Andrew Plum

Professor Hickman

Econ 453

12/15/2023

**FORECASTING THE FINANCIAL FUTURE**

**INTRODUCTION**

I am trying to build a regression model to best approximate the stock market. The specific stock market index I am trying to predict is the S&P 500 which comprises of the largest publicly traded companies in the United States. I chose this one because it is a stock market index that is both more well known among most people but also isn’t weighted by the stock price like the Dow Jones Industrial Average (arguably the most followed stock index among the masses). Although it is not a perfect reflection, it can be useful to follow the stock market because it does give some indication of how the US and even the global economy is doing. Even if you don’t invest in the S&P 500 or any other stock indices, it can be useful to follow and know where the market is heading as its trajectory can have policy implications which affect everyone including you. This is how I crafted my model to attempt to capture the market trends.

**SUMMARY**

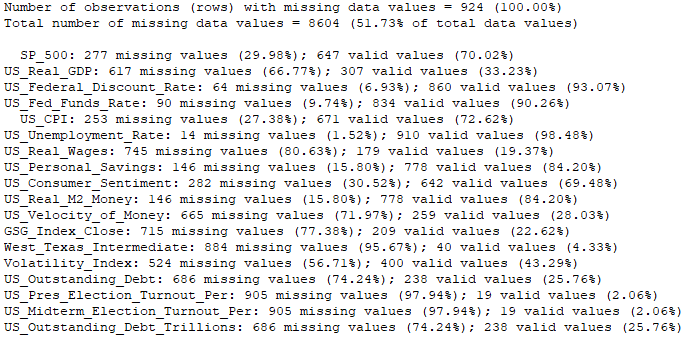
* Moving forward, I would add more complete data to the dataset so that less rows are dropped when each regression is run.
  + I would find more data for the variables I already have and the variables I don’t have.
* I also want to continue and analyze the data I already have as quarterly data.
* The S&P 500 is difficult to predict.
  + It can be difficult to address serial correlation and heteroscedasticity while pursuing high adjusted R^2 with the model and variable significance.
  + I believe more data will solve this problem.

A screenshot of a computer

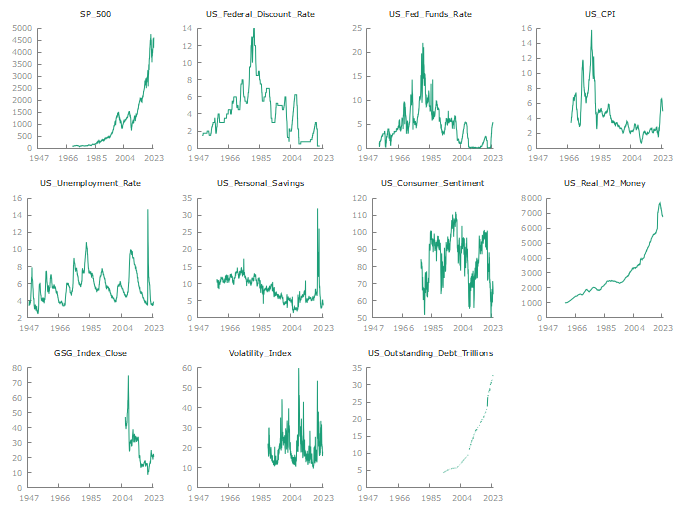
Description automatically generated**DATA**

The data obtained is very credible as the sources of the data were the St. Louis Fed (FRED), the US Treasury, the US elections project, and Yahoo Finance. I do not know where I could collect better data for free. Because the data is the best available, the conclusions that come from it are well supported if the statistical analysis is performed correctly.

Adding to the dataset in gretl, I created US\_Outstanding\_Debt\_Trillions out of US\_Outstanding\_Debt. From now on I will use the debt adjusted for trillions variable instead of US\_Outstanding\_Debt. I adjusted this variable while leaving other variables with large numbers unadjusted because it was represented in scientific notation, and it would be easier to interpret by adjusting it. I also created an autoregressive term of SP\_500 for when I use it later. Any future variables that are created will be mentioned later when they are used.



