Journal of Mathematics and Statistics

Original Research Paper

Model Comparison for the Prediction of Stock Prices in the NYSE

[1]Victoria Switlyk and [2]Junfeng Shang

[1]PartnerShip LLC, Westlake, OH 44145, USA
[2]Department of Mathematics and Statistics, Bowling Green State University, OH 43403, USA

Article history
Received: 23-07-2019
Revised: 14-09-2019
Accepted: 24-09-2019

Corresponding Author:
Junfeng Shang
Department of Mathematics and Statistics Bowling Green State University, OH 43403, USA
Tel: (419)372-7457
Email: jshang@bgsu.edu

Abstract: The stock market is an integral part of investments as well as the economy. The prediction of stock prices is an exciting and challenging problem that has been considered by many due to the complexity and noise within the market and to the potential profit that can be yielded from accurate predictions. We aim to construct and compare models used for the prediction of weekly closing prices for some of the top stocks in the New York Stock Exchange (NYSE) and to discuss the relationship between stock prices and the predictor variables. Relationships explored in the study include that with macroeconomic variables such as the Federal Funds Rate and the M1 money supply and 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 involving 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.

Keywords: Model Comparison, Regression, Time Series Analysis, Varying Intercept, Stock Market

Introduction

Introduction to Data

Data Description

When it comes to creating an investment portfolio to build up held assets, there are several different asset classes to choose including bonds, cash equivalents, and equities. Equities, or stocks, are the most volatile but can generate a large profit. It is crucial to understand patterns in equity prices as it is vital to those who wish to invest in the stock market.
The study focuses on stocks within a specific exchange rather than an index such as the S&P 500, so that the values estimated can 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. Out of all possible exchanges, the NYSE is chosen 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. When looking at daily stock data, there are missing data for weekends and holidays. Excluding these data points would make the intervals between time points unequal, which would not be appropriate for time series analysis. Instead, this study uses weekly time points. There can be a great amount of variation among stock prices within a day, which can be largely attributed to white noise. Instead, this study considers trends over longer periods of time.

New York Stock Exchange 100

There are approximately 2,800 companies that have equities listed in the NYSE. The New York Stock Exchange 100, or NYSE 100, composes of a list of very promising, high achieving, and popular stocks. Since this group consists of successful and well-established stocks, it

provides a sample of stocks for this model that is relevant in application. The study investigates weekly closing stock prices in the NYSE 100 for each Friday from January 1, 2000 through December 23, 2017 and 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.

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 Yit for each stock i = 1,..., 85 for each week t = 1,..., 939.

One variable that is used as a predictor is the volume of stock sold each week. The volume sold in the previous week is used to predict the volume sold in the current week. This study defines the volume sold as V_{it} for each stock i = 1,..., 85 for each week t = 1,..., 939.

The second predictor 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 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.

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. These variables tend to be recorded monthly or annually rather than weekly and are useful for observing trends over long periods of time.

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. The 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. If the funds rate is low, the cost of exchanging funds decreases, encouraging borrowing and spending, leading to increased stock prices. The data is monthly, 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). The study defines *DFF* as DFF_t for each week t = 1,..., 939 (Online (a), 2018).

M1 is the entire supply of physical money in the United States and is composed of federal notes and coins and 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 and 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. The study defines M1 as M1t for each week t = 1,..., 939 (Online (b), 2018).

Market Indexes

Unlike the macroeconomic factors which give a broad look on the state of the economy, market indexes can be utilized to inspect 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), 2018).

TNX is the CBOE's index that measures the interest rate for 10-year T-notes and bonds. Equities 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 study defines TNX as TNX_t for each week t = 1,..., 939.

The Wilshire 5000 Total Market Full Cap 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, and 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), 2018).

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 are distributed in 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. The 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 also retrieved from Yahoo! Finance. This study defines HUI as HUI_t for each week t = 1,..., 939 (Online (d), 2018).

Introduction to the Study

In literature, a 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) seen in studies such as one by Moghaddama *et al.* (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 black box building relations between the stock prices and the predictors making it difficult to interpret the model and the relationships therein. Instead, this study turns to more classical methods including multiple linear regression and time series analysis to provide more meaningful interpretations.

Chang *et al*. (2012) study the daily stock trends using another type of neural network, and the Evolving Partially Connected Neural Network (EPCNN) explains 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 this study considers a longer time interval using weekly data to account for this short term environment. Chang *et al.* (2012) also expresses the strong relationship between stock trends and other outside factors. This study therefore considers many other economic variables as described in last subsection.

Previous studies such as those by (Al-Tamimi *et al*., 2011; Sharif *et al*., 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. 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.

Section 2 is dedicated to the pooled multiple linear regression model which is based on the median weekly stock price over all indexes. Section 3 discusses the time series model which takes advantage of the time series nature of the data. Section 4 is a multiple linear regression model that considers each of the 85 selected stocks in the NYSE individually. This study is concluded in Section 5 which gives comparisons of the three models.

Pooled Regression Model

Multiple Linear Regression

Regression models are used to show how the variation of stock prices are related to the predictor variables.
Since the data are time series, time is used as a predictor.

