Polling and Finance: An Initial Comparison

Kevin Sanders
PY538
Boston University

In this project, certain attributes of political polling data are compared with stock pricing data. Polling data was gathered from the 2016 U.S. primaries and U.S. approval ratings. Stock data was gathered from the largest companies in the S&P500 based off market capitalization. Both sets of data follow a power law with similar exponents. Volatility clustering is found to exist for polling data. Cross correlations show differences between the two types of data. Primary polling data demonstrates negative correlations between candidates.

I. INTRODUCTION

A. Motivation

My motivation for pursuing this topic can be seen by examining the plots in FIG. 1.

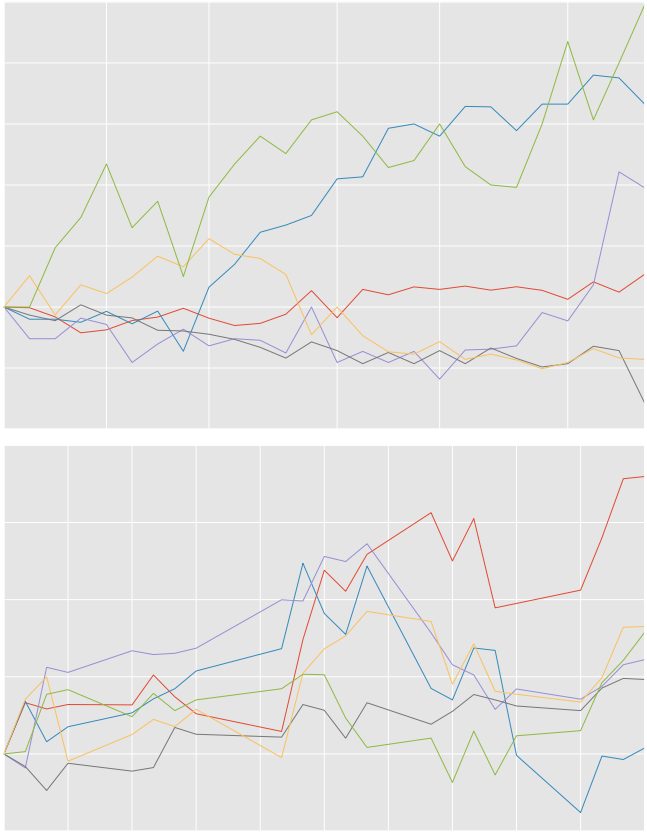

FIG. 1. Above is a simple comparison to show the similarity between the 2016 GOP primary results from September to March and the price of the top six stocks in the S&P500 for the past month, respectively.

At a first glance, the plots follow a very similar course; they both oscillate over time, and the final values appear to be distributed somewhat evenly. However, the plot of stocks contains much larger jumps in shorter periods of time and it also lacks data points for weekends.

My goal for this analysis is to examine the finer details in a comparison of these two types of time series data.

B. Limitations

Most of the limitations in this analysis surround the nature of polling data. There is very limited data in comparison to the amount of stock data that is available. The polls are also less accurate.

Because polls are taken from a sample of people, they do not necessarily represent the actual 'value' of a candidate. A solution to this could be using the actual results for each state because they would give a time series representation (at least for primary results). But each state typically has its own regional influences that separate it from national polling trends, making it difficult to connect primary results with previously gathered data.

Some other accuracy issues arise from the fact that certain polling groups tend to perform better than others. There are also varying types of polls, electronic and over the phone, each presenting their own slight bias to the types of answers given. There can be disparity between different polls as well—the percentages for a candidate will not necessarily be the same for all polls on the same day.

Because results of poll results are typically published documents, they tend not to be in electronic format, which makes the available data for analysis more limited. It is necessary to find groups that have compiled all the paper format data. Polls also tend to bunch around certain time periods (elections), which causes an inconsistent influx of data.

