Price Prediction in the Stock Market using Economic Indicators

And a comparison of the pre- and post- COVID-19 market environment

Emet Bethany

Texas State University

Abstract

The stock market experienced a sharp selloff as investors panicked about the COVID-19 virus. However, it quickly recovered from the lows and this price action is mainly a result of the unprecedented monetary policy that was enacted by the Federal Reserve in March 2020. Using multiple regression, the future price of the S&P 500 – the most commonly followed stock index – can be predicted using several economic indicators as dependent variables. Comparing economic data between a time period before COVID-19 and after and training a regression model with these datasets, there seems to be a difference in the model's ability to predict the future price of the S&P 500. It is possible that the economy is becoming disconnected from the stock market as a result of the actions taken by the Federal Reserve. In the past, the health of the economy was very much related to the stock market.

1. Introduction

The motivation for this project is highlighted by a few reasons. The first is that price prediction of stock prices has the potential to be profitable and there are many algorithms that have been implemented and tested in an attempt to predict the future price of stocks. Some algorithms are successful, but those ones would not be readily available to the public and if they were consistently profitable, they would be worth a lot of money. Algorithms like this are typically owned by large funds or brokers and would not be a study or project.

Secondly, economic indicators provide data that is relevant and may be able to explain and possibly predict future prices. Some examples of economic indicators are GDP, inflation numbers, and employment data. The stock market often reacts to changes in these economic indicators positively or negatively depending on the implications of the data to the broader economy. For example, if the number of new weekly jobless claims rises more than what is expected, then it signals a possible downturn in the economy as people are being laid off and unable to work or produce goods and services. A bad report on these numbers often leads to a selloff in the stock market as a reaction to price in a possibly contracting economy.

Another motivation for this project is how unprecedented the reaction of the stock market was to the COVID-19 virus. This includes things like how quickly the Federal Reserve was to act on monetary policy changes like lowering interest rates and quantitative easing programs. These monetary policy tools fueled a rebound in the sharp selloff that happened in the stock market initially in response to the uncertainty of the COVID-19 virus. The importance of these unprecedented monetary policy changes is that, since they were implemented, the stock market has likewise seen an unprecedented recovery from the lows made in late March. The stock market and the economy have certainly been related in the past, but now there is clearly a disconnect. The S&P 500 closed at a new all-time high on August 24, despite the unemployment

rate still being at historic highs and many Americans still struggling financially. The cause for stocks to be at such high valuations despite the still bad economic environment is the previously mentioned monetary policy. The loose monetary policy makes it inviting to continue buying stocks. So, this leads to the question and problem: **Is the stock market becoming separated from the real economy now?**

2. Methodology

The machine learning concept used in this project is multiple linear regression. Multiple regression is similar to simple linear regression except that it uses at least 2 independent variables to help explain the dependent variable. In this case, I have 4 independent variables. The independent variables are economic indicators and are the GDP, the Unemployment rate, CPI, and Corporate bonds. The dependent variable in this case is the monthly closing price of the S&P500 Index. The equation at the center of multiple regression takes the form of a linear combination of the independent variables used to predict the dependent variable. This equation is generalized below.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \cdots$$

In the case of this project, the multiple linear regression formula will look like the following equation:

$$Monthly\ Closing\ Price\ of\ S\&P\ 500 =$$
$$\beta_0 + \beta_1(GDP) +$$
$$\beta_2(Unemployment\ Rate) +$$
$$\beta_3(CPI) + \beta_4(Corporate\ Bonds)$$

Multiple linear regression is useful for determining the influence of independent variables (the economic indicators) on the dependent variable (price of S&P). This methodology is also helpful in understanding how much the price of the S&P 500 will change when there is a change in the economic indicators. This regression can also be used to predict future values of the dependent variable with the equation.

There are specific reasons why I chose the independent and dependent variables that are used in this project. A brief explanation of these variables is now provided.

The dependent variable is the monthly closing price of the S&P 500 Index. This was chosen as the dependent variable because it is the most commonly followed basket of stock and includes 500 large companies listed on the stock exchanges in the United States. This index makes a good overall representation of the stock market in general as it spans companies from multiple sectors and provides a single price that represents them all. In this report, when stock market is mentioned it is typically referring to the S&P 500 Index. **[See figure 1]**

The first independent variable that was chosen was GDP, or growth domestic product. This economic indicator gives a number that indicates the value of all the goods and services made in a country. For this project, I am using the United States data. GDP is considered by many economists to be one of the best single indicators for the size and health of a country's economy, so including this as an independent variable in the linear regression model was almost a given. **[See figure 2]**