The pooled model considers predicting Y_t which represents the median closing stock price for each week t. When pooling the closing price over all stock indexes, the median is used rather than the average because the average merges the variability within the data over time while the median will retain the patterns of variation. 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 significantly higher stock prices than other companies.

The volume of stock i sold at time t is also a variable that depends on the stock index. Similarly, with the closing price, this model considers the median volume sold at time t over each stock i, denoted by V_{it}.

When conducting regression modeling, the data are partitioned into two parts: Training and testing data. The training data is used in the process of fitting the model. The testing data is used to check the fit of the 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.

We comment that these two percentages are a common choice for splitting the data and the size of 939 observations is big enough for the estimation of the models. The other possible choices could be 80%-20% and 50%-50% splits.

For the model fitting, it is assumed that the relationship between the median closing stock price Y_t at week t and the selected set of predictors variables, X_{t1},..., X_{tk}, roughly follows the linear regression model:

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

where $\in_t$ is a random variable that represents the error. Let β_0 be the intercept and β_1,...,β_k be the parameter coefficients for the predictors. All β's are defined as fixed and unknown parameters. The regression model uses the least squares estimates of β_1, ...,β_k which are the values that minimize the residual sum of squares.

The notation for a multiple regression model can be simplified by writing the model in terms of vectors and matrices. Y is set as the vector of median closing prices Y_{tj} from t = 1,..., 657 and X as the (k+1)× 657 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. β is the vector of the unknown parameters β_1, ...,β_k and $\in$ is the vector of error terms from t = 1,..., 657. The model can then be rewritten as $Y = X\beta + \in$.

Stepwise Variable Selection

Classically, there are three popular methods of variable selection: forward selection, backwards elimination, and stepwise selection. Each of these methods determines which predictor variables should be included in the model with the goal of minimizing a selection criterion such as the Akaike Information Criterion (AIC), which is defined as:

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

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 $Y \sim N(X\beta, \sigma^2)$.

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 by maximizing the likelihood function. The added 2k penalizes the complexity of the model. 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 $Y = X\beta + \in$. 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 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 meaning that each variable increased the maximum likelihood function so as to outweigh the cost of adding an additional variable. There are 9 iterations, and for each iteration a variable was introduced to the model. For each iteration, the AIC decreases at a slower rate, indicating that there are diminishing returns for the reduction of AIC due to the cost of adding an additional variable. Finally, note that the variables introduced into the model first are the variables that increase the likelihood function the most.

Model Assumptions

To obtain the final mode, the significance of the predictor and possible multicollinearity between predictors must be analyzed.

Multicollinearity in a model occurs when predictors are highly correlated with one another. 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 1 gives the correlation between each of the variables in the model rounded to the nearest tenth. Some of the strongest correlations involving predictors 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. The result of inflation over time can be an underlying effect for correlation between the predictors and time.

To deal with multicollinearity, it is best to remove predictors that are highly correlated with other predictors. The Variance Inflation Factor (VIF) for each variable, as shown in Table 2, can be used to determine which predictors should be removed from the model and is defined as:

