided by their respective degrees of freedom. Therefore, the adjusted coefficent of determination considers both the sample size, n, and the number of predictor variables in the model, K. The adjusted value R^2_{adj} will never be larger than R^2. While the unadjusted value will always increase with each additional predictor variables added to the model, the adjusted value will only increase if the additional variation explained by the added predictor variable is greater considering the added complexity to the model. The interpretation of

the adjusted coefficient is nearly the same as the unadjusted value. Given the two values, R^2_{adj} is preferred given the consideration of model complexity. The adjusted coefficient of determination for the pooled model is $R^2_{adj} = 0.9661$ and indicates that 96.61% of the total variance in the weekly median stock price for the top stocks in the NYSE 100 is linearly associated with the variance in the Wilshire 5000 Index (WIL), the Oil and Gas Index (XOI), the Volatility Index (VIX), Gold Bugs Index (HUI), the interest rate for 10-year T-notes and bonds (TNX), and the Federal Funds Rate (DFF). Compared to the unadjusted value, the adjusted value is only slightly smaller. This indicates that all of the predictors give additional explanation in the price variation meaning that their inclusion within the model is beneficial to the fit of the model considering the increased complexity. The interpretation yields a similar conclusion as the unadjusted value and indicates that the model is successful in explaining the variation in the weekly median stock prices. Furthermore, we then expect the model to provide accurate price forecasts which will be tested later using the testing data.

Considering the coefficient of determination is useful for determining the fit of the model and how well the predictors overall explain the variation in the response, however, we are interested to consider each predictor individually to gain information on the relationship between each of the macro and microeconomic variables in the model and the weekly median closing price. This is done by interpreting the beta coefficients for the predictor variables from the final pooled model.

The coefficients give the relationship between each predictor and the median closing price by analyzing the change or variability in stock price related to the change in our predictors. Table 3.4 lists the independent variables in the final model along with the corresponding beta coefficients which are estimates of the model parameters built from the least squares regression model.

The estimate for the parameter β_0 or the y-intercept is 22.4656 and can be interpreted as the predicted median stock price when all predictors take a value of zero. The intercept does not have a meaningful interpretation in this model since it does not make sense for any of the predictors to take a value of zero.

The regression coefficient for the variables WIL is 0.3539 and indicates that holding all other

Table 3.4: Estimates of Model Parameters

Variable	Parameter	Estimate
Intercept	β_0	22.4656
WIL	β_1	0.3539
XOI	β_2	0.0136
VIX	β_3	-0.1587
HUI	β_4	-0.0094
TNX	β_5	-1.3197
DFF	β_6	0.5498

variables constant, when the Wilshire 5000 Index increases by 1 point, it is predicted that the median closing stock price for the top stocks is NYSE 100 will increase by an average of $0.3539 or approximately 35 cents. As discussed earlier, the Wilshire 5000 is used as a method to estimate the state of the stock market as a whole, so it is expected that we see a positive relation with the median closing price and the Wilshire Index.

The regression coefficient for XOI is 0.3539 and indicates that holding all other variables constant, when the NYSE ARCA Oil and Gas Index increases by 1 point, it is predicted that the median closing stock price for the top stocks in the NYSE 100 will increase by an average of approximately 35 cents. Since oil and gas are commodity goods, their price has a positive relationship with the price of stocks across the board.

The regression coefficient for VIX is -0.1587 and indicates that holding all other variables constant, when the CBOE Volatility Index increases by 1 point, it is estimated that the median closing stock price for the top stocks in the NYSE 100 will decrease by an average of approximately 16 cents. Here we see the negative relationship between expected volatility and stock price. When it is expected that stock prices will become more volatile, people are more hesitant to invest in the market relating to a decrease in price.

The coefficient for HUI is -0.0094 and indicates that holding all other variables constant, when the Gold Index increases by 1 point, it is predicted that the median closing stock price for the top stocks in the NYSE 100 will decrease by an average of approximately 1 cent. The price of gold and the price of stocks have a negative relationship. This is because although there is

a correlation between the two, they are not considered to be equivalent assets. This means that if the prices of stock equities are down, investors tend to choose to move their holds into gold instead, hoping to gain higher returns rather than continue to invest in a declining asset. For a well balanced investment portfolio, it is safest to invest in a variety of assets such as stocks as well as gold because different assets can hold different price trends.

The coefficient for TNX is -1.3197 and indicates that holding all other variables constant, when the CBOE interest rate for 10 year T-note bonds increases by 1 percent, it is estimated that the median closing stock price for the top stocks in the NYSE 100 will decrease by an average of approximately \$1.32. Similarly with gold, bonds are a separate type of asset from stocks. Bonds are commonly found along with other equities in an investment portfolio. When the rates on bonds are increased, then there is a higher yield for bonds value meaning that investors will tend to invest more in bonds as opposed to stocks.

The regression coefficient for DFF is 0.5498 and indicates that holding all other variables constant, when the Federal Funds Rate increases by 1 percent, it is predicted that the median closing stock price for the top stocks in the NYSE 100 will increase by an average of approximately 55 cents. As explained earlier, the Federal Funds Rate is the interest that companies and banks must pay when borrowing from the federal reserve. From an economic standpoint, when the funds rate increases, it becomes more costly for businesses to invest or expand on their business, and higher costs are generally related to lower profit. Therefore, we might predict an inverse relationship between the funds rate and stock prices, however, we see a positive relationship instead. When the economy is suffering, the Federal Reserve lowers the borrowing rate in order to promote borrowing and spending. For example, during the Great Recession, there are both low stock prices as well as low Federal Funds rates.

Now that we have considered the fit and interpretation of the model based on the adjusted R^2 and the estimates of the model parameters, we look to see the performance of the model for predictions and forecasting. At the beginning of the modeling process, the data are partitioned such that the testing data are not used in the creation of the model. This means that if the pooled model

is used to make predictions for the testing data, we can compare the actual values to the predicted values. Using data that was not used in the creation of the model is important to make sure that the model is not over-fit to the training data.

The predicted values are obtained by plugging in the values for each of the predictor variables for each week into the model. The training data include the values for 282 randomly selected weeks from the original data set. The figure below is a plot of the training data representing the actual median closing prices and the predicted closing prices obtained from the model.

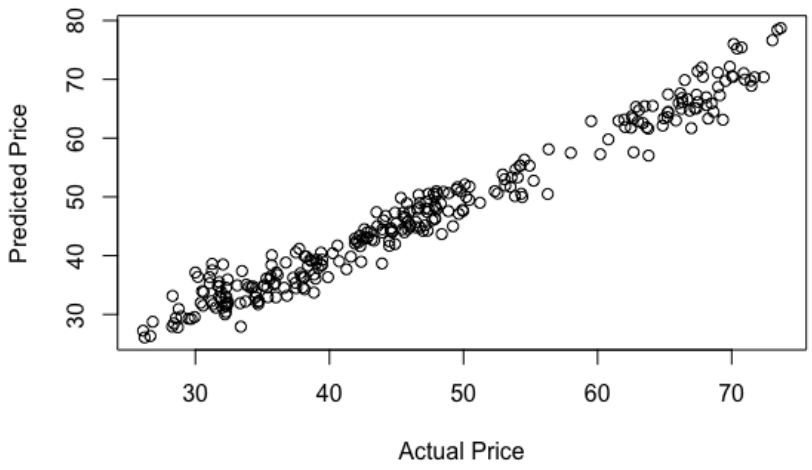

Figure 3.6: Fit of Testing Data vs Predicted Values

If the model is successful, we should expect to see predicted values that are close to the actual values. From the plot, there is a very strong correlation between the predicted and training values indicating that the model has created accurate predictions. Since these predictions were made on data not used in the creation of the model, we conclude that there is not a problem of overfitting of the model to the training data.

CHAPTER 4 TIME SERIES ANALYSIS

4.1 Exploratory Data Analysis

Time series analysis is available uniquely to data that occur sequentially over intervals of time. The closing price of stocks are time series data because the price can be recorded over various time intervals. Data can be recorded in any time interval including daily, weekly, monthly, and yearly. The data however must be recorded in a way such that the time passed between each observation is equal. Stock prices are recorded and updated constantly during open market hours so that stock brokers and stock traders can make quick buying and selling decisions. Therefore, data for daily stock prices are easily obtainable for any stock in the market. Despite the availability, this study prefers to consider weekly prices for two reasons.

The first reason is the requirement that the time series data have equal intervals between observations. Since the New York Stock Exchange does not have trading hours on weekends and holidays, there are not updated prices on these days meaning that daily stock data would have missing values. There are methods such as bootstrapping for dealing with missing data, however there is a second reason for why we simply consider weekly data.

The second reason is that considering shorter time intervals such as daily data means that there is more variation within our data as stock prices can have periods of increase and decrease from day to day as well as throughout a single day. Changes during very short time periods can be the result of white noise or unpredictable anomalies. So although there are stock brokers or traders who time deals to the minute or second based on price trends in the immediate short run, for the purposes of predictive analysis, considering longer time intervals such as weekly data give more stable predictions and clearer price trends.

So why consider a time series analysis as opposed to simply using a regression model such as the one constructed in the previous chapter? Since time series analysis considers data that are observed in sequential intervals, a time series model is built with the knowledge that the data

observations are not independent from each other. This means that the closing price at the end of a week is correlated to the closing price at the end of the previous week. Even though there are differences in prices each week, the prices in consecutive weeks tend to be similar to each other unless there is an anomaly in the state of the economy that causes prices to change drastically. The fact that consecutive observations in time series data are dependent of each other causes issues in a regression model where independence is assumed. Recall the regression model from the previous chapter where our model fails to assume independence between the residuals. A time series analysis of the data is our way of dealing with dependence in the data.

Before we begin our time series analysis, let us clarify the data that is being considered. In this analysis we are looking to forecast or predict the weekly median closing price for the top stocks in the NYSE as is done in the pooled regression model from the previous chapter. At the end of this section, we consider what would change in our analysis if we were to use the weekly closing price for each index rather than the median.

The first step in the time series analysis is a first look at the data to consider any patterns, trends, cycles, or abnormalities that occur in the weekly closing stock price over time. This consideration is the exploratory data analysis of the time series data before modeling can occur. We visualize these patterns through a time series plot of the data as shown in Figure 4.1 which includes the median closing price for the NYSE for each week from January 01, 2000 through December 23, 2017.

Based on the time series plot, there is an overall increasing trend in the median stock price from January 01, 2000 through December 23, 2017. This indicates that over time, the median weekly closing price tends to increase. Based on economic theory, this makes sense because over time, inflation will drive prices higher. The data analyzed in this study has been recorded over 17 years which is a substantial period of time such that the effects of inflation are visible. Any sequence of random variables, Y_t, is defined as a stochastic process and includes times series data such as the weekly median closing stock price. For time series or a stochastic process, the mean function or expected value at time t is $E(Y_t) = \mu_t$. This indicates that the expected median closing stock

Figure 4.1: Closing Price for each Week

price is dependent on the time period. This conclusion intuitively appears natural for time series data since we know that the data are dependent on time. In addition, the mean or expected median closing price is generally different for each week which is why we consider μ_t for each time point.

