Creating a Predictive Model for Valero Energy Corporation Stock Value

By: Ian Scarff

Directed Studies

STAT 485-300

Summer II-2018

Creating a Predictive Model for Valero Energy Corporation Stock Value

for

Dr. Alan Dabney
Associate Professor
Texas A&M University
College Station, TX

by

Ian Scarff
STAT 485-300

August 14, 2018

Table of Contents

Abstract

The stock market has always been a place where people try to make money quickly. However, there is incredible risk when buying stocks. Modeling stocks can help reduce this risk. This allows us to make better and more confident choices, especially when choosing what stock to purchase. The purpose of this project was to make a predictive model for Valero Energy Corporation (VLO) stock. This project resulted in the creation of two model types for VLO stock, an Autoregressive Integrated Moving Average (ARIMA) model and a linear regression model.

Introduction

Motivation

The stock market was founded on March 8, 1817 on Wall Street in New York City, New York. Ever since then, people have been buying and selling stock in order to make a profit. People use past stock values in order to make quick, and sometimes unconfident, predictions about future values. These predictions tell people when it is a good time to sell or buy stocks. If a certain stock has been rising in value over the past few days, one would think that it will continue to rise. However, there are many influencers on the stock market that can make these quick predictions unreliable. Having an unreliable prediction brings great risk when purchasing stock. You can either make money or lose money quickly.

Purpose

The goal of this project was to create a predictive model for VLO stock. Since I was working with stock values over time, I needed to create a model that worked with time series data. While researching for ways to model time series data, along with linear regression, I found that an ARIMA model is one tool that is used to model and predict time series. For the purposes of this project, I decided to create both types of models in Rstudio.

Methods and Procedures

ARIMA Model

For an ARIMA model, only one data set was needed. R packages that were used to create this model are *pdfetch*, *data.table*, *MASS*, *timeSeries*, *forecast*, *zoom*, and *lubridate*. Daily, weekly, and monthly closing stock values for VLO stock were used to create three separate models. When creating each model, the same methods and procedures were used. Using the R package *pdfetch*, data starting from the first weekday of 2010 to the current date was taken from Yahoo Finance. Stock are based on returns, and returns are based on percentages. Transforming the data into a logarithmic data set captures these qualities in the time series.

To have confidence that an ARIMA model would fit the data well, I first created a test model using all but the last 100 data entries. Using the *auto.arima* function, an ARIMA model was created for the data sets. The *auto.arima* function automatically chooses the number of autoregressors, the number of differences, and the order of the moving average model. For each individual model, an ARIMA(0,1,0) model was chosen. This means it chose to take a first difference of the data. This allowed for stationary data.

When modeling time series, stationary data is needed. Having stationary data allows for constant mean, constant variance, and constant autocorrelation throughout the data set. Partial autocorrelation plots are an informal way to test for stationary data. The Dickey-Fuller Test is a formal way to test for stationarity. The null hypothesis for this test is that the data is not stationary, while the alternative hypothesis is that the data is stationary. It was found that a first difference of the logarithmic data set was stationary at a $p\text{-value} < 0.05$. This confirms that the data in the ARIMA(0,1,0) model is stationary. Along with the Dickey-Fuller Test, the Ljung-Box Test was used to test whether the residuals were random. The null hypothesis is that the residuals are random, and the alternative hypothesis is that the residuals are not random. Using arbitrary lags of 5, 10, and 15, the test showed that the residuals were random for the daily, weekly, and monthly models with $p\text{-values} > 0.05$.

After the test models were created, the *forecast* function was used to forecast the final 100 days of the full data set. By comparing the predicted last 100 values with the real last 100 values, it was found that there are a percentage error of approximately 9% on average (Figures showing the ARIMA model vs. raw logarithmic data can be found in the Appendix). This low percentage error gives confidence that the data was being modeled well. After testing the models, using the same methods, full models were created with the full data sets. Each model created was an ARIMA(0,1,0) model. With the full models, I forecasted the next 10 time entries. Figures 1-3 and tables 1-3 show the results of these forecasts.

Predicted Future Daily VLO Value

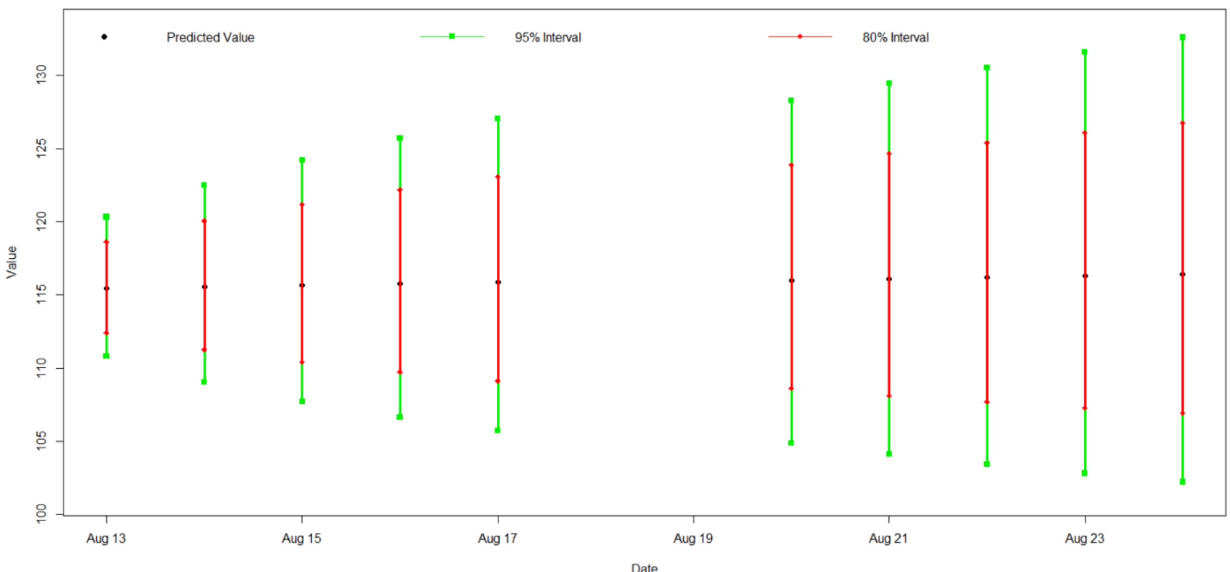


Figure 1.