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

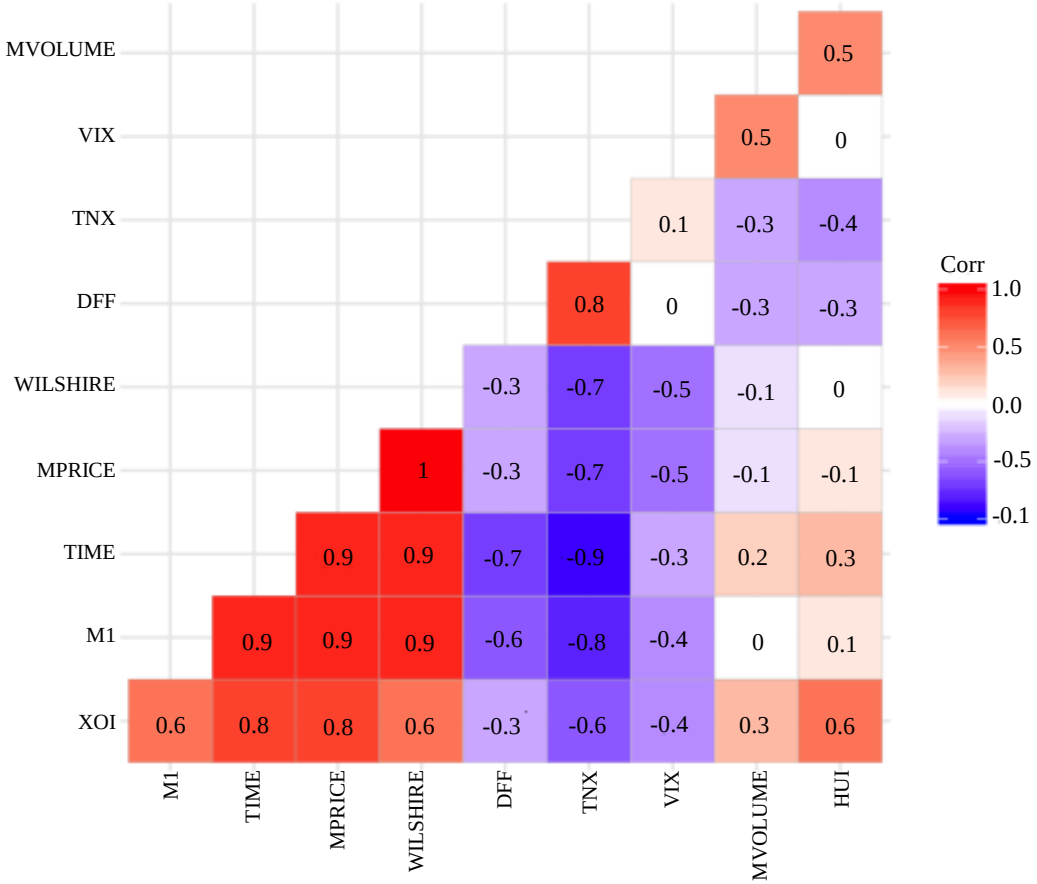

Fig. 1: Correlation Matrix

Table 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	+V IX	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

Table 2: VIF of predictor variables

Variable	VIF	VIF
WIL	53.12	5.79
XOI	7.52	4.35
V IX	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

For the predictors $X_1,..., X_k$, R_j^2 is the correlation coefficient for the fit of X_j on the remaining k-1 variables. The correlation coefficient 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.

M1 is the most highly correlated with the other predictor variables with a *VIF* of 133.05. Considering a reduced model with *M1* removed solves the multicollinearity problem.

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. The significance is tested as:

$$H_0: E(Y) = \beta_0 + \beta_1 X_1 + ... + \beta_7 X_7$$
$$H_\alpha: E(Y) = \beta_0 + \beta_1 X_1 + ... + \beta_6 X_6.$$

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 remaining predictor variables are included. 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. The corresponding p-value results for the tests are shown in Table 3.

Most *p*-values for the *F* test statistics (Table 3) are close to zero indicating that for each of those predictors is significant in the model. The only insignificant predictor is the median volume sold in week *t* (*p*-value is 0.7298) and it is concluded that the median volume is not significant for predicting the median closing price when also considering the other predictors.

After removing volume, the final pooled model is defined as:

$$\hat{Y}_t = \hat{\beta}_0 + \beta_1 WIL_t + \hat{\beta}_2 XOI_t + \hat{\beta}_3 VIX_t + \hat{\beta}_4 HUI_t + \hat{\beta}_5 TNX_t + \hat{\beta}_6 DFF_t,$$

where $\hat{Y}$ is the predicted median closing price at week *t*, $\hat{\beta}_0, ..., \beta_6$ are the parameter estimates for the 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*).

The four main assumptions of the model as shown by Dielman (2005), are linearity between the closing price and the predictor variables, independence of the error terms or residuals, constant variance or homoscedasticity of the residuals, and normality for the distribution of the residuals.

Based on the *Q-Q* plot shown in Fig. 2, points near zero tend to follow a straight line while points farther out tend to stray farther from the line and create a curved shape on the end points. This indicates that values closer to the zero are closer to the theoretical values following the normal. Values on the end points of the graph indicate values that differ greatly from the normal model. The unusual prices can be due to external effects or anomalies in the market or across the economy as a group.

Interpretation and Fit of the Model

The most common way to assess the fit or the predictive capabilities of the model is with the coefficient of determination which 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:

Table 3: Significance of the predictor

Variable	F-Statistic	p-value
WIL	17493.0100	<0.0001
XOI	941.1400	<.0001
V IX	123.0200	<.0001
HUI	53.8500	<.0001
TNX	6.7200	.0098
DFF	35.5000	<.0001
V	.1194	.7298