Although there is a clear increasing trend in the closing price over time, there is a clear abnormality in this trend that occurs between late 2007 through mid 2009. Starting in late 2007, there is a break in the increasing trend. At this point in time, the median closing price drastically decreases compared to the prices previously observed in the data. The visibility of this anomaly is not coincidental, but instead reflects the period that covers the Great Recession which officially occurred from December 2007 through June 2009. When the recession began in 2007, it was a result of a crash in the United States real estate market which then brought repercussions to a global recession. Since the top stocks in the NYSE 500 mainly include companies based in the United States, the median closing prices for these stocks decrease sharply at this initial shock. During the early to mid 2000's, the housing market in the United States was booming which led to increased investment in mortgage-backed securities. These securities were issued at high rates which were not strictly regulated during this time. Because of the booming housing market, the value of the securities were high, but because of loose regulations, the mortgage-backed securities were issued at high-risk rates. When the housing market crashed in late 2007, these securities drastically

decreased in value causing many financial institutions invested in these securities unable to meet financial obligations or file for bankruptcy. The financial instability in major institutions expanded on the economic shock of the market crash. The recession led to decreased GDP, increased unemployment, decreased spending and other negative economic repercussions including a dramatic decrease in stock prices across the board. Even though the effects of the Great Recession on GDP and unemployment was not a large as the effects of the Great Depression, the effects of the former lasted for such a significant period of time and then spread across the world that it still holds the title "Great". Response to the crisis included federal funds rates to be set at minimum levels by the Federal Reserve. This action was to stimulate the economy by promoting spending and borrowing to increase liquidity of assets in the economy.

We have pointed out the overall increasing trend as well as the abnormality in the median price that occurred during the Great Recession. Lastly we use the time series plot to analyze the variation in the closing price over time. When we consider variation, we are interested in the change in the median closing price. Based on the plot, we see that the variation does not appear to be constant. Over short periods of time, there can be small or large changes in the stock prices. In the short run, stock prices can vary greatly leading to unpredictable prices. These unpredictable short run changes are referred to as white noise. For example, a policy change such as the change in the federal funds rate can influence the economy including stock prices. Short term effects can be stabilized in the long run which is a reason why we consider longer periods of time in this study. Here we consider weekly stock prices rather than daily data. First we will consider the variation in price for a single week or time point t. For the analysis we consider the value or price at each time point as a random variable. The variance for the median closing price at time t can be defined in the same way as any random variable. Here we define the variance as

$$V(Y_t) = E(Y_t - \mu_t)^2,$$

which is the squared expected value of the difference between the observed closing price and the

true closing price at time t. For time series data, we also consider the variation between time periods, or the covariance. The auto-covariance function (ACVF) between the median closing price at times t and $t+h$ is defined as

$$\begin{aligned}\gamma_{t,t+h} &= Cov(Y_t, Y_{t+h}) \\ &= E[(Y_t - \mu_t)(Y_{t+h} - \mu_{t+h})],\end{aligned}$$

which represents the expected value of the product of the differences between the closing price and the true mean closing price for weeks t and $t+h$. h represents a time period or lag after time week t. The auto-covariance functions can be used to measure the linear dependence for median closing price at various weeks. From the covariance, we can define the autocorrelation function (ACF) between the median closing price at times t and $t+h$ as

$$\rho_{t,t+h} = Corr(Y_t, Y_{t+h}) = \frac{\gamma_{t,t+h}}{\sqrt{V(Y_t)V(Y_{t+h})}}.$$

The correlation between the median closing price at times t and $t+h$ represents the amount of variation in each variable that can be explained by the other. We see from the formula that the correlation is a ratio of the covariance and the product of the standard deviations which represents the overall variations of both variables. We note that $\gamma_{t,t} = V(Y_t)$ and $\gamma_{t+h,t+h} = V(Y_t + h)$ since the covariance between the same variable is simply the variance.

Next we discuss some of the properties of the correlation. First notice that $|\gamma_{t,t+h}| \leq \sqrt{\gamma_{t,t}\gamma_{s,s}}$. Since the covariance, $\gamma_{t,t+h}$, measures the variation between the median closing prices at weeks t and $t+h$, it is bounded by the variation of the two variables. Therefore, the correlation between the two random variables are bounded between -1 and 1. In other words, $|\rho_{t,t+h}| \leq 1$. When $|\rho_{t,t+h}|$ is closer to one, the covariance or interaction between the two variables is close to the overall variations each. This indicates that the overall variance can be closely explained by the variation between the variables indicating that there is a strong relationship between the variables. In other words, the median closing stock price at week $t+h$ is related to the closing price at a previous

week t. A value of $\rho_{t,t+h}$ close to zero indicates that the variation between the two variables is small relative to the overall variations of both variables. In other words, the closing prices would be considered uncorrelated.

Before we can model the time series process, we need to be able to make an assumption about the behavior or structure of the process over time. If we are to make a model based on the observed time series data with the purpose of predicting future values, we must assume that the structure of the process remains the same. This type of process is considered to be stationary. With stationarity, there is strict stationarity, however we simply wish to model a weakly stationary process which is weaker mathematically but holds some similar assumptions. For a stochastic process to be weakly stationary, the mean function must be constant over time, and the variance must be constant over time. Both of these requirements indicate that if met, the process maintains the same structure over time. If the mean function is constant over time, then the expected value of the closing price at any time point is the same. In other words, $E(Y_t) = \mu$. This is stronger than the definition of the expected value or mean function that we have previously defined. For the second requirement, if the variances are equal, then $V_t = V_{t+h} = V_0$ such that the variance is not a function of time but is a constant for all weeks. We can write the constant variance in terms of covariance as γ_0. Also, when we consider constant variance over time, the covariance between random variables that are equal distances apart should be equal. In other words, $Cov(Y_t, Y_{t+h}) = Cov(Y_{t+k}, Y_{t+k+h})$ for some lag h and some time period k. This is the same as saying $\gamma_{h,t} = \gamma_{h,t+k} = \gamma_h$. Therefore, the variance is constant and not dependent on time. Instead the covariance between the closing prices Y_t and Y_{t+h} is instead based on the time lag difference between them, h. We can also extend the implications of constant variance to the correlation between two random variables in the stationary time series process. Suppose we consider the correlation between weeks t and $t + h$ for a stationary process. Then, based on our redefined values for the variance and covariance we obtain, $\rho_{t,t+h} = \frac{\gamma_{t,t+h}}{\gamma_0}$. Based on these redefined formulas, notice that for a stationary process, the variance is a constant value, and the covariance and correaltion between two variables is dependent on the lag h between them.

Based on our observations of the time series plot, we first noticed that due to inflation, the median closing stock price increases over time. This indicates that rather than being constant over time, the function is instead dependent on time. This would then indicate that the median stock price over time is not a stationary process because the requirement of a constant mean is not met. We also noticed periods where the change in the closing stock price is either larger or smaller. The most obvious example of this was the extreme decrease in price observed during the Great Recession. Because the occurrence of the recession is an anomaly in the economy, not only can it be related to a non-stationary process, it can also have a potential of creating a model that is overfitted to this anomaly. For the sake of providing a model that better meets the underlying assumptions and makes accurate predictions, we will consider time series models that are based on the only the data after the the Great Recession as well as the entirety of the data. In other words, we will consider the modeling process based on the data from January 1, 2000 through December 23, 2017 as well as the data from June 6, 2009 through December 23, 2017.

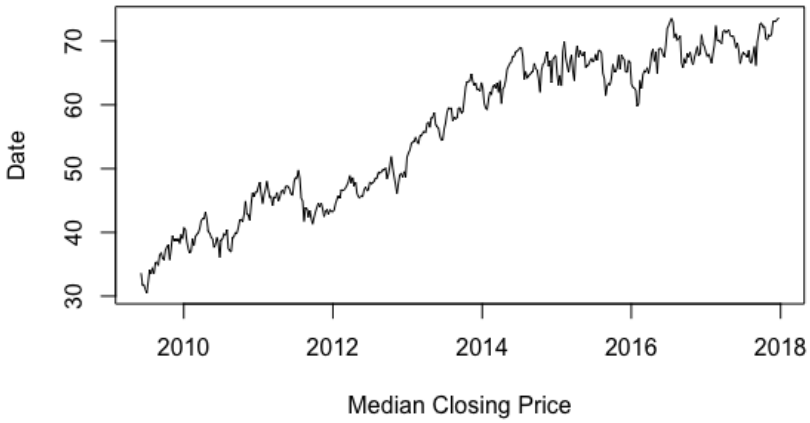

Figure 4.2: Closing Price for each Week: Post Great Recession

Figure 4.2 shows the time series plot of the median closing price after the end of the Great Recession. Based on the plot, we see that the anomaly in the data has been removed. However, there is still the same increasing trend due to inflation. There is still some visible changes in variation over time. For example, it appears that from 2009 to 2014 the increasing trend is steeper

or more drastic than from 2014 through 2017. Both of these observations indicate that the time series data after the Great Recession are also non-stationary.

Stationarity can also be formally tested using the Augmented Dickey-Fuller (ADF) test which tests the null hypothesis that the data are not stationary against the alternative hypothesis that the data are stationary. For the entire set of data for all 939 weeks, the test statistic is -2.0018 with a p-value of 0.5776, which indicates that there is not sufficient evidence to reject the null hypothesis and we conclude that the data are non-stationary. For the data occurring after the Great Recession, the test statistic is -2.5896 with a p-value of 0.3282, which indicates that there is not sufficient evidence to reject the null hypothesis and we conclude that the data are non-stationary. Both of these conclusions align with those constructed from the time series plots. Also, the p-value for the data after the Great Recession is smaller than that for the entire data set which indicates that removing the drastic drop in price that occurred during the recession has reduced non-stationarity in the data.

Similarly with the construction of the pooled model, the data for both time intervals must be partitioned into training and testing data. Since the data are being retained as time series, the data are not partitioned randomly. Instead, the data for the last year will be reserved for testing so that the actual data can be compared with the model predictions. For the entire data set, the 888 weeks from January 1, 2000 through December 31, 2016 make up the training data while the 51 weeks from January 7, 2017 through December 23, 2017 make up the testing data. For the post-recession data, the 396 weeks from June 6, 2009 through December 31, 2016 make up the training data while the year of 2017 still represents the testing data set. Both models use the same testing data, which means that it is simpler to compare the fits of the models based on the accuracy of their predictions.