The next independent variable chosen is an index that follows the amount of corporate bonds issued. Corporate bonds are debt securities that are issued by a company so that it can raise capital. It is essentially lending money to a company in return for interest when bonds are purchased. I chose this indicator since the Federal Reserve committed to buying billions of dollars in corporate bonds in response to the economic downturn and stock market selloff experienced because of COVID-19. Typically, the Fed is not involved in the corporate bond market like this. **[See figure 3]**

The third independent variable is CPI, or the Consumer Price Index. This indicator puts a numeric value calculated from a weighted average of prices of consumer goods and services. I chose this as an independent variable

since it is used by many as a measure of inflation and the value of the US Dollar's purchasing power. Since inflation causes asset prices, like stocks, to rise in nominal price, I included it in the regression model. The Federal Reserve has made a commitment to targeting a higher inflation rate. The Federal Reserve's new quantitative easing program has also introduced over $3 trillion worth of new money into the money supply – which also causes inflation. QE is a monetary policy tool where the central bank purchases long term securities to introduce new money into the money supply. This is supposed to encourage investment and money lending. The Fed already committed to buying $200B in mortgage-backed securities and $500B in Treasury securities before March 23. On March 23, the Fed announced that it would continue purchasing securities in any amount that would be needed to support the market and broader economy. Since there was no defined limit on how much securities the Fed would buy now, it was coined 'QE infinity'. **[See figure 4]**

The last independent variable I chose is the Unemployment rate. This is important since it shows how many people are working and producing products or services for the economy. This rate skyrocketed when the pandemic hit and may be able to help explain or influence the stock market. **[See figure 5]**

Figure 1: S&P 500 Index

Figure 2: Growth Domestic Product

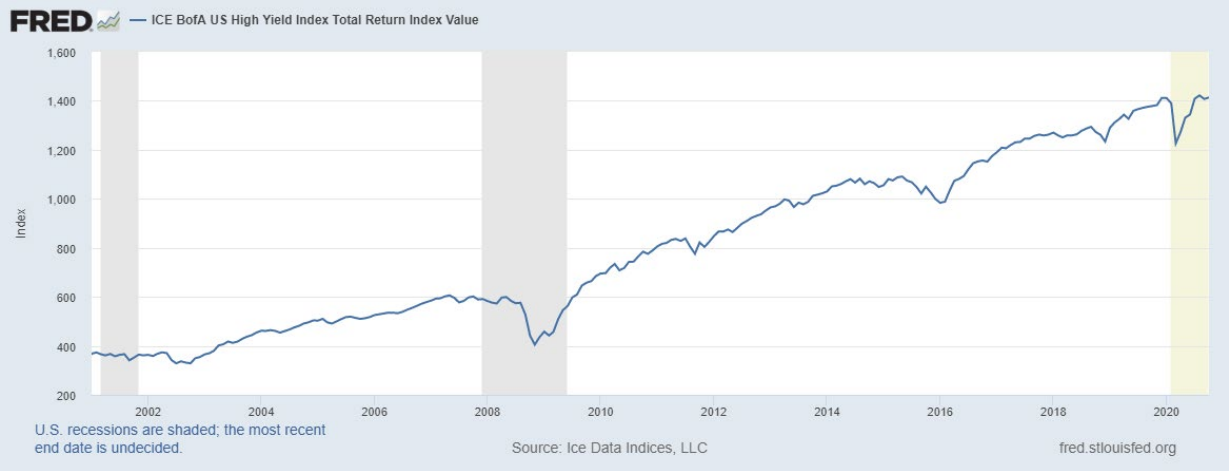


Figure 3: Corporate Bonds Index

Figure 4: Consumer Price Index

Figure 5: Unemployment Rate

3. Code and Model

This section will provide a brief overview of the thought process and implementation of the code.[1] The code has two parts: pre-COVID and post-COVID

```
#import necessary libraries/packages
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd

from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn import datasets, linear_model
```

At the beginning of the code, I import all the necessary libraries and packages needed to run the regression experiments. I read in the dataset which includes all of the economic indicators and the S&P prices and assign X and y to the independent and dependent variables, respectively. Using a function in the sci-kitLearn package, I separate the data into testing and training data. Then I create regression object to perform methods on. Then I create the predictions.

```
regressor = linear_model.LinearRegression()
regressor.fit(X_train, y_train)

y_pred = regressor.predict(X_test)
```