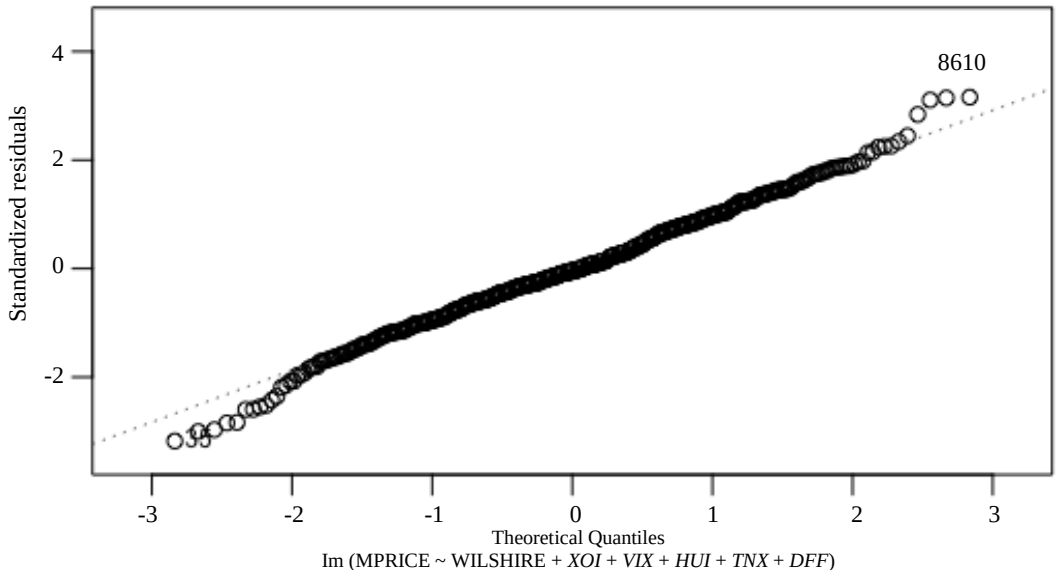

Fig. 2: Normal Q-Q Plot

Table 4: Estimates of model parameters

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

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

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. Therefore, R^2 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 explains more of the variation in the stock price. The coefficient of determination for the pooled model is R^2 = 0: 9664, indicating 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 predictors are added to the model. This means that based solely on the R^2 a better model would be a model with more predictors. 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. The adjusted coefficient of determination accounts for the complexity of the model and is defined as:

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

While the unadjusted value will always increase with each additional predictor added to the model, the adjusted value will only increase if the additional variation explained by the added predictor is greater considering the added complexity to the model. Given the two values, R_{adj}^2 is preferred given the consideration of model complexity. The adjusted coefficient of determination for the pooled model is R_{adj}^2 = 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 coefficients feature 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 4 lists the independent variables in the final model along with the corresponding 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 *WIL* is 0.3539 and indicates that holding all other 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 and 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, 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. Although there is a correlation between the two, they are not considered to be equivalent assets. That is, 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, 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. 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.

The predicted values are obtained by plugging in the values for each of the predictors for each week into the model. The testing data include the values for 282 randomly selected weeks from the original data set. Figure 3 features a plot of the testing data representing the actual median closing prices and the predicted closing prices obtained from the model.

If the model is successful, the predicted values are close to the actual values. 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, it is concluded that there is not a problem of overfitting of the model to the training data.

Time Series Analysis

Exploratory Data Analysis

Time series analysis is available uniquely to data that occur sequentially over intervals of time. The purpose of this analysis is to forecast the weekly median closing price for the top stocks in the NYSE as is done in the pooled regression model from the previous chapter.

The consideration of exploratory data analysis of the time series data is required before modeling can occur. Patterns are visualized through a time series plot of the data as shown in Fig. 4 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 over time, the median weekly closing price tends to increase, since inflation will drive prices higher. The data analyzed in this study has been recorded over 17 years which is a substantial period such that the effects of inflation are visible.

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 values 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 were not a large as the effects of the Great Depression, the effects of the former lasted for such a significant period and then spread across the world that it still holds the title Great.