Predicted Future Daily VLO Value

Starting from 2018-08-10 the predicted values for the next 10 days are:

	Date	Value	95% Upper bound	95% Lower bound	80% Upper bound	80% Lower bound
1	2018-08-13	115.4741	120.3213	110.8222	118.6209	112.4108
2	2018-08-14	115.5783	122.4986	109.0490	120.0576	111.2662
3	2018-08-15	115.6826	124.2221	107.7302	121.1972	110.4189
4	2018-08-16	115.7870	125.7117	106.6459	122.1837	109.7253
5	2018-08-17	115.8915	127.0524	105.7110	123.0726	109.1294
6	2018-08-20	115.9961	128.2880	104.8819	123.8925	108.6029
7	2018-08-21	116.1008	129.4442	104.1328	124.6604	108.1289
8	2018-08-22	116.2055	130.5379	103.4468	125.3872	107.6962
9	2018-08-23	116.3104	131.5807	102.8123	126.0806	107.2973
10	2018-08-24	116.4154	132.5812	102.2207	126.7462	106.9265

Table 1.

Predicted Future Weekly VLO Value

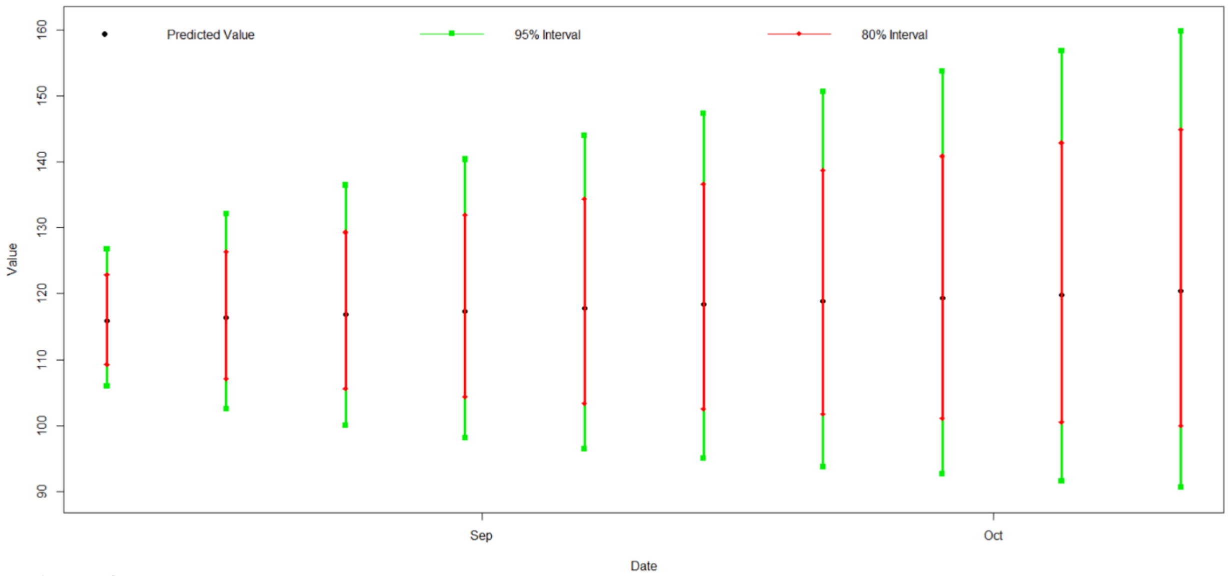


Figure 2.

Predicted Future Weekly VLO Value

Starting from 2018-08-10 the predicted values for the next 10 weeks are:

	Date.w	Value.w	95% Upper bound	95% Lower bound	80% Upper bound	80% Lower bound
1	2018-08-10	115.8583	126.7130	105.93337	122.8453	109.2686
2	2018-08-17	116.3486	132.0583	102.50768	126.3941	107.1015
3	2018-08-24	116.8410	136.4463	100.05262	129.3135	105.5714
4	2018-08-31	117.3354	140.3517	98.09364	131.9144	104.3677
5	2018-09-07	117.8320	143.9572	96.44801	134.3167	103.3705
6	2018-09-14	118.3307	147.3561	95.02251	136.5814	102.5187
7	2018-09-21	118.8315	150.6037	93.76206	138.7449	101.7761
8	2018-09-28	119.3344	153.7358	92.63096	140.8305	101.1194
9	2018-10-05	119.8394	156.7769	91.60456	142.8546	100.5322
10	2018-10-12	120.3466	159.7452	90.66497	144.8289	100.0028

Table 2.