I started out by formatting the data as monthly because before I had all my data sources chosen, most of the data was monthly. I also wanted data that was recorded more frequently because I wanted the model to be closer to real time since most people who invest in the stock market do not invest on a less frequent basis but rather a more frequent basis especially because most data related to the stock market is readily available. I finally thought that more values would be preserved by treating it as monthly; this may be true initially, but it also meant I could create fewer first difference variables with valid values. Because I initially formatted the data as monthly data and not as quarterly data, some of their variables are missing quite a few observations. I decided to move forward with the monthly formatted dataset I had rather than reformatting it, as I had already spent far longer compiling and formatting the dataset than I wanted to, and I was eager to start running regressions.



Some takeaways from the graphs are the SP\_500, US\_Real\_M2\_Money, US\_Outsanding\_Debt\_Trillions can all likely be represented well by as quadratic equation. Most all the data is very cyclical. Note that not all the graphs are shown as some of the variables couldn’t be visualized when graphed against time because the values were too few and sparse. One of these graphs for example, is the US\_Real\_GDP variable. The other variables that are not shown as graphed against time are not graphed because even though they could visually be seen, they either lacked data or still had a lot of gaps in time making them more difficult to interpret. The West\_Texas\_Intermediate graph against time was one of these graphs that could visually be seen but lacked enough data for me to justify showing.

A screenshot of a graph

Description automatically generated

In these first difference graphs, there is a lot of variation in these graphs. Most of the graphs also have one or more time periods where the magnitude spikes dramatically.

Summary statistics of the variables.

A table of numbers with numbers

Description automatically generatedIt seems that there are a few issues with the provided statistical summary. Here are some noteworthy problems:

* Some of the variables are not unadjusted for large numbers which means you will need to be more careful when interpreting them.
* Comparing the mean and median as well as the minimum and the maximum can provide insight into the distribution of the data. Because the mean is quite different from the median, it could indicate the presence of outliers.
  + The range of considerations may be important when interpreting and analyzing the data for these variables; later on we will look at the context of the values for d\_US\_Federal\_Discount\_Rate and d\_US\_Real\_M2\_Money when using them in a regression equation interpretation.

**EMPIRICAL STRATEGY**

This section explains what I initially planned to do before running any regressions. I expect that most of the variables I have chosen to examine are at least somewhat related to the S&P 500. However, because I am working with time series data, I need to be wary of serial correlation. I will need to first experiment around with time trend models, first difference models, and autoregressive models so I am confident serial correlation is not a present issue in the model. I predict that there will be several variables that explain the variation in the S&P 500; these variables will cover distinctly different factors that in combination best predict the stock index. The strategy I plan to start with is to categorize each variable into groups where they share similarities between the other variables in the group, and I plan to run regressions with these groups. The groupings of variables I am planning on using are: macroeconomic indicators like real GDP and unemployment, monetary policy like the federal discount rate, consumer behavior which includes consumer sentiment, commodity and energy models with an index like the West Texas Intermediate, and election related data like the turnout in presidential and midterm elections. From there, I will make a new regression using the most significant variables of each of the previous regression groupings. After this, I would expect a good starting model to further experiment with will have been achieved.

Because I didn’t realize until after looking at the total missing observations of each variable, I don’t know if I will be able to start out by running all these regressions in these variable groupings. If that is the case I will attempt to run the best combination of as many variables as possible at the start in one regression and clean up the model by narrowing down the number of variables used in the model. This means I expect to run a regression(s) like this:

Y = *β*0 ​+ *β*1 \* X1 + … + *β*i \* Xi

Where Y = S&P 500 and X1 to Xi represents any distinct group of 1 or more variables and their respective first difference variants.

The general goal is to try and reach a model with the best adjusted R^2 and statistical significance of its regressors while simultaneously not displaying any problematic results like issues with multicollinearity, serial correlation, or heteroscedasticity.