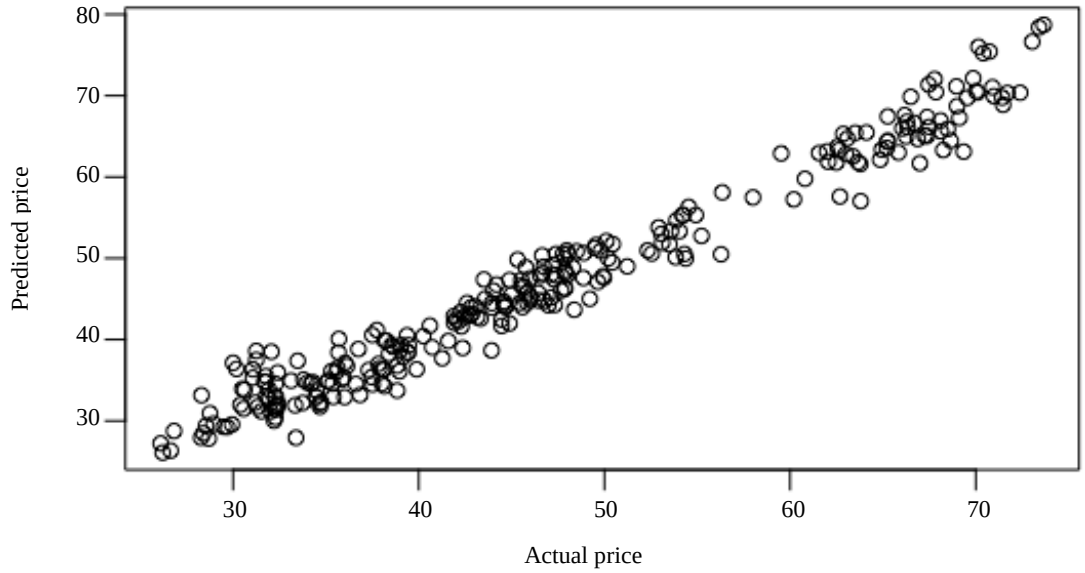

Fig. 3: Fit of testing data vs predicted values

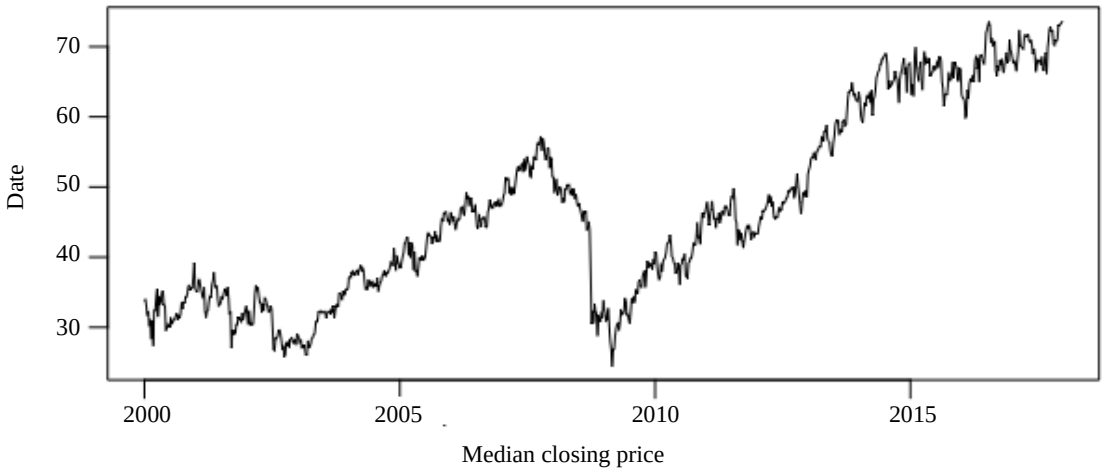

Fig. 4: Closing price for each week

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.

From the time series plot in Fig. 4, there are 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 over fitted to this anomaly. Therefore, only the data after the Great Recession is used.

Stationarity can 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 it is concluded that the data are non-stationary. 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.

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 which uses the difference or change in price over each week. In other words, the response of model is represented as $\Delta Y_t = Y_t - Y_{t-1}$.

Similarly, with the construction of the pooled model, the data for two-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. 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.

Model Identification and Selection

It has been determined that the differenced data, $\Delta Y_t = Y_t - Y_{t-1}$ are stationary. The differenced data can be expressed as:

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

where $\in_t$ represents the white noise that is assumed to be normally distributed with mean 0 and variance $\sigma_{\in}^2$. 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, the notation can be simplified. The mean is written as $E(\Delta Y_t) = \mu_t = \mu$ since there is a constant mean difference. The variance is written 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 γh since the covariance is a function of the lag. Based on these observations, the correlation between any two observations ΔY_i and ΔY_j where $|i-j| = h$, is defined as $\rho h = \frac{\gamma_h}{\gamma_0}$ which, is the Autocorrelation Function (ACF).

The auto-regressive model of degree p, AR (p), is defined as:

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

The process is based on the previous p observations. Consider a simple AR (1) model. If the equation for the AR (1) is multiplied by ΔY_{t-h} and the expected values are taken, 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 Cryer and Chan (2008). Based on the theoretical values of the autocorrelation function for an AR (1) model, the process can be identified 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, the Partial Autocorrelation Function (PACF) is utilized. 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, $Y_{t-1},...,Y_{t-k+1}$. So we have ϕ_{kk} = Corr (Y_t, $Y_{t-k}|Y_{t-1},....,Y_{t-k+1}$).

As shown in 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, the moving average model of degree q, MA (q), is considered and defined as:

$$\Delta Y_t = \in_t + \theta_1 \in_{t-1} + ... + \in_{t-q.}$$

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. The simplest moving average model is one of degree one where $\Delta Y_t = \in_t + \theta_1 \in_t + \theta_1 \in_{t-1}$. 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.

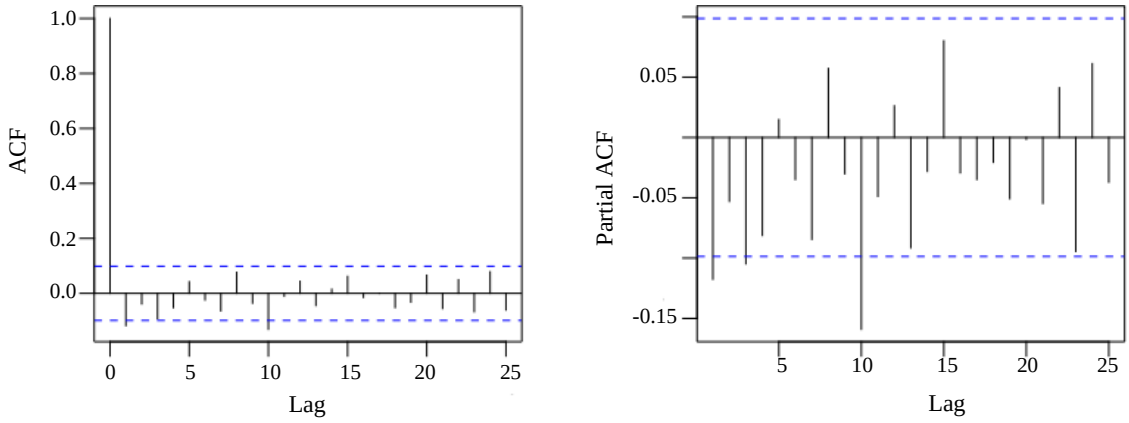

Fig. 5: SACF and SPACF of post-recession price difference

Table 5: Post-recession candidate models and AIC

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

The Sample Autocorrelation Functions (SACF) and Sample Partial Autocorrelation Functions (SPACF) of the differenced data can be analyzed in order to identify appropriate *ARIMA* (*p*, 1, *q*) candidate models.