C. Expectations

Some basic expectations can be inferred from FIG. 1. Beyond being time series data, it is likely that both types of data exhibit some form of bias with their trend. However, the direction of bias is likely to differ. Stocks tend

to go up together, while when a certain candidate goes up in the polls, his/her opponents are likely to go down.

Because stock data returns follow a power law for deviations from zero, the polling data returns will likely also exhibit some type of power law. But the exponents must be different because the polling data does not show the same severity of spikes.

These expectations are what drove me to create the following plots for analysis.

II. DATA USED

A. Political Polling Data

The Huffington Post has a collection of political polls dating back to 2004 in its Pollster database. This data can be downloaded as CSV files for analysis. I used the 2016 GOP primary polling results as well as various approval ratings for more long term data analysis.

B. Stock Data

The daily adjusted close prices for stocks were gathered using Yahoo Finance. Returns for both types of data are defined as:

$$R_i = \frac{\log(P_i) - \log(P_{i-1})}{\sigma} \tag{1}$$

where R_i is the return for a certain time, P_i is the adjusted close price or percentage for that time, and P_{i-1} is the previous value. σ is the standard deviation for the returns.

III. DISTRIBUTION OF RETURNS

It is well known that normalized stock returns follow an inverse cubic law when binned in a cumulative distribution.

A. Short Time Frame

My first step in this analysis was to examine the normalized returns. These can be seen in FIG. 2 and FIG. 3.

As expected, the stock data has spikes above a standard deviation and is certainly not normally distributed. But the GOP primary polling data also exhibits spikes above a standard deviation, some even more drastic than the stock data.

Larger and more frequent deviations from zero indicate a potentially larger absolute exponent if a power law exists.

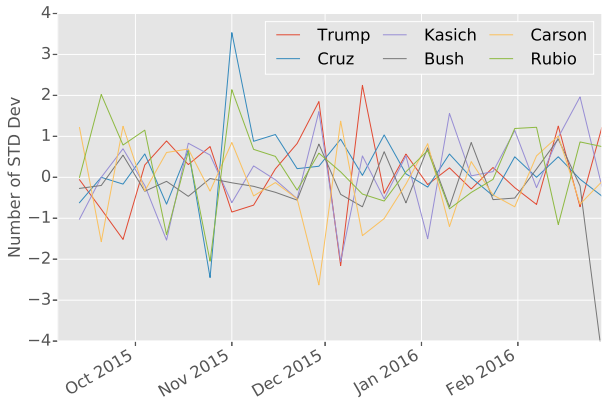

FIG. 2. Normalized weekly returns for the top 6 candidates in the 2016 GOP primary from September to March.

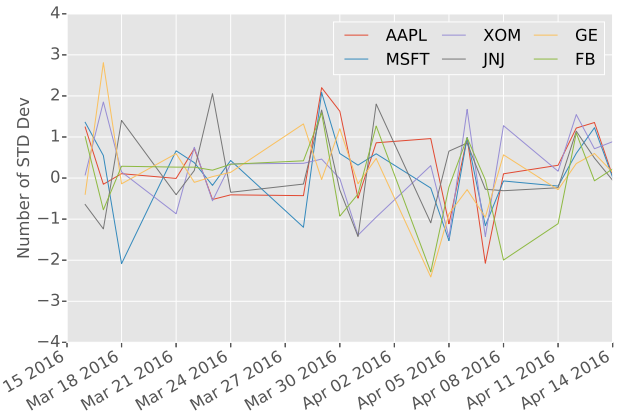

FIG. 3. Normalized daily returns for the top 6 stocks in the S&P500 for the past month.

To further investigate this hypothesis, I binned the returns in FIG. 4 and FIG. 5.

These histograms appear to have slightly different kurtoses. The stock data has the typical fat tails associated price returns, while the polling data has a fatter peak but with a rather sharp cutoff.

This would lead one to believe that these sets of data follow different laws. To confirm, the cumulative distribution of these returns were plotted.

Both sets of data resembled power laws, however the exponent varied between 1 and 2. This discrepancy with the cubic law is unexpected for the stock data. This is likely an artifact of the small/short time frame of the data set.