Since it has been established that the data are non-stationary, it is necessary to make the data stationary before modeling can be done. The most common way to obtain stationarity is through differencing. This means that instead of modeling the price at week t which is increasing over time, we consider the difference or change in price over each week. In other words, we wish to model $\Delta Y_t = Y_t - Y_{t-1}$. Figure 4.3 shows the time series plots for the first order difference for the

entire data and for the post-recession data.

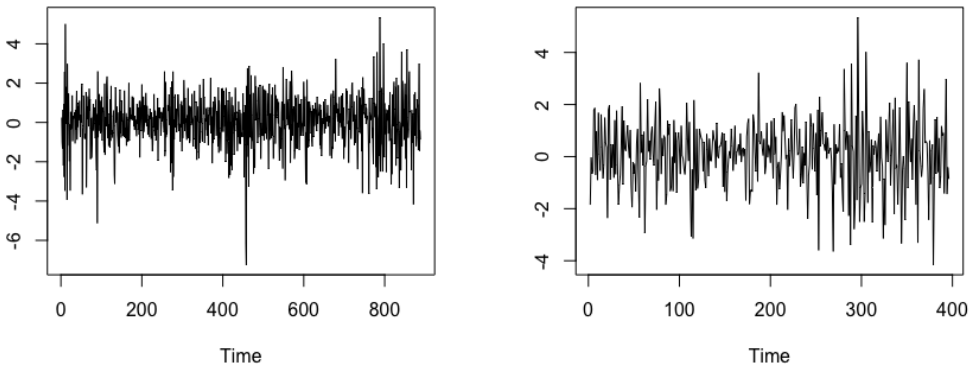

Figure 4.3: Time Series Plot of First Order Difference

From the graph, we see that for both time period, the average value appears to be centered around zero. This indicates that the median closing price tends to be similar to the median closing price for the previous week. In other words, the stock price does not differ greatly on a weekly basis. Furthermore, this indicates that the first order difference are stationary data because the mean is not dependent on time but is rather a constant. By looking at the variation, we see that it tends to be constant for both data sets. One exceptions would be for the entire data set, the impact of the Great Recession is still visible although not as drastic as the data was before the differencing. At the point of the recession there is an extremely small difference illustrating the catastrophic decrease in price when the housing market crashed. Some change in variation is also slightly visible for the data occurring after the recession. It appears that there is greater price variability for more recent weeks which is illustrated by the larger and smaller differenced values for the more current data. In conclusion, based on the time series plots, it appears that the first order difference appears to be stationary.

We can formally test the hypothesis that the first order difference are not stationary against the null hypothesis that the first order difference are stationary using the ADF test. Performing the test for the entire data set yields a statistic of -10.818 with a corresponding p-value of less than 0.01 which means that there is sufficient evidence to reject the null hypothesis and conclude that the

differenced data for the entire data set are stationary. The test statistic for the first order difference of the post-recession data is -7.7475 with a corresponding p-value of less than 0.01 which means that there is sufficient evidence to reject the null hypothesis and conclude that the differenced post-recession data are stationary. Since the data are stationary, we continue to the modeling process using the first order differences for both data sets.

4.2 Model Identification and Selection

It has been determined that the differenced data, $\Delta Y_t = Y_t - Y_{t-1}$ are stationary. When fitting a model for a stationary time series, we model the data based on past observations and past errors. In other words, the differenced data can be expressed as

$$\Delta Y_t = \sum_{j=1}^{p} \phi_j \Delta Y_{t-j} + \sum_{i=0}^{q} \theta_i \epsilon_{t-i},$$

where ϵ_t represent the white noise that is assumed to be normally distributed with mean 0 and variance σ_ϵ^2. The past p observations included in the model and the corresponding coefficients of ϕ represent the components of an auto-regressive process of order p. The past q white noise terms and the corresponding coefficients of θ represent the components of a moving average process of order q. These types of models are called Univariate Box-Jenkins (UBJ) models. They are also referred to as $ARIMA(p, d, q)$ models where $AR(p)$ indicates the auto-regressive component, $MA(q)$ indicates the moving average component, and d is the degree of differencing for non-stationary data. As discussed on the previous section, first degree differencing is satisfactory meaning that we consider the median closing stock price to follow an $ARIMA(p, 1, q)$. The goal is to identify possible candidate models of auto-regressive and moving average components and then select the model with the best fit.

Since the the differenced data are stationary, we can simply the notation. We can write the mean as $E(\Delta Y_t) = \mu_t = \mu$ since there is a constant mean difference. We write the variance as $V(\Delta Y_t) = \gamma_{t,t} = \gamma_0$ since the variance is constant over time. The covariance between any two observations ΔY_i and ΔY_j where $|i - j| = h$ can be simply written as γ_h since the covariance

is a function of the lag. Based on these observations, we define the correlation between any two observations ΔY_i and ΔY_j where $|i-j| = h$, as $\rho_h = \frac{\gamma_h}{\gamma_0}$ which is the autocorrelation function (ACF).

First we consider the auto-regressive model of degree p, $AR(p)$, which we define as

$$\Delta Y_t = \phi_1 \Delta Y_{t-1} + ... + \phi_p \Delta Y_{t-p} + \epsilon_t.$$

From the equation, we see that the process is based on the previous p observations. Let us consider a simple AR(1) model. If we consider the equation for the AR(1) and then multiply each side of the equation by ΔY_{t-h} and take the expected values, the autocorrelation function, ρ_h, can be derived as

$$\rho_h = \frac{\gamma_h}{\gamma_0} = \frac{\phi \gamma_{h-1}}{\gamma_0} = \phi^h,$$

as shown in the text by Cryer and Chan (2008) Based on the theoretical values of the autocorrelation function for an AR(1) model, we can identify an process as being auto-regressive if ACF experiences exponential decay. However, the autocorrelation function does not allow us to identify the degree of the auto-regressive process. For this, we turn to the partial autocorrelation function (PACF). The parital autocorrelation at lag k is defined as the correlation of two observation ΔY_t and ΔY_{t-k} accounting for the effect of the variables in between, $Y_{t-1}, ..., Y_{t-k+1}$. In other words,

$$\phi_{kk} = Corr(Y_t, Y_{t-k} | Y_{t-1}, ..., Y_{t-k+1}).$$

As shown in the text by Cryer and Chan (2008) this means that for an auto-regressive process of degree p, $\phi_{k,k} = 0$ for $k > 1$. Furthermore, an $AR(p)$ model can be identified by a dampening of the PACF after lag p.

Next we discuss the moving average model of degree q, $MA(q)$, which we define as

$$\Delta Y_t = \epsilon_t + \theta_1 \epsilon_{t-1} + ... + \epsilon_{t-q}.$$

The moving average process defines the differenced median closing price as a function of the current random error and previous random error terms. Since the error terms are assumed to be normally distributed with mean zero, the expected value of the differenced data would also be zero. Lets consider the simplest moving average model which is one of degree one where $\Delta Y_t = \epsilon_t + \theta_1 \epsilon_{t-1}$. Then the variance is $\gamma_0 = \sigma^2(1 + \theta^2)$. The autocorrelation function is then $\rho_h = \frac{-\theta}{1+\theta^2}$ for $h = 1$ and $\rho_h = 0$ for $h > 1$. This can be expanded to the general case for any $MA(q)$ model. If a process follows an $MA(q)$ model, the autocorrelation is zero for any lag greater than q. Therefore, the moving average component can be identified from a plot of the ACF.

Now that we have defined the auto-regressive and moving average components of the UBJ model, we can analyze the sample autocorrelation functions (SACF) and sample partial autocorrelation functions (SPACF) of the differenced data in order to identify appropriate $ARIMA(p, 1, q)$ candidate models.

We begin by observing the autocorrelation for the first order difference of the entire data set which is plotted in Figure 4.4. From the autocorrelation function, we first see that the the correlation for a lag of zero is 1. This is because any point is 100% correlated with itself. Also, there is a significant correlation when the lag is 1. This means that there is a significant relationship between the difference in price between two weeks and the difference in price for the week prior. In other words, there is a significant correlation between ΔY_t and Δ_{t-1} for any week t. After lag 2, the correlation becomes insignificant. In other words, there is not a significant relationship between ΔY_t andΔ_{t-2} for any week t. These observations indicate the presence of a moving average component of order 1 within the model. Therefore, the first candidate model that we consider is an $ARIMA(0, 1, 1)$. This means that we consider a model where the price difference is a function of the previous error term. In other words, we consider $\Delta Y_t = \epsilon_t + \theta \epsilon_{t-1}$.

Next, we analyze the sample partial autocorrelation function. From the plot, we notice that the SPACF is significant for a lag of 1, and then becomes insignificant for any lag greater than 1. This means that when accounting for the effect of all intervening variables, there is only a significant relationship between the difference in price and the difference in price for the previous week. These

observations indicate the presence of an auto-regressive component of degree 1 within the model. Therefore, the second candidate model that we consider is an $ARIMA(1,1,0)$. This means that we consider a model where the difference in price is a function of the previous price difference. In other words, we consider $\Delta Y_t = \phi \Delta Y_{t-1} + \epsilon_t$.

Since the correlation plots indicate the presence of a moving average component and an auto-regressive component, the third candidate model that we include is an $ARIMA(1,1,1)$ where the price difference is a function of the previous price difference as well as the previous error term. In other words, we include the model $\Delta Y_t = \phi \Delta Y_{t-1} + \epsilon_t + \theta \epsilon_{t-1}$.

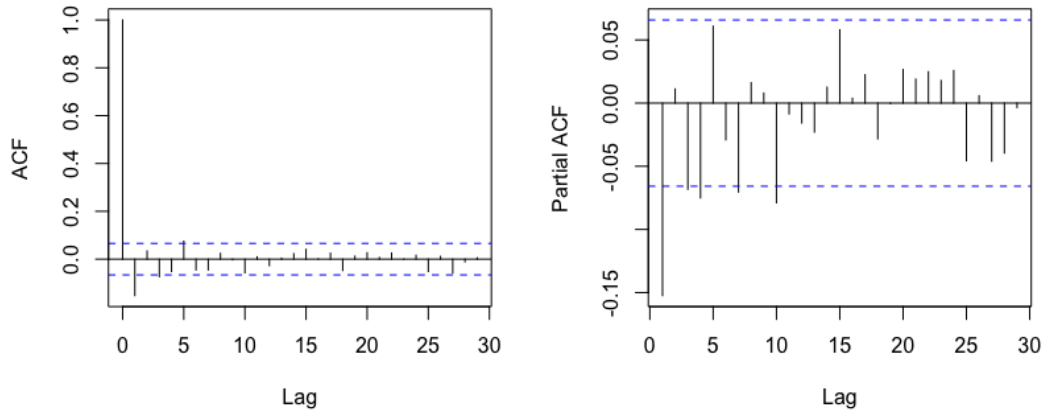

Figure 4.4: SACF and SPACF of Price Difference

Now that we have identified candidate models when considering the entirety of the data, we now turn to the SACF and SPACF of the price difference for the post-recession data which is plotted in Figure 4.5. From the autocorrelation function, we see that again there is a correlation of 1 for lag 0 indicating the absolute correlation with a value to itself. There is also significant correlation for lags 1 and 10 while the correlation for all other lags are insignificant. Based on these observations, the first candidate model that we consider is an $ARIMA(0,1,1)$ which includes the moving average component of degree 1. Even though a lag of 10 is significant, we will not include an $MA(10)$ component in our candidate model because it would include the all intervening lags which we find to be insignificant. In other words, the first candidate model is $\Delta Y_t = \epsilon_t + \theta \epsilon_{t-1}$.