From the autocorrelation function, the correlation for a lag of zero is 1. This is because any point is 100% correlated with itself. There is a significant correlation when the lag is 1, meaning that there is a significant relationship between the difference in price between two weeks and the difference in price for the week prior, also meaning that there is a significant correlation between ΔY_t and ΔY_{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 and ΔY_{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 is an *ARIMA* (0, 1, 1) or $\Delta Y_t = \in_t + \theta \in_{t-1}$.

From Fig. 5, the sample partial autocorrelation function 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 is an *ARIMA* (1, 1, 0) or $\Delta Y_t = \phi \Delta Y_{t-1} + \in_t$.

For consistency, the Akaike Information Criterion (AIC) is used as the selection criterion. Table 5 provides the AIC values for the candidate models.

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.

Diagnostics

The model has been described as $\Delta Y_t = \phi \Delta Y_{t-1} + \in_t + \theta \in_{t-1}$, however the error for the current week is not obtainable. Therefore, to use the model, it is written in terms as the predicted price difference:

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

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 6.

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 also significant.

Next, it is necessary to check to see if the selected model meets the model assumptions of normality, constant variance, and independence of the residuals. For normality, Anderson-Darling tests 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.

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 in Fig. 4.

Independence can 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. Figure 6 shows the time series model diagnostics plots.

Forecasting and Model Interpretation

Since the estimated values of the model parameters have been obtained, the final equation of the time series model is written as:

$$\hat{Y}_t = 1.7038 Y_{t-1} - 0.8240 \in_{t-1}.$$

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

From the forecasting plot, 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 without considering short term changes or the "white noise" that occurs. The model therefore appears successful in predicting the overall trend in price over time, but 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 8 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.