Predicted Future Monthly VLO Value

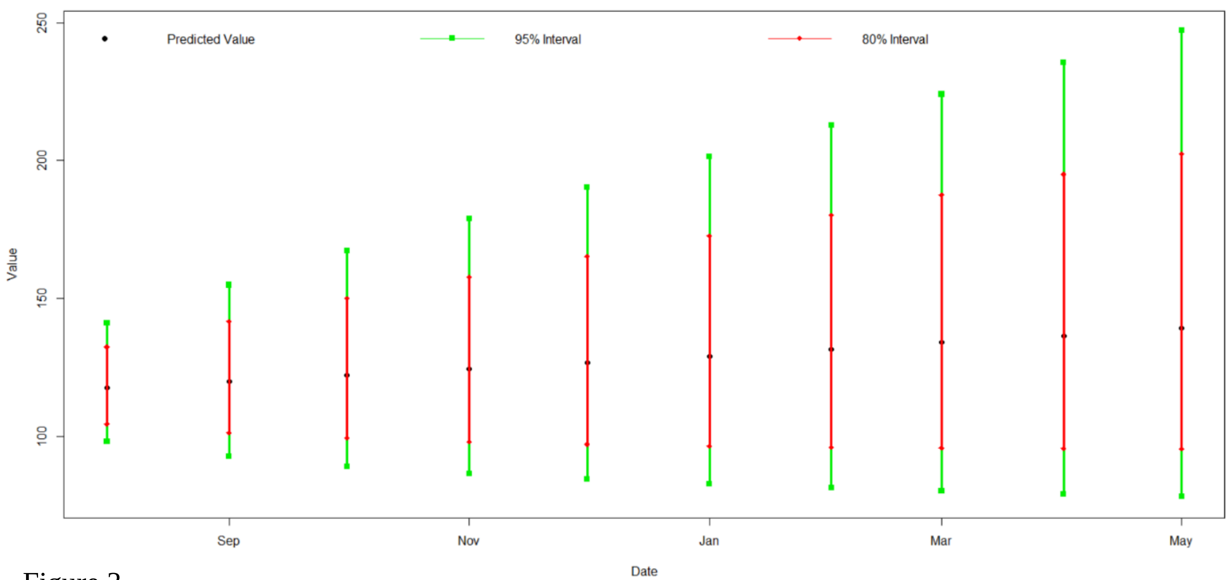


Figure 3.

Predicted Future Monthly VLO Value

Starting from 2018-08-01 the predicted values for the next 10 months are:

	Date.mo	Value.mo	95% Upper bound	95% Lower bound	80% Upper bound	80% Lower bound
1	2018-08-01	117.5459	140.9712	98.01329	132.3770	104.37654
2	2018-09-01	119.7629	154.8588	92.62087	141.6781	101.23769
3	2018-10-01	122.0217	167.1611	89.07156	149.9061	99.32420
4	2018-11-01	124.3231	178.8123	86.43836	157.6745	98.02625
5	2018-12-01	126.6679	190.1706	84.37041	165.2184	97.11245
6	2019-01-01	129.0570	201.4197	82.69155	172.6581	96.46640
7	2019-02-01	131.4911	212.6699	81.29922	180.0652	96.02025
8	2019-03-01	133.9711	223.9949	80.12792	187.4871	95.73057
9	2019-04-01	136.4978	235.4474	79.13300	194.9576	95.56776
10	2019-05-01	139.0723	247.0677	78.28256	202.5019	95.51067

Table 3.

Linear Regression

For the linear regression model, multiple data sets were needed. R packages that were used to create this model are *pdfetch*, *data.table*, *MASS*, *TTR*, *timeSeries*, *Quandl*, and *forecast*. The preliminary predictor variables for this model were the crude oil price from West Texas Intermediate (WTI), the S&P 500 (which is a measure of the overall stock market), the 50 day moving average of the S&P 500 (SP500_MA), the VIX (which is a measure of market volatility), and the 50 day moving average of VLO (VLO_MA). The response variable for this model is the VLO stock value. Using the R package *pdfetch*, data starting from the first weekday in 2010 to the last weekday of 2017 of VLO, the S&P 500, and VIX were taken from Yahoo Finance. Using the R package *Quandl*, data of the same time frame for WTI was taken from Quandl.

These data sets were merged into one time series data frame. Some data entries had NAs induced into them because the WTI had more data entries than the stock market values did (e.g. on certain weekends and holidays). To avoid this, there merger did not include dates with NAs in the data frame. This only removed at most 10 data entries, which is insignificant compared to the full data set of above 1,900 data entries. This also allowed for a continuous data set.

Using the *SMA* function, 50 day moving averages were calculated for VLO and the S&P 500. All data sets were then assigned as time series data sets using the *ts* function. I decided to remove the first 49 data entries for all data sets because the 50 day moving averages did nothing to model the first 49 days of VLO. Removing the first 49 data entries only removed approximately 2.4% of the data. I considered this removal of data insignificant compared to the rest of the data.

As in the ARIMA model, the Dickey-Fuller Test was used to test for stationary data. The test concluded all but the VIX and VLO_MA were not stationary. To make the VLO, S&P 500, WTI, and SP500_MA stationary, I took a first difference of each data set and then shifted them by the absolute value of their minimums plus one, respectively. This allowed for stationary data and non-negative, non-zero data. Performing the Dickey-Fuller Test on the new data showed that they were now stationary. However, these new data sets were one data entry shorter than the VIX and VLO_MA data sets. To make all the data sets the same length, I removed to first values from the VIX and VLO_MA. Now the data was ready to be modeled.

When modeling the data, I used the *tslm* function. This function creates linear models for time series data. It works similar to the *glm* function. The first preliminary model included all the predictor variables and interaction terms between the WTI and VLO_MA, the S&P 500 and SP500_MA, and VIX and the S&P 500:

$$\begin{aligned} VLO_{diff.shift} = {} & \beta_0 + \beta_1 WTI_{diff.shift} + \beta_2 VLO\_MA + \beta_3 S\&P\ 500_{diff.shift} \\ & + \beta_4 S\&P\ 500\_MA_{diff.shift} + \beta_5 VIX + \beta_6 \left(WTI_{diff.shift} \times VLO\_MA\right) \\ & + \beta_7 \left(S\&P\ 500_{diff.shift} \times S\&P\ 500\_MA_{diff.shift}\right) \\ & + \beta_8 \left(VIX \times S\&P\ 500_{diff.shift}\right) + \varepsilon \end{aligned}$$

Variables would be removed if their p-values were greater than 0.05. Interaction terms would be removed first if predictors alone had high p-values. To formally test for the removal of a predictor variable, the ANOVA Partial F-Test was used. The null hypothesis states that the full model and the reduced model do not significantly differ, and the alternative hypothesis states that the full model is significantly better. This process continued until a model was reached where all the predictors were significant:

$$VLO_{diff.shift} = \beta_0 + \beta_1 S\&P\ 500_{diff.shift} + \beta_2 S\&P\ 500\_MA_{diff.shift} + \beta_3 VIX + \beta_4 \left(S\&P\ 500_{diff.shift} \times S\&P\ 500\_MA_{diff.shift}\right) + \beta_5 \left(VIX \times S\&P\ 500_{diff.shift}\right) + \varepsilon$$

Figure 4 shows the plot of this first preliminary model.