From the sample partial autocorrelation function, we notice that similarly with the SACF, there

is a significant correlation for lags 1 and 10 and the correlation for all other lags are insignificant. Based on these observations, we choose an $ARIMA(1,1,0)$ model as the second candidate model which includes an auto-regressive component of degree one. This means that the model defines the price difference as a function of the previous price difference. In other words, the second candidate model is $\Delta Y_t = \phi \Delta Y_{t-1} + \epsilon_t$. An $AR(10)$ component is not included in the candidate model since all intervening lags are insignificant.

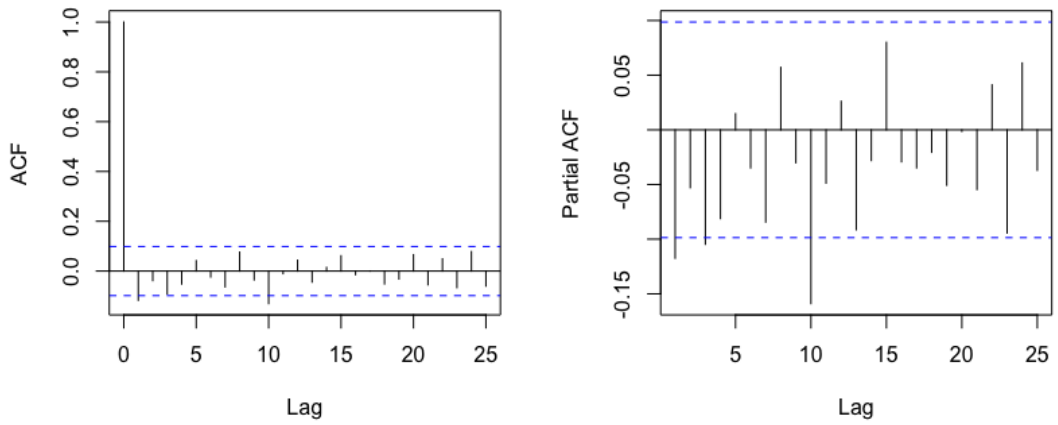

Figure 4.5: SACF and SPACF of Post-Recession Price Difference

Finally, the third candidate model is an $ARIMA(1,1,1)$ model which combines the auto-regressive and moving average components of the previous two candidate models. In other words, the third candidate model is $\Delta Y_t = \phi \Delta Y_{t-1} + \epsilon_t + \theta \epsilon_{t-1}$. Notice that the candidate models chosen are the same when considering the entirety of the data and only the post-recession data. However, this does not mean that the models will be the same for both data sets. Because the models will be built using different training data, the model fits will differ as well as the parameter estimates.

Three candidate UBJ models have been identified for both sets of training data which means that a final model can be selected for each using a selection criterion. For consistency, we use the same criterion used for regression modeling. We again use the Akaike Information Criterion (AIC) defined as $AIC = 2k - 2log(L)$ where k is the number of unknown parameters and L is the likelihood function. As discussed previously, the model with the smallest AIC is selected since a model that is least complex with the best fit is desirable. Table 4.1 gives the value of the selection

criterion for each of the candidate models.

Table 4.1: Candidate Models and AIC

Model	Equation	AIC
ARIMA(0,1,1)	$\Delta Y_t = \epsilon_t + \theta\epsilon_{t-1}$	2947.85
ARIMA(1,1,0)	$\Delta Y_t = \phi\Delta Y_{t-1} + \epsilon_t$	2947.35
ARIMA(1,1,1)	$\Delta Y_t = \phi\Delta Y_{t-1} + \epsilon_t + \theta\epsilon_{t-1}$	2948.98

Based on the AIC, all of the models have similar fits. The largest AIC is for the $ARIMA(1,1,1)$ model meaning that the adding the additional complexity to the model with the additional unknown parameter does not improve on the fit. The $ARIMA(0,1,1)$ and $ARIMA(1,1,0)$ are very similar with AIC values of 2947.85 and 2947.35 respectively. However, since the $ARIMA(1,1,0)$ has the smallest AIC, it is chosen as the best fit. This means that the difference in closing price is best modeled as a function of the previous price difference when considering the entire data set.

Next we consider model selection process for the candidate models based on the post-recession data. Table 4.2 gives the value of the AIC for each of the candidate models. Even though the candidate models are the same for the post-recession data, the fits will differ since different training data are involved.

Table 4.2: Post-Recession Candidate Models and AIC

Model	Equation	AIC
ARIMA(0,1,1)	$\Delta Y_t = \epsilon_t + \theta\epsilon_{t-1}$	1337.058
ARIMA(1,1,0)	$\Delta Y_t = \phi\Delta Y_{t-1} + \epsilon_t$	1337.651
ARIMA(1,1,1)	$\Delta Y_t = \phi\Delta Y_{t-1} + \epsilon_t + \theta\epsilon_{t-1}$	1333.520

The first thing that we can notice is that the AIC are significantly lower for each of the candidate models when the post-recession data are used rather than the entirety of the data. In other words, we are able to produce models with better fits when only modeling based on the data occurring after the recession. This is what we would expect since the recession is an anomaly and does not represent the typical trend for the closing stock prices. When considering the post-recession data, the model with the best fit is an $ARIMA(1,1,1)$ since it has the smallest AIC with a value of 1333.520. This is interesting since this model contains more unknown parameters than the

other candidate models meaning that the added complexity of the model is outweighed by the improvement of the fit. The $ARIMA(1,1,1)$ is chosen as the final model since it is the model with the smallest AIC.

4.3 Diagnostics

Before we interpret the final model, we first must consider the diagnostics of the models including the significance of the parameter estimates and the assumptions of the model. We have described the model as $\Delta Y_t = \phi \Delta Y_{t-1} + \epsilon_t + \theta \epsilon_{t-1}$, however the error for the current week is not obtainable. Therefore, to use the model, we write it in terms as the predicted price difference

$$\Delta \hat{Y}_t = \hat{\phi} \Delta Y_{t-1} + \hat{\theta} \epsilon_{t-1},$$

where the error is represented by the difference between the predicted price difference and the actual value. The values of the estimates for the unknown parameters are given in table 4.3.

Table 4.3: Coefficients for the ARIMA(1,1,1) Model

Parameter	Estimate	p-value
ϕ	0.70381	$<$0.0001
θ	-0.82398	$<$0.0001

From the table we see that the coefficient for the auto-regressive component is 0.70381 with a corresponding p-value of less than 0.0001 which means that the auto-regressive component, or the previous price difference, is significant in the prediction of the price difference. The coefficient for the moving average component is -0.82398 with a corresponding p-value of less than 0.0001 which means that the moving average component, or the previous error term, is significant in the prediction of the price difference.

Next, we check to see if the selected model meets the model assumptions. The assumptions are normality, constant variance, and independence of the residuals. First we will check to see if the residuals are normally distributed. Normality can be checked visually using a Q-Q Normal Plot and tested formally using the Anderson-Darling test. Figure 4.6 shows the Q-Q plot of the residuals. As

discussed previously, the plot gives the values of the standardized residuals against the theoretical values based on a normal distribution. Therefore, if the residuals come from a normal distribution, the plot should follow a straight line. Based on the plot, we see that the residuals tend to follow the normal model except for extremely large or small values that occur on the endpoints.

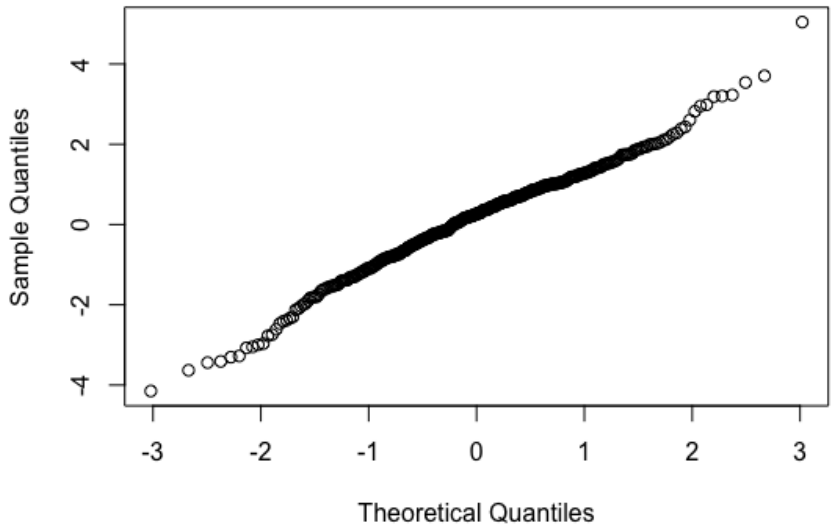

Figure 4.6: Normal Q-Q Plot for Residuals of Time Series Model

We use the Anderson-Darling test to test the null hypothesis that the standardized residuals are normally distributed against the alternative hypothesis that the residuals are not normally distributed. The value of the test statistic is $A = 1.5122$ with a corresponding p-value of 0.0007 which indicates that there is sufficient evidence to reject the null hypothesis and conclude that the residuals are not normally distributed. This means that the assumption of normality is not satisfied.

To check for the independence and constant variance of the residuals, we look to some diagnostics plots shown in Figure 4.7.

The second model assumption is that the variance of the residuals is constant over time. From the plot of the standardized residuals, we see that they tend to be centered around zero or the mean. The variance over time is represented by how large or small the residuals are. It appears that there tends to be a larger amount of variance for more current periods of time which was an observation that was also noticeable from the original time series plot. This indicates a possible problem with the assumption of constant variance in the model.