Table 6: Coefficients for the ARIMA(1,1,1) model

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

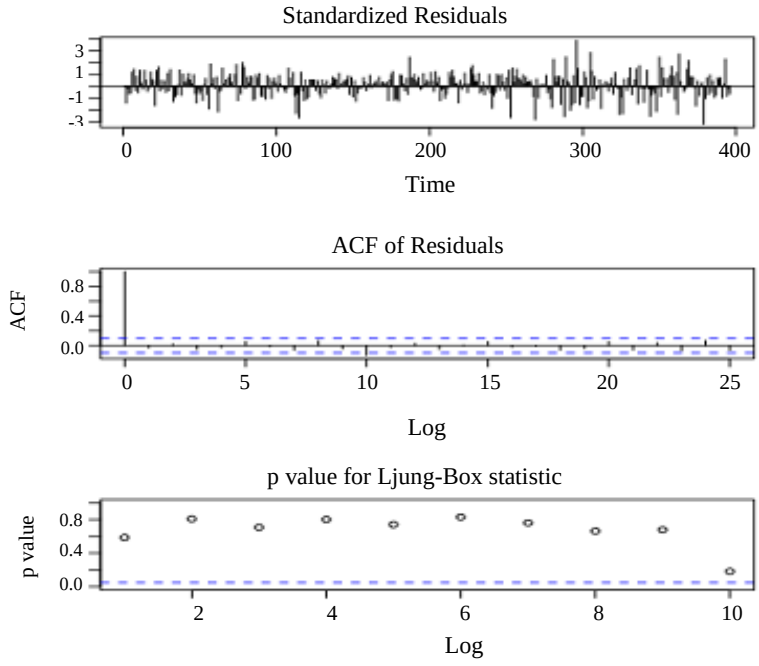

Fig. 6: Time series model diagnostics plots

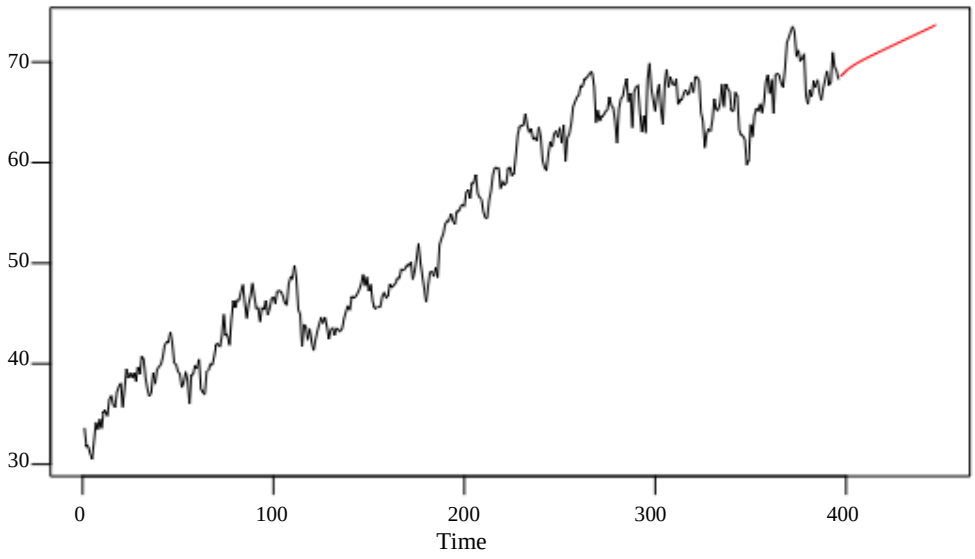

Fig. 7: Time series forecasting

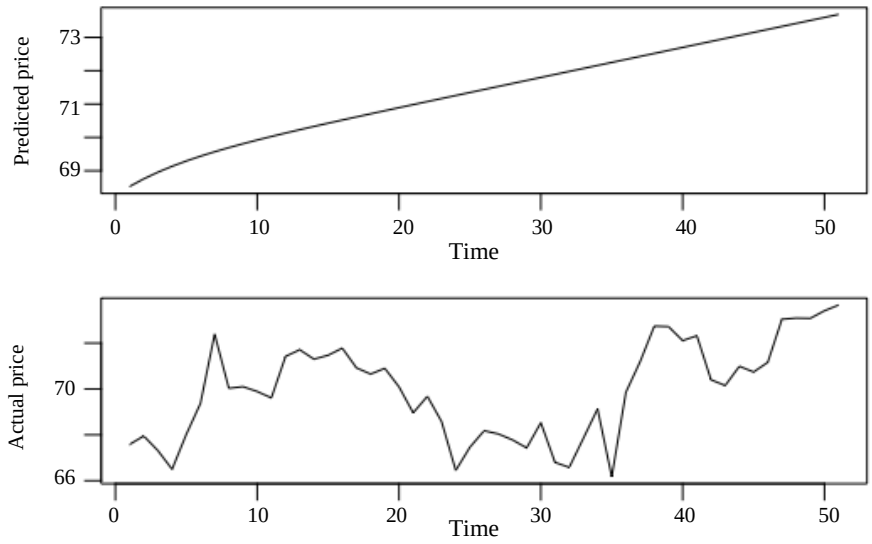

Fig. 8: Time series plots for predicted and actual prices

Varying Intercept Regression Model

Regression Revisited

Previously, this study has considered a pooled regression model and a time series model. Instead of considering each stock individually, the median is used to describe the closing price for the top stocks of the NYSE as one. 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 Fig. 8: Time Series Plots for Predicted and Actual Prices for individual stocks. If each stock is considered separately, the model then can be used by anyone who is invested into a stock within the top stocks in 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 o investors considering various stocks and which stocks tend to have higher or steadier prices.

This model 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 so that out of the 79,815 observations, 60,882 are in the training data and 18,993 are in the testing data. 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.

Table 7: 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	+ V IX	-96901.5
8	+ DFF	-96935.42
9	+ TNX	-96937.78

Table 8: 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	-
V IX	1.2708	1.1033
DFF	2.2827	1.8453
TNX	3.5399	2.1564

The varying intercept model is written as:

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

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. Note that essentially, this varying intercept model is similar to a standard regression model with a dummy variable associated with the different stocks because these two models all use a constant to adjust the difference between two distinct stocks. However, this varying intercept model is more straightforward and easier to interpret than a standard regression model with a dummy variable.

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. 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 is volume which represents the weekly volume sold for each index individually.

Table 7 shows each of the iterations of the stepwise process for the training data. The process 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. Some stocks can see increasing prices and other stocks can see decreasing prices over time depending on the financial health of its corresponding company.

Table 8 gives the *VIF* for each predictor in the model selected by the stepwise process and shows that 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 altogether while M1 and time are both related to the effects of inflation. After removing the highly correlated predictors, there no longer appears to be a problem of multicollinearity within the model.

Diagnostics and Model Assumptions

Since the variables for the final model have been chosen, it is possible to run diagnostics on the model.

If a variable is significant within the model, the slope or coefficient for that variable is significantly different from zero. The null hypothesis that the slope coefficient is equal to zero is tested against the alternative hypothesis that the slope coefficient is not equal to zero. The test statistics and corresponding *p*-values for each of the predictor variables are calculated, and they are not presented, but will be discussed in next section. 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 estimated coefficients are significantly greater than zero. Among all variables, the only exception is the variable corresponding to the stock index DVN. However, the variable is still included since it represents one category for the stock index variable.

Next, it is necessary to 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 in Fig. 9. 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, but not too much.