First Preliminary Model Plots

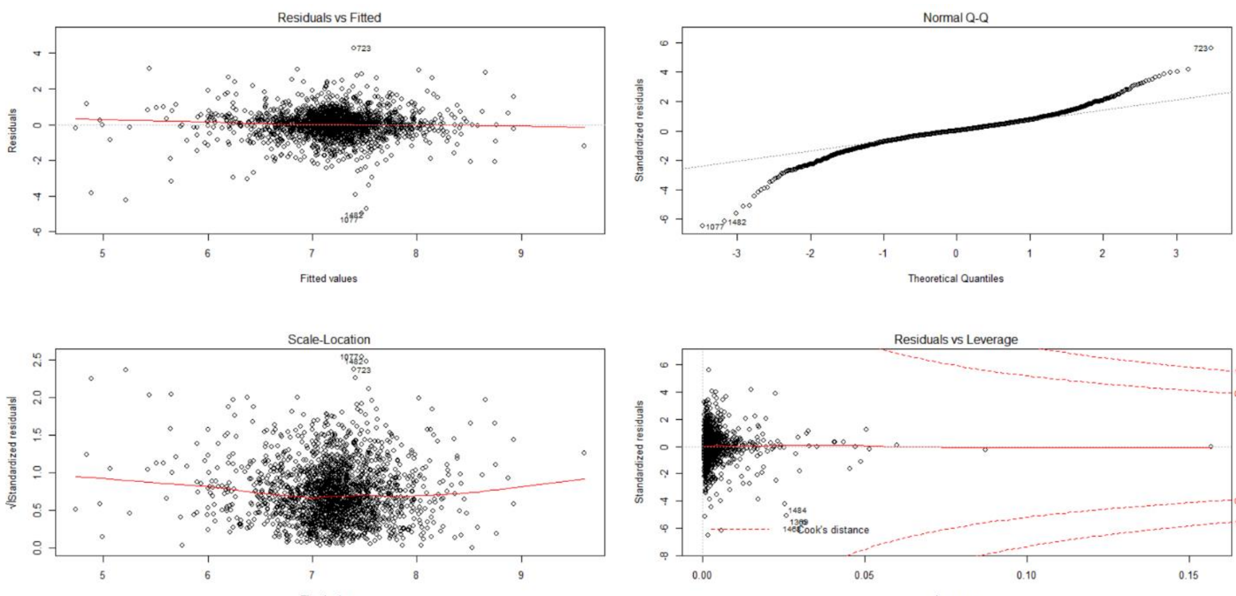


Figure 4.

The Normal Q-Q plot shows that the residuals appear not to follow the normal distribution. However, according to the Central Limit Theorem, this does not matter because the data set is large (over 1,900). The Scale-Location plot shows that the variance may not be constant. To fix this problem, using the *powerTransform* function, power transformations for the predictor variables and response variables were introduced.

After fitting a new model with transformations in both the response and predictor variables, a marginal model plot was used to see if the data was modeled well. Added variable plots were used to check for co-linearity. Unfortunately, the plots showed that the data was not modeled well and that there were many problems with co-linearity. A model with only transformations in response variable was then fitted to the data. This new model was chosen as the final model.

The final model had close to constant variance according to the model plots (Figure 5), does model the data well, and shows little to no problems with co-linearity (marginal model plot and added variable plots for final model can be found in the Appendix). The final model chosen was:

$$
\begin{aligned}
VLO_{diff.shift}^{1.5} = {} & \beta_1 S\&P\ 500_{diff.shift} + \beta_2 S\&P\ 500\_MA_{diff.shift} + \beta_3 VIX \\
& + \beta_4\left(S\&P\ 500_{diff.shift} \times S\&P\ 500\_MA_{diff.shift}\right) \\
& + \ \beta_5\left(VIX \times S\&P\ 500_{diff.shift}\right) + \ \varepsilon
\end{aligned}
$$

Note that the intercept has been removed. This says that the shifted difference between two VLO stock values is zero when all other predictors are zero. Table 4 shows the coefficients for each predictor.

Final Model Beta Values

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
SP500.diff.shift	0.2428650	0.0067216	36.132	< 2e-16	***
SP500_MA.diff.shift	0.8679854	0.1045909	8.299	< 2e-16	***
VIX	0.2006978	0.0264854	7.578	5.39e-14	***
SP500.diff.shift:SP500_MA.diff.shift	-0.0108705	0.0013991	-7.769	1.26e-14	***
SP500.diff.shift:VIX	-0.0026671	0.0003616	-7.376	2.39e-13	***

Table 4.