Something else of note, is that for the polling data, candidates with lower exponents tended to fair better in the actual voting results. Perhaps candidates should look into campaign strategies that minimize large or consistent jumps in the polls to improve long term performance.

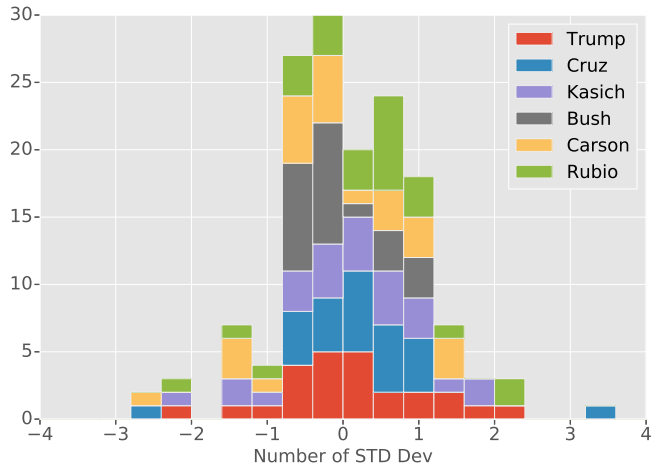

FIG. 4. Distribution of normalized weekly returns for the top 6 candidates in the 2016 GOP primary from September to March.

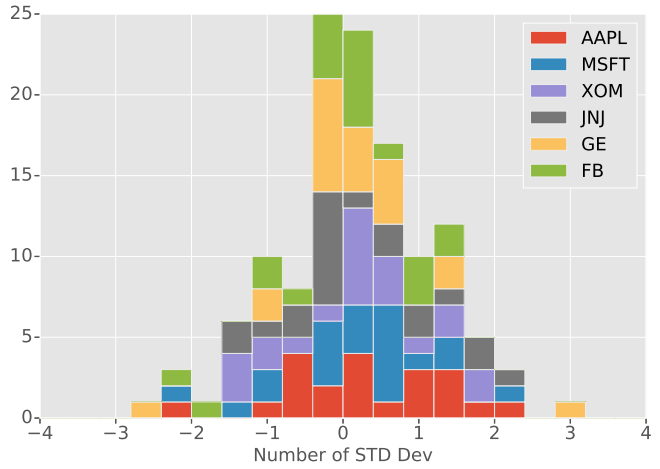

FIG. 5. Distribution of normalized daily returns for the top 6 stocks in the S&P500 for the past month.

B. Long Time Frame

In attempt to avoid small data sets while still using polling data, I performed the same analysis using the past seven years of congressional approval ratings, and compared this to the S&P500 over the same time period. The results for the cumulative distribution can be seen in FIG. 6 and FIG. 7.

In this case, the expected inverse cubic law for the stock data is realized. But surprisingly, a very similar law holds for the Congressional approval rating.

IV. TIME CORRELATION OF RETURNS

It is also well known that stock data shows signs of 'volatility clustering'. That is to say that returns have a very short correlation time period while volatility, defined as,

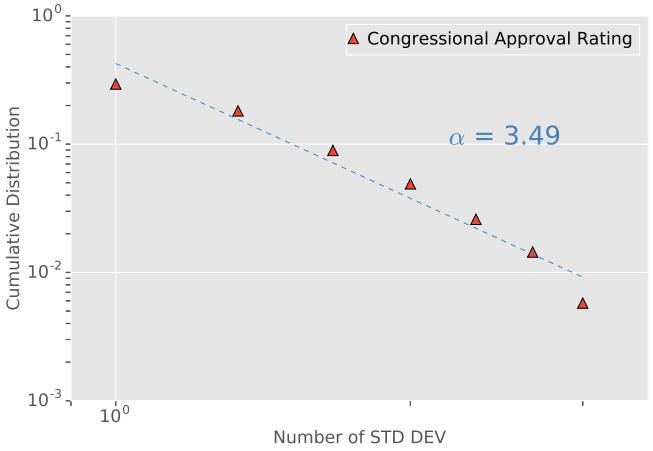

FIG. 6. Cumulative distribution of returns for Congressional approval rating polls from 2009 to 2016.