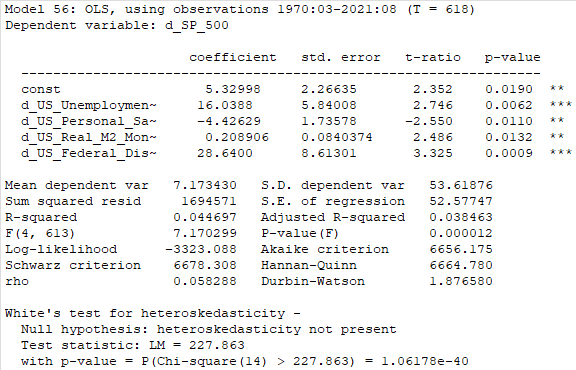
**RESULTS**

* I first tried to run a regression of all variables.
  + I couldn’t because there were no entries with no missing values which meant I had to drop some variables to run regressions.
    - I then looked at the table with the number of observations missing from each variable (this was done by clicking on “Data” and then “Count missing values” in gretl). I removed these variables because they were severely limiting how many records could be used in each regression:
      * Real\_GDP
      * US\_Real\_Wages
      * US\_Velocity\_of\_Money
      * GSG\_Index\_Close
      * West\_Texas\_Intermediate
      * Volatility\_Index
      * US\_Presidential\_Election\_Turnout\_Per
      * US\_Midterm\_Election\_Turnout\_Per
      * US\_Debt\_Outstanding\_Trillions
* After running a regression with the variables, I had left, I noticed that the adjusted R^2 seemed to be quite inflated. Because of this, I wanted to check for serial correlation but realized the results did not display the value of the Durbin-Watson statistic.
  + Through trial and error, I removed some of the variables with relatively high numbers of missing observations and finally got the Durbin-Watson statistic to display after removing US\_Consumer\_Sentiment.
* The Durbin-Watson statistic was quite low at 0.127483 causing me to replace all the variables with their first difference variant of each of them.
  + I the added the first difference of US\_Consumer\_Sentiment back into the model which gave me these results:
  + Before making the first difference model, I initially tried an autoregressive model and a model with a linear time trend.

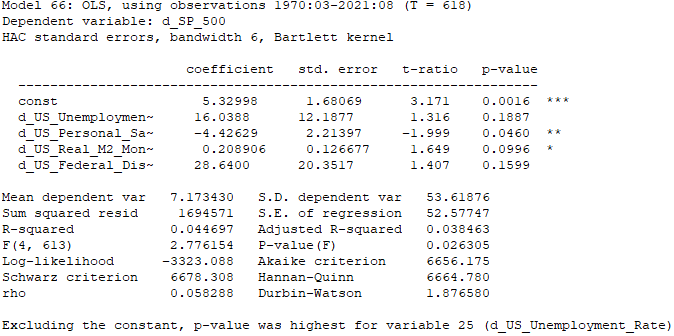
A screenshot of a computer

Description automatically generated

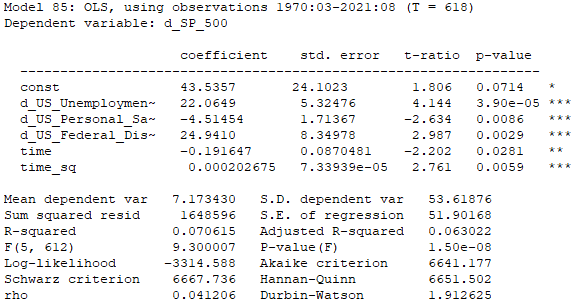
* The Durbin-Watson statistic is drastically better than it was before.
* The adjusted R^2 is much lower now, and I decided the model could be cleaned up because not all the variables were significant.
  + I tested for linear equivalence and d\_US\_CPI, d\_US\_Consumer\_Sentiment, d\_US\_Fed\_Funds\_Rate could be taken out of the model as they test results indicate that they are equivalent to d\_US\_Federal \_Discount\_Rate which gave me the results to this model:



* Regression equation of previous model:
  + Change in S&P 500 = *β*0 ​+ *β*1 \* Change in US Unemployment + *β*2​ \* Change in US Personal Savings Rate + *β*3 \* Change in US Real M2 Money + Change in *β*4 ​\* Federal Discount Rate + *ε*
* All the variables in the model are significant and the R^2 is slightly higher than it previously was.
* I then noticed my model has a problem with heteroscedasticity.
  + The residuals against time have an oscillating pattern where the magnitude is increasing.
    - This is concerning as it likely indicates heteroscedasticity is present.
      * I decided to look at White’s test and heteroscedasticity seemed to be present in the model as it pvalue for the test was under 0.05.
* I need to modify my model to deal with the heteroscedasticity.