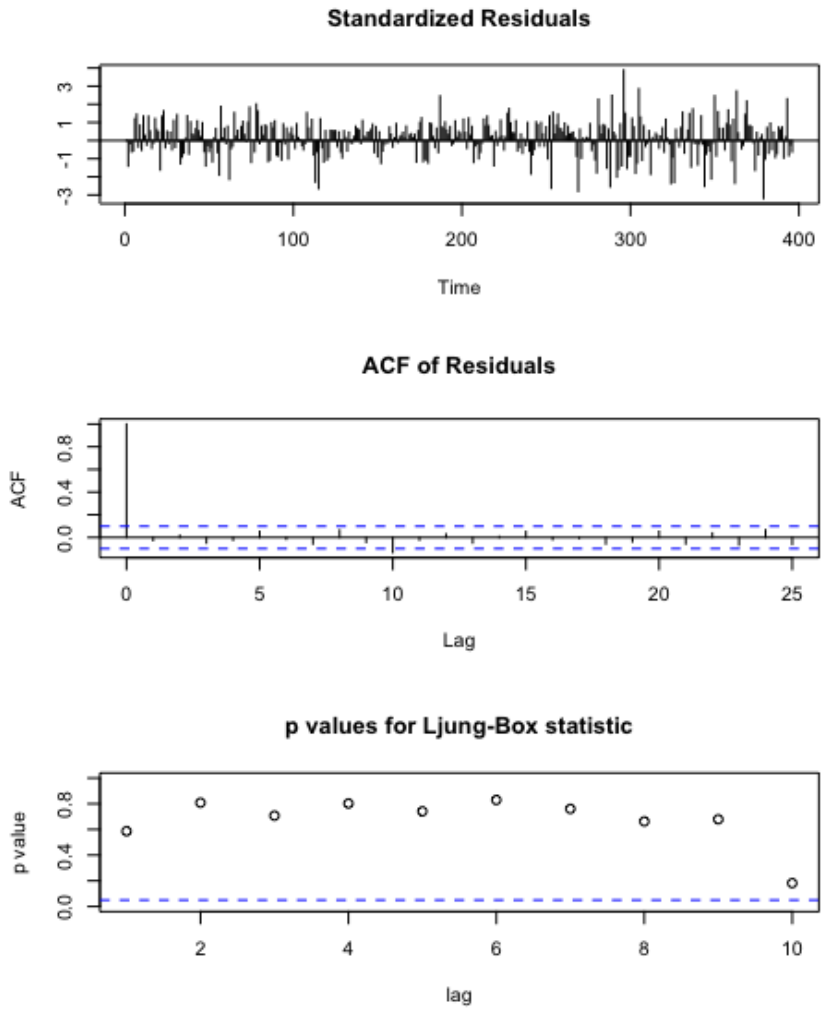

Figure 4.7: Time Series Model Diagnostics Plots

The third model assumption is that the residuals are independent from each other. To check this assumption, we first look at the autocorrelation function of the residuals. Based on the plot, it appears that the correlation is insignificant for all lag values except for zero which is expected. Based on the ACF, it would appear that the assumption of independence is satisfied. Independence can also be formally tested using the Ljung-Box test which tests the null hypothesis that the residuals are independent against the alternative hypothesis that the residuals are correlated. The p-values for the test are visualized in the plot since the test is performed for every lag value. For each value of the lag, the p-value corresponding to the Ljung-Box test is significantly greater than 0.05 indicating that at a 5% significance level, there is not sufficient evidence to reject the null hypothesis

and we conclude that the residuals are independent. Therefore, we conclude that the independence assumption is satisfied. The results of the diagnostics are significant because the time series model, unlike the regression model, is able to provide independent residuals.

4.4 Forecasting and Model Interpretation

Now that the final model is selected, we now wish to interpret the model, test the fit of the model, and check the accuracy of the model's predictions. We have defined the model in terms of the first order difference in price, however, the model can be rewritten in terms of the median closing price by simply substituting Δ_t with $Y_t - Y_{t-1}$. The substitution yields:

$$\begin{aligned} \Delta Y_t &= \phi \Delta Y_{t-1} + \theta \epsilon_{t-1} \\ Y_t - Y_{t-1} &= \phi(Y_{t-1} - Y_{t-2}) + \theta \epsilon_{t-1} \\ Y_t &= (1+\phi) Y_{t-1} + \theta \epsilon_{t-1}. \end{aligned}$$

Since we have the estimated values of the model parameters, we obtain the final equation of the time series model as

$$\hat{Y_t} = 1.7038 Y_{t-1} - 0.8240 \epsilon_{t-1}.$$

The value $1 + \phi = 1.7038$ indicates that holding all other variables constant, when the median closing price increases by \$1, the median closing price for the next week is predicted to increase by an average of approximately \$1.70. The value $\theta = -0.8240$ indicates that holding all other variables constant, when the error increases by 1, the median closing price for the next week is predicted to decrease by an average of approximately 82 cents.

Finally, we use the time series model to make predictions for the weekly median closing price from January 7, 2017 through December 31, 2017. Figure 4.8 shows a plot of the training data along with the values predicted from the model.

From the forecasting plot, it is clear that the model retains the increasing trend of the median closing price over time. However, the model does not capture the periods of increase and decrease in price that occur in the short run. Instead, the time series model is better for describing the price

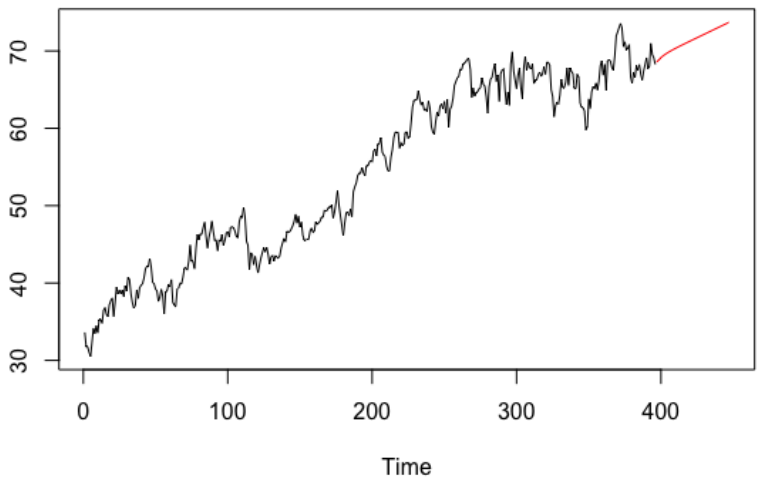

Figure 4.8: Time Series Forecasting

without considering short term changes or the 'white noise' that occurs. In other words, the model appears successful in predicting the overall trend in price over time, but is not useful for predicting short term anomalies.

The fit of the model can also be assessed by comparing the predicted values for the weekly closing price from January 7, 2017 through December 31, 2017 with the actual closing prices. These actual values represent the testing data set and were not used in the building of the model. Figure 4.9 shows a comparison of the time series plots for the predicted values and the actual values which illustrates the variation visible in the weekly closing price opposed to the steady rising trend provided by the model. In conclusion, the time series model can predict the increasing trend in price but cannot predict short-run variation.

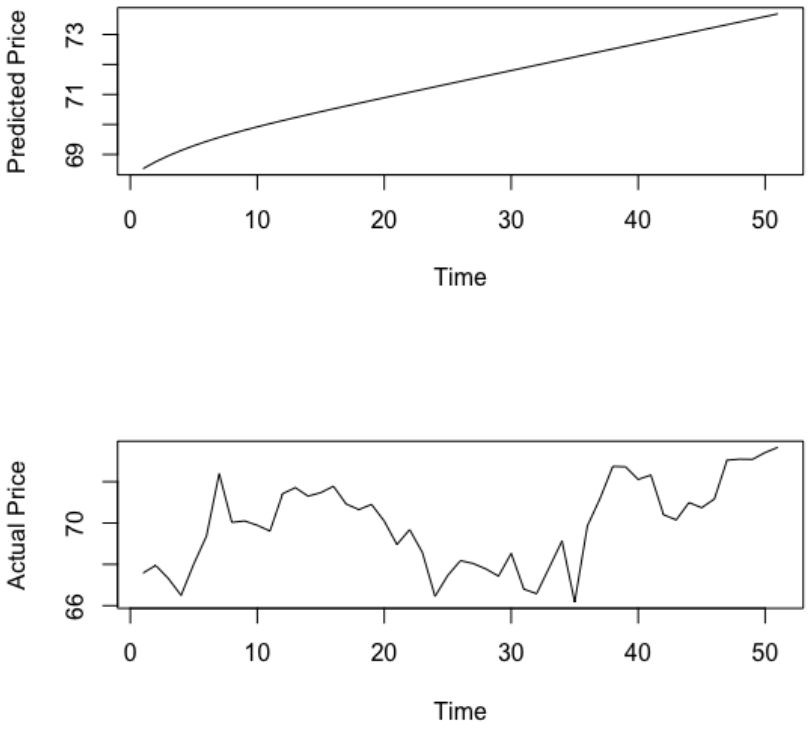

Figure 4.9: Time Series Plots for Predicted and Actual Prices

CHAPTER 5 VARYING INTERCEPT REGRESSION MODEL

5.1 Regression Revisited

Previously in this study, we have considered a pooled regression model and a time series model. These two models differ greatly since the regression model considers the effects of other macro and microeconomic factors that are related to stock prices while the times series model simply considers past observations based on the dependence of stock prices to historical prices. However, both models share an important commonality. Both models predict the median stock price at a specific week t. Instead of considering each stock individually, the median is used to describe the closing price for the top stocks of the NYSE as a whole. These types of models are useful to predict the overall state of the stock market and can be used as a market index for popular stocks in the NYSE. However, these models cannot be used to predict the prices for individual stocks, which is why we also consider a regression model with a varying intercept. If we consider each stock separately, the model then can be used by anyone who is invested into a particular stock within the top stocks in the the NYSE. This type of model is also useful for comparing the differences in prices over time for different stock types. Analyzing these patterns can be helpful to investors considering various stocks and which stocks tend to have higher or more steady prices.

First we look at the data that will be used for the modeling process. The pooled models used the median price meaning that there was only one observation for each of the 939 weeks collected. Since we are interested in each individual stock type, this model will use the entire data set collected which includes the closing stock price for each of the 939 weeks for each of the 85 stocks chosen from the NYSE 100. For the modeling process, the data set is randomly partitioned at a 70%-30% split for the training data which is used in the creation of the model and the testing data which is used to analyze the fit of the model. In other words, out of the 79,815 observations, 60,882 are in the training data and 18,993 are in the testing data. As described previously when discussing the pooled model, the distribution of stock prices is skewed right due to a minority of companies

that on average have high stock values. To normalize the distribution, the log transformation of the stock price will be used as shown in Figure 5.1.