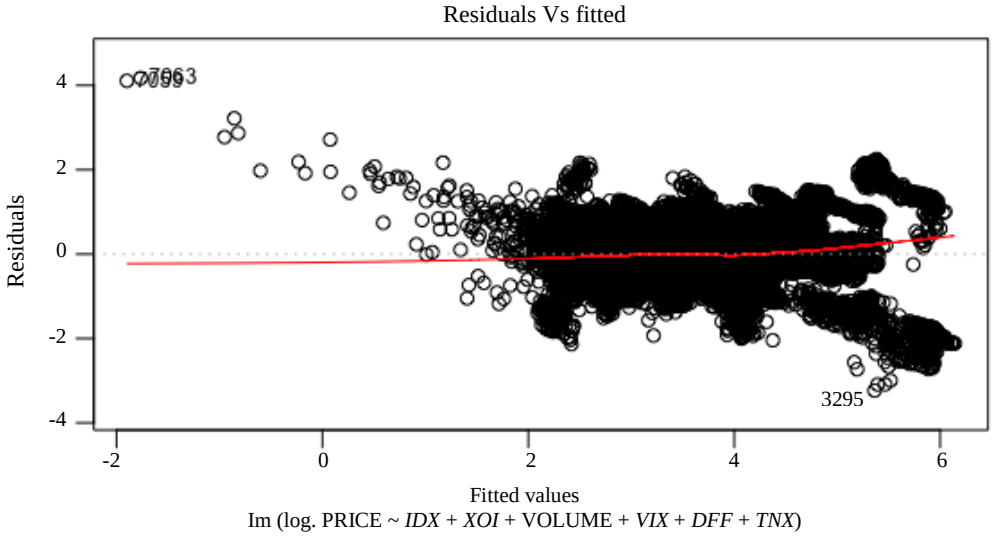

Fig. 9: Residuals Vs fitted values

The second assumption is the constant variance of the residuals. Constant variance can be checked visually by looking at the fitted residual plot in Fig. 9. For the variance to be constant, the residuals should be spread out equally around zero for all fitted values. 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. So, the independence assumption has been violated.

The final assumption is that the residuals are normally distributed. Normality is 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.

Model Fit and Interpretation

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

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

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 predictors 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. 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, there is a negative relationship.

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

Conclusion

Now that each of the three models have been thoroughly explored, it is important to compare the benefits and drawbacks of the models. Table 9 gives the equations for the pooled regression, time series, and varying intercept regression models.

Here, we comment that the approach of the three models for predicting stock prices can also be utilized in other markets of the same or similar size, even for smaller markets, provided that the data size of observed stock prices is large enough to conduct statistical inferences and the stock market across the economy as a group is not unusually affected by external effects or anomalies.

First, 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. It means that the model only requires knowledge of the values for the week of interest, not the time relative to other data points.

Next, 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.

Table 9: Model equations

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

Acknowledgment

The authors would like to express their appreciation to the editor and referees for providing thoughtful and insightful comments which helped to improve the original version of this manuscript.

Authors' Contributions

The first author managed the data, developed the methods, performed the computation, analyzed the results and wrote the manuscript. The second author conceived of the presented idea, encouraged the first author to investigate this analysis, supervised the development of the methods, supervised the findings of this work, discussed the results, revised and finalized the manuscript.

Ethics

The authors declare that there is no conflict of interests regarding the publication of this article.

References

Al-Tamimi, H.A.H., A.A. Alwan and A.A.A. Rahman, 2011. Factors affecting stock prices in the uae financial markets. J. Trans. Manage., 16: 3-19.

Chang, P.C., D.D. Wang and C.l. Zhou, 2012. A novel model by evolving partially connected neural network for stock price trend forecasting. Expert Syst. Applic., 39: 611-620.

Cryer, J.D. and K.S. Chan, 2008. Time Series Analysis with Applications in R. 2nd Edn. Springer Science & Business Media, ISBN-10: 038775959X, pp: 491.

Dielman, T.E., 2005. Applied Regression Analysis. 4th Edn. Brooks/Cole Thomson Learning, ISBN-10: 053446548X, pp: 608.

Moghaddama, A.H., M.H. Moghaddamb and M. Esfandyari, 2016. Stock market index prediction using artificial neural network. J. Econo., Finance Adm. Sci., 21: 89-93.

Online (a), 2018. Effective federal funds rate [dff]. https://fred.stlouisfed.org/series/DFF

Online (b), 2018. M1 money stock [m1]. Board of Governors of the Federal Reserve System (US). https://fred.stlouisfed.org/series/M1.

Online (c), 2018. Cboe volatility index: Vix [vixcls]. Chicago Board Options Exchange https://fred.stlouisfed.org/series/VIXCLS.

Online (d), 2018. Crude oil prices: West texas intermediate (wti). U.S. Energy Information Administration. https://fred.stlouisfed.org/series/WCOILWTICO.

Online (e), 2018. Wilshire 5000 total market full cap index. Wilshire Associates. Retrieved from

Sharif, T., H. Purohit and R. Pillai, 2015. Analysis of factors affecting share prices: The case of bahrain stock exchange. Int. J. Econ. Finance, 7: 207-216.