Final Model Plots

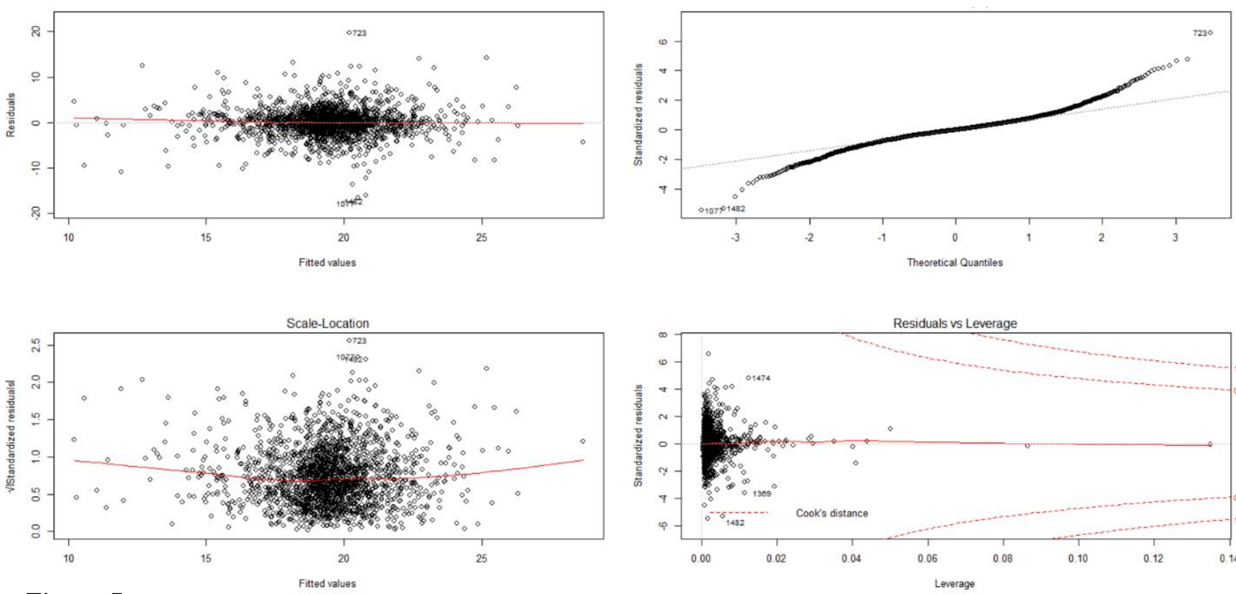


Figure 5.

To shows that this model approximately fits the data, figures 6-8 show the first, middle, and final 50 day prediction intervals for the data set respectively.

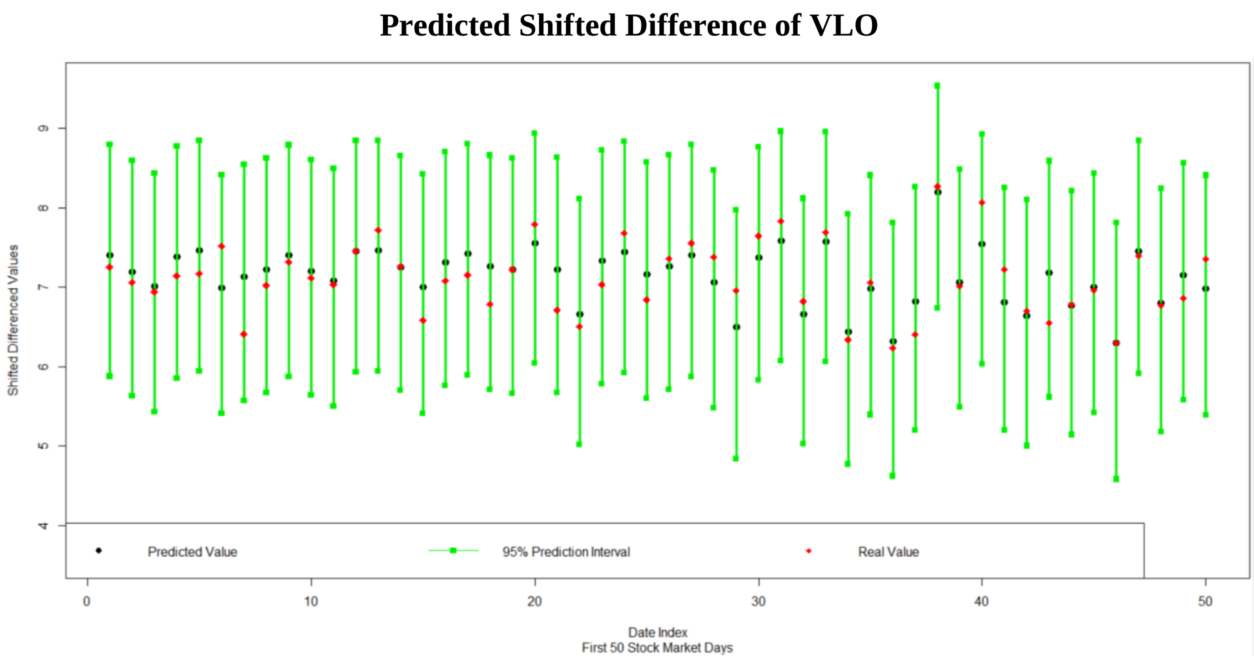


Figure 6.

