Andrew Plum

Professor Hickman

Econ 453

12/15/2023

**Forecasting the Financial Future**

**INTRODUCTION**

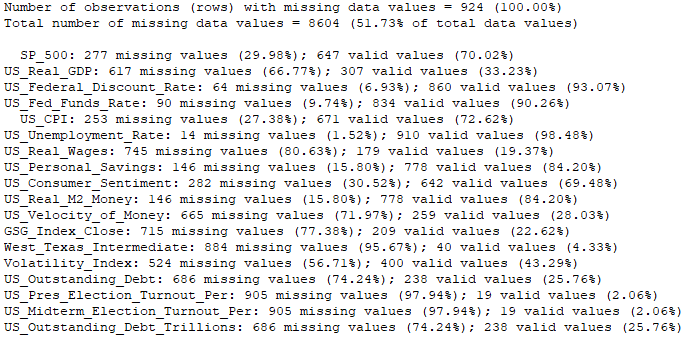
* Problem with serial correlation

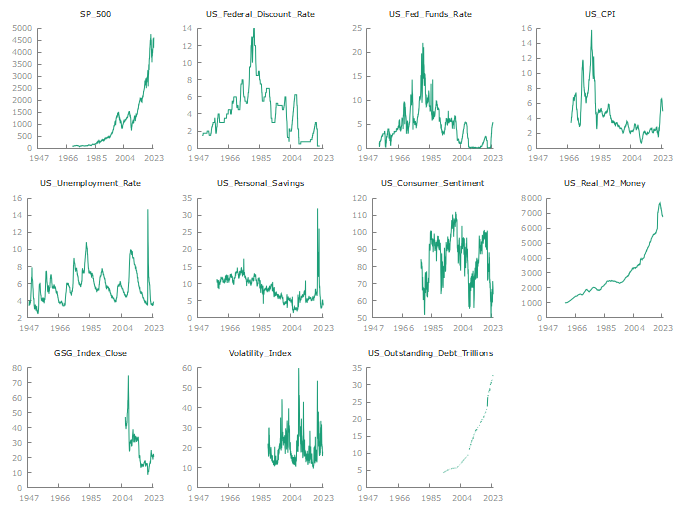
**DATA**

A screenshot of a computer screen

Description automatically generated

* I created US\_Outstanding\_Debt\_Trillions out of US\_Outstanding\_Debt\_Trillions.
  + From now on I will use the debt adjusted for trillions instead of US\_Outstanding\_Debt.



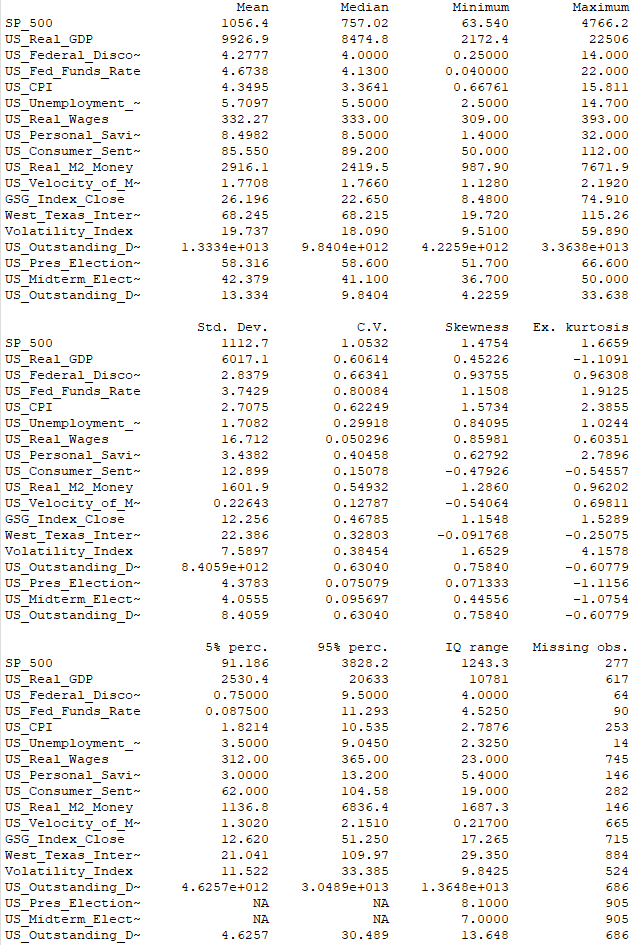


* Takeaways
  + ANSWER STILL!!!!!!!!!!
* Some of the variables couldn’t be visualized when graphed against time because the values were too few and sparse.
  + An example of this is the US\_Real\_GDP variable
* The other variables that are not shown as graphed against time because their graphs were to
  + The West\_Texas\_Intermediate graph against time was one of these graphs

A screenshot of a graph

Description automatically generated

* Takeaways
  + Theres is a lot of variation in these graphs.
    - Most of the graphs have one or more time periods where the magnitude spikes dramatically.



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

1. Formatting inconsistencies: The columns containing the variable names and the statistical measures have inconsistent formatting, making it difficult to read and interpret the data.

2. Missing values: The "Missing obs." row indicates that there are missing observations for some variables, which can affect the accuracy of the statistical summary.