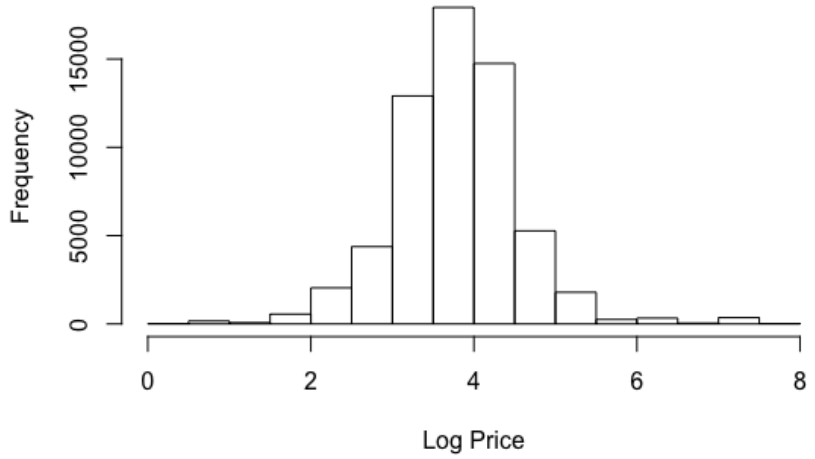

Figure 5.1: Distribution of Log Price

In the previous chapter, the specifics of the format of a multiple linear regression model were discussed. Here we highlight the differences in the multiple regression model when considering the median price over all stocks versus the closing price for each stock individually. The pooled regression model followed the following format

$$Y_t = \beta_0 + \beta_1 X_{t1} + ... + \beta_k X_{tk} + \epsilon_t,$$

where β_0 is the intercept and k is the number of predictor variables. For the varying intercept model, we consider

$$log(Y_{it}) = \sum_{i=1}^{84} [\alpha_i I_i] + \beta_0 + \beta_1 X_{t1} + ... + \beta_k X_{tk} + \epsilon_t,$$

where $i = 1, ..., 85$ represent the 85 stock indexes in the NYSE 100 that we are considering and I_i is an indicator function relating to each index i. In other words, the stock index is a dummy variable which corresponds to a unique coefficient α_i such that $\alpha_i + \beta_0$ represents a different intercept for each stock. It is for this reason that the model is referred to as having a varying intercept. Notice

that the model does not include a coefficient for the 85th index so that there is not an issue of multicollinearity among the variables. The intercept for the 85th stock index is represented simply by β_0. Also notice that the equation is being used to model $log(Y_{it})$ which is the log of the closing price for each index i for each week t.

5.2 Stepwise Variable Selection

Similarly with the modeling process for both the pooled regression model and the time series model, the Akaike Information Criterion (AIC) is again used as the selection criterion. The stepwise selection process begins with the null model and each step of the process then adds or removes variables until the smallest AIC is obtained. All variables that are considered in the process are identical to those considered in the pooled model except for two variables. The first variable is the dependent variable that is being modeled which is the log closing price for each stock index for each week. This difference is discussed in the previous section. The second variable that differs from the pooled model is the volume. In the pooled model, volume represents the weekly median volume sold across all stocks while in the varying intercept model, volume represents the weekly volume sold for each index individually. The other candidate predictor variables (Wilshire 5000 Index (WIL), Oil and Gas Index (XOI), Volatility Index (VIX), Gold Bugs Index (HUI), the interest rate for 10-year T-notes and bonds (TNX), Federal Funds Rate (DFF), Money Supply (M1), and TIME) remain the same since they are not related to specific stock therefore do not differ for each stock index.

Table 5.1 shows each of the iterations of the stepwise process for the training data. There are 9 iterations in the process and for each step a variable was added to the model. Unlike the process for the pooled model which yielded a full model, when considering each stock individually the Gold Index (HUI) was not added into the model. Also, if we compare the iterations to that of the pooled model, we notice that the order in which variables are added into the model differ. This indicates that the relationship between the closing price and the various micro and macroeconomic predictor variables differ when considering each stock individually as opposed to considering the median across all stocks.

Table 5.1: Iterations and AIC for the Stepwise Process

Iteration	Add/Remove	AIC
0	-	-31651.85
1	$+IDX$	-74701.11
2	$+WIL$	-90274.54
3	$+XOI$	-93120.6
4	$+V$	-96448.91
5	$+M1$	-96608.61
6	$+TIME$	-96829.34
7	$+VIX$	-96901.5
8	$+DFF$	-96935.42
9	$+TNX$	-96937.78

The process also shows that the categorical variable representing the stock index is the first variable added into the model. This indicates that the stock index is the variable most related to the closing price. This makes sense since some stocks can see increasing prices and other stocks can see decreasing prices over time depending on the financial health of its corresponding company.

Before selecting the final model, we must first analyze the relationship between the predictor variables to check for any possible problem of multicollinearity. As discussed previously, to remove the effects of multicollinearity, it is necessary to remove predictor variables that are highly correlated to other predictors. Again we rely on the Variance Inflation Factor (VIF) to quantify the correlation of the each predictor relative to the other predictors. Table 3.2 gives the VIF for each predictor in the model selected by the stepwise process. The table shows the the predictors with the highest VIF are the Wilshire 5000, M1, and time. It is not surprising that these variables would be highly correlated with other predictors since the Wilshire is representative of stock prices as a whole while M1 and time are both related to the effects of inflation. To remove the effects of multicollinearity, these three variables are removed from the model. The VIF for the remaining variables are also listed in the table. After removing the highly correlated predictors, we see that there no longer appears to be a problem of multicollinearity within the model.

Table 5.2: VIF of predictor variables

Variable	VIF	VIF
IDX	1.0033	1.0032
WIL	7.1368	-
XOI	2.7199	1.3988
V	1.3496	1.3332
$M1$	9.8433	-
$TIME$	6.5781	-
VIX	1.2708	1.1033
DFF	2.2827	1.8453
TNX	3.5399	2.1564

5.3 Diagnostics and Model Assumptions

Now that the variables for the final model are chosen, we run some diagnostics on the model. First, we look to see if each of the variables within the model are significant, then we check to see if the model meets the assumptions for linear regression.

If a variable is significant within the model, then that means that the slope or coefficient for that variable is significantly different from zero. We test the null hypothesis that the slope coefficient is equal to zero against the alternative hypothesis that the slope coefficient is not equal to zero. Tables 5.3, 5.4 and 5.5 show the test statistics and corresponding p-values for each of the predictor variables. For each variable, the corresponding p-value is less than 0.05 which indicates that there is sufficient evidence to reject the null hypothesis and conclude that the slope coefficients are significantly greater than zero. The only exception is the variable corresponding to the stock index DVN. However, we still include the variable since it represents one category for the stock index variable.

Now we check to see if the model meets the four underlying assumptions of the linear regression model. The first assumption is that there is a linear relationship between the log of the weekly closing stock price and the predictor variables. The relationship can be visualized from the plot of residuals versus the fitted values. If there is a linear relationship, the residuals should have a distribution centered around a straight line across all fitted values. The plot shows some curvature

of the relationship which indicates possible problems with linearity.

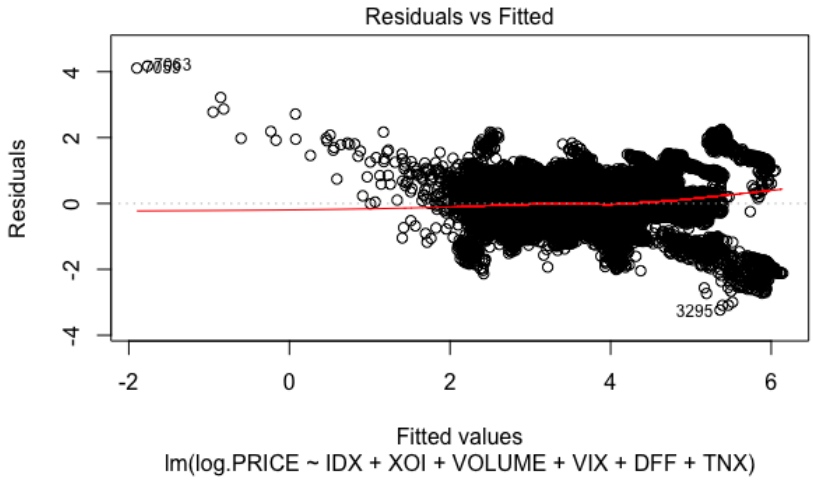

Figure 5.2: Residuals vs Fitted Values

The second assumption that we assess is the constant variance of the residuals. We can check for constant variance visually by also looking at the fitted residual plot. For the variance to be constant, the residuals should be spread out equally around zero for all fitted values. From the plot we see that this is not the case. For fitted values below 2, the values of the residuals are much higher than for other fitted values. This indicates a problem of heteroscedasticity among the residuals.

The third assumption is that the residuals are independent of one another. Independence is tested formally using the Durbin-Watson test where the null hypothesis that the autocorrelation among the residuals is zero is testing against the alternative hypothesis that the autocorrelation is greater than zero. The test statistic for the varying intercept model is $d = 0.037461$ with a corresponding p-values of less than 0.0001. This indicates that there is sufficient evidence to reject the null hypothesis and conclude that the autocorrelation among the residuals is greater than zero. In other words, we conclude that the independence assumption has been violated.

The final assumption is that the residuals are normally distributed. Normally can be tested visually using the Q-Q normal plot which plots the standardized residuals against the theoretical values from the normal model. If the residuals follow the normal model, the plot should follow a straight line. From the plot, we see that very large and small values tend to stray from the

theoretical quantities.

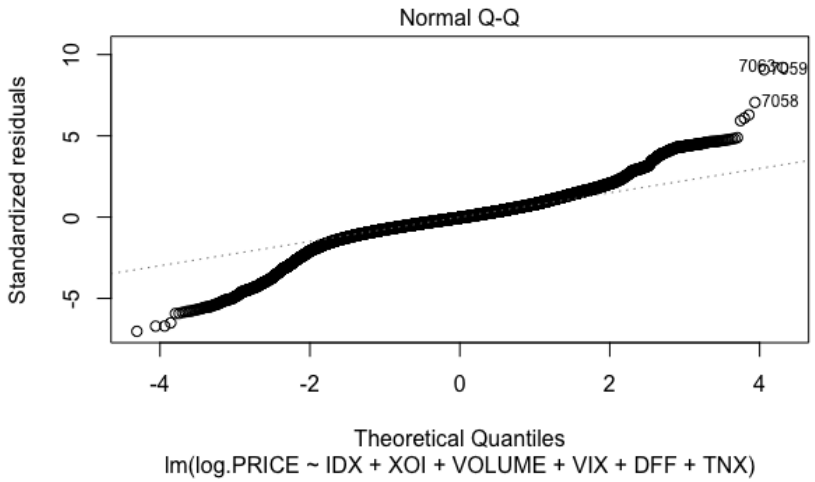

Figure 5.3: Normal Q-Q Plot

Normality is also tested formally using the Anderson-Darling test where the null hypothesis that the residuals are normally distributed is tested against the alternative hypothesis that the residuals are not normally distributed. The test statistic for the varying intercept model is $A = 686.87$ with a corresponding p-value of less than 0.0001. This indicates that there is sufficient evidence to reject the null hypothesis and conclude that the residuals are not normally distributed. This means that the normality assumption has been violated.