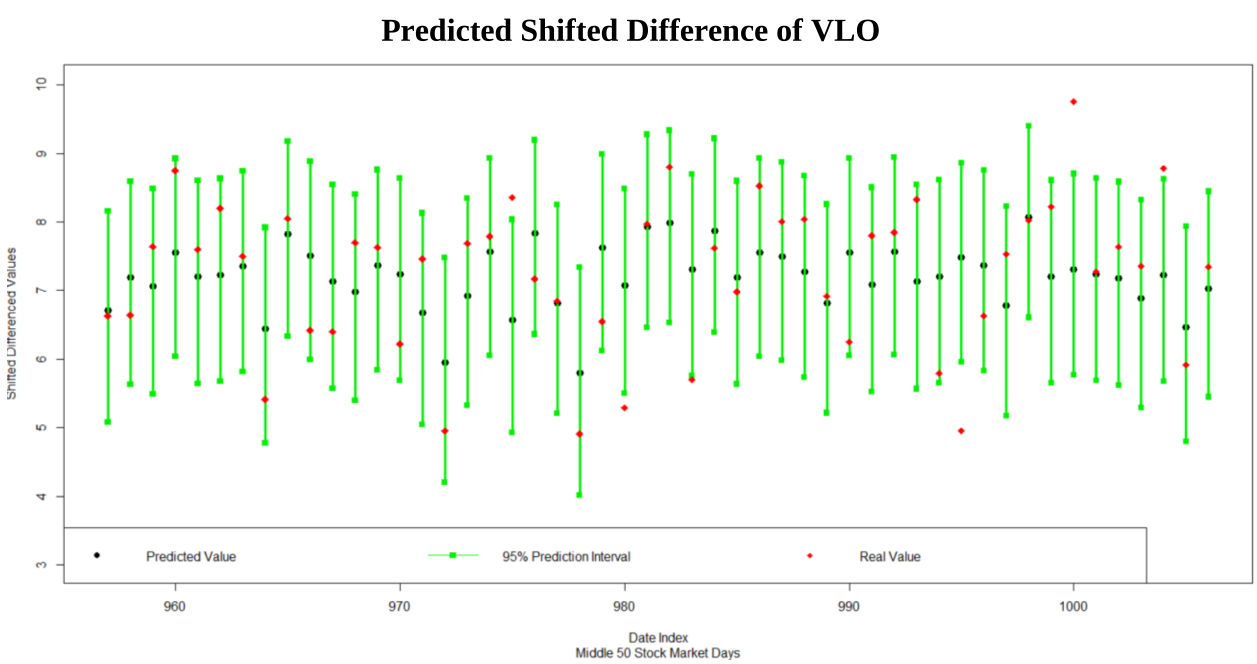


Figure 7.

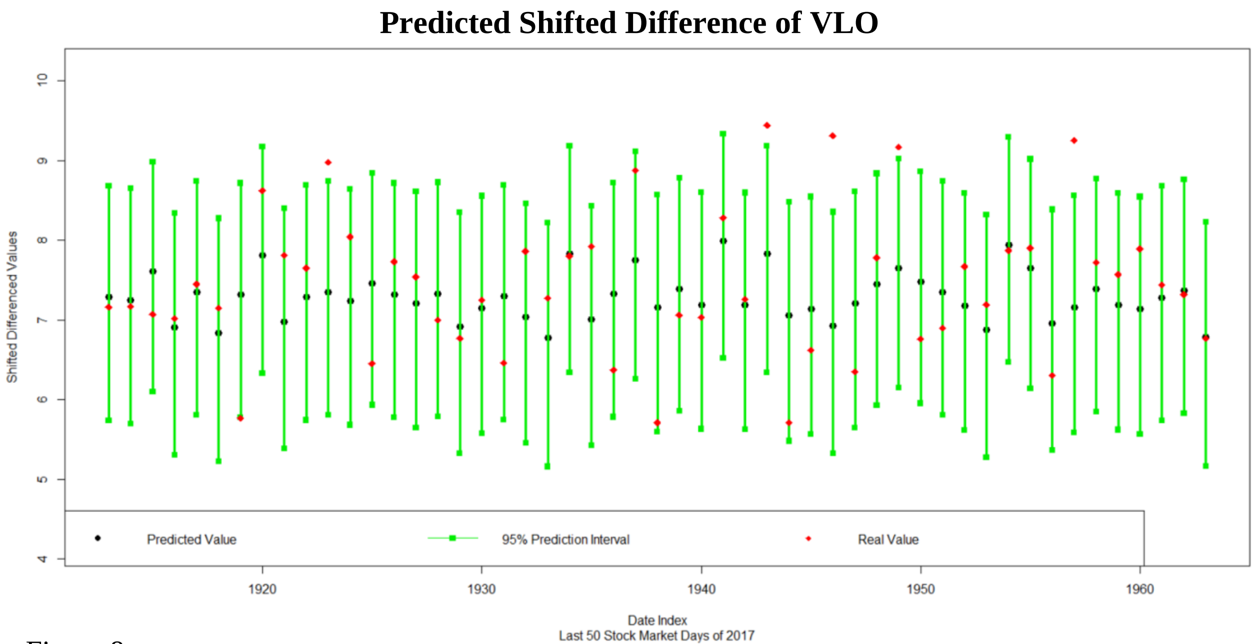


Figure 8.

Similar results were found when modeling with the *glm* function.

Discussion

There are some limitations to this project. The ARIMA model is a graduate level method that I as an undergraduate had never heard of before this project. My knowledge of this tool only extends to what I researched for this project. I may not fully understand what an ARIMA model really is. There are limitations to the data as well. Having to remove data entries may or may not cause unseen problems. For the linear regression model, further research into the *tslm* function is needed to fully understand its purpose. How does it differ from *glm*? Also, there may be other predictor variables that are important that were not included, such as news stories. Furthermore, there may be other methods to reduce variance and co-linearity further. Full knowledge of how the stock market operates is also necessary for this project. Future work may include predictive models for multiple stocks and the inclusion of more predictor variables.

Acknowledgements

I would like to thank Dr. Alan Dabney for all his support and helpful comments throughout the duration of this project.

Bibliography

Athanasopoulos, George, et al. "Package ‘forecast’." *Forecasting Functions for Time Series and Linear Models*. 21 June 2018. <http://pkg.robjhyndman.com/forecast, https://github.com/robjhyndman/forecast>.

Barbu, Corentin M. and Sebastian Gibb. "Package ‘zoom’." *A spatial data visualization tool*. 19 February 2015. <https://github.com/cbarbu/R-package-zoom>.

Dowle, Matt, et al. "Package ‘data.table’." *Extension of `data.frame`*. 27 May 2018. <http://r-datatable.com>.

Fox, John, et al. "Package ‘car’." *Companion to Applied Regression*. 2 April 2018. <https://r-forge.r-project.org/projects/car/, https://CRAN.R-project.org/package=car, http://socserv.socsci.mcmaster.ca/jfox/Books/Companion/index.html>.

McTaggart, Raymond, Gergely Daroczi and Clement Leung. "Package ‘Quandl’." *API Wrapper for Quandl.com*. 29 August 2016. <https://github.com/quandl/quandl-r>.