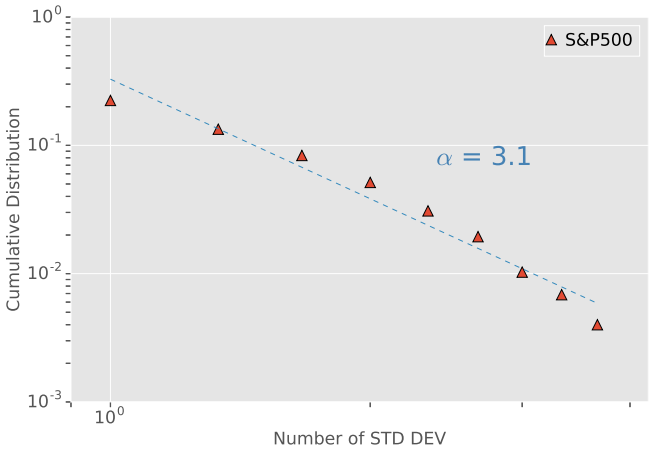

FIG. 7. Cumulative distribution of returns for S&P500 prices from 2009 to 2016.

$$V_i = R_i^2 \tag{2}$$

has a much longer correlation period. Once again to avoid the artifacts of a short data set, I use congressional approval rating and S&P500 for this analysis. The autocorrelation function represents these time correlations and can be seen in FIG. 8 and FIG. 9.

Once again the stock data and the polling data exhibit very similar properties. The approval rating for Congress demonstrates signs of volatility clustering as well.

V. CROSS CORRELATION OF RETURNS

Another comparison of interest is the cross correlation of stocks with the cross correlation of various candidates.

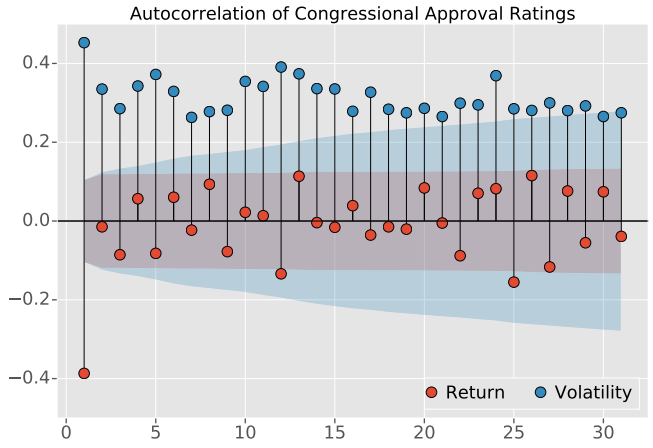

FIG. 8. Autocorrelation function of returns and volatility for Congressional approval rating polls from 2009 to 2016.

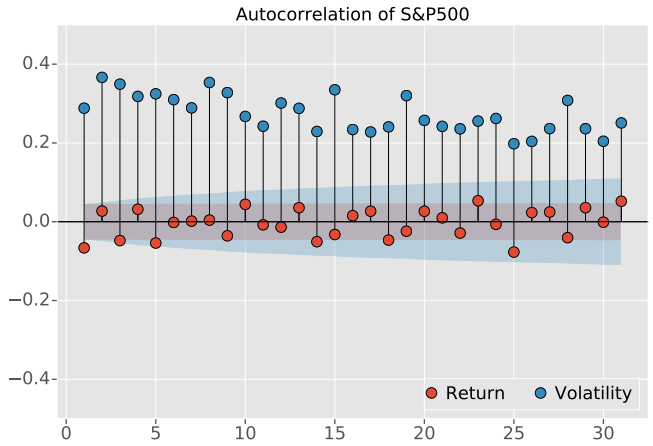

FIG. 9. Autocorrelation function of returns and volatility for S&P500 prices from 2009 to 2016.