From the diagnostics, it is shown that each of the predictor variables are significant in the model and that there are serious issues with the assumptions of the model. In the next section, we look to see how useful the model is for making predictions given these violations.

5.4 Model Fit and Interpretation

The final model obtained for the varying intercept regression can be written as

$$log(\hat{Y_{it}}) = \sum_{i=1}^{84}[\hat{\alpha}_i I_i] + \hat{\beta}_0 + \hat{\beta}_1 XOI_t + \hat{\beta}_2 V_{it} + \hat{\beta}_3 VIX_t + \hat{\beta}_4 DFF_t + \hat{\beta}_5 TNX_t.$$

In the section we consider the fit of the model analyzing the coefficient of determination and by comparing the values predicted by the model to the testing data. We also look at the interpretation

of the model as a way to explain the relationship between the closing stock price and the predictor variables.

The coefficient of determination for the varying intercept model is $R^2 = 0.6433$ and indicates that 65.33% of the total variance in the log of the weekly closing stock price for the top stocks in the NYSE 100 is linearly associated with the variation in the weekly traded volume for each stock index (V), the Oil and Gas Index (XOI), the Volatility Index (VIX), the interest rate for 10-year T-notes and bonds (TNX), and the Federal Funds Rate (DFF). The percentage of explained variation is relatively high indicating that the model provides a good fit for the log weekly closing stock price. The coefficient of determination adjusted for the degrees of freedom is $R^2_{adj} = 0.6428$ and indicates that considering the complexity and sample size used in the model, 64.28% of the total variance in the log of the weekly closing stock price for the top stocks in the NYSE 100 is linearly associated with the variation in the weekly traded volume for each stock index (V), the Oil and Gas Index (XOI), the Volatility Index (VIX), the interest rate for 10-year T-notes and bonds (TNX), and the Federal Funds Rate (DFF). Even accounting for the degrees of freedom in the model, the explained variance is still high which indicates that the model is a good fit.

Interpreting the coefficients for the predictor variables gives insight on the relationship between the closing price for each stock and each predictor individually. The coefficient for XOI is 0.00051 and indicates that holding all other variables constant, when the NYSE ARCA Oil and Gas Index increases by one point, it is predicted that the closing stock price will increase by 0.0510%. This means that there is a positive relationship between the overall oil and gas prices and the prices of individual stocks in the NYSE.

The regression coefficient for V is -1.608e-09 and indicates that holding all other variables constant, when the weekly volume increases by one unit, it is estimated that the closing stock price will decrease by a percentage that is near zero, or 1.608e-07%. This means that when more volume of a stock is sold, the price tends to be cheaper. This makes sense because people tend to buy more when prices are lower. Also notice that the coefficient for volume is extremely small yet still significant based on the p-value. This is because the weekly volume of stock sold is extremely

high, therefore a change in one unit is very small. However, when considering larger changes in volume yields a more substantial predicted decrease in the price.

The slope coefficient for VIX is -0.007607 and indicates that holding all other variables constant, when the CBOE Volatility Index increases by one point, it is estimated that the closing stock price will decrease by an average of approximately 0.7578%. This indicates that there is a negative relationship between volatility and price meaning that when there is more volatility or uncertainty for the future, buyers tend to hold off and prices decrease.

The coefficient for DFF is 0.03756 and indicates that holding all other variables constant, when the Federal Funds Rate increases by one percent, it is predicted that the closing stock price will increase by an average of 3.8274%. This illustrates a positive relationship between the interest rate and price as also seen in the pooled model.

The coefficient for TNX is -0.1047 and indicates that holding all other variables constant, when the CBOE interest rate for 10 year T-note bonds increases by 1%, it is estimated that the closing stock price will decrease by an average of 9.94%. Similarly to the pooled model, we again see a negative relationship.

Lastly, we consider the intercepts of the model represented by the dummy variables for each individual stock. Since the coefficients for each dummy variable is significant but one, this indicates a significant difference in the closing stock prices for each variable.

Before the modeling process was started, the data are split into the training data, which the model is built off of, and the testing data. Since the testing data was not used in the building of the model, we compare the predicted values for the testing data based on the chosen varying intercept regression model against the actual testing data. From Figure 5.4, we see that there is indeed a rather strong correlation between the predicted and actual values. However, the relationship does not appear to be as strong as the one seen for the pooled regression model. This makes sense since the varying intercept model has a smaller coefficient of determination.

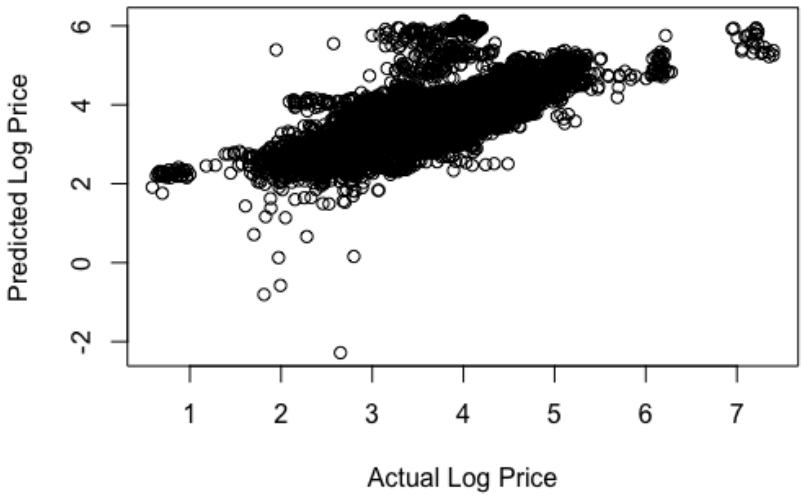

Figure 5.4: Predicted Data vs Actual Data

Table 5.3: Model Estimates I

Variable	Estimate	Test Statistic	p-value
Intercept	3.815e+00	165.346	$< .0001$
IDXABT	-5.100e-01	-22.212	$< .0001$
IDXAGN	3.220e-01	13.675	$< .0001$
IDXAIG	1.843e+00	78.965	$< .0001$
IDXAPA	1.883e-01	7.958	$< .0001$
IDXAPX	9.476e-02	4.000	$< .0001$
IDXBA	4.703e-01	19.635	$< .0001$
IDXBAC	-6.773e-02	-2.532	.0113
IDXBAX	-5.031e-01	-22.123	$< .0001$
IDXBEN	-5.226e-01	-22.509	$< .0001$
IDXBK	-2.584e-01	-10.978	$< .0001$
IDXBMY	-2.447e-01	-10.112	$< .0001$
IDXBRKB	4.976e-01	20.769	$< .0001$
IDXC	1.297e+00	55.567	$< .0001$
IDXCAT	1.993e-01	8.293	$< .0001$
IDXCCL	-2.156e-01	-8.869	$< .0001$
IDXCL	-1.635e-01	-6.806	$< .0001$
IDXCOP	-1.010e-01	-4.407	$< .0001$
IDXCVS	-2.281e-01	-9.442	$< .0001$
IDXCVX	4.561e-01	18.853	$< .0001$
IDXD	-9.544e-02	-3.963	$< .0001$
IDXDE	8.030e-02	3.344	$< .0001$
IDXDHR	-5.056e-01	-22.514	$< .0001$
IDXDIS	-1.553e-01	-6.630	$< .0001$
IDXDUK	1.851e-01	7.874	$< .0001$
IDXDVN	1.797e-02	0.739	.4598
IDXDWDP	-2.067e-01	-8.494	$< .0001$
IDXEMR	-1.065e-01	-4.403	$< .0001$
IDXEXC	-1.881e-01	-7.796	$< .0001$
IDXF	-1.141e+00	-46.816	$< .0001$
IDXFCX	-7.737e-01	-32.552	$< .0001$
IDXFDX	6.492e-01	27.391	$< .0001$
IDXGD	4.007e-01	16.783	$< .0001$
IDXGE	-2.213e-01	-9.038	$< .0001$
IDXGIS	-3.572e-01	-14.715	$< .0001$
IDXGLW	-9.511e-01	-39.500	$< .0001$
IDXGS	1.073e+00	46.136	$< .0001$
IDXHAL	-4.763e-01	-19.805	$< .0001$

Table 5.4: Model Estimates II

Variable	Estimate	Test Statistic	p-value
IDXHD	1.153e-01	4.771	$< .0001$
IDXHIG	-7.709e-02	-3.215	.0013
IDXHON	1.122e-01	4.955	$< .0001$
IDXHPQ	-1.005e+00	-43.466	$< .0001$
IDXIBM	9.789e-01	41.790	$< .0001$
IDXITW	1.136e-01	4.689	$< .0001$
IDXJMP	8.976e-02	3.673	$< .0001$
IDXJNJ	4.242e-01	17.738	$< .0001$
IDXKBM	3.926e-01	16.998	$< .0001$
IDXKO	-3.703e-01	-15.338	$< .0001$
IDXLLY	2.019e-01	8.416	$< .0001$
IDXLMT	6.368e-01	26.832	$< .0001$
IDXLOW	-3.907e-01	-16.257	$< .0001$
IDXMCD	1.888e-01	7.758	$< .0001$
IDXMDT	8.891e-02	3.677	.0002
IDXMMM	6.406e-01	27.096	$< .0001$
IDXMO	-6.111e-02	-2.539	.0111
IDXMRK	5.357e-02	2.231	.0257
IDXMRO	-9.647e-01	-41.726	$< .0001$
IDXMS	-9.943e-02	-4.135	$< .0001$
IDXNEM	-2.802e-01	-11.536	$< .0001$
IDXNKE	-1.016e+00	-43.930	$< .0001$
IDXNOV	-4.585e-01	-20.117	$< .0001$
IDXOXY	-5.327e-02	-2.367	.0179
IDXPEP	3.311e-01	13.631	$< .0001$
IDXPFE	-3.247e-01	-13.294	$< .0001$
IDXPG	2.825e-01	11.804	$< .0001$
IDXPNC	3.071e-01	12.716	$< .0001$
IDXPX	3.165e-01	13.055	$< .0001$
IDXRIG	7.307e-02	3.130	.0018
IDXSCCO	-1.252e+00	-53.332	$< .0001$
IDXSLB	1.766e-01	7.338	$< .0001$
IDXSO	-3.067e-01	-12.752	$< .0001$
IDXSPG	4.977e-01	22.136	$< .0001$
IDXT	-2.533e-01	-10.461	$< .0001$
IDXTGT	9.282e-02	3.808	.0001
IDXTRV	1.836e-01	7.646	$< .0001$
IDXTWX	1.893e-01	8.150	$< .0001$
IDXUNP	-3.432e-01	-14.616	$< .0001$
IDXUPS	4.530e-01	18.774	$< .0001$
IDXUSB	-4.123e-01	-17.083	$< .0001$

Table 5.5: Model Estimates III

Variable	Estimate	Test Statistic	p-value
IDXUTX	3.060e-01	12.823	$< .0001$
IDXVLO	-5.957e-01	-25.974	$< .0001$
IDXVZ	-1.207e-01	-5.198	$< .0001$
IDXWFC	-2.233e-01	-9.227	$< .0001$
IDXWMT	2.629e-01	10.814	$< .0001$
IDXXOM	4.266e-01	17.574	$< .0001$
XOI	5.100e-04	68.014	$< .0001$
VOLUME	-1.608e-09	-56.565	$< .0001$
VIX	-7.607e-03	-32.367	$< .0001$
DFF	3.756e-02	22.100	$< .0001$
TNX	-1.047e-01	-32.590	$< .0001$

CHAPTER 6 CONCLUSION

6.1 Model Comparisons

Now that each of the three models have been thoroughly explored, we now look to compare the benefits and drawbacks of the models. Table 6.1 gives the equations for the pooled regression, time series, and varying intercept regression models.