Quandl. *Cushing, OK WTI Spot Price FOB, Daily*. 2018. <https://www.quandl.com/data/EIA/PET_RWTC_D-Cushing-OK-WTI-Spot-Price-FOB-Daily>.

Reinhart, Abiel. "Package ‘pdfetch’." *Fetch Economic and Financial Time Series Data from Public*. 17 October 2017. <https://github.com/abielr/pdfetch>.

Ripley, Brian, et al. "Package ‘MASS’." *Support Functions and Datasets for Venables and Ripley's MASS*. 30 April 2018. <http://www.stats.ox.ac.uk/pub/MASS4/>.

Rougier, Jonathan. "Package ‘Oarray’." *Arrays with Arbitrary Offsets*. 20 March 2018.

Spinu, Vitalie, et al. "Package ‘lubridate’." *Make Dealing with Dates a Little Easier*. 11 April 2018. <http://lubridate.tidyverse.org, https://github.com/tidyverse/lubridate>.

Tillier, Martin. *What Is The Stock Market, And How Does It Work?* 22 December 2017. 13 August 2018. <https://www.nasdaq.com/article/what-is-the-stock-market-and-how-does-it-work-cm895748>.

Trapletti, Adrian, Kurt Hornik and Blake LeBaron. "Package ‘tseries’." *Time Series Analysis and Computational Finance*. 4 June 2018.

Yahoo! *CBOE Volatility Index (^VIX)*. 2018. <https://finance.yahoo.com/quote/%5EVIX/history?p=%5EVIX>.

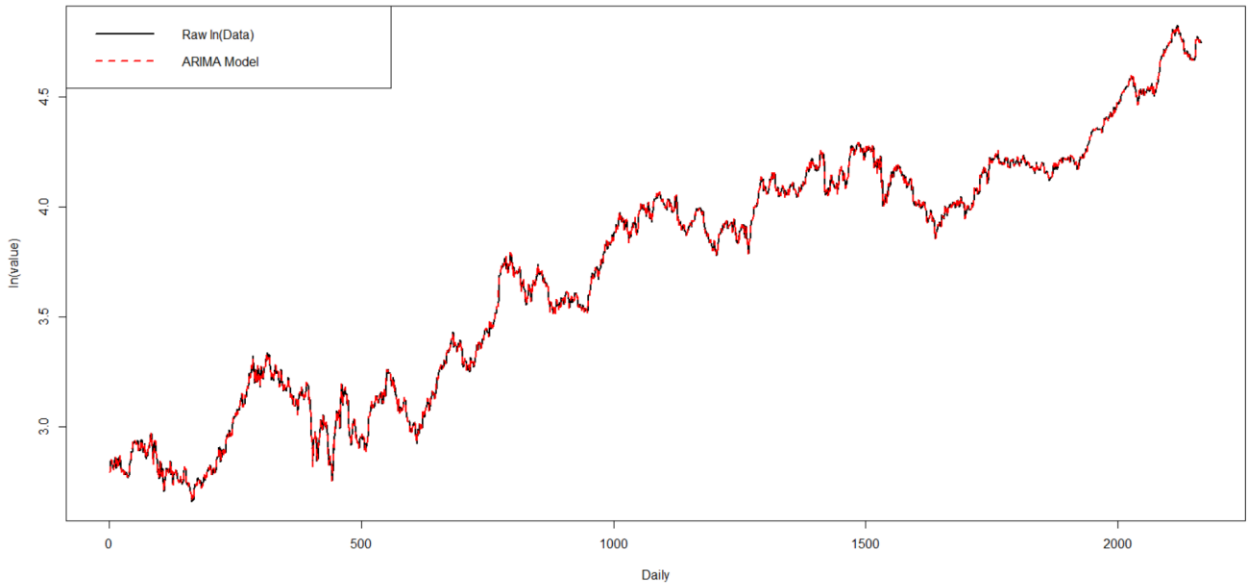
—. *S&P 500 (^GSPC)*. 2018. <https://finance.yahoo.com/quote/%5EGSPC/history?p=%5EGSPC>.

—. *Valero Energy Corporation (VLO)*. 2018. <https://finance.yahoo.com/quote/VLO/history?p=VLO>.