3. Abbreviations: Some variable names have been abbreviated, which can make it difficult to understand the meaning of each variable.

4. Typos and special characters: There are some typos and special characters in the data, such as "ΝΑ" instead of "NA", which may need to be corrected for clarity.

5. Outliers: The summary statistics do not specifically mention any outliers, which could affect the interpretation of the data.

To improve the usability and accuracy of the statistical summary, it would be beneficial to ensure consistent formatting, correct any typos, provide a key for abbreviations, clarify the treatment of missing values, and address any potential outliers in the data.

There are several noteworthy aspects to consider about the summary statistics for these variables:

1. Missing Observations: Some variables have a high number of missing observations, specifically US\_CPI with 745 missing observations, which may impact the reliability of the statistics for this variable.

2. Skewness and Excess Kurtosis: Skewness and excess kurtosis measures indicate the departure from normal distribution of the variables. Several variables have high skewness and kurtosis, which suggests that these variables may not follow a normal distribution.

3. Coefficient of Variation (C.V.): Looking at the coefficient of variation may be helpful when comparing the variability of different variables as it is a measure of dispersion relative to the mean.

4. Outliers: The large difference between the 5th and 95th percentiles, as well as the large range of the variables, suggests the possibility of outliers.

5. Mean and Median: Comparing the mean and median can provide insight into the distribution of the data. If the mean is substantially different from the median, it could indicate the presence of outliers.

These considerations may be important when interpreting and analyzing the data for these variables.

It appears that US\_Real\_GDP, US\_Real\_Wages, US\_Real\_M2\_Money, and GSG Index Close have relatively high skewness and excess kurtosis. These variables may have non-normal distributions.

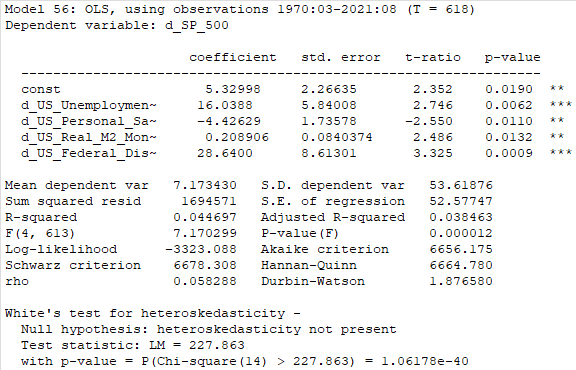
**Results**

* 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 weren’t displaying 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 of the variables with their first difference variant of each of them.
  + I actually the added the first difference of US\_Consumer\_Sentiment back into the model which gave me these results:

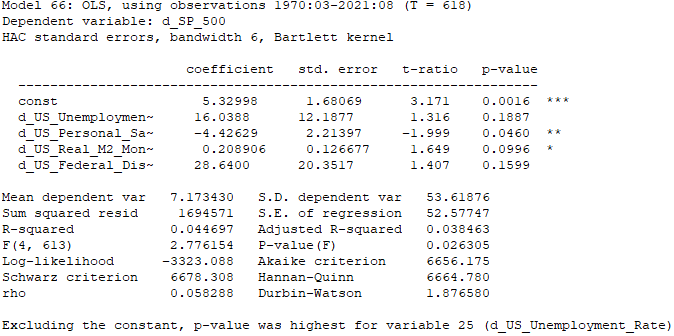
A screenshot of a computer

Description automatically generated

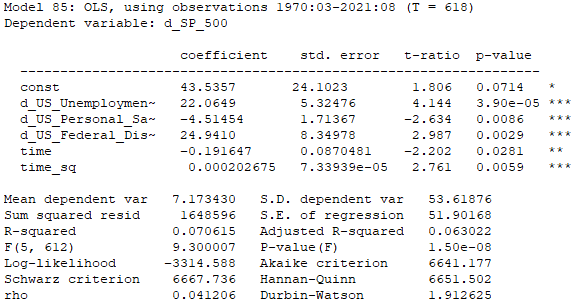
* The Durbin-Watson statistic is drastically better than it was before.
* The adjusted R^2 was way lower now, and I decided the model could be cleaned up because not all of the variables were significant.
  + I tested for linear equivalence and d\_US\_CPI, d\_US\_Consumer\_Sentiment, d\_US\_Fed\_Funds\_Rate 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:



* All the variables in the model are significant and the adjusted R^2 is slightly higher from what 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:



* 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 add back in some of the previous variable.
* 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 actually 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 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 make sense when taking in context the summary statistics of each of the variables.
    - CHECK THIS!!!!!!!!!!!!!!!!!!!
  + The formula which was modeled with the best adjusted R^2 I arrived at is:
    - logged change in S&P 500 = *β*0 ​+ *β*1 ​⋅ Time Trend + *β*2​ ⋅ US Federal Discount Rate + *β*3 ​⋅ US Real M2 Money Stock + *ε*

**CONCLUSIONS**

* Going forward
  + Need more variables with complete data in my dataset
    - I could re-explore the dataset as quarterly data.
      * Would need to reformat all of the data in excel.
* SP500 is very difficult to model and predict.
  + I think the biggest issue I had with my models was heteroscedasticity caused by omitted variable bias.
* MAYBE: SP500 moves most with drastic changes in sentiment
  + Likes certainty
  + More volatility with uncertainty