Table 6.1: Model Equations

Pooled Regression
$\hat{Y}_t = \hat{\beta}_0 + \hat{\beta}_1 WIL_t + \hat{\beta}_2 XOI_t + \hat{\beta}_3 VIX_t + \hat{\beta}_4 HUI_t + \hat{\beta}_5 TNX_t + \hat{\beta}_6 DFF_t$
Time Series
$Y_t = (1+\phi)Y_{t-1} + \theta\epsilon_{t-1}$
Varying Intercept Regression
$log(\hat{Y}_{it}) = \sum_{i=1}^{84}[\hat{\alpha}_i I_i] + \hat{\beta}_0 + \hat{\beta}_1 XOI_t + \hat{\beta}_2 V_{it} + \hat{\beta}_3 VIX_t + \hat{\beta}_4 DFF_t + \hat{\beta}_5 TNX_t$

First lets consider the time series model against the regression models. The benefit of the regression models over the time series model is that multiple regression allows the consideration of other variables as predictors and provides insight on the relationship between the closing stock price and these additional factors. A benefit of the time series model over the regression models is that the time series better fits the nature of the data where the closing price is highly correlated with the closing price of the previous week. For this data, more underlying assumptions of the time series model are satisfied over the underlying assumptions of linear regression. The time series model also uses previous observations which are more readily available information that current data which is used for prediction in the regression models. A benefit of the regression models is that time is not used as a variable but rather as an index. In other words, the model only requires knowledge of the values for the week of interest, not the time relative to other data points.

Now that we have compared the time series against the regression models, we consider the differences between the two regression models. The pooled regression model uses the median or 'pooled' weekly closing stock price over all of the stock indexes considered from the NYSE. The benefit of this is that the model can be used to give a comprehensive overview of the trends

of these selected stocks. The drawback to pooling the data is that the model cannot be used to predict individual stocks. On the other hand, the varying intercept model considers each stock individually which allows for investors interested in specific NYSE stock to compare the trends of each. However, the varying intercept model includes the index as a categorical variable which adds 84 dummy variables to the model making it a much more complex model than the pooled regression. Finally, the regression models can be compared by their predictive ability by considering the coefficient of determination or the amount of variation in the closing price that is explained by each model. The amount of variation explained by the pooled model is 96.61% and the amount of variation explained by the varying intercept model is much lower with 64.28%. In conclusion, the pooled model gives a general comprehensive view of the NYSE stocks overall with high accuracy in predictive power while the varying intercept model gives more in depth information on individual stocks at the cost of lower predictive capabilities.

6.2 Further Considerations

In the future, it would be interesting to consider some different models such as neural networks to predict the weekly closing stock price for each index since these types of models have more lenient model assumptions. This could be a more comprehensive analysis since the regression models do not tend to meet the underlying model assumptions.

It would also be interesting to consider individual time series models for each stock index rather than the pooled median price. This would be a study interesting for investors interested in specific stock within the ones selected from the NYSE.

BIBLIOGRAPHY

Al-Tamimi, H. A. H., A. A. Alwan, and A. A. A. Rahman (2011). Factors affecting stock prices in the uae financial markets. *Journal of Transnational Management 16*(1), 3–19.

Chang, P.-C., D.-d. Wang, and C.-l. Zhou (2012, January). A novel model by evolving partially connected neural network for stock price trend forecasting. *Expert Systems with Applications 39*(1), 611–620.

Cryer, J. D. and K.-S. Chan (2008). *Time Series Analysis With Applications in R* (2 ed.). Springer Texts in Statistics. Springer.

Dielman, T. E. (2005). *Applied Regression Analysis* (4 ed.). South-Western Cengage Learning.

Moghaddama, A. H., M. H. Moghaddamb, and M. Esfandyari (2016). Stock market index prediction using artificial neural network. *Journal of Economic, Finance and Administrative Science 21*, 89–93.

Online, a. Effective federal funds rate [dff]. Retrieved from https://fred.stlouisfed.org/series/DFF.

Online, b. M1 money stock [m1]. Board of Governors of the Federal Reserve System (US). Retrieved from https://fred.stlouisfed.org/series/M1.

Online, c. Cboe volatility index: Vix [vixcls]. Chicago Board Options Exchange. Retrieved from https://fred.stlouisfed.org/series/VIXCLS.

Online, d. Crude oil prices: West texas intermediate (wti). U.S. Energy Information Administration. Retrieved from https://fred.stlouisfed.org/series/WCOILWTICO.

Online, e. Wilshire 5000 total market full cap index. Wilshire Associates. Retrieved from https://fred.stlouisfed.org/series/WILL5000INDFC.

Sharif, T., H. Purohit, and R. Pillai (2015). Analysis of factors affecting share prices: The case of bahrain stock exchange. *International Journal of Economics and Finance 7*(3), 207–216.

APPENDIX SELECTED R PROGRAMS

Pooled Regression Model

- Calculate the Medians for Price and Volume

```
MPRICE<-sapply(split(Stock$PRICE, Stock$TIME), median)
MVOLUME<-sapply(split(Stock$VOLUME, Stock$TIME), median)
```

- Partition the Data

```
pooldata<-data.frame(cbind(MPRICE,MVOLUME, Pool))
set.seed(123)
ind<-sample.split(Y=pooldata$MPRICE, SplitRatio = 0.7)
train.pooldata<-pooldata[ind,]
valid.pooldata<-pooldata[!ind,]
```

- Stepwise Regression

```
null<-lm(MPRICE~1, data=train.pooldata)
full<-lm(MPRICE~., data=train.pooldata)
poolmodel<-step(null, data=train.pooldata, scope =
list(upper=full), direction = "both")
summary(poolmodel)
```

- Multicollinearity

```
vif(poolmodel)
# Remove Correlated Variables
poolmodel1<-lm(formula = MPRICE ~ WILSHIRE + XOI + VIX +
HUI + TNX + DFF + MVOLUME, data = train.pooldata)
vif(poolmodel1)
```

- Variable Significance

```
poolmodel2<-lm(formula = MPRICE ~ WILSHIRE + XOI + VIX +
HUI + TNX + DFF, data = train.pooldata)
summary(poolmodel2)
anova(poolmodel2)
```

- Check Model Assumptions

```
plot(poolmodel2)
#Durbin Watson Test
dwtest(poolmodel2, alternative = "greater")
#Anderson Darling Test
ad.test(poolmodel2$residuals)
```

- Testing Data in the Model

```
yhat<-predict.lm(poolmodel2, valid.pooldata)
plot(valid.pooldata$MPRICE, yhat, ylab = "Predicted Price", xlab =
"Actual Price")
```

Time Series Model

- Data Partition

```
tstrain<-MPRICE[-(889:939)]
tstest<-MPRICE[-(1:888)]
tstrain<-ts(tstrain)
tstest<-ts(tstest)
```

- Time Series, ACF, and PACF Plots

```
plot(DATE,MPRICE, type="n", xlab="Median Closing Price", ylab="Date")
lines(DATE,MPRICE, pch=16)
acf(tstrain)
pacf(tstrain)
```

- First Order Difference

```
lagdif1<-diff(tstrain,lag=1)
par(mfrow=c(1,1))
plot(lagdif1)
par(mfrow=c(1,2))
acf(lagdif1)
pacf(lagdif1)
acf(lagdif1,plot=F)
pacf(lagdif1,plot=F)
```

- Candidate Models

```
tsmodel1<-arima(tstrain,order=c(1,1,0))
tsmodel2<-arima(tstrain,order=c(0,1,1))
tsmodel3<-arima(tstrain,order=c(1,1,1))
AIC(tsmodel1,tsmodel2,tsmodel3)
```

- Diagnostic Testing

```
# Significance of Model Parameters
library(lmtest)
coeftest(tsmodel1)
#Anderson Darling Test
ad.test(tsmodel1$residuals)
qqnorm(tsmodel1$residuals)
# Constant Variance/ Independence
tsdiag(tsmodel1)
```

- Forecasting

```
nobs=length(tstrain)
tsfit=arima(tstrain, order=c(1,1,0), xreg=1:nobs)
```

```
fore=predict(tsfit , 51, newxreg=(nobs+1):(nobs+51))
ts.plot(tstrain ,fore$pred ,col=1:2)
```

Varying Intercept Regression Model

- Data Partition and Transformation

```
# Set seed to get the same results each time
set.seed(123)
# Set categorical variables (IDX and SECTOR) as factor
Stock[,1]=as.factor(Stock[,1])
Stock[,2]=as.factor(Stock[,2])
# Create Train and Validation data
ind<-sample.split(Y=Stock$PRICE,SplitRatio = 0.7)
train.Stock<-Stock[ind,]
valid.Stock<-Stock[!ind,]
# Transform Price
log.PRICE<-log(train.Stock$PRICE)
train.Stock1<-data.frame(cbind(train.Stock[,-3],log.PRICE))
```

- Model Selection

```
# Stepwise Regression
null<-lm(log.PRICE~1,data=train.Stock1)
full<-lm(log.PRICE~.,data=train.Stock1)
varmodel1<-step(null,data=train.Stock1,scope =
list(upper=full),direction = "both")
summary(varmodel1)
# Multicollinearity
vif(varmodel1)
vif(varmodel2)
```

```
# Final Model
varmodel2<-lm(log.PRICE ~ IDX + XOI + VOLUME + VIX + DFF +
TNX, data = train.Stock1)
summary(varmodel2)
```

- Checking Model Assumptions

```
plot(varmodel2)
# Durbin Watson Test
dwtest(varmodel2,alternative = "greater")
# Anderson Darling Test
ad.test(varmodel2$residuals)
```