This is where I expect to see a difference. Because the polling percentages for candidates are restricted to add up to 100%, when one candidate goes up, another must go down. This is contrary to the stock market which does not have the same kind of bound.

The cross correlation matrices for the 2016 GOP primary polling results and the top 6 stocks in the S&P500 can be seen in FIG. 10 and FIG 11.

As anticipated the candidates show substantially more negative correlation than the stock data.

VI. CONCLUSION

There is a surprising amount of similarity between the polling data and the stock data. Looking at some of the major differences—polling data not necessarily accurately describing value, having a cap at 100%, and only allowing percentage point changes—it is impressive that both sets of data show inverse cubic laws for deviations away from zero.

It is interesting to see that volatility clustering also exists for polling data. But the cap of 100% certainly has a measurable effect by examining the cross correlations that represent a major difference between the two sets of data (even though a different selection of stocks will likely show negative correlations).

More data sets and different types of comparisons are needed to further analyze the relationships between these two types of data. However, a more thorough examination should shed some light on the causes of these characteristics.

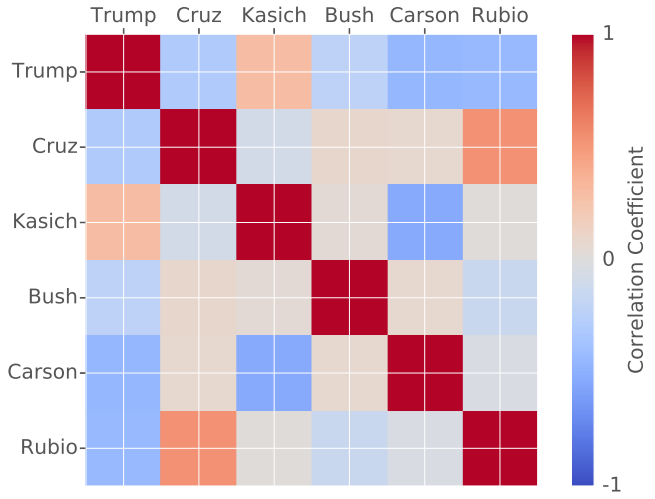

FIG. 10. Cross correlation for top 6 candidates in 2016 GOP primary from September to March.

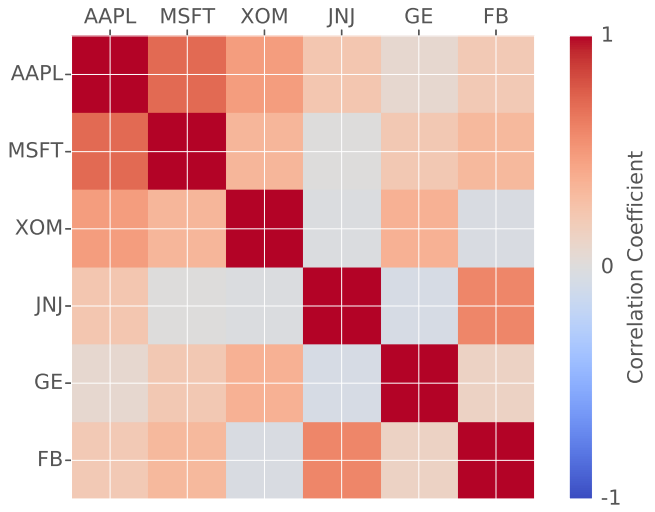

FIG. 11. Cross correlation for top 6 stocks in S&P500 for the past month.

[1] *Huffington Post Pollster*, 2016.

[2] Parameswaran Gopikrishnan, Vasiliki Plerou, Yan Liu, LA Nunes Amaral, Xavier Gabaix, and H Eugene Stanley. Scaling and correlation in financial time series. *Physica A: Statistical Mechanics and its Applications*, 287(3):362–373, 2000.