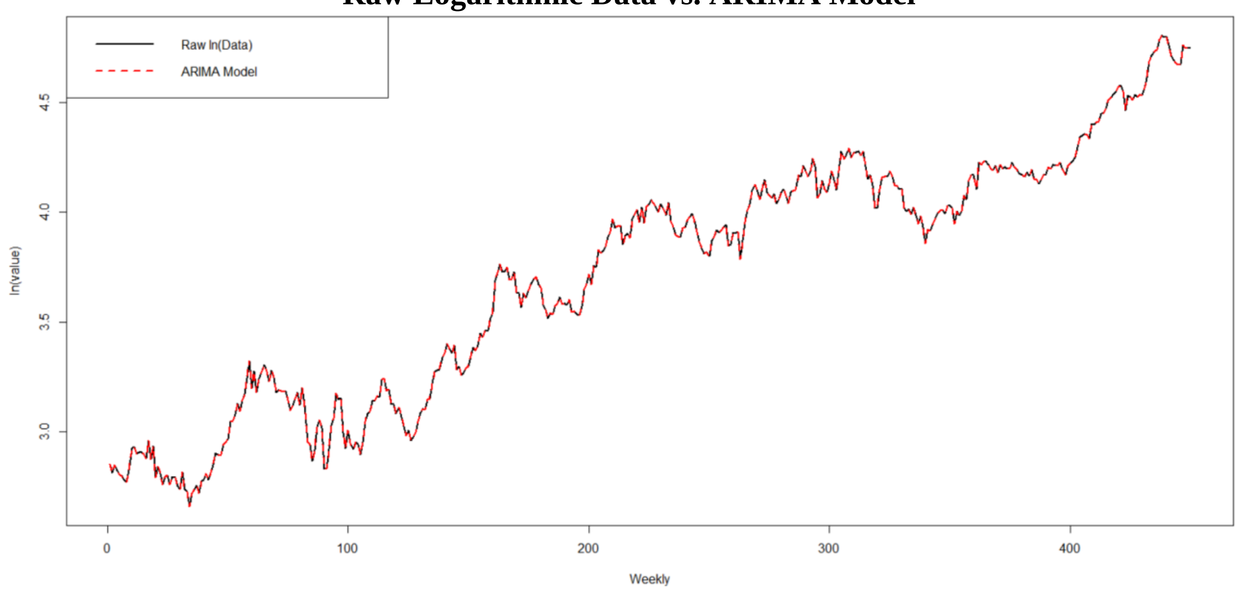
Appendix

ARIMA Model

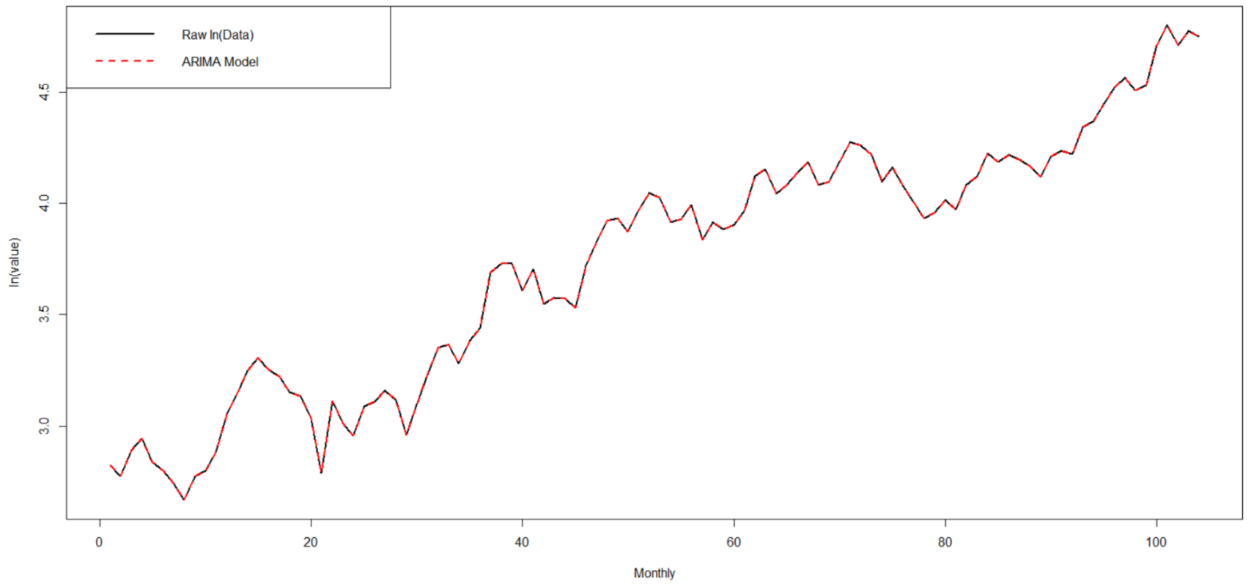
Raw Logarithmic Data vs. ARIMA Model



Raw Logarithmic Data vs. ARIMA Model

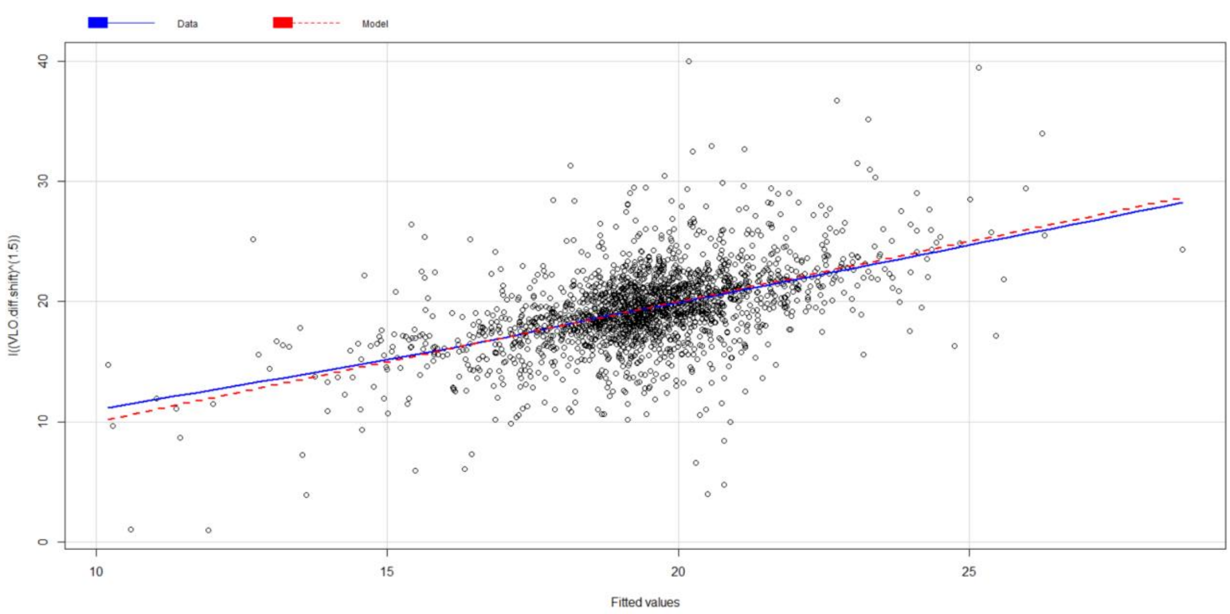


Raw Logarithmic Data vs. ARIMA Model



Linear Regression

Final Model Marginal Model Plot



Final Model Added Variable Plots