* In this model, I used robust standard errors to deal with the heteroscedasticity, however, it was still present after the change. The variables in my model were also less significant than they were before.
* I tried to get a bit more creative and added a time trend. That did not solve the heteroscedasticity issue, so I added a squared time trend giving me this model which also failed to remedy the situation:



* Regression equation of most recent model:
  + Change in S&P 500 = *β*0 ​+ *β*1 \* Change in US Unemployment + *β*2​ \* Change in US Personal Savings Rate + *β*3 ​\* Change in Federal Discount Rate + *β*4 ​\* Time Trend + *β*5 ​\* (Time Trend)^2 + *ε*
* Both time trends are significant and increase the adjusted R^2 of the model.
* Afterwards, because heteroscedasticity was still present, I tried using the same model except this time with robust standard errors although the problem persisted.
  + I decided next to log the y variable and experiment by adding back in one of the previous variables.
* Eventually I ended with this model:

A screenshot of a computer

Description automatically generated

* The logarithm and the robust standard errors removed the heteroscedasticity from the model.
* I also used robust standard errors which increased the significance of the d\_US\_Federal\_Discount\_Rate and d\_US\_Real\_M2\_Money.
  + The time trend is quite significant. d\_US\_Real\_M2\_Money is also significant while d\_US\_Federal\_Discount\_Rate.
  + The adjusted R^2 of the model is quite high.
    - Most of the reason is because of the time trend. When the time trend is taken out, the adjusted R^2 is 0.050423.
* Interpreting the model is not the most straightforward; it is saying:
  + For each month that goes by, the change in the S&P 500 increases by 0.585318% all else equal.
  + For each point increase in the US federal discount rate, the change in the S&P 500 decreases by 43.2054 % all else equal.
  + For each increase of 1000 billion dollars in US real M2 money, the change in the S&P 500 increases by 0.336328% all else equal.
  + These interpretations make sense when taking in context the summary statistics of each of the variables.
    - The range of values for US federal discount rate is several orders of magnitudes smaller than US real M2 money.
  + The formula which was modeled last with the best adjusted R^2 that I achieved at is:
    - Logged Change in S&P 500 = *β*0 ​+ *β*1 ​\* Time Trend + *β*2​ \* Change in US Federal Discount Rate + *β*3 ​\* Change in US Real M2 Money Stock + *ε*

**CONCLUSIONS**

Going forward, if I had more time to continue working with this dataset, I would see if there is any more data to utilize both of the variables that already exist in my dataset and new variables that could be added to the dataset. I also would like to continue and analyze the data I already have as quarterly data to see if my analysis could be improved by that at all. I believe it would improve my analysis because then I could utilize more of the first difference variables that I created because more of them would have values, and I could utilize more data from variables when running a regression because less rows would be dropped for lack of observations. In terms of gathering more data, I wanted to measure the impact US elections had on the S&P 500, but I was unable to do this because of how little elections data I had. To explore this in the future, I will likely have to get my hands on polling data, although this can be difficult to come by. As demonstrated by the low adjusted R^2 in my models, clearly the S&P 500 is very difficult to predict. Serial correlation was present in the initial data, but most effectively addressed by solutions like making a first differences model and trying to fit the dependent SP\_500 variable to a different curve fit. There is much to still be improved upon, however. By adding more complete data to my models, I think the problems of heteroscedasticity present in my regressions would be less of an issue as they would be addressed by minimizing the amount of omitted variable bias in the models.