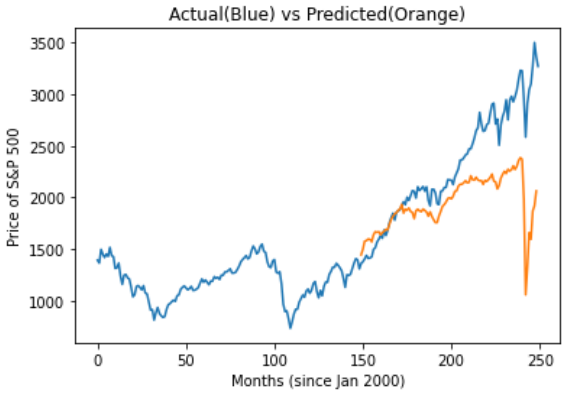
Code is used for formatting and setting values like the percentage of data that is being used for testing and training. At the end of the code, I print out relevant statistical numbers for the regression model like the R-squared score. The R-squared score is a goodness-of-fit measure for the regression model and describes the strength of the relationship between the model and the dependent variable (S&P 500 monthly closing price). I also put in code to output a graph of the actual vs predicted price of the S&P 500. More discussion on this in the results section.

The pre- and post- COVID-19 code sections of the entire program are similar in many ways except namely the datasets that are used. The datasets are over different time periods that correspond with each part.

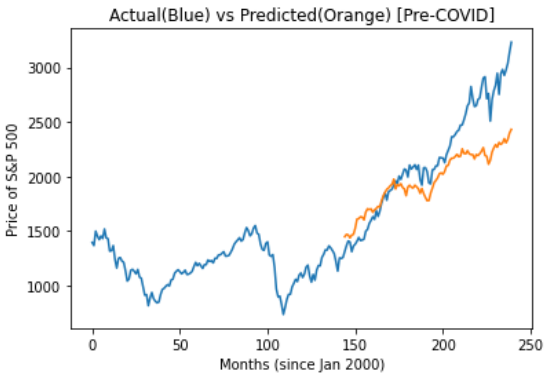
[1] All shown/unshown code and datasets used in this project: https://github.com/eyb6/CS4347_Final.

4. Results

Below are snippets of output from the post- and pre- COVID code segments.




```
R-Squared Score:                    0.7115036504502594
GDP Coefficient:                    -0.1872684561810091
Corporate Bonds Coefficient:        1.542058517640268
CPI Coefficient:                    18.01662744578599
UnemploymentRate Coefficient:       -129.92763923016392
Intercept-Value:                    51.66586864395936
```




```
R-Squared Score:                    0.8586599256690068
GDP Coefficient:                    -0.20378641381840226
Corporate Bonds Coefficient:        1.6360864695388506
CPI Coefficient:                    19.52168925433537
UnemploymentRate Coefficient:       -133.36828231469076
Intercept-Value:                    -56.84091133873562
```

The graphs both show an orange and a blue line. The orange line represents the predicted values based on the training of the linear regression model. This orange line starts only partway through the graph since the entire graph includes the entire dataset used. The part of the graph in the beginning where there is no orange line would be the training data, so there is no predicted line there. The blue line represents the actual price of the S&P 500 since January 2000. The R-squared score and the coefficients that make up the multiple linear regression equation are listed under each respective graph.

Just from looking at the graphs themselves, it is easy to tell that the predicted does not stray too far away from the actual in the pre-COVID experiment. In the post-COVID experiment, however, there is a bit more of a disconnect and there it appears especially off in the last 8-9 months of the graph. These last 8-9 months is the period where there was a major stock market selloff and when the Federal Reserve stepped in. By my hypothesis, it does seem like the model is becoming more disconnected because of the conflicting economic data with the actual price of the S&P 500.

Looking at the R-squared scores for each experiment, it further solidifies the question that arises at the beginning of the report. The R-squared score is a goodness-of-fit measure for the regression model. Post-COVID, this is saying that the strength of relationship between my model and the S&P 500 price is about 71%. The model seems to predict better pre-COVID, with an R-squared percent of nearly 86%. This is saying that the strength of the relationship between the model and the dependent variable, the S&P 500 price is about 86%.

5. Conclusion

In conclusion, an interpretation of this difference in the output of the code pre- and post-COVID may show that the stock market has become more disconnected from the economic reality. The model had much more trouble predicting price when you include economic data after the reaction to COVID-19 than it did before.

In the future, it would likely be useful and interesting to test out different independent variables and see how it affects the outcome. Whether that means adding in more independent variables or switching out the ones that I used. Other possibilities would be to switch unemployment rate with labor participation rate, or switching the Consumer Price Index with Manufacturing and Production Indices. It may also be interesting to add other possible economic indicators entirely like personal spending and savings among consumers.

The contributions to this project are entirely my own and I accept full credit for the accomplishments, results, possible errors, or flaws in the project. I have learned a lot from this project and find linear regression to be an incredibly interesting topic for use in determining relationships of data. The application of statistical methods, which is my focus in my Mathematics degree, also makes this project and methodology interesting and I plan to continue my study in machine learning as it relates to the